

Request from the editor:

In addition, please register any new software application in the bio.tools and SciCrunch.org databases to receive RRID (Research Resource Identification Initiative ID) and biotoolsID identifiers, and include these in your manuscript. This will facilitate tracking, reproducibility and re-use of your tool.

Regarding this request, our manuscript does not include any software application. We provide a short script to adapt Arima Hi-C data to the Juicer program (Supplementary Protocol S2), but this is not going to be updated in the future and guarantees full reproducibility and re-use as it is.

Reviewer #2:

Summary: I was impressed with the authors' tests on PCR overamplification and assembly quality. These have addressed many of my concerns with the previous manuscript, so my remaining concerns are minor.

We appreciate the reviewer's repeated assessment of our manuscript.

Line 323: The authors' tests of PCR overamplification bias have allayed many of my concerns. I still think that the interpretation of the data in this sentence could be couched in more caution. The Arima libraries had 10% more valid interaction pairs than the Icon-HI-C prep. Why was this?

We have not reached any understanding of what exactly contributed to the larger proportion of valid interaction pairs with the Arima kit, but it is possible that the most obvious characteristic of the Arima kit, namely the multiplicity of restriction enzymes, contributed to the larger proportion of valid interaction pairs. To suggest this possibility, we have inserted a sentence below in front of the sentence in question.

'The increase of restriction site combinations might have possibly contributed to the larger proportion of valid interaction pairs (Fig. 8).'

Line 435: I still believe that this paragraph is gratuitous. I would be satisfied if the authors shortened this by two sentences and made the point that Hi-C scaffolding software does not provide consistent gap lengths for gaps of unknown length.

We have deleted the second half of this paragraph to satisfy this reviewer's suggestion as below.

'The total length of assembly sequences is expected to increase after Hi-C scaffolding, because scaffolding programs simply insert a stretch of the unassigned base 'N' with a uniform length between input sequences in most cases (500 bp as a default in both 3d-dna and SALSA2). However, this has a minor impact on the total length of assembled sequences. ~~In fact, the insertion of 'N' stretches with an arbitrary length has been an implicit, rampant practice even before Hi-C scaffolding prevailed—~~

~~example, the most and second most frequent lengths of the 'N' stretch in the publicly available zebrafish genome assembly Zv10 are 100 and 10-bp, respectively.'~~

Supplementary table S8: Please provide captions that explain the difference between libraries "g" and "h" in the table as this is not immediately clear without referring to the main text.

In the previously submitted manuscript, we included a line showing the number of PCR cycles in this table (shown in red below). Also, we have inserted a guide to Figure 7A in the footnote as shown below in green.

Supplementary Table S8: Effect of variable degrees of PCR amplification

Library preparation condition	Library ID	
	g	h
Tissue type	Liver	
Restriction enzyme	Sau3AI	
Number of PCR cycles	15	11

Hi-C Pro results	
Number of input read pairs	200,000,000
Category	Proportion of valid interaction (%)
Valid interaction after removing duplicates	55.1
	70.4

See Figure 7A for the detail of the library preparation procedure. Note that 'trans' and 'cis' interactions mean contacts between scaffolds and those within scaffolds,

1 **Multifaceted Hi-C benchmarking: what makes a difference in**
2 **chromosome-scale genome scaffolding?**

3

4 Mitsutaka Kadota^{1*}, Osamu Nishimura^{1*}, Hisashi Miura², Kaori Tanaka^{1,3}, Ichiro
5 Hiratani², and Shigehiro Kuraku¹

6

7 ¹Laboratory for Phyloinformatics, RIKEN Center for Biosystems Dynamics Research
8 (BDR), Kobe, 650-0047, Japan, ²Laboratory for Developmental Epigenetics, RIKEN
9 BDR, Kobe, 650-0047, Japan, ³Present address: Division of Transcriptomics, Medical
10 Institute of Bioregulation, Kyushu University, Fukuoka, 812-0054, Japan

11

12 *These authors contributed equally to this study.

13

14 Correspondence address. Shigehiro Kuraku, Laboratory for Phyloinformatics, RIKEN
15 BDR, Japan. Tel: +81 78 306 3048; Fax: +81 78 306 3048; E-mail:
16 shigehiro.kuraku@riken.jp

17

18

19 **Abstract**

20 **Background:** Hi-C is derived from chromosome conformation capture (3C) and targets
21 chromatin contacts on a genomic scale. This method has also been used frequently in
22 scaffolding nucleotide sequences obtained by *de novo* genome sequencing and
23 assembly, in which the number of resultant sequences rarely converges to the
24 chromosome number. Despite its prevalent use, the sample preparation methods for Hi-
25 C have not been intensively discussed, especially from the standpoint of genome
26 scaffolding.

27 **Results:** To gain insight into the best practice of Hi-C scaffolding, we performed a
28 multifaceted methodological comparison using vertebrate samples and optimized
29 various factors during sample preparation, sequencing, and computation. As a result, we
30 identified several key factors that helped improve Hi-C scaffolding, including the choice
31 and preparation of tissues, library preparation conditions, the choice of restriction
32 enzyme(s), and the choice of scaffolding program and its usage.

33 **Conclusions:** This study provides the first comparison of multiple sample preparation
34 kits/protocols and computational programs for Hi-C scaffolding by an academic third
35 party. We introduce a customized protocol designated ‘inexpensive and controllable Hi-
36 C (iconHi-C) protocol’, which incorporates the optimal conditions identified in this
37 study, and demonstrated this technique on chromosome-scale genome sequences of the
38 Chinese softshell turtle *Pelodiscus sinensis*.

39

40 **Keywords:** Hi-C, genome scaffolding, chromosomes, proximity-guided assembly,
41 softshell turtle

42

43 **Background**

44 Chromatin, a complex of nucleic acids (DNA and RNA) and proteins, exhibits a
45 complex three-dimensional organization in the nucleus, which enables the intricate
46 regulation of the expression of genome information via spatio-temporal control
47 (reviewed in [1]). To characterize chromatin conformation on a genomic scale, the Hi-C
48 method was introduced as a derivative of chromosome conformation capture (3C) (Fig.
49 1A; [2]). This method detects chromatin contacts on a genomic scale via the digestion
50 of cross-linked DNA molecules with restriction enzymes, followed by proximity
51 ligation of the digested DNA molecules. Massively parallel sequencing of the library
52 containing ligated DNA molecules enables the comprehensive quantification of contacts
53 both within and between chromosomes, which is presented in a heatmap that is
54 conventionally called the ‘contact map’ [3].

55 Analyses of chromatin conformation using Hi-C have revealed more frequent
56 contacts between more closely linked genomic regions, which has recently prompted the
57 use of this method in scaffolding *de novo* genome sequences [4-6]. In *de novo* genome
58 sequencing, the number of assembled sequences is usually far larger than the number of
59 chromosomes in the karyotype of the species of interest, regardless of the sequencing
60 platform chosen [7]. The application of Hi-C scaffolding enabled a remarkable
61 enhancement of sequence continuity to reach a chromosome scale, and the integration
62 of fragmentary sequences into longer sequences, which are similar in number to that of
63 chromosomes in the karyotype.

64 In early 2018, commercial Hi-C library preparation kits were introduced (Fig.
65 1B), and *de novo* genome assembly was revolutionized by the release of versatile
66 computational programs for Hi-C scaffolding (Table 1), namely LACHESIS [4], HiRise

67 [8], SALSA [9, 10], and 3d-dna [11] (reviewed in [12]). These movements assisted the
68 rise of mass sequencing projects targeting a number of species, such as the Earth
69 BioGenome Project (EBP) [13], the Genome 10K (G10K)/Vertebrate Genome Project
70 (VGP) [14], and the DNA Zoo Project [15]. Optimization of Hi-C sample preparation,
71 however, has been limited [16], which leaves room for the improvement of efficiency
72 and the reduction of required sample quantity. Thus, the specific factors that are key for
73 Hi-C scaffolding remain unexplored, mainly because of the costly and resource-
74 demanding nature of this technology.

75 In addition to performing protocol optimization using human culture cells, we
76 focused on the softshell turtle *Pelodiscus sinensis* (Fig. 2). This species has been
77 adopted as a study system for evolutionary developmental biology (Evo-Devo),
78 including the study of the formation of the dorsal shell (carapace) (reviewed in [17]).
79 Access to genome sequences of optimal quality by relevant research communities is
80 desirable in this field. In Japan, live materials (adults and embryos) of this species are
81 available through local farms mainly between May and August, which implies its high
82 utility for sustainable research. A previous cytogenetic report revealed that the
83 karyotype of this species consists of 33 chromosome pairs including Z and W
84 chromosomes ($2n = 66$) that show a wide variety of sizes (conventionally categorized as
85 macrochromosomes and microchromosomes) [18]. Despite the moderate global GC-
86 content in its whole genome at around 44%, the intragenomic heterogeneity of GC-
87 content between and within the chromosomes has been suggested [19]. A wealth of
88 cytogenetic efforts on this species led to the accumulation of fluorescence *in situ*
89 hybridization (FISH)-based mapping data for 162 protein-coding genes covering almost
90 all chromosomes [18-22], which serve as structural landmarks for validating genome

91 assembly sequences.

92 A draft sequence assembly of the softshell turtle genome was built using short
93 reads and was released in 2013 [23]. This sequence assembly achieved the N50 scaffold
94 length of >3.3 Mb but remains fragmented into approximately 20,000 sequences (see
95 Supplementary Table S1). The longest sequence in this assembly is only slightly larger
96 than 16 Mb, which is much shorter than the largest chromosome size estimated from the
97 karyotype report [18]. The total size of the assembly is approximately 2.2 Gb, which is
98 a moderate size for a vertebrate species. Because of the affordable genome size,
99 sufficiently complex structure, and availability of validation methods, we reasoned that
100 the genome of this species is a suitable target for our methodological comparison, and
101 its improved genome assembly is expected to assist a wide range of genome-based
102 studies of this species.

103

104

105 **Results**

106

107 **Stepwise QC prior to large-scale sequencing**

108 The assessment of the quality of prepared libraries before engaging in costly sequencing
109 would be ideal. According to the literature [16, 24], we routinely control the quality of
110 Hi-C DNAs and Hi-C libraries by observing DNA size shifts via digestion targeting the
111 restriction sites in properly prepared samples (Fig. 3). More concretely, a successfully
112 ligated Hi-C DNA sample should exhibit a slight increase in the length of its restricted
113 DNA fragments after ligation (QC1), which serves as an indicator of qualified samples
114 (e.g., Sample 1 in Fig. 3B). In contrast, an unsuccessfully prepared Hi-C DNA does not

115 exhibit this length recovery (e.g., Sample 2 in Fig. 3B). In a subsequent step, DNA
116 molecules in a successfully prepared HindIII-digested Hi-C library should contain the
117 NheI restriction site at a high probability. Thus, the length distribution observed after
118 NheI digestion of the prepared library serves as an indicator of qualified or disqualified
119 products (QC2; Fig. 3C). This series of QCs is incorporated into our protocol by default
120 (Supplementary Protocol S1) and can also be performed in combination with sample
121 preparation using commercial kits if it employs a single restriction enzyme.

122 Some of the libraries prepared by us passed the QC steps performed before
123 sequencing but yielded an unfavourably large proportion of invalid read pairs. To
124 identify such libraries, we routinely performed small-scale sequencing for quick and
125 inexpensive QC (designated ‘QC3’) using the HiC-Pro program [25] (see Fig. 4 for the
126 read pair categories assigned by HiC-Pro). Our test using variable input data sizes (500
127 K to 200 M read pairs) resulted in highly similar breakdowns into different categories of
128 read pair properties (Supplementary Table S2) and guaranteed QC3 with an extremely
129 small data size of 1 M or fewer reads. These post-sequencing QC steps, which do not
130 incur a large cost, are expected to help avoid the large-scale sequencing of unsuccessful
131 libraries that have somehow passed through the QC1 and QC2 steps. Importantly,
132 libraries that have passed QC3 can be further sequenced with greater depth, as
133 necessary.

134

135 **Optimization of sample preparation conditions**

136 We identified overt differences between the sample preparation protocols of published
137 studies and those of commercial kits, especially regarding the duration of fixation and
138 enzymatic reaction as well as the library preparation method used. (Fig. 1B). Therefore,

139 we first sought to optimize the conditions of several of these steps using human culture
140 cells.

141 To evaluate the effect of the degree of cell fixation, we prepared Hi-C libraries
142 from GM12878 cells fixed for 10 and 30 minutes. Our comparison did not detect any
143 marked differences in the quality of the Hi-C DNA (QC1; Fig. 5A) and Hi-C library
144 (QC2; Fig. 5B). However, libraries that were prepared with a longer fixation time
145 exhibited a larger proportion of dangling end read pairs and religation read pairs, as well
146 as a smaller proportion of valid interaction reads (Fig. 5C). The increase in the duration
147 of cell fixation also reduced the proportion of long-range (>1 Mb) interactions among
148 the overall captured interactions (Fig. 5D).

149 The reduced preparation time of commercial Hi-C kits (up to two days
150 according to their advertisement) is attributable mainly to shortened restriction and
151 ligation times (Fig. 1B). To monitor the effect of shortening these enzymatic reactions,
152 we first analysed the progression of restriction and ligation in a time-course experiment
153 using GM12878 cells. We observed the persistent progression of restriction up to 16
154 hours and of ligation up to 6 hours (Fig. 6). To scrutinize further the possible adverse
155 effects of the prolonged reaction, Hi-C libraries of GM12878 cells were prepared with
156 variable durations of restriction digestion (1 hour and 16 hours) and ligation (15
157 minutes, 1 hour, and 6 hours). We found that the proportions of dangling end and
158 religation read pairs were reduced in cases with an extended duration of restriction
159 digestion (Supplementary Table S4). The yield of the library, which can be estimated
160 from the number of PCR cycles, increased with the extended duration of ligation
161 without any effect on the proportion of valid interaction read pairs (Supplementary
162 Table S4). The proportion of valid interaction read pairs containing the proper DpnII

163 junction sequence ‘GATCGATC’ also remained unchanged, suggesting that the
164 prolonged reaction times did not induce any adverse effects, such as star activity of the
165 restriction enzyme.

166

167 **Multifaceted comparison using softshell turtle samples**

168 Based on the detailed optimization of the sample preparation conditions described
169 above, we built an original protocol, designated the ‘iconHi-C protocol’, that included a
170 10 minute-long cell fixation, 16 hour-long restriction, 6 hour-long ligation, and
171 successive QC steps (Methods; also see Supplementary Protocol S1; Fig. 1B).

172 We performed Hi-C sample preparation and scaffolding using tissues from a
173 female Chinese softshell turtle which has both Z and W chromosomes [18]. We
174 prepared Hi-C libraries using various tissues (liver or blood cells), restriction enzymes
175 (HindIII or DpnII), and protocols (our iconHi-C protocol, the Arima kit in conjunction
176 with the KAPA Hyper Prep Kit, or the Phase kit), as outlined in Fig. 7A (see
177 Supplementary Table S5; Supplementary Fig. S1). As in some of the existing protocols
178 (e.g. [26]), we performed T4 DNA polymerase treatment in our iconHi-C protocol
179 (Library a–d), expecting reduced proportions of ‘dangling end’ read pairs that contain
180 no ligated junction, and thus do not contribute to Hi-C scaffolding. We also
181 incorporated this T4 DNA polymerase treatment into the workflow of the Arima kit
182 (Library e vs. Library f without this additional treatment). Furthermore, we tested a
183 lesser degree of PCR amplification (11 cycles) together with the use of the Phase kit
184 which recommends as many as 15 cycles by default (Library h vs. Library g; Fig. 7A).

185 All samples prepared using the iconHi-C protocol passed both controls, QC1
186 and QC2 (Fig. 7B). The prepared Hi-C libraries were sequenced to obtain one million

187 127 nt-long read pairs and were subjected to QC3 using the HiC-Pro program (Fig. 8).
188 As a result of this QC3, the largest proportion of ‘valid interaction’ pairs was observed
189 for Arima libraries (Library e and f). Regarding the iconHi-C libraries (Library a–d),
190 fewer ‘unmapped’ and ‘religation’ pairs were detected for the DpnII libraries compared
191 with HindIII libraries. It should be noted that the QC3 of the softshell turtle libraries
192 generally produced lower proportions of the ‘valid interaction’ category and larger
193 proportions of ‘unmapped pairs’ and ‘pairs with singleton’ than with the human
194 libraries. This cross-species difference may be attributable to the use of incomplete
195 genome sequences as a reference for Hi-C read mapping (Supplementary Table S1).
196 This invokes a caution when comparing QC results across species.

197

198 **Scaffolding using variable input and computational conditions**

199 In this study, only well-maintained open-source programs, i.e., 3d-dna and SALSA2,
200 were used in conjunction with variable combinations of input libraries, input read
201 amounts, input sequence cut-off lengths, and number of iterative misjoin correction
202 rounds (Fig. 9A). As a result of scaffolding, we observed a wide spectrum of basic
203 metrics, including the N50 scaffold length (0.6–303 Mb), the largest scaffold length
204 (8.7–703 Mb), and the number of chromosome-sized (>10 Mb) sequences (0–65) (Fig.
205 9; Supplementary Table S6).

206 First, using the default parameters, 3d-dna consistently produced more
207 continuous assemblies than did SALSA2 (see Assembly 1 vs. 5, 3 vs. 6, 9 vs. 10, and 11
208 vs. 12 in Fig. 9). Second, the increase in the number of iterative corrections (‘-r’ option
209 of 3d-dna) resulted in relatively large N50 lengths, but with more missing orthologues
210 (see Assembly 3 and 13–14). Third, a smaller input sequence cut-off length (‘-i’ option

211 of 3d-dna) resulted in a smaller number of scaffolds but again, with more missing
212 orthologues (see Assembly 3 and 15–17). Fourth, the use of the liver libraries
213 consistently resulted in a higher continuity than the use of the blood cell libraries (see
214 Assembly 1 vs. 2 and 3 vs. 4 in Fig. 9).

215 Assembly 8, which resulted from input Hi-C reads derived from both liver and
216 blood, exhibited an outstandingly large N50 scaffold length (303 Mb) but a larger
217 number of undetected reference orthologues (141 orthologues) than most of the other
218 assemblies. The largest scaffold (scaffold 5) in this assembly is approximately 703 Mb
219 long, causing a large N50 length, and accounts for approximately one-third of the whole
220 genome in length, as a result of possible chimeric assembly that bridged 14 putative
221 chromosomes (see Supplementary Fig. S4).

222 The choice of restriction enzymes has not been discussed in depth in the
223 context of genome scaffolding. Here, we prepared Hi-C libraries separately with HindIII
224 and DpnII. We did not mix multiple enzymes in the same reaction (other than using the
225 Arima kit which originally employs two enzymes); rather, we performed a single
226 scaffolding run with both HindIII-based and DpnII-based reads (see Assembly 7 in Fig.
227 9). As expected, our comparison of multiple metrics yielded a more successful result
228 with DpnII than with HindIII (see Assembly 1 vs. 3 as well as 2 vs. 4; Fig. 9). However,
229 the mixed input of HindIII-based and DpnII-based reads did not necessarily yield a
230 better scaffolding result (see Assembly 3 vs. 7).

231 To gain additional insight regarding the evaluation of the scaffolding results,
232 we assessed the contact maps constructed upon the Hi-C scaffolds (Supplementary Fig.
233 S5). The comparison of Assembly 3, 9 and 11, which represent the three different
234 preparation methods, revealed anomalous patterns, particularly for Assembly 11, with

235 intensive contact signals separated from the diagonal line that indicate the presence of
236 errors in the scaffolds [15]. We also performed genome-wide alignments between the
237 Hi-C scaffolds obtained. The comparison of Assembly 3, 9, and 11 revealed a high
238 similarity between Assembly 3 and 9, while Assembly 11 exhibited a significantly
239 larger number of inconsistencies against either of the other two assemblies
240 (Supplementary Fig. S6). These observations are consistent with the evaluation based
241 on sequence length and gene space completeness, which alone does not, however,
242 provide a reliable metric for the assessment of the quality of scaffolding.

243

244 **Validation of scaffolding results using transcriptome and FISH data**

245 In addition to the above-mentioned evaluation of the scaffolding results, we assessed the
246 sequence continuity using independently obtained data. First, we mapped assembled
247 transcript sequences onto our Hi-C scaffold sequences (see Methods). This did not show
248 any substantial differences between the assemblies (Supplementary Table S7), probably
249 because the sequence continuity after Hi-C scaffolding exceeded that of RNA-seq
250 library inserts, even when the length of intervening introns in the genome was
251 considered. The present analysis with RNA-seq data did not provide an effective source
252 of continuity validation.

253 Second, we referred to the fluorescence *in situ* hybridization (FISH) mapping
254 data of 162 protein-coding genes from published cytogenetic studies [18-22], which
255 allowed us to check the locations of those genes with our resultant Hi-C assemblies. In
256 this analysis, we evaluated Assembly 3, 7, and 9 (see Fig. 9A) that showed better
257 scaffolding results in terms of sequence length distribution and gene space completeness
258 (Fig. 9D). As a result, we confirmed the positioning of almost all genes and their

259 continuity over the centromeres, which encompassed not only large but also small
260 chromosomes (conventionally called ‘macrochromosomes’ and ‘microchromosomes’;
261 Fig. 10). Two genes that were not confirmed by Assembly 7 (*UCHL1* and *COX15*; Fig.
262 10) were found in separate scaffold sequences that were shorter than 1 Mb, which
263 indicates insufficient scaffolding. Conversely, the gene array including *RBM5*, *TKT*,
264 *WNT7A*, and *WNT5A*, previously shown by FISH, was consistently unconfirmed by all
265 three assemblies (Fig. 10), which did not provide any clues for among-assembly
266 evaluation or perhaps indicates an erroneous interpretation of FISH data in a previous
267 study.

268

269

270 **Discussion**

271

272 **Starting material: not genomic DNA extraction but *in situ* cell fixation**

273 In genome sequencing, best practices for high molecular weight DNA extraction have
274 often been discussed (e.g. [27]). This factor is fundamental to building longer contigs,
275 regardless of the use of short-read or long-read sequencing platforms. Moreover, the
276 proximity ligation method using Chicago libraries provided by Dovetail Genomics
277 which is based on *in vitro* chromatin reconstruction [8], uses genomic DNA as starting
278 material. In contrast, proximity-guided assembly enabled by Hi-C employs cellular
279 nuclei with preserved chromatin conformation, which brings a new technical challenge
280 regarding appropriate sampling and sample preservation in genomics.

281 In the preparation of the starting material, it is important to optimize the degree
282 of cell fixation depending on sample choice, to obtain an optimal result in Hi-C

283 scaffolding (Fig. 5). Another practical indication of tissue choice was obtained by
284 examining Assembly 8 (Fig. 9A). This assembly was produced by 3d-dna scaffolding
285 using both liver and blood libraries (Library b and d), which led to an unacceptable
286 result possibly caused by over-assembly (Fig. 9B–D; also see Results). It is likely that
287 increased cellular heterogeneity, which possibly introduces excessive conflicting
288 chromatin contacts, did not allow the scaffolding program to group and order the input
289 genome sequences properly. In brief, we recommend the use of samples with modest
290 cell-type heterogeneity that are amenable to thorough fixation.

291

292 **Considerations regarding sample preparation**

293 In this study, we did not test all commercial Hi-C kits available in the market. This was
294 partly because the Dovetail Hi-C kit specifies the non-open source program HiRise as
295 the only supported downstream computation solution and does not allow a direct
296 comparison with other kits, namely those from Phase Genomics and Arima Genomics.

297 According to our calculations, the preparation of a Hi-C library using the
298 iconHi-C protocol would be at least three times cheaper than the use of a commercial
299 kit. Practically, the cost difference would be even larger, either when the purchased kit
300 is not fully consumed or when the post-sequencing computation steps cannot be
301 undertaken in-house, which implies additional outsourcing costs.

302 The genomic regions that are targeted by Hi-C are determined by the choice of
303 restriction enzymes. Theoretically, 4-base cutters (e.g. DpnII), which potentially have
304 more frequent restriction sites on the genome, are expected to provide a higher
305 resolution than 6-base cutters (e.g., HindIII) [16]. Obviously, the use of restriction
306 enzymes that were not employed in this study might be promising in the adaptation of

307 the protocol to organisms with variable GC-content or methylation profiles. However,
308 this might not be so straightforward when considering the interspecies variation in GC-
309 content and the intra-genomic heterogeneity. The use of multiple enzymes in a single
310 reaction is a promising approach; however, from a computational viewpoint, not all
311 scaffolding programs are compatible with multiple enzymes (see Table 1 for a
312 comparison of the specification of scaffolding programs). Another technical downside
313 of this approach is the incompatibility of DNA ends restricted by multiple enzymes,
314 with restriction-based QCs, such as the QC2 step of our iconHi-C protocol (Fig. 3).
315 Therefore, in this study, DpnII and HindIII were used separately in the iconHi-C
316 protocol, which resulted in a higher scaffolding performance with the DpnII library
317 (Figs. 8 and 9), as expected. In addition, we input the separately prepared DpnII and
318 HindIII libraries together in scaffolding (Assembly 7), but this approach did not lead to
319 higher scaffolding performance (Figs. 9B–D and 10). The Arima kit employs two
320 different enzymes that can produce a much greater number of restriction site
321 combinations, because one of these two enzymes recognizes the nucleotide stretch
322 ‘GANTC’. The increase of restriction site combinations might have possibly
323 contributed to the larger proportion of valid interaction pairs (Fig. 8). Scaffolding with
324 the libraries prepared using this kit resulted in one of the most acceptable assemblies
325 (Assembly 9). However, this result did not explicitly exceed the performance of
326 scaffolding with the iconHi-C libraries, including the one that used a single enzyme
327 (DpnII; Library d).

328 Overamplification by PCR is a concern regarding the use of commercial kits
329 (with the exception of the Arima kit used with the Arima-QC2) because their manuals
330 specify the use of a certain number of PCR cycles *a priori* (15 cycles for the Phase kit

331 and 11 cycles for the Dovetail Hi-C kit) (Supplementary Table S8). In our iconHi-C
332 protocol, an optimal number of PCR cycles is estimated by means of a preliminary real-
333 time PCR using a small aliquot (Step 11.25 to 11.29 in Supplementary Protocol S1), as
334 done traditionally for other library types (e.g., [28]). This procedure allowed us to
335 reduce the number of PCR cycles, down to as few as five cycles (Supplementary Table
336 S5). The Dovetail Hi-C kit recommends the use of larger amounts of kit components
337 than that specified for a single sample, depending on the genome size, as well as the
338 degree of genomic heterozygosity and repetitiveness, of the species of interest. In
339 contrast, with our iconHi-C protocol, we always prepared a single library, regardless of
340 those species-specific factors, which seemed to suffice in all the cases tested.

341 Commercial Hi-C kits, which usually advertise easiness and quickness of use,
342 have largely shortened the protocol down to two days, compared with the published
343 non-commercial protocols (e.g., [16, 26]). Such time-saving protocols are achieved
344 mainly by shortening the duration of restriction enzyme digestion and ligation (Fig. 1B).
345 Our assessment, however, revealed unsaturated reaction within the shortened time
346 frames employed in the commercial kits (Fig. 6), which was accompanied by an
347 unfavorable composition of read pairs (Supplementary Table S4). Our attempt to insert
348 a step of T4 DNA polymerase treatment in the sample preparation of the Arima kit
349 protocol resulted in reduced ‘dangling end’ reads (Library e vs. f in Fig. 8). Regarding
350 the Phase kit, transposase-based library preparation contributes largely to its shortened
351 protocol, but this does not allow flexible control of library insert lengths. Recent
352 protocols (versions 1.5 and 2.0) of the Phase kit instruct users to employ a largely
353 reduced DNA amount in the tagmentation reaction, which should mitigate the difficulty
354 in controlling insert length but require excessive PCR amplification. The Arima and

355 Phase kits assume that the quality control of Hi-C DNA is based on the yield, and not
356 the size, of DNA (see Fig. 1B). Nevertheless, quality control based on DNA size
357 (equivalent to QC1 in iconHi-C) is feasible by taking aliquots at each step of sample
358 preparation. In particular, if preparing a small number of samples for Hi-C, as practised
359 typically for genome scaffolding, one should opt to consider these points, even when
360 using commercial kits, to improve the quality of the prepared libraries and scaffolding
361 products.

362

363 **Considerations regarding sequencing**

364 The quantity of Hi-C read pairs to be input for scaffolding is critical because it accounts
365 for the majority of the cost of Hi-C scaffolding. Our protocol introduces a thorough
366 safety system to prevent sequencing unsuccessful libraries, first by performing pre-
367 sequencing QCs for size shift analyses (Fig. 3) and second via small-scale (down to 500
368 K read pairs) sequencing (see Results; also see Supplementary Tables S2 and S9).

369 Our comparison showed a dramatic decrease in assembly quality in cases in
370 which <100 M read pairs were used (see the comparison of Assembly 18–22 described
371 above; Fig. 9; also see [29]). Nevertheless, we obtained optimal results with a smaller
372 number of reads (ca. 160 M per 2.2 Gb of genome) than that recommended by the
373 manufacturers of commercial kits (e.g., 100 M per 1 Gb of genome for the Dovetail Hi-
374 C kit and 200 M per Gb of genome for the Arima kit). As generally and repeatedly
375 discussed [29][29], the proportion of informative reads and their diversity, rather than
376 just the overall number of obtained reads, is critical.

377 In terms of read length, we did not perform any comparisons in this study.

378 Longer reads may enhance the fidelity of the characterization of the read pair properties

379 and allow precise QC. Nevertheless, the existing Illumina sequencing platform has
380 enabled the less expensive acquisition of 150 nt-long paired-end reads, which did not
381 prompt us to vary the read length.

382

383 **Considerations regarding computation**

384 In this study, 3d-dna produced a more reliable scaffolding output than did SALSA2,
385 whether sample preparation employed a single or multiple enzyme(s) (Fig. 9B–D). On
386 the other hand, 3d-dna required a greater amount of time for the completion of
387 scaffolding than did SALSA2. Apart from the choice of program, several points should
388 be considered if successful scaffolding for a smaller investment is to be achieved. In
389 general, Hi-C scaffolding results should not be taken for granted, and it is necessary to
390 improve them by referring to contact maps using an interactive tool, such as Juicebox
391 [15]. In this study, however, we compared raw scaffolding output to evaluate sample
392 preparation and reproducible computational steps.

393 We used various parameters of the scaffolding programs (Fig. 9A). First, the
394 Hi-C scaffolding programs that are available currently have different default length cut-
395 off values for input sequences (e.g., 15000 bp for the ‘-i’ parameter in 3d-dna and 1000
396 bp for the ‘-c’ parameter in SALSA2). Only sequences that are longer than the cut-off
397 length value contribute to sequence scaffolding towards chromosome sizes, while
398 sequences shorter than the cut-off length are implicitly excluded from the scaffolding
399 process and remain unchanged. Typically, when using the Illumina sequencing
400 platform, genomic regions with unusually high frequencies of repetitive elements and
401 GC-content are not assembled into sequences with a sufficient length (see [30]). Such
402 genomic regions tend to be excluded from chromosome-scale Hi-C scaffolds because

403 their length is smaller than the threshold. Alternatively, these regions may be excluded
404 because few Hi-C read pairs are mapped to them, even if they exceed the cut-off length.
405 The deliberate setting of a cut-off length is recommended if particular sequences with
406 relatively small lengths are the target of scaffolding. It should be noted that lowering the
407 length threshold can result in frequent misjoins in the scaffolding output (Fig. 9B–D) or
408 in overly long computational times. Regarding the number of iterative misjoin
409 correction rounds (the ‘-r’ parameter in 3d-dna and ‘i’ parameter in SALSA2), our
410 attempts of using increased values did not necessarily yield favourable results (Fig. 9B–
411 D). This did not provide a consistent optimal range of values but rather suggests the
412 importance of performing multiple scaffolding runs with varying parameters.

413

414 **Considerations regarding the assessment of chromosome-scale genome sequences**

415 Our assessment using cytogenetic data confirmed the continuity of gene linkage over
416 the obtained chromosome-scale sequences (Fig. 10). This validation was required by the
417 almost saturated scores of typical gene space completeness assessment tools such as
418 BUSCO (Supplementary Table S6) and by transcript contig mapping (Supplementary
419 Table S7), neither of which provided an effective metric for evaluation.

420 For further evaluation of our scaffolding results, we referred to the sequence
421 length distributions of the genome assemblies of other turtle species that are regarded as
422 being chromosome-scale data. This analysis yielded values of the basic metrics that
423 were comparable to those of our Hi-C scaffolds of the softshell turtle, i.e. an N50 length
424 of 127.5 Mb and a maximum sequence length of 344.5 Mb for the genome assembly of
425 the green sea turtle (*Chelonia mydas*) released by the DNA Zoo Project [15] and an N50
426 length of 131.6 Mb and a maximum length of 370.3 Mb for the genome assembly of the

427 Goode's thornscrub tortoise (*Gopherus evgoodei*) released by the Vertebrate Genome
428 Project (VGP) [14]. Scaffolding results should be evaluated by referring to the
429 estimated N50 length and the maximum length based on the actual value and to the
430 length distribution of chromosomes in the intrinsic karyotype of the species in question,
431 or of its close relative. Turtles tend to have an N50 length of approximately 130 Mb and
432 a maximum length of 350 Mb, while many teleost fish genomes exhibit an N50 length
433 as low as 20–30 Mb and a maximum length of <100 Mb [31]. If these values are
434 excessive, the scaffolded sequences harbour overassembly, which erroneously boosts
435 length-based metrics. Thus, higher values, which are conventionally regarded as signs
436 of successful sequence assembly, do not necessarily indicate higher precision.

437 The total length of assembly sequences is expected to increase after Hi-C
438 scaffolding, because scaffolding programs simply insert a stretch of the unassigned base
439 'N' with a uniform length between input sequences in most cases (500 bp as a default in
440 both 3d-dna and SALSA2). However, this has a minor impact on the total length of
441 assembled sequences. ~~In fact, the insertion of 'N' stretches with an arbitrary length has~~
442 ~~been an implicit, rampant practice even before Hi-C scaffolding prevailed—for~~
443 ~~example, the most and second most frequent lengths of the 'N' stretch in the publicly~~
444 ~~available zebrafish genome assembly Zv10 are 100 and 10 bp, respectively.~~

445

446 **Conclusions**

447 In this study, we introduced the iconHi-C protocol which implements successive QC
448 steps. We also assessed potential key factors for improving Hi-C scaffolding. Overall,
449 our study showed that small variations in sample preparation or computation for
450 scaffolding can have a large impact on scaffolding output, and that any scaffolding

451 output should ideally be validated using independent information, such as cytogenetic
452 data, long reads, or genetic linkage maps. The present study aimed to evaluate the
453 output of reproducible computational steps, which in practice should be followed by the
454 modification of the raw scaffolding output by referring to independent information or
455 by analysing chromatin contact maps. The study employed limited combinations of
456 species, sample prep methods, scaffolding programs, and its parameters, and we will
457 continue to test different conditions for kits/programs that did not necessarily perform
458 well here using our specific materials.

459

460 **Methods**

461

462 **Initial genome assembly sequences**

463 The softshell turtle (*Pelodiscus sinensis*) assembly published previously [23] was
464 downloaded from NCBI GenBank (GCA_000230535.1), whose gene space
465 completeness and length statistics were assessed by gVolante [32] (see Supplementary
466 Table S1 for the assessment results). Although it could be suggested to remove
467 haplotigs before Hi-C scaffolding [33], we omitted this step because of the low
468 frequency of the reference orthologues with multiple copies (0.72%; Supplementary
469 Table S1), indicating a minimal degree of haplotig contamination.

470

471 **Animals and cells**

472 We sampled tissues (liver and blood cells) from a female purchased from a local farmer
473 in Japan, because the previous whole genome sequencing used the whole blood of a
474 female [23]. All experiments were conducted in accordance with the Guideline of the

475 Institutional Animal Care and Use Committee of RIKEN Kobe Branch (Approval ID:
476 A2017-12).

477 The human lymphoblastoid cell line GM12878 was purchased from the Coriell
478 Cell Repositories and cultured in RPMI-1640 medium (Thermo Fisher Scientific)
479 supplemented with 15% FBS, 2 mM L-glutamine, and a 1× antibiotic-antimycotic
480 solution (Thermo Fisher Scientific), at 37 °C, 5% CO₂, as described previously [34].

481

482 **Hi-C sample preparation using the original protocol**

483 We have made modifications to the protocols that are available in the literature [3, 26,
484 35] (Fig. 1B). The full version of our ‘inexpensive and controllable Hi-C (iconHi-C)’
485 protocol is described in Supplementary Protocol S1 and available at Protocols.io
486 (<https://www.protocols.io/private/950FFCBDE7C46D1598CA7DDFE7441C9F>).

487

488 **Hi-C sample preparation using commercial kits**

489 The Proximo Hi-C kit (Phase Genomics) which employs the restriction enzyme Sau3A1
490 and transposase-based library preparation [36] (Fig. 1B) was used to prepare a library
491 from 50 mg of the softshell turtle liver according to the official ver. 1.0 animal protocol
492 provided by the manufacturer (Library g in Fig. 7A) and a library from 10 mg of the
493 liver that was amplified with a reduced number of PCR cycles based on a preliminary
494 real-time qPCR using an aliquot (Library h; see [28] for the details of the pre-
495 determination of the optimal number of PCR cycles). The Arima-HiC kit (Arima
496 Genomics), which employs a restriction enzyme cocktail (Fig. 1B), was used in
497 conjunction with the KAPA Hyper Prep Kit (KAPA Biosystems), protocol ver.
498 A160108 v00, to prepare a library using the softshell turtle liver, according to its official

499 animal vertebrate tissue protocol (ver. A160107 v00) (Library f) and a library with an
500 additional step of T4 DNA polymerase treatment for reducing ‘dangling end’ reads
501 (Library e). This additional treatment is detailed in Step 8.2 (for DpnII-digested
502 samples) of Supplementary Protocol S1.

503

504 **DNA sequencing**

505 Small-scale sequencing for library QC (QC3) was performed in-house to obtain 127 nt-
506 long paired-end reads on an Illumina HiSeq 1500 in the Rapid Run Mode. For
507 evaluating the effects of variable duration of the restriction digestion and ligation
508 reactions, sequencing was performed on an Illumina MiSeq using the MiSeq Reagent
509 Kit v3 to obtain 300 nt-long paired-end reads. Large-scale sequencing for Hi-C
510 scaffolding was performed to obtain 151 nt-long paired-end reads on an Illumina HiSeq
511 X. The obtained reads underwent quality control using FastQC ver. 0.11.5
512 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), and low-quality regions
513 and adapter sequences in the reads were removed using Trim Galore ver. 0.4.5
514 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with the parameters
515 ‘-e 0.1 -q 30’.

516

517 **Post-sequencing quality control (QC3) of Hi-C libraries**

518 For post-sequencing library QC, one million trimmed read pairs for each Hi-C library
519 were sampled using the ‘subseq’ function of the program seqtk ver. 1.2-r94
520 (<https://github.com/lh3/seqtk>). The resultant sets of read pairs were processed using
521 HiC-Pro ver. 2.11.1 [25] with bowtie2 ver. 2.3.4.1 [37] to evaluate the insert structure
522 and mapping status onto the softshell turtle genome assembly PelSin_1.0

523 (GCF_000230535.1) or the human genome assembly hg19. This resulted in
524 categorization as valid interaction pairs and invalid pairs, with the latter being divided
525 further into ‘dangling end’, ‘religation’, ‘self circle’, and ‘single-end’ pairs (Fig. 4). To
526 process the read pairs derived from the libraries prepared using either HindIII or DpnII
527 (Sau3AI) with the iconHi-C protocol (Library a–d) and the Phase kit (Library g and h),
528 the restriction fragment file required by HiC-Pro was prepared according to the script
529 ‘digest_genome.py’ of HiC-Pro. To process the reads derived from the Arima kit
530 (Library e and f), all restriction sites (‘GATC’ and ‘GANTC’) were inserted into the
531 script. In addition, the nucleotide sequences of all possible ligated sites generated by
532 restriction enzymes were included in a configuration file of HiC-Pro. The details of this
533 procedure and the sample code used are included in Supplementary Protocol S2.

534

535 **Computation for Hi-C scaffolding**

536 To control our comparison with intended input data sizes, a certain number of trimmed
537 read pairs were sampled for each library with seqtk, as described above. Scaffolding
538 was processed with the following methods employing two program pipelines, 3d-dna
539 and SALSA2.

540 Scaffolding via 3d-dna was performed using Hi-C read mapping onto the
541 genome with Juicer ver. 20180805 [38] using the default parameters with BWA
542 ver.0.7.17-r1188 [39]. The restriction fragment file required by Juicer was prepared by
543 the script ‘generate_site_positions.py’ script of Juicer. By converting the restriction
544 fragment file of HiC-Pro to the Juicer format, an original script that was compatible
545 with multiple restriction enzymes was prepared (Supplementary Protocol S2).
546 Scaffolding via 3d-dna ver. 20180929 was performed using variable parameters (see

547 Fig. 9A).

548 Scaffolding via SALSA2 using Hi-C reads was preceded by Hi-C read pair
549 processing with the Arima mapping pipeline ver. 20181207
550 (https://github.com/ArimaGenomics/mapping_pipeline) together with BWA, SAMtools
551 ver. 1.8-21-gf6f50ac [40], and Picard ver. 2.18.12
552 (<https://github.com/broadinstitute/picard>). The mapping result in the binary alignment
553 map (bam) format was converted into a BED file by bamToBed of Bedtools ver. 2.26.0
554 [41], the output of which was used as the input of scaffolding using SALSA2 ver.
555 20181212 with the default parameters.

556

557 **Completeness assessment of Hi-C scaffolds**

558 gVolante ver. 1.2.1 [32] was used to perform an assessment of the sequence length
559 distribution and gene space completeness based on the coverage of one-to-one reference
560 orthologues with BUSCO v2/v3 employing the one-to-one orthologue set ‘Tetrapoda’
561 supplied with BUSCO [42]. No cut-off length was used in this assessment.

562

563 **Continuity assessment using RNA-seq read mapping**

564 Paired-end reads obtained by RNA-seq of softshell turtle embryos at multiple stages
565 were downloaded from NCBI SRA (DRX001576) and were assembled using Trinity
566 ver. 2.7.0 [43] with default parameters. The assembled transcript sequences were
567 mapped to the Hi-C scaffold sequences with pblat [44], and the output was assessed
568 with isoblat ver. 0.31 [45].

569

570 **Comparison with chromosome FISH results**

571 Cytogenetic validation of Hi-C scaffolding results was performed by comparing the
572 gene locations on the scaffold sequences with those provided by previous chromosome
573 FISH for 162 protein-coding genes [18-22]. The nucleotide exonic sequences for those
574 162 genes were retrieved from GenBank and aligned with Hi-C scaffold sequences
575 using BLAT ver. 36x2 [46], followed by the analysis of their positions and orientation
576 along the Hi-C scaffold sequences.

577

578 **Availability of supporting data**

579 All sequence data generated in this study have been submitted to the DDBJ Sequence
580 Read Archive (DRA) under accession IDs DRA008313 and DRA008947. The datasets
581 supporting the results of this article are available in FigShare
582 (<https://figshare.com/s/6ea495a65fc231a74458>).

583

584 **Additional files**

585 Supplementary Figure S1. DNA size distribution of the softshell turtle Hi-C libraries.

586

587 Supplementary Figure S2. Pre-sequencing quality control of softshell turtle blood Hi-C
588 libraries (Library a and b).

589

590 Supplementary Figure S3. Pre-sequencing quality control (QC2) of the Hi-C libraries
591 generated using the Phase kit (Library g and h).

592

593 Supplementary Figure S4. Structural analysis of the possibly chimeric scaffold in
594 Assembly 8.

595

596 Supplementary Figure S5. Hi-C contact maps for selected softshell turtle Hi-C

597 scaffolds.

598

599 Supplementary Figure S6. Pairwise alignment of Hi-C scaffolds.

600

601 Supplementary Table S1. Statistics of the Chinese softshell turtle draft genome

602 assembly before Hi-C.

603

604 Supplementary Table S2. HiC-Pro results for the human GM12878 HindIII Hi-C library

605 with reduced reads.

606

607 Supplementary Table S3. Quality control of the human GM12878 Hi-C libraries.

608

609 Supplementary Table S4. Effect of the duration of restriction enzyme digestion and

610 ligation.

611

612 Supplementary Table S5. Quality control of Hi-C libraries.

613

614 Supplementary Table S6. Scaffolding results with variable input data and computational

615 parameters.

616

617 Supplementary Table S7. Mapping results of assembled transcript sequences onto Hi-C

618 scaffolds.

619

620 Supplementary Table S8. Effect of variable degrees of PCR amplification.

621

622 Supplementary Table S9. HiC-Pro results for the softshell turtle liver libraries (Library
623 d, e, and h) with reduced reads.

624

625 Supplementary Protocol S1. iconHi-C protocol.

626

627 Supplementary Protocol S2. Computational protocol to support the use of multiple
628 enzymes.

629

630

631

632 **Abbreviations**

633 PCR: polymerase chain reaction; FISH, fluorescence *in situ* hybridization; BUSCO,
634 benchmarking universal single-copy orthologs; NCBI, National Center for
635 Biotechnology Information; NGS, next generation DNA sequencing.

636

637 **Funding**

638 This work was supported by intramural grants within RIKEN to S.K. and I.H. and by a
639 Grant-in-Aid for Scientific Research on Innovative Areas from the Ministry of
640 Education, Culture, Sports, Science, and Technology (MEXT) to I.H. (18H05530).

641

642 **Competing interests**

643 The authors declare that they have no competing interests.

644

645 **Acknowledgements**

646 The authors acknowledge Naoki Irie, Juan Pascual Anaya and Tatsuya Hirasawa in
647 Laboratory for Evolutionary Morphology, RIKEN BDR for suggestions for sampling,
648 Rawin Poonperm for comments and discussion on the iconHi-C protocol, Olga
649 Dudchenko, Erez Lieberman-Aiden, Arang Rhie, Sergey Koren, and Jay Ghurye for
650 their technical suggestions for sample preparation and computation, Yoshinobu Uno for
651 guidance in the cytogenetic data interpretation, and Anthony Schmitt of Arima
652 Genomics and Stephen Eacker of Phase Genomics for providing information about the
653 Hi-C kits. The authors also thank the other members of the Laboratory for
654 Phyloinformatics and Laboratory for Developmental Epigenetics in RIKEN BDR for
655 technical support and discussion.

656

657 **Author contributions**

658 S.K., I.H., H.M., and M.K. conceived the study. M.K. and K.T. performed laboratory
659 works, and O.N. performed bioinformatic analysis. M.K., O.N., and H.M. analyzed the
660 data. S.K., M.K., and O.N. drafted the manuscript. All authors contributed to the
661 finalization of the manuscript.

662

663 **References**

- 664 1. Rowley MJ and Corces VG. Organizational principles of 3D genome architecture.
665 Nat Rev Genet. 2018;19 12:789-800. doi:10.1038/s41576-018-0060-8.

- 666 2. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling
667 A, et al. Comprehensive mapping of long-range interactions reveals folding
668 principles of the human genome. *Science*. 2009;326 5950:289-93.
669 doi:10.1126/science.1181369.
- 670 3. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et
671 al. A 3D map of the human genome at kilobase resolution reveals principles of
672 chromatin looping. *Cell*. 2014;159 7:1665-80. doi:10.1016/j.cell.2014.11.021.
- 673 4. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO and Shendure J.
674 Chromosome-scale scaffolding of de novo genome assemblies based on chromatin
675 interactions. *Nat Biotechnol*. 2013;31 12:1119-25. doi:10.1038/nbt.2727.
- 676 5. Marie-Nelly H, Marbouty M, Cournac A, Flot JF, Liti G, Parodi DP, et al. High-
677 quality genome (re)assembly using chromosomal contact data. *Nat Commun*.
678 2014;5 1:5695. doi:10.1038/ncomms6695.
- 679 6. Kaplan N and Dekker J. High-throughput genome scaffolding from in vivo DNA
680 interaction frequency. *Nat Biotechnol*. 2013;31 12:1143-7. doi:10.1038/nbt.2768.
- 681 7. Sedlazeck FJ, Lee H, Darby CA and Schatz MC. Piercing the dark matter:
682 bioinformatics of long-range sequencing and mapping. *Nat Rev Genet*. 2018;19
683 6:329-46. doi:10.1038/s41576-018-0003-4.

- 684 8. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al.
685 Chromosome-scale shotgun assembly using an in vitro method for long-range
686 linkage. *Genome Res.* 2016;26 3:342-50. doi:10.1101/gr.193474.115.
- 687 9. Ghurye J, Pop M, Koren S, Bickhart D and Chin CS. Scaffolding of long read
688 assemblies using long range contact information. *BMC Genomics.* 2017;18 1:527.
689 doi:10.1186/s12864-017-3879-z.
- 690 10. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-
691 C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol.*
692 2019;15 8:e1007273. doi:10.1371/journal.pcbi.1007273.
- 693 11. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De
694 novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length
695 scaffolds. *Science.* 2017;356 6333:92-5. doi:10.1126/science.aal3327.
- 696 12. Ghurye J and Pop M. Modern technologies and algorithms for scaffolding
697 assembled genomes. *PLoS Comput Biol.* 2019;15 6:e1006994.
698 doi:10.1371/journal.pcbi.1006994.
- 699 13. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al.
700 Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci*
701 USA. 2018;115 17:4325-33. doi:10.1073/pnas.1720115115.

- 702 14. Koepfli KP, Paten B, Genome KCoS and O'Brien SJ. The Genome 10K Project: a
703 way forward. *Annu Rev Anim Biosci.* 2015;3:57-111. doi:10.1146/annurev-animal-
704 090414-014900.
- 705 15. Dudchenko O, Shamim MS, Batra SS, Durand NC, Musial NT, Mostofa R, et al.
706 The Juicebox Assembly Tools module facilitates de novo assembly of mammalian
707 genomes with chromosome-length scaffolds for under \$1000. *bioRxiv.*
708 2018:254797. doi:10.1101/254797.
- 709 16. Belaghzal H, Dekker J and Gibcus JH. Hi-C 2.0: An optimized Hi-C procedure for
710 high-resolution genome-wide mapping of chromosome conformation. *Methods.*
711 2017;123:56-65. doi:10.1016/j.ymeth.2017.04.004.
- 712 17. Kuratani S, Kuraku S and Nagashima H. Evolutionary developmental perspective
713 for the origin of turtles: the folding theory for the shell based on the developmental
714 nature of the carapacial ridge. *Evol Dev.* 2011;13 1:1-14. doi:10.1111/j.1525-
715 142X.2010.00451.x.
- 716 18. Matsuda Y, Nishida-Umehara C, Tarui H, Kuroiwa A, Yamada K, Isobe T, et al.
717 Highly conserved linkage homology between birds and turtles: bird and turtle
718 chromosomes are precise counterparts of each other. *Chromosome Res.* 2005;13
719 6:601-15. doi:10.1007/s10577-005-0986-5.

- 720 19. Kuraku S, Ishijima J, Nishida-Umehara C, Agata K, Kuratani S and Matsuda Y.
721 cDNA-based gene mapping and GC3 profiling in the soft-shelled turtle suggest a
722 chromosomal size-dependent GC bias shared by sauropsids. *Chromosome Res.*
723 2006;14 2:187-202. doi:10.1007/s10577-006-1035-8.
- 724 20. Uno Y, Nishida C, Tarui H, Ishishita S, Takagi C, Nishimura O, et al. Inference of
725 the protokaryotypes of amniotes and tetrapods and the evolutionary processes of
726 microchromosomes from comparative gene mapping. *PLoS One.* 2012;7
727 12:e53027. doi:10.1371/journal.pone.0053027.
- 728 21. Kawai A, Nishida-Umehara C, Ishijima J, Tsuda Y, Ota H and Matsuda Y.
729 Different origins of bird and reptile sex chromosomes inferred from comparative
730 mapping of chicken Z-linked genes. *Cytogenet Genome Res.* 2007;117 1-4:92-102.
731 doi:10.1159/000103169.
- 732 22. Kawagoshi T, Uno Y, Matsubara K, Matsuda Y and Nishida C. The ZW micro-sex
733 chromosomes of the Chinese soft-shelled turtle (*Pelodiscus sinensis*, Trionychidae,
734 Testudines) have the same origin as chicken chromosome 15. *Cytogenet Genome*
735 *Res.* 2009;125 2:125-31. doi:10.1159/000227837.
- 736 23. Wang Z, Pascual-Anaya J, Zadissa A, Li W, Niimura Y, Huang Z, et al. The draft
737 genomes of soft-shell turtle and green sea turtle yield insights into the development

- 738 and evolution of the turtle-specific body plan. *Nat Genet.* 2013;45 6:701-6.
739 doi:10.1038/ng.2615.
- 740 24. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y and Dekker J. Hi-C: a
741 comprehensive technique to capture the conformation of genomes. *Methods.*
742 2012;58 3:268-76. doi:10.1016/j.ymeth.2012.05.001.
- 743 25. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an
744 optimized and flexible pipeline for Hi-C data processing. *Genome Biol.*
745 2015;16:259. doi:10.1186/s13059-015-0831-x.
- 746 26. Sofueva S, Yaffe E, Chan WC, Georgopoulou D, Vietri Rudan M, Mira-Bontenbal
747 H, et al. Cohesin-mediated interactions organize chromosomal domain architecture.
748 *Embo j.* 2013;32 24:3119-29. doi:10.1038/emboj.2013.237.
- 749 27. Mayjonade B, Gouzy J, Donnadieu C, Pouilly N, Marande W, Callot C, et al.
750 Extraction of high-molecular-weight genomic DNA for long-read sequencing of
751 single molecules. *Biotechniques.* 2016;61 4:203-5. doi:10.2144/000114460.
- 752 28. Tanegashima C, Nishimura O, Motone F, Tatsumi K, Kadota M and Kuraku S.
753 Embryonic transcriptome sequencing of the ocellate spot skate *Okamejei kenojei*.
754 *Sci Data.* 2018;5:180200. doi:10.1038/sdata.2018.200.
- 755 29. DeMaere MZ and Darling AE. bin3C: exploiting Hi-C sequencing data to

756 accurately resolve metagenome-assembled genomes. *Genome Biol.* 2019;20 1:46.
757 doi:10.1186/s13059-019-1643-1.

758 30. Botero-Castro F, Figuet E, Tilak MK, Nabholz B and Galtier N. Avian Genomes
759 Revisited: Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds.
760 *Mol Biol Evol.* 2017;34 12:3123-31. doi:10.1093/molbev/msx236.

761 31. Hotaling S and Kelley JL. The rising tide of high-quality genomic resources. *Mol*
762 *Ecol Resour.* 2019;19 3:567-9. doi:10.1111/1755-0998.12964.

763 32. Nishimura O, Hara Y and Kuraku S. gVolante for standardizing completeness
764 assessment of genome and transcriptome assemblies. *Bioinformatics.* 2017;33
765 22:3635-7. doi:10.1093/bioinformatics/btx445.

766 33. Roach MJ, Schmidt SA and Borneman AR. Purge Haplotigs: allelic contig
767 reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics.*
768 2018;19 1:460. doi:10.1186/s12859-018-2485-7.

769 34. Kadota M, Hara Y, Tanaka K, Takagi W, Tanegashima C, Nishimura O, et al.
770 CTCF binding landscape in jawless fish with reference to Hox cluster evolution.
771 *Sci Rep.* 2017;7 1:4957. doi:10.1038/s41598-017-04506-x.

772 35. Ikeda T, Hikichi T, Miura H, Shibata H, Mitsunaga K, Yamada Y, et al. Srf
773 destabilizes cellular identity by suppressing cell-type-specific gene expression

774 programs. *Nat Commun.* 2018;9 1:1387. doi:10.1038/s41467-018-03748-1.

775 36. Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, et al. Rapid, low-
776 input, low-bias construction of shotgun fragment libraries by high-density in vitro
777 transposition. *Genome Biol.* 2010;11 12:R119. doi:10.1186/gb-2010-11-12-r119.

778 37. Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat*
779 *Methods.* 2012;9 4:357-9. doi:10.1038/nmeth.1923.

780 38. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, et al. Juicer
781 Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments.
782 *Cell Syst.* 2016;3 1:95-8. doi:10.1016/j.cels.2016.07.002.

783 39. Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler
784 transform. *Bioinformatics.* 2009;25 14:1754-60.
785 doi:10.1093/bioinformatics/btp324.

786 40. Li H. A statistical framework for SNP calling, mutation discovery, association
787 mapping and population genetical parameter estimation from sequencing data.
788 *Bioinformatics.* 2011;27 21:2987-93. doi:10.1093/bioinformatics/btr509.

789 41. Quinlan AR and Hall IM. BEDTools: a flexible suite of utilities for comparing
790 genomic features. *Bioinformatics.* 2010;26 6:841-2.
791 doi:10.1093/bioinformatics/btq033.

792 42. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM.
793 BUSCO: assessing genome assembly and annotation completeness with single-
794 copy orthologs. *Bioinformatics*. 2015;31 19:3210-2.
795 doi:10.1093/bioinformatics/btv351.

796 43. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-
797 length transcriptome assembly from RNA-Seq data without a reference genome.
798 *Nat Biotechnol*. 2011;29 7:644-52. doi:10.1038/nbt.1883.

799 44. Wang M and Kong L. pblat: a multithread blat algorithm speeding up aligning
800 sequences to genomes. *BMC Bioinformatics*. 2019;20 1:28. doi:10.1186/s12859-
801 019-2597-8.

802 45. Ryan JF. Baa.pl: A tool to evaluate de novo genome assemblies with RNA
803 transcripts. *arXiv e-prints*. 2013;arXiv:1309.2087.

804 46. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res*. 2002;12 4:656-64.
805 doi:10.1101/gr.229202.

806 47. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR,
807 et al. Iterative correction of Hi-C data reveals hallmarks of chromosome
808 organization. *Nat Methods*. 2012;9 10:999-1003. doi:10.1038/nmeth.2148.
809
810

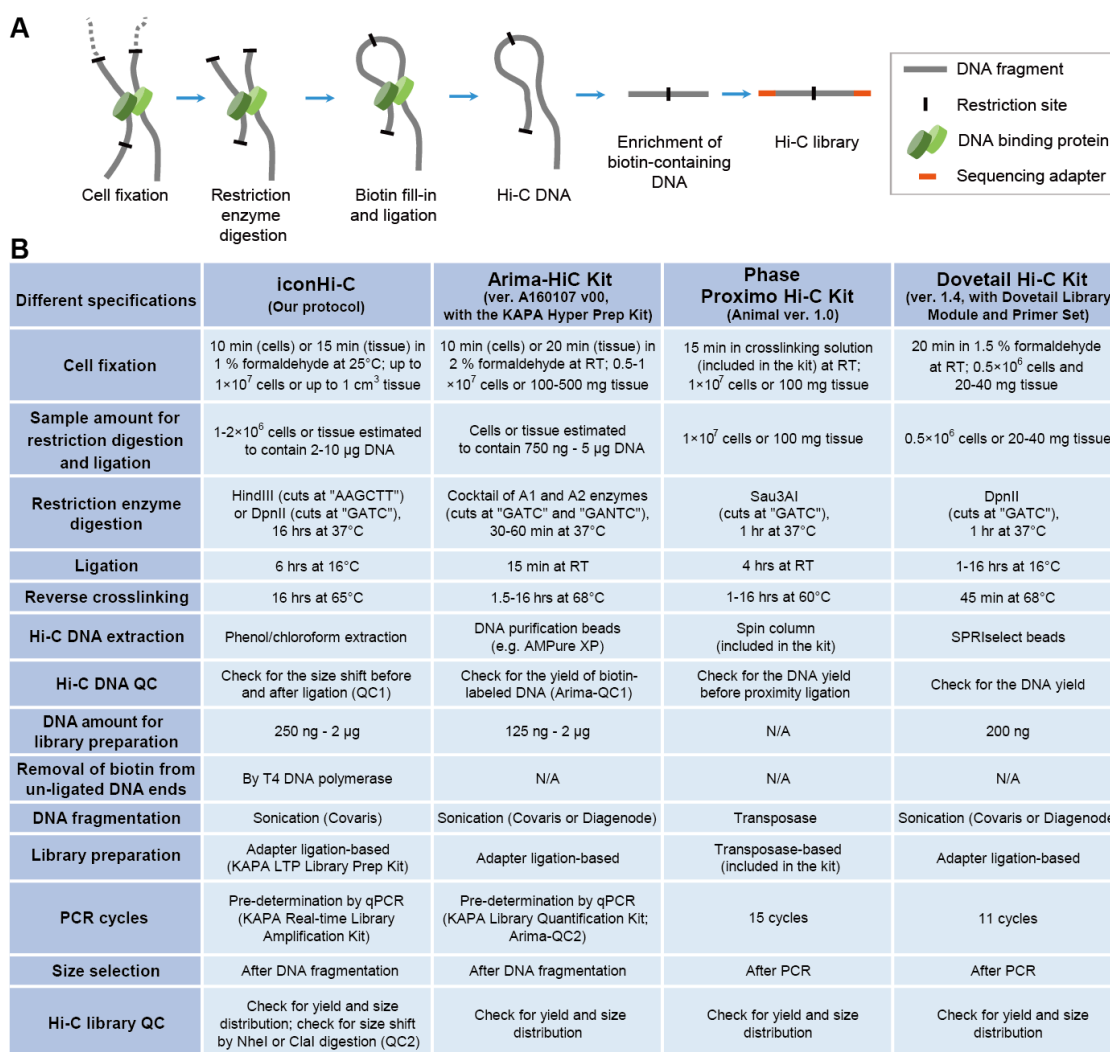
811 **Table 1:** Overview of the specification of major scaffolding programs.

Program	Support and availability	Input data requirement	Other information	Literature
LACHESIS	Developer's support discontinued; intricate installation	Generic bam format	No function to correct scaffold misjoins	[4]
HiRise	Open source version at GitHub not updated since 2015	Generic bam format	Employed in Dovetail Chicago/Hi-C service. Default input sequence length cut-off=1000 bp	[8]
3d-dna	Actively maintained and supported by the developer	Not compatible with multiple enzymes; Accept only Juicer mapper format	Default parameters: -t 15000 (input sequence length cut-off), -r 2 (no. of iterations for misjoin correction)	[11, 38]
SALSA2	Actively maintained and supported by the developer	Compatible with multiple enzymes; generic bam (bed) file, assembly graph, unitig, 10x link files	Default parameters: -c 1000 (input sequence length cut-off), -i 3 (no. of iterations for misjoin correction)	[9, 10]

812

813

814 **Figures**



815

816 **Figure 1: Hi-C library preparation.** (A) Basic procedure. (B) Comparison of Hi-C
 817 library preparation methods. Only the major differences between the methods are
 818 included here. The versions of the Arima and Phase kits used in this study are presented.
 819 The KAPA Hyper Prep Kit (KAPA Biosystems) is assumed to be conjunctly used with
 820 Arima Hi-C Kit, among the several specified kits. See Supplementary Protocol S1 for
 821 the full version of the iconHi-C protocol which was derived from the protocols
 822 published previously [3, 26, 35].

823

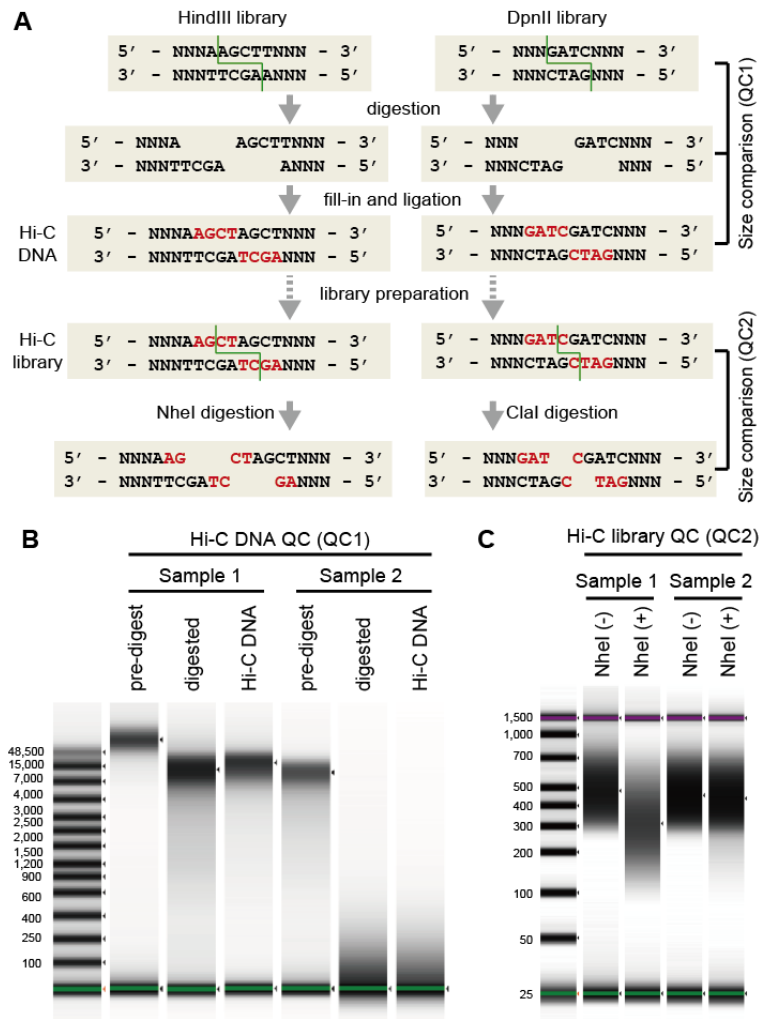


824

825 **Figure 2:** A juvenile softshell turtle *Pelodiscus sinensis*.

826

827



828

829 **Figure 3:** Structure of the Hi-C DNA and principle of the quality controls. (A)

830 Schematic representation of the library preparation workflow based on HindIII or DpnII

831 digestion. The patterns of restriction are indicated by the green lines. The nucleotides

832 that are filled in are indicated by the letters in red. (B) Size shift analysis of HindIII-

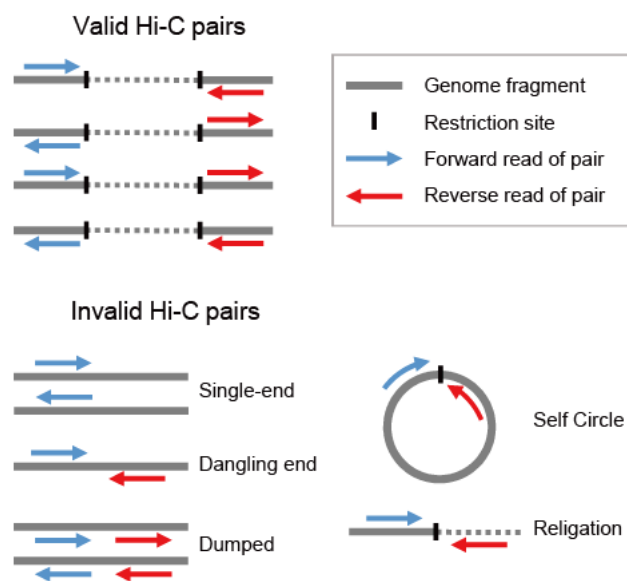
833 digested Hi-C DNA (QC1). Representative images of qualified (Sample 1) and

834 disqualified (Sample 2) samples are shown. (C) Size shift analysis of the HindIII-

835 digested Hi-C library (QC2). Representative images of the qualified (Sample 1) and

836 disqualified (Sample 2) samples are shown. Size distributions were measured with

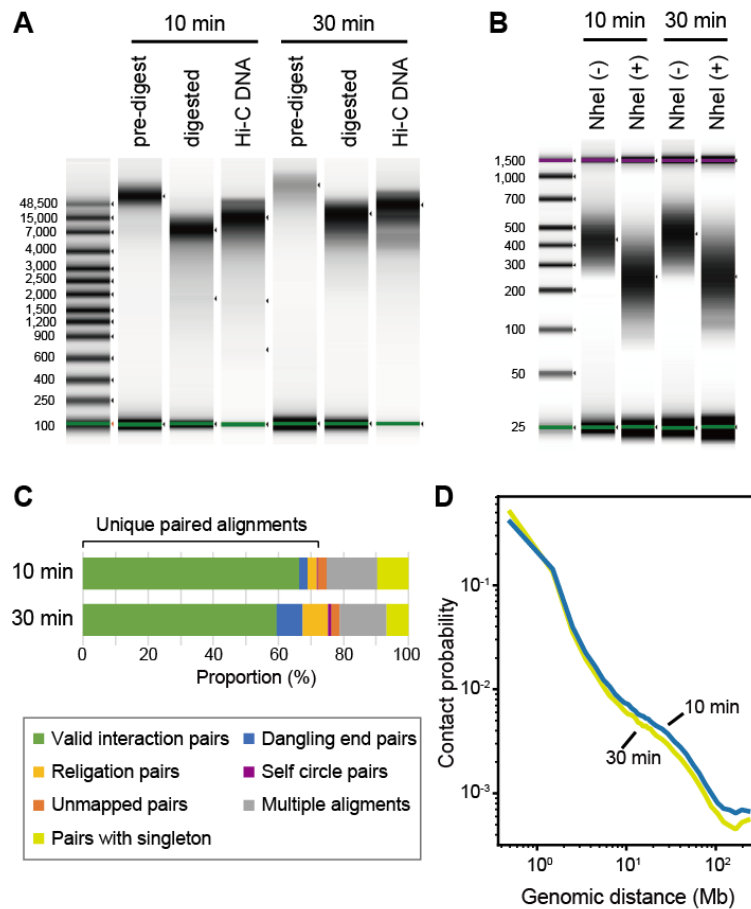
837 Agilent 4200 TapeStation.



838

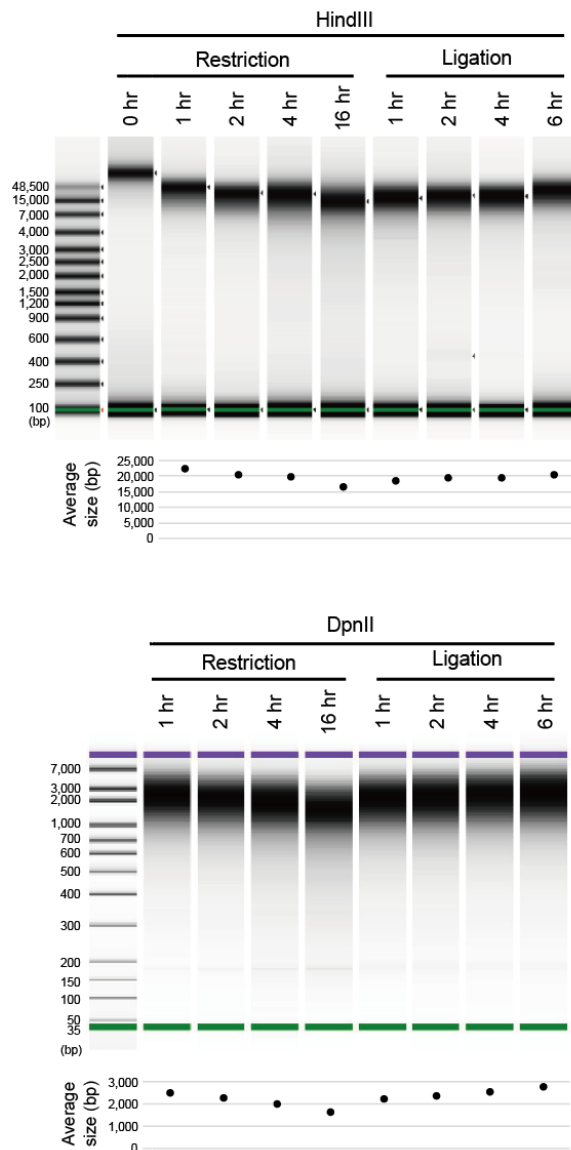
839 **Figure 4:** Post-sequencing quality control of Hi-C reads. Read pairs were categorized
 840 into valid and invalid pairs by HiC-Pro, based on their status in the mapping to the
 841 reference genome (see Methods). This figure was adapted from the article that described
 842 HiC-Pro originally [25].

843



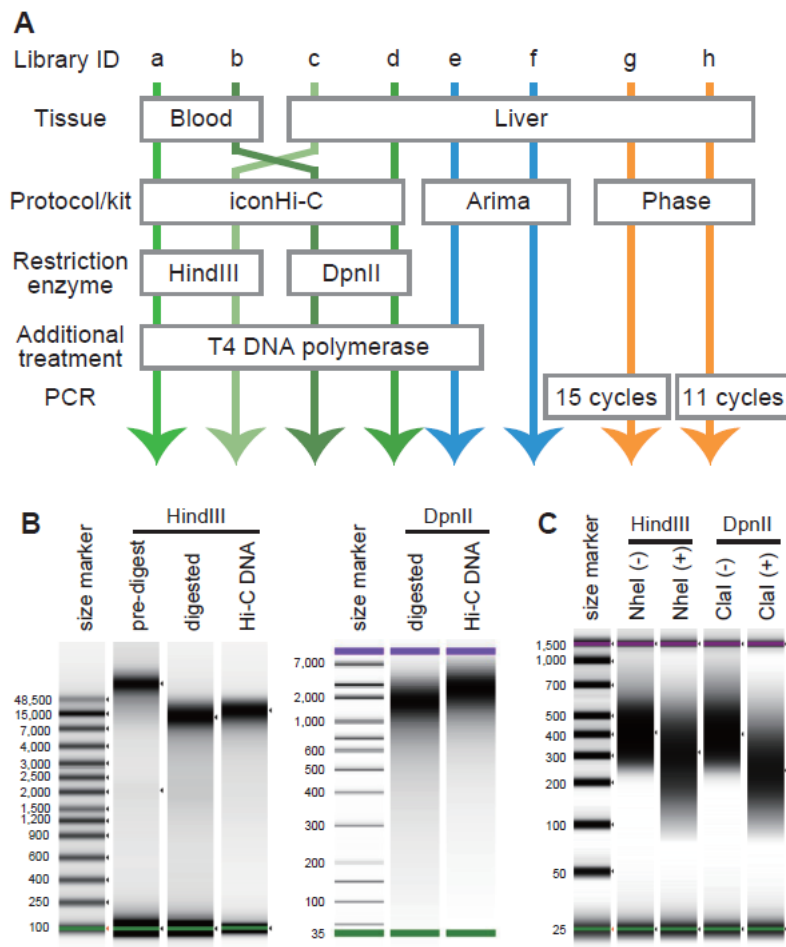
844

845 **Figure 5:** Effect of cell fixation duration. (A) QC1 of the HindIII-digested Hi-C DNA
 846 of human GM12878 cells fixed for 10 or 30 minutes in 1% formaldehyde. (B) QC2 of
 847 the HindIII-digested library of human GM12878 cells. (C) Quality control of the
 848 sequence reads by HiC-Pro using 1 M read pairs. See Fig. 4 for the details of the read
 849 pair categorization. See Supplementary Table S3 for the actual proportion of the reads
 850 in each category. (D) Contact probability measured by the ratio of observed and
 851 expected frequencies of Hi-C read pairs mapped along the same chromosome [47].



852

853 **Figure 6:** Testing varying durations of restriction and ligation. The length distributions
 854 of the DNA molecules prepared from human GM12878 cells after restriction and
 855 ligation of variable duration are shown. The size distributions of the HindIII-digested
 856 samples (top) and DpnII-digested samples (bottom) were measured with an Agilent
 857 4200 TapeStation and an Agilent Bioanalyzer, respectively.



858

859 **Figure 7:** Softshell turtle Hi-C libraries prepared for our methodological comparison.

860 (A) Lineup of the prepared libraries. This chart includes only the conditions in

861 preparation methods that varied between these libraries, and the remainder preparation

862 workflows are described in Supplementary Protocol S1 for the non-commercial

863 ('iconHi-C') protocol and in the manuals of the commercial kits. (B) Quality control of

864 Hi-C DNA (QC1) for Library c and d. The Hi-C DNA for the Chinese softshell turtle

865 liver sample was prepared with either HindIII or DpnII digestion. (C) Quality control of

866 Hi-C libraries (QC2). The HindIII library prepared from the softshell turtle liver was

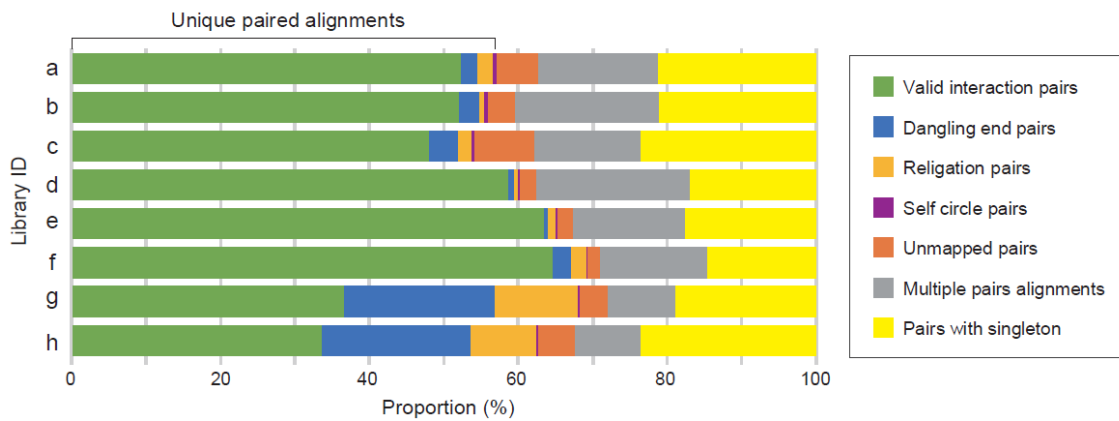
867 digested by NheI, and the DpnII library was digested by ClaI (see Fig. 3 for the

868 technical principle). See Supplementary Fig. S2 for the QC1 and QC2 results of the

869 samples prepared from the blood of this species. See Supplementary Fig. S3 for the

870 QC2 result of the Phase libraries.

871

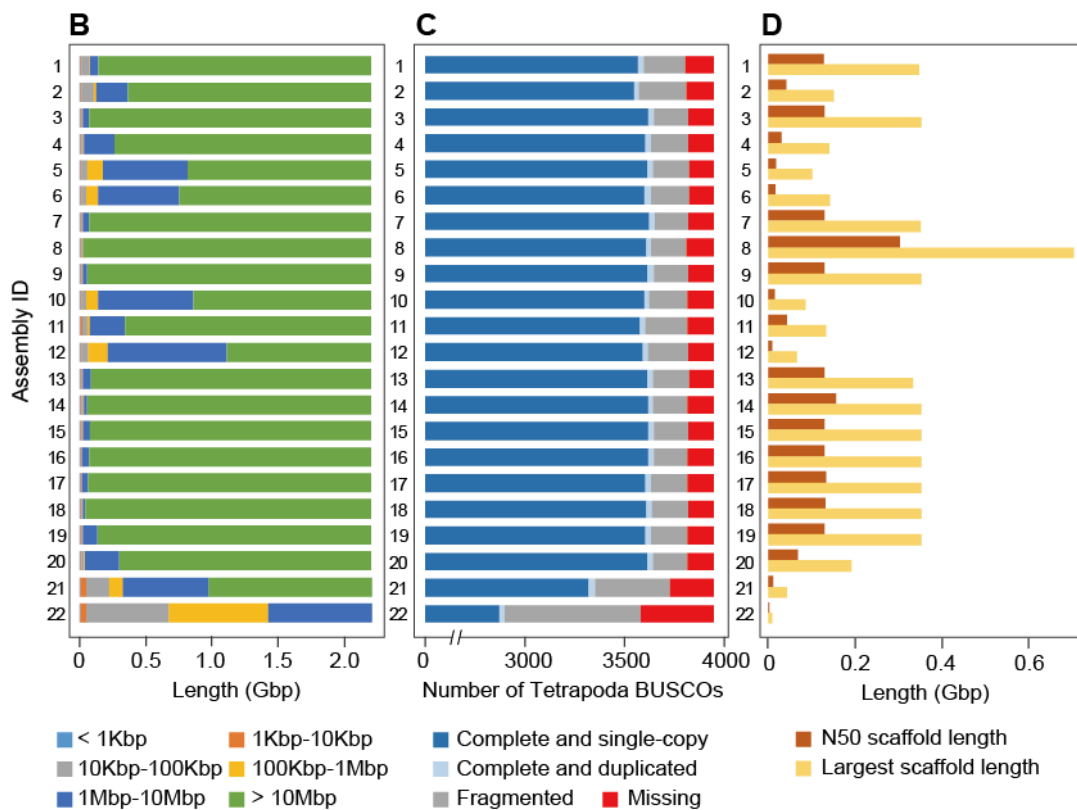


872

873 **Figure 8:** Results of the post-sequencing quality control with HiC-Pro. One million read
874 pairs were used for computation with HiC-Pro. See Fig. 7A for the preparation
875 conditions of Library a-h, Fig. 4 for the categorization, and Supplementary Table S5 for
876 the actual proportion of the reads in each category. The post-sequencing quality control
877 using variable read amounts (500 K to 200 M pairs) for one of these softshell turtle
878 libraries (Supplementary Table S9) and human GM12878 libraries (Supplementary
879 Table S2) shows the validity of this quality control with as few as 500 K read pairs.

A

Assembly ID	Library ID	Scaffolding program	Input sequence length cutoff (nt)	Number of iterative misjoin correction rounds	Number of read pairs input
1	c	3d-dna	15000	2	200 M
2	a				
3	d				
4	b				
5	c	SALSA2	1000	3	
6	d	3d-dna	15000	2	
7	c + d				
8	b + d				
9	e				
10	e	SALSA2	1000	3	
11	h	3d-dna	15000	2	
12	h	SALSA2	1000	3	
13	d	3d-dna	15000	4	
14			10000	6	
15			5000	2	
16			3000		
17			15000		
18			15000		
19			280 M		
20	160 M				
21	80 M				
22	20 M				
					10 M



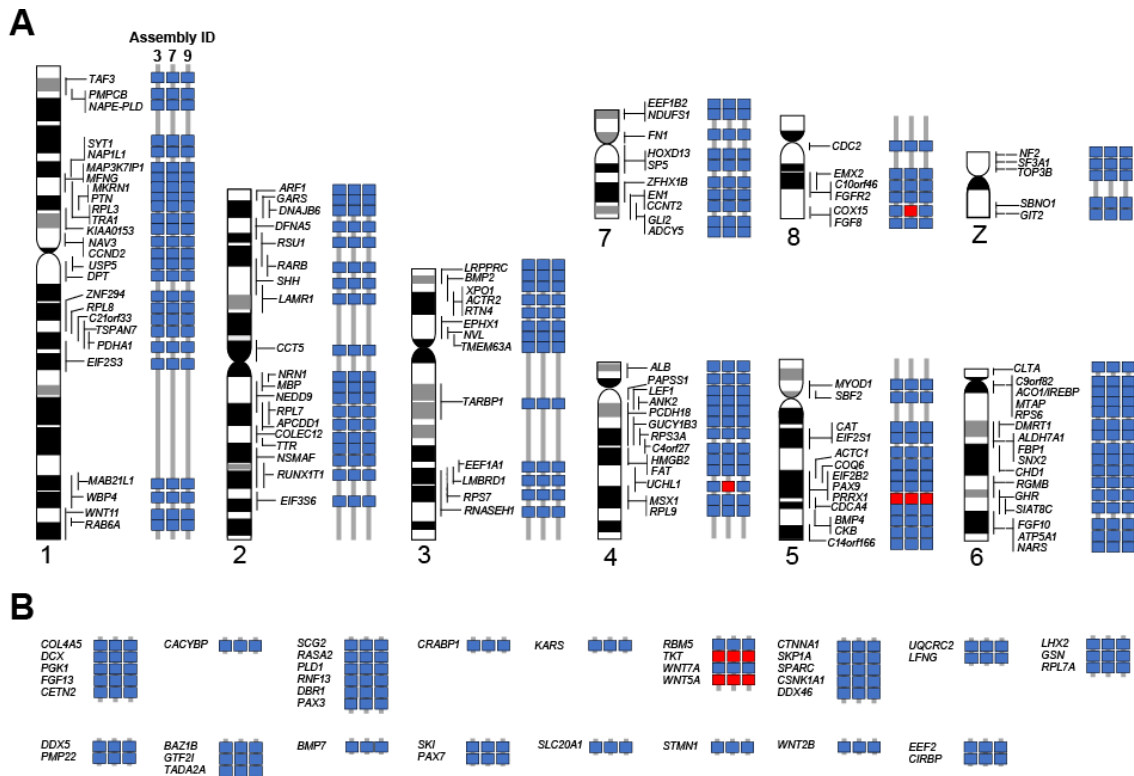
880

881 **Figure 9:** Comparison of Hi-C scaffolding products. (A) Scaffolding conditions used to

882 produce Assembly 1 to 22. The default parameters are shown in red. (B) Scaffold length

883 distributions. (C) Gene space completeness. (D) Largest and N50 scaffold lengths. See
884 the panel A for Library IDs and Supplementary Table S6 for raw values of the metrics
885 shown in B–D.

886



887

888 **Figure 10:** Cytogenetic validation of Hi-C scaffolding results. For the scaffolded
 889 sequences of Assembly 3, 7, and 9, we evaluated the consistency of the positions of the
 890 selected genes that were previously localized on eight macrochromosomes and Z
 891 chromosome (A) and microchromosomes (B) by chromosome FISH [18-22] (see
 892 Results). Concordant and discordant gene locations on individual assemblies are
 893 indicated with blue and red boxes, respectively. The arrays of genes without ideograms
 894 in B were identified on chromosomes that are cytogenetically indistinguishable from
 895 each other.

892 distributions. (C) Gene space completeness. (D) Largest and N50 scaffold lengths. See
893 the panel A for Library IDs and Supplementary Table S6 for raw values of the metrics
894 shown in B–D.

895

896

897