GigaScience, 9, 2020, 1–15

doi: 10.1093/gigascience/giz158
Research

RESEARCH

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

Mitsutaka Kadota[1,†], Osamu Nishimura[1,†], Hisashi Miura[3], Kaori Tanaka[1,2], Ichiro Hiratani[3] and Shigehiro Kuraku[1,*]

[1]Laboratory for Phyloinformatics, RIKEN Center for Biosystems Dynamics Research (BDR), Kobe 650-0047, Japan; [2]Present Address: Division of Transcriptomics, Medical Institute of Bioregulation, Kyushu University, Fukuoka 812-0054, Japan and [3]Laboratory for Developmental Epigenetics, RIKEN BDR, Kobe 650-0047, Japan

*Correspondence address. Shigehiro Kuraku, Laboratory for Phyloinformatics, RIKEN BDR, Japan. Tel: +81 78 306 3048; Fax: +81 78 306 3048; E-mail: shigehiro.kuraku@riken.jp http://orcid.org/0000-0003-1464-8388

†These authors contributed equally to this study.

Abstract

Background: Hi-C is derived from chromosome conformation capture (3C) and targets chromatin contacts on a genomic scale. This method has also been used frequently in scaffolding nucleotide sequences obtained by *de novo* genome sequencing and assembly, in which the number of resultant sequences rarely converges to the chromosome number. Despite its prevalent use, the sample preparation methods for Hi-C have not been intensively discussed, especially from the standpoint of genome scaffolding. **Results:** To gain insight into the best practice of Hi-C scaffolding, we performed a multifaceted methodological comparison using vertebrate samples and optimized various factors during sample preparation, sequencing, and computation. As a result, we identified several key factors that helped improve Hi-C scaffolding, including the choice and preparation of tissues, library preparation conditions, the choice of restriction enzyme(s), and the choice of scaffolding program and its usage. **Conclusions:** This study provides the first comparison of multiple sample preparation kits/protocols and computational programs for Hi-C scaffolding by an academic third party. We introduce a customized protocol designated “inexpensive and controllable Hi-C (iconHi-C) protocol,” which incorporates the optimal conditions identified in this study, and demonstrate this technique on chromosome-scale genome sequences of the Chinese softshell turtle *Pelodiscus sinensis*.

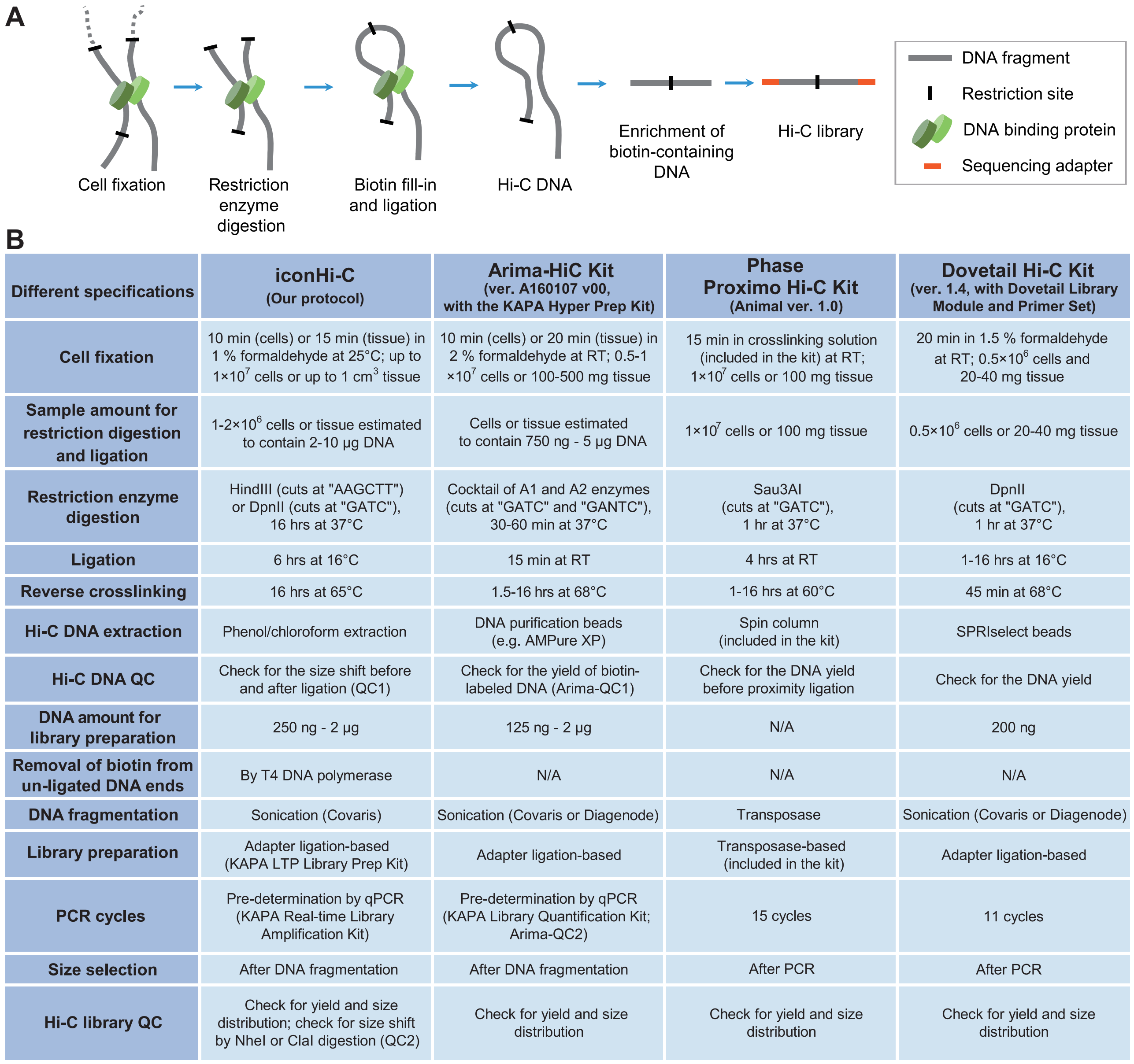
Keywords: Hi-C; genome scaffolding; chromosomes; proximity-guided assembly; softshell turtle

Background

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

Analyses of chromatin conformation using Hi-C have revealed more frequent contacts between more closely linked genomic regions, which has recently prompted the use of this method in scaffolding *de novo* genome sequences [4–6]. In *de novo* genome sequencing, the number of assembled sequences is usually far larger than the number of chromosomes in the

Received: 6 June 2019; **Revised:** 23 October 2019; **Accepted:** 2 December 2019



Different specifications	iconHi-C (Our protocol)	Arima-HiC Kit (ver. A160107 v00, with the KAPA Hyper Prep Kit)	Phase Proximo Hi-C Kit (Animal ver. 1.0)	Dovetail Hi-C Kit (ver. 1.4, with Dovetail Library Module and Primer Set)
Cell fixation	10 min (cells) or 15 min (tissue) in 1 % formaldehyde at 25°C; up to 1×10^{7} cells or up to 1 cm^{3} tissue	10 min (cells) or 20 min (tissue) in 2 % formaldehyde at RT; 0.5-1 $\times10^{7}$ cells or 100-500 mg tissue	15 min in crosslinking solution (included in the kit) at RT; 1×10^{7} cells or 100 mg tissue	20 min in 1.5 % formaldehyde at RT; 0.5×10^{6} cells and 20-40 mg tissue
Sample amount for restriction digestion and ligation	1-2×10^{6} cells or tissue estimated to contain 2-10 µg DNA	Cells or tissue estimated to contain 750 ng - 5 µg DNA	1×10^{7} cells or 100 mg tissue	0.5×10^{6} cells or 20-40 mg tissue
Restriction enzyme digestion	HindIII (cuts at "AAGCTT") or DpnII (cuts at "GATC"), 16 hrs at 37°C	Cocktail of A1 and A2 enzymes (cuts at "GATC" and "GANTC"), 30-60 min at 37°C	Sau3AI (cuts at "GATC"), 1 hr at 37°C	DpnII (cuts at "GATC"), 1 hr at 37°C
Ligation	6 hrs at 16°C	15 min at RT	4 hrs at RT	1-16 hrs at 16°C
Reverse crosslinking	16 hrs at 65°C	1.5-16 hrs at 68°C	1-16 hrs at 60°C	45 min at 68°C
Hi-C DNA extraction	Phenol/chloroform extraction	DNA purification beads (e.g. AMPure XP)	Spin column (included in the kit)	SPRIselect beads
Hi-C DNA QC	Check for the size shift before and after ligation (QC1)	Check for the yield of biotin-labeled DNA (Arima-QC1)	Check for the DNA yield before proximity ligation	Check for the DNA yield
DNA amount for library preparation	250 ng - 2 µg	125 ng - 2 µg	N/A	200 ng
Removal of biotin from un-ligated DNA ends	By T4 DNA polymerase	N/A	N/A	N/A
DNA fragmentation	Sonication (Covaris)	Sonication (Covaris or Diagenode)	Transposase	Sonication (Covaris or Diagenode)
Library preparation	Adapter ligation-based (KAPA LTP Library Prep Kit)	Adapter ligation-based	Transposase-based (included in the kit)	Adapter ligation-based
PCR cycles	Pre-determination by qPCR (KAPA Real-time Library Amplification Kit)	Pre-determination by qPCR (KAPA Library Quantification Kit; Arima-QC2)	15 cycles	11 cycles
Size selection	After DNA fragmentation	After DNA fragmentation	After PCR	After PCR
Hi-C library QC	Check for yield and size distribution; check for size shift by NheI or ClaI digestion (QC2)	Check for yield and size distribution	Check for yield and size distribution	Check for yield and size distribution

Figure 1: Hi-C library preparation. (A) Basic procedure. (B) Comparison of Hi-C library preparation methods. Only the major differences between the methods are included here. The versions of the Arima and Phase kits used in this study are presented. The KAPA Hyper Prep Kit (KAPA Biosystems) is assumed to be conjointly used with the Arima-HiC Kit, among the several specified kits. See Supplementary Protocol S1 for the full version of the iconHi-C protocol, which was derived from the protocols published previously [3, 7, 8]. N/A: not applicable; QC: quality control; RT: room temperature.

karyotype of the species of interest, regardless of the sequencing platform chosen [9]. The application of Hi-C scaffolding enabled a remarkable enhancement of sequence continuity to reach a chromosome scale, and the integration of fragmentary sequences into longer sequences, which are similar in number to that of chromosomes in the karyotype.

In early 2018, commercial Hi-C library preparation kits were introduced (Fig. 1B), and *de novo* genome assembly was revolutionized by the release of versatile computational programs for Hi-C scaffolding (Table 1), namely, LACHESIS [4], HiRise [10], SALSA [11, 12], and 3d-dna [13] (reviewed by Ghurye and Pop [14]). These movements assisted the rise of mass sequencing projects targeting a number of species, such as the Earth BioGenome Project [15], the Genome 10 K/Vertebrate Genome Project [16], and the DNA Zoo Project [17]. Optimization of Hi-C sample preparation, however, has been limited [18], which leaves room for the improvement of efficiency and the reduction of required sample quantity. Thus, the specific factors that are key for Hi-C scaffolding remain unexplored, mainly because of the costly and resource-demanding nature of this technology.

In addition to performing protocol optimization using human culture cells, we focused on the softshell turtle *Pelodiscus sinensis* (Fig. 2). This species has been adopted as a study system for evolutionary developmental biology, including the study of the formation of the dorsal shell (carapace) (reviewed by Kuratani et al. [19]). Access to genome sequences of optimal quality by relevant research communities is desirable in this field.

Table 1: Overview of the specifications 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	Used in Dovetail Chicago/Hi-C service. Default input sequence length cut-off = 1,000 bp	[10]
3d-dna	Actively maintained and supported by the developer	Not compatible with multiple enzymes; accepts only Juicer mapper format	Default parameters: -t 15000 (input sequence length cut-off), -r 2 (No. of iterations for misjoin correction)	[13, 20]
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)	[11, 12]

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

In Japan, live materials (adults and embryos) of this species are available through local farms mainly between May and August, which implies its high utility for sustainable research. A previous cytogenetic report revealed that the karyotype of this species consists of 33 chromosome pairs including Z and W chromosomes (2n = 66) that show a wide variety of sizes (conventionally categorized as macrochromosomes and microchromosomes) [21]. Despite the moderate global GC content in its whole genome at ∼44%, the intragenomic heterogeneity of GC content between and within the chromosomes has been suggested [22]. A wealth of cytogenetic efforts on this species led to the accumulation of fluorescence *in situ* hybridization (FISH)-based mapping data for 162 protein-coding genes covering almost all chromosomes [21–25], which serve as structural landmarks for validating genome assembly sequences.

A draft sequence assembly of the softshell turtle genome was built using short reads and was released in 2013 [26]. This sequence assembly achieved the N50 scaffold length of >3.3 Mb but remains fragmented into ∼20,000 sequences (see Supplementary Table S1). The longest sequence in this assembly is only slightly larger than 16 Mb, which is much shorter than the largest chromosome size estimated from the karyotype report [21]. The total size of the assembly is ∼2.2 Gb, which is a moderate size for a vertebrate species. Because of the affordable genome size, sufficiently complex structure, and availability of validation methods, we reasoned that the genome of this species is a suitable target for our methodological comparison, and its improved genome assembly is expected to assist a wide range of genome-based studies of this species.

Results

Stepwise QC prior to large-scale sequencing

It would be ideal to be able to assess the quality of prepared libraries before engaging in costly sequencing. Based on the literature [18, 27], we routinely control the quality of Hi-C DNAs and Hi-C libraries by observing DNA size shifts via digestion targeting the restriction sites in properly prepared samples (Fig. 3). More concretely, a successfully ligated Hi-C DNA sample should exhibit a slight increase in the length of its restricted DNA fragments after ligation (quality control 1 [QC1]), which serves as an indicator of qualified samples (e.g., Sample 1 in Fig. 3B). In contrast, an unsuccessfully prepared Hi-C DNA does not exhibit this length recovery (e.g., Sample 2 in Fig. 3B). In a subsequent step, DNA molecules in a successfully prepared HindIII-digested Hi-C library should contain the NheI restriction site at a high probability. Thus, the length distribution observed after NheI digestion of the prepared library serves as an indicator of qualified or disqualified products (QC2; Fig. 3C). This series of QCs is incorporated into our protocol by default (Supplementary Protocol S1) and can also be performed in combination with sample preparation using commercial kits if it employs a single restriction enzyme.

Some of the libraries that we prepared passed the QC steps performed before sequencing but yielded an unfavourably large proportion of invalid read pairs. To identify such libraries, we routinely performed small-scale sequencing for quick and inexpensive QC (designated "QC3") using the HiC-Pro program [28] (see Fig. 4 for the read pair categories assigned by HiC-Pro). Our test using variable input data sizes (500,000–200,000,000 read pairs) resulted in highly similar breakdowns into different categories of read pair properties (Supplementary Table S2) and guaranteed QC3 with an extremely small data size of ≤1,000,000 reads. These post-sequencing QC steps, which do not incur a large cost, are expected to help avoid the large-scale sequencing of unsuccessful libraries that have somehow passed through the QC1 and QC2 steps. Importantly, libraries that have passed QC3 can be further sequenced with greater depth, as necessary.

Optimization of sample preparation conditions

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

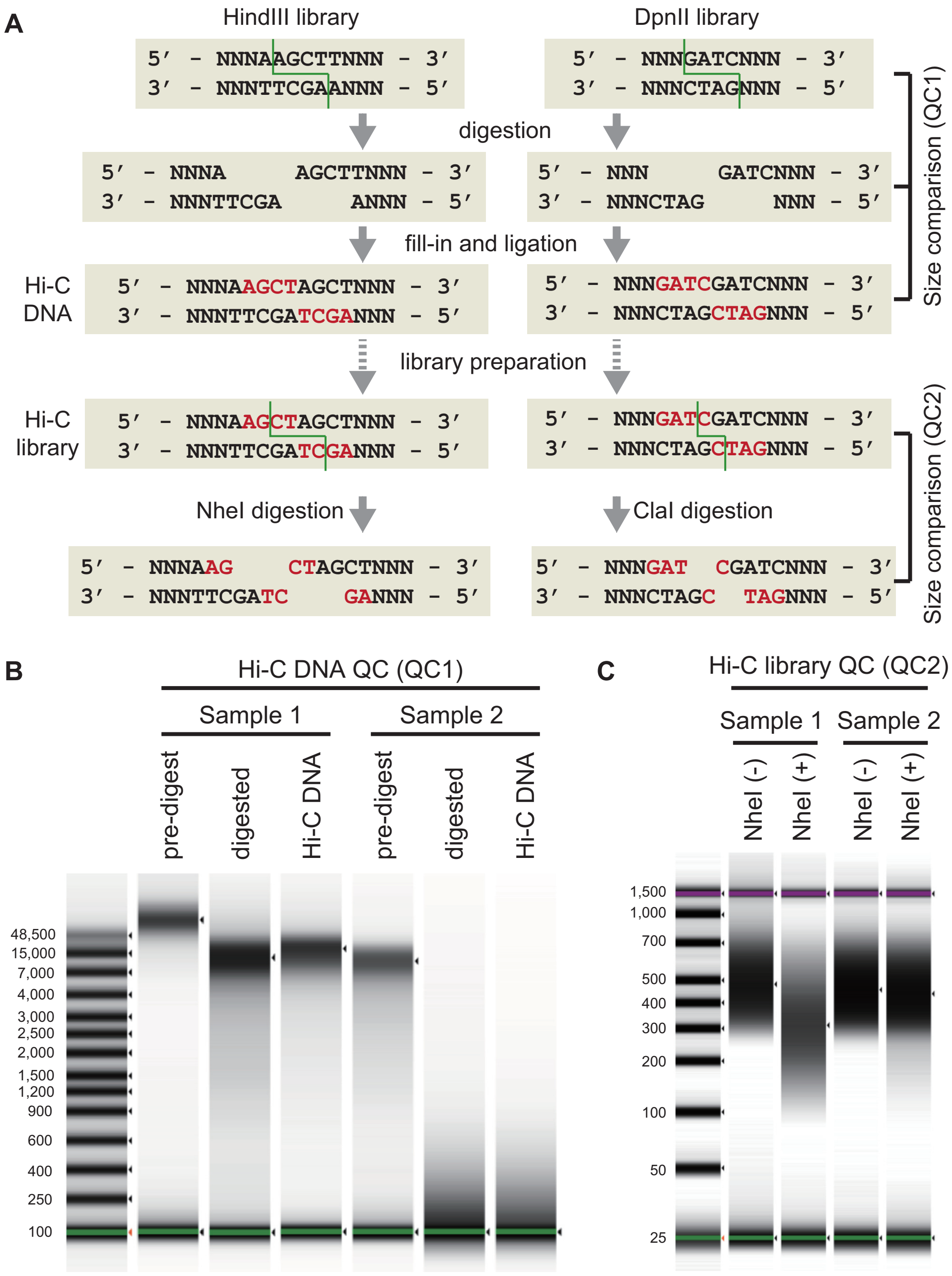


Figure 3: Structure of the Hi-C DNA and principle of the quality controls. (A) Schematic representation of the library preparation workflow based on HindIII or DpnII digestion. The patterns of restriction are indicated by the green lines. The nucleotides that are filled in are indicated by the letters in red. (B) Size shift analysis of HindIII-digested Hi-C DNA (QC1). Representative images of qualified (Sample 1) and disqualified (Sample 2) samples are shown. (C) Size shift analysis of the HindIII-digested Hi-C library (QC2). Representative images of the qualified (Sample 1) and disqualified (Sample 2) samples are shown. Size distributions were measured with Agilent 4200 TapeStation.

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

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

The reduced preparation time of commercial Hi-C kits (≤2 days according to their advertisement) is attributable mainly to shortened restriction and ligation times (Fig. 1B). To monitor the effect of shortening these enzymatic reactions, we first analysed the progression of restriction and ligation in a time course experiment using GM12878 cells. We observed the persistent progression of restriction up to 16 hours and of ligation up to 6 hours (Fig. 6). To scrutinize further the possible adverse effects of the prolonged reaction, Hi-C libraries of GM12878 cells were prepared with variable durations of restriction digestion (1 and 16 hour(s)) and ligation (15 minutes, 1 hour, and 6 hours). We found that the proportions of dangling end and religation read

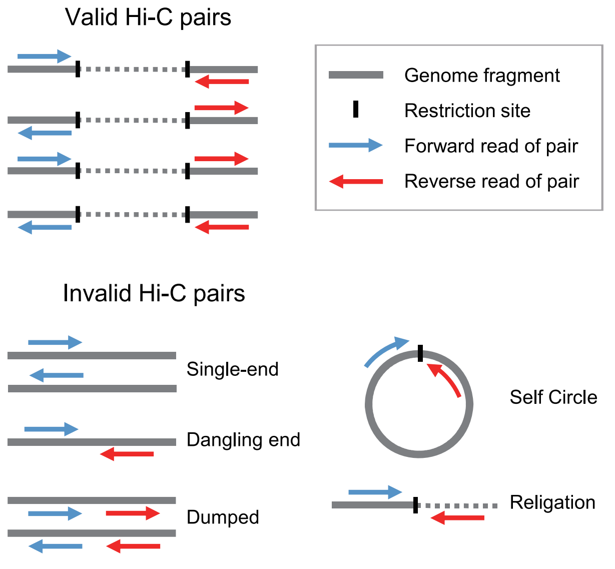


Figure 4: Post-sequencing quality control of Hi-C reads. Read pairs were categorized into valid and invalid pairs by HiC-Pro on the basis of their status in the mapping to the reference genome (see Methods). This figure was adapted from the article that described HiC-Pro originally [28].

pairs were reduced in cases with an extended duration of restriction digestion (Supplementary Table S4). The yield of the library, which can be estimated from the number of PCR cycles, increased with the extended duration of ligation without any effect on the proportion of valid interaction read pairs (Supplementary Table S4). The proportion of valid interaction read pairs containing the proper DpnII junction sequence "GATCGATC" also remained unchanged, suggesting that the prolonged reaction times did not induce any adverse effects, such as star activity of the restriction enzyme.

Multifaceted comparison using softshell turtle samples

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

We performed Hi-C sample preparation and scaffolding using tissues from a female Chinese softshell turtle, which has both Z and W chromosomes [21]. We prepared Hi-C libraries using various tissues (liver or blood cells), restriction enzymes (HindIII or DpnII), and protocols (our iconHi-C protocol, the Arima kit in conjunction with the KAPA Hyper Prep Kit, or the Phase kit), as outlined in Fig. 7A (see Supplementary Table S5; Supplementary Fig. S1). As in some of the existing protocols (e.g., [8]), we performed T4 DNA polymerase treatment in our iconHi-C protocol (Libraries a–d), expecting reduced proportions of "dangling end" read pairs that contain no ligated junction and thus do not contribute to Hi-C scaffolding. We also incorporated this T4 DNA polymerase treatment into the workflow of the Arima kit (Library e vs Library f without this additional treatment). Furthermore, we tested a lesser degree of PCR amplification (11 cycles) together with the use of the Phase kit, which recommends as many as 15 cycles by default (Library h vs Library g; Fig. 7A).

All samples prepared using the iconHi-C protocol passed both controls, QC1 and QC2 (Fig. 7B and 7C). The prepared Hi-C libraries were sequenced to obtain one million 127 nt-long read pairs and were subjected to QC3 using the HiC-Pro program (Fig. 8). As a result of this QC3, the largest proportion of "valid interaction" pairs was observed for Arima libraries (Libraries e and f). Regarding the iconHi-C libraries (Libraries a–d), fewer "unmapped" and "religation" pairs were detected for the DpnII libraries compared with HindIII libraries. It should be noted that the QC3 of the softshell turtle libraries generally produced lower proportions of the "valid interaction" category and larger proportions of "unmapped pairs" and "pairs with singleton" than with the human libraries. This cross-species difference may be attributable to the use of incomplete genome sequences as a reference for Hi-C read mapping (Supplementary Table S1). This invokes a caution when comparing QC results across species.

Scaffolding using variable input and computational conditions

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

First, using the default parameters, 3d-dna consistently produced more continuous assemblies than did SALSA2 (see Assembly 1 vs 5, 3 vs 6, 9 vs 10, and 11 vs 12 in Fig. 9). Second, the increase in the number of iterative corrections ("-r" option of 3d-dna) resulted in relatively large N50 lengths, but with more missing orthologues (see Assembly 3 and 13–14). Third, a smaller input sequence cut-off length ("-i" option of 3d-dna) resulted in a smaller number of scaffolds but again, with more missing orthologues (see Assembly 3 and 15–17). Fourth, the use of the liver libraries consistently resulted in a higher continuity than the use of the blood cell libraries (see Assembly 1 vs 2 and 3 vs 4 in Fig. 9).

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

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

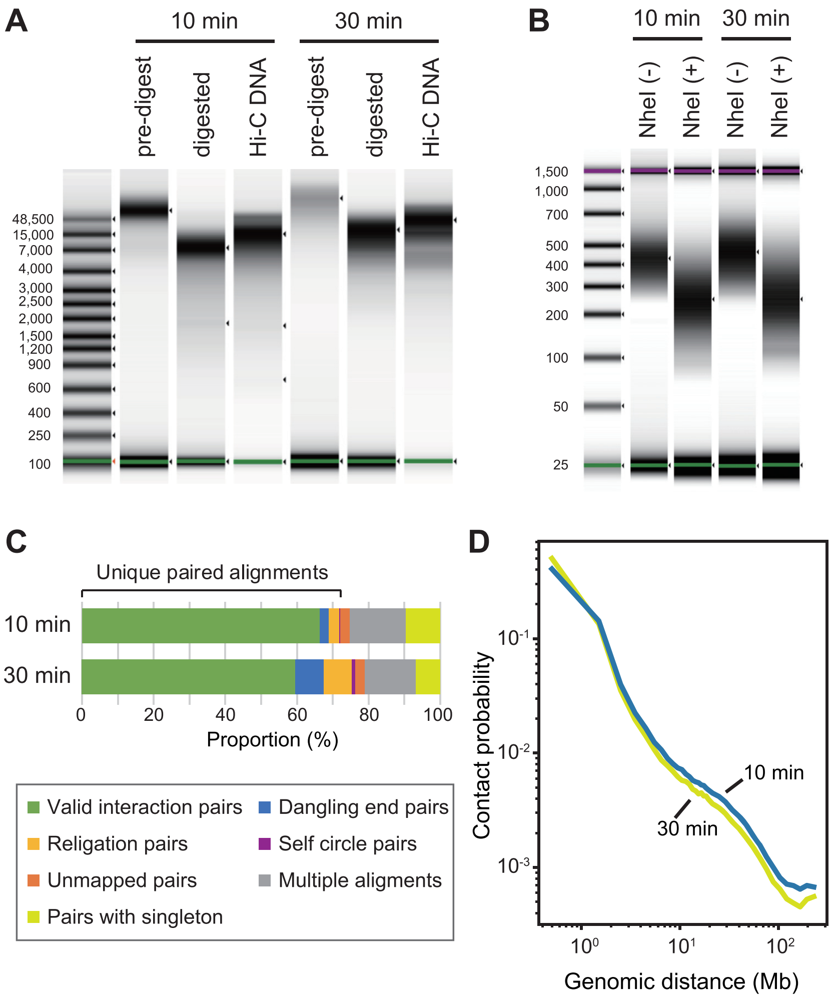


Figure 5: Effect of cell fixation duration. (A) QC1 of the HindIII-digested Hi-C DNA of human GM12878 cells fixed for 10 or 30 minutes in 1% formaldehyde. (B) QC2 of the HindIII-digested library of human GM12878 cells. (C) Quality control of the sequence reads by HiC-Pro using 1,000,000 read pairs. See Fig. 4 for the details of the read pair categorization. See Supplementary Table S3 for the actual proportion of the reads in each category. (D) Contact probability measured by the ratio of observed and expected frequencies of Hi-C read pairs mapped along the same chromosome [29].

To gain additional insight regarding the evaluation of the scaffolding results, we assessed the contact maps constructed upon the Hi-C scaffolds (Supplementary Fig. S5). The comparison of Assembly 3, 9, and 11, which represent the 3 different preparation methods, revealed anomalous patterns, particularly for Assembly 11, with intensive contact signals separated from the diagonal line that indicates the presence of errors in the scaffolds [17]. We also performed genome-wide alignments between the Hi-C scaffolds obtained. The comparison of Assembly 3, 9, and 11 revealed a high similarity between Assembly 3 and 9, while Assembly 11 exhibited a significantly larger number of inconsistencies against either of the other 2 assemblies (Supplementary Fig. S6). These observations are consistent with the evaluation based on sequence length and gene space completeness, which alone does not, however, provide a reliable metric for the assessment of the quality of scaffolding.

Validation of scaffolding results using transcriptome and FISH data

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

Second, we referred to the FISH mapping data of 162 protein-coding genes from published cytogenetic studies [21–25], which allowed us to check the locations of those genes with our resultant Hi-C assemblies. In this analysis, we evaluated Assembly 3, 7, and 9 (see Fig. 9A), which showed better scaffolding results in terms of sequence length distribution and gene space completeness (Fig. 9D). As a result, we confirmed the positioning of almost all genes and their continuity over the centromeres, which encompassed not only large but also small chromosomes (conventionally called “macrochromosomes” and “microchromosomes”; Fig. 10). Two genes that were not confirmed by Assembly 7 (*UCHL1* and *COX15*; Fig. 10) were found in separate scaffold sequences that were shorter than 1 Mb, which

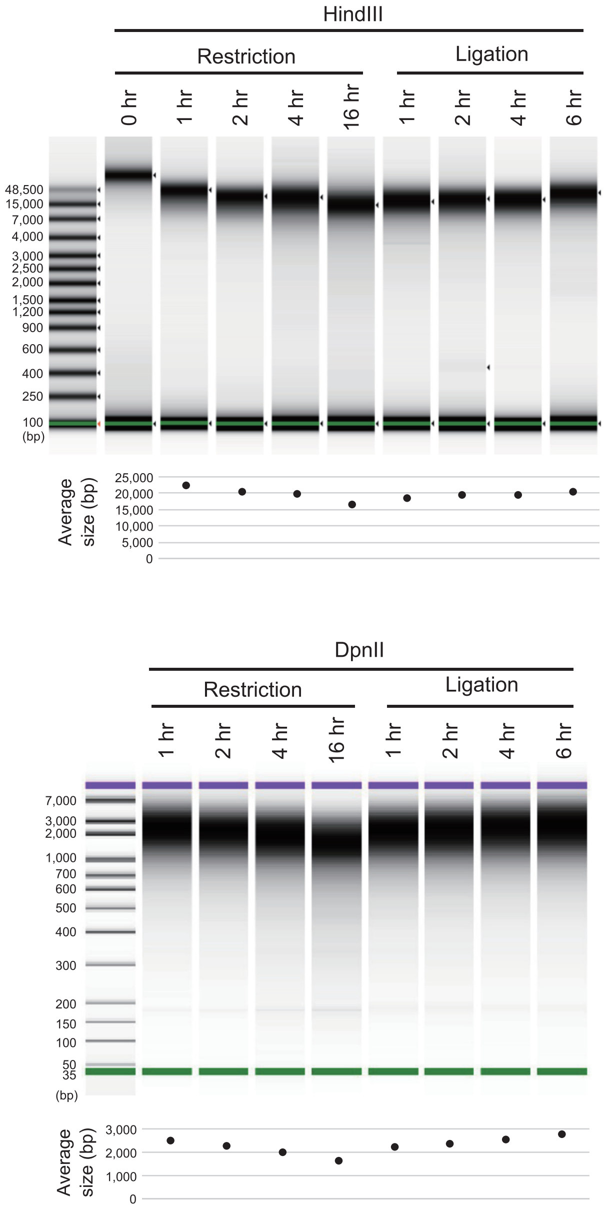


Figure 6: Testing varying durations of restriction and ligation. The length distributions of the DNA molecules prepared from human GM12878 cells after restriction and ligation of variable duration are shown. The size distributions of the HindIII-digested samples (top) and DpnII-digested samples (bottom)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

indicates insufficient scaffolding. Conversely, the gene array including *RBM5*, *TKT*, *WNT7A*, and *WNT5A*, previously shown by FISH, was consistently unconfirmed by all 3 assemblies (Fig. 10), which did not provide any clues for among-assembly evaluation or perhaps indicates an erroneous interpretation of FISH data in a previous study.

Discussion

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

In genome sequencing, best practices for high molecular weight DNA extraction have often been discussed (e.g., [30]). This factor is fundamental to building longer contigs, regardless of the use of short-read or long-read sequencing platforms. Moreover, the proximity ligation method using Chicago libraries provided by Dovetail Genomics, which is based on *in vitro* chromatin reconstruction [10], uses genomic DNA as starting material. In contrast, proximity-guided assembly enabled by Hi-C uses cellular nuclei with preserved chromatin conformation, which brings a new technical challenge regarding appropriate sampling and sample preservation in genomics.

In the preparation of the starting material, it is important to optimize the degree of cell fixation depending on sample choice, to obtain an optimal result in Hi-C scaffolding (Fig. 5). Another practical indication of tissue choice was obtained by examining Assembly 8 (Fig. 9A). This assembly was produced by 3d-dna scaffolding using both liver and blood libraries (Libraries b and d), which led to an unacceptable result possibly caused by over-assembly (Fig. 9B–D; also see Results). It is likely that increased cellular heterogeneity, which possibly introduces excessive conflicting chromatin contacts, did not allow the scaffolding program to group and order the input genome sequences properly. In brief, we recommend the use of samples with modest cell-type heterogeneity that are amenable to thorough fixation.

Considerations regarding sample preparation

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

According to our calculations, the preparation of a Hi-C library using the iconHi-C protocol would be ≥3 times cheaper than the use of a commercial kit. Practically, the cost difference would be even larger, either when the purchased kit is not fully consumed or when the post-sequencing computation steps cannot be undertaken in-house, which implies additional outsourcing costs.

The genomic regions that are targeted by Hi-C are determined by the choice of restriction enzymes. Theoretically, 4-base cutters (e.g., DpnII), which potentially have more frequent restriction sites on the genome, are expected to provide a higher resolution than 6-base cutters (e.g., HindIII) [18]. Obviously, the use of restriction enzymes that were not used in this study might be promising in the adaptation of the protocol to organisms with variable GC content or methylation profiles. However, this might not be so straightforward when considering the interspecies variation in GC content and the intra-genomic heterogeneity. The use of multiple enzymes in a single reaction is a promising approach; however, from a computational viewpoint, not all scaffolding programs are compatible with multiple enzymes (see Table 1 for a comparison of the specifications of scaffolding programs). Another technical downside of this approach is the incompatibility of DNA ends restricted by multiple enzymes, with restriction-based QCs, such as the QC2 step of our iconHi-C protocol (Fig. 3). Therefore, in this study, DpnII and HindIII were used separately in the iconHi-C protocol, which resulted in a higher scaffolding performance with the DpnII library (Figs 8 and 9), as expected. In addition, we input the separately prepared DpnII and HindIII libraries together in scaffolding (Assembly 7), but this approach did not lead to higher scaffolding performance (Figs 9B–D and 10). The Arima kit uses 2 different enzymes that can produce a much greater number of restriction site combinations because 1 of these 2 enzymes recognizes the nucleotide stretch “GANTC.” The increase of

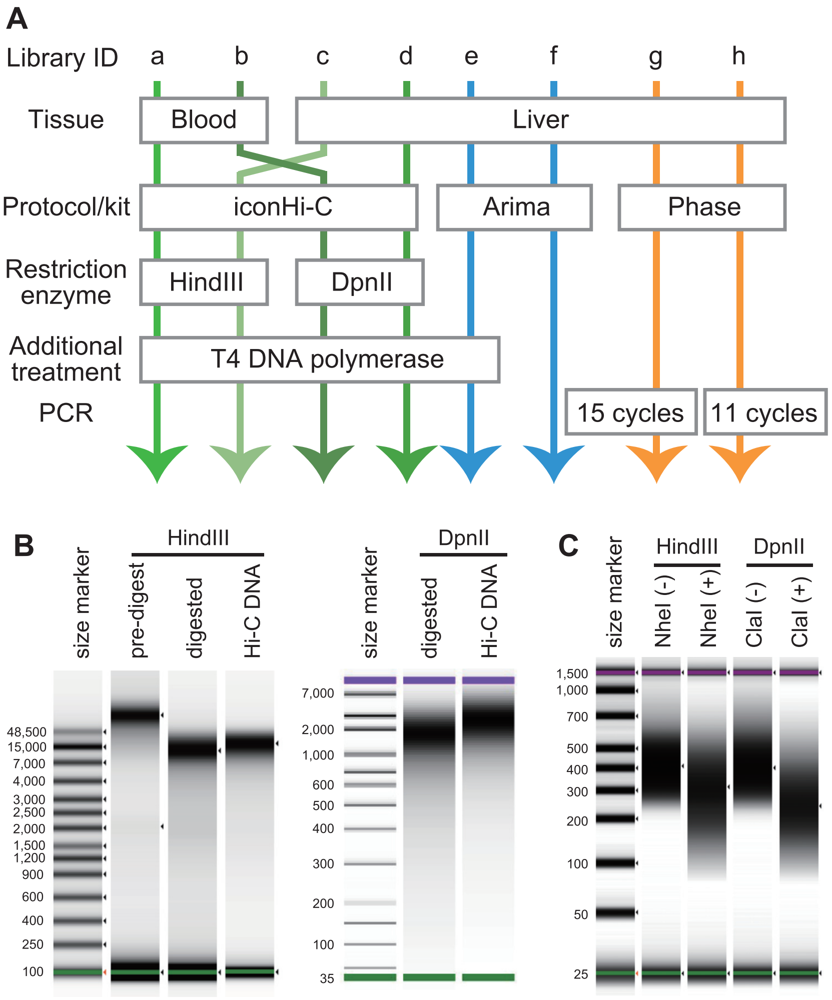


Figure 7: Softshell turtle Hi-C libraries prepared for our methodological comparison. (A) Line-up of the prepared libraries. This chart includes only the conditions in preparation methods that varied between these libraries, and the remainder of the preparation workflows are described in Supplementary Protocol S1 for the non-commercial ("iconHi-C") protocol and in the manuals of the commercial kits. (B) Quality control of Hi-C DNA (QC1) for Libraries c and d. The Hi-C DNA for the Chinese softshell turtle liver sample was prepared with either HindIII or DpnII digestion. (C) Quality control of Hi-C libraries (QC2). The HindIII library prepared from the softshell turtle liver was digested by NheI, and the DpnII library was digested by ClaI (see Fig. 3 for the technical principle). See Supplementary Fig. S2 for the QC1 and QC2 results of the samples prepared from the blood of this species. See Supplementary Fig. S3 for the QC2 result of the Phase libraries.

restriction site combinations might have possibly contributed to the larger proportion of valid interaction pairs (Fig. 8). Scaffolding with the libraries prepared using this kit resulted in one of the most acceptable assemblies (Assembly 9). However, this result did not explicitly exceed the performance of scaffolding with the iconHi-C libraries, including the one that used a single enzyme (DpnII; Library d).

Overamplification by PCR is a concern regarding the use of commercial kits (with the exception of the Arima kit used with the Arima-QC2) because their manuals specify the use of a certain number of PCR cycles *a priori* (15 cycles for the Phase kit and 11 cycles for the Dovetail Hi-C kit) (Supplementary Table S8). In our iconHi-C protocol, an optimal number of PCR cycles is estimated by means of a preliminary real-time PCR using a small aliquot (Step 11.25 to 11.29 in Supplementary Protocol S1), as done traditionally for other library types (e.g., [31]). This procedure allowed us to reduce the number of PCR cycles, down to as few as 5 cycles (Supplementary Table S5). The Dovetail Hi-C kit recommends the use of larger amounts of kit components than that specified for a single sample, depending on the genome size, as well as the degree of genomic heterozygosity and repetitiveness, of the species of interest. In contrast, with our iconHi-C protocol, we always prepared a single library, regardless of those species-specific factors, which seemed to suffice in all the cases tested.

Commercial Hi-C kits, which usually advertise ease and speed of use, have largely shortened the protocol down to 2 days, compared with the published non-commercial protocols (e.g., [18]). Such time-saving protocols are achieved mainly by shortening the duration of restriction enzyme digestion and ligation (Fig. 1B). Our assessment, however, revealed unsaturated reaction within the shortened time frames used in the commercial kits (Fig. 6), which was accompanied by an unfavourable composition of read pairs (Supplementary Table S4). Our attempt to insert a step of T4 DNA polymerase treatment in the sample preparation of the Arima kit protocol resulted in reduced "dangling end" reads (Library e vs f in Fig. 8). Regarding the Phase kit, transposase-based library preparation

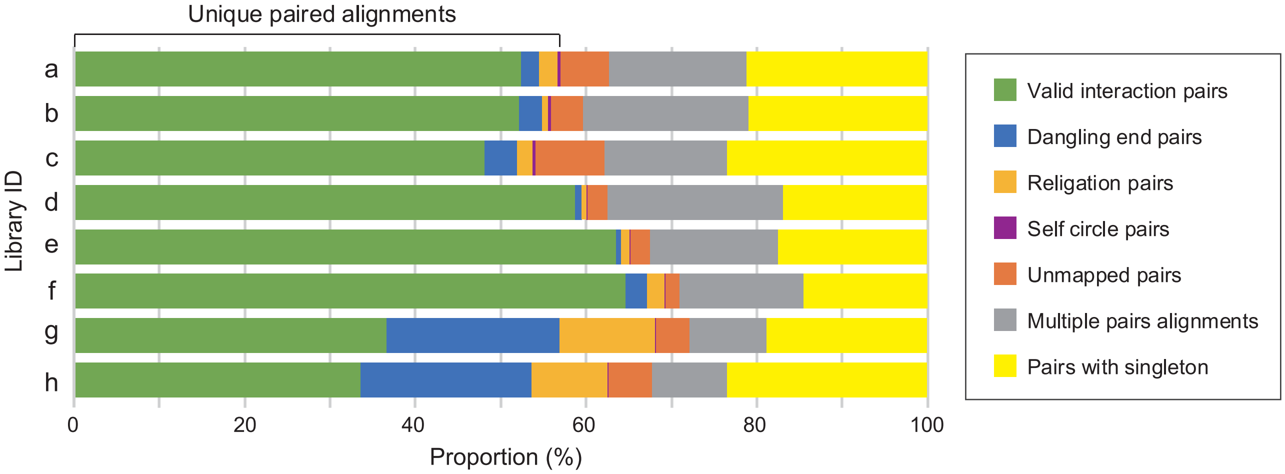


Figure 8: Results of the post-sequencing quality control with HiC-Pro. One million read pairs were used for computation with HiC-Pro. See Fig. 7A for the preparation conditions of Libraries a–h, Fig. 4 for the categorization, and Supplementary Table S5 for the actual proportion of the reads in each category. The post-sequencing quality control using variable read amounts (500,000–200,000,000 pairs) for one of these softshell turtle libraries (Supplementary Table S9) and human GM12878 libraries (Supplementary Table S2) shows the validity of this quality control with as few as 500,000 read pairs.

contributes largely to its shortened protocol, but this does not allow flexible control of library insert lengths. Recent protocols (versions 1.5 and 2.0) of the Phase kit instruct users to employ a greatly reduced DNA amount in the tagmentation reaction, which should mitigate the difficulty in controlling insert length but require excessive PCR amplification. The Arima and Phase kits assume that the QC of Hi-C DNA is based on the yield, and not the size, of DNA (see Fig. 1B). Nevertheless, QC based on DNA size (equivalent to QC1 in iconHi-C) is feasible by taking aliquots at each step of sample preparation. In particular, if preparing a small number of samples for Hi-C, as practised typically for genome scaffolding, one should opt to consider these points, even when using commercial kits, to improve the quality of the prepared libraries and scaffolding products.

Considerations regarding sequencing

The quantity of Hi-C read pairs to be input for scaffolding is critical because it accounts for the majority of the cost of Hi-C scaffolding. Our protocol introduces a thorough safety system to prevent sequencing unsuccessful libraries, first by performing pre-sequencing QCs for size shift analyses (Fig. 3) and second via small-scale (down to 500,000 read pairs) sequencing (see Results; also see Supplementary Tables S2 and S9).

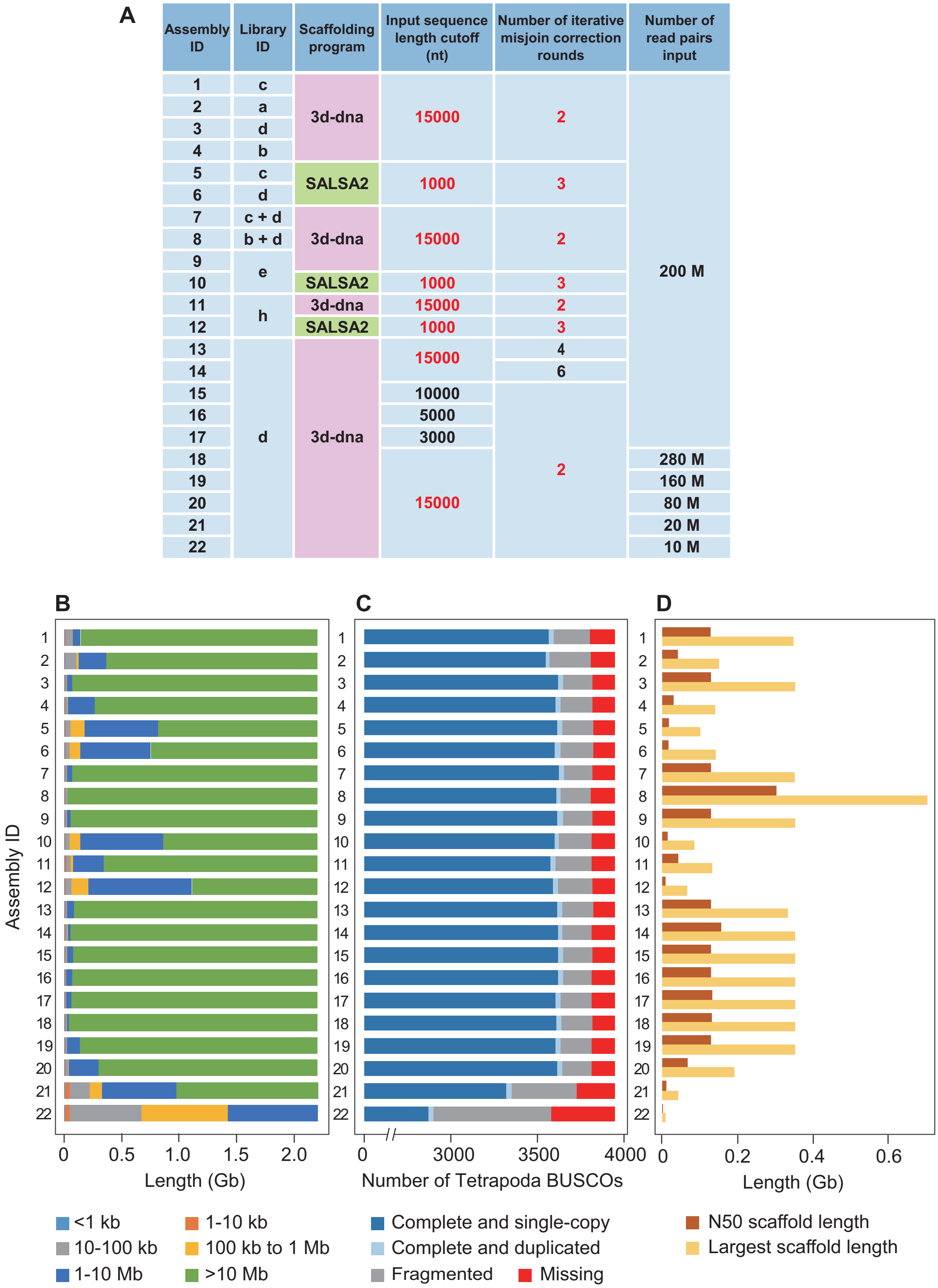
Our comparison showed a dramatic decrease in assembly quality in cases in which <100,000,000 read pairs were used (see the comparison of Assembly 18–22 described above; Fig. 9; also see [32]). Nevertheless, we obtained optimal results with a smaller number of reads (~160,000,000 per 2.2 Gb of genome) than that recommended by the manufacturers of commercial kits (e.g., 100,000,000 per 1 Gb of genome for the Dovetail Hi-C kit and 200,000,000 per 1 Gb of genome for the Arima kit). As generally and repeatedly discussed [32], the proportion of informative reads and their diversity, rather than just the overall number of obtained reads, are critical.

In terms of read length, we did not perform any comparisons in this study. Longer reads may enhance the fidelity of the characterization of the read pair properties and allow precise QC. Nevertheless, the existing Illumina sequencing platform has enabled the less expensive acquisition of 150 nt-long paired-end reads, which did not prompt us to vary the read length.

Considerations regarding computation

In this study, 3d-dna produced a more reliable scaffolding output than did SALSA2, whether sample preparation used a single or multiple enzyme(s) (Fig. 9B–D). On the other hand, 3d-dna required a longer time for the completion of scaffolding than did SALSA2. Apart from the choice of program, several points should be considered if successful scaffolding for a smaller investment is to be achieved. In general, Hi-C scaffolding results should not be taken for granted, and it is necessary to improve them by referring to contact maps using an interactive tool, such as Juicebox [17]. In this study, however, we compared raw scaffolding output to evaluate sample preparation and reproducible computational steps.

We used various parameters of the scaffolding programs (Fig. 9A). First, the Hi-C scaffolding programs that are available currently have different default length cut-off values for input sequences (e.g., 15 kb for the "-i" parameter in 3d-dna and 1 kb for the "-c" parameter in SALSA2). Only sequences that are longer than the cut-off length value contribute to sequence scaffolding towards chromosome sizes, while sequences shorter than the cut-off length are implicitly excluded from the scaffolding process and remain unchanged. Typically, when using the Illumina sequencing platform, genomic regions with unusually high frequencies of repetitive elements and GC content are not assembled into sequences with a sufficient length (see [33]). Such genomic regions tend to be excluded from chromosome-scale Hi-C scaffolds because their length is smaller than the threshold. Alternatively, these regions may be excluded because few Hi-C read pairs are mapped to them, even if they exceed the cut-off length. The deliberate setting of a cut-off length is recommended if particular sequences with relatively small lengths are the target of scaffolding. It should be noted that lowering the length threshold can result in frequent misjoins



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				
7	c + d	3d-dna	15000	2	
8	b + d				
9	e				
10		SALSA2	1000	3	
11	h	3d-dna	15000	2	
12		SALSA2	1000	3	
13	d	3d-dna	15000	4	
14				6	
15			10000	2	
16			5000		
17			3000		
18			15000		280 M
19					160 M
20					80 M
21					20 M
22					10 M

Figure 9: Comparison of Hi-C scaffolding products. (A) Scaffolding conditions used to produce Assembly 1–22. The default parameters are shown in red. (B) Scaffold length distributions. (C) Gene space completeness. (D) Largest and N50 scaffold lengths. See panel A for Library IDs and Supplementary Table S6 for raw values of the metrics shown in B–D. nt: nucleotides.

in the scaffolding output (Fig. 9B–D) or in overly long computational times. Regarding the number of iterative misjoin correction rounds (the "-r" parameter in 3d-dna and "-i" parameter in SALSA2), our attempts at using increased values did not necessarily yield favourable results (Fig. 9B–D). This did not provide a consistent optimal range of values but rather suggests the importance of performing multiple scaffolding runs with varying parameters.

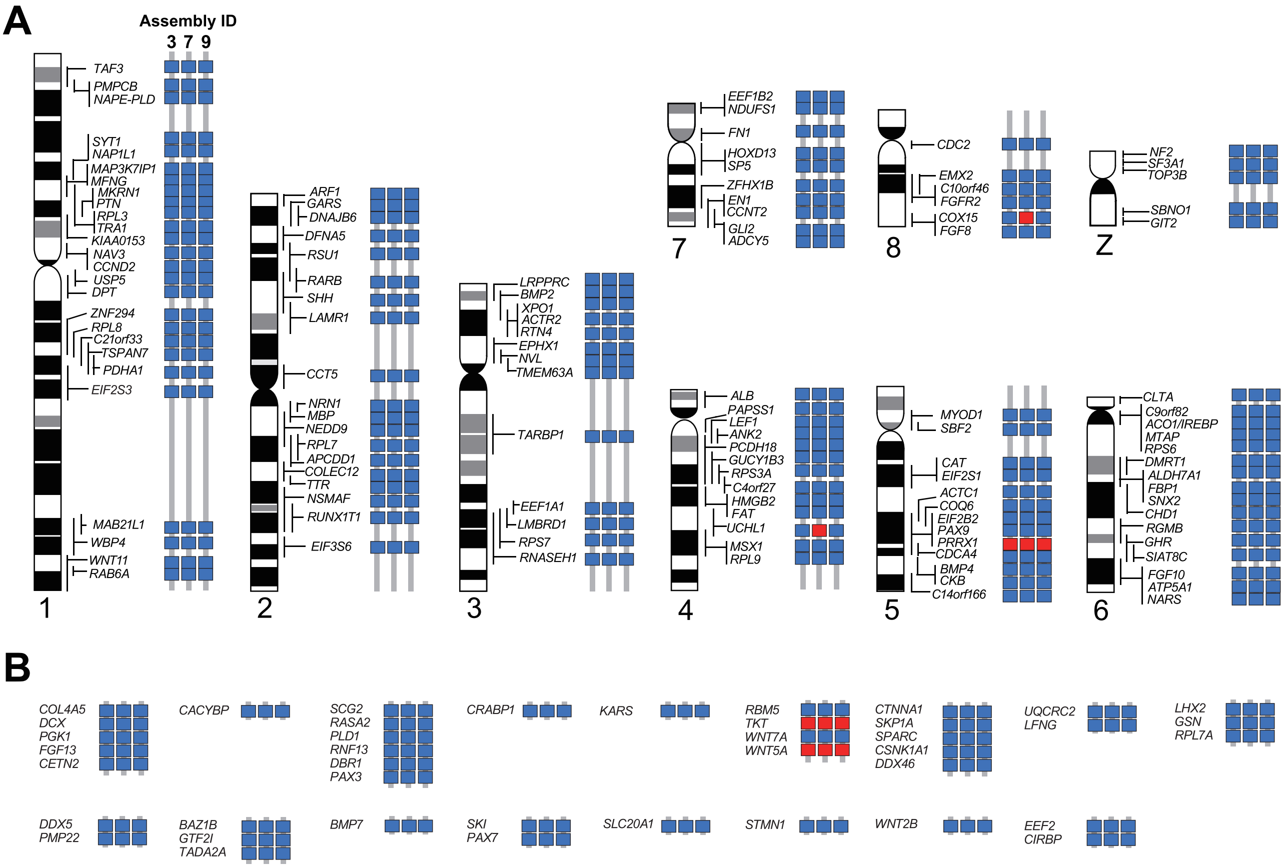


Figure 10: Cytogenetic validation of Hi-C scaffolding results. For the scaffolded sequences of Assembly 3, 7, and 9, we evaluated the consistency of the positions of the selected genes that were previously localized on 8 macrochromosomes and Z chromosome (A) and microchromosomes (B) by chromosome FISH [21–25] (see Results). Concordant and discordant gene locations on individual assemblies are indicated with blue and red boxes, respectively. The arrays of genes without idiograms in B were identified on chromosomes that are cytogenetically indistinguishable from each other.

Considerations regarding the assessment of chromosome-scale genome sequences

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

For further evaluation of our scaffolding results, we referred to the sequence length distributions of the genome assemblies of other turtle species that are regarded as being chromosome-scale data. This analysis yielded values of the basic metrics that were comparable to those of our Hi-C scaffolds of the softshell turtle, i.e., an N50 length of 127.5 Mb and a maximum sequence length of 344.5 Mb for the genome assembly of the green sea turtle (*Chelonia mydas*) released by the DNA Zoo Project [17] and an N50 length of 131.6 Mb and a maximum length of 370.3 Mb for the genome assembly of the Goode's thornscrub tortoise (*Gopherus evgoodei*) released by the Vertebrate Genome Project [16]. Scaffolding results should be evaluated by referring to the estimated N50 length and the maximum length based on the actual value and to the length distribution of chromosomes in the intrinsic karyotype of the species in question, or of its close relative. Turtles tend to have an N50 length of ∼130 Mb and a maximum length of 350 Mb, while many teleost fish genomes exhibit an N50 length as low as 20–30 Mb and a maximum length of <100 Mb [34]. If these values are excessive, the scaffolded sequences harbour overassembly, which erroneously boosts length-based metrics. Thus, higher values, which are conventionally regarded as signs of successful sequence assembly, do not necessarily indicate higher precision.

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 effect on the total length of assembled sequences.

Conclusions

In this study, we introduced the iconHi-C protocol, which implements successive QC steps. We also assessed potential key factors for improving Hi-C scaffolding. Overall, our study showed that small variations in sample preparation or computation for scaffolding can have a large effect on scaffolding output, and that any scaffolding output should ideally be validated using independent information, such as cytogenetic data, long reads, or genetic linkage maps. The present study aimed to evaluate the output of reproducible computational steps, which in practice should be followed by the modification of the raw scaffolding output by referring to independent information or by analysing chromatin contact maps. The study used limited combinations of species, sample preparation methods, scaffolding programs, and their parameters, and we will continue to test different conditions for kits/programs that did not necessarily perform well here using our specific materials.

Methods

Initial genome assembly sequences

The Chinese softshell turtle (*Pelodiscus sinensis*) assembly published previously [26] was downloaded from NCBI GenBank (GCA_000230535.1), whose gene space completeness and length statistics were assessed by gVolante [35] (see Supplementary Table S1 for the assessment results). Although it could be suggested to remove haplotigs before Hi-C scaffolding [36], we omitted this step because of the low frequency of reference orthologues with multiple copies (0.72%; Supplementary Table S1), indicating a minimal degree of haplotig contamination.

Animals and cells

We sampled tissues (liver and blood cells) from a female purchased from a local farmer in Japan because the previous whole-genome sequencing used the whole blood of a female [26]. All experiments were conducted in accordance with the Guideline of the Institutional Animal Care and Use Committee of RIKEN Kobe Branch (Approval ID: A2017–12).

The human lymphoblastoid cell line GM12878 (Coriell Cat# GM12878, RRID:CVCL_7526) was purchased from the Coriell Cell Repositories and cultured in RPMI-1640 medium (Thermo Fisher Scientific, Waltham, MA) supplemented with 15% fetal bovine serum, 2 mM L-glutamine, and a 1× antibiotic-antimycotic solution (Thermo Fisher Scientific), at 37 °C, 5% CO_2, as described previously [37].

Hi-C sample preparation using the original protocol

We have made modifications to the protocols that are available in the literature [3, 8, 9] (Fig. 1B). The full version of our "inexpensive and controllable Hi-C (iconHi-C)" protocol is described in Supplementary Protocol S1 and available at Protocols.io [38].

Hi-C sample preparation using commercial kits

The Proximo Hi-C Kit (Phase Genomics, Seattle, WA) which uses the restriction enzyme Sau3A1 and transposase-based library preparation [39] (Fig. 1B) was used to prepare a library from 50 mg of the softshell turtle liver according to the official ver. 1.0 animal protocol provided by the manufacturer (Library g in Fig. 7A) and a library from 10 mg of the liver that was amplified with a reduced number of PCR cycles based on a preliminary real-time qPCR using an aliquot (Library h; see [31] for the details of the pre-determination of the optimal number of PCR cycles). The Arima-HiC Kit (Arima Genomics, San Diego, CA), which employs a restriction enzyme cocktail (Fig. 1B), was used in conjunction with the KAPA Hyper Prep Kit (KAPA Biosystems, Cape Town, South Africa), protocol ver. A160108 v00, to prepare a library using the softshell turtle liver, according to its official animal vertebrate tissue protocol (ver. A160107 v00) (Library f) and a library with an additional step of T4 DNA polymerase treatment for reducing "dangling end" reads (Library e). This additional treatment is detailed in Step 8.2 (for DpnII-digested samples) of Supplementary Protocol S1.

DNA sequencing

Small-scale sequencing for library QC (QC3) was performed in-house to obtain 127 nt-long paired-end reads on a HiSeq 1500 (Illumina, San Diego, CA) in the Rapid Run Mode. For evaluating the effects of variable duration of the restriction digestion and ligation reactions, sequencing was performed on a MiSeq (Illumina) using the MiSeq Reagent Kit v3 to obtain 300 nt-long paired-end reads. Large-scale sequencing for Hi-C scaffolding was performed to obtain 151 nt-long paired-end reads on a HiSeq X (Illumina). The obtained reads underwent QC using FastQC ver. 0.11.5 (FastQC, RRID:SCR_014583; [40]), and low-quality regions and adapter sequences in the reads were removed using Trim Galore ver. 0.4.5 (TrimGalore, RRID:SCR_011847; [41]) with the parameters "-e 0.1 -q 30."

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

For post-sequencing library QC, 1,000,000 trimmed read pairs for each Hi-C library were sampled using the "subseq" function of the program seqtk ver. 1.2-r94 [42]. The resultant sets of read pairs were processed using HiC-Pro ver. 2.11.1 [28] with bowtie2 ver. 2.3.4.1 [43] to evaluate the insert structure and mapping status onto the softshell turtle genome assembly PelSin_1.0 (GCF_000230535.1) or the human genome assembly hg19. This resulted in categorization as valid interaction pairs and invalid pairs, with the latter being divided further into "dangling end," "religation," "self circle," and "single-end" pairs (Fig. 4). To process the read pairs derived from the libraries prepared using either HindIII or DpnII (Sau3AI) with the iconHi-C protocol (Libraries a–d) and the Phase kit (Libraries g and h), the restriction fragment file required by HiC-Pro was prepared according to the script "digest_genome.py" of HiC-Pro. To process the reads derived from the Arima kit (Libraries e and f), all restriction sites ("GATC" and "GANTC") were inserted into the script. In addition, the nucleotide sequences of all possible ligated sites generated by restriction enzymes were included in a configuration file of HiC-Pro. The details of this procedure and the sample code used are included in Supplementary Protocol S2.

Computation for Hi-C scaffolding

To control our comparison with intended input data sizes, a certain number of trimmed read pairs were sampled for each library with seqtk, as described above. Scaffolding was processed with the following methods using 2 program pipelines, 3d-dna and SALSA2.

Scaffolding via 3d-dna was performed using Hi-C read mapping onto the genome with Juicer ver. 20180805 (Juicer, RRID:SCR_017226) [19] using the default parameters with BWA ver.0.7.17-r1188 (BWA, RRID:SCR_010910) [44]. The restriction fragment file required by Juicer was prepared by the script "generate_site_positions.py" script of Juicer. By converting the restriction fragment file of HiC-Pro to the Juicer format, an original script that was compatible with multiple restriction enzymes was prepared (Supplementary Protocol S2). Scaffolding via 3d-dna ver. 20180929 was performed using variable parameters (see Fig. 9A).

Scaffolding via SALSA2 using Hi-C reads was preceded by Hi-C read pair processing with the Arima mapping pipeline ver. 20181207 [45] together with BWA, SAMtools ver. 1.8–21-gf6f50ac (SAMTOOLS, RRID:SCR_002105) [46], and Picard ver. 2.18.12 (Picard, RRID:SCR_006525) [47]. The mapping result in the binary alignment map (bam) format was converted into a BED file by bamToBed of Bedtools ver. 2.26.0 (BEDTools, RRID:SCR_006646) [48], the output of which was used as the input of scaffolding using SALSA2 ver. 20181212 with the default parameters.

Completeness assessment of Hi-C scaffolds

gVolante ver. 1.2.1 [35] was used to perform an assessment of the sequence length distribution and gene space completeness based on the coverage of 1-to-1 reference orthologues with BUSCO v2/v3 employing the 1-to-1 orthologue set "Tetrapoda" supplied with BUSCO (BUSCO, RRID:SCR_015008) [49]. No cut-off length was used in this assessment.

Continuity assessment using RNA-sequencing read mapping

Paired-end reads obtained by RNA sequencing of softshell turtle embryos at multiple stages were downloaded from NCBI SRA (DRX001576) and were assembled using Trinity ver. 2.7.0 (Trinity, RRID:SCR_013048) [50] with default parameters. The assembled transcript sequences were mapped to the Hi-C scaffold sequences with pblat [51], and the output was assessed with isoblat ver. 0.31 [52].

Comparison with chromosome FISH results

Cytogenetic validation of Hi-C scaffolding results was performed by comparing the gene locations on the scaffold sequences with those provided by previous chromosome FISH for 162 protein-coding genes [21–25]. The nucleotide exonic sequences for those 162 genes were retrieved from GenBank and aligned with Hi-C scaffold sequences using BLAT ver. 36x2 (BLAT, RRID:SCR_011919) [53], followed by the analysis of their positions and orientation along the Hi-C scaffold sequences.

Availability of Supporting Data and Materials

All sequence data generated in this study have been submitted to the DDBJ Sequence Read Archive (DRA) under accession IDs DRA008313 and DRA008947. The datasets supporting the results of this article are available in FigShare [54] and the *GigaScience* GigaDB database [55].

Additional Files

Supplementary Figure S1. DNA size distribution of the softshell turtle Hi-C libraries.
Supplementary Figure S2. Pre-sequencing quality control of softshell turtle blood Hi-C libraries (Libraries a and b).
Supplementary Figure S3. Pre-sequencing quality control (QC2) of the Hi-C libraries generated using the Phase kit (Libraries g and h).
Supplementary Figure S4. Structural analysis of the possibly chimeric scaffold in Assembly 8.
Supplementary Figure S5. Hi-C contact maps for selected softshell turtle Hi-C scaffolds.
Supplementary Figure S6. Pairwise alignment of Hi-C scaffolds.
Supplementary Table S1. Statistics of the Chinese softshell turtle draft genome assembly before Hi-C.
Supplementary Table S2. HiC-Pro results for the human GM12878 HindIII Hi-C library with reduced reads.
Supplementary Table S3. Quality control of the human GM12878 Hi-C libraries.
Supplementary Table S4. Effect of the duration of restriction enzyme digestion and ligation.
Supplementary Table S5. Quality control of Hi-C libraries.
Supplementary Table S6. Scaffolding results with variable input data and computational parameters.
Supplementary Table S7. Mapping results of assembled transcript sequences onto Hi-C scaffolds.
Supplementary Table S8. Effect of variable degrees of PCR amplification.
Supplementary Table S9. HiC-Pro results for the softshell turtle liver libraries (Libraries d, e, and h) with reduced reads.
Supplementary Protocol S1. iconHi-C protocol.
Supplementary Protocol S2. Computational protocol to support the use of multiple enzymes.

Abbreviations

3C: chromosome conformation capture; BLAT: BLAST-like alignment tool; bp: base pairs; BUSCO: Benchmarking Universal Single-Copy Orthologs; BWA: Burrows-Wheeler Aligner; FISH: fluorescence *in situ* hybridization; Gb: gigabase pairs; GC: guanine-cytosine; Mb: megabase pairs; NCBI: National Center for Biotechnology Information; NGS: next-generation sequencing; QC: quality control; SRA: Sequence Read Archive.

Funding

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

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

S.K., I.H., H.M., and M.K. conceived the study. M.K. and K.T. performed laboratory work, and O.N. performed bioinformatic analysis. M.K., O.N., and H.M. analysed the data. S.K., M.K., and O.N. drafted the manuscript. All authors contributed to the finalization of the manuscript.

Acknowledgements

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

References

1. Rowley MJ, Corces VG. Organizational principles of 3D genome architecture. Nat Rev Genet 2018;**19**(12):789–800.
2. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009;**326**(5950):289–93.
3. Rao SS, Huntley MH, Durand NC, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 2014;**159**(7):1665–80.
4. Burton JN, Adey A, Patwardhan RP, et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol 2013;**31**(12):1119–25.
5. Marie-Nelly H, Marbouty M, Cournac A, et al. High-quality genome (re)assembly using chromosomal contact data. Nat Commun 2014;**5**(1):5695.
6. Kaplan N, Dekker J. High-throughput genome scaffolding from in vivo DNA interaction frequency. Nat Biotechnol 2013;**31**(12):1143–7.
7. Sofueva S, Yaffe E, Chan WC, et al. Cohesin-mediated interactions organize chromosomal domain architecture. EMBO J 2013;**32**(24):3119–29.
8. Ikeda T, Hikichi T, Miura H, et al. Srf destabilizes cellular identity by suppressing cell-type-specific gene expression programs. Nat Commun 2018;**9**(1):1387.
9. Sedlazeck FJ, Lee H, Darby CA, et al. Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. Nat Rev Genet 2018;**19**(6):329–46.
10. Putnam NH, O'Connell BL, Stites JC, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res 2016;**26**(3):342–50.
11. Ghurye J, Pop M, Koren S, et al. Scaffolding of long read assemblies using long range contact information. BMC Genomics 2017;**18**(1):527.
12. Ghurye J, Rhie A, Walenz BP, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol 2019;**15**(8):e1007273.
13. Dudchenko O, Batra SS, Omer AD, et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 2017;**356**(6333):92–5.
14. Ghurye J, Pop M. Modern technologies and algorithms for scaffolding assembled genomes. PLoS Comput Biol 2019;**15**(6):e1006994.
15. Lewin HA, Robinson GE, Kress WJ, et al. Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci U S A 2018;**115**(17):4325–33.
16. Koepfli KP, Paten B, Genome KCoS, et al. The Genome 10 K Project: A way forward. Annu Rev Anim Biosci 2015;**3**:57–111.
17. Dudchenko O, Shamim MS, Batra SS, et al. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv 2018:254797.
18. Belaghzal H, Dekker J, Gibcus JH. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. Methods 2017;**123**: 56–65.
19. Kuratani S, Kuraku S, Nagashima H. Evolutionary developmental perspective for the origin of turtles: The folding theory for the shell based on the developmental nature of the carapacial ridge. Evol Dev 2011;**13**(1):1–14.
20. Durand NC, Shamim MS, Machol I, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst 2016;**3**(1):95–8.
21. Matsuda Y, Nishida-Umehara C, Tarui H, et al. Highly conserved linkage homology between birds and turtles: Bird and turtle chromosomes are precise counterparts of each other. Chromosome Res 2005;**13**(6):601–15.
22. Kuraku S, Ishijima J, Nishida-Umehara C, et al. cDNA-based gene mapping and GC3 profiling in the soft-shelled turtle suggest a chromosomal size-dependent GC bias shared by sauropsids. Chromosome Res 2006;**14**(2):187–202.
23. Uno Y, Nishida C, Tarui H, et al. Inference of the protokaryotypes of amniotes and tetrapods and the evolutionary processes of microchromosomes from comparative gene mapping. PLoS One 2012;**7**(12):e53027.
24. Kawai A, Nishida-Umehara C, Ishijima J, et al. Different origins of bird and reptile sex chromosomes inferred from comparative mapping of chicken Z-linked genes. Cytogenet Genome Res 2007;**117**(1–4):92–102.
25. Kawagoshi T, Uno Y, Matsubara K, et al. The ZW micro-sex chromosomes of the Chinese soft-shelled turtle (*Pelodiscus sinensis*, Trionychidae, Testudines) have the same origin as chicken chromosome 15. Cytogenet Genome Res 2009;**125**(2):125–31.
26. Wang Z, Pascual-Anaya J, Zadissa A, et al. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nat Genet 2013;**45**(6):701–6.
27. Belton JM, McCord RP, Gibcus JH, et al. Hi-C: A comprehensive technique to capture the conformation of genomes. Methods 2012;**58**(3):268–76.
28. Servant N, Varoquaux N, Lajoie BR, et al. HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. Genome Biol 2015;**16**:259.
29. Imakaev M, Fudenberg G, McCord RP, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods 2012;**9**(10):999–1003.
30. Mayjonade B, Gouzy J, Donnadieu C, et al. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. BioTechniques 2016;**61**(4):203–5.
31. Tanegashima C, Nishimura O, Motone F, et al. Embryonic transcriptome sequencing of the ocellate spot skate *Okamejei kenojei*. Sci Data 2018;**5**:180200.
32. DeMaere MZ, Darling AE. bin3C: Exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. Genome Biol 2019;**20**(1):46.
33. Botero-Castro F, Figuet E, Tilak MK, et al. Avian genomes revisited: Hidden genes uncovered and the rates versus traits paradox in birds. Mol Biol Evol 2017;**34**(12):3123–31.
34. Hotaling S, Kelley JL. The rising tide of high-quality genomic resources. Mol Ecol Resour 2019;**19**(3):567–9.
35. Nishimura O, Hara Y, Kuraku S. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. Bioinformatics 2017;**33**(22):3635–7.
36. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 2018;**19**(1):460.
37. Kadota M, Hara Y, Tanaka K, et al. CTCF binding landscape in jawless fish with reference to Hox cluster evolution. Sci Rep 2017;**7**(1):4957.
38. Kadota M, Nishimura O, Miura H, et al. iconHi-C Protocol (ver. 1.0). protocols.io. 2019. http://dx.doi.org/10.17504/protocols.io.4mjgu4n.
39. Adey A, Morrison HG, Asan, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol 2010;**11**(12): R119.

40. FastQC: A quality control tool for high throughput sequence data. Babraham Bioinformatics. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 20 July 2018.
41. Trim Galore: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. Babraham Bioinformatics. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. Accessed 21 July 2018.
42. seqtk: Toolkit for processing sequences in FASTA/Q formats. https://github.com/lh3/seqtk. Accessed 13 November 2018.
43. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;**9**(4):357–9.
44. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;**25**(14):1754–60.
45. Arima Hi-C mapping pipeline. https://github.com/ArimaGenomics/mapping_pipeline. Accessed 7 December 2018.
46. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011;**27**(21):2987–93.
47. Picard: A set of command line tools for manipulating high-throughput sequencing data and formats such as SAM/BAM/CRAM and VCF. Broad Institute. http://broadinstitute.github.io/picard/. Accessed 7 December 2018.
48. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 2010;**26**(6):841–2.
49. Simao FA, Waterhouse RM, Ioannidis P, et al. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;**31**(19):3210–2.
50. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 2011;**29**(7):644–52.
51. Wang M, Kong L. pblat: A multithread blat algorithm speeding up aligning sequences to genomes. BMC Bioinformatics 2019;**20**(1):28.
52. Ryan JF. Baa.pl: A tool to evaluate de novo genome assemblies with RNA transcripts. arXiv 2013:1309.2087.
53. Kent WJ. BLAT–the BLAST-like alignment tool. Genome Res 2002;**12**(4):656–64.
54. Kadota M, Nishimura O, Miura H, et al. Softshell turtle genome assemblies scaffolded with Hi-C data. Figshare 2019, doi:10.6084/m9.figshare.8024858.v2.
55. Kadota M, Nishimura O, Miura H, et al. Supporting data for "Multifaceted Hi-C benchmarking: What makes a difference in chromosome-scale genome scaffolding?" GigaScience Database 2019, http://dx.doi.org/10.5524/100675.