Supplemental Information for:

Technical considerations in Hi-C scaffolding and evaluation of chromosome-scale genome assemblies

Kazuaki Yamaguchi, Mitsutaka Kadota, Osamu Nishimura, Yuta Ohishi, Yuki Naito, Shigehiro Kuraku

Supplemental Methods

Sample source
Adult P. picta was provided by the Laboratory for Animal Resources and Genetic Engineering, RIKEN BDR. Animal breeding and experiments were conducted in accordance with the guidelines of the Institutional Animal Care and Use Committee (IACUC), RIKEN Kobe Branch.

Sequencing and assembly using short shotgun reads
To obtain a new Illumina shotgun sequencing library of short reads, 500 ng of genomic DNA (gDNA) was sheared using an E220 Focused-ultrasonicator (Covaris), to obtain DNA fragments of variable length distributions. The gDNA was extracted from a previously reported individual (Hara et al., 2018). The sheared DNA was denatured at 90°C for 5 min, size-selected by Agencourt AMPure XP (Beckman Coulter), and enriched for GC-rich regions of gDNA (Tilak, Botero-Castro, Galtier, & Nabholz, 2018). The GC-enriched DNA was used for paired-end library preparation with a KAPA LTP Library Preparation Kit (KAPA Biosystems). The optimal number of PCR cycles for the library was determined using a Real-Time Library Amplification Kit (KAPA Biosystems) by preliminary qPCR-based quantification using an aliquot of adaptor-ligated DNAs. To amplify and enrich the library, 6 cycles of PCR amplification were performed. Small molecules in the prepared libraries were removed by size selection using Agencourt AMPure XP. The library was sequenced on a HiSeq 1500 platform (Illumina), operated...
by the HiSeq Control Software v2.0.12.0 using a HiSeq SR Rapid Cluster Kit v2 (Illumina) and HiSeq Rapid SBS Kit v2 (Illumina), to obtain 127-nt-long paired-end reads (PE127). Base calling was performed with RTA v1.17.21.3, and the fastq files were generated by bcl2fastq v1.8.4 (Illumina). Registration information for the library is given in Table S1. Removal of low-quality bases from paired-end reads was processed by TrimGalore v0.3.3 with the options “--stringency 2 --quality 20 --length 25 --paired --retain_unpaired”. Short-read assembly of the Madagascar ground gecko (Paroedura picta) was performed using Platanus v1.2.4, as described previously (Hara et al., 2018), using the newly obtained GC-rich reads and previously reported reads (Hara et al., 2018) (Table S1). These reads were all derived from a single individual of an unknown sex, and the total sequencing coverage amounted to 77.7-fold the genome size quantified using flow cytometry (1.84 pg). Scaffolding with mate-pair reads and gap closure were also performed using the program Platanus v1.2.4. The resulting genome assembly is designated as Assembly 13 in this article (see Figure S2).

**Sequencing and assembly using Chromium linked-reads**

Ultra high molecular weight gDNA was extracted from female blood for the preparation of the Chromium linked-read library using the CHEF Mammalian Genomic DNA Plug Kit (BioRad, Cat. No. #1703591). Next, 7.2 ng of the gDNA was used for the preparation of the Chromium genome library. The library was constructed according to the user guide of the Chromium Genome Library Kit v2 Chemistry using the Chromium Genome Library Kit & Gel Bead Kit v2 (10x Genomics, Cat. No. #120258) and the Chromium Genome Chip Kit v2 (10x Genomics, Cat. No. #120257). The library was sequenced on a HiSeq X (Illumina) platform to obtain 151 nt-long paired-end reads (PE151). Registration information for the library is given in Table S1. De novo assembly was performed by the program Supernova v2.0 using Chromium linked-reads, with the input read number set to 336 million. The sequencing coverage was 56 times the genome size. The resulting genome assembly is designated as Assembly 7 in this article (see Figure S2).

**Scaffolding using RNA-seq reads and Chromium linked-reads**

The genome assembly sequences generated by the program Platanus were used as a reference for RNA-seq read mapping using the program HISAT2 v2.1.0 (Figure S2). The mapping output was used in the scaffolding with the program P_RNA_scaffolder
The RNA-seq library used in this step was previously reported (Hara et al., 2015) (Table S2). The output of the scaffolding using RNA-seq reads mentioned above was used as a reference in mapping Chromium linked-reads using the program BWA v0.7.17. Based on the linked-read mapping information, misconnections in the sequences scaffolded by the programs Platanus and P_RNA_scaffolder were corrected by the program Tigmint v1.1.2 (Jackman et al., 2018). The output sequences were scaffolded by the program ARKS v1.1.0 with the linked-read mapping information (Coombe et al., 2018). The resulting genome assembly is designated as Assembly 1 in this article (see Figure S1).

**Hi-C library preparation and sequencing**
A Hi-C library was constructed as previously reported (Kadota et al., 2020) using the restriction enzyme HindIII. A whole embryo of *P. picta* at stage 28 that had been kept frozen at –80°C after dissection and after snap freezing in liquid nitrogen was cryopulverized and fixed in 1% formaldehyde solution. Fixed tissue containing 10 μg of DNA was used for the preparation of Hi-C DNA via *in situ* restriction digestion and ligation. The Hi-C library was constructed from 2 μg of post-ligated DNA with 5 cycles of PCR amplification. Quality controls of the post-ligated DNA (QC1) and the Hi-C library (QC2) were performed as described previously (Kadota et al., 2020). In each of these QCs, the expected shift of DNA length distributions was confirmed in the samples (Figure S1). The prepared Hi-C library was sequenced with PE127 on a HiSeq 1500 platform, and 101 million read pairs were obtained (Table S1).

**Hi-C scaffolding**
Each of the three genome assemblies (Assemblies 1, 7, and 13 in Figure S2) was used as a reference for Hi-C read mapping and chromosome-scale scaffolding, which were performed using the program 3d-dna (v180922) (Dudchenko et al., 2017). In the scaffolding, five different input sequence length thresholds were applied (Figure S2). The resulting 15 scaffolding outputs, as well as the original assemblies (before scaffolding), were assessed based on sequence length distribution (Figure S3) and protein-coding gene completeness. The contact map of the Hi-C reads link was checked to the plausible assembly, and some unnatural contacts were manually curated using Juicebox v1.11.08 (Durand et al., 2016). For Assembly 6, which was judged to be the best among all assemblies, the “review” of the scaffolding results was
performed with the chromatin contact map on Juicebox by referring to the two following types of independent information: 1) gene mapping results with FISH for *Gekko hokouensis* (Srikulnath, Uno, Nishida, Ota, & Matsuda, 2015) and 2) nucleotide sequence similarity between different scaffolding outputs visualized by SyMAP (Soderlund, Bomhoff, & Nelson, 2011). After Assembly 6 was chosen as the optimal output, contaminated sequences were removed from this genome assembly, for public release, as reported previously (Hara et al., 2018). The resulting sequences were deposited in DDBJ/NCBI under the BioProject PRJDB5392.

**Assessment of BUSCO results**

The telomere-to-telomere human genome assembly CHM13 v1.0 ([https://github.com/nanopore-wgs-consortium/chm13](https://github.com/nanopore-wgs-consortium/chm13)) was assessed using the version (5.0.0) of BUSCO (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) and the Tetrapoda ortholog set on the gVolante webserver (Nishimura, Hara, & Kuraku, 2019). Based on the output of BUSCO, we identified OrthoDB entries for the 79 genes that were recognized as “missing” by BUSCO, as well as their human ortholog sequences in NCBI RefSeq (as of April 5, 2021). For each of these 79 cases, the presence of a continuous nucleotide sequence harboring a complete intact open reading frame was scanned with BLASTN 2.5.0+ or TBLASTN 2.5.0+ (Altschul et al., 1997) in the CHM13 genome assembly, and manually confirmed later (Table S3).
Figure S1. Pre-sequencing quality control of the Madagascar ground gecko Hi-C library. (a) Quality control of the Hi-C DNA (QC1) comparing pre-digestion, digested, and ligated DNA (Hi-C DNA) with a TapeStation Genomic DNA ScreenTape (Agilent Technologies). (b) Quality control of the Hi-C library (QC2) comparing library DNA with and without NheI restriction enzyme digestion with TapeStation High Sensitivity D1000 ScreenTape (Agilent Technologies).
**Figure S2. Madagascar ground gecko genome assemblies.** The bold numbers 1 to 18 are the identifiers of the constructed genome assemblies. Assembly 1 was constructed using the short-read assembly program Platanus (Kajitani et al., 2014), the transcript-based scaffold P_RNA_Scaffold (Zhu et al., 2018), and ARKS (Coombe et al., 2018), a scaffold that processes 10x Genomics Chromium linked-reads and was used as the input assembly for Hi-C scaffolding with the program 3d-dna, with five different input sequence length thresholds (1,000, 3,000, 5,000, 10,000, and 15,000 bp), to produce Assemblies 2–6. Assembly 7 was initially constructed with Supernova, a program for processing 10x Genomics Chromium linked-reads, and was later scaffolded with 3d-dna, to produce Assemblies 8–12. Assembly 13 was initially constructed by Platanus (Hara et al., 2018) and was later scaffolded with 3d-dna, to produce Assemblies 14–18. See Supplemental Methods for details of the assembly and scaffolding processes.
Figure S3. Basic sequence length statistics of Hi-C scaffolding output for the Madagascar ground gecko genome. (a) Length distribution of the scaffold sequences in the individual genome assemblies. The numbers 1–18 indicated on the left correspond to the identifiers of the genome assemblies shown in Figure S2. (b) Proportion of the sum length of scaffold sequences >1 Mbp compared with the total sequence length in the assembly. (c) N50 scaffold length of the individual genome assemblies. (d) Length of the largest scaffold in the individual genome assemblies.
## Supplemental Tables

### Table S1. Libraries prepared for Madagascar ground gecko genome sequencing

| Library ID | Library type                    | Accession ID                     | Starting DNA amount (ng) | Peak insert length (bp) | Number of PCR cycles | Read length (bp) |
|------------|---------------------------------|----------------------------------|--------------------------|-------------------------|----------------------|-----------------|
| P079_01_1’ | Paired-end shotgun              | DRR089867, DRR089870             | 1,000                    | 379                     | 0                    | 151, 171        |
| P079_01_2’ | Paired-end shotgun              | DRR089868, DRR089871             | 2,000                    | 481                     | 0                    | 151, 171        |
| P079_01_3’ | Paired-end shotgun              | DRR089869, DRR089872, DRR089873 | 4,000                    | 615                     | 0                    | 151, 171, 301   |
| P091_01_1’ | Mate-pair                       | DRR089874, DRR089875             | 4,000                    | 663                     | 10                   | 301, 171        |
| P101_01_1’ | Mate-pair                       | DRR089878, DRR089879             | 4,000                    | 487                     | 8                    | 301, 171        |
| P101_01_2’ | Mate-pair                       | DRR089880, DRR089881             | 4,000                    | 485                     | 10                   | 301, 171        |
| P101_01_3’ | Mate-pair                       | DRR089876, DRR089877             | 4,000                    | 494                     | 8                    | 301, 171        |
| P104_01_5’ | Mate-pair                       | DRR089882, DRR089883             | 12,000                   | 411                     | 10                   | 301, 171        |
| P377_04_1  | Paired-end shotgun (GC-rich)    | DRR288659                        | 500                      | 319                     | 6                    | 127             |
| P461_02_1  | Chromium linked-read            | DRR288658                        | 0.9**                    | 478                     | 10†                  | 151             |
| P312_02_1  | Hi-C                            | DRR288660                        | 2,000                    | 317                     | 5                    | 127             |

*Reported in a previous publication (Hara et al., 2018). **Amount of gDNA loaded onto the Chromium Genome Chip. †Number of PCR cycles using the Chromium i7 Sample index.
| Library ID  | Library type | Accession ID | Starting RNA amount (ng) | Peak insert length (bp) | Number of PCR cycles | Read length (bp) |
|------------|--------------|--------------|--------------------------|-------------------------|----------------------|-----------------|
| P061_01_2  | RNA-seq      | DRR047248    | 40                       | 226                     | 6                    | 171             |
| P061_01_1  | RNA-seq      | DRR047249    | 40                       | 542                     | 6                    | 171             |
| P076_01_2  | RNA-seq      | DRR047250    | 40                       | 542                     | 6                    | 251             |
| P076_01_1  | RNA-seq      | DRR047251    | 1,100                    | 243                     | 4                    | 151             |
| P080_01_1  | RNA-seq      | DRR047252    | 2,200                    | 662                     | 2                    | 151             |
| P080_03_1  | RNA-seq      | DRR047253    | 1,000                    | 365                     | 6                    | 151             |
| P080_02_1  | RNA-seq      | DRR047254    | 1,000                    | 376                     | 6                    | 151             |
Table S3. Manual assessment of automated completeness assessment results by BUSCO

| OrthoDB group ID | Human gene symbol (Gene ID) | Presence of the full ORF in the human genome |
|------------------|-----------------------------|----------------------------------------------|
| 89910at32523     | ABCG1 (9619)                | Yes                                          |
| 163158at32523    | ACAT1 (38)                  | Yes                                          |
| 146229at32523    | ADIPOR2 (79602)             | Yes                                          |
| 119446at32523    | ALG9 (79796)                | Yes                                          |
| 103667at32523    | ANKRD6 (22881)              | Yes                                          |
| 147451at32523    | ARSB (411)                  | Yes                                          |
| 37139at32523     | BOC (91653)                 | Yes                                          |
| 266661at32523    | C3orf38 (285237)            | Yes                                          |
| 267504at32523    | CACFD1 (11094)              | Yes                                          |
| 142624at32523    | CCDC13 (152206)             | Yes                                          |
| 84377at32523     | COL1A2 (1278)               | Yes                                          |
| 176420at32523    | COL3A1 (1281)               | Yes                                          |
| 129804at32523    | COL4A1 (1282)               | Yes                                          |
| 75528at32523     | COL4A2 (1284)               | Yes                                          |
| 127708at32523    | COL4A3 (1285)               | Yes                                          |
| 96920at32523     | COL4A4 (1286)               | Yes                                          |
| 68937at32523     | COL5A2 (1290)               | Yes                                          |
| 184179at32523    | COL9A1 (1297)               | Yes                                          |
| 242674at32523    | COL9A3 (1299)               | Yes                                          |
| 245810at32523    | COLQ (8292)                 | Yes                                          |
| 170550at32523    | CPB2 (1361)                 | Yes                                          |
| 255236at32523    | CPSF7 (79869)               | Yes                                          |
| 280970at32523    | CSRP1 (1465)                | Yes                                          |
| 290134at32523    | CTNNBIP1 (56998)            | Yes                                          |
| 62739at32523     | DLL1 (28514)                | Yes                                          |
| 200948at32523    | DUS4L (11062)               | Yes                                          |
| 79372at32523     | E4F1 (1877)                 | Yes                                          |
| 232302at32523    | ECHS1 (1892)                | Yes                                          |
| 115831at32523    | EDEM2 (55741)               | Yes                                          |
| Chromosome Position | Gene Symbol | Gene Name | Verified |
|---------------------|-------------|-----------|----------|
| 154353at32523       | EIF2AK1     | EIF2AK1 (27102) | Yes |
| 190153at32523       | EIF3G       | EIF3G (8666) | Yes |
| 113562at32523       | EMILIN3     | EMILIN3 (90187) | Yes |
| 283389at32523       | ENDOG       | ENDOG (2021) | Yes |
| 137199at32523       | EXO1        | EXO1 (9156) | Yes |
| 132426at32523       | FAM126B     | FAM126B (285172) | Yes |
| 133800at32523       | FAM149A     | FAM149A (25854) | Yes |
| 66945at32523        | FBLN1       | FBLN1 (2192) | Yes |
| 113162at32523       | HAS3        | HAS3 (3038) | Yes |
| 265292at32523       | HNRNPH      | HNRNPH (3189) | Yes |
| 133315at32523       | HYAL2       | HYAL2 (8692) | Yes |
| 16711at32523        | JAG2        | JAG2 (3714) | Yes |
| 215217at32523       | KERA        | KERA (11081) | Yes |
| 89901at32523        | KLHL17      | KLHL17 (339451) | Yes |
| 255366at32523       | LSM11       | LSM11 (134353) | Yes |
| 92894at32523        | MAML3       | MAML3 (55534) | Yes |
| 195716at32523       | MC5R        | MC5R (4161) | Yes |
| 259403at32523       | MMAB        | MMAB (326625) | Yes |
| 295209at32523       | MPZL2       | MPZL2 (10205) | Yes |
| 40350at32523        | NDST1       | NDST1 (3340) | Yes |
| 254990at32523       | NHEJ1       | NHEJ1 (79840) | Yes |
| 150210at32523       | NPRL2       | NPRL2 (10641) | Yes |
| 155481at32523       | NPTX2       | NPTX2 (4885) | Yes |
| 180371at32523       | OPTN        | OPTN (10133) | Yes |
| 207120at32523       | PDZD3       | PDZD3 (79849) | Yes |
| 54439at32523        | PIK3R6      | PIK3R6 (146850) | Yes |
| 109060at32523       | PPARG       | PPARG (5468) | Yes |
| 85764at32523        | PRDM5       | PRDM5 (11107) | Yes |
| 134671at32523       | PRDM6       | PRDM6 (93166) | Yes |
| 130632at32523       | PYROXD2     | PYROXD2 (84795) | Yes |
| 208381at32523       | REN         | REN (5972) | Yes |
| 213362at32523       | ROGDI       | ROGDI (79641) | Yes |
| 177255at32523       | SGK2        | SGK2 (10110) | Yes |
| 122940at32523       | SLC13A1     | SLC13A1 (6561) | Yes |
| 195601at32523       | SLC25A19    | SLC25A19 (60386) | Yes |
| Gene Symbol  | Description               | Presence |
|--------------|---------------------------|----------|
| SLC25A42     | Yes                       |
| SLC38A4      | Yes                       |
| SPAST        | Yes                       |
| SPICE1       | Yes                       |
| SUFU         | Yes                       |
| TFCP2L1      | Yes                       |
| TM4SF19      | Yes                       |
| TMEM177      | Yes                       |
| TMEM216      | Yes                       |
| TMEM237      | Yes                       |
| TMEM82       | Yes                       |
| TPH1         | Yes                       |
| VSIG10       | Yes                       |
| WIPF1        | Yes                       |
| ZBTB43       | Yes                       |
References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research, 25*(17), 3389-3402. doi:10.1093/nar/25.17.3389

Coombe, L., Zhang, J., Vandervalk, B. P., Chu, J., Jackman, S. D., Birol, I., & Warren, R. L. (2018). ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinformatics, 19*(1), 234. doi:10.1186/s12859-018-2243-x

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., . . . Aiden, E. L. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science, 356*(6333), 92-95. doi:10.1126/science.aal3327

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst, 3*(1), 95-98. doi:10.1016/j.cels.2016.07.002

Hara, Y., Takeuchi, M., Kageyama, Y., Tatsumi, K., Hibi, M., Kiyonari, H., & Kuraku, S. (2018). Madagascar ground gecko genome analysis characterizes asymmetric fates of duplicated genes. *BMC Biology, 16*(1), 40. doi:10.1186/s12915-018-0509-4

Hara, Y., Tatsumi, K., Yoshida, M., Kajikawa, E., Kiyonari, H., & Kuraku, S. (2015). Optimizing and benchmarking de novo transcriptome sequencing: from library preparation to assembly evaluation. *BMC Genomics, 16*, 977. doi:10.1186/s12864-015-2007-1

Jackman, S. D., Coombe, L., Chu, J., Warren, R. L., Vandervalk, B. P., Yeo, S., . . . Birol, I. (2018). Tigmint: correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics, 19*(1), 393. doi:10.1186/s12859-018-2425-6

Kadota, M., Nishimura, O., Miura, H., Tanaka, K., Hiratani, I., & Kuraku, S. (2020). Multifaceted Hi-C benchmarking: what makes a difference in chromosome-scale genome scaffolding? *Gigascience, 9*(1). doi:10.1093/gigascience/giz158

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., . . . Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research, 24*(8), 1384-1395. doi:10.1101/gr.170720.113

Nishimura, O., Hara, Y., & Kuraku, S. (2019). Evaluating Genome Assemblies and Gene Models Using gVolante. *Methods Mol Biol, 1962*, 247-256. doi:10.1007/978-1-4939-9173-0_15
Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31*(19), 3210-3212. doi:10.1093/bioinformatics/btv351

Soderlund, C., Bomhoff, M., & Nelson, W. M. (2011). SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Research, 39*(10), e68. doi:10.1093/nar/gkr123

Srikulnath, K., Uno, Y., Nishida, C., Ota, H., & Matsuda, Y. (2015). Karyotype Reorganization in the Hokou Gecko (Gekko hokouensis, Gekkonidae): The Process of Microchromosome Disappearance in Gekkota. *PLoS One, 10*(8), e0134829. doi:10.1371/journal.pone.0134829

Tilak, M. K., Botero-Castro, F., Galtier, N., & Nabholz, B. (2018). Illumina Library Preparation for Sequencing the GC-Rich Fraction of Heterogeneous Genomic DNA. *Genome Biol Evol, 10*(2), 616-622. doi:10.1093/gbe/evy022

Zhu, B. H., Xiao, J., Xue, W., Xu, G. C., Sun, M. Y., & Li, J. T. (2018). P_RNA_scaffolder: a fast and accurate genome scaffolder using paired-end RNA-sequencing reads. *BMC Genomics, 19*(1), 175. doi:10.1186/s12864-018-4567-3