STAR ✦ Protocols

CellPress

Mapping the three-dimensional organization of dinoflagellate genomes

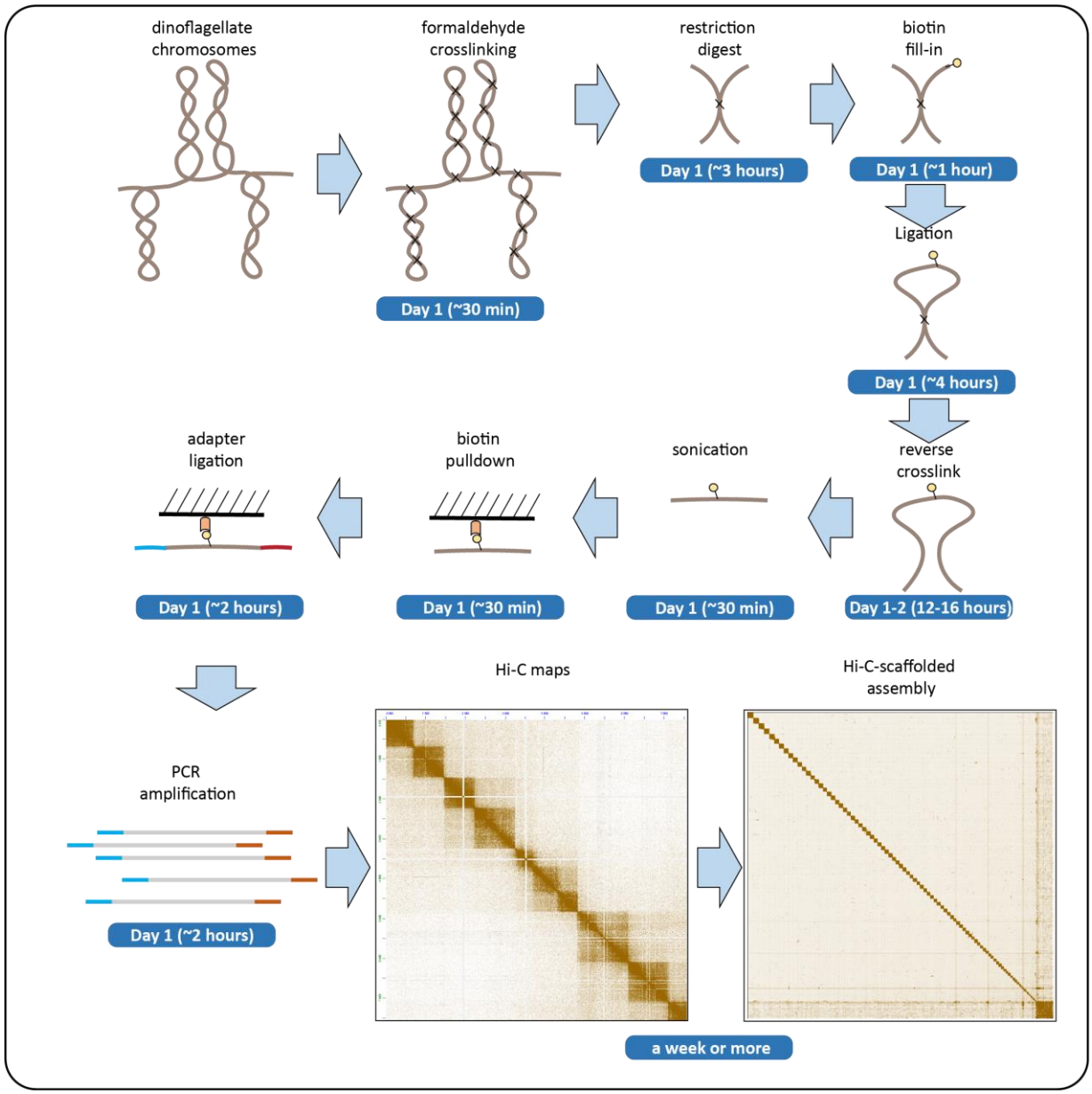
Georgi K. Marinov[1,7,8,*], Anshul Kundaje[1,2], William J. Greenleaf[1,3,4,5], Arthur R. Grossman[6],
[1]Department of Genetics, Stanford University, Stanford, California 94305, USA
[2]Department of Computer Science, Stanford University, Stanford, California 94305, USA
[3]Department of Applied Physics, Stanford University, Stanford, California 94305, USA
[4]Center for Personal Dynamic Regulomes, Stanford University, Stanford, California 94305, USA
[5]Chan Zuckerberg Biohub, San Francisco, California, USA
[6]Carnegie Institution for Science, Department of Plant Biology, Stanford, CA 94305, USA
[7]Technical contact
[8]Lead contact
*Correspondence: marinovg@stanford.edu

Summary

Dinoflagellate genomes represent perhaps the most unique and drastic deviations from the conventional eukaryotic state among all known lineages. Dinoflagellates have lost core features of eukaryote genome biology such nucleosomal packaging of DNA and transcriptional regulation, and their genomes are organized into long unidirectional gene arrays while also featuring abundant highly unusual DNA modifications such as 5-hydroxymethyluracile. They are also often very large and difficult to assemble, which has until recently precluded their analysis with modern functional genomic tools. Here we describe the protocols we have developed for mapping three-dimensional genome organization in dinoflagellates and using it for scaffolding their genome assemblies.

For complete details on the use and execution of this protocol, please refer to Marinov et al. (2021).

Graphical abstract



Before you begin

The protocol below describes the specific steps for generating Hi-C libraries for dinoflagellates and then carrying out Hi-C-assisted assembly/scaffolding. However, the protocol is also extendable to most other single-celled organisms and suspensions cells while in the same time certain steps may

have to be modified for specific dinoflagellate species. As described here, it has been successfully tested in Symbiodiniaceae dinoflagellates (Marinov et al. 2021), which will be used as an example for genome assembly further below, and members of several other dinoflagellate orders as well as many other algae and protozoans.

Prepare cells for crosslinking

Timing: [weeks]

1. Different dinoflagellate species have widely differing growth condition requirements reflecting their diverse trophic strategies and nutrient preferences. It is desirable to obtain at least 2-5 million cells in order to make sure final Hi-C libraries are sufficiently complex, with larger input cell numbers and generating multiple libraries in parallel from them being ideal for maximizing the number of contacts mapped.

Order reagents and kits

Timing: [Days to weeks]

2. Order the reagents and kits listed in the Key Resources Table below and store them at the appropriated temperatures upon receipt.

Prepare buffer solutions

Timing: [Day 1, 3h]

3. Prepare the buffers and solutions listed below:
    a. Hi-C Lysis Buffer
    b. TWB Buffer (Tween Wash Buffer)
    c. 2x Binding Buffer (BB)
    d. IP Elution Buffer
    e. 2.5M Glycine

4. Install the software packages listed in the Key Resources Table

Key Resources Table

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Chemicals and Recombinant Proteins		
Nuclease-free H2O	Thermo Fisher	AM9916
1M Tris-HCl pH 7.5	Thermo Fisher	15567027
1M Tris-HCl pH 8.0	Thermo Fisher	15568025
1M MgCl2	Thermo Fisher	AM9530G

5M NaCl	Thermo Fisher	AM9759
0.5M EDTA, pH 8.0	Thermo Fisher	15575020
20% SDS	Thermo Fisher	AM9820
1xPBS Buffer Solution pH 7.4	Thermo Fisher	10010023
1x TE buffer	Thermo Fisher	12090015
IGEPAL CA-630 detergent	Sigma	I8896-50ML
Tween-20 detergent 10%	Sigma	11332465001
10% Triton X-100	Sigma	93443
37% formaldehyde	Sigma	252549-100ML
0.1M NaHCO3	Sigma	S6014-1KG
Glycine	Sigma	1005901000
MboI restriction enzyme	NEB	R0147
10x CutSmart Buffer	NEB	B7204
0.4 mM biotin-14-dATP	Thermo Fisher	19524-016
100 mM dCTP/dGTP/dTTP	Promega	U1330
DNA Polymerase I Large (Klenow) Fragment	NEB	M0210
10x NEB T4 DNA ligase buffer	NEB	B0202
Bovine Serum Albumin (BSA) 100x	NEB	B9000
T4 DNA Ligase	NEB	M0202
Proteinase K	Promega	MC5005
Dynabeads MyOne Streptavidin T1	Thermo Fisher	65602
General supplies		
AMPure XP beads	Beckman Coulter	A63881
100% EtOH	Sigma	493546-1L
200-µL PCR tubes	Thermo Fisher	AB1182
1.5-mL DNA LoBind microcentrifuge tubes	Eppendorf	022431021
15/50 mL tubes		
Commercial assays		
NEBNext Ultra II DNA Library Prep Kit for Illumina	NEB	E7645L
NEBNext Multiplex Oligos for Illumina	NEB	E7600S
MinElute PCR Purification Kit	Qiagen	28004/28006
Covaris milliTUBE 1 mL	Covaris	520135
QuBit tubes	Thermo Fisher	Q32856

QuBit dsDNA HS Assay Kit	Thermo Fisher	Q32854
TapeStation D1000 tape	Agilent	5067-5582
TapeStation D1000 reagents	Agilent	5067-5583
Equipment		
TapeStation	Agilent	
Covaris E220	Covaris	
QuBit fluorometer	Thermo Fisher	Q33327
Tabletop centrifuge		
Thermomixer	Eppendorf	5382000023
PCR Thermocycler		
magnetic rack		
tube rotator		
Software		
Juicer		https://github.com/aidenlab/juicer
Juicebox		https://github.com/aidenlab/Juicebox
3D-DNA		https://github.com/aidenlab/3d-dna
bwa		https://github.com/lh3/bwa
sra-tools		https://github.com/ncbi/sra-tools
Publicly deposited datasets		
Fugacium kawagutii Hi-C		SRA: SRR25948349 and SRR25948348
Fugacium kawagutii draft assembly		http://sampgr.org.cn/index.php/download

Materials and equipment

Buffer recipes

Hi-C Lysis Buffer		
Stock	**Final concentration**	**Amount**
1M Tris-HCl pH 8.0	10 mM	500 µL
5M NaCl	10 mM	500 µL
Igepal CA630	0.20%	100 µL
Nuclease-free H2O	-	48.9 mL
Total	**-**	**50 mL**

TWB (Tween Washing Buffer)

Stock	Final concentration	Amount
1M Tris-HCl pH 7.5	5 mM	250 µL
0.5M EDTA	0.5 mM	50 µL
5M NaCl	1 M	10 mL
10% Tween 20	0.05%	250 µL
Nuclease-free H2O	-	39.5 mL
Total	**-**	**50 mL**

2x BB (Binding Buffer)		
Stock	**Final concentration**	**Amount**
1M Tris-HCl pH 7.5	10 mM	500 µL
0.5M EDTA	1 mM	100 µL
5M NaCl	2 M	20 mL
Nuclease-free H2O	-	29.4 mL
Total	**-**	**50 mL**

IP Elution Buffer		
Stock	**Final concentration**	**Amount**
20% SDS	1%	2.5 mL
NaHCO3	0.1M	420 mg
Nuclease-free H2O	-	47.5 mL
Total	**-**	**50 mL**

2.5M Glycine		
Stock	**Final concentration**	**Amount**
Glycine	2.5M	9.375g
Nuclease-free H2O	-	50 mL
Total	**-**	**50 mL**

Notes: Store the Hi-C Lysis Buffer and the TWB Buffer at 4°C. Store the 2x BB Buffer at room temperature, as well as the IP Elution Buffer and the 2.5M Glycine solution. Do not store the IP Elution Buffer at 4°C as the SDS will precipitate (though that is reversible by heating it up to 37°C).

Step-by-step method details

The protocol below describes a modification of the *in situ* version (Rao et al. 2014) of the Hi-C (Lieberman-Aiden et al. 2009) assay. Briefly, the method involves crosslinking of cells, followed by denaturation of chromatin (in order to uniformly expose the genome to restriction enzymes), restriction digestion, biotin fill-in, ligation *in situ,* purification of DNA, DNA shearing, streptavidin pull-down of biotin-labeled fragments, and sequencing library generation on beads.

We also describe computational processing using the Juicer pipeline (Durand et al. 2016a), and Hi-C-based automated genome scaffolding and correction using the 3D-DNA pipeline (Dudchenko et al. 2017) and manual correction in Juicebox (Durand et al. 2016b)

Crosslinking

Timing: [Day 1, 30 minutes]

This step fixes chromatin in place by covalently crosslinking proteins to DNA and proteins to proteins when they are in very close physical proximity to each other. This allows mapping long-range *cis*- and *trans*- genomic contacts in subsequent steps that generate bridged chimeric DNA fragments.

1. Pellet at least 2 million cells in a centrifuge (time and speed varying depending on the properties of the cells; usually 2 minutes at 1000 *g* is sufficient for dinoflagellate cells)
2. Resuspend in 10 mL room-temperature 1x PBS buffer.
3. Add 270 µL 37% formaldehyde for a final concentration of 1%.
4. Incubate at room temperature for 15 minutes.
5. Quench the reaction by adding 1 mL 2.5M Glycine solution and incubate at room temperature for 5 minutes.
6. Centrifuge cells and discard the supernatant.
7. Resuspend cells in 1 mL cold 1x PBS buffer.
8. Centrifuge cells and discard the supernatant.
9. Store crosslinked chromatin at -80°C.

Pause point: Crosslinked chromatin is stable at -80°C almost indefinitely.

CRITICAL: It is preferable to not crosslink in growth media, in order to standardize crosslinking between different conditions as certain components in some media recipes may interfere with the crosslinking reaction. Do not use Tris-based buffers for crosslinking as Tris reacts with formaldehyde and reduces crosslinking efficiency. Make sure the crosslinking buffer is not cold as that can affect both the regulatory state of the cells and the crosslinking reaction.

Nuclei lysis

Timing: [Day 1, 30 minutes]

In this step, nuclei are permeabilized in order for the reagents used in the subsequent *in situ* reactions to more easily access chromatin (some permeabilization has already happened during crosslinking).

10. Resuspend crosslinked cells in 250 µL cold Hi-C Lysis Buffer.
11. Incubate on ice for 15 minutes.
12. Centrifuge cells and remove the supernatant.
13. Wash with 500 µL cold Hi-C Lysis Buffer.
14. Centrifuge cells and remove the supernatant.

Denaturation

Timing: [Day 1, 30 minutes]

In this step, chromatin is denatured in order to expose DNA to restriction digestion. Non-denatured chromatin prevents restriction digestion where DNA is occluded by protein occupancy, resulting in biased restriction digestion and ligation profiles, which is undesirable for general Hi-C purposes.

15. Resuspend nuclei in 50 µL 0.5% SDS.
16. Incubate at 62°C for 10 minutes.
17. Quench by adding 145 µL nuclease-free H_2O and 25 µL 10% Triton X-100
18. Incubate at 37°C for 15 minutes.

CRITICAL: Do not incubate at 62°C for too long or crosslink reversal may commence.

Restriction digest

Timing: [Day 1, 2.5 hours]

In this step, chromatin is digested with a restriction enzyme, usually MboI, which is a 4-cutter recognizing the GATC sequence. The free DNA ends will later be ligated with other fragments in physical proximity. The restriction enzyme is then heat-inactivated in order to block its activity during subsequent steps in the protocol.

19. Add 25 µL 10x CutSmart Buffer and 100 U of the MboI restriction enzyme.
20. Incubate at 37°C for 2 hours or longer in a Thermomixer with shaking at 1000 rpm.
21. Incubate at 62°C for 20 minutes to inactivate the restriction enzyme.

CRITICAL: As above, do not incubate at 62°C for too long or crosslink reversal may commence.
Optional: Additional or alternative restriction enzymes may be used to improve resolution. However, attention needs to be paid to their sensitivity to DNA modifications such as 5-methylcytosine and N^6-methyladenosine as well as whether they are efficiently heat-inactivated.

Restriction end fill-in

Timing: [Day 1, 1 hour]

In this step, restrictions ends are filled in and labeled with biotin using biotin-14-dATP, which allows ligated fragments to be later pulled down with streptavidin.

22. Add 37.5 µL 0.4 mM biotin-14-dATP, 1.5 µL 10 mM dCTP, 1.5 µL 10 mM dGTP, 1.5 µL 10 mM dTTP, and 8 µL 5U/µL DNA Polymerase I Large (Klenow) Fragment.
23. Incubate at 37°C for 45 minutes in a Thermomixer at 1000 rpm.

Fragment end ligation

Timing: [Day 1, 4.5 hours]

In this step, filled in restriction ends are ligated, creating chimeric DNA fragments that correspond to short- and long-range and *cis*- and *trans*- interactions between genomic regions.

24. Add 663 µL H_2O, 120 µL 10x T4 DNA Ligase Buffer, 100 µL 10% Triton X-100, 12 µL BSA, and 5 µL 400 U/µL T4 DNA Ligase.
25. Incubate at room temperature on a rotator for at least 4 hours.

Reverse crosslinking

Timing: [Day 1 and Day 2, 12-16 hours]

In this step, crosslinked and ligated chromatin is reverse crosslinked so that DNA can be purified and sequenced. This is accomplished through prolonged exposure to high temperature.

26. Pellet nuclei by centrifugation.
27. Resuspend in 200 µL IP Elution Buffer.
28. Add 20 µL Proteinase K.
29. Incubate at 65°C for 12-16 hours (e.g., in a Thermomixer) overnight to reverse crosslinks.

CRITICAL: Always use safe-lock tubes during reverse crosslinking as the exposure to heat occasionally leads to spontaneous opening of regular tubes, sample evaporation and complete degradation of DNA, and total experiment loss.

Pause point: reverse crosslinked material can be safely stored at -20°C or -80°C for days to weeks.

DNA shearing and purification

Timing: [Day 2, 2 hours]

In this step, reverse crosslinked DNA is sheared and purified. Shearing is most efficiently done using a Covaris instrument, or an equivalent automated mid- to high-throughput system, which allows parallelization over multiple tubes, but probe sonicators can also be used.

30. Add 660 µL 1x TE Buffer to the reverse crosslinked Hi-C samples.
31. Shear using a Covaris in 1-mL Covaris miliTubes down to 200-400 bp.
32. Load DNA onto a MinElute column, by mixing 200-µL sample with 600 µL PB buffer (from the MinElute kit) and centrifuging for 1 minute at 13,000 rpm and discarding the flowthrough. Repeat until the sample is exhausted.
33. Add 600 µL PE Buffer (from the MinElute kit), and centrifuge for 1 minute at 13,000 rpm.
34. Centrifuge for 1 minute at 13,000 rpm to dry the column.

35. Elute in 300 µL EB Buffer (from the MinElute kit) by centrifuging for 1 minute at 13,000 rpm.
36. Measure DNA concentration using QuBit
37. Evaluate fragment distribution using TapeStation (or equivalent, e.g. BioAnalyzer)

CRITICAL: Measuring DNA concentration and evaluating fragment distribution are very important at this step. Make sure that a sufficient amount of DNA (i.e., at least a microgram starting with in the neighborhood of 10^6-10^7 cells) is recovered and that it is properly sheared. Optimize sonication condition as needed to achieve the desired fragment distribution.

Note: Use DNA LoBind tubes for elution and long-term storage of DNA, in order to minimize sample loss.

Pause point: Purified DNA can be stored at -20°C or -80°C and biotin pull down and library generation can be generated at any time after that.

Alternative: Physical shearing can be replaced with enzymatic shearing, i.e. using the NEBNext Ultra FS fragmentation reagents or an equivalent kit. In that alternative protocol, DNA is first purified and then fragmented enzymatically, then directly used for biotin pull-down.

Biotin pull-down and sequencing library generation

Timing: [Day 2, 8 hours]

We carry out biotin pull down using magnetic streptavidin beads and then make libraries while ligated DNA fragments are on beads using the NEBNext Ultra II DNA Library Prep kit, for maximum efficiency and convenience of handling.

Biotin pull-down

38. Pipette 20 µL Streptavidin T1 Dynabeads into 1.5 mL DNA LoBind tubes.
39. Separate beads on magnet and remove supernatant.
40. Resuspend beads in 180 µL TWB Buffer.
41. Separate beads on magnet and remove supernatant.
42. Resuspend beads in 300 µL 2x BB Buffer.
43. Add 300 µL fragmented DNA.
44. Incubate at room temperature on a rotator for at least 15 minutes.
45. Separate beads on magnet and remove supernatant.
46. Resuspend beads in 180 µL TWB Buffer.
47. Incubate at 55°C in a Thermomixer with shaking at 1000 rpm for 2 minutes.
48. Separate beads on magnet and remove supernatant.
49. Resuspend beads in 180 µL TWB Buffer.
50. Incubate at 55°C in a Thermomixer with shaking at 1000 rpm for 2 minutes.
51. Separate beads on magnet and remove supernatant.

End Repair

52. Resuspend beads in 50 µL EB or 0.1x TE Buffer.
53. Add 7 µL NEB End Repair Buffer.
54. Add 3 µL NEB End Repair Enzyme.
55. Incubate at 20°C for 30 minutes in a Thermomixer with shaking at 1000 rpm.
56. Incubate at 65°C for 30 minutes to inactivate enzymes.

Adaptor Ligation

57. Add 2.5 µL NEB Adaptor.
58. Add 1 µL NEB Ligation Enhancer.
59. Add 30 µL NEB Ligation Mix.
60. Incubate at 20°C for 20 minutes in a Thermomixer with shaking at 1000 rpm.
61. Add 3 µL NEB USER enzyme.
62. Incubate at 37°C for 15 minutes in a Thermomixer with shaking at 1000 rpm.
63. Separate beads on magnet and remove supernatant.
64. Resuspend beads in 180 µL TWB Buffer.
65. Incubate at 55°C in a Thermomixer with shaking at 1000 rpm for 2 minutes.
66. Separate beads on magnet and remove supernatant.
67. Resuspend beads in 100 µL 0.1x TE Buffer
68. Separate beads on magnet and remove supernatant.
69. Resuspend beads in 15 µL 0.1x TE Buffer
70. Transfer beads to PCR tubes

PCR amplification

71. Incubate beads at 98°C for 10 minutes.
72. Prepare PCR amplification reactions as follows:

Reagent	**Amount**
Beads with DNA	15 µL
NEB i5 primer	5 µL
NEB i7 primer	5 µL
NEBNext Ultra II Q5 Master Mix (2x)	25 µL

73. Carry out amplification as follows:

PCR cycling conditions

Steps	Temperature	Time	Cycles
Initial Denaturation	98°C	30 seconds	1
Denaturation	98°C	10 seconds	12-15 cycles
Annealing	65°C	30 seconds	
Extension	72°C	30 seconds	

Final extension	72 °C	5 minutes	1
Hold	4 °C	forever	

Pause point: Amplified libraries can be stored for some time at -20°C before clean up and quality evaluation.

PCR amplification clean up

74. Separate beads on magnet and transfer the supernatant to new PCR tubes.
75. Mix the reaction with 50 µL AMPure XP beads.
76. Incubate at room temperature for 15 minutes to bind DNA to beads.
77. Separate the beads from the supernatant on magnet.
78. Add 200 µL freshly made 80% EtOH to beads (while on magnet)
79. Remove the EtOH
80. Add 200 µL freshly made 80% EtOH to beads (while on magnet)
81. Remove the EtOH
82. Resuspend beads in 33 µL 0.1X TE beads. Mix well and incubate at room temperature for 2 minutes.
83. Place on magnet to separate the liquid from the beads.
84. Transfer the supernatant to 1.5mL DNA LoBind tubes.
85. Measure library concentration using a QuBit.
86. Evaluate fragment distribution using TapeStation.

CRITICAL: Always use QuBit (or equivalent highly accurate DNA quantification assays) to measure final library concentration. Large errors in library quantification can result in suboptimal sequencing outcomes. It is also vital to characterize fragment length distribution for quality evaluation purposes.

Pause point: Finished libraries can be stored indefinitely at -20°C or -80°C.

Library sequencing

87. Sequence final libraries on Illumina NextSeq, NovaSeq or equivalent platforms in a paired-end format

CRITICAL: It is vital to obtain sufficient coverage in order to allow properly powered downstream analysis. For Hi-C libraries this means more sequencing than any other functional genomic assay. For example, for a genome of size around 1 Gbp (which is the general size of most Symbiodiniaceae species genomes), a billion read pairs or more is required. For larger genomes, targeted coverage increases as the square of genome size.

Initial Hi-C read mapping

In this step, reads are mapped to a draft genome assembly that is assumed to already exist. We will use as an example for the individual processing steps Hi-C data for *Fugacium kawagutii* that was published previously by Li et al. (Li et al. 2020). We use Juicer (Durand et al. 2016a) for mapping and processing of Hi-C datasets.

88. Download genome assembly:

```
wget http://sampgr.org.cn/downloads/Fugacium_kawagutii_V3_genome_scaffold.fasta.gz
```

89. Uncompress the FASTA file:

```
gunzip Fugacium_kawagutii_V3_genome_scaffold.fasta.gz
```

90. Prepare the genomic index folder:

```
mkdir bwa-indexes; ln -s Fugacium_kawagutii_V3_genome_scaffold.fa bwa-
indexes/Fugacium_kawagutii_V3_genome_scaffold.fa
```

91. Make genomic index:

```
bwa index bwa-indexes/Fugacium_kawagutii_V3_genome_scaffold.fa
```

92. Download *F. kawagutii* Hi-C sequencing reads from the Short Read Archive (SRA). In this case we use the `fasterq-dump` tool from the `sra-tools` package:

```
fasterq-dump SRR25948349 --split-spot --split-files --threads 20 -O SRR25948348-out
fasterq-dump SRR25948348 --split-spot --split-files --threads 20 -O SRR25948349-out
```

93. Combine reads into single files for input into Juicer:

```
cat SRR25948348-out/SRR25948349_1.fastq SRR25948349-out/SRR25948348_1.fastq | gzip
> reads_R1.fastq.gz
cat SRR25948348-out/SRR25948349_2.fastq SRR25948349-out/SRR25948348_2.fastq | gzip
> reads_R2.fastq.gz
```

94. Prepare Juicer input folders as follows:

```
mkdir juicer-Fugacium_kawagutii_V3-Hi-C
cd juicer-Fugacium_kawagutii_V3-Hi-C
ln -s juicer-1.6/CPU scripts
ln -s juicer_tools.2.13.07/juicer_tools.jar juicer_tools.jar
ln -s ../bwa-indexes references
mkdir fastq
cd fastq
ln -s ../../reads_R1.fastq.gz
ln -s ../../reads_R2.fastq.gz
cd ..
```

95. Set up the Juicer running environment:

```
export PATH=bwa-0.7.17:$PATH; export PATH=$PATH:juicer_tools.2.13.07/;
```

96. Run Juicer as follows:

```
scripts/juicer.sh -t 20 -D ../juicer-Fugacium_kawagutii_V3-Hi-C -d ../juicer-
Fugacium_kawagutii_V3-Hi-C -p ../Fugacium_kawagutii_V3_genome_scaffold.chrom.sizes
-y none -s none -z ../bwa-indexes/Fugacium_kawagutii_V3_genome_scaffold.fa
```

The `chrom.sizes` file contains one tab-separated line with the name and size (in bp) of each contig in the assembly

97. Juicer will produce a "`aligned`" folder, which contains a `inter.hic` file, which can be used for visualization in Juicebox (Durand et al. 2016b), and a "`merged_no_dups.txt`", which can be used as input to the 3D-DNA scaffolding pipeline.

3D-DNA scaffolding/assembly correction

In this step, we take the initial Hi-C map and the available draft assembly, and use the Hi-C proximity information to rescaffold the assembly and correct assembly errors. We use the 3D-DNA pipeline (Dudchenko et al. 2017) for this purpose.

98. Prepare the 3D-DNA running folder as follows:

```
mkdir juicer-Fugacium_kawagutii_V3-Hi-C-3D_DNA
cd juicer-Fugacium_kawagutii_V3-Hi-C-3D_DNA
cp -R 3d-dna-master/* .
chmod +x *.sh
chmod +x */*.sh
chmod +x */*.awk
```

99. Run 3D-DNA as follows:

```
./run-asm-pipeline.sh --sort-output --build-gapped-map -r 10 -i 1000 ../bwa-
indexes/Fugacium_kawagutii_V3_genome_scaffold.fa ../juicer-Fugacium_kawagutii_V3-
Hi-C/aligned/merged_nodups.txt
```

In this case 3D-DNA is run with 10 iterative rounds of rescaffolding and with a maximum resolution of edits of 1000 bp.

3D-DNA scaffolding/assembly correction

While the automated 3D-DNA assembly correction can achieve great results on its own, it is often not perfect (this problem is the more severe the worse the initial assembly is). Often, especially when working with complex genomes such as those of dinoflagellates, manual correction is needed. This is carried out in Juicebox using the `*.rawchrom.hic` and `*.rawchrom.assembly` 3D-DNA output files (in this case

`Fugacium_kawagutii_V3_genome_scaffold.rawchrom.hic` and `Fugacium_kawagutii_V3_genome_scaffold.rawchrom.assembly`).

100. Examples of the typical issues to be corrected and how this is to be done interactively are shown in Figure 3. Users are advised to consult with the 3D-DNA manual and video tutorials for more details.

101. Save the manually corrected assembly in another file, e.g. as `Fugacium_kawagutii_V3_genome_scaffold.review.assembly`, and use it as input for the generation of the final assembly as follows:

```
./run-asm-pipeline-post-review.sh --sort-output -r ../juicer-Fugacium_kawagutii_V3-Hi-C-3D_DNA/Fugacium_kawagutii_V3_genome_scaffold.review.assembly ../Fugacium_kawagutii_V3_genome_scaffold.fa ../juicer-Fugacium_kawagutii_V3-Hi-C/aligned/merged_nodups.txt
```

This step will produce a new `FINAL.fasta` FASTA file, e.g. in this case `Fugacium_kawagutii_V3_genome_scaffold.FINAL.fasta`.

Final map generation

After creating a manually corrected assembly, it is necessary to remap the raw Hi-C reads to generate final maps. This is done following the same procedure as described above but using the new assembly FASTA file as a reference.

Expected outcomes

Hi-C is usually a very reliable and robust assay. It is expected that the pre-library generation steps will result in obtaining a large amount of DNA (several micrograms), and it should be fragmented down to 200-400 bp range. A TapeStation profile of such a sample is shown in Figure 1a. Library generation should also produce robust libraries with abundant amount of DNA. An example is shown in Figure 1b. Weak libraries with little product are unlikely to sequence well and not exhibit serious complexity issues.

Examples of initial Juicer maps for *F. kawagutii* are shown in Figure 2a and 2b. These will vary depending on the specifics of the species and assembly one is working with, but in this case several common issues are to be noted. The *F. kawagutii* used here has already been Hi-C scaffolded using an automated procedure. However, the global map (Figure 2a) shows clear errors in terms of chromosome definition. A particularly striking example Is to be found towards the lower right corner, where multiple distinct chromosomes are lumped into the same extremely long pseudochromosome. In addition, there are still many missassemblies, visible as strong ``interchromosomal interactions'' in the global map, and much more clearly in the local maps for many individual pseudochromsomes (example shown in Figure 2b). These need to be corrected.

In addition, Figure 2c shows the common quality control statistics relevant to Hi-C experiments. The key ones are the percentage of uniquely mapping reads (the higher the better, although it need not necessarily reflect a failure of the experiment, but rather the presence of too many repeats in the genome), the library complexity estimate (again, the higher the better), and the number of Hi-C contacts. When working with fully and reliably assembled genomes, it is also relevant to look at the ratios of interchromosomal versus intrachromosomal and short-range versus long-range contacts (too few long-range contacts is undesirable), but these do not have the same relevance for preliminary maps against highly fragmented draft assemblies.

Figures 3a-d shows the typical manual assembly correction operations. Figures 3e-f show the final outcome of assembly correction after remapping of reads to the corrected assembly. In this case, chromosomes have been properly defined and the bulk of misjoins and other missassemblies have been rectified. However, notice the large square of high-density interactions in the lower right corner. These are the debris remaining after scaffolding that cannot be placed into the main chromosome scaffolds. It is common for these to correspond to collapsed repeats in draft dinoflagellate assemblies, i.e. repetitive elements contigs that are artificially present in only a single copy in the assembly, but in fact exist in a very large number of copies in the actual genome. Long read-based draft assemblies are needed to resolve this problem.

Limitations

The Hi-C assays is a very powerful tool for mapping the 3D organization of genomes. However, when working with extremely repetitive and complex genomes, such as those of many dinoflagellates with larger genomes than those of Symbiodiniaceae, some limitation can become apparent. One of them is the use of short reads, which, when the genome is full of highly similar repeats, makes it difficult to place the short reads uniquely, and results in sparse and difficult to interpret maps. Versions of Hi-C using long read sequencing (PacBio or nanopore; Deshpande et al. 2022) can help in such cases, but those are beyond the scope of the current manuscript.

Troubleshooting

Problem 1:

There is no structure visible in the Hi-C maps other than a very narrow main diagonal band.

Potential solution:

This indicates a total failure of the experiment. There are several possible reasons for it – failure to crosslink properly, failure to ligate, or degradation of the sample during reverse crosslinking. The experiment needs to be redone in such cases.

Problem 2:

Very few Hi-C contacts are observed relative to the number of sequenced reads and the total estimated library complexity is low (see the quality control statistics that Juicer outputs in Figure 2C regarding where these numbers are to be found).

Potential solution:

There are two possible reasons for such an outcome – either the experiment started with too few cells or the subsequent biotin capture was not efficient/was accompanied by too much DNA loss.

Increase the number of input cells and the amount of DNA going into the pull-down. It is also a good idea to generate multiple different libraries, then pool the contacts from each of them in order to maximize the density of the final maps.

Problem 3:

The assembly is still too fragment after 3D-DNA scaffolding and/or there are still too many obvious misjoins and inversions.

Potential solution:

One common reason for poor scaffolding results is the poor quality of the original assembly used. While Hi-C scaffolding is a very powerful method for achieving chromosome-level assemblies, it is not all-powerful and is limited by the contiguity of the original assembly, especially for highly repetitive genomes such as those of dinoflagellates. In our experience, original dinoflagellate assemblies with N_{50} values in the neighborhood of 10 kbp or worse are nearly impossible to scaffold automatically and must be improved. This can be done by obtaining more long-read data, either using PacBio or nanopore sequencing, but this topic is beyond the scope of the current protocol.

If the original assembly is sufficiently contiguous, but the scaffolding still contains too many obvious errors, running 3D-DNA with more rounds of rescaffolding can improve it substantially. The default setting is to do two such rounds, but that can be increased arbitrarily with the "-r N" parameter of the 3D-DNA pipeline. If assembly errors persist, correct them manually in Juicebox.

Resource availability

Lead contact

Georgi K. Marinov (marinovg@stanford.edu)

Materials availability

No new materials were generated associated with this protocol.

Data and code availability

No new data was generated fir this study. The public datasets used for illustration purposes are listed in the Key Resources Table.

The updated assembly for *F. kawagutii* has been deposited to Zenodo under doi: 10.5281/zenodo.10035644.

Acknowledgments

This work was supported by NIH grants (P50HG007735, RO1 HG008140, U19AI057266 and UM1HG009442 to W.J.G., 1UM1HG009436 to W.J.G. and A.K., 1DP2OD022870-01 and 1U01HG009431 to A.K.), the Rita Allen Foundation (to W.J.G.), the Baxter Foundation Faculty Scholar Grant, and the Human Frontiers Science Program grant RGY006S (to W.J.G). W.J.G. is a Chan Zuckerberg Biohub investigator and acknowledges grants 2017-174468 and 2018-182817 from the Chan Zuckerberg Initiative. Fellowship support provided by the Stanford School of Medicine Dean's Fellowship (G.K.M.). This work is also supported by NSF-IOS EDGE Award 1645164 to A.R.G. and Carnegie Venture grant 10907 (to G.K.M.).

The authors would like to thank Tingtin Xiang, Alexandro E. Trevino, Lucas Phillip, Alex Leffell, Olga Dudchenko, Erez Lieberman Aiden, and members of the Greenleaf, Kundaje, Pringle and Grossman laboratories for helpful discussion and suggestions regarding this work.

Author contributions

G.K.M. wrote the manuscript with input from all authors.

Declaration of interests

The authors declare no competing interests.

References

Deshpande AS, Ulahannan N, Pendleton M, Dai X, Ly L, Behr JM, Schwenk S, Liao W, Augello MA, Tyer C, Rughani P, Kudman S, Tian H, Otis HG, Adney E, Wilkes D, Mosquera JM, Barbieri CE, Melnick A, Stoddart D, Turner DJ, Juul S, Harrington E, Imieliński M. 2022. Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. *Nat Biotechnol* **40**(10):1488--1499.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**(6333):92--95.

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. 2016a. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**(1):95-98.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. 2016b. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**(1):99-101.

Li T, Yu L, Song B, Song Y, Li L, Lin X, Lin S. 2020. Genome Improvement and Core Gene Set Refinement of *Fugacium kawagutii*. *Microorganisms* **8**(1):102

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**(5950):289-293.

Marinov GK, Trevino AE, Xiang T, Kundaje A, Grossman AR, Greenleaf WJ. 2021. Transcription-dependent domain-scale three-dimensional genome organization in the dinoflagellate *Breviolum minutum*. *Nature Genetics* **53**:613-617.

Rao SSP, Huang SC, Glenn St Hilaire B, Engreitz JM, Perez EM, Kieffer-Kwon KR, Sanborn AL, Johnstone SE, Bascom GD, Bochkov ID, Huang X, Shamim MS, Shin J, Turner D, Ye Z, Omer AD, Robinson JT, Schlick T, Bernstein BE, Casellas R, Lander ES, Aiden EL. 2017. Cohesin Loss Eliminates All Loop Domains. *Cell* **171**(2):305-320.e24.

Figure legends

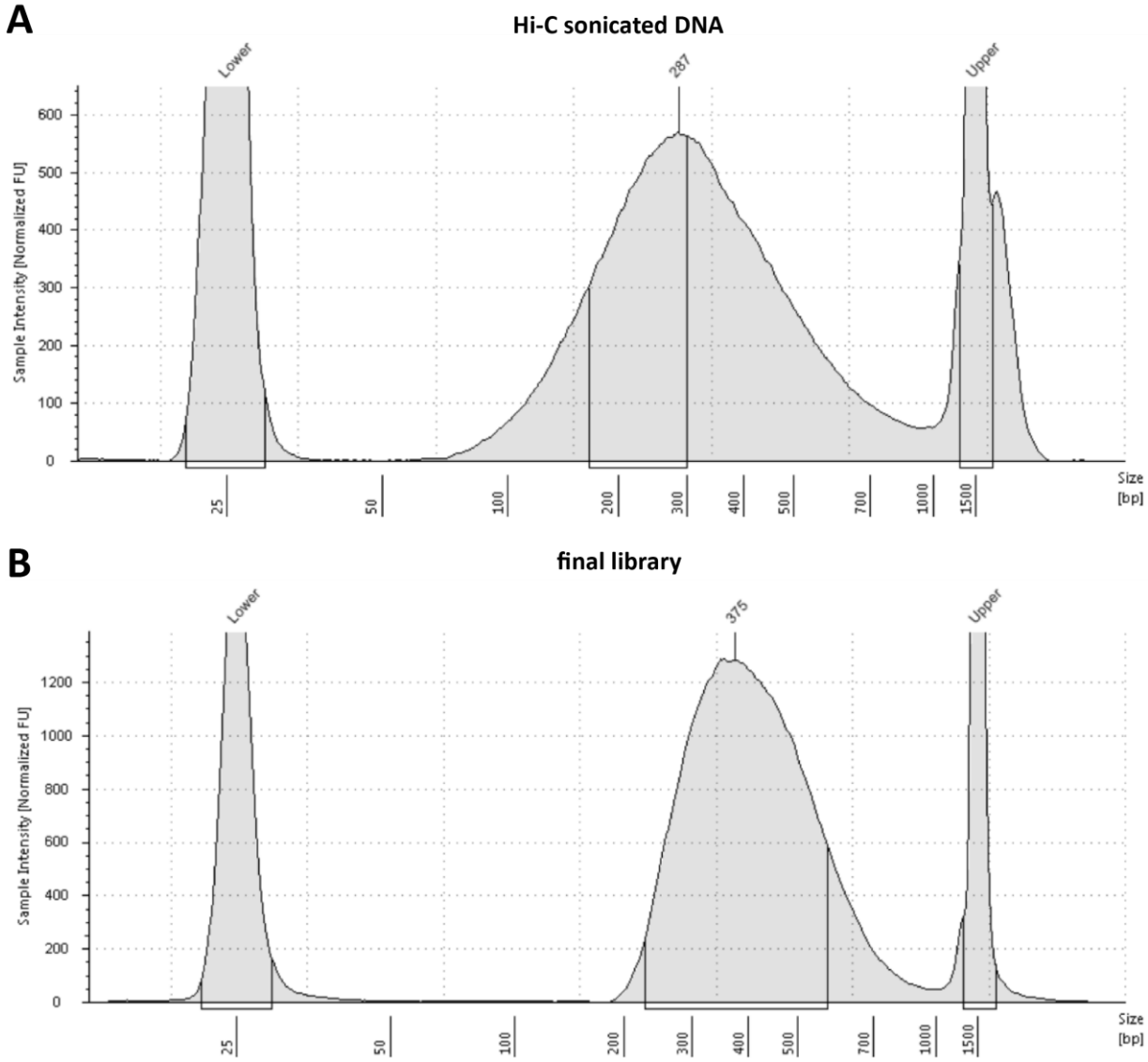
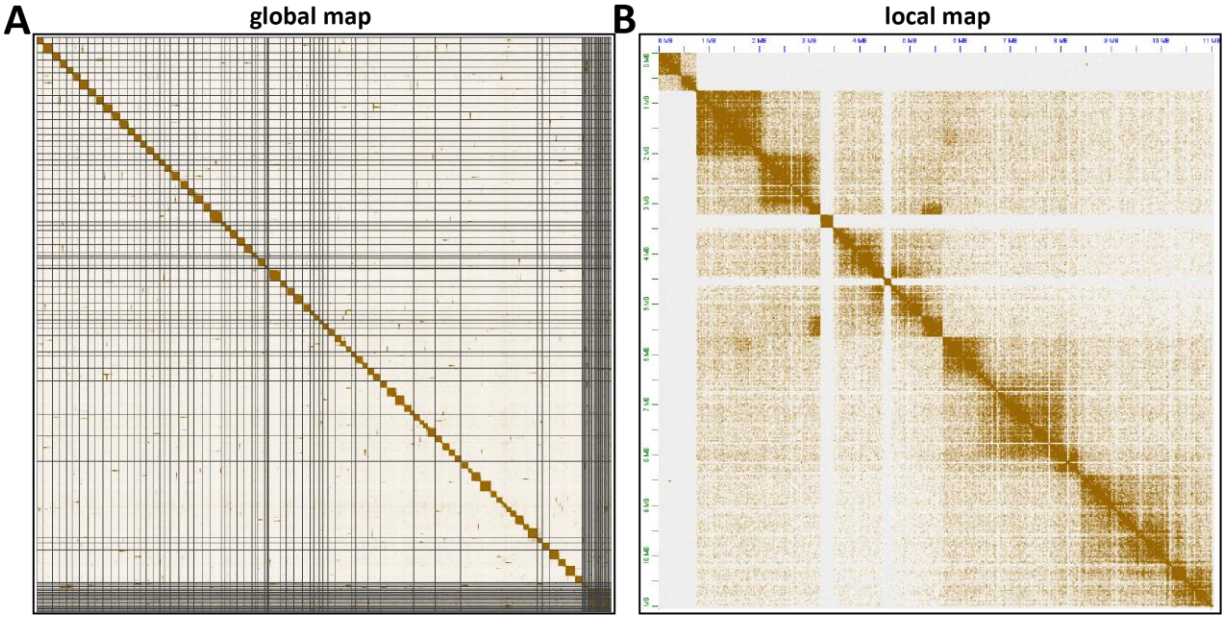


Figure 1: Typical outcomes of ligated DNA sonication and sequencing library generation. (A) TapeStation profile of sheared Hi-C DNA before pull down. (B) TapeStation profile of a final Hi-C library.



C

```
Sequenced Read Pairs:  1,286,459,855
      Normal Paired: 1,018,372,681 (79.16%)
      Chimeric Paired: 60,074,530 (4.67%)
      Chimeric Ambiguous: 8,566,117 (0.67%)
      Unmapped: 199,446,527 (15.50%)
      Ligation Motif Present: 0 (0.00%)
Alignable (Normal+Chimeric Paired): 1,078,447,211 (83.83%)
Unique Reads: 497,811,875 (38.70%)
PCR Duplicates: 580,635,336 (45.13%)
Optical Duplicates: 0 (0.00%)
Library Complexity Estimate: 594,891,611
Intra-fragment Reads: 0 (0.00% / 0.00%)
Below MAPQ Threshold: 103,250,929 (8.03% / 20.74%)
Hi-C Contacts: 394,560,946 (30.67% / 79.26%)
      Ligation Motif Present: 0  (0.00% / 0.00%)
      3' Bias (Long Range): 0% - 0%
      Pair Type %(L-I-O-R): 25% - 25% - 25% - 25%
Inter-chromosomal: 64,198,890  (4.99% / 12.90%)
Intra-chromosomal: 330,362,056  (25.68% / 66.36%)
Short Range (<20Kb): 297,195,039  (23.10% / 59.70%)
Long Range (>20Kb): 33,136,363  (2.58% / 6.66%)
```

Figure 2: Typical results for dinoflagellate Hi-C. (A) Global Hi-C map for *F. kawagutii*. (B) Locl Hi-C map for *F. kawagutii.* Missassembly and incorrect chromosome definition issues are clearly visible in both snapshots. (C) Typical Juicer mapping and quality control statistics. The key parameters to watch for are the unique reads percentage, the library complexity estimate, and number of Hi-C contacts, the pair type distribution (which should be uniform). Note that the ratios of inter- and intrachromosomal contacts and short-range and long-range contacts depend greatly on how fragmented the assembly against which reads are mapped is.

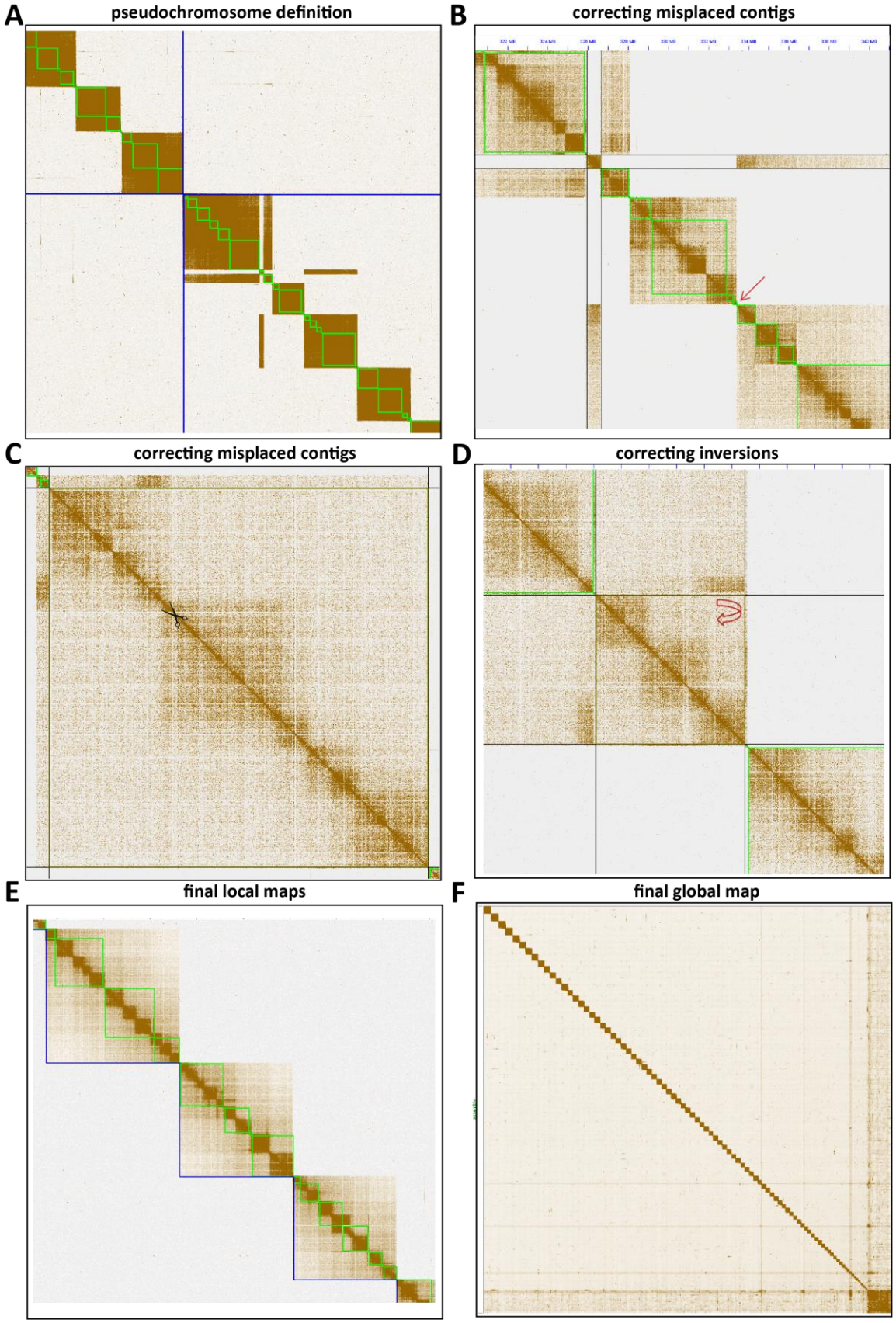

A
pseudochromosome definition
B
correcting misplaced contigs
C
correcting misplaced contigs
D
correcting inversions
E
final local maps
F
final global map

Figure 3: Improving dinoflagellate assemblies using manual correction in Juicebox. (A) Manual definition of chromosome boundaries. (B) Correction of misplaced contigs through direct rearrangement. (C) Correction of misjoins through excision. (D) Correction of inversions. (E and F) Final *F. kawagutii* local and global maps after manual correction.