

# Revision of the *Capsaspora* genome using read mating information adjusts the view on premetazoan genome

Seitaro Denbo<sup>1</sup> | Katsutoshi Aono<sup>1</sup> | Takaaki Kai<sup>1</sup> | Rei Yagasaki<sup>2</sup> |  
 Ñaki Ruiz-Trillo<sup>3,4,5</sup>  | Hiroshi Suga<sup>1</sup> 

<sup>1</sup>Faculty of Life and Environmental Sciences, Prefectural University of Hiroshima, Shobara, Japan

<sup>2</sup>Department of Zoology, Division of Biological Sciences, Graduate School of Science, Kyoto University, Kyoto, Japan

<sup>3</sup>Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Barcelona, Spain

<sup>4</sup>Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona (UB), Barcelona, Spain

<sup>5</sup>ICREA, Barcelona, Spain

## Correspondence

Hiroshi Suga

Email: hsuga@pu-hiroshima.ac.jp

## Funding information

JSPS KAKENHI, Grant/Award Number: 16K07468; NOVARTIS Foundation; ITOH Science Foundation; Naito Foundation; Prefectural University of Hiroshima JUTEN Grant, European Research Council Consolidator Grant, Grant/Award Number: ERC-2012-Co-616960, BFU2014-57779-P and BFU2017-90114-P; Ministerio de Economía y Competitividad (MINECO); Agencia Estatal de Investigación (AEI); Fondo Europeo de Desarrollo Regional (FEDER)

## Abstract

The genome sequences of unicellular holozoans, the closest relatives to animals, are shedding light on the evolution of animal multicellularity, shaping the genetic contents of the putative premetazoans. However, the assembly quality of the genomes remains poor compared to the major model organisms such as human and fly. Improving the assembly is critical for precise comparative genomics studies and further molecular biological studies requiring accurate sequence information such as enhancer analysis and genome editing. In this report, we present a new strategy to improve the assembly by fully exploiting the information of Illumina mate-pair reads. By visualizing the distance and orientation of the mapped read pairs, we could highlight the regions where possible assembly errors exist in the genome sequence of *Capsaspora*, a lineage of unicellular holozoans. Manual modification of these errors repaired 590 assembly problems in total and reassembled 84 supercontigs into 55. Our telomere prediction analysis using the read pairs containing the pan-eukaryotic telomere-like sequence identified at least 13 chromosomes. The resulting new assembly posed us a re-annotation of 112 genes, including 15 putative receptor protein tyrosine kinases. Our strategy thus provides a useful approach for improving assemblies of draft genomes, and the new *Capsaspora* genome offers us an opportunity to adjust the view on the genome of the unicellular animal ancestor.

## KEYWORDS

*Capsaspora* genome, mate-pair reads, next generation sequencing, origin of multicellularity, unicellular holozoans

## 1 | INTRODUCTION

Unicellular holozoans are the protists closely related to metazoans (Hehenberger et al., 2017; Lang, O'Kelly, Nerad, Gray, & Burger, 2002; Torruella et al., 2015). Recent studies on these unicellular descendants of the animal ancestor are shedding light on the evolution of animal multicellularity. Notably, the comprehensive genome sequencing of major unicellular holozoan clades (the Choanoflagellata, the Filasterea, the Ichthyosporea, and the Corallochytraea) provided

us with an in-depth knowledge on the presence and absence of genes that are considered essential for the evolution of multicellularity, reshaping the putative premetazoan genome. For instance, the genome of *Monosiga brevicollis*, a choanoflagellate, revealed that the unicellular ancestor of metazoans was already equipped with a rich repertoire of genes that are involved in cell adhesion and intercellular communication (King et al., 2008). The genome of *Capsaspora owczarzaki* (from herein *Capsaspora*), a filasterean, indicated an even richer premetazoan genome with the full integrin

machinery, transcription factors that are in animals used for developmental control, and the players for organ growth control such as the Hippo pathway genes (De Mendoza et al., 2013; Seb e-Pedr os, Roger, Lang, King, & Ruiz-Trillo, 2010; Seb e-Pedr os, Zheng, Ruiz-Trillo, & Pan, 2012; Suga et al., 2013). The six additional genomes of the Ichthyosporea and the Corallochytraea pushed down the origin of some of those "multicellularity genes" close to the emergence of the Holozoa (De Mendoza, Suga, Permanyer, Irimia, & Ruiz-Trillo, 2015; Grau-Bov e et al., 2017).

However, the quality of the unicellular holozoan genomes is still far less than those of predominant model organisms. Assembly problems such as large gaps, broken (discontinuous) sequences, and over-scaffoldings have hampered the full exploitation of genetic information such as promoter and enhancer sequences. Moreover, gene annotation could also be affected by the assembly problems, especially for the gene families comprising many members produced by frequent gene duplication and domain shuffling. For example, the comparative genomics analyses have shown that the receptor tyrosine kinase family, which is used for the intercellular communication in the multicellular context, greatly expanded in each lineage of unicellular holozoans (Manning, Young, Miller, & Zhai, 2008; Suga, Torruella, Burger, Brown, & Ruiz-Trillo, 2014; Suga et al., 2008, 2012). However, the heavily duplicated protein domains in the extracellular (receptor) regions often disturb the sequence assembly, leading to a possible gene mis-prediction and an under- or over-estimation of family expansion. Continuous improvement of the genome sequences is critical for further in-depth analyses of unicellular holozoans.

Here, we present a new strategy to improve the assembly quality by the use of insert length information of Illumina mate-pair reads. By visualizing the distance between two paired reads of 5.8 kb insert library and their mapping orientations along the supercontigs, we efficiently highlighted the assembly problems of the *Capsaspora* genome. The 84 supercontigs of the original (version 3) assembly were integrated into 55 supercontigs in the version 4, of which 19 major supercontigs cover 98.7% of the whole sequence. We also amended 590 assembly problems such as gaps and mis-inserted sequences. In addition, our telomere prediction analysis successfully recovered at least 13 chromosomes of *Capsaspora*.

These modifications affected the predictions of 112 genes, which include those involved in intercellular communication such as receptor protein tyrosine kinases (RTKs). Our new strategy is quite efficient for manually improving draft genome assemblies. Moreover, the new *Capsaspora* genome should allow us to perform further investigation that is in need of an accurate genome sequence, and fine-tune our view on the putative premetazoan genome.

## 2 | MATERIALS AND METHODS

### 2.1 | Genomic DNA extraction

We used the same *Capsaspora owczarzaki* culture as in the previous study reporting the genome sequence (Suga et al., 2013). The

genomic DNA of *Capsaspora* was extracted with Blood & Cell Culture DNA Midi Kit (QIAGEN), from three 75 cm<sup>2</sup> flasks with 50 ml ATCC 1034 PYNFH medium. The culture was grown at 23°C for 5 days until it reached confluency. The Qubit (Thermo) quantification found an 82.6 µg DNA in total.

### 2.2 | Mate-pair sequencing

High-quality Illumina 5 kb mate-pair reads (150 bp) were produced by BGI (Shenzhen, China) with HiSeq sequencer (Illumina). The library was prepared using the company's in-house protocol. After removing the adapter sequences, low-quality sequences, and obviously contaminated sequences by SOAPnuke program (Chen et al., 2018), we obtained 7,876,540 reads corresponding to 1,181,481,000 bases. The averaged read coverage on the original *Capsaspora* genome (version 3; 27,967,784 bp) (Suga et al., 2013) is thus 42×. The average insert length ± SD was estimated to be 5,852 ± 309 bp by calculating the distance between the forward and reverse reads that were independently mapped to the genome. The mapping was done with SMALT program (<http://www.sanger.ac.uk/science/tools/smalt-0>) version 0.7.5 with -x option (independent mapping of two reads in a pair). The raw read data are deposited in the Short Read Archive (SRA) of DNA Databank of Japan (DDBJ) under the BioProject PRJDB7484.

### 2.3 | Visualizing the read mating information

The reads of the mate-pair library were mapped on the whole draft genome sequence by the use of SMALT program with -x option. The directions of mate-pair reads were changed in advance so that the two reads in a pair point to each other. The Sanger reads of the 40 kb fosmid library used to assemble the version 3 genome (Suga et al., 2013) were also mapped and utilized for assessing the assembly integrity in a larger span.

To detect the assembly problems, we visualize the mating information (i.e. the distance between the paired reads and their orientations) of the mapped reads by Gnuplot program (<http://www.gnuplot.info/>), which offers a simple interactive user interface including scrolling and scaling. The mating information was also used for gap filling by using the reads that are expected to jump into the gap regions from the distance of the insert length, a strategy similar to the one used in published software including GapFiller (Boetzer & Pirovano, 2012). The algorithm is called DIMP (Draft genome Improvement by Mate Pair library). DIMP was implemented in a series of scripts written by Ruby, which is available on request. Although the script implements an algorithm to detect assembly problems automatically, we detected them by eye in this study.

To validate the efficacy of the strategy, we used the 22nd supercontig (the largest supercontig) of the oyster *Crassostrea gigas* genome (Genbank AFTI00000000) (Zhang et al., 2012). The raw reads were downloaded from the Genbank SRA under the accession number SRA040229. The improved supercontig 22 is downloadable at <http://www.pu-hiroshima.ac.jp/~hsuga/research/>.

We applied DIMP to the version 3 *Capsaspora* genome sequence (BioProject PRJNA20341; also available at [https://figshare.com/authors/Multicellgenome\\_Lab/2628379](https://figshare.com/authors/Multicellgenome_Lab/2628379)), generating the version 4 assembly. It should be noted that the *Capsaspora* version 3 genome shares the same DNA sequence with the unpublished version 2 assembly (only the annotations differ), and thus the supercontigs are designated as 2.N (where N is the supercontig number) also in the version 3 genome. The read mapping viewer Tablet (Milne et al., 2013) was also used to precisely localize the error position. The version 4 *Capsaspora* genome sequence is deposited at <http://www.pu-hiroshima.ac.jp/~hsuga/research/>. The improvement in the version 4 genome was evaluated by manual inspection of the alignment with the version 3 genome done by Geneious program (Digital Biology). The *Capsaspora* genes whose sequences would be affected by these modifications were sought manually on the same alignment. The identities of the affected genes were assessed by the use of Blast and HMMER search (Eddy, 2009), with a special attention paid on the similarity to the receptor tyrosine kinases of *Capsaspora* (Suga et al., 2012).

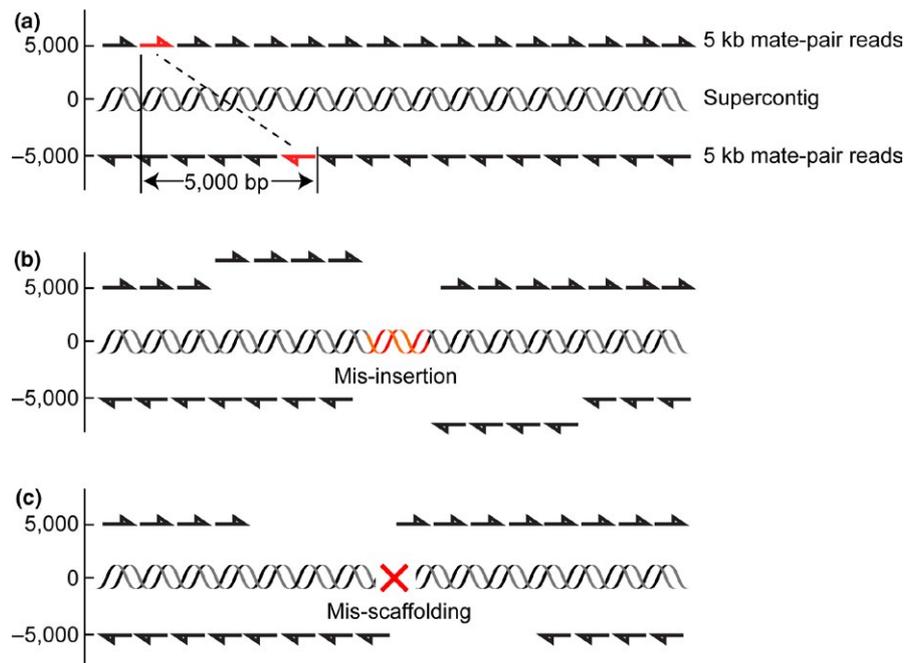
## 2.4 | Verification of new inter-supercontig connections by PCR

We validated the new connections between supercontigs 17, 68, and 6 of the version 3 genome suggested by DIMP by performing a

genomic PCR. We used KOD FX Neo kit (TOYOBO) and the manufacturer's Step-down protocol. For the connection between 68 and 6, a nested PCR was carried out. The used primers are as follows: for the connection between 17 and 68, 5'-ttcagatggggtgaatgggc-3' (forward) and 5'-ttctcggacgttgatca-3' (reverse), and for the connection between 68 and 6, 5'-tgatgcttcacgggcaag-3' (forward primary), 5'-ggaggtcaggtggcaagaa-3' (reverse primary), 5'-gcctcacacctaccagtc-3' (forward nested), and 5'-gttggcaagaacctctgcg-3' (reverse nested).

## 2.5 | Telomere detection

To seek for the putative chromosomal ends, we extracted reads containing five repeats of TTAGGG motif, the major eukaryote telomere sequence, and mapped them to the version 4 assembly together with their counterpart reads. By using the "depth" command of Samtools program (Li et al., 2009), we identified the peaks of the mapping depth along the supercontigs. Prior to the mapping, the extracted read pairs containing the putative telomere sequence are re-oriented so that the telomere positions can be predicted by the orientations of the mapped reads. For the peaks with more than 15 mapping depth, we examined the sequence within the 12 kb region ( $\approx 5.8 + 5.8$  kb) around the peak. We considered a peak positive when it was associated with the supercontig edge or a gap (i.e., a stretch of N) more than 2,000 bp. The mapping of the whole



**FIGURE 1** Strategy of assembly improvement. Our strategy to detect assembly problems by using 5 kb mate-pair reads is schematically shown for three examples: (a) assembly without problem, (b) assembly with a mis-inserted sequence (shown in red), and (c) mis-assembly of two contigs (red cross). Arrows indicate the reads mapped to a supercontig. Reads that have the mating counterparts (shown as another arrow in the reversed direction) in the downstream (right in this scheme) of the supercontig (depicted as a double helix) are shown above, and their counterparts are shown below. Note that the mate-pair read sequences are reversed prior to the mapping so that they point to each other. The distance between the paired reads is reflected also by their vertical position. The vertical distance value is negative when the counterpart read is located in the upstream (left in this scheme). Dotted line in (a) indicates a pair of reads (in red) as an example. All the reads line up at  $\pm 5,000$  position when there is no assembly problem as in (a). By aligning the reads in such two dimensional geometry, an assembly problem is easily detectable as shown in (b) and (c)

mate-pair reads was also examined, and possible mis-scaffoldings were sought near the peaks by searching regions where reads were not mapped in a pair with the correct distance (5.8 kb). If a peak was associated with such putative mis-scaffolded positions, it was considered positive as well.

### 3 | RESULTS AND DISCUSSION

#### 3.1 | Strategy of assembly improvement

The core of our strategy DIMP is to plot the distance between the mapped mate-pair reads on the axis perpendicular to the axis of the reference sequence (Figure 1a). In the plot, only the pairs that both reads are mapped in the same supercontig are shown. By using this approach, we convert the problem of assembly error detection to a pattern recognition problem. For instance, a sequence mis-insertion (Figure 1b) can be recognized as a small heap of the line composed by the points (designated by arrows in Figure 1) representing the mapped position and the distance to the counterpart reads. The most significant advantage of this strategy is the sensitivity. In particular, small indels and mis-scaffoldings are easily detectable by searching the bumps of the line. Another advantage is the low requirement for the read depth. Usually an approximately 10 $\times$  coverage is sufficient as is the case for the 5 kb mate-pair library of the oyster genome. Even with the coverage close to 1 $\times$  (e.g. the 40 kb fosmid reads of *Capsaspora* genome), the assembly problems are detectable, although the sensitivity drops.

By applying this strategy to the region where the assembly suffers a mis-scaffolding, we can detect the problem as two gaps in the lines (Figure 1c). We break such mis-connected sequences into fragments and try to reconnect them to other supercontigs. A similar strategy is also used in the published tool REAPR (Hunt et al., 2013), which uses the mating information for assessing the assembly quality. In our implementation, the script searches the supercontigs in a 1,000 bp window for a region where more than a certain number (e.g., 15) of reads suggest a mis-scaffolding. We then manually inspected the new inter-supercontig connection suggested by the program and reconnect the fragments when the link was validated by the reads mapped to the new supercontig.

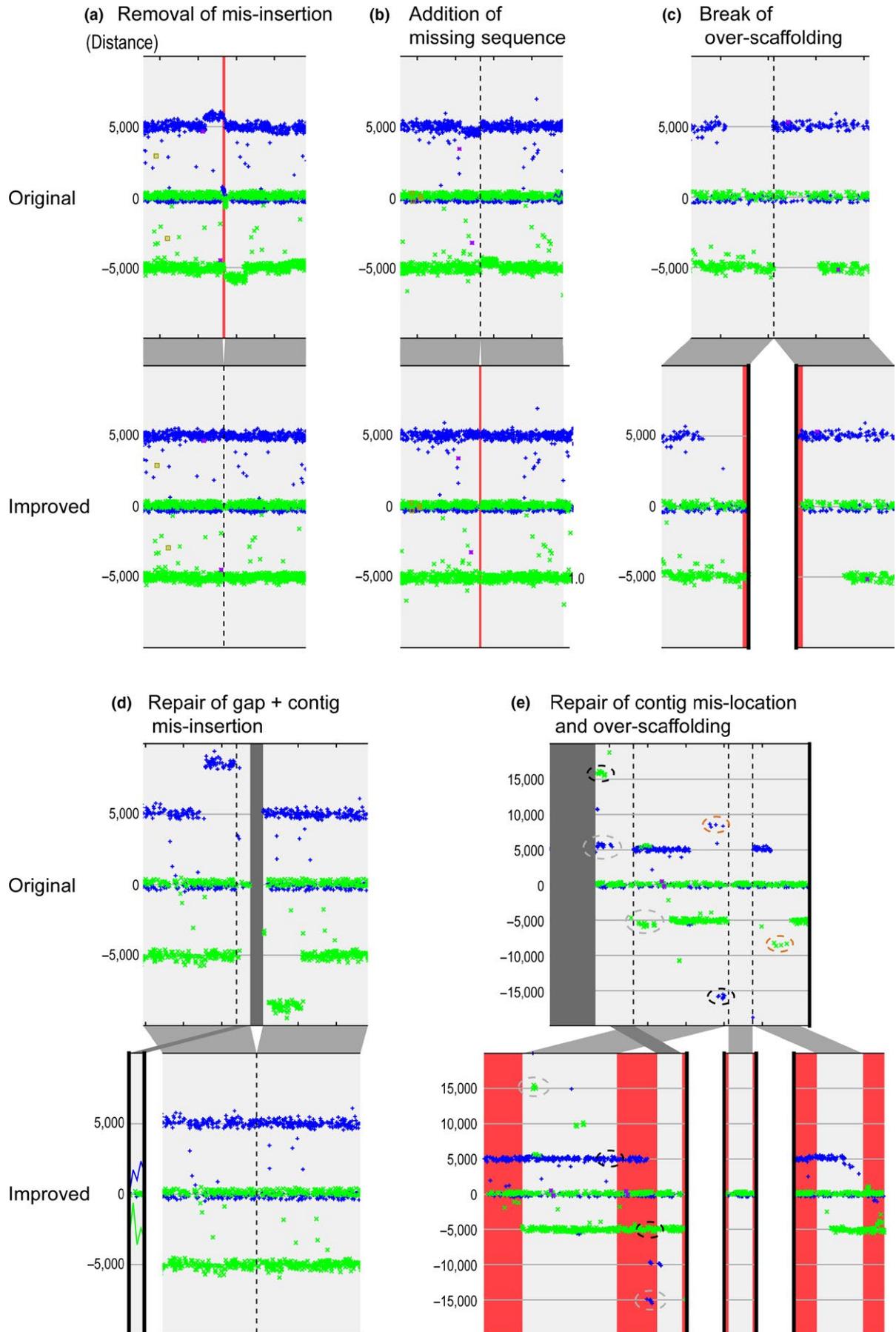
A limitation of this approach is that the assembly discrepancy spanning over a region more than the insert length is hardly detectable. To overcome this limitation, we combine multiple libraries with larger insert lengths if available.

#### 3.2 | Validating the strategy

To validate the strategy, we used the draft oyster genome and the 5 kb mate-pair reads used for the published assembly (Zhang et al., 2012). Our approach efficiently highlighted the assembly problems in the supercontig 22, the largest supercontig of the genome (Figure 2). The manual assessment and modification of the detected problems improved the assembly with high fidelity. A short mis-insertion was removed and the resulting sequence breakage was amended with a small Phrap (<http://www.phrap.org/>) assembly of the reads that were supposed to jump in the gap (Figure 2a). We could also pinpoint the sequence position where a short sequence was missing by carefully examining reads mapped to the region where a deviation was found in the straight line at  $\pm 5,000$  bp (Figure 2b). An over-scaffolding problem, which was caused by a possible misuse of the reads from a larger insert library, is clearly visible as well (Figure 2c). We broke down such supercontigs after confirming that neither the 10 kb nor 20 kb mate-pair reads strongly support this false connection. When a contig mis-insertion is combined with a gap less than 5.8 kb, we fixed the region by removing the gap and the mis-inserted contig (Figure 2d). An even more complicated problem (Figure 2e), such as a combination of translocation and multiple mis-scaffoldings, was amendable by carefully inspecting the positions of the abnormally placed read groups (dotted ovals). It is also possible to detect a sequence inversion by tracing the read pairs that are mapped in an identical direction (Figure 2 violet and olive points), although such pattern did not frequently appear in the supercontig 22. Our modification broke the original supercontig 22 into 41 sub-supercontigs, implying the existence of severe over-scaffoldings in the draft oyster genome.

The current DIMP implementation does not offer a fully automatic improvement. We consider DIMP useful for fixing a relatively small-scaled assembly problem by manual inspection.

**FIGURE 2** Strategy validations using the oyster genome. The mate-pair reads were mapped to the supercontig 22 of the draft oyster genome and the distance between the reads were plotted along the supercontig. The distance value is converted to negative (thus plotted below the reference sequence) when the counterpart read is mapped in the upstream (i.e., left). In each pair, reads that were mapped in forward and reverse directions are represented by blue and green points, respectively, while those that were mapped in an identical direction (suggesting an inversion) are represented by violet and olive green. Five examples (a–e) of assembly error detection are shown. The detected errors were manually corrected, and the improved new supercontigs are shown below the original supercontigs. The reads are independently mapped to the original and improved supercontigs. Red bars indicate the sequences that were manually added to fill the gaps or to extend the edges of supercontig breaks. Corresponding regions between the original and improved sequences are indicated by grey shades between upper and lower panels. Black dotted oval shows the reads indicating the translocation in the original assembly, while gray dotted oval shows the reads incorrectly supporting the original scaffolding at the same place. Orange dotted ovals indicate the reads supporting a wrong scaffolding, which disappear after breaking the connection. +, forward read; x, reverse read; \*, unidirectional forward read; □, unidirectional reverse read; ---, mis-assembly edge/fixed position; ■, gap (Ns); ■, mis-insertion/newly added sequence; |, Scaffold break



### 3.3 | Improving the *Capsaspora* genome

We applied DIMP to the version 3 *Capsaspora* genome sequence (Suga et al., 2013) by using the newly generated Illumina mate-pair reads with 5.8 kb inserts, and produced the version 4 genome (the detailed metrics are compared in Table 1). The *Capsaspora* genome is compact (28 Mb) compared to those of other unicellular holozoans, and the quality of the published version 3 assembly is already high (supercontig N50 = 1.6 Mb). Nevertheless, we detected many assembly discrepancies by applying DIMP to the assembly. We manually filled 353 gaps by collecting the reads that are expected to jump into the gap region (summarized in Table 2; Table S1 for the detail). We also inserted 62 new gaps and 67 new sequences to repair the detected assembly problems. In addition, to amend the mis-insertions and wrongly duplicated regions, we removed 108 sequences, most of which were re-integrated to other parts of the sequence or to different supercontigs.

We also performed reconnections or integrations of 29 out of 84 supercontigs to different supercontigs by using the pairs in which the reads were mapped to different supercontigs (Figure 3a and Table S2). The obtained version 4 genome comprises 55 supercontigs in total. The number of supercontigs spanning more than 50,000 bp regions is 19, which cover 98.7% of the whole sequence.

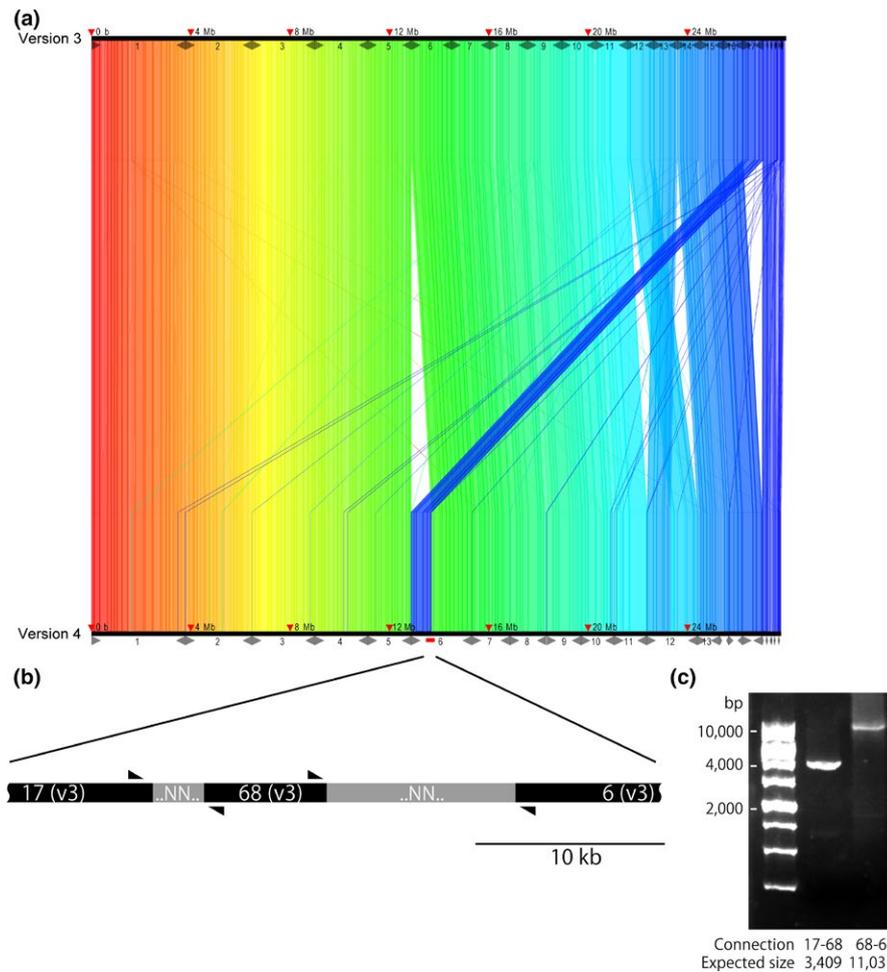
**TABLE 1** Metrics of the *Capsaspora* genome assemblies

Metrics	Version 3	Version 4
Total sequence length	27,967,784 bp	27,768,722 bp
Number of supercontigs	84	55
Supercontig N50	1,617,775 bp	2,049,511 bp
Longest supercontig	3,794,338 bp	3,784,708 bp
Proportion of total >1 Mb supercontig length	81.0%	87.8%

**TABLE 2** Numbers of modifications applied to the *Capsaspora* genome sequence

Modification*	
Gap completely filled	300
Gap partially filled	53
New gap (≥10 bases) inserted	62
New sequence (≥10 bases) inserted	67
Sequence (≥10 bases) removed	108
Inter-supercontig connection/integration	29

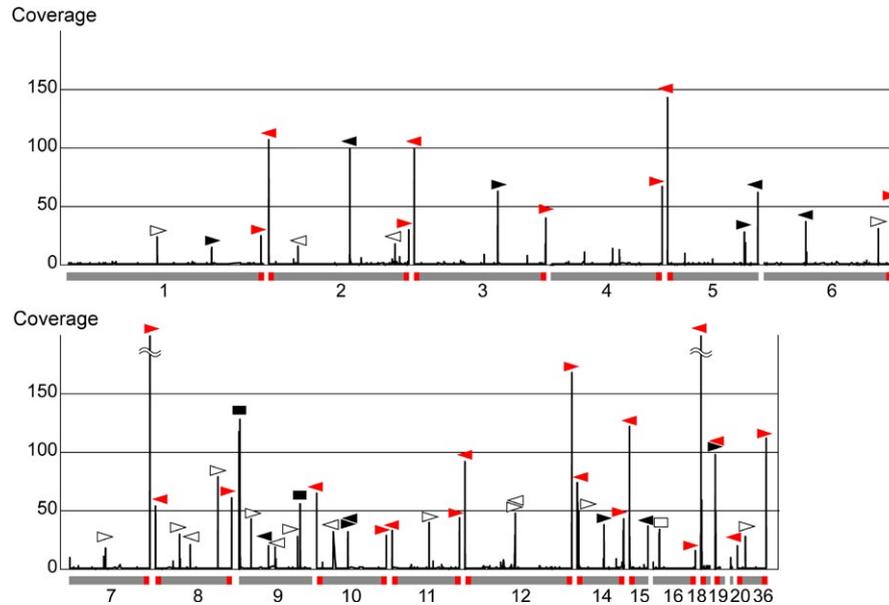
\*The detailed numbers for each supercontig are in Table S1.



**FIGURE 3** Supercontig reconnection and relocations in the *Capsaspora* genome. (a) The version 3 assembly (top) and the version 4 assembly (bottom) are compared using the Murasaki program (Popendorf, Hachiya, Osana, & Sakakibara, 2010) and GMV viewer. The lower threshold of tf-idf (term frequency–inverse document frequency) anchors was set at around 100,000. The numbers and sizes of supercontig are shown at the edges of the diagram. The corresponding regions are connected by color lines. Red line at the bottom indicates a region of supercontig 6 of the version 4 genome, where the supercontigs 17, 68, and 6 of the version 3 genome are connected. (b) The structure of the supercontig 6 of the version 4 genome is partially drawn. Black and grey bars represent the version 3 (v3) supercontigs and the gaps connecting them, respectively. PCR primer positions are indicated by triangles. (c) Result of the genomic PCR to validate the two connections indicated in (b). The expected amplicon sizes are shown at the bottom

**TABLE 3** Frequency of major telomere sequences in the raw reads

Sequence	5× repeats (%)	10× repeats (%)	Typical clade
TTAGGG	3.96	2.91	Vertebrates, many eukaryotes
TTTAGGG	0	0	Plants, <i>Creolimax</i>
TYAGG	0.00254	0.00038	Insects

**FIGURE 4** Telomere prediction on the major supercontigs. The mapping depth (coverage) of the read pairs containing 5'-(TTAGGG)<sub>5</sub>-3' sequence was plotted along the largest 19 supercontigs of the version 4 assembly. The 19 supercontigs are graphically depicted by grey bars with the numbers. Red squares represent the chromosomal termini identified in this study. The peaks with more than 15× coverage were examined in detail. Arrow head indicates the predicted direction of the telomere location, which is ambiguous at the square. Red arrowheads indicate the presence of telomere sequences correctly predicted at the edges of supercontigs. Black and open arrowheads suggest the presence of telomere sequences in the middle of supercontig sequence, implying over-scaffolding during the assembly process, although the open arrowheads are considered false positives (see section 2.5)

To validate the new connections between the version 3 supercontigs, we performed PCR on *Capsaspora* genomic DNA. The 5' end of the new scaffold 6 of version 4 genome comprises the scaffolds 17, 68, and 6 of the version 3 genome, which are connected to each other with long intervening gaps (Figure 3b). The lengths of those two gaps were estimated by the distance between mapped read pairs of the 5.8 kb mate-pair library or the 40 kb fosmid library. We successfully amplified these regions including the gaps (Figure 3c). Although the actual size of the amplicon (~4,000 bp) bridging the supercontigs 17 and 68 is slightly larger than expected (3,409 bp), our PCR results are satisfactorily consistent with the DIMP estimation.

The improvement of *Capsaspora* genome affected the predicted open reading frames of 112 genes (1.3% of total predicted genes) and one pseudogene (Table S3). Fifteen out of these 112 genes were removed from the *Capsaspora* genome, because their nucleotide sequences were fully or partially integrated into other genomic regions after moving to other places, suggesting some redundant information was successfully removed in the new assembly. It appears that the assembler was disturbed by

polymorphic sequences, sometimes regarding two alleles as two duplicated genes.

The predicted functions of affected genes are diverse. Eighty-four out of 112 genes were annotated "hypothetical" in the version 3 assembly. Interestingly, 15 genes (\*, #, † in Table S3) are included in, or similar to specific subfamilies of RTKs, exhibiting the characteristic domain organization seen in *Capsaspora* RTKs such as an array of leucine rich repeat sequences, Sushi domains, and Calx-beta motifs (Suga et al., 2012). It is therefore possible that our modification of *Capsaspora* genome critically influences the annotations of RTK genes and our view on the RTK expansions prior to the evolution of animal multicellularity (Suga et al., 2012, 2014).

### 3.4 | Chromosome termini prediction by searching telomeres

Our strategy successfully recovered 19 major supercontigs covering 98.7% of the whole *Capsaspora* genome sequence. However, it remains unclear whether the supercontigs correspond to the chromosomes because the version 3 assembly does not contain any clear

telomere sequence on the edges of larger supercontigs. Therefore, we performed a prediction of the telomere position in the 19 supercontigs, using the 5.8 kb mate-pair reads.

First, we searched the raw reads for repeats of known telomere sequences. The TTAGGG sequence, which is used in many eukaryotes (Podlevsky, Bley, Omana, Qi, & Chen, 2008), was found as 5× repeats in 0.4% reads (Table 3). Contrary, the TTTAGGG and TYAGG sequences (Podlevsky et al., 2008) that are frequently used in plants and insects, respectively, are virtually absent, although in the ichthyosporean *Creolimax* TTTAGGG is used. Thus, the *Capsaspora* telomere sequence is most likely 5'-(TTAGGG)<sub>n</sub>-3'.

Next, we mapped the read pairs containing the putative telomere sequence on the version 4 assembly (Figure 4). The coverage peak corresponds to the sequence position where telomere sequence is expected in close proximity, within the 5.8 kb upstream or downstream region. By inspecting the mapping orientations, whether the telomere is on the upstream or downstream of the peak is predictable. A peak situated close to the supercontig edges (and pointing to the supercontig periphery) strongly suggests the presence of a telomere at the edge of the supercontig. In addition, by assessing the assembly problems such as mis-scaffoldings and large gaps close to the peak, we were able to extract positive peaks even in the middle of sequences, which imply an over-scaffolding of two different chromosomes.

We successfully recovered 39 positive peaks suggesting the presence of telomeres. All except two (black squares in Figure 4) of them clearly pointed the directions where the telomeres would exist. The 25 out of 38 termini of the 19 largest supercontigs are predicted to be the actual chromosomal ends (red arrowheads). This indicates that the *Capsaspora* genome comprises at least 13 chromosomes. Indeed, a previous study found at least 12 chromosomes by a pulse field gel electrophoresis (Ruiz-Trillo, Lane, Archibald, & Roger, 2006). However, we also detected, in the middle of supercontig sequences, 14 locations where telomeres may exist ectopically. It is possible that the version 4 assembly still suffers over-scaffolding problems, which can be solved by a super long read technique such as MinION (Oxford Nanopore).

### 3.5 | Enhanced quality of the *Capsaspora* genome assembly affects the view on the evolution of multicellularity

In this report, we successfully improved the quality of *Capsaspora* genome sequence by fully exploiting the information of the newly generated 5.8 kb mate-pair Illumina reads. The improved assembly should allow us to perform further studies of molecular biology and genetics requiring accurate genetic information, such as enhancer analysis and genome editing.

Even in the version 4 assembly, however, there remain some incongruous places. For instance, our telomere prediction ectopically found 14 possible chromosome ends, suggesting over-scaffoldings among different chromosomes. Nevertheless, the mapping of 40 kb fosmid paired reads still supports their connections (data not shown).

Sequence polymorphism may account for this inconsistency. In addition, the version 4 genome still contains 36 small (<30,000 bp) supercontigs. It is likely that the assembler failed to integrate them into larger supercontigs due to the sequence polymorphism, because most of them show strong (~90%) sequence similarity to the major 19 supercontigs (data not shown). Further improvement using the super long read device would solve these problems.

For obtaining a better assembly, we could also have started the assembler de novo, using both the old Sanger reads and the new Illumina mate-pair reads together. In this study, however, we decided to improve the old Sanger read assembly with the help of new Illumina reads. One of the main reasons is the consistency of the gene identifiers used in the past studies. For improving the assembly with the consistency between studies secured, DIMP strategy should provide a decent option.

Recent studies on the unicellular holozoans have completely reshaped our traditional view on the genetic contents of the unicellular metazoan ancestor. The putative premetazoans should have a rich repertoire of “multicellularity genes” that was later co-opted for the development and maintenance of animal multicellular system (Fairclough et al., 2013; King et al., 2008; Richter, Fozouni, Eisen, & King, 2018; Sebé-Pedrós, Degnan, & Ruiz-Trillo, 2017; Suga et al., 2013). Such genes include RTKs, which are involved in intercellular communication in animals. Our modifications applied to the *Capsaspora* genome altered amino acid sequences of 112 genes including 15 possible RTK genes, exerting an adjustment of our view on the RTK expansion in the premetazoan history (Suga et al., 2013, 2014). It is possible that a heavily expanded gene family with many repeats of protein domains such as RTKs in unicellular holozoans accounts for the severe disturbance of the correct assembly. Therefore, further DIMP improvements on the available genomes of unicellular holozoans should offer an opportunity to fine-tune the scenario how the unicellular ancestor expanded and co-opted their “multicellularity genes” at the transition to the multicellular system.

### ACKNOWLEDGMENTS

This work was partially supported by JSPS KAKENHI 16K07468 and research grants from the NOVARTIS foundation for the Promotion of Science, ITOH Science Foundation, Naito Foundation, and Prefectural University of Hiroshima (JUTEN) to H.S. IRT was supported by a European Research Council Consolidator Grant (ERC-2012-Co-616960) grant, and grants (BFU2014-57779-P and BFU2017-90114-P) from Ministerio de Economía y Competitividad (MINECO), Agencia Estatal de Investigación (AEI), and Fondo Europeo de Desarrollo Regional (FEDER).

### ORCID

Iñaki Ruiz-Trillo  <https://orcid.org/0000-0001-6547-5304>

Hiroshi Suga  <https://orcid.org/0000-0003-1795-2174>

## REFERENCES

- Boetzer, M., & Pirovano, W. (2012). Toward almost closed genomes with GapFiller. *Genome Biology*, *13*, R56.
- Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., ... Chen, Q. (2018). SOAPnuke: A MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience*, *7*, 1–6.
- De Mendoza, A., Sebé-Pedrós, A., Sestak, M. S., Matejčić, M., Torruella, G., Domazet-Lošo, T., Ruiz-Trillo, I. (2013). Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, E4858–E4866.
- De Mendoza, A., Suga, H., Permanyer, J., Irimia, M., & Ruiz-Trillo, I. (2015). Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. *Elife*, *4*, e08904.
- Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Informatics*, *23*, 205–211.
- Fairclough, S. R., Chen, Z., Kramer, E., Zeng, Q., Young, S., Robertson, H. M., ... King, N. (2013). Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. *Genome Biology*, *14*, R15.
- Grau-Bové, X., Torruella, G., Donachie, S., Suga, H., Leonard, G., Richards, T. A., Ruiz-Trillo, I. (2017). Dynamics of genomic innovation in the unicellular ancestry of animals. *Elife*, *6*, e26036.
- Hehenberger, E., Tikhonenkov, D. V., Kolisko, M., del Campo, J., Esaulov, A. S., Mylnikov, A. P., Keeling, P. J. (2017). Novel predators Reshape Holozoan phylogeny and reveal the presence of a two-component signaling system in the ancestor of animals. *Current Biology*, *27*, 2043–2050.e2046.
- Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., & Otto, T. D. (2013). REAPR: A universal tool for genome assembly evaluation. *Genome Biology*, *14*, R47.
- King, N., Westbrook, M. J., Young, S. L., Kuo, A., Abedin, M., Chapman, J., ... Rokhsar, N. (2008). The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature*, *451*, 783–788.
- Lang, B. F., O'Kelly, C., Nerad, T., Gray, M. W., & Burger, G. (2002). The closest unicellular relatives of animals. *Current Biology*, *12*, 1773–1778.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079.
- Manning, G., Young, S. L., Miller, W. T., & Zhai, Y. (2008). The protist, *Monosiga brevicollis*, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 9674–9679.
- Milne, I., Stephen, G., Bayer, M., Cock, P. J. A., Pritchard, L., Cardle, L., ... Marshall, D. (2013). Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*, *14*, 193–202.
- Podlevsky, J. D., Bley, C. J., Omana, R. V., Qi, X., & Chen, J. J. (2008). The telomerase database. *Nucleic Acids Research*, *36*, D339–D343.
- Popendorf, K., Hachiya, T., Osana, Y., & Sakakibara, Y. (2010). Murasaki: A fast, parallelizable algorithm to find anchors from multiple genomes. *PLoS ONE*, *5*, e12651.
- Richter, D. J., Fozouni, P., Eisen, M. B., & King, N. (2018). Gene family innovation, conservation and loss on the animal stem lineage. *Elife*, *7*, e34226.
- Ruiz-Trillo, I., Lane, C. E., Archibald, J. M., & Roger, A. J. (2006). Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts *Capsaspora owczarzakii* and *Sphaeroforma arctica*. *Journal of Eukaryotic Microbiology*, *53*, 379–384.
- Sebé-Pedrós, A., Degnan, B. M., & Ruiz-Trillo, I. (2017). The origin of Metazoa: A unicellular perspective. *Nature Reviews Genetics*, *18*, 498–512.
- Sebé-Pedrós, A., Roger, A. J., Lang, F. B., King, N., & Ruiz-Trillo, I. (2010). Ancient origin of the integrin-mediated adhesion and signaling machinery. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 10142–10147.
- Sebé-Pedrós, A., Zheng, Y., Ruiz-Trillo, I., & Pan, D. (2012). Premetazoan origin of the Hippo signaling pathway. *Cell Reports*, *1*, 13–20.
- Suga, H., Chen, Z., De Mendoza, A., Sebé-Pedrós, A., Brown, M. W., Kramer, E., ... Ruiz-Trillo, I. (2013). The genome of *Capsaspora* reveals a complex unicellular prehistory of animals. *Nature Communications*, *4*, 2325.
- Suga, H., Dacre, M., De Mendoza, A., Shalchian-Tabrizi, K., Manning, G., & Ruiz-Trillo, I. (2012). Genomic survey of pre-metazoans shows deep conservation of cytoplasmic tyrosine kinases and multiple radiations of receptor tyrosine kinases. *Science Signaling*, *5*, ra35.
- Suga, H., Sasaki, G., Kuma, K., Nishiyori, H., Hirose, N., Su, Z.-H., ... Miyata, T. (2008). Ancient divergence of animal protein tyrosine kinase genes demonstrated by a gene family tree including choanoflagellate genes. *FEBS Letters*, *582*, 815–818.
- Suga, H., Torruella, G., Burger, G., Brown, M. W., & Ruiz-Trillo, I. (2014). Earliest holozoan expansion of phosphotyrosine signaling. *Molecular Biology and Evolution*, *31*, 517–528.
- Torruella, G., De Mendoza, A., Grau-Bové, X., Antó, M., Chaplin, M. A., del Campo, J., ... Ruiz-Trillo, I. (2015). Phylogenomics reveals convergent evolution of lifestyles in close relatives of animals and fungi. *Current Biology*, *25*, 2404–2410.
- Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., ... Wang, J. (2012). The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, *490*, 49–54.

## SUPPORTING INFORMATION

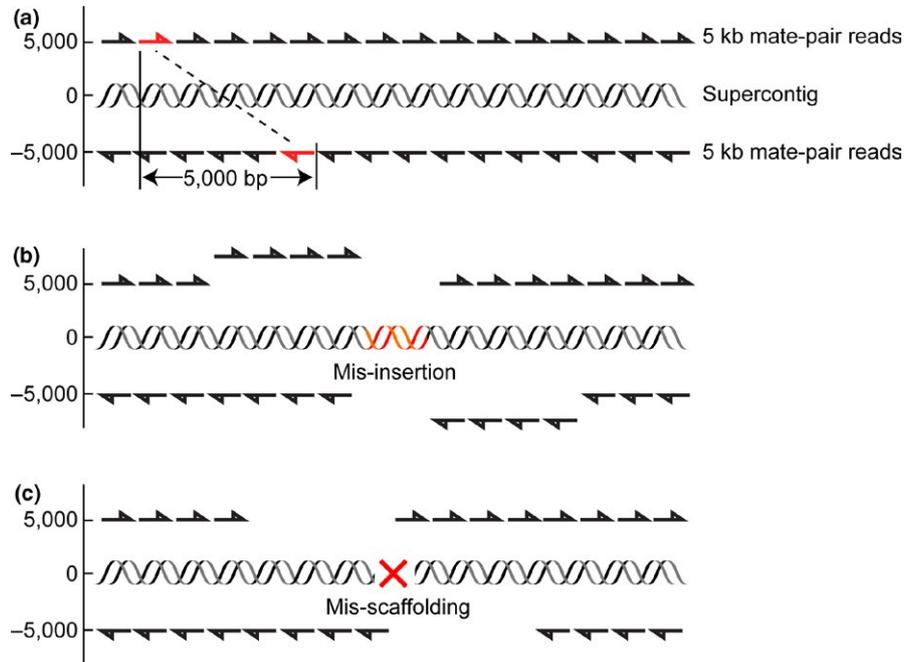
Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Denbo S, Aono K, Kai T, Yagasaki R, Ruiz-Trillo I, Suga H. Revision of the *Capsaspora* genome using read mating information adjusts the view on premetazoan genome. *Develop. Growth Differ.* 2018;00:1–9. <https://doi.org/10.1111/dgd.12587>

# Graphical Abstract

The contents of this page will be used as part of the graphical abstract of html only.

It will not be published as part of main article.



The article describes a novel approach to improve draft genome assembly by the use of Illumina mate-pair reads. We first validated the strategy by improving the published oyster genome, and then applied it to the genome of *Capsaspora*, a unicellular holozoan, by using a newly-generated 5.8 kb Illumina mate-pair reads. Our improvement of *Capsaspora* genome should give an impact on our view to the mechanisms of how the animal multicellular system evolved. Moreover, our new approach will offer a decent option for improving the published genomes of non-model organisms, enabling us more accurate comparative genomics studies and molecular biological studies that are in need of precise genomic sequences, such as promoter analysis and genome editing.