Computational and Structural Biotechnology Journal 19 (2021) 1928–1932

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

A strategy for complete telomere-to-telomere assembly of ciliate macronuclear genome using ultra-high coverage Nanopore data

Guangying Wang [a], Su Wang [a,b,c], Xiaocui Chai [a], Jing Zhang [a], Wentao Yang [a,c], Chuanqi Jiang [a], Kai Chen [a], Wei Miao [a,d,e,*], Jie Xiong [a,*]

[a] Key Laboratory of Aquatic Biodiversity and Conservation, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China
[b] College of Fisheries and Life Science, Dalian Ocean University, Dalian 116023, China
[c] University of Chinese Academy of Sciences, Beijing 100049, China
[d] CAS Center for Excellence in Animal Evolution and Genetics, Kunming 650223, China
[e] State Key Laboratory of Freshwater Ecology and Biotechnology of China, Wuhan 430072, China

ARTICLE INFO

Article history:
Received 3 November 2020
Received in revised form 30 March 2021
Accepted 3 April 2021
Available online 09 April 2021

Keywords:
Ciliated protozoa
Tetrahymena
Complete genome
Nanopore sequencing
Telomeres

ABSTRACT

Ciliates contain two kinds of nuclei: the germline micronucleus (MIC) and the somatic macronucleus (MAC) in a single cell. The MAC usually have fragmented chromosomes. These fragmented chromosomes, capped with telomeres at both ends, could be gene size to several megabases in length among different ciliate species. So far, no telomere-to-telomere assembly of entire MAC genome in ciliate species has been finished. Development of the third generation sequencing technologies allows to generate sequencing reads up to megabases in length that could possibly span an entire MAC chromosome. Taking advantage of the ultra-long Nanopore reads, we established a simple strategy for the complete assembly of ciliate MAC genomes. Using this strategy, we assembled the complete MAC genomes of two ciliate species *Tetrahymena thermophila* and *Tetrahymena shanghaiensis*, composed of 181 and 214 chromosomes telomere-to-telomere respectively. The established strategy as well as the high-quality genome data will provide a useful approach for ciliate genome assembly, and a valuable community resource for further biological, evolutionary and population genomic studies.

1. Introduction

Ciliate separates its germline and somatic genetic information by maintaining two kinds of functionally distinct nuclei: the diploid micronucleus (MIC), and the polyploid macronucleus (MAC) [1,2]. The MIC, like other eukaryotes, contains long chromosomes with centromeres and capped by telomeres. In general, the MAC genome comes from the MIC genome through a MAC differentiation process in the sexual stage (conjugation) of ciliate [3]. During MAC differentiation, the MIC-like chromosomes are fragmented into small pieces at the chromosome breakage sites, and the internal eliminated sequences, which often contain transposable elements, are removed [3]. This process finally results in the MAC containing fragmented chromosomes with length range from gene size to several megabases, and capped by telomere sequences at both ends but without centromeres.

Development of the third generation sequencing technologies, e.g. the Nanopore sequencing, allows to generate sequencing reads up to megabases in length [4], and thus could sometimes sequence an entire MAC chromosome of ciliate by a single read. The generation of such long sequencing reads gives the opportunity to assemble complete MAC genomes of ciliates.

Here, we reported a simple strategy which was used to assemble the complete genome of *Tetrahymena thermophila* and *Tetrahymena shanghaiensis* using high coverage Nanopore sequencing data. The established strategy and genomic data presented in this study will be highly valuable for ciliate research community.

2. Materials and methods

2.1. Cell culture and DNA extraction

T. thermophila SB210 and *T. shanghaiensis* (ATCC accession: 205039) cells were grown in SPP medium [5] and harvested at a density of 250,000 cells/ml. The total DNA was extracted using the Blood & Cell Culture DNA Midi Kit (Q13343, Qiagen, CA, USA)

* Corresponding authors at: Institute of Hydrobiology, Chinese Academy of Sciences, Donghu South Road 7, Wuhan 430072, Hubei Province, China.
E-mail addresses: miaowei@ihb.ac.cn (W. Miao), xiongjie@ihb.ac.cn (J. Xiong).

https://doi.org/10.1016/j.csbj.2021.04.007

following the manufacturer's protocol. The DNA was then purified using the Agencourt AMPure XP beads (A63881, BECKMAN), and the DNA quality and quantity were tested using both NanoDrop One UV–Vis spectrophotometer (Thermo Fisher Scientific, USA) and Qubit 3.0 Fluorometer (Invitrogen, USA).

2.2. Nanopore sequencing

Approximately 10 μg of DNA was size-selected (10–50 Kb) using Blue Pippin (Sage Science, Beverly, MA), and sequencing libraries were constructed using the Ligation sequencing 1D kit (SQK-LSK108, ONT, UK) according to the manufacturer's instructions. Each library was sequenced on R9.4 FlowCells using the PromethION sequencer (ONT, UK) for 48 h at the Genome Center of Nextomics (Wuhan, Hubei, China). Base calling was subsequently performed on fast5 files using the ONT Guppy software (v1.8), and the "passed filter" reads (high quality data) were used for downstream analysis.

2.3. Next generation sequencing

We used Illumina short reads to polish the assembly of Nanopore reads. To this end, DNA libraries were prepared using a TruSeq Nano DNA Library Prep Kit (Illumina, San Diego, CA) according to the manufacturer's protocol. To assist gene prediction, total RNA was extracted from cells in growth and starvation conditions using an RNeasy Protect Cell Mini Kit (Qiagen, Valencia, CA, USA), as described in the *Tetrahymena* Functional Genomics Database [6,7]. DNA and RNA libraries were then sequenced with paired-end reads on a HiSeq 2000 or HiSeq 4000 instrument (Illumina, San Diego, CA).

2.4. Genome assembling and polishing

Genome assembling was performed using 60X Nanopore datasets (Fig. 1). Assemblers, including CANU [8], NECAT (https://github.com/xiaochuanle/NECAT), SHASTA (https://github.com/chanzuckerberg/shasta), Flye [9], and wtdbg2 [10], were used. The parameters for the assemblers are listed as follows: 1) CANU, -fast genomeSize = 100 m; 2) NECAT, GENOME_SIZE = 100,000,00 0 MIN_READ_LENGTH = 3000; 3) SHASTA, default settings; 4) Flye, -g 100 m; 5) wtdbg2, default settings. The performance of CANU and NECAT are far better than the other three assemblers in assembling the MAC chromosomes capped with telomere sequences at both ends. Comparing to CANU, the time cost of NECAT was far less than CANU, and thus NECAT was recommended. Quickmerge (https://github.com/mahulchak/quickmerge) was used to merge the un-closed scaffolds to the 60X genome assemblies (command line: merge_wrapper.py un-closed_scaffolds 60X_assembly). After each round of merging, the closed scaffolds (MAC chromosomes) were extracted, and the left un-closed scaffolds were used to perform the next round of merging. After that, an additional round of merging between the un-closed scaffolds and error corrected telomere-sequences-containing reads was performed using mini-asm (-1-2 -c 1) [11]. Final genome polishing was performed based on the Illumina sequencing data using Pilon (https://github.com/broadinstitute/pilon). The base-level accuracy and completeness of the assembled genomes were assessed by Merqury [12]. Base accuracy was measured using k = 18, and the completeness value was estimated by calculating the fraction of reliable k-mers (genomic substrings of length k) in Illumina reads that were also found in the assembly.

2.5. Gene prediction and repeat analysis

Protein-coding gene models were predicted following our previously established pipeline [13] that combines *de novo* and homology-based methods as well as RNA sequencing (RNA-Seq) data. Briefly, assembled RNA-Seq data were used to generate training gene sets for *ab initio* gene predictions and were also incorporated as cDNA evidence. Evidence Modeler was used to generate a set of gene models combining evidences from all gene prediction programs. Final predicted gene sets were generated after a few manual corrections. Repeat consensus sequences were *de novo* identified and annotated using RepeatModeler (http://www.repeatmasker.org/RepeatModeler/) with default settings, and the output consensus sequence library of RepeatModeler was used to mask the genomes using RepeatMasker (-div 20).

2.6. Data availability

The complete genome sequences and gene annotation files of *T. thermophila* and *T. shanghaiensis* can be accessed from *Tetrahymena* Comparative Genomics Database [14]. The raw sequence data have been deposited in the Genome Sequence Archive in National Genomics Data Center, China National Center for Bioinformation under accession number CRA003720 and the NCBI Gene Expression Omnibus under accession number GSE27971.

3. Results and discussion

T. thermophila is a very useful unicellular model organism for molecular and cellular biology [15]. In 2006, the MAC genome of *T. thermophila* has been sequenced using the Sanger method [16,17], which greatly accelerated the studies using *Tetrahymena* system. The current MAC genome assembly (103.0 Mb, http://ciliate.org/index.php/home/downloads) of *T. thermophila* has 1158 scaffolds, among which 128 (~58.9 Mb) were capped by telomeres with (CCCCAA)n repeats at 5′-end and (GGGGTT)n repeats at 3′-end (hereafter defined as closed scaffolds) and could be regarded as complete MAC chromosomes. However, about a half of genome sequences, composed of 1030 scaffolds, are still not assembled as complete MAC chromosomes (hereafter defined as un-closed scaffolds).

About 1000X Nanopore sequencing data (total DNA of both MAC and MIC, reads N50: 25.8Kb) were obtained to finish the MAC genome assembly. Comparison of different third-generation sequencing data assemblers, including CANU, NECAT, SHASTA, Flye and wtdbg2, were performed. In practice, CANU and NECAT showed better performance on assembling closed scaffolds compared to other assemblers. We divided the ~1000X Nanopore data into different parts, each with ~60X data, and individually assembled them (Fig. 1). We have two reasons to do this division: 1) The MIC reads (contaminations) could be limited below 3X (the copy number ratio between MAC and MIC is 45:2), which will usually be filtered by genome assemblers [4]; 2) At 60X coverage, CANU and NECAT already have good assembling performance and the time cost of assembling could be greatly reduced.

We started from the 1158 scaffolds in current genome assembly of *T. thermophila* (Fig. 1), and divided these scaffolds into two parts: 1) 128 closed scaffolds which assembled as complete MAC chromosomes; 2) 1030 un-closed scaffolds which have not been assembled as MAC chromosomes. For the 128 closed scaffolds, three of them still have gaps (one per each). These gaps were easily closed by aligning the three scaffolds to the 60X Nanopore data assemblies. The left 1030 un-closed scaffolds were iteratively merged with each assembled genome using 60X Nanopore data (Fig. 1). After six rounds of merging using quickmerge, 34 closed scaffolds

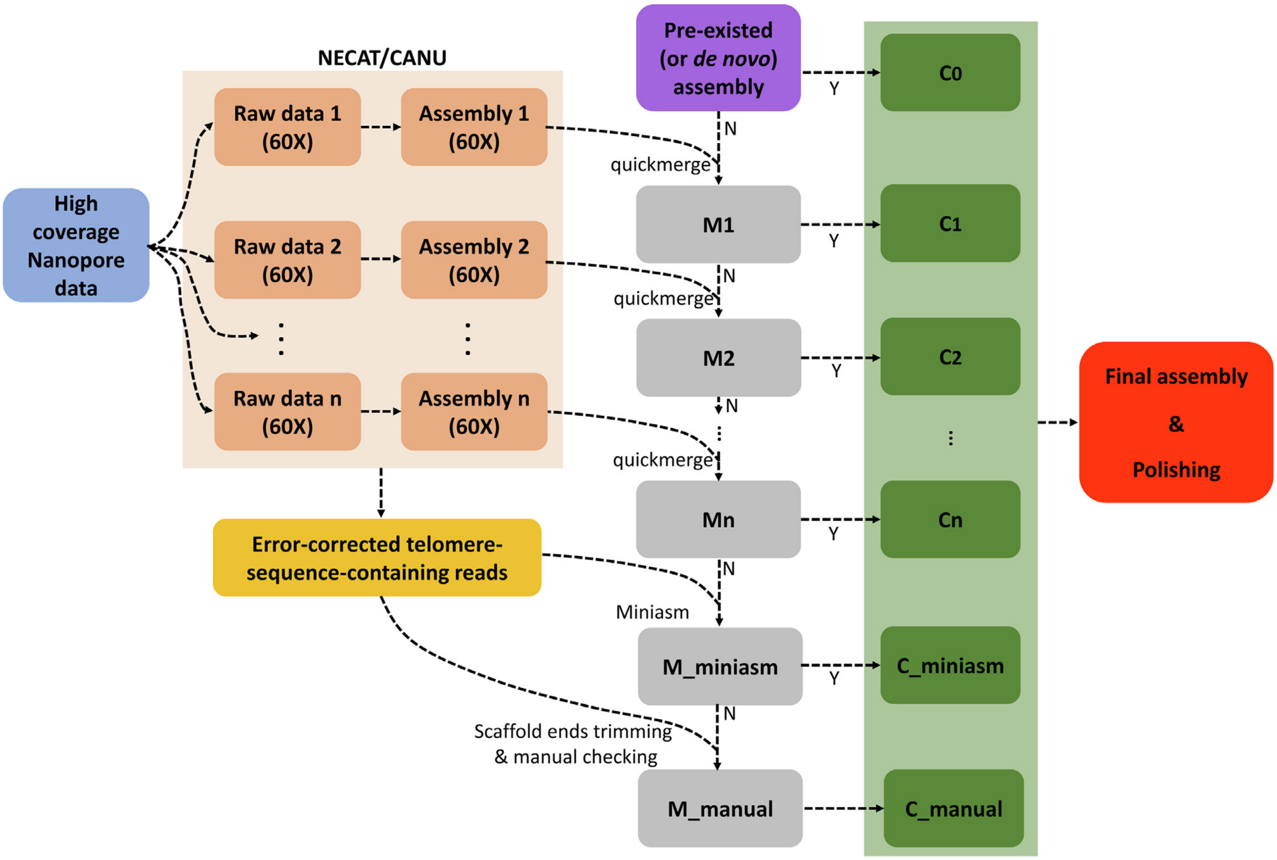

Fig. 1. Diagram showing the strategy to assemble complete MAC genome of ciliate. M1 to Mn, the un-closed scaffolds in each round (1 to n) which do not have telomere sequences at both ends. M_miniasm means the un-closed scaffolds after merging using miniasm. C1 to Cn, the closed scaffolds (MAC chromosomes) in each round (1 to n) which have telomere sequences at both ends. C_miniasm means the closed scaffolds (MAC chromosomes) after merging using miniasm. C_manual means the closed scaffolds after the manual checking of overlaps between TSCR and the un-closed scaffolds (trimmed).

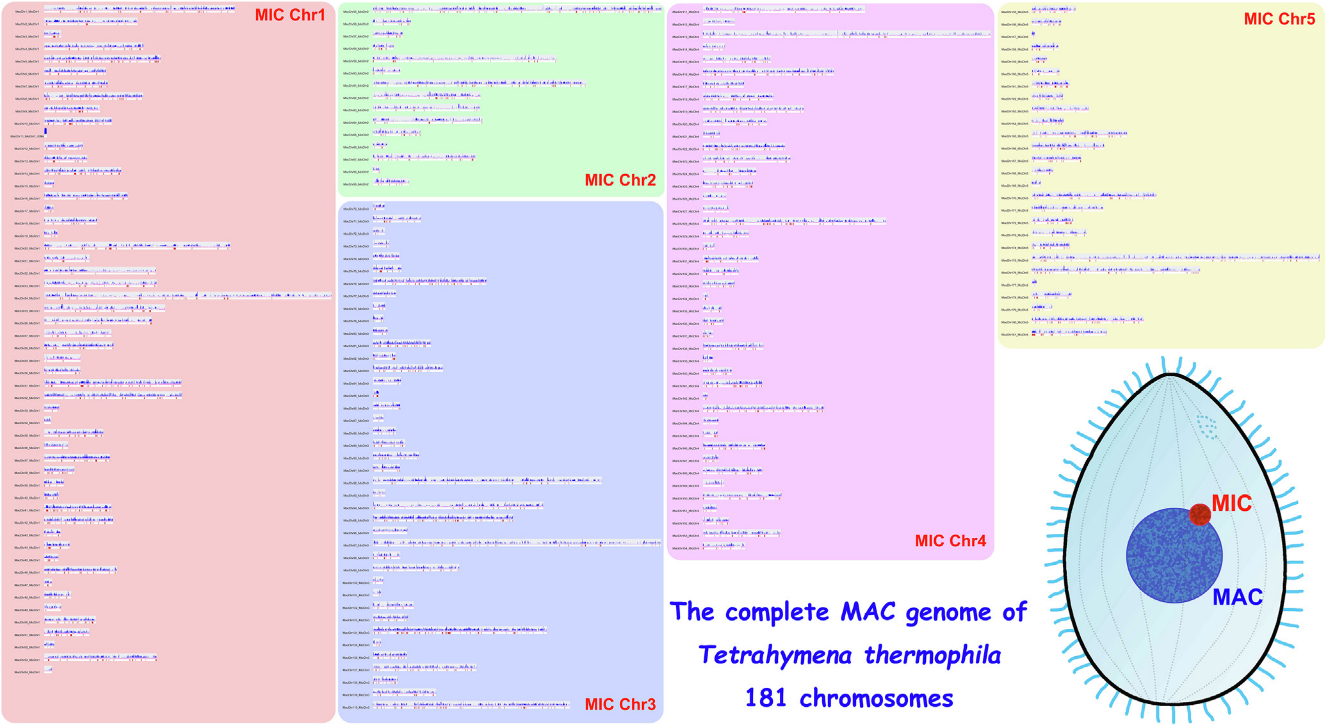

Fig. 2. A full panel of 181 MAC chromosomes of *Tetrahymena thermophila*. For each MAC chromosome, the pink boxes represents the predicted genes; the red boxes represent all the genes that have been named in TGD wiki (http://ciliate.org/); the blue histogram represents the gene expression profile across the chromosome in vegetative growth. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

were newly obtained. After that, we extracted the 256,181 raw telomere-sequence-containing reads (TSCR, reads N50, 28.5 Kb) from Nanopore data (Fig. 1), and sequencing errors were corrected using NECAT. These error corrected TSCR were aligned to the left scaffolds using minimap2, and followed by a new round of assembly using miniasm (Fig. 1), and additional 12 scaffolds with telomere sequences capped at both ends were obtained, and only six scaffolds (3.3 Mb) could not be resolved. To close these six scaffolds, we manually checked the overlaps between TSCR and these scaffolds (Fig. 1), and all of them were closed by trimming their terminal sequences and re-merging with TSCR.

In summary, the complete MAC genome (102.9 Mb) with a total of 181 MAC chromosomes (including rDNA mini-chromosome) were obtained. Previous study has also shown that the MAC genome is contained in 181 chromosomes through physical and genetic mapping [18]. These MAC chromosomes were re-named from 1 to 181 by their order along the five MIC chromosomes. Fig. 2 showed the full panel of the 181 MAC chromosomes. The longest MAC chromosome is 3.26 Mb in length, and the shortest one (excluding rDNA mini-chromosome) is 38 Kb in length. The real N50 of the MAC genome is about 891 Kb. A total of 22 classes of repetitive sequences, which masked 5.2% MAC genome, were identified by RepeatModeler. Among these repetitive sequences, 3.3% were annotated as long interspersed nuclear elements (mainly composed of Leucine-rich repeat genes [13]) while the rest were unclassified. The repetitive sequences in the MAC are not randomly distributed, most of them are enriched in the MAC chromosomes and derived from the pericentromeric and subtelomeric regions of MIC chromosomes [13,19]. In particular, we also found some new genes which missed in the current genome assembly, for example, the alpha 2 subunit of the proteasome.

To test the applicability of this strategy, we generated ~900X Nanopore sequencing data (reads N50: 30.8 Kb) of *T. shanghaiensis*. Instead of using pre-existed assembly, we started from a 60X *de novo* assembly by NECAT, and then followed the strategy showing in Fig. 1. After eight rounds of merging using quickmerge and a round of assembly using miniasm, and followed by additional manual checking, we finally got the complete genome of *T. shanghaiensis* with 214 MAC chromosomes (92.0 Mb) which capped with telomere sequences at both ends. Genome assembly statistics of the two *Tetrahymena* species are shown in Table 1. For *T. shanghaiensis*, we found about 3.6 Mb difference in length between the newly assembled genome and the current version [13]. Closer inspection revealed that the 3.6 Mb sequences in the current version are composed of short scaffolds, 87.4% of which have length less than 5 Kb. Furthermore, after mixing the 3.6 Mb sequences

Table 1
Genome assembly statistics of *T. thermophila* and *T. shanghaiensis.*

	Assemblies in this study		Current assemblies	
	T. thermophila	*T. shanghaiensis*	*T. thermophila*	*T. shanghaiensis*
Assembly size (Mb)	102.9	92.0	103.0	95.6
Number of scaffolds	181	214	1158	2660
Closed Scaffolds	181	214	128	31
N50 (Kb)	891.3	620.0	520.9	153.6
Longest scaffold size (Mb)	3.26	1.98	2.22	0.79
Mean scaffold size (Kb)	568.5	430.0	89.0	36.0
“N” gaps (Kb)	0	0	63.7	90.0

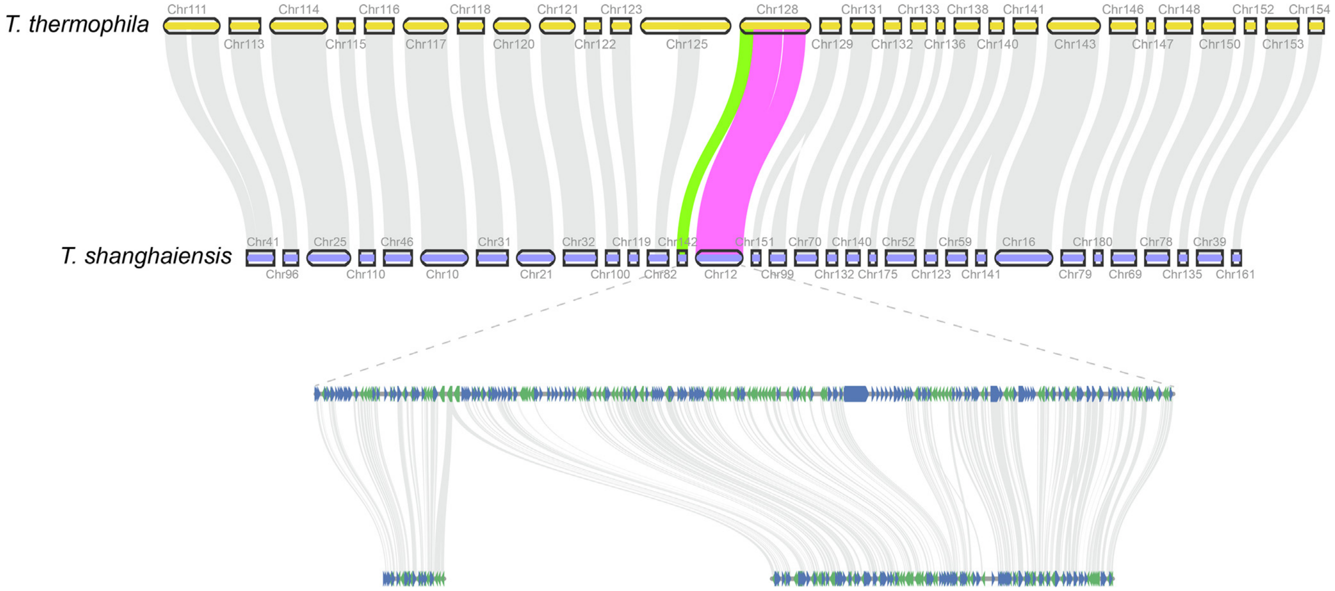

Fig. 3. An example of synteny maps between the two *Tetrahymena* species. The upper panel shows the synteny relationships between *T. thermophila* MAC chromosomes (from MIC chromosome 4) and the corresponding *T. shanghaiensis* chromosomes. The highlighted synteny blocks (with colors green and magenta) show a single chromosome in *T. thermophila* was broken into two separate chromosomes in *T. shanghaiensis*. Note that *T. shanghaiensis* chromosomes are numbered according to their length in descending order because its MIC genome is still not available. The lower panel shows the gene-level matches of the highlighted blocks. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and the newly assembled genome and mapping Illumina reads onto them, we found that the read densities on these short scaffolds were significantly lower than those on other chromosomes (t test, $P < 2.2e-16$). These results indicated that the 3.6 Mb sequences in the current version should probably be the contamination of MIC genome sequences due to the inability to completely separate the MIC from the MAC prior to DNA isolation [13].

We then used Merqury [12] to estimate the assembly accuracy and completeness by comparing k-mers in the assemblies to those found in the high-accuracy Illumina reads. The results show that the per-base error rates for *T. thermophila* and *T. shanghaiensis* assemblies are 3.5e−05 and 2.3e−05, respectively, and the completeness values are 98.9% and 98.3%, respectively. However, because the Illumina reads were sequenced from the mixed MAC and MIC DNA, a small fraction of k-mers in reads were derived from MIC-specific genome sequences. Therefore, the completeness values should be underestimated. All these results suggested that our assembled genomes are highly accurate and complete. Based on the complete assemblies, we could study gene synteny between the two species at the chromosome level using MCScanX [20]. Interestingly, we found some cases that a single chromosome in *T. thermophila* was evolutionarily broken into two separate chromosomes in *T. shanghaiensis*, which can help understanding why *T. shanghaiensis* contains relatively more chromosomes (Fig. 3). We anticipate that the established strategy can probably be used directly or with a slight adaptation to assemble complete MAC genomes of other ciliate species.

CRediT authorship contribution statement

Guangying Wang: Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Su Wang:** Formal analysis, Investigation. **Xiaocui Chai:** Methodology, Formal analysis. **Jing Zhang:** Formal analysis, Writing - review & editing. **Wentao Yang:** Data curation. **Chuanqi Jiang:** Writing - review & editing. **Kai Chen:** Writing - review & editing. **Wei Miao:** Conceptualization, Methodology, Supervision, Writing - review & editing. **Jie Xiong:** Conceptualization, Methodology, Formal analysis, Investigation, Supervision, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China, grant number 31672281, 31872221 (to J.X.), 31900316 (to G.W.) and 31525021 (to W.M.); Knowledge Innovation Program of the Chinese Academy of Sciences (to J.X.); Youth Innovation Promotion Association, Chinese Academy of Sciences (to J.X.). The bioinformatics analysis was aided by the computing resources of the Wuhan Branch, Supercomputing Center, Chinese Academy of Sciences, China.

References

[1] Lynn D. The ciliated protozoa: characterization, classification, and guide to the literature. 3rd ed. Springer; 2008.
[2] Gorovsky MA. Macro- and micronuclei of *Tetrahymena pyriformis*: a model system for studying the structure and function of eukaryotic nuclei. J Protozool 1973;20(1):19–25.
[3] Orias E. Toward sequencing the Tetrahymena genome: exploiting the gift of nuclear dimorphism. J Eukaryot Microbiol 2000;47(4):328–33.
[4] Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 2018;36:338–45.
[5] Cassidy-Hanley DM. *Tetrahymena* in the laboratory: strain resources, methods for culture, maintenance, and storage. Methods Cell Biol 2012;109:237–76.
[6] Xiong J, Lu X, Zhou Z, Chang Y, Yuan D, Tian M, et al. Transcriptome analysis of the model protozoan, *Tetrahymena thermophila*, using deep RNA sequencing. PLoS One 2012;7(2):e30630.
[7] Xiong J, Lu YM, Feng JM, Yuan DX, Tian M, Chang Y, et al. *Tetrahymena* functional genomics database (TetraFGD): an integrated resource for *Tetrahymena* functional genomics. Database 2013;2013. bat008.
[8] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 2017;27(5):722–36.
[9] Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 2019;37:540–6.
[10] Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods 2019.
[11] Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 2016;32(14):2103–10.
[12] Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 2020;21(1):245.
[13] Xiong J, Yang WT, Chen K, Jiang CQ, Ma Y, Chai XC, et al. Hidden genomic evolution in a morphospecies – the landscape of rapidly evolving genes in *Tetrahymena*. PLoS Biol 2019;17(6).
[14] Yang W, Jiang C, Zhu Y, Chen K, Wang G, Yuan D, et al. Tetrahymena comparative genomics database (TCGD): a community resource for *Tetrahymena*. Database (Oxford) 2019;2019.
[15] Ruehle MD, Orias E, Pearson CG. *Tetrahymena* as a unicellular model eukaryote: genetic and genomic tools. Genetics 2016;203(2):649–65.
[16] Coyne RS, Thiagarajan M, Jones KM, Wortman JR, Tallon LJ, Haas BJ, et al. Refined annotation and assembly of the *Tetrahymena thermophila* genome sequence through EST analysis, comparative genomic hybridization, and targeted gap closure. BMC Genomics 2008;9:562.
[17] Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, et al. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol 2006;4(9):1620–42.
[18] Coyne RS, Stover NA, Miao W. Whole genome studies of Tetrahymena. Methods Cell Biol 2012;109:53–81.
[19] Hamilton EP, Kapusta A, Huvos PE, Bidwell SL, Zafar N, Tang HB, et al. Structure of the germline genome of *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome. Elife 2016;5:e19090.
[20] Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 2012;40(7).