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

Guangying Wang[1], Xiaocui Chai[1, 2], Jing Zhang[1], Wentao Yang[1, 2], Chuanqi Jiang[1], Kai Chen[1], Wei Miao[1, 3, 4]*, Jie Xiong[1]*

[1]Key Laboratory of Aquatic Biodiversity and Conservation, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China

[2]University of Chinese Academy of Sciences, Beijing 100049, China

[3]CAS Center for Excellence in Animal Evolution and Genetics, Kunming 650223, China

[4]State Key Laboratory of Freshwater Ecology and Biotechnology of China, Wuhan 430072, China

*Corresponding author

Email:

miaowei@ihb.ac.cn

xiongjie@ihb.ac.cn

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 is 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 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.

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) (Gorovsky, 1973; Lynn, 2008). The MIC, like other eukaryotes, usually contains long chromosomes with centromeres and capped by telomeres. In general, the MAC genome comes from the MIC genome through a so-called MAC differentiation process in the sexual stage (conjugation) of ciliate (Orias, 2000). In MAC differentiation, the MIC-like chromosomes are fragmented into small pieces at the chromosome breakage sites (CBSs), and the internal eliminated sequences (IESs), which contain transposable elements, are removed (Orias, 2000). 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 (Jain et al., 2018), 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 more complete MAC genomes of ciliates.

Here, we reported a simple strategy which was used to assemble the complete genome of *T. thermophila* and *T. shanghaiensis* using high coverage Nanopore sequencing data.

Materials and Methods

Cell culture and DNA extraction

T. thermophila SB210 and *T. shanghaiensis* (ATCC accession: 205039) cells were grown in SPP medium (Cassidy-Hanley, 2012) 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) following the manufacturer's protocol. The DNA was then purified using the Agencourt AMPure XP beads (A63881, BECKMAN), and the DNA quality and quantity was tested using both NanoDrop One UV-Vis spectrophotometer (Thermo Fisher Scientific, USA) and Qubit 3.0 Fluorometer (Invitrogen, USA).

Nanopore sequencing

Approximately 10 µg of DNA was size-selected (10-50 Kb) using Blue Pippin (Sage Science, Beverly, MA), and sequencing library was 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 hours. Base calling was subsequently performed on fast5 files using the ONT Albacore software (v0.8.4), and the “passed filter” reads (high quality data) were used for downstream analysis.

Genome assembling and polishing

Genome assembling was performed using 60X Nanopore datasets. Assemblers, including CANU (Koren et al., 2017), NECAT (https://github.com/xiaochuanle/NECAT), SHASTA (https://github.com/chanzuckerberg/shasta), Flye (Kolmogorov, Yuan, Lin, & Pevzner, 2019), and wtdbg2 (Ruan & Li, 2019), were used. The parameters for the assemblers are listed as follows: 1) CANU, -fast genomeSize=100m; 2) NECAT, GENOME_SIZE=100000000 MIN_READ_LENGTH=3000; 3) SHASTA, default settings; 4) Flye, -g 100m; 5) wtdbg2, default settings. The performance of CANU and NECAT far better than three other assemblers in assembling the MAC chromosomes capped with telomere sequences in 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 addition round of merging between the un-closed scaffolds and error corrected telomere-sequences-containing reads was performed using miniasm (-1 -2 -c 1) (Li, 2016). Final genome polishing was performed based on the Illumina sequencing data using Pilon (https://github.com/broadinstitute/pilon).

Results and Discussion

T. thermophila is a very useful unicellular model organism for molecular and cellular biology (Ruehle, Orias, & Pearson, 2016). In 2006, the MAC genome of *T. thermophila* has been sequenced using the Sanger method (Coyne et al., 2008; Eisen et al., 2006), 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 C4A2 repeats at 5'-end and G4T2 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: 25.8 Kb) 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 (Figure 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 (Jain et al., 2018); 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* (Figure 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 (Figure 1). After six rounds of merging using quickmerge, 34 closed scaffolds were newly obtained. After that, we extracted the 256,181 raw telomere-sequence-containing reads (TSCR, reads N50, 28.5 Kb) from Nanopore data (Figure 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 (Figure 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 (Figure 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. These MAC chromosomes were re-named from 1 to 181 by their order along the five MIC chromosomes. Figure 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 ~891 Kb. A total of 22 classes of repetitive sequences, which masked 5.2% MAC genome, were identified by RepeatModeler. 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 (Hamilton et al., 2016; Xiong et al., 2019). 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 using pre-existed assembly, we started from a 60X *de novo* assembly by NECAT, and then followed the strategy showing in Figure 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. 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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China Grant no. 31525021, 91631303 (to W.M.), 31672281, 31872221 (to J.X.) and 31900316 (to G.W.); 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.

Author contributions

W.M. and J.X. designed the project. J.X., G.W. and W.Y. assembled and annotated the genome. X.C., J.Z., C.J. and K.C. prepared DNA samples for sequencing. J.X. and G.W. wrote the manuscript. All authors read, revised and approved the final manuscript.

Data accessibility

The complete genome sequences of *T. thermophila* and *T. shanghaiensis* can be accessed from http://ciliate.ihb.ac.cn/tcgd/download.html.

References

Cassidy-Hanley, D. M. (2012). *Tetrahymena* in the laboratory: strain resources, methods for culture, maintenance, and storage. *Methods in Cell Biology, 109*, 237-276. doi:10.1016/B978-0-12-385967-9.00008-6

Coyne, R. S., Thiagarajan, M., Jones, K. M., Wortman, J. R., Tallon, L. J., Haas, B. J., . . . Methe, B. A. (2008). Refined annotation and assembly of the *Tetrahymena thermophila* genome sequence through EST analysis, comparative genomic hybridization, and targeted gap closure. *BMC Genomics, 9*, 562. doi:10.1186/1471-2164-9-562

Eisen, J. A., Coyne, R. S., Wu, M., Wu, D. Y., Thiagarajan, M., Wortman, J. R., . . . Orias, E. (2006). Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PloS Biology, 4*(9), 1620-1642. doi:10.1371/journal.pbio.0040286

Gorovsky, M. A. (1973). Macro - and micronuclei of *Tetrahymena pyriformis*: A model system for studying the structure and function of eukaryotic nuclei. *The Journal of Protozoology, 20*(1), 19-25. doi:10.1111/j.1550-7408.1973.tb05995.x

Hamilton, E. P., Kapusta, A., Huvos, P. E., Bidwell, S. L., Zafar, N., Tang, H. B., . . . Coyne, R. S. (2016). Structure of the germline genome of *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome. *Elife, 5*. doi:10.7554/eLife.19090.001

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., . . . Fiddes, I. T. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology, 36*, 338-345. doi:10.1038/nbt.4060

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology, 37*, 540-546. doi:10.1038/s41587-019-0072-8

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research, 27*(5), 722-736. doi:10.1101/gr.215087.116

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics, 32*(14), 2103-2110. doi:10.1093/bioinformatics/btw152

Lynn, D. (2008). *The ciliated protozoa: characterization, classification, and guide to the literature* (3rd ed.). Berlin: Springer.

Orias, E. (2000). Toward sequencing the Tetrahymena genome: exploiting the gift of nuclear dimorphism. *Journal of Eukaryotic Microbiology, 47*(4), 328-333. doi:10.1111/j.1550-7408.2000.tb00057.x

Ruan, J., & Li, H. (2019). Fast and accurate long-read assembly with wtdbg2. *Nature Methods*. doi:10.1038/s41592-019-0669-3

Ruehle, M. D., Orias, E., & Pearson, C. G. (2016). *Tetrahymena* as a Unicellular Model Eukaryote: Genetic and Genomic Tools. *Genetics, 203*(2), 649-665. doi:10.1534/genetics.114.169748

Xiong, J., Lu, X., Zhou, Z., Chang, Y., Yuan, D., Tian, M., . . . Orias, E. (2012). Transcriptome analysis of the model protozoan, *Tetrahymena thermophila*, using deep RNA sequencing. *PloS one, 7*(2), e30630. doi:10.1371/journal.pone.0030630

Xiong, J., Yang, W., Chen, K., Jiang, C., Ma, Y., Chai, X., . . . Miao, W. (2019). Hidden genomic evolution in a morphospecies-The landscape of rapidly evolving genes in *Tetrahymena*. *PloS Biology, 17*(6), e3000294. doi:10.1371/journal.pbio.3000294

Figure legends

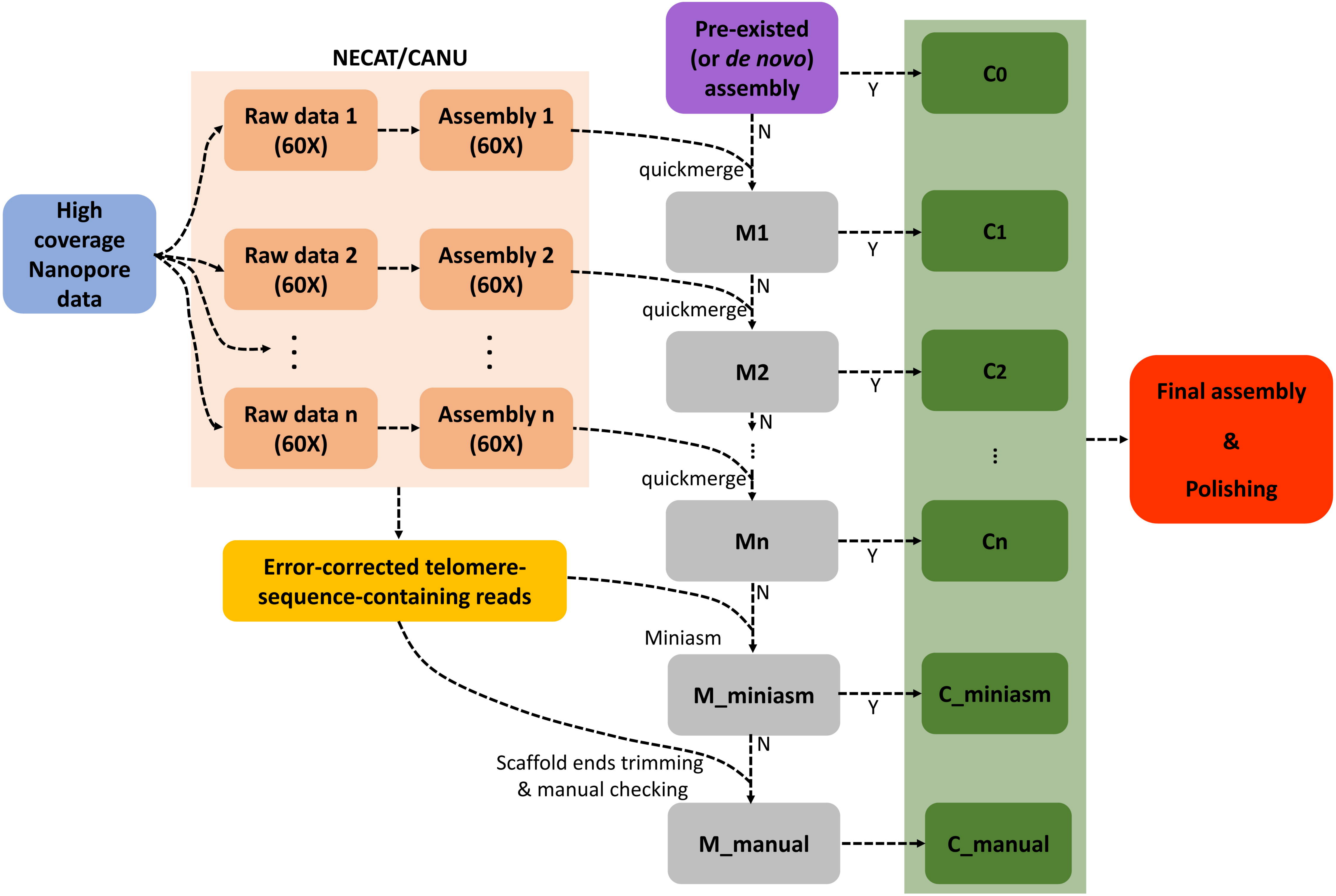
Figure 1. Diagram showing the strategy to assemble complete MAC genome of ciliate. M1 to Mn, the un-closed scaffolds in each round (1 to n) which do not have telomere sequences in both ends. M_miniasm means the un-closed scaffolds after merging using miniasm. C1 to Cn, the closed scaffolds (MAC chromosomes)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

Figure 2. A full panel of 181 MAC chromosomes of *T. 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 (Xiong et al., 2012).

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

	Assembly in this study		Current assembly	
	T. thermophila	*T. shanghaiensis*	*T. thermophila*[†]	*T. shanghaiensis*[‡]
Assembly size (Mb)	102.9	92.0	103.0	95.6
Number of scaffolds	181	214	1,158	2,660
Closed Scaffolds[§]	181	214	128	31
Scaffold 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

[†] Genome data from website: http://ciliate.org/index.php/home/downloads
[‡] Genome data from (Xiong et al., 2019)
[§] Scaffolds capped with telomeres at both ends
[¶] Sum of all "N" nucleotides present in the genome assembly


NECAT/CANU
High coverage Nanopore data
Raw data 1 (60X)
Assembly 1 (60X)
Raw data 2 (60X)
Assembly 2 (60X)
Raw data n (60X)
Assembly n (60X)
Error-corrected telomere-sequence-containing reads
Pre-existed (or *de novo*) assembly
Y
N
quickmerge
M1
quickmerge
M2
quickmerge
Mn
Miniasm
M_miniasm
Scaffold ends trimming & manual checking
M_manual
C0
C1
C2
Cn
C_miniasm
C_manual
Final assembly & Polishing

MIC Chr1
MacChr1_MicChr1
MacChr2_MicChr1
MacChr3_MicChr1
MacChr4_MicChr1
MacChr5_MicChr1
MacChr6_MicChr1
MacChr7_MicChr1
MacChr8_MicChr1
MacChr9_MicChr1
MacChr10_MicChr1
MacChr11_MicChr1_rDNA
MacChr12_MicChr1
MacChr13_MicChr1
MacChr14_MicChr1
MacChr15_MicChr1
MacChr16_MicChr1
MacChr17_MicChr1
MacChr18_MicChr1
MacChr19_MicChr1
MacChr20_MicChr1
MacChr21_MicChr1
MacChr22_MicChr1
MacChr23_MicChr1
MacChr24_MicChr1
MacChr25_MicChr1
MacChr26_MicChr1
MacChr27_MicChr1
MacChr28_MicChr1
MacChr29_MicChr1
MacChr30_MicChr1
MacChr31_MicChr1
MacChr32_MicChr1
MacChr33_MicChr1
MacChr34_MicChr1
MacChr35_MicChr1
MacChr36_MicChr1
MacChr37_MicChr1
MacChr38_MicChr1
MacChr39_MicChr1
MacChr40_MicChr1
MacChr41_MicChr1
MacChr42_MicChr1
MacChr43_MicChr1
MacChr44_MicChr1
MacChr45_MicChr1
MacChr46_MicChr1
MacChr47_MicChr1
MacChr48_MicChr1
MacChr49_MicChr1
MacChr50_MicChr1
MacChr51_MicChr1
MacChr52_MicChr1
MacChr53_MicChr1
MacChr54_MicChr1
MacChr55_MicChr2
MacChr56_MicChr2
MacChr57_MicChr2
MacChr58_MicChr2
MacChr59_MicChr2
MacChr60_MicChr2
MacChr61_MicChr2
MacChr62_MicChr2
MacChr63_MicChr2
MacChr64_MicChr2
MacChr65_MicChr2
MacChr66_MicChr2
MacChr67_MicChr2
MacChr68_MicChr2
MacChr69_MicChr2
MIC Chr2
MacChr70_MicChr3
MacChr71_MicChr3
MacChr72_MicChr3
MacChr73_MicChr3
MacChr74_MicChr3
MacChr75_MicChr3
MacChr76_MicChr3
MacChr77_MicChr3
MacChr78_MicChr3
MacChr79_MicChr3
MacChr80_MicChr3
MacChr81_MicChr3
MacChr82_MicChr3
MacChr83_MicChr3
MacChr84_MicChr3
MacChr85_MicChr3
MacChr86_MicChr3
MacChr87_MicChr3
MacChr88_MicChr3
MacChr89_MicChr3
MacChr90_MicChr3
MacChr91_MicChr3
MacChr92_MicChr3
MacChr93_MicChr3
MacChr94_MicChr3
MacChr95_MicChr3
MacChr96_MicChr3
MacChr97_MicChr3
MacChr98_MicChr3
MacChr99_MicChr3
MacChr100_MicChr3
MacChr101_MicChr3
MacChr102_MicChr3
MacChr103_MicChr3
MacChr104_MicChr3
MacChr105_MicChr3
MacChr106_MicChr3
MacChr107_MicChr3
MacChr108_MicChr3
MacChr109_MicChr3
MacChr110_MicChr3
MIC Chr3
MacChr111_MicChr4
MacChr112_MicChr4
MacChr113_MicChr4
MacChr114_MicChr4
MacChr115_MicChr4
MacChr116_MicChr4
MacChr117_MicChr4
MacChr118_MicChr4
MacChr119_MicChr4
MacChr120_MicChr4
MacChr121_MicChr4
MacChr122_MicChr4
MacChr123_MicChr4
MacChr124_MicChr4
MacChr125_MicChr4
MacChr126_MicChr4
MacChr127_MicChr4
MacChr128_MicChr4
MacChr129_MicChr4
MacChr130_MicChr4
MacChr131_MicChr4
MacChr132_MicChr4
MacChr133_MicChr4
MacChr134_MicChr4
MacChr135_MicChr4
MacChr136_MicChr4
MacChr137_MicChr4
MacChr138_MicChr4
MacChr139_MicChr4
MacChr140_MicChr4
MacChr141_MicChr4
MacChr142_MicChr4
MacChr143_MicChr4
MacChr144_MicChr4
MacChr145_MicChr4
MacChr146_MicChr4
MacChr147_MicChr4
MacChr148_MicChr4
MacChr149_MicChr4
MacChr150_MicChr4
MacChr151_MicChr4
MacChr152_MicChr4
MacChr153_MicChr4
MacChr154_MicChr4
MIC Chr4
MIC Chr5
MacChr155_MicChr5
MacChr156_MicChr5
MacChr157_MicChr5
MacChr158_MicChr5
MacChr159_MicChr5
MacChr160_MicChr5
MacChr161_MicChr5
MacChr162_MicChr5
MacChr163_MicChr5
MacChr164_MicChr5
MacChr165_MicChr5
MacChr166_MicChr5
MacChr167_MicChr5
MacChr168_MicChr5
MacChr169_MicChr5
MacChr170_MicChr5
MacChr171_MicChr5
MacChr172_MicChr5
MacChr173_MicChr5
MacChr174_MicChr5
MacChr175_MicChr5
MacChr176_MicChr5
MacChr177_MicChr5
MacChr178_MicChr5
MacChr179_MicChr5
MacChr180_MicChr5
MacChr181_MicChr5
MIC
MAC

The complete MAC genome of

Tetrahymena thermophila

181 chromosomes