Insights into an Extensively Fragmented Eukaryotic Genome: De Novo Genome Sequencing of the Multinuclear Ciliate *Uroleptopsis citrina*

Weibo Zheng[1,2,†], Chundi Wang[1,†], Ying Yan[1,†], Feng Gao[1,3,*], Thomas G. Doak[4,5,*], and Weibo Song[1,6]

[1]Laboratory of Protozoology, Institute of Evolution & Marine Biodiversity, Ocean University of China, Qingdao, China

[2]Center for Mechanisms of Evolution, Arizona State University, Tempe, USA

[3]Key Laboratory of Mariculture, Ocean University of China, Ministry of Education, Qingdao, China

[4]Department of Biology, Indiana University, Bloomington

[5]National Center for Genome Analysis Support, Indiana University, Bloomington

[6]Laboratory for Marine Biology and Biotechnology, Qingdao National Laboratory for Marine Science and Technology, Qingdao, China

[†]These authors contributed equally to this work.

*Corresponding authors: E-mails: gaof@ouc.edu.cn; tdoak@iu.edu.

Accepted: March 1, 2018

Data deposition: The genome of *Uroleptopsis citrina* has been deposited at GenBank under the accession GCA_001653735.1.

Abstract

Ciliated protists are a large group of single-celled eukaryotes with separate germline and somatic nuclei in each cell. The somatic genome is developed from the zygotic nucleus through a series of chromosomal rearrangements, including fragmentation, DNA elimination, de novo telomere addition, and DNA amplification. This unique feature makes them perfect models for research in genome biology and evolution. However, genomic research of ciliates has been limited to a few species, owing to problems with DNA contamination and obstacles in cultivation. Here, we introduce a method combining telomere-primer PCR amplification and high-throughput sequencing, which can reduce DNA contamination and obtain genomic data efficiently. Based on this method, we report a draft somatic genome of a multimacronuclear ciliate, *Uroleptopsis citrina*. 1) The telomeric sequence in *U. citrina* is confirmed to be C4A4C4A4C4 by directly blunt-end cloning. 2) Genomic analysis of the resulting chromosomes shows a "one-gene one-chromosome" pattern, with a small number of multiple-gene chromosomes. 3) Amino acid usage is analyzed, and reassignment of stop codons is confirmed. 4) Chromosomal analysis shows an obvious asymmetrical GC skew and high bias between A and T in the subtelomeric regions of the sense-strand, with the detection of an 11-bp high AT motif region in the 3′ subtelomeric region. 5) The subtelomeric sequence also has an obvious 40 nt strand oscillation of nucleotide ratio. 6) In the 5′ subtelomeric region of the coding strand, the distribution of potential TATA-box regions is illustrated, which accumulate between 30 and 50 nt. This work provides a valuable reference for genomic research and furthers our understanding of the dynamic nature of unicellular eukaryotic genomes.

Key words: *Uroleptopsis citrina*, fragmented genome, ciliates, genome amplification.

Introduction

Ciliates are highly differentiated, morphologically complex, and widespread unicellular eukaryotes (Song et al. 2009; Chen et al. 2017; Liu et al. 2017). They are distinguished by nuclear dimorphism: each cell contains two types of nucleus, micronuclei (MIC) and macronuclei (MAC), which serve as germline and somatic genomes, respectively (Prescott 1994; Morgens et al. 2013; Wang YR et al. 2017). Due to this unique genome structure, ciliates have been models for researches in genetics and genomics, evolution and cell biology (Bracht et al. 2013; Chen et al. 2015; Gao et al. 2016; Xiong et al. 2016; Yi et al. 2016; Wang C et al. 2017). Studies in *Tetrahymena* and *Paramecium* have contributed to important biological discoveries, including catalytic RNA (Bass and Cech 1984; Greider and Blackburn 1985), telomeric repeats (Greider and Blackburn 1989), and histone modifications

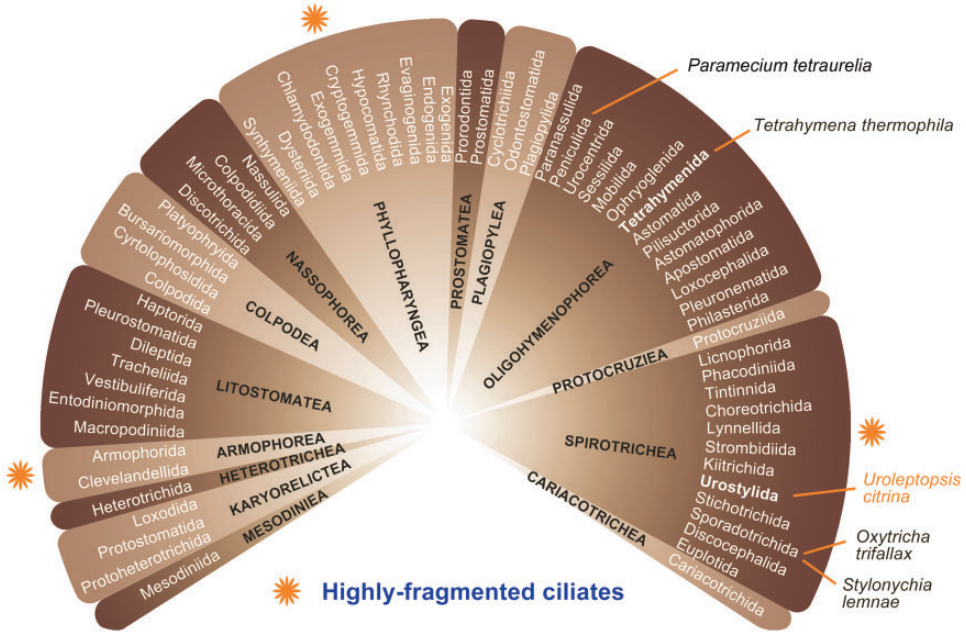


Fig. 1.—Outline of the classes in the phylum Ciliophora. Species involved in our research are highlighted.

(Gao et al. 2013; Wang, Chen, et al. 2017; Wang, Sheng, et al. 2017; Zhao et al. 2017), whereas studies in the hypotichs *Oxytricha* and *Stylonychia* have revealed scrambled genes, and the small RNAs and transposases that guide gene unscrambling during macronuclear differentiation (Landweber et al. 2000; Bracht et al. 2013; Chen et al. 2014). Telomerase was first identified by biochemical purification from the ciliate *Euplotes aediculatus* (Lingner and Cech 1996).

Despite the vast morphological diversity of ciliates, genomic research has been limited to only a few species (Swart et al. 2013; Aeschlimann et al. 2014; Slabodnick et al. 2017), but with modern sequencing methods it is now possible to greatly expand these studies. More than 4,500 species of ciliates dispersed in 11 classes have been morphologically described, and this number is growing (Foissner et al. 2008; Dong et al. 2016; Fan et al. 2016; Wang et al. 2016; Luo et al. 2017). Among these, only 18 species from 3 classes (10 Oligohymenophorea, 7 Spirotrichea, 1 Heterotrichea) have genomic information in GenBank. There are two main obstacles in ciliate genomic research: 1) cultivation is a precondition for most genomic research (Zheng et al. 2015; Bełżecki et al. 2016), whereas most ciliates, especially those living in extreme conditions, are difficult to cultivate in the lab; 2) ciliates are heterotrophs, feeding on other small organisms, such as bacteria and algae (Wolf 2014), which can lead to extensive contamination of genomic DNA preparations. In addition, many ciliates carry intracellular bacteria as parasites or symbionts, another source of contamination (Görtz 1996; Fokin 2004; Xiong et al. 2015).

Ciliates can be divided into two groups based on the structure of their MAC chromosomes (fig. 1). One group has relatively normal eukaryotic MAC chromosomes, each long chromosome carrying tens to hundreds of genes (although MAC chromosomes are acentric); examples include the Oligohymenophorea (*Tetrahymena thermophila* and *Paramecium tetraurelia*) and Heterotichea (*Stentor coeruleus*) (Eisen et al. 2006; Arnaiz et al. 2007; Slabodnick et al. 2017). In contrast, ciliate belonging to Phyllopharyngea, Spirotrichea, and Armophorea have extremely short MAC chromosomes, each chromosome carrying one or a very few genes with flanking telomeres (fig. 1) (Riley and Katz 2001; Swart et al. 2013; Aeschlimann et al. 2014; Gao et al. 2014, 2015; Huang et al. 2016). In addition, ciliates with long MAC chromosomes retain most of the MIC genome structure in the differentiated MAC, whereas ciliates with extensively fragmented chromosomes tend to lose a large portion of MIC sequence when forming their MAC (Arnaiz et al. 2012; Coyne et al. 2012; Bracht et al. 2013; Chen et al. 2014; Hamilton et al. 2016), although this may reveal limited sampling: *Paramecium caudatum* has a compact MAC genome (Mcgrath et al. 2014) but an extremely large MIC (Arnaiz et al. 2012), which has not yet been sequenced. In *Paramecium aurelia* species, MAC chromosomes are 100–1,000 kb long (perhaps each representing a MIC chromosome arm), and only 25% of MIC sequence complexity is lost during MAC differentiation (Arnaiz et al. 2012). In contrast, in *Oxytricha trifallax*, which has an extensively fragmented MAC genome, the mean chromosome length is 3.2 kb, containing only one or a few genes, and capped by telomeres with (C4A4)2C4 at the 5′-end and a longer G4(G4T4)2∼7 at 3′-end forming G-quadruplex structure, and 90% of the MIC complexity is lost (Bracht et al. 2013; Swart et al. 2013; Aeschlimann et al. 2014; Chen et al. 2014).

Given the small size and telomere caps of extensively fragmented chromosomes, we apply telomere-primer PCR to amplify the MAC genome from a limited number of cells: telomere-anchored oligonucleotides are used as PCR primers, and long-range high-fidelity DNA polymerases easily amplify most chromosomes, whereas reducing the influence of contaminating DNA from food and symbionts (Ricard et al. 2008). Using this method, we report the first assembly and analysis of the genome from the halobiotic hypotrichous ciliate *Uroleptopsis citrina* (Ciliophora, Spirotrichea, Hypotrichia). Although other hypotrichous ciliates with sequenced genomes inhabit freshwater and have two macronuclei (e.g., *Oxytricha trifallax* and *Stylonychia lemnae*), *Uroleptopsis citrina* has numerous dispersed macronuclei and is a marine species. This research generates novel insights into genomic structure of a unicellular eukaryote and facilitates genomic research of further ciliates.

Materials and Methods

Cell Isolation and DNA Extraction

Uroleptopsis citrina was collected from a harbor of Qingdao China (36°05′N, 120°33′E) in 2011. Cells were maintained in Petri dishes in filtered, autoclaved seawater with rice grains added to promote bacterial growth as food. *Uroleptopsis citrina* was identified by its morphological features and SSU-rRNA gene sequence, which differed by one nucleotide from the sequence of *U. citrine* in the NCBI database (Accession no.: GU437211) (Huang et al. 2010; Baek et al. 2011). Before DNA isolation for genomics studies, this strain had been cultured in lab for 2 years, maintained by subculture with only three to eight cells transferred; and only vegetative propagation was observed in these subcultures. DNA samples, of 10 cells (sample A) and 5,000 cells (sample B) were isolated, washed in autoclaved seawater, and extracted using the DNeasy Blood & Tissue Kit (QIAGEN, cat. no. 69506), following the manufacturer's protocol.

Chromosome Cloning, Genomic Amplification, High-Throughput Sequencing, and Assembly

DNA of sample B was treated with T4 DNA polymerase (TAKARA, cat. no. 2040 A) to generate blunt chromosome ends. These blunt products were cloned with the Mighty Cloning Reagent Set (TAKARA, cat. no. 6027) and sequenced on an ABI 3700 sequencer (GENEWIZ Incorporated Company, Beijing, China). The PCR primer for whole genome amplification was synthesized based on the chromosome telomeres found in the Sanger sequences: "5′-CCCCAAAA CCCCAAAACCCC-3′." DNA of sample A was telomere amplified using Q5 Hot Start High-Fidelity DNA Polymerase (NEB, cat. no. M0493L) with the elongation time of 5 minutes. The PCR product of sample A was purified with the EasyPure Quick Gel Extraction Kit (Transgen, cat. no. EG101-02); purified and amplified sample A DNA was used to generate a 500 bp average insert size PE library and sequenced on the Illumina Miseq platform (read length: 300 bp). About 1 Gb of Illumina PE data were obtained after quality control. Assembly of the genome was performed with SPAdes (default parameters and -k = 21, 33, 55, 77), ABySS (default parameters and -k = 63) and SOA Pdenovo2 (default parameters and -k = 63) (Simpson et al. 2009; Bankevich et al. 2012; Luo et al. 2012). QUAST was used to evaluate the resulting contigs (default parameters and -M = 0, means that all contigs are evaluated) (Gurevich et al. 2013). Bowtie2 was used to map the reads back to the genome to calculate the sequencing depth of each contig.

Reference Genome and Data Location of This Research

Five reference genomes were downloaded from NCBI GenBank: *Saccharomyces cerevisiae* (GenBank assembly accession number: GCA_000146045.2), *Oxytricha trifallax* (GenBank assembly accession number: GCA_000295675.1), *Stylonychia lemnae* (GenBank assembly accession number: GCA_000751175.1), *Paramecium tetraurelia* (GenBank assembly accession number: GCA_000165425.1), and *Tetrahymena thermophila* (GenBank assembly accession number: GCA_000189635.1). The assembled genomic data set of *U. citrina* can be accessed at NCBI (GenBank assembly accession number: GCA_001653735.1).

Gene Prediction

Reference-based gene prediction was performed by aligning all the contigs with the SWISS-PROT database using BLASTx (evalue = 1e-5, querygenecode = 6) (Catania and Lynch 2010). Because we lacked RNA-seq data, some BLASTx alignments (nucleotide against protein) included multiple HSPs (high similarity positions) separated by introns that cannot be divided by 3. A perl script (blast_xml_protein_coding_sequence_extract_v2.pl in the Supplementary Material online) was developed to analyze the BLASTx XML file: with this script, local alignments without start and stop codons were extended by concatenating the separate protein HSPs after deleting potential introns. The final gene prediction based on SWISS-PROT has been confirmed as in a consistent coding frame comparing with known proteins.

De novo gene prediction was performed using AUGUSTUS (version 2.5.5) (Stanke et al. 2004). An AUGUSTUS gene model for *U. citrina* was trained using the SWISS-PROT database, based on our gene prediction results above. In the parameter file, TAA and TAG codon were modified to encode for glutamine, instead of a translation stop. Because the minimum intron length in AUGUSTUS was not suitable for ciliates, the min_intron_len in the source code of types.cc file in AUGUSTUS was modified from 39 to 15 before recompiling the software. The prediction result of AUGUSTUS was filtered to discard genes without an annotated start or stop codon using a custom perl script (gff_analysis.pl in the

Supplementary Material online). The remaining protein coding genes from prediction of AUGUSTUS were blasted against the SWISS-PROT database to add functional information.

Predicted genes from AUGUSTUS and those from the BLASTx search against SWISS-PROT database are combined to form the final prediction files. Another perl script (combine_database_and_denovo_annotation.pl in the Supplementary Material online) was designed to compare the genes predicted by the SWISS-PROT database and those predicted by AUGUSTUS. For genes shared by both results, the detailed information from the SWISS-PROT database based prediction was maintained in the final prediction file.

Motif and Potential TATA-Box Searching

After confirming the telomere sequence (C4A4C4), the Perl regular expressions/(CCCCAAAA)+CCCC/and/GGGG(TTTTGGGG)+/were used to identify and remove telomere sequences from each 2-telomere contig. All 7,498 2-telomere contigs were compared against the SWISS-PROT database (-query_genecode = 6, -evalue = 1e-10): sequences with a match on the reverse strand were reverse complimented. About 5,380 of 7,498 contigs could be strand-confirmed.

MEME (Bailey et al. 2009) was used for motif searching in the 5′ and 3′ subtelomeric regions (-dna, -nmotifs 5, -maxw 25, -maxsize 500000). A perl script which identifies the longest consecutive pure AT region was used to identify potential TATA-boxs. This script recorded the position and length of all pure AT regions and output the longest one after scanning the whole target area. About 200 bp areas in the 5′ and 3′ subtelomeric regions of telomere-removed and strand-confirmed contigs were searched. The location of the longest consecutive pure AT region of each sequence was recorded and plotted.

Results and Discussion

Telomere Sequence Confirmation and Genomic Amplification

At present, all telomeric sequences reported in hypotrichous ciliates are (C4A4)2C4. We confirmed this in *Uroleptopsis citrina* by direct blunt-end cloning of chromosomes into plasmid vectors. About 44 cloned chromosomes were sequenced and each contained the telomeric sequence 5′-CCCCAAAACCCCAAAACCCC-3′, as reported in other hypotrichous ciliates (Swart et al. 2013; Aeschlimann et al. 2014). We used this sequence to synthesize telomere primers and amplify the MAC genome of *U. citrina*. PCR products range from 500 to 4,000 bp with a peak at ~1,000 bp (fig. 2). This amplified genome was purified (concentration: 218 ng/ul, OD260/280: 1.88), and subjected to high-throughput sequencing.

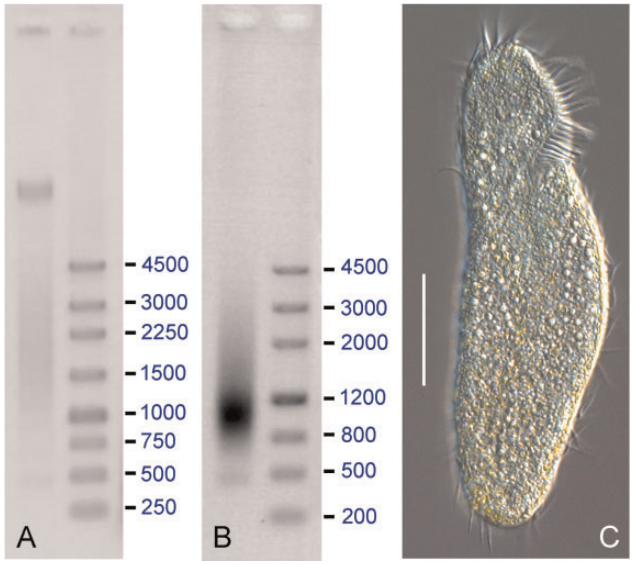


Fig. 2.—Total genomic DNA (*A*) and telomere-primer PCR amplified DNA (*B*) of *Uroleptopsis citrina*. (*A*) and (*B*) Marker unit: bp. (*C*) Living cell of *U. citrina* under DIC microscope. Scale bar: 40 μm.

Genome Assembly and Evaluation

Genome Assembly

To achieve the best assembly possible, we performed de novo assemblies with three different assemblers: the data set was assembled independently with SPAdes, ABySS, and SOAPdenovo2 (Simpson et al. 2009; Bankevich et al. 2012; Luo et al. 2012). The assemblies' quality was evaluated by N50, number of contigs and assembly size of contigs with zero, one, or two (0/1/2) telomeres—contigs with two telomeres were assumed to be complete MAC chromosomes (table 1). As table 1 shows, SPAdes has the best performance, yielding the longest N50, fewer contigs and more two-telomere caped contigs, and this is chosen for downstream analysis. This assembly was comprised of 33,912 contigs (31.15 Mb), with 5,490/20,924/7,498 0/1/2-telomere contigs, respectively.

Because of the PCR bias of the amplified chromosomes and relatively low sequencing depth (~20×), it is hard to obtain the whole genome. However, after removing bacterial contamination (discussed below), 31,754 contigs remained. The final data set for analysis is 29,661,436 bp (fig. 3); about one-third of the contigs (7,498) are completed MAC chromosomes. The sequencing coverage of each complete chromosome was calculated by mapping the reads to 2-telomere contigs, and is ~20× (fig. 4*A*). Chromosomes around ~400 and 3,900 bp have the highest coverage (50×). Chromosomes around 500–700 bp, 3,500–3,800 bp and >4,800 bp have comparatively low coverage (fig. 4*A*). This suggests that the coverage of each chromosome is a result of both PCR amplification and original chromosome copy number.

Table 1
Comparison of *Uroleptopsis citrina* Genome Assemblies

Assembler	SPAdes	AByss	SOAPdenovo
Assembly size (bp)	34,689,849	41,495,597	46,732,849
Number of contigs	33,912	205,876	182,603
Largest contig	6,746	3,142	2,678
N50 (bp)	1,007	224	293
Number of 2-telomere contigs	7,498	342	383
2-telomere contigs size/assembly size (%)	31.09	0.60	0.58
Number of 1-telomere contigs	20,924	29,867	22,414
1-telomere contigs size/assembly size (%)	42.98	18.38	17.08
Number of 0-telomere contigs	5,490	175,667	159,806
0-telomere contigs size/assembly size (%)	25.93	81.02	82.34

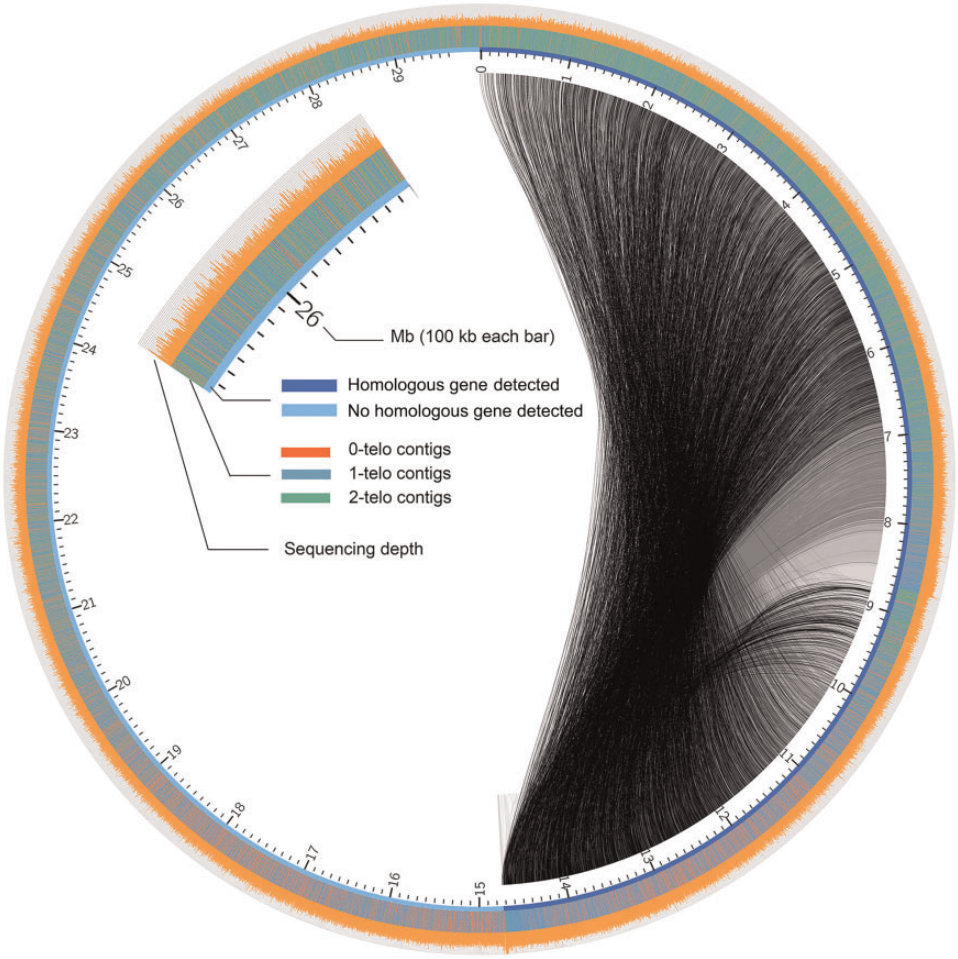


Fig. 3.—Whole assembly circle plot. About 31,754 contigs are separated into two groups based on the presence/absence of a homologous genes. Contigs are joined together and arranged by length (clockwise, from long to short). Homologous genes are linked by black lines within the circle. The outermost layer shows sequencing depth (calculated as TPM value). Middle layer shows distribution of 2/1/0-telomere contigs.

Self-to-self alignment with BLASTn searches (e-value: 1e-10) shows the existence of homologous genes that have assembled separately (fig. 3): 47.54% of contigs are found to have at least one homologous gene (average identity 92.03%) in the genome. Most consist of pairs, suggestive of allele pairs that have assembled separately, consistent with a diploid *U. citrina* germline genome. However, because of the MAC extreme fragmentation, it is impossible to identify synteny in this "one-gene one-chromosome" genome, and to distinguish paralogous genes from gene alleles.

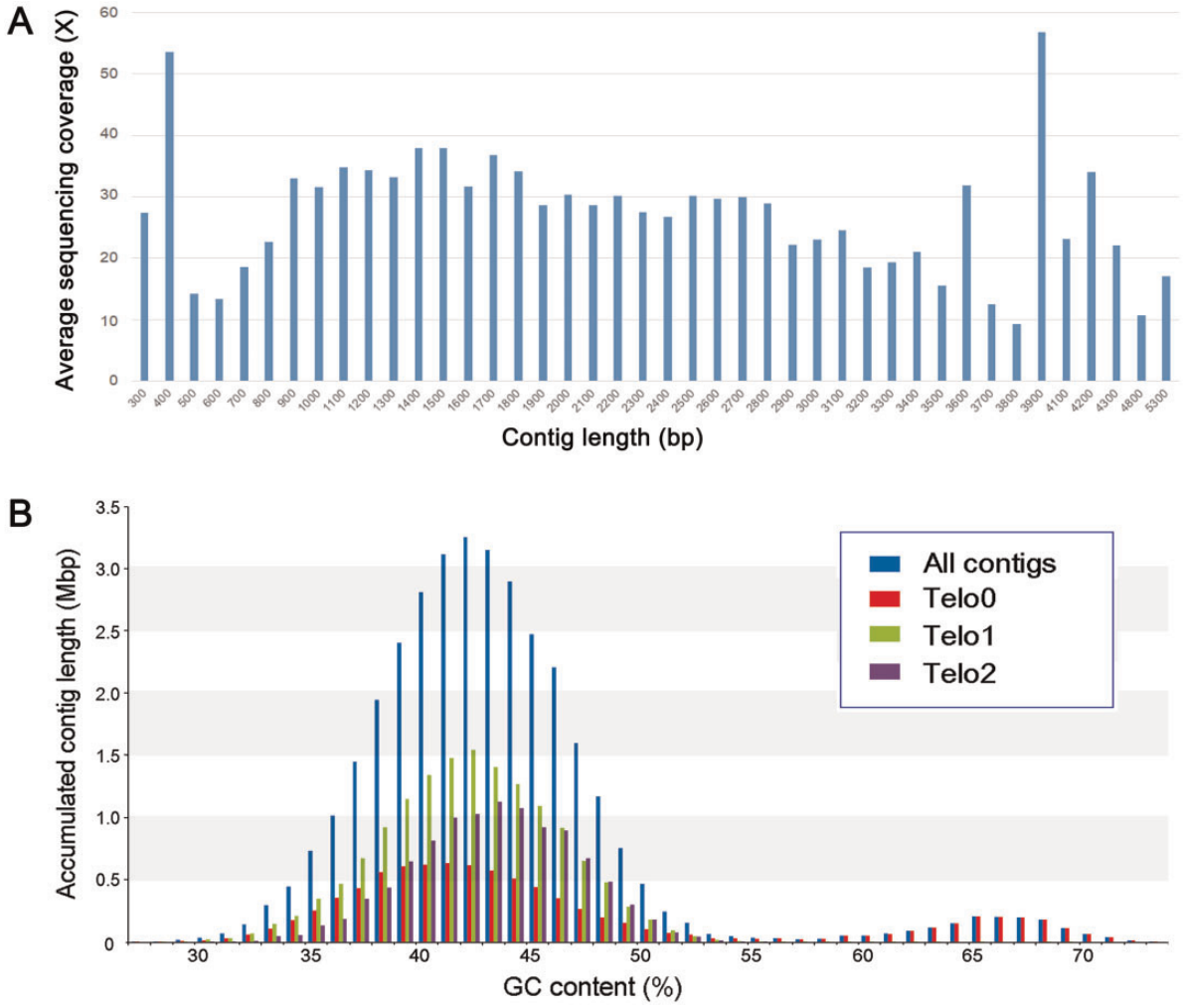


Fig. 4.—Accumulated contig length distribution based on GC content of different data sets from *Uroleptopsis citrina*. Telo2: Contigs with telomeres on both 5′- and 3′-ends. Telo1: Contigs with one telomere on the 5′- or 3′-end. Telo0: Contigs without telomeres. All contigs: Combination of all contigs.

Limited Bacteria Contamination

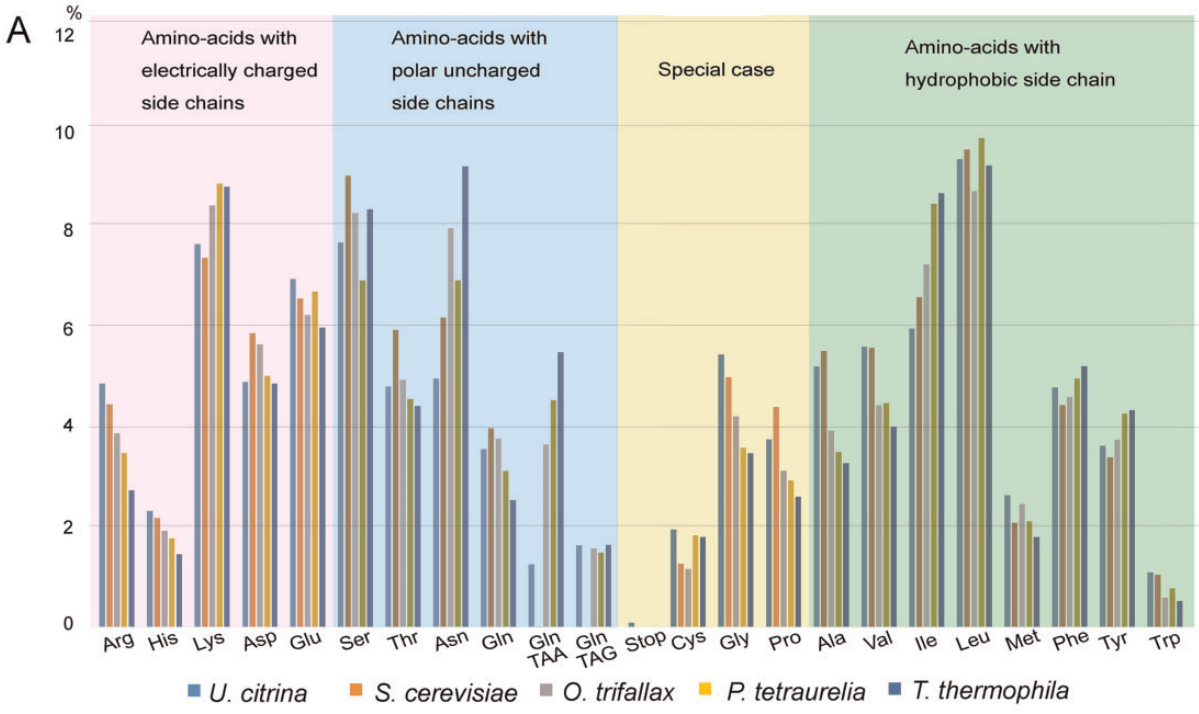
After assembly, contigs are clustered by GC content (fig. 4). For a specific species, GC content is rather stable and in some case may be a reliable standard for separating two species, especially for ciliates which have rather low GC content compared with bacteria. In the reported genome of hypotrichous ciliate *Oxytricha*, the GC content is ~40%. As shown in figure 4B, contigs are divided into two groups based on GC content: a major peak (95.46% of contigs) at 41% GC content and a minor set (4.54% of all contigs) at 65%. The minor peak indicates the presence of nonciliate contamination. This is confirmed by the presence of telomere capped contigs in only the low GC set (fig. 4). BLASTn searches (e-value: 1e-5) of the high GC set against the NCBI GenBank nucleotide database finds that 2,158 contigs with high GC, 91% (1,968 contigs) have strong matches to bacterial genomes. This result supports the contention that the low GC data set belongs to *U. citrina*, whereas the minor high GC set comes from bacterial contamination. It also indicates that the final assembly restricts the bacterial to a low and tractable level.

We believe that with deeper sequencing, more genomic data and more completed chromosomes will be obtained. Taken together, our results show that combining telomere-primer PCR amplification and high-throughput sequencing is a viable way to reduce DNA contaminations and obtain the genomic data efficiently for the ciliates with extensively fragmented genomes.

Protein Coding Gene and tRNA Predictions

The *U. citrina* data set, comprised of 31,754 contigs, is split into three subsets based on the number of telomeres/contigs (2, 1, or 0). SWISS-PORT BLAST search and AUGUSTUS de novo gene prediction are combined to predict genes in the three data sets. About 15,345 genes are predicted, and 9,529 of these with reliable matches to entries in the SWISS-PROT database were annotated. Most contigs (13,432) encode a

A



B

Genome GC content:	41%	38%	31%	28%	22%
	U. citrina	*S. cerevisiae*	*O. trifallax*	*P. tetraurelia*	*T. thermophila*
U. citrina		0.944	0.901	0.872	0.832
S. cerevisiae	0.944		0.97	0.934	0.904
O. trifallax	0.901	0.97		0.978	0.961
P. tetraurelia	0.872	0.934	0.978		0.986
T. thermophila	0.832	0.904	0.961	0.986	

Fig. 5.—(*A*) Amino acids usage of *Uroleptopsis citrina*, *Saccharomyces cerevisiae*, *Oxytricha trifallax*, *Paramecium tetraurelia*, and *Tetrahymena thermophile*. Codons TAA and TAG were calculated individually. (*B*) Spearman correlation coefficient of amino acid usage between each two species.

single gene, with an average length of 1,131 bp. Contigs coding for two (830 contigs), three (76 contigs), and four genes (3 contigs) are much longer, 1,972 bp, 2,535 bp, and 3,575 bp, respectively. No contig is predicted to contain five or more genes. This "one-gene one-chromosome" pattern is consistent with findings from other fragmented ciliates (Ricard et al. 2008; Swart et al. 2013; Aeschlimann et al. 2014).

We performed a tRNA scan to analyze the tRNA gene set (Lowe and Eddy 1997). About 90 contigs are detected that encode tRNAs, included all 20 standard amino acids. We also find two tRNA genes carrying the "TAA" codon. This provides proof of stop codon reassignment in *U. citrina* genome, which will be discussed in the next section.

Amino Acids Usage and Stop Codon Reassignment

Based on the 15,345 genes predicted using SWISS-PROT matches and AUGUSTUS, the amino acid usage of *U. citrina* is analyzed (fig. 5*A*). As comparisons, the CDS data sets of *Saccharomyces cerevisiae*, *Oxytricha trifallax*, *Paramecium tetraurelia*, and *Tetrahymena thermophila* are applied to the same analysis. *Saccharomyces cerevisiae* uses standard stop codons, including UGA, UAA, and UAG. The other three ciliate species have stop codon reassignment, with only UGA serving as a stop codon, whereas UAA and UAG are reassigned to glutamine (Gln) (Swart et al. 2016).

Uroleptopsis citrina appears to have the same stop-codon reassignment as *O. trifallax*, *P. tetraurelia*, and *T. thermphila* (Swart et al. 2016). As shown in figure 5*A*, the presence of in-frame UAA and UAG codons in *U. citrina* is much higher than in-frame UGA (UAA: 1.24%, UAG: 1.62%, UGA: 0.11%), similar to the other three ciliate species, and quite different from *S. cerevisiae*. The presence of in-frame UGA stop codons may be a consequence of sequencing error, potential mRNA frame-shifting, or the existence of selenocysteine.

Correlation between amino acid usage and GC content has been widely reported (Knight et al. 2001). Unlike other

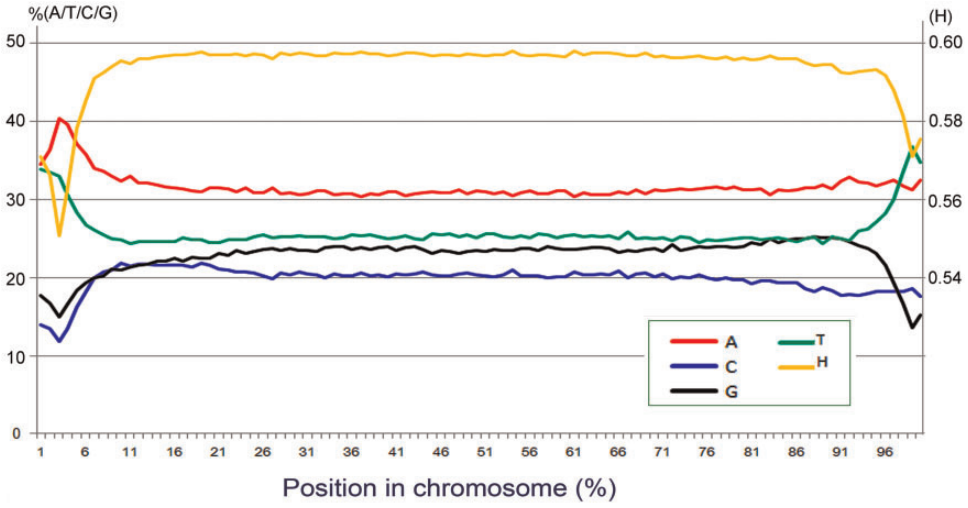


Fig. 6.—Sliding-window analysis for Shannon entropy (H) and nucleotide distribution of complete chromosomes.

ciliates in this analysis (*O. trifallax*: 31%, *P. tetraurelia*: 28%, *T. thermophila*: 22%), *U. citrina* has a relatively high GC content (41%), which is closer to that in *S. cerevisiae* (38%) (Wood et al. 2001; Eisen et al. 2006; Arnaiz et al. 2007; Swart et al. 2013). After removing the stop codon and reassigning UAA and UGA codons, the Spearman correlation coefficient of all the 20 amino acids was calculated and compared between each pair of species (fig. 5*B*). The results show that the amino acid usage is more divergent as the difference in GC content increasing. Amino acid usage of *U. citrina* is more similar to that in yeast than to other ciliates, which have much lower GC content.

Biased Nucleotide Distribution of Coding Chromosomes

To examine the nucleotide distribution across chromosomes, we apply a sliding-window analysis to telomere-removed complete chromosomes, which included 5,380 sense-stranded single-gene coding contigs (N50: 1,710 bp). The window size for each contig is equal to 1% of the contig length (17 bp in average) (fig. 6). There is an obvious AT bias for the first and last 9% of the full length of chromosomes, including most or all of the predicted 5′- and 3′-UTSs. Across the transcribed region, the AT/GC content is stable but reveals an obvious adenine richness for the coding strand (A: 31.6%; T: 25.8%; G: 22.8%; C: 19.7%). A gradually increasing guanine (G) and cytosine (C) bias is observed. This bias starts from the AT rich region of the 5′-end and reaches a peak (~8% G over C) at the 3′-end of the sense-strand. It seems to be strand sensitive: the GC bias in the sense-strand is asymmetric. This phenomenon has also been reported in the fragmented genomes of the ciliates *Nyctotherus ovalis* and *Oxytricha trifallax* (Ricard et al. 2008; Swart et al. 2013). The biological function of this symmetrical GC skew in the subtelomeric regions of the sense-strand is uncertain. A reasonable explanation is that the 3′-end of chromosomes is used as the leading strand in DNA replication, and is exposed as single stranded DNA for a greater time than the lagging strand. During the exposure as a single strand, cytosine is prone to deamination to uracil, which will result in error in the newly synthesized strand (Frank and Lobry 1999; Cavalcanti et al. 2004). This suggests that the 3′ subtelomeric region of the sense-strand may be the initial area where DNA replication starts. The replisome of this genome-fragmented ciliate may be strand sensitive.

We calculate the Shannon entropy (H) of the nucleotide composition for each window of the above analysis (Shannon equation $H = -\sum_{i=1}^{4} P_i \log P_i$) (fig. 6) (Shannon 1948). The entropy of the four nucleotides is high and stable in the transcribed region of chromosomes, and is obviously lower at both the 5′ and 3′ subtelomeric ends of chromosomes, indicating the potential existence of motifs. We perform MEME analysis of subtelomeric regions and found a motif TTGATTTC [A/T] TT in the 3′-end of 222 contigs among 5,380 contigs. No motif is detected in the 5′-end subtelomeric regions. This motif is within 11–39 nt (avg. 25 nt) of the telomere addition site of each chromosome (see next section), and may be related with DNA replication.

Base Composition of Subtelomeric Regions

In previous genomic characterization of other ciliates with extensively fragmented chromosomes, a unique subtelomeric nucleotide strand bias was detected (Prescott and Dizick 2000; Cavalcanti. et al. 2004; Ricard et al. 2008). Based on the above analysis, we further focus on the nucleotide composition for each site of the 200-nt subtelomeric region adjacent to 5′ and 3′ telomere of two-telomere chromosomes (fig. 7).

As shown in figure 7*A*, the first ~40 nt of both strands shows an A-over-T bias oscillation and consistent A-richness

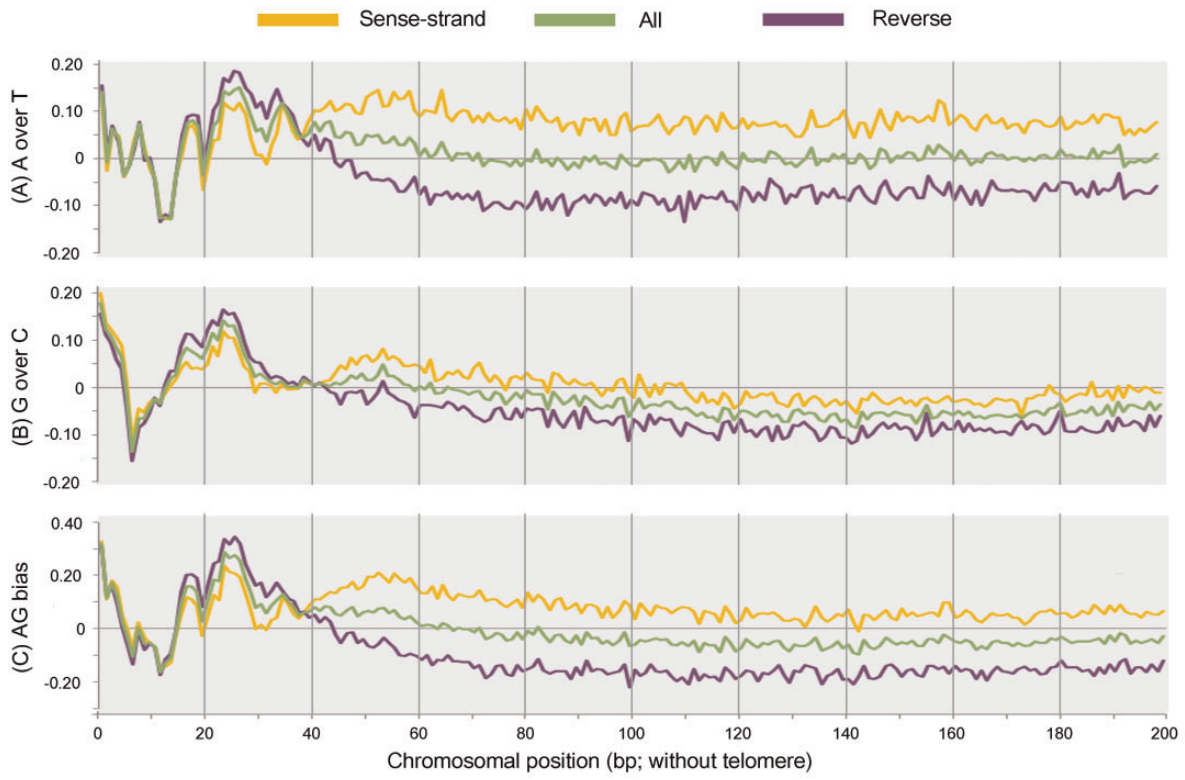


Fig. 7.—(*A*) A over T skew, (*B*) G over C skew, (*C*) AG over CT skew for the first 200 nt of sense- and antisense-strand of coding contigs in *Uroleptopsis citrina*. Sense-strand refers to the 5′ subtelomeric region of coding strands, reverse to the 5′ region of antisense-strands, which is the reverse complementary strand of 3′ subtelomeric region of coding strands.

between 20∼40 nt. In the 40–200 nt region, the sense-strand is A-rich while the antisense-strand is T-rich. A similar pattern of G-over-C (G-C) bias oscillation in the first 40 nt of both strands is observed (fig. 7*B*). For the remaining 160 nt, no obvious G-over-C bias is detected in the sense-strand, whereas the antisense-strand shows an obvious C-over-G bias. As a combined effect, there is consistent oscillation in the first 40 nt of purine skew $((A+G)-(C+T)/(A+G)+(C+T))$. In the 40–200 nt region, the sense-strand is slightly purine rich (∼8%), and the antisense-strand is pyridine rich (∼18%). This result indicates that the subtelomeric region starts to lose its special strand-nonsensitive oscillation 40 nt away from each telomere addition site and conforms to the patterns of normal coding sequence.

Overall, this result reveals that the sequence of extensively fragmented chromosomes can be divided into two main regions based on the nucleotide distribution: a 40-nt subtelomeic region with stable strand-nonsensitive oscillation at both the 5′- and 3′-ends and a relatively longer coding region which has a different nucleotide distribution in two strands. This feature agrees with the similar discovery in *Oxytricha trifallax* (Cavalcanti et al. 2004), and indicates that this may be a common feature among all ciliates with nanochromosomes. Currently we do not have a clear explanation for this strand-nonsensitive oscillation in the subtelomeric regions. Considering the special location of the subtelomeric regions, they may serve regulatory and structural functions for chromosome fragmentation, addition of telomeres, and initiation of transcription and replication (Helftenbein et al. 1989; Prescott 1994; Johnson et al. 1999; Johnson 2001).

Presence of Potential TATA-Box in Subtelomeric Regions

The above sections "Biased nucleotide distribution of coding chromosomes" and "Base composition of subtelomeric regions" both point to the special structure of the subtelomeric regions (e.g., Low Shannon entropy and different base composition compared with coding region), indicating that subtelomeric regions may have special functions. The TATA-box structure, which is the binding site of the TATA-binding protein (TBP) and responsible for the initiation of transcription, is as an obvious target for further analyses (Zhang et al. 2016).

Because transcription only starts at 5′ regions of coding strands, the 5,380 single gene coding chromosomes, which have confirmed strand information, are used in this analysis. By searching for pure-AT regions at least 9 nt in length, we find a strong pattern in the 200-nt subtelomeric regions of coding strands (fig. 8). The 5′ and 3′ subtelomeric regions of the coding strand divergence, with the 5′ subtelomeric regions obviously more pure-AT, which might serve as the potential TATA-boxes.

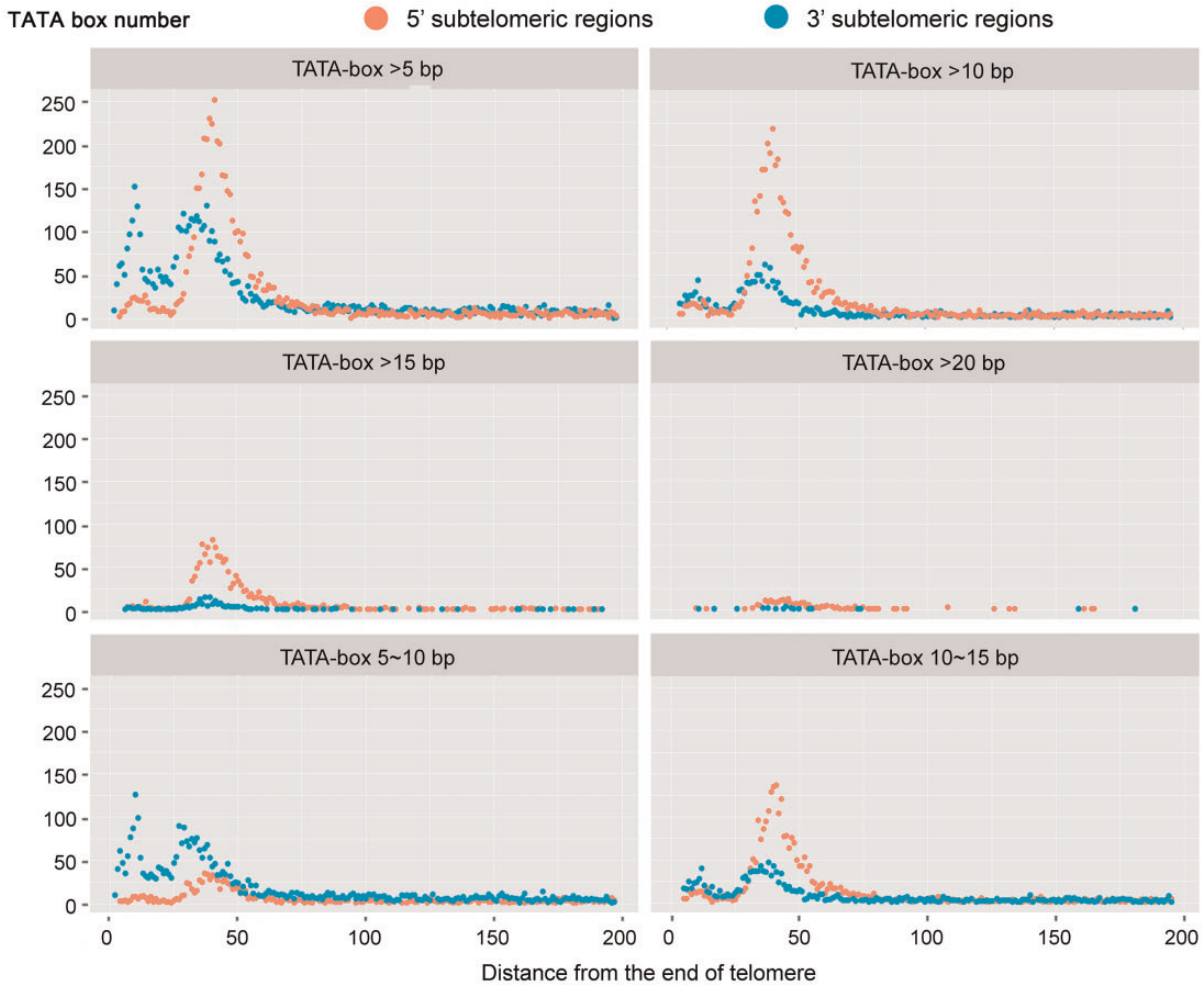


Fig. 8.—Distribution of potential TATA-box regions. The *X*-axis is the distance from the proximal end of telomere. The *Y*-axis is the number of putative TATA-box regions, which observed as pure-AT. Both 5′ and 3′ subtelomeric regions are involved and analyzed.

The distribution of pure-AT regions presents an obvious peak between 30 and 50 nt and a small peak in the first 25 nt. As the minimum length in identifying the pure-AT region increasing from > 5 nt to > 20 nt, the occurrence of potential TATA-boxs is maintained at the 30–50 nt in the 5′ subtelomeric region. The slight TATA-box enrichment in 3′ subtelomeric regions may be the resulted of false positive signals, incorrect strand confirmation, or failures in detecting two-gene chromosomes. By comparing the 5–10 nt and 10–15 nt potential TATA-boxs, the major TATA-box of 3′ subtelomeric region is 5–10 nt and relatively erratic in distribution, but that of 5′ subtelomeric region is 10–15 nt and stable in distribution. It indicates that false positive signals account for a large proportion of TATA-box enrichment in 3′ subtelomeric regions, but can be easily decreased by increasing the length restriction of pure-AT region. It also indicates that the minimum length of TATA-boxs is smaller than 10 nt. The subtelomeric regions are likely to possess TATA-boxs in a specific area and have tight relation of the transcription initiation of nanochromosomes.

TBP plays an important role in the preinitiation complex (PIC) recognizing the TATA-box and allowing transcription initiation (Geiger et al. 1996; Nikolov et al. 1996). Two TBP genes (TBP1a, 500 aa and TBP1b, 502 aa) are found in *Oxytricha trifallax* (Swart et al. 2013), along with *U. citrina* a stichotrich (Huang et al. 2010). We find orthologous genes of both TBP1a and TBP1b in two 2-telomere chromosomes of *U. citrina* (Contig 6476, 1,346 bp and Contig 9119, 1,141 bp). The high similarity of both alignments (TBP1a vs. Contig 6476 with the similarity of 79.4% and TBP1b vs. Contig 9119 with the similarity of 82.1%) are found at the last 200 aa of the C-terminal, which is known as the functional core and conserved region of TBP protein.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by Natural Science Foundation of China (project No. 31772428 and 31430077), Fundamental Research Funds for the Central Universities (project No. 201762017 and 201562029), and National Center for Genome Analysis Support, Indiana University (NSF grant DBI 1062432 and 1458641). We are grateful to Dr Xiaotian Luo, OUC, for her kind help with species identification. We appreciate Dr Wei-Jen Chang, Hamilton College, USA, and two reviewers for their helpful comments.

Literature Cited

Aeschlimann SH, et al. 2014. The draft assembly of the radically organized *Stylonychia lemnae* macronuclear genome. Genome Biol Evol. 6(7):1707–1723.

Arnaiz O, Cain S, Cohen J, Sperling L. 2007. ParameciumDB: a community resource that integrates the *Paramecium tetraurelia* genome sequence with genetic data. Nucleic Acids Res. 35(Database issue):D439–D444.

Arnaiz O, et al. 2012. The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. PloS Genet. 8(10):e1002984.

Baek Y-S, Jung J-H, Min G-S. 2011. Redescription of two marine ciliates (Ciliophora: Urostylida: Pseudokeronopsidae), *Pseudokeronopsis carnea* and *Uroleptopsis citrina*, from Korea. Korean J Syst Zool. 27:220–227.

Bailey TL, et al. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37(Web Server):W202–W208.

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455–477.

Bass BL, Cech TR. 1984. Specific interaction between the self-splicing RNA of *Tetrahymena* and its guanosine substrate: implications for biological catalysis by RNA. Nature 308(5962):820–826.

Bełżecki G, Miltko R, Michałowski T, McEwan NR. 2016. Methods for the cultivation of ciliated protozoa from the large intestine of horses. FEMS Microbiol Lett. 363(2):fnv233.

Bracht JR, et al. 2013. Genomes on the edge: programmed genome instability in ciliates. Cell 152(3):406–416.

Catania F, Lynch M. 2010. Evolutionary dynamics of a conserved sequence motif in the ribosomal genes of the ciliate *Paramecium*. BMC Evol Biol. 10:129.

Cavalcanti ARO, et al. 2004. Sequence features of *Oxytricha trifallax* (class Spirotrichea) macronuclear telomeric and subtelomeric sequences. Protist 155(3):311–322.

Cavalcanti ARO, Stover NA, Orecchia L, Doak TG, Landweber LF. 2004. Coding properties of *Oxytricha trifallax* (*Sterkiella histriomuscorum*) macronuclear chromosomes: analysis of a pilot genome project. Chromosoma 113(2):69–76.

Chen L, Zhao X, Shao C, Miao M, Clamp JC. 2017. Morphology and phylogeny of two new ciliates, *Sterkiella sinica* sp. nov. and *Rubrioxytricha tsinlingensis* sp. nov. (Protozoa, Ciliophora, Hypotrichia) from north-west China. Syst Biodivers. 15:131–142.

Chen X, et al. 2014. The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. Cell 158(5):1187–1198.

Chen X, et al. 2015. Phylogenomics of non-model ciliates based on transcriptomic analyses. Protein Cell 6(5):373–385.

Coyne RS, Stover N, Miao W. 2012. Whole genome studies of *Tetrahymena*. Methods Cell Biol. 109:53–81.

Dong J, Lu X, Shao C, Huang J, Al-Rasheid KA. 2016. Morphology, morphogenesis and molecular phylogeny of a novel saline soil ciliate, *Lamtostyla salina* n. sp. (Ciliophora, Hypotricha). Eur J Protistol. 56:219–231.

Eisen JA, et al. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol. 4(9):e286.

Fan Y, Lu X, Huang J, Hu X, Warren A. 2016. Redescription of two little-known urostyloid ciliates, *Anteholosticha randani* (Groliere, 1975) Berger, 2003 and *A. antecirrata* Berger, 2006 (Ciliophora, Urostylida). Eur J Protistol. 53:96–108.

Foissner W, Chao A, Katz LA. 2008. Diversity and geographic distribution of ciliates (Protista: Ciliophora). Biodivers Conserv. 17(2):345–363.

Fokin SI. 2004. Bacterial endocytobionts of ciliophora and their interactions with the host cell. Int Rev Cytol. 236:181–249.

Frank A, Lobry J. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238(1):65–77.

Gao F, et al. 2016. The all-data-based evolutionary hypothesis of ciliated protists with a revised classification of the phylum Ciliophora (Eukaryota, Alveolata). Sci Rep. 6:24874.

Gao F, Roy SW, Katz LA. 2015. Analyses of alternatively processed genes in ciliates provide insights into the origins of scrambled genomes and may provide a mechanism for speciation. mBio 6(1):e01998-14–e01914.

Gao F, Song W, Katz LA. 2014. Genome structure drives patterns of gene family evolution in ciliates, a case study using *Chilodonella uncinata* (Protista, Ciliophora, Phyllopharyngea). Evolution 68(8):2287–2295.

Gao S, et al. 2013. Impaired replication elongation in *Tetrahymena* mutants deficient in histone H3 Lys 27 monomethylation. Genes Dev. 27(15):1662–1679.

Geiger JH, Hahn S, Lee S, Sigler PB. 1996. Crystal structure of the yeast TFIIA/TBP/DNA complex. Science 272(5263):830–836.

Görtz H. 1996. Symbiosis in Ciliates. In: Hausmann PCB K., editor. Ciliates: cells as organisms. Gustav Fischer, Stuttgart: Spektrum Akademischer Verlag. p. 441–462.

Greider CW, Blackburn EH. 1985. Identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts. Cell 43(2 Pt 1):405–413.

Greider CW, Blackburn EH. 1989. A telomeric sequence in the RNA of *Tetrahymena* telomerase required for telomere repeat synthesis. Nature 337(6205):331–337.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29(8):1072–1075.

Hamilton EP, et al. 2016. Structure of the germline genome of *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome. Elife 5:e19090.

Helftenbein E, Conzelmann KK, Becker KF, Fritzenschaf H. 1989. Regulatory structures of gene expression, DNA-replication and DNA-rearrangement in macronuclear genes of *Stylonychia lemnae*, a hypotrichous ciliate. Eur J Protistol. 25(2):158–167.

Huang JA, Yi Z, Al-Farraj SA, Song W. 2010. Phylogenetic positions and taxonomic assignments of the systematically controversial genera, *Spirotrachelostyla*, *Uroleptopsis* and *Tunicothrix* (Protozoa, Ciliophora, Stichotrichia) based on small subunit rRNA gene sequences. Syst Biodivers. 8:409–416.

Huang JB, Luo X, Bourland WA, Gao F, Gao S. 2016. Multigene-based phylogeny of the ciliate families Amphisiellidae and Trachelostylidae (Protozoa: Ciliophora: Hypotrichia). Mol Phylogenet Evol. 101:101–110.

Johnson BF, Sinclair DA, Guarente L. 1999. Molecular biology of aging. Cell 96(2):291–302.

Johnson BF. 2001. The *Saccharomyces cerevisiae* WRN homolog Sgs1p participates in telomere maintenance in cells lacking telomerase. EMBO J. 20(4):905–913.
Knight RD, Freeland SJ, Landweber LF. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. 2:research0010.0011.
Landweber LF, Kuo T-C, Curtis EA. 2000. Evolution and assembly of an extremely scrambled gene. Proc Natl Acad Sci USA. 97(7):3298–3303.
Lingner J, Cech TR. 1996. Purification of telomerase from *Euplotes aediculatus*: requirement of a primer 3'overhang. Proc Natl Acad Sci USA. 93(20):10712–10717.
Liu W, et al. 2017. Diversity of free-living marine ciliates (Alveolata, Ciliophora): faunal studies in coastal waters of China during the years 2011–2016. Eur J Protistol. 61(Pt B):424–438.
Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25(5):955–964.
Luo R, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:1–6.
Luo X, et al. 2017. Morphologic and phylogenetic studies of two hypotrichous ciliates, with notes on morphogenesis in *Gastrostyla steinii* Engelmann, 1862 (Ciliophora, Hypotrichia). Eur J Protistol. 60:119–133.
Mcgrath CL, Gout JF, Doak TG, Yanagi A, Lynch M. 2014. Insight into three whole-genome duplications gleaned from the *Paramecium caudatum* genome sequence. Genetics 197(4):1417–1428.
Morgens DW, Lindbergh KM, Adachi M, Radunskaya A, Cavalcanti AR. 2013. A model for the evolution of extremely fragmented macronuclei in ciliates. PLoS One 8(5):e64997.
Nikolov DB, et al. 1996. Crystal structure of a human TATA box-binding protein/TATA element complex. Proc Natl Acad Sci USA. 93(10):4862–4867.
Prescott DM. 1994. The DNA of ciliated protozoa. Microbiol Rev. 58(2):233–267.
Prescott DM, Dizick SJ. 2000. A unique pattern of intrastrand anomalies in base composition of the DNA in hypotrichs. Nucleic Acids Res. 28(23):4679–4688.
Ricard G, et al. 2008. Macronuclear genome structure of the ciliate *Nyctotherus ovalis*: single-gene chromosomes and tiny introns. BMC Genomics 9:587–601.
Riley JL, Katz LA. 2001. Widespread distribution of extensive chromosomal fragmentation in ciliates. Mol Biol Evol. 18(7):1372–1377.
Shannon CE. 1948. A mathematical theory of communication. Bell Syst Tech J. 27:379–423, 623–656.
Simpson JT, et al. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19(6):1117–1123.
Slabodnick MM, et al. 2017. The macronuclear genome of *Stentor coeruleus* reveals tiny introns in a giant cell. Curr Biol. 27(4):569–575.
Song W, Warren A, Hu X. 2009. Free-living ciliates in the Bohai and Yellow Seas. Beijing: Science Press.
Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 32(Web Server issue):W309–W312.
Swart EC, et al. 2013. The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. PLoS Biol. 11(1):e1001473.
Swart EC, Serra V, Petroni G, Nowacki M. 2016. Genetic codes with no dedicated stop codon: context-dependent translation termination. Cell 166(3):691–702.
Wang C, et al. 2017. Disentangling sources of variation in SSU rDNA sequences from single cell analyses of ciliates: impact of copy number variation and experimental error. Proc R Soc B: Biol Sci 284:20170425.
Wang J, Lyu Z, Warren A, Wang F, Shao C. 2016. Morphology, ontogeny and molecular phylogeny of a novel saline soil ciliate, *Urosomoida paragiliformis* n. sp. (Ciliophora, Hypotrichia). Eur J Protistol. 56:79–89.
Wang YR, et al. 2017. A comparative study of genome organization and epigenetic mechanisms in model ciliates, with an emphasis on *Tetrahymena*, *Paramecium* and *Oxytricha*. Eur J Protistol. 61:376–387.
Wang YY, Chen X, Sheng Y, Liu Y, Gao S. 2017. N6-adenine DNA methylation is associated with the linker DNA of H2A. Z-containing well-positioned nucleosomes in Pol II-transcribed genes in *Tetrahymena*. Nucleic Acids Res. 45:11594–11606.
Wang YY, Sheng Y, et al. 2017. N 6-methyladenine DNA modification in the unicellular eukaryotic organism *Tetrahymena thermophila*. Eur J Protistol. 58:94–102.
Wolf M. 2014. Cilia and Flagella. Ciliates and Flagellates. Ultrastructure and Cell Biology, Function and Systematics, Symbiosis and Biodiversity. Stuttgurt: Schweizerbart Science Publisher.
Wood V, Rutherford K, Ivens A, Rajandream M-A, Barrell B. 2001. A re-annotation of the *Saccharomyces cerevisiae* genome. Comp Funct Genomics 2(3):143–154.
Xiong J, et al. 2015. Genome of the facultative scuticociliatosis pathogen *Pseudocohnilembus persalinus* provides insight into its virulence through horizontal gene transfer. Sci Rep. 5:15470.
Xiong J, et al. 2016. Dissecting relative contributions of cis-and trans-determinants to nucleosome distribution by comparing *Tetrahymena* macronuclear and micronuclear chromatin. Nucleic Acids Res. 44(21):10091–10105.
Yi Z, Huang L, Yang R, Lin X, Song W. 2016. Actin evolution in ciliates (Protist, Alveolata) is characterized by high diversity and three duplication events. Mol Phylogenet Evol. 96:45–54.
Zhang Z, et al. 2016. Rapid dynamics of general transcription factor TFIIB binding during preinitiation complex assembly revealed by single-molecule analysis. Genes Dev. 30(18):2106–2118.
Zhao X, Wang Y, Wang Y, Liu Y, Gao S. 2017. Histone methyltransferase TXR1 is required for both H3 and H3. 3 lysine 27 methylation in the well-known ciliated protist *Tetrahymena thermophila*. Sci China Life Sci. 60(3):264–270.
Zheng W, Gao F, Warren A. 2015. High-density cultivation of the marine ciliate *Uronema marinum* (Ciliophora, Oligohymenophorea) in axenic medium. Acta Protozool. 54:325–330.

Associate editor: Rebecca Zufall