Architecture and evolution of subtelomeres in the unicellular green alga *Chlamydomonas reinhardtii*

Frédéric Chaux-Jukic¹,†, Samuel O’Donnell ⊕¹,†, Rory J. Craig², Stephan Eberhard³, Olivier Vallon³,*† and Zhou Xu ⊕¹,*†

¹Sorbonne Université, CNRS, UMR7238, Institut de Biologie Paris-Seine, Laboratory of Computational and Quantitative Biology, 75005 Paris, France, ²Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, EH9 3FL, Edinburgh, UK and ³Sorbonne Université, CNRS, UMR7141, Institut de Biologie Physico-Chimique, Laboratory of Chloroplast Biology and Light-Sensing in Microalgae, 75005 Paris, France

ABSTRACT

In most eukaryotes, subtelomeres are dynamic genomic regions populated by multi-copy sequences of different origins, which can promote segmental duplications and chromosomal rearrangements. However, their repetitive nature has complicated the efforts to sequence them, analyse their structure and infer how they evolved. Here, we use recent genome assemblies of *Chlamydomonas reinhardtii* based on long-read sequencing to comprehensively describe the subtelomere architecture of the 17 chromosomes of this model unicellular green alga. We identify three main repeated elements present at subtelomeres, which we call *Sultan*, *Subtile* and *Suber*, alongside three chromosome extremities with ribosomal DNA as the only identified component of their subtelomeres. The most common architecture, present in 27 out of 34 subtelomeres, is a heterochromatic array of *Sultan* elements adjacent to the telomere, followed by a transcribed *Spacer* sequence, a G-rich microsatellite and transposable elements. Sequence similarity analyses suggest that *Sultan* elements underwent segmental duplications within each subtelomere and rearranged between subtelomeres at a much lower frequency. Analysis of other green algae reveals species-specific repeated elements that are shared across subtelomeres, with an overall organization similar to *C. reinhardtii*. This work uncovers the complexity and evolution of subtelomere architecture in green algae.

INTRODUCTION

The extremities of linear chromosomes in eukaryotes are essential to maintain stable genomes (1). At their very end, repeated sequences called telomeres recruit specific factors that collectively prevent detection of the extremities as double-strand breaks and avoid deleterious effects caused by repair attempts by the cell (2,3). Telomeres also counteract the end-replication problem, which would otherwise lead to replicative senescence and cell death. In most organisms, this is achieved by recruiting the reverse-transcriptase telomerase, which processively adds *de novo* telomere sequences. Instead of telomerase, some species of the Diptera order use other maintenance mechanisms, such as retrotransposons in *Drosophila melanogaster* or recombination-dependent mechanisms in *Chironomus* or *Anopheles* (4–6). Homology-directed recombination can also be used to maintain telomeres in a number of cancer cells and in models where telomerase is experimentally inactivated, as for example in *Arabidopsis thaliana* or *Saccharomyces cerevisiae* (7). Next to the telomere, the subtelomere is usually a gene-poor region comprising repeated elements, such as transposable elements (TEs), satellite sequences, ribosomal DNA (rDNA), or paralogous genes, which are often shared between different subtelomeres (8–15). The described gene families are involved in diverse life cycle and adaptive processes such as metabolism in *S. cerevisiae* (11,13), surface antigen repertoires in *Plasmodium falciparum* (16), resistance to pathogens in common bean (14), or olfactory receptors and cytoskeleton (proteins of the WASP family) in human (17,18).

In a number of organisms, including *D. melanogaster*, *S. cerevisiae*, *Schizosaccharomyces pombe* and mammals, subtelomeres are generally considered heterochromatic (19–24). This is less clear in plant subtelomeres where signatures of both euchromatin and heterochromatin have been reported (25,26). An emerging view proposes that subtelom-
subtelomeres might be associated with a specific type of chromatin, alongside canonical heterochromatin (27), which can lead to reversible transcriptional silencing, a property called telomere position effect. Additionally, specific non-coding transcripts have been detected in these regions, including the telomeric repeat-containing RNA (TERRA), which plays multiple roles in telomere biology (28,29). Importantly, subtelomeres can regulate telomere length, telomere-associated chromatin and replicative senescence, and can help maintain telomere and genome integrity (11,23,28,30–34).

Subtelomeres are rapidly evolving regions and can vary greatly in structure and composition between closely related species and even individuals of the same species, as shown recently in humans (16,35–42). Several mechanisms have been shown or proposed to explain subtelomeric variations. The repetitive nature of the region promotes homologous recombination (HR), unequal sister chromatid exchange (SCE), break-induced replication (BIR) and replication slippage (8,14,35,40,43–48). Transposition also contributes to subtelomere variations (10,14,39,46). These mechanisms, along with others such as non-homologous end-joining (NHEJ)-mediated translocations and fusions, have been described in a variety of species and can lead to segmental duplications and amplification of repeated elements (14,44,46,47). Consistently, mutation rates and chromosomal rearrangements are elevated at chromosome ends, even more so in the absence of telomerase, as reported in yeast, Drosophila, mammals and plants (35,38,49–52).

While telomeres and subtelomeres are of critical importance for both genome stability and evolution, they are often misassembled or altogether absent in reference genomes of most species because of their intrinsically complex and repetitive nature. For example, although its assembly has recently been improved, the human reference genome still lacks a comprehensive and accurate representation of its repetitive nature. For example, although its assembly has recently been improved, the human reference genome still lacks a comprehensive and accurate representation of its subtelomeres (41,53–55). With the advent of long read sequencing technologies (16,40,42,56), we can now gain access to better assemblies and descriptions of subtelomeres, enabling the mechanisms underlying their structural variations and evolution to be inferred for a diverse range of organisms.

We recently characterized telomere structure and telomerase mutants in the unicellular green alga Chlamydomonas reinhardtii (57), a major model for photosynthesis and cilia research. The discovery of blunt ends at a subset of telomeres and a wide range of telomere length distributions in different reference strains prompted us to further explore how chromosome ends have evolved and are structured. Here, we provide a comprehensive description of the architecture of the subtelomeres in C. reinhardtii and a comparative analysis with other green algae. An early study evidenced a high level of similarity in the sequences adjacent to a few cloned C. reinhardtii telomeres (58). Subtelomere architecture has also been partially outlined in a limited number of plant species, including Arabidopsis thaliana, Silene latifolia and Phaseolus vulgaris (12,14,46,47,59,60), and the green alga Coccomyxa subellipsoidea (61). To probe the structure of these repetitive regions, we recently generated a contiguous de novo assembly from published Oxford Nanopore Technologies long reads (62,63), which we analyze alongside newly released PacBio-generated assemblies (https://raba.ibpc.fr/home/ovallon@ibpc.fr/Briefcase/Chlamy_genomes). We show that most C. reinhardtii subtelomeres are composed, reading from the telomere toward the centromere, of an array of repeated elements that we call Sultan (for Subtelomeric Long TANdem repeats), a Spacer sequence, a G-rich microsatellite sequence of variable length and various types of TEs. Sequence homology analysis of the Sultan elements suggests that they mostly propagated within a subtelomere through segmental duplications and less frequently between different subtelomeres. Subtelomeres in other green algae also contain specific repeated sequences, unrelated to the Sultan element, suggesting a common organization that has possibly evolved independently for subtelomere functions.

MATERIALS AND METHODS

Genome assemblies

All genome assemblies used in this study can be downloaded from https://raba.ibpc.fr/home/ovallon@ibpc.fr/Briefcase/Chlamy_genomes. The assemblies for Chlamydomonas incerta, Chlamydomonas scloesseri and Edaphochlamys debaryana are described in (64), that from CC-2931 was obtained by assembly of PacBio reads (Craig et al., bioRxiv, doi: https://doi.org/10.1101/2021.04.23.441226). The CC-4532 and CC-503 (v6) assemblies are available on the Phytozome website (https://phytozome-next.jgi.doe.gov/). For strain CC-1690 (“21gr”), recently released Nanopore raw sequencing data (62) were base-called and de novo assembled into chromosomes as described in (63). GenBank accession:JABWPN000000000). For the present work, we used a version prior to the Illumina polishing step and used linkage data (65) to further scaffold the last unplaced contig (unplaced_1) to the end of chromosome 15, forming its right arm. Compared to our released genome (63), we corrected a mistake in the assembly of subtelomere 9_R (a replacement contig is appended to the genome), which was distorted at the telomere-proximal side of the Sultan array by reads from 15_R. To do this, reads were first mapped against the whole genome using minimap2 (66), then extracted if they mapped to the 9_R and had a mapping quality of 60. This subsample of reads was then used for re-assembly with Canu (V2) using default settings. Additionally, the 1_R end, which did not contain a telomeric sequence nor Sultan repeats at its apparent terminus, was analysed by read mapping and we were able to recover a few reads extending beyond the assembly and containing both telomere sequences and 14 Sultan repeats.

A curated library of Volvocales TEs (64) was used to identify mobile and repetitive genetic elements, using RepeatMasker (http://repeatmasker.org/).

Search for tandem repeats

We use the term ‘repeat’ to refer to the finite pattern found in a repetitive sequence, ‘copy’ to a specific instance of the repeat and ‘array’ to a series of copies. Copies that are found in an array in the same orientation and are not separated from each other by unrelated DNA sequences are called ‘in tandem’.
We extracted and analysed the first 30 kb of the chromosome ends (300 kb for class C subtelomeres). Sequences from the right extremities were reverse complemented, so that both left and right chromosome ends started with telomeric repeats in the form of 5′-(CCCTAA AA)$_n$-3′ tracts. Sequences were aligned using Tandem Repeat Finder (v4.04, parameters ‘3 5 5 80 20 100 2000’) and X-STREAM (variable sequence tandem repeats extraction and architecture modelling, https://amnewmanlab.stanford.edu/xstream/) (67,68). X-STREAM was run with default parameters, except that ‘TR significance’ was disabled and ‘Minimal word match’ and ‘Minimum Consensus match’ were adjusted in the ranges 0.1–0.7 and 0.7–0.95, respectively, to allow detection of incomplete repeats at extremities of tandem arrays. Repeat consensus sequences were phased and used as blast queries to retrieve the coordinates of the repeat copies. Sequences were extracted using EMBOSS (v6.4.0) seqret (Supplemental File F2). Multiple sequence alignments were generated with MAFFT (v7.130) with iterative refinement method G-INS-i. Pairwise distances were calculated using EMBOSS distmat with the Jukes-Cantor substitution model.

Phylogenetic analyses and trees were generated using PhyML with the generalized time-reversal (GTR) model for nucleotide evolution and drawn using Interactive Tree Of Life (https://itol.embl.de/) (69). JAL-view (70) and Bioedit (71) were used for data visualization and calculation of consensus and logo sequences. Consensus sequences were computed with Advanced Consensus Maker (https://www.hiv.lanl.gov/cgi-bin/CONSENSUS_TOOL/consensus.cgi).

Transcriptomics

The transcript dataset from (72) (accession number: GSE112394; strain CC-5390) was searched using each Spacer sequence as BLAST queries on NCBI server. Duplicate hits were discarded and coverage was computed using bedtools coverage (v2.29.2) (73).

Iso-Seq data (accession number: PRJNA670202; multiple laboratory strains) and ChIP-seq data (accession number: PRJNA681680; strain CC-5390) were used to assess transcription and H3K4me3 marks. Circular consensus sequence Iso-Seq reads were mapped against the CC-1690 assembly using minimap2 (parameters: -ax splice:ht -m 200000) (74). ChIP-seq reads were mapped using bwa-mem (74), duplicates were removed using the Picard tool MarkDuplicates (http://broadinstitute.github.io/picard/), and peaks were called with MACS v2 (parameters: callpeak -g 9.1e7 -B —fix-bimodal —extsize 150) (75).

A ChIP-seq score was calculated for H3K4me3 marks relative to the input control sample based on the approach in (76). Gene promoters were defined as 500 bp downstream of the transcription start site.

Genomic reads mapping

Illumina data for each strain (Supplemental Table ST3) were mapped against the whole genome of CC-1690 using bwa-mem (74). The bam file was used to calculate the average whole genome coverage and extract all reads mapping to Sultan arrays. This read subset was then aligned against all Sultan consensus sequences from the same strain. The fold increase in median coverage within each consensus, compared to the whole genome, was used as a measure of the number of reads within each array from which the consensus was derived (Supplemental File F3).

5-Methylcytosine detection

Raw 40× Nanopore reads previously used for genome assembly were downsampled to 10× using filtlong (https://github.com/rrwick/Filtlong); fast5 files were subset equally. The subsets of fastq and fast5 were then used in the default deepsignal pathway for 5mC-CpG prediction using model ‘model.CpG.R9.4_1D.human_hx1.bn17.sn360.v0.1.7+’ (77) (Supplemental File F4).

Genomic DNA extraction and Southern blot

Cells were grown to early stationary phase under continuous illumination (8 µE m$^{-2}$ s$^{-1}$) in agitated 200 ml liquid cultures in Tris-acetate-phosphate (TAP) medium. Genomic DNA extraction and Southern blot were performed as in (57), except that instead of radioactive probes, oligonucleotide probes biotinylated at both ends were used (Sultan_oZX076: 5′-GGCTGCGTGGCTG GACTGCTGCACT-3′, Suber_oZX179: 5′-GCCACAG GGGAAAGTCGAGAATCTG-3′, rDNA_oZX178: 5′-ACCGCATGCACCTCAGACGTC-3′ and telomere_oT959: 5′-CTAAACCCCTAAAACCCCTAA ACCTTTAAC-3′; Eurofins Genomics) and detected by chemiluminescence. After hybridization of the probe, the membrane was washed 3× 5 min in wash buffer (58 mM Na$_2$HPO$_4$, 17 mM NaH$_2$PO$_4$, 68 mM NaCl, 0.1% SDS). The membrane was next processed for detection with three successive incubations (5, 5 and 30 min) in blocking buffer (Thermo Scientific, Nucleic Acid Detection Blocking Buffer) before a 30 min incubation with alkaline phosphatase-conjugated streptavidin (Invitrogen) diluted in blocking buffer (0.4 µg/ml). The membrane was then washed again 3× 5 min in wash buffer, incubated 2× 2 min in assay buffer (0.1 M Tris, 0.1 M NaCl pH9.5) and 5 min in CDP-Star substrate (Applied Biosystems) before imaging with a G:BOX Gel Doc system (Syngene).

RESULTS

*Chlamydomonas* subtelomeres comprise arrays of specific tandemly repeated sequences

Because the publicly available *C. reinhardtii* reference genome version 5 (v5; https://phytozome-next.jgi.doe.gov/) at the time of this work was incompletely assembled near the chromosome extremities, we took advantage of our recent release of a *de novo* genome assembly (63) based on long-read sequencing data of strain CC-1690 (62), a commonly used laboratory strain also known as 21gr. Briefly, the raw Oxford Nanopore Technologies (hereafter referred to as ‘Nanopore’) electrical signal was base-called and subsets of the longest reads ($N_{50} \simeq 55$ kb) were assembled independently using various protocols, after which assemblies were combined to create a 21-contig genome assembly,
readily scaffolded onto the 17 chromosomes. The chromosome arms, and therefore their termini, are labelled ‘left’ and ‘right’ (or _L and _R) based on the orientation used in the reference genome and the sequences and features are generally presented reading from the telomere towards the centromere.

Arrays of the 8-bp telomeric repeat motif (5’-CCCTAAAA-3’/5’-TTTAGGG-3’) previously described in _C. reinhardtii_ (57,58,78) were found at the extremity of 33 out of the 34 chromosome ends (Figure 1, black segments). In the genome assembly, telomeric repeats had a median length of 311 ± 125 bp (median ± SD), at the shorter end of the 300–700 bp range observed previously by terminal restriction fragment analysis (57).

Alignments systematically revealed extensive homology between subtelomeres, usually covering several kilobases in the form of long repeated arrays. To identify the repeated elements in subtelomeres, we scanned the last 30 kb of each chromosome end using XSTREAM (67) and Tandem Repeats Finder (TRF) (68). Figure 1 shows a map of the subtelomeres of strain CC-1690, depicting their repetitive architecture and shared elements. The most widespread arrays were composed of a ~850 bp element, repeated in direct orientation without interspersed sequence and absent from the rest of the genome, which we thus called _Sultan_ for _SUBtelomeric Long TANdem_ repeat (Figure 1, green boxes).

We categorized all subtelomeres into four classes. The 27 subtelomeres containing _Sultan_ arrays adjacent to the telomeres belonged either to class A, if the _Sultan_ elements overall closely matched the most common ~850 bp sequence, or class B, if all the _Sultan_ elements carried large insertions (Supplemental File F1). The number of _Sultan_ copies in an array was highly variable, with an overall median of 14 repeats (Supplemental Table ST1). In the four class C subtelomeres, _Sultan_ arrays are separated from the telomeres by large arrays of other repetitive sequences (Figure 1, pale pink boxes). Finally, class D extremities _1_L, 8_R and 14_R contained rDNA as the only subtelomeric element (Figure 1, purple segments; Supplemental Figure S1).

The tandem repeat arrangement of the _Sultan_ arrays was confirmed experimentally by Southern blot using restriction sites within the _Sultan_ element and probing with a _Sultan_-specific oligonucleotide, revealing a major ~850 bp band as well as bands corresponding to _Sultan_ elements with insertions (Supplemental Figure S2). Uncut _Sultan_ arrays migrated as a smeared due to the variable number of _Sultans_ across subtelomeres and shearing of long DNA molecules during extraction (Supplemental Figure S2). Importantly, to verify that the number of _Sultan_ repeats was correctly assessed in class A and B subtelomeres, we manually verified the colinearity between individual long reads from the raw unassembled dataset (Supplemental Figure S3). Southern blotting and analysis of read mapping were also used to confirm the structure of class C and D subtelomeres (see below).

In 29 out of the 31 _Sultan_-containing subtelomeres, we found a non-repetitive sequence adjacent to the most centromere-proximal _Sultan_ that we called ‘Spacer’ (Figure 1, blue boxes), since it seemed to connect the _Sultan_ array to a (GGGA)_n_ microsatellite (Figure 1, yellow boxes).

Most _Spacer_ sequences were 450–550 bp long and on the _Sultan_ repeat side the first dozen nucleotides were highly conserved across subtelomeres (Supplemental File F1, TG GTGAGAGCAAC found in 24 subtelomeres and TG GTG GCCGGGCAACCATTT found in 4, the two least conserved nucleotides are in bold). Three _Spacers_ were different: the one in subtelomere 12_L lost homology to the others after the first 40 nt, 13_R was truncated on the _Sultan_ repeat side and 10_R displayed a 140-bp insertion just downstream of the highly conserved start described above.

Since it is shared by 27 out of 34 chromosome extremities, we propose that the canonical architecture of a subtelomere in _C. reinhardtii_ is, from telomeres inward, an array of _Sultan_ repeats, a _Spacer_ sequence and a _G-rich microsatellite_ array.

### Sultan element organization

To further examine subtelomere architecture, we compared the sequences of all 483 _Sultan_ elements pair-wise (Figure 2A; Supplemental File F1). We found that _Sultan_ copies were systematically more conserved within a given subtelomere than between them. This analysis also revealed that some subtelomeres, such as _2_L, 2_R, 8_L and 12_L, shared highly similar _Sultan_ elements (Figure 2A).

Most _Sultan_ elements contained a telomere-like sequence (CCTAAA, CCTAA or CTAAA) on their left border (Figure 2B). Interestingly, this sequence served as a seamless transition into the telomeric tracts (CCCTAAAA) on most subtelomeres (Figure 2D). Exceptions are shown by black diamonds in Figure 1 and exemplified by _2_R in Figure 2D, where the telomere-proximal _Sultan_ lacked >500 bp as compared to the following repeats. The 3’ side of _Sultan_ was also well conserved between chromosomes (Figure 2B). The last 10–12 nucleotides were truncated in the most centromere-proximal _Sultan_ repeat (Figure 2D), except in rare cases where an insertion or deletion modified the transition from the _Sultan_ array to the _Spacer_ (Figure 1, blue diamonds; Figure 2D, subtelomere _2_R).

We found that the _Sultan_ element was GC poor (average of 53% while the genome-wide average is 64%), so their large arrays formed significant regions with lower GC content at the genome-wide level (Supplemental Figure S4). The central part of _Sultan_ sequences was less conserved but showed similarity to the minisatellite _MSAT2_CR_ (Figure 2B), composed of a 184-bp monomer (https://www.girinst.org/2005/vol5/issue3/MSAT-2_CR.html). _MSAT2_CR_ was not restricted to subtelomeres and was present in two arrays >10 kb located immediately upstream of the putatively centromeric _Zepplike repeats of chromosomes 11 and 13_ (64). The _Sultan_ repeat itself is not a TE: it was detected neither in a previous large-scale survey of TEs, including the terminal repeat in miniature (TRIM) retrotransposons that are shorter than 1000 bp and form long arrays (79), nor in a recent annotation of _Chlamydomonas_ TEs (64), nor in a search against Pfam databases.

In class B subtelomeres, _Sultan_ repeats were longer than in class A, due to the presence of large insertions homologous to various TEs (Figure 2C and Supplemental Table ST2). On a given class B subtelomere, all _Sultan_ repeats
Figure 1. Architecture of subtelomeres in *C. reinhardtii* strain CC-1690. Left and right ends (L and R, respectively) of CC-1690 chromosomes are depicted with telomeres on the left-hand side, sorted by class and number of subtelomeric elements, which are displayed as boxes drawn at scale. The most common architecture, class A subtelomeres, comprises a telomere tract (black), a tandem array of Sultan repeats (green; numbering starts on the telomere side), a Spacer sequence (blue) and a G-rich microsatellite (yellow). Distinct large DNA insertions (brown) found in the Sultan repeats define the class B subtelomeres. Other repeats (pink) are found upstream of the Sultan array in class C subtelomeres (see Figure 4). Arrays of ribosomal DNA (purple) compose class D subtelomeres (Supplemental Figure S1). The display of the longest class C and D subtelomeres is not at scale and interrupted by ‘//’, while ‘?’ marks the only elusive molecule end due to assembly collapse. Diamonds denote junctions with telomere or Spacer sequences interrupting a Sultan element (see Supplemental Table ST1).

(shared the same inserted element with only minor variations in sequence. The inserted elements were different for each class B subtelomere.

To obtain insights into their propagation, we analysed the similarity between Sultan elements within a subtelomere. On most subtelomeres, individual Sultan repeats contain very few variations as compared to the local consensus sequence (>99.5% identity). Because single-nucleotide variants (SNVs) might result from sequencing and assembly errors, we only used INDELs found in at least two repeats to infer Sultan similarity within a given array. The class A subtelomeres 4_L and 7_R harboured in a subset of their Sultan elements an insertion of 245 bp and a duplication of 12 bp, respectively (Figure 3). Since in these examples the modified Sultan repeats were not contiguous and an identical pattern of modified and standard Sultans was found at least twice, we inferred that duplication of Sultan elements could involve multiple copies in a single event (Figure 3B and D, dotted brackets).

The 4 class C subtelomeres display distinct repeat arrays composed of Subtile and Suber elements

As depicted in Figure 1, Sultan arrays were not adjacent to telomeres in class C subtelomeres 3_R, 5_R, 15_L and 16_L. Using repeat detectors (XSTREAM and TRF) and BLAST, we found the sequence of variable length on the telomeric side to contain two new types of repeats described below, unrelated to the Sultan element, as well as vast low-complexity regions and short repeats.

All class C subtelomeres contained a ~190 bp repeat that we named Subtile for SUBTelomeric repeat of Intermediate Length (Figure 4A, B and D). The 133 Subtile copies found in the CC-1690 assembly formed 29 tandem repeat arrays of
Figure 2. The Sultan element. All individual Sultan sequences were aligned using MAFFT (Supplemental File F1). (A) Pairwise distance heatmap of 483 individual Sultan copies. The colour scale of the distances, with Jukes-Cantor correction for multiple substitutions, ranges from 0 (green) to 164 (red). Lower left: scheme depicting the numbering of Sultan elements within a subtelomere. (B) Conservation of nucleotides (black dots) in the consensus sequence. In addition, the moving 40 bp average is plotted as a line (red to green gradient). The sequence logos are shown for the most conserved regions. The telomere-like sequence at start is highlighted in yellow. Similarity with satellite MSAT2.CR is indicated in dark blue. (C) Location of the largest insertions (triangle) and deletion (X) in Sultan repeats from class A, B and C subtelomeres (Supplemental Table ST2). (D) Alignment of the first and last nucleotides of representative Sultan arrays showing phased (3L and 6R) and unphased transitions (2R) to telomere repeats. 5‘ telomere-like sequences are shown in yellow, a 10–12-bp sequence in the 3‘ region absent in the Sultan closest to the Spacer is highlighted in green and the 5‘ region of Spacer in blue. A 22-bp insertion in the transition from Sultan to Spacer in subtelomere 2.R is shown in red.

We further identified a third type of repeat in the 5.R, 3.R and 15.L subtelomeres, up to 2450 bp in length, that we called Suber for SUBtelomeric Extra-long Repeats. The 147 Subers assembled into massive arrays and analysis of the Suber variants indicated that they were also generated by segmental duplication. Four large INDELs (>400 bp) allowed us to define five main types of Suber (Figure 4C; Supplemental File F1), which formed a homogeneous array on subtelomere 15.L (52 kb) and two similar hybrid arrays on subtelomeres 5.R and 3.R (108 and 66 kb respectively) (Figure 4D). In addition, subtelomere 3.R carried individual Suber copies between the Subile arrays found in the centromere-proximal region. As for Subile repeats, the similarity between subtelomeres 5.R and 3.R suggested an inter-chromosomal recombination, but the different numbers of 2461-bp Suber copies (green, 10 in 5.R versus 16 in 3.R) and 1475-bp Suber copies (red, 68 versus 35) suggested that Suber repeats continued to propagate in situ after the recombination event. The tandem repeat organization of the Suber arrays was verified experimentally by Southern blot using a restriction enzyme that cut once in the

various lengths, each containing between 1 and 12 Subile copies. Several types of INDELs were detected upon alignment of Subile copies (Figure 4A; Supplemental File F1). For example, the last copy of an array on the centromere side was always truncated at the 3‘ end side, by either 47 nt (Figure 4A, blue) or 57 nt (Figure 4A, dark green); in six arrays, the telomere-proximal copy was 5‘-truncated by 134 nt (Figure 4A, red); in six other arrays, the sixth copy was larger due to an unrelated extra sequence of 146 nt (Figure 4A, orange). Various combinations of these variants create 6 main types of Subile arrays (Figure 4B) and the dotted lines suggest possible routes for their generation. The number of arrays per subtelomere also varied greatly, from 1 (5.R, 16.L) to 21 (3.R). The structural alignment of the 29 arrays (Figure 4D, simplified as plain boxes as shown in Figure 4B) and their pattern of localisation in the subtelomeres suggested that full arrays and even series of arrays duplicated and propagated between chromosome arms. The analysis of non-repetitive sequences found between the arrays (Other DNA in Figure 4D) suggested that they were likely duplicated along with the Subile arrays.
Figure 3. Evidence for multiple-copy duplication events in Sultan arrays. (A) and (C) Alignment of Sultan repeat sequences from class A subtelomeres 4_L and 7_R, from the telomere-proximal (top) to the Spacer-proximal (bottom). Conservation of nucleotide across repeats is indicated in dark blue. Vertical black bars and ‘//’ denote sequence portions not shown. Sultan repeats highlighted in black present a large insertion, represented in brown. Orange frames highlight duplications (including a putative hAT-N2 8-bp target-site duplication in 4_L). In (C), the end of the telomere is highlighted in yellow. (B) and (D) Sketch of 4_L and 7_R Sultan arrays with the conserved insertions (brown). Dotted brackets indicate multi-Sultan segments likely duplicated ‘en bloc’ on each subtelomere.

Suber element, revealing the 3 main types migrating at sizes ~1.5, ~1.9 and ~2.5 kb (Supplemental Figure S2). Additionally, BLAST and Conserved Domain searches indicate homology with bacterial HNH endonucleases, which belong to the homing endonuclease superfamily and can code for self-splicing introns and inteins (http://pfam.xfam.org/family/HNH). Only a few Subers contained these HNH-like regions in putative open-reading frames. Each Suber element also contained a telomere-like sequence at its 3’ side, corresponding to up to 10 degenerated repeats (~80 bp). In Southern blots, the telomeric probe indeed hybridized with the bands corresponding to the Suber element (Supplemental Figure S2).

The 3 class D subtelomeres are entirely composed of ribosomal DNA

In 1_L, only one partial and one complete rDNA copy (which was disrupted by a retrotransposon) were present, which were capped by a telomere (Supplemental Figure S1A, B and D). To estimate the size of the rDNA arrays, we calculated the coverage of Illumina sequencing reads from CC-1690 (80) mapped on a full rDNA unit, normalized it to the average whole genome coverage and estimated a total of 350 copies in the genome, corresponding to ~3 Mb, consistent with previous estimates (81,82). The rDNA arrays in 8_R and 14_R subtelomeres were therefore predicted to be much longer than average Nanopore reads, explaining why the genome assembly could not reach the actual end of the chromosomes. Nevertheless, we found in the raw data numerous reads carrying both rDNA and telomere repeats of >300 bp, which we were able to cluster into three distinct telomere-rDNA junctions (Supplemental Figure S1B and C). One was linked to the non-transcribed spacer (NTS) region of the rDNA unit and easily identified as corresponding to subtelomere 1_L based on its perfect match to the assembly. In another junction, the sequence transitioned seamlessly from the telomere repeats into a NTS through
Figure 4. Arrays of Subtile and Suber repeats populate 4 subtelomeres in CC-1690. (A) Diagram of putative insertion/deletion steps depicting Subtile repeat variants, based on sequence alignment (Supplemental File F1). (B) Diagram of structural variations in Subtile arrays (left) and their simplified plain box representation (right). (C) Alignment (Supplemental File F1) and distance matrix of all Suber repeats. Light to dark blue colour scale indicates increasing conservation. Subtypes colored on the diagonal of the distance matrix are depicted in the lower right diagram. (D) Map of Subtile and Suber arrays in class C subtelomeres drawn at scale following colour code shown in (B) and (C). Telomeres are shown as black boxes, Sultan repeat arrays as pale green boxes, other DNA sequences in white.

a telomere-like sequence present within the NTS. In the last junction, the telomere interrupted the 18S gene. While we could not unambiguously ascribe the latter two junctions to the 8_R or the 14_R subtelomeres, this analysis demonstrated that all 3 rDNA subtelomeres were directly capped by telomeres with no other sequence in between. As for the Suber element, Southern blots using the telomeric probe readily identified the telomere-like sequence in the NTS-derived bands (Supplemental Figure S2). We also characterised the centromere-proximal sequence immediately following the rDNA arrays, which consisted of a common sequence we termed RAE (rDNA-associated element) followed by a minisatellite repeat (MSAT-10_cRei) (Supplemental Figure S1B).

Transposable elements populate subtelomeres downstream of the Spacer sequence

TEs are quite common in the subtelomeres of many organisms and can even function as telomeres in D. melanogaster (4). In C. reinhardtii, we found that, downstream of the G-rich repeats, a region most often spanning 5–15 kb and reaching ~50 kb on some chromosome arms was generally populated by TEs, with exon density increasing progressively beyond these regions towards the centromeres (Supplemental Figure S5). The L1 LINE element L1-5_cRei was found to specifically target the (GGGA)_n motif and its copy number was enriched more than 50-fold in the 20 kb immediately downstream of Spacers relative to the rest of the genome (Supplemental Figure S7C). It is possible that L1-5_cRei has evolved such a targeted insertion sequence as a result of the abundance of the G-rich repeat in subtelomeres, which may serve as a safe haven that minimizes any deleterious effects of insertion.

Chromatin modifications and transcription at subtelomeres

To investigate chromatin modification at C. reinhardtii subtelomeres, we analysed methylation of cytosines (5-methylcytosines, 5mC) in Cpg contexts, an epigenetic mark primarily associated with transcriptional silencing (83), detected directly from the Nanopore sequencing reads using DeepSignal, a deep-learning-based method (77). We first confirmed that this method was able to detect regions within previously identified hypermethylated loci (84), characterised by the most common element LI-1_cRei (also known as ZeppL-1_cRei) which was recently shown to be the major constituent repeat of centromeres (64) (Supplemental Figure S6A). We found that Sultan arrays were clearly hypermethylated with their boundary at the Spacer coinciding very well with a decline of the methylation level (Figure 5A and Supplemental Figure S6B). Suber arrays showed an intermediate level of 5mC whereas Subtile elements could be found associated with either hyper or hypomethylation. Interestingly, rDNA repeats on the centromeric side of 8_R and 14_R subtelomeres were largely hypomethylated while the smaller rDNA locus at 1_L subtelomere was hypermethylated (Figure 5A and Supplemental Figure S6B). Inspection of methylation of the 8_R/14_R rDNA repeats next to the telomeres showed a high level of 5mC close to the telomeres and, moving away, a pro-
Figure 5. Chromatin modifications and transcription at subtelomeres. (A) DNA methylation (5mC at CpG context) at subtelomeres representative of the 4 A to D classes, detected from the Nanopore reads. The location of the Sultan, Suber, Subtile and rDNA arrays (with the NTS marked as an open circle) are depicted in colour on the x-axis. The 14 R subtelomere being truncated, DNA methylation at the telomere-rDNA junctions is shown in (Supplemental Figure S6), along with other subtelomeres. (B) Plot of nucleotide coverage of RNA-seq reads along the 11 most expressed Spacers, starting from the Sultan/Spacer junction (0) towards the G-rich repeats. Transcript dataset from (72). (C) H3K4me3 enrichment values extracted for the Spacer sequences, all gene promoters and 20 000 random genomic regions of 500 bp (size similar to the Spacers), using H3K4me3 ChIP-seq data (accession number: PRJNA681680) mapped to the CC-1690 assembly. *p-value < 10⁻³, Wilcoxon–Mann–Whitney test.

We then asked if subtelomeres were transcribed downstream of the Sultan arrays and we found 100%-matching reads in nearly all Spacer sequences in the deep transcriptome data published by (72) (Figure 5B and Supplemental Figure S7A and B). The only exception was the 5′-truncated 13_R Spacer. Using Iso-Seq data we observed full-length polyadenylated transcripts originating from the Spacer towards the centromere at 14 subtelomeres (Supplemental Figure S7C). Furthermore, we observed peaks in H3K4me3 ChIP-seq coverage at the Spacers (Figure 5C and Supplemental Figure S7C), which are highly indicative of transcription start sites and active promoters in C. reinhardtii (87). The enrichment of H3K4me3 marks observed at Spacers was significantly higher than the background signal at random genomic sites of equal length, but lower than the one at gene promoters (Figure 5C), suggesting that the Spacers might have lower promoter activity or that their activity might be variable across different subtelomeres or from cell to cell (e.g. if cells are not synchronized, see Supplemental Figure S7B). The transcripts were generally characterized by a conserved 5′-splice site (G\∧GTAG), with the (GGGA)n repeat positioned at the beginning of the first, usually extremely long, intron. Sequence similarity was limited to the first exon. The transcripts originating from the Spacer only contained very short putative coding sequences (average of 124 aa). BLASTx-based comparison of the transcribed sequences to Chlorophytes proteins only yielded some hits to TE genes, suggesting that they may represent long non-coding RNAs. Interestingly, the expression of Spacer sequences peaked at dusk in synchronous diurnal cultures (Supplemental Figure S7B), correlating with tran-
scription of genes associated with DNA replication (72). We also observed transcription from the rDNA sequences following the rDNA arrays on 8_R and 14_R (Supplemental Figure S1B).

Overall, *C. reinhardtii* subtelomeres appeared to be largely heterochromatic over the arrays of repetitive elements, with the exception of the two major rDNA clusters at 8_R and 14_R. In contrast, the *Spacer* sequences appear to act as active promoters for transcription toward the centromere.

**Interstrain variations provide insights into subtelomere evolution**

All common laboratory strains were derived from two parental genotypes in the 1940s, and as a result their genomes are mosaics of two divergent haplotypes, known as haplotype 1 and 2 (88). As such, any differences within-haplotype are expected to be the result of evolution during approximately 75 years of laboratory culture, while between-haplotype differences reflect ancestral variation in the population the strains were derived from. To investigate the evolution of subtelomeres in *C. reinhardtii*, we compared the *Sultan* repeats in the CC-1690 genome with those in other laboratory strains, CC-503 and CC-4532, for which assemblies were recently generated from PacBio sequencing data. The former has served initially as the reference strain for the genome (89), but has been recently replaced by the latter. The subtelomeres of CC-503 all belong to haplotype 1, while 8_L in CC-1690 and 6_L in CC-4532 belong to haplotype 2.

A maximum likelihood phylogenetic tree was built from the aligned *Sultan* consensus sequences from each subtelomere (Figure 6A). When the sequence data were available, the *Sultan* repeats of the same chromosome end but from the different laboratory strains generally grouped together, suggesting that most *Sultan* arrays have not relocated since these strains were genetically separated in the laboratory. A notable exception is the grouping of CC-503 2_R with 9_L in other strains, due to a documented reciprocal translocation between these chromosome arms that occurred in the laboratory history of CC-503. We also found that the 6_L *Sultan* of CC-4532 (haplotype 2) was highly similar to that found on 11_L in all three laboratory strains, including CC-4532 itself. This implies the existence of two 6_L subtelomere alleles, with the allele in haplotype 2 possibly formed by copying that of 11_L. In contrast, the 8_L *Sultan* of CC-1690 (haplotype 2) was closely related to that in the other two strains, in spite of their belonging to haplotype 1.

We also analysed a field isolate from North Carolina (CC-2931), which belongs to a highly divergent population relative to laboratory strains (80,90). Here, a clear grouping with laboratory strains was observed only for 9 out of 24 available extremities (Figure 6A). Some of this variation can be attributed to larger chromosomal rearrangements: the standard laboratory strain chromosome arms 1_R, 6_L, and 10_R are equivalent to 10_R, 1_R and 6_L, respectively, in CC-2931. However, many of the chromosome arms displaying variable subtelomeres were otherwise entirely syntenous, implying extensive allelic variation in *Sultans* as evolutionary distances increase.

We found that the phylogenetic tree of the *Spacer* sequences from different subtelomeres was poorly concordant with the phylogeny of the *Sultan* element, as shown for CC-1690 (Supplemental Figure S8), which might indicate that the *Spacers* evolve at a faster rate outside the highly conserved motifs.

To quantify the variability in the number of copies in a given *Sultan* array between strains, we mapped Illumina sequencing reads of laboratory strains (80) against the *Sultan* consensus sequences from CC-1690. We normalized the median nucleotide coverage of each *Sultan* consensus by the average whole genome coverage (Figure 6B and C). As a control, plotting the results from CC-1690 Illumina sequencing against the number of *Sultan* repeats observed in the CC-1690 end-to-end chromosomal assembly (Figure 6B, blue) showed a linear relationship with only a slight overestimation of the repeat number. The same approach was then applied to CC-503 (Figure 6B, orange) and other strains. The overall distribution of *Sultan* copy number across a panel of laboratory strains is shown as a boxplot of repeat counts for each subtelomere consensus (Figure 6C) and a detailed comparison is displayed in Supplemental Figure S9. Repeat counts were generally close to that of CC-1690 for most subtelomeres (median coefficient of variation = 20%). Several of the major differences were in agreement with the expected distribution of the two alternative haplotypes amongst strains (88): CC-1009 and CC-408 had shorter subtelomeres at 6_L, 9_L and 12_R, in accordance with their carrying haplotype 2 at these loci. The shorter 6_L and 12_R subtelomeres were also found in CC-124 (also haplotype 2); at 8_L, strains with haplotype 1 (CC-503, CC-125, CC-1009, CC-408) had longer arrays than those with haplotype 2 (CC-1690, CC-1010), except for CC-124. This length variation between the haplotypes again supports substantial subtelomere allelic variation within the species. For *Suber* and *Subtile* repeats, we observed mapped Illumina reads from all laboratory strain datasets, but only in some of the available wild isolates, indicating that their presence might not be fully conserved in the species. Accordingly, the CC-2931 assembly lacked both *Suber* and *Subtile* repeats. Interestingly, we found that the 1_L subtelomere in the CC-2931 assembly featured an rDNA array with similar organization to 8_R and 14_L, i.e. complete rDNA copies followed by a full-length rDNA-associated element (RAE) and *MSAT-10_c Rei*. In contrast, CC-503 and CC-4532 showed a truncated array as in CC-1690 (Supplemental Figure S1D). The unusual 1_L rDNA organization is thus a characteristic feature of *C. reinhardtii* laboratory strains.

Overall, the interstrain analyses of *Sultan* sequence variations and copy number indicates that subtelomeres have been remarkably stable during laboratory culture. However, differences between laboratory strain haplotypes and between the laboratory strains and a divergent field isolate supports extensive subtelomere allelic variation within the species.

*Subtelomeres in other green algae*

We wondered whether a subtelomeric organization similar to that in *C. reinhardtii* would be found in other algae. We concentrated on the few algal genomes that present the degree of completeness and accuracy that was needed for
eral chromosomes appear almost fully conserved with (84% and 83% of their genome length, respectively). Sev-
erous long read genome assemblies were recently produced. Interestingly, we did not find by BLAST any trace of the C. reinhardtii are C. incerta and C. schloesseri, for which highly contigu-
ous long read genome assemblies were recently produced (64). They show a high degree of synteny with C. reinhardtii (84% and 83% of their genome length, respectively). Several chromosomes appear almost fully conserved with C. reinhardtii, and they putatively share a centromeric structure based on arrays of Zepp-like retrotransposons. Inter-
estingly, we did not find by BLAST any trace of the Sultan, Subtile and Suber repeats described in C. reinhardtii. In C. incerta, the 4 contigs showing terminal arrays of the 8-bp telomeric repeats shared a well-conserved 350 nt repeat forming immediately-subtelomeric arrays (Figure 7). We called this repeat Subrin, for SUBtelomic Repeat of C. INCerta. Subrin arrays were found in 29 additional contigs lacking telomeres, but in an orientation generally consistent with a subtelomeric position. Some arrays were very extensive and we counted a total of 1819 Subrin copies in the assembly. Subrin copies were more similar within an array than across arrays, again indicating preferential local tandem duplications. In 29 of the 33 contigs, we could collect the sequence immediately upstream of the Subrin ar-
ray, and found that 24 of them started with a homologous spacer sequence, generally spanning ~1.2 kb. No G-rich repeat region was observed. We conclude that in C. incerta also, the majority of the chromosomes comprise a repetitive subtelomeric sequence anchored on a shared spacer, even though the sequences themselves were unrelated to those in C. reinhardtii.

Subjected to the same analysis, C. schloesseri revealed a similar type of subtelomere organization (Figure 7), but again based on repeats unrelated to either the Sultan or the Subrin. We called these repeats Subrecs, for SUBtelomic REPeat of C. Schloesseri. However, they displayed more heterogeneity than in C. reinhardtii or C. incerta. We distin-
guished two types, unrelated in sequence, called Subrecs-I (319–321 bp, 233 copies) and Subrecs-II (266–327 bp, 298 copies). They formed arrays in respectively 14 and 15 contigs but immediately adjoined the terminal telomeric repeats only in respectively 2 and 6 cases. This is because many contigs carried one or even two non-terminal telomeric repeat arrays, often adjoined by or embedded in a Subrecs array. Noticeably, a contig carried only type-I or type-II Sub-
recs, never a mixture. As in C. reinhardtii, the centromere-
proximal Subrecs were adjacent to a shared spacer, again with a short 3'-truncation (2 nt for Subrecs-I, 7 for Subrecs-II). Two types of spacers could be identified, one associated with Subrecs-I or -II arrays, the other one exclusively with Subrecs-II.

In C. incerta and C. schloesseri, the contigs terminating in rDNA arrays are syntenous with 1L, 8R and 14R in C. reinhardtii (64). On the centromeric side, the rDNA arrays were followed by a RAE-like sequence and a stretch of satellite DNA, as in C. reinhardtii. As in CC-2931, the con-
tigs syntenous with \( I_L \) in both species did not appear to harbour disrupted/truncated rDNA arrays as was observed in the \( C. \ reinhardtii \) laboratory strains.

\textit{Edaphochlamys debaryana} is a more distant relative of \( C. \ reinhardtii \), but also groups within the core-Reinhardtinia clade of the Chlamydomonadales. The synteny with \( C. \ reinhardtii \) is less marked (46%) and the assembly is less contiguous. Here telomeric repeats of 7 nt (CCCTAAA) were observed, but in only two telomeres were they associated with a subtelomere-specific repeat which we called \textit{Subred} (SUBtelomeric Repeat of \textit{Edaphochlamys Debaryana}) (Figure 7). The rDNA arrays in \textit{E. debaryana} were not sufficiently well-assembled for comparison.

We were also able to characterize subtelomeres in a number of distantly related algal species that likely shared a last common ancestor with \textit{Chlamydomonas} in the pre-Cambrian. The genome of \textit{Chromochloris zofingiensis} (class Chlorophyceae, order Sphaeropleales), also showed 7-nt telomeric repeats but we did not observe any subtelomere-specific repeats. In the almost fully assembled genome of \textit{Coccomyxa subellipsoidea} (class Trebouxiophyceae), the 20 chromosomes carry 7-nt telomere repeats at both extremities.
ities. In 19 extremities, the subtelomere comprises what we call a Subrel-I repeat (SUBtelomeric Repeat of Coccomyxa subEllipsopidea) of 166 nt (1 to 11 copies per extremity, 86 in total) (Figure 7). Only in 3 cases was the array adjacent to the telomere. Again, a conserved spacer sequence of ~180 nt was found on the centromeric side of every Subrel array, and in one case S′-truncated and abutting the telomeric array, suggestive of a deletion of the intervening sequence and part of the Subrel array. In addition, other repeats called Subrel-II (~90 nt) and Subrel-III (~19 nt) were found in respectively 3 and 2 subtelomeres.

In Ostreococcus lucimarinus (class Mamiellophyceae), with 21 chromosomes, no subtelomeric repeat could be identified. However, many extremities shared a homologous sequence immediately after the telomere (Figure 7). Type-I (up to 411 nt) was found in 21 extremities, Type-II (up to 844 nt) in 15. In both groups, especially Type-II, some subtelomeres were truncated at the S′ end and the junction with the telomeric repeat was in various phases. Combined with the presence of fragments of the Type-I sequence at the S′ of a Type-II element, this suggests a history of partial deletions and repair.

DISCUSSION

Subtelomeres are notoriously difficult to assemble due to their repetitive nature. Previous reference genomes of C. reinhardtii failed to provide a clear picture of the subtelomeres and also lacked telomere sequences at most extremities. Using long read sequencing data (PacBio and Oxford Nanopore Technology) and de novo genome assemblies (62–64) (Craig et al., bioRxiv, doi: https://doi.org/10.1101/2021.04.23.441226), we now provide a nearly complete map of all chromosome extremities in C. reinhardtii, including telomere sequences at 33 out of 34 extremities. Given the mean read length and N50, both equal to 55 kb, of the Nanopore reads, and the contiguity of our assembly, we are confident that our description of the subtelomeres is accurate, especially for the exact number of repeated elements in each subtelomere. We describe three new types of repeated elements present in C. reinhardtii subtelomeres and group subtelomeres into 4 classes based on their composition. We further provide functional and mechanistic insights into their chromatin status and their evolution at the intra and inter species level.

Subtelomere chromatin and transcription

While heterochromatin is thought to be a general property of subtelomeres in many organisms, chromatin organization at subtelomeres might be more complex (27) and for instance, its nature in plants is still a matter of debate (25,26,91). Here, we find that chromatin in C. reinhardtii might follow a simple organization. Sultan and Suber arrays show DNA methylation, likely associated with H3K9me1 for the Sultan arrays (86), both hallmarks of heterochromatin. A previous study examining C. reinhardtii methylome (84) found highly methylated DNA at 23 major loci in the nuclear genome correlating with highly repetitive and non-protein-coding DNA, 2 of which were located near chromosome ends (5′R and 11′R) and likely corresponded to Sultan elements. However, most of the subtelomeric 5mC signal we describe here was not detected in that work because the sequences were aligned to the v5 genome that lacks proper assembly of the subtelomeres.

The heterochromatin domain has a clear and sharp boundary at the Spacer sequence, marked by H3K4me3. Its S′ part functions as a promoter, active essentially at dusk and during the first phase of night in a light-dark cycle, concomitantly with replication and histone deposition (72). The downstream region of the subtelomere is transcribed and no longer methylated, thus defining a euchromatic domain, which encompasses all TEs and distal subtelomeric genes. The putatively non-coding but spliced and polyadenylated transcripts emanating from the Spacer are similar to sub-TERRA and other subtelomeric transcripts, as described in multiple organisms (29,92), with potential functions in telomere maintenance that remain to be investigated.

The three subtelomeric rDNA clusters are directly capped by telomere sequences with no other sequence in between, as in A. thaliana (93). A subtelomeric localization for the rDNA is a wide-spread feature found in many eukaryotes, including plants, animals and fission yeast (15,94–97). Epigenetic regulation can limit the number of active rDNA genes, depending on the context. In C. reinhardtii, we find that only the laboratory strain-specific truncated copy of rDNA at 1′L shows hypermethylation while the two major rDNA arrays at 8′R and 14′R, representing a total of ~3 Mb, are both hypomethylated and presumably transcribed, except for the ~60 kb adjacent to the telomeres. This stands in contrast to the subtelomeric clusters of rDNA genes in A. thaliana found in two equally sized nucleolar organizer regions (NORs), NOR2 and NOR4, with NOR2 subjected to chromatin-mediated silencing and NOR4 being active (93). Whether, in C. reinhardtii, the rDNA clusters can be heterochromatinized to regulate the physiological requirement for ribosomes in specific conditions will be interesting to explore.

We conclude that subtelomeres in C. reinhardtii overall show a clear two-domain organization: (i) a telomere-proximal heterochromatin domain, including hypermethylated Sultan and Suber arrays and the distal copies of the rDNA arrays, (ii) followed by a transcriptionally active euchromatin domain. When present, the Spacer sequence, acting as a promoter in the centromeric direction, forms the boundary.

Segmental duplication and contraction within subtelomeres

An important finding is that Sultan elements show higher similarity within a subtelomere than between subtelomeres, suggesting a very low frequency of rearrangements involving different extremities. This observation is consistent with the relatively low efficiency of homology-based recombination in vegetative C. reinhardtii cells (98). It is however in contrast with what is known in other species where subtelomeric regions show signatures of frequent interchromosomal recombination between repeated sequences (14,37,44). Nevertheless, although infrequent, rearrangements between subtelomeres did occur, as evidenced by the propagation of the Sultan elements on almost every sub-
telomere and by the similarities between the arrangement of *Subtile* and *Suber* repeats in different subtelomeres. The high sequence similarity between *Sultan* elements belonging to the same subtelomere suggests that at some point, only one *Sultan* element was present at a given subtelomere, or maybe sometimes two for the *Sultan* arrays composed of two slightly different types of repeat (e.g. 4_L or 7_R, Figure 3). Another argument in this direction is that the *Sultan* elements in each class B subtelomere contained a single type of insertion.

Segmental duplication of one or several tandem *Sultan* elements may be explained by a number of mechanisms. Based on the above, we favour mechanistic models that do not require other chromosome ends. We thus propose that *Sultan* elements propagated essentially through two post replicative mechanisms, unequal sister chromatid exchange (SCE) and/or break-induced replication (BIR) using the sister chromatid. SCE was shown to occur at high rates near chromosome ends and to allow segmental duplication of subtelomeric elements by unequal exchange (43,45,99). SCE at chromosome ends is also found at increased frequency in alternative lengthening of telomeres (ALT) cancer cells (51,100). BIR is a recombination-based replication mechanism used when only one side of a double-strand break is available for homology search. It is therefore preferred at subtelomeres over gene conversion because the terminal fragment can be rapidly lost (101). Additionally, segmental duplication in yeast has been shown to heavily rely on Pol32, a non-essential subunit of polymerase Pol6, required for BIR (102). BIR also serves as an alternative mechanism for telomere maintenance, as shown in yeast and ALT cancer cells (103,104), leading to the amplification of the subtelomeric Y′ elements in one subtype of telomerase-independent survivors in yeast. Both unequal SCE and BIR using the sister chromatid can be promoted by the short telomere sequence at the 5′ end of *Sultan* elements and the annealing to this sequence at different positions can lead to segmental duplication.

It is also conceivable that the *Sultan* arrays might occasionally decrease in repeat copy number, as it was shown for Y′ tandem repeats in *S. cerevisiae* subtelomeres (43). These putative contraction events might also be promoted by the telomere sequence present at the 5′ end of the *Sultan* element. Indeed, since telomeres and repeated elements are difficult to replicate, which leads to replication fork stalling and collapse (105,106), resulting DNA breaks might be repaired by telomere healing primed by this seed sequence. This possibility is supported by the fact that in most cases the telomere seed sequence of the *Sultan* closest to the telomere is in phase with and transitions seamlessly into the telomeric tract. Such a mechanism would lead to the terminal deletion of a variable number of *Sultan* elements. Telomere-like sequences are also found in each *Suber* element and in the NTS of the rDNA, suggesting that contraction events based on telomere healing might occur on all classes of subtelomeres.

The current architecture of the *C. reinhardtii* subtelomeres can thus be described as resulting from a complex history of both expansion and collapse of *Sultan* elements, with some rare interchromosomal rearrangements.

**Evolution of subtelomeres within *C. reinhardtii* and beyond**

To provide an idea of how dynamic the subtelomeres of *C. reinhardtii* are, we compared the sequences of the *Sultan* and *Spacer* elements, as well as the copy number of the *Sultan* elements in each subtelomere, in different strains, including laboratory strains and field isolates. We found strong conservation of the *Sultan* sequences with no evidence for subtelomere-specific rearrangements during decades of laboratory culture. We also found relatively stable copy number of *Sultans* amongst strains, and the few exceptions to this could in almost all cases be traced to the strains carrying a distinct ancestral haplotype (88). We observed a mixed pattern in the genome of the divergent wild isolate CC-2931, with both conserved subtelomeres and evidence for polymorphic alleles, some obviously created by the translocation of *Sultan* elements between chromosomes. In addition, *Suber* and *Subtile* elements were totally absent from the CC-2931 assembly, suggesting a different organization of the subtelomeres that form class C in laboratory strains. Thus, substantial polymorphism of subtelomeres exists at the population and species-wide level. It remains to be seen whether such variation can cause genomic incompatibilities, which could potentially pave a way towards speciation.

At a larger evolutionary scale, we found specific repeated elements for most species of green algae we investigated, spanning almost a billion years of evolution. Interestingly, subtelomere organization in these algae seemed to follow a structure similar to *C. reinhardtii*, with an array of repeated elements adjoining the telomere and a spacer sequence conserved across subtelomeres, but the repeated element and the spacer sequence were unrelated across species. This study was limited by the small number of available chromosome level assemblies for green algae, but it suggests that subtelomeres in many green algal lineages have converged to strikingly similar organizations, with shared species-specific (usually repeated) elements populating the subtelomere. Further investigation of the underlying properties that drove the propagation of these elements, such as heterochromatin formation, binding of particular factors, replication problems or transcription from the spacer sequences, might contribute to a deeper understanding of subtelomere functions and evolution in algae.

**DATA AVAILABILITY**

All alignments can be found in Supplemental File F1. The scripts used in this work are included in Supplemental files F2, F3 and F4. The genomes used in this study are accessible at this secure self-certified URL: https://raba.ibpc.fr/home/ovallon@ibpc.fr/Briefcase/Chlamy_genomes

They include the assembly for CC-1690 (GenBank accession: JABWP000000000) with a correction for subtelomere 9_R, for CC-4532 and CC-503 (v6) (https://phytozome-next.jgi.doe.gov/), for *C. incertae* (https://phylocosm.jgi.doe.gov/Chlin1; NCBI accession: GCA_016834605.1), *C. schoenessleri* (https://phylocosm.jgi.doe.gov/Chls1; NCBI accession: GCA_016834595.1) and *E. debaryana* (https://phylocosm.jgi.doe.gov/Edade1; NCBI acces-
sion: GCA_016858145.1), and for CC-2931 (Craig et al., bioRxiv, doi: https://doi.org/10.1101/2021.04.23.441226).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the Department of Energy Joint Genome Institute and collaborators for pre-installation access to the genomes of Chlamydomonas reinhardtii strains CC-503 and CC-4532 in support of this analysis. We thank Maria Teresa Teixeira and Gilles Fischer for their critical reading of the manuscript.

FUNDING

Agence Nationale de la Recherche grants ‘AlgaTelo’ [ANR-17-CE20-0002-01], ‘PhenoVar’ [ANR-16-CE12-0019]; ‘Initiative d’Excellence’ program from the French State [‘DY-NAMO’, ANR-11-LABX-0011-01]; Ville de Paris (Programme Émergence(s)); R.C. is supported by a BBRC EASTBIO Doctoral Training Partnership grant. Funding for open access charge: Agence Nationale de la Recherche [ANR-17-CE20-0002-01]. Conflict of interest statement. None declared.

REFERENCES

1. Jain,D. and Cooper,J.P. (2010) Telomeric strategies: means to an end. Annu. Rev. Genet., 44, 243–269.
2. de Lange,T. (2018) Shelterin-mediated telomere protection. Annu. Rev. Genet., 52, 223–247.
3. Wellinger,R.J. and Zakian,V.A. (2012) Everything you ever wanted to know about Saccharomyces cerevisiae telomeres: beginning to end. Genetics, 191, 1073–1105.
4. Pardue,M.L. and DeBaryshe,P.G. (2011) Retrotransposons that maintain chromosome ends. Proc. Natl. Acad. Sci. U.S.A., 108, 20317–20324.
5. Roth,C.W., Kobeski,F., Walter,M.F. and Biessmann,H. (1997) Chromosome end elongation by recombination in the mosquito. Annu. Rev. Genet., 31, 191–207.
6. Cohn,M. and Edstrom,J.E. (1992) Telomere-associated repeats in Anopheles gambiae. Chromosome end elongation by recombination in the mosquito. Annu. Rev. Genet., 31, 191–207.
7. Cesare,A.J. and Reddel,R.R. (2010) Alternative lengthening of telomeres: models, mechanisms and implications. Nat. Rev. Genet., 11, 319–330.
8. Corcoran,L.M., Thompson,J.K., Walliker,D. and Kemp,D.J. (1988) Homologous recombination within subtelomeric repeat sequences generates chromosome size polymorphisms in F. falciparum. Cell, 53, 807–813.
9. Louis,E.J. (1995) The chromosome ends of Saccharomyces cerevisiae. Yeast, 11, 1553–1573.
10. Kim,J.M., Vanguri,S., Boeke,J.D., Gabriel,A. and Voytas,D.F. (1998) Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. Genome Res., 8, 464–478.
11. Fabre,E., Muller,H., Therizols,P., Lafontaine,I., Dujon,B. and Fairhead,C. (2005) Comparative genomics in homoeosaccae yeasts: evolution of sex, silencing, and subtelomeres. Mol. Biol. Evol., 22, 856–873.
12. Richard,M.M., Chen,N.W., Thareau,V., Pfieger,S., Blanchet,S., Pedrosa-Harand,A., Iwata,A., Chavarro,C., Jackson,S.A. and Geoffroy,V. (2013) The subtelomeric khipu satellite repeat from Phaseolus vulgaris: lessons learned from the genome analysis of the Andean genotype G19833. Front. Plant Sci., 4, 109.
13. Brown,C.A., Murray,A.W. and Verstrepen,K.J. (2010) Rapid expansion and functional divergence of subtelomeric gene families in yeasts. Curr. Biol., 20, 895–903.
14. Chen,N.W.G., Thareau,V., Ribeiro,T., Magdelanat,G., Ashfield,T., Innes,R.W., Pedrosa-Harand,A. and Geoffroy,V. (2018) Common bean subtelomeres are hot spots of recombination and favor resistance gene evolution. Front Plant Sci., 9, 1185.
15. Sochorova,J., Garcia,S., Galvez,F., Symonova,R. and Kovarik,A. (2018) Evolutionary trends in animal ribosomal DNA loci: introduction to a new online database. Chromosoma, 127, 141–150.
16. Otto,T.D., Bohme,U., Sanders,M., Reid,A., Bruske,E.L., Duffy,C.W., Bull,P.C., Pearson,R.D., Ahdi,A., Dimonte,S. et al. (2018) Long read assemblies of geographically dispersed Plasmodium falciparum isolates reveal highly structured subtelomeres. Wellcome Open Res., 3, 52.
17. Trask,B.J., Friedman,C., Martin-Gallardo,A., Rowen,L., Akinbami,C., Blankenship,J., Collins,C., Giorgi,D., Iadonato,S., Johnson,F. et al. (1998) Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. Hum. Mol. Genet., 7, 13–26.
18. Linardopoulou,E.V., Parghi,S.S., Friedman,C., Osborn,G.E., Parkhurst,S.M. and Trask,B.J. (2007) Human subtelomeric WASH genes encode a new subclass of the WASP family. PLoS Genet., 3, e237.
19. Elgin,S.C. and Reuter,G. (2013) Position-effect variation, heterochromatin formation, and gene silencing in Drosophila. Cold Spring Harb. Perspect. Biol., 5, a017780.
20. Koering,C.E., Pollice,A., Zibella,M.P., Bauwens,S., Puisieux,A., Vrbsky,J., Akimcheva,S., Watson,J.M., Turner,T.L., Daxinger,L., Fairhead,C. and decondensed silent chromatin in fission yeast. Nat. Commun., 6, 7753.
21. Pedram,M., Sprung,C.N., Gao,Q., Lo,A.W., Reynolds,G.E. and Murnane,J.P. (2006) Telomere position effect and silencing of transgenes near telomeres in the mouse. Mol. Cell. Biol., 26, 1865–1878.
22. Matsuda,A., Shikahige,Y., Ding,D.Q., Ohtsuki,S., Puisieux,A., Bruni,M., Brun,C., Martins,L., Sabatier,L., Peltzer,J.F. et al. (2002) Human telomeric position effect is determined by chromosomal context and telomeric chromatin integrity. EMBO Rep., 3, 1055–1061.
23. Gottschling,D.E., Aparicio,O.M., Billington,B.L. and Zakian,V.A. (1990) Position effect at S. cerevisiae telomeres: reversible repression of Pol II transcription. Cell, 63, 751–762.
24. Baur,J.A., Zou,Y., Shay,J.W. and Wright,W.E. (2001) Telomere position effect in human cells. Science, 292, 2075–2077.
25. Vrbsky,J., Akimcheva,S., Watson,J.M., Turner,T.L., Daxinger,L., Vysokt,B., Aufsatz,W. and Riha,K. (2010) siRNA-mediated methylation of Arabidopsis telomeres. PLoS Genet., 6, e1000986.
26. Vazquez-Sedas,M.I., Gomez-Arjona,F.M. and Vega-Polás,M.A. (2011) Arabidopsis thaliana telomeres exhibit euchromatic features. Nucleic Acids Res., 39, 2007–2017.
27. Hocheder,A. and Taddei,A. (2020) Subtelomeres as specialized chromatin domains. Bioessays, 42, e1900205.
28. Azzalin,C.M., Reichenback,P., Khoriauli,L., Giulotto,E. and Lingner,J. (2007) Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. Science, 318, 798–801.
29. Azzalin,C.M. and Lingner,J. (2015) Telomere functions grounding telomerase by telomerase in yeast. Trends Cell Biol., 25, 29–36.
30. Craven,R.J. and Peters,T.D. (1999) Dependence of the regulation of telomere length on the type of subtelomeric repeat in the yeast Saccharomyces cerevisiae. Genetics, 152, 1531–1541.
31. Jolivet,P., Serhal,K., Graf,M., Eberhard,S., Xu,Z., Luke,B. and Teixeira,M.T. (2019) A subtelomeric region affects telomerase-negative replicative senescence in Saccharomyces cerevisiae. Sci. Rep., 9, 1845.
32. Azzalin,C.M. and Lingner,J. (2007) Tel1 kinase and subtelomere-bound Tbf1 mediate preferential elongation of short telomeres by telomerase in yeast. EMBO Rep., 8, 1080–1085.
33. Schoeftner,S. and Blasco,M.A. (2008) Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. Nat. Cell Biol., 10, 228–236.
