The complete mitochondrial genome sequence of the green microalga Lobosphaera (Parietochloris) incisa reveals a new type of palindromic repetitive repeat

Nicolas J. Tourasse, Nastassia Shtaida, Inna Khozin-Goldberg, Sammy Boussiba, Olivier Vallon

Tourasse et al. BMC Genomics (2015) 16:580
DOI 10.1186/s12864-015-1792-x

BMC Genomics

RESEARCH ARTICLE Open Access

The complete mitochondrial genome sequence of the green microalga *Lobosphaera* (*Parietochloris*) *incisa* reveals a new type of palindromic repetitive repeat

Nicolas J. Tourasse[1,2,4*], Nastassia Shtaida[3], Inna Khozin-Goldberg[3], Sammy Boussiba[3] and Olivier Vallon[1]

Abstract

Background: *Lobosphaera incisa*, formerly known as *Myrmecia incisa* and then *Parietochloris incisa*, is an oleaginous unicellular green alga belonging to the class Trebouxiophyceae (Chlorophyta). It is the richest known plant source of arachidonic acid, an ω-6 poly-unsaturated fatty acid valued by the pharmaceutical and baby-food industries. It is therefore an organism of high biotechnological interest, and we recently reported the sequence of its chloroplast genome.

Results: We now report the complete sequence of the mitochondrial genome of *L. incisa* from high-throughput Illumina short-read sequencing. The circular chromosome of 69,997 bp is predicted to encode a total of 64 genes, some harboring specific self-splicing group I and group II introns. Overall, the gene content is highly similar to that of the mitochondrial genomes of other Trebouxiophyceae, with 34 protein-coding, 3 rRNA, and 27 tRNA genes. Genes are distributed in two clusters located on different DNA strands, a bipartite arrangement that suggests expression from two divergent promoters yielding polycistronic primary transcripts. The *L. incisa* mitochondrial genome contains families of intergenic dispersed DNA repeat sequences that are not shared with other known mitochondrial genomes of Trebouxiophyceae. The most peculiar feature of the genome is a repetitive palindromic repeat, the LIMP (*L. Incisa* Mitochondrial Palindrome), found 19 times in the genome. It is formed by repetitions of an AACCA pentanucleotide, followed by an invariant 7-nt loop and a complementary repeat of the TGGTT motif. Analysis of the genome sequencing reads indicates that the LIMP can be a substrate for large-scale genomic rearrangements. We speculate that LIMPs can act as origins of replication. Deep sequencing of the *L. incisa* transcriptome also suggests that the LIMPs with long stems are sites of transcript processing. The genome also contains five copies of a related palindromic repeat, the HyLIMP, with a 10-nt motif related to that of the LIMP.

Conclusions: The mitochondrial genome of *L. incisa* encodes a unique type of repetitive palindromic repeat sequence, the LIMP, which can mediate genome rearrangements and play a role in mitochondrial gene expression. Experimental studies are needed to confirm and further characterize the functional role(s) of the LIMP.

Keywords: Chlorophyta, Trebouxiophyceae, *Myrmecia*, Replication origin, Genome rearrangement, LIMP, HyLIMP, DNA cruciform, Palindromic repeat, Transcript processing

* Correspondence: nicolas.tourasse@ibpc.fr
[1]Institut de Biologie Physico-Chimique, UMR CNRS 7141 - Université Pierre et Marie Curie, Paris, France
[2]Institut de Biologie Physico-Chimique, FRC CNRS 550, Université Pierre et Marie Curie, Paris, France
Full list of author information is available at the end of the article

Background

Lobosphaera incisa (Reisigl) comb. nov. is a unicellular green alga belonging to the class Trebouxiophyceae (phylum Chlorophyta), which includes coccoid or pseudo-filamentous species from subaerial, soil, or freshwater habitats, and lichen photobionts. *L. incisa* was originally assigned to the genus *Myrmecia*, but subsequently reclassified as *Parietochloris* and then as *Lobosphaera* [1–3]. Based on chloroplast genome sequences, *L. incisa* was recently placed in clade C, the most derived of the Core Trebouxiophyceae [4]. *L. incisa* is an oleaginous alga and a target organism of high biotechnological interest because it is the richest known plant source of arachidonic acid, a pharmaceutically and nutraceutically valuable ω-6 long-chain polyunsaturated fatty acid that accumulates in considerable amounts when cells are cultivated under specific conditions such as nitrogen starvation [5, 6]. Recently, a nuclear transformation system has been developed for *L. incisa* [7] and the complete sequence of its chloroplast genome has been released [8]. Here we report the complete sequence of the mitochondrial genome of *L. incisa* determined from high-throughput Illumina short-read sequencing. We compare it with the nine other mitochondrial genomes reported for Trebouxiophyceae. We focus on the analysis of the main feature of the genome, a novel type of palindromic repeat sequence.

Results

Feature content and organization of the L. incisa *mitochondrial genome*

Starting with scaffolds and contigs built from paired-end (PE) and mate-pair (MP) sequencing libraries using several assembly programs run with various parameter settings, the assembly of the mitochondrial genome of *L. incisa* was manually closed into a single, continuous, circular-mapping sequence of 69,997 bp (Fig. 1). The assembly presented here is based on ~5000X sequencing coverage (3,889,871 mapped reads) and therefore represents a high-quality and high-confidence sequence. It is the most massively consistent with the high-throughput sequencing reads, but mapping of the reads revealed three loci, all intergenic, with noticeable heterogeneity. They correspond to short deletions (173, 110 and 57 nt; red arrows in Fig. 1) in a fraction of the molecules, due to short direct repeats (respectively 9, 12 and 11 nt long). In addition, minor changes in the length of short poly-A or poly-T tracts were observed at 68 positions.

The *L. incisa* mitochondrial genome exhibits a G + C content of ~36 %, and is predicted to encode 64 genes: 34 protein-coding genes (total coding capacity, 42 %), 3 rRNAs, and 27 tRNAs (Fig. 1), a gene repertoire similar to that found in the nine known Trebouxiophyceae mitochondrial genomes [9–15]. tRNAs are present for all amino acids and three of them are interrupted by group II introns (*trnC*-GCA, *trnF*-GAA and *trnK*-UUU). A self-splicing group I intron, possibly mobile as it encodes a LAGLIDADG homing endonuclease, is inserted in the 23S rRNA gene. The noncoding part of the intron shows similarity to introns found in the mitochondrial *rrnL* genes of *Chlorella variabilis* and *Trebouxia aggregata*. But its homing endonuclease belongs to subfamily 2 and rather resembles similar enzymes encoded in chloroplast *rrnL* introns of other Trebouxiophyceae, which suggests inter-organelle intron trafficking.

In the mitochondrial genome of *L. incisa*, the genes are distributed in two clusters read on opposite DNA strands, comprising 53 and 11 genes, respectively (Fig. 1). From this bipartite organization, one can presume that the genome is expressed in two transcriptional units produced from divergent promoters, as found in animal mitochondria and experimentally demonstrated in the Trebouxiophyceae *Prototheca wickerhamii* [16]. Between *cox1* and *trnC-GCA*, we found two divergent motifs (TATATAGAA and TTTATAGGA at positions 69264 and 69295, on the minus and plus strands, respectively) resembling one of the *P. wickerhamii* promoter sequences TATATAG**G**A (where the first nt of the transcript is underlined). In support of this assignment, RNA-Seq transcriptome coverage is almost nil between these positions and increases a few nucleotide downstream.

The *L. incisa* mitochondrial genome harbors seven families of various dispersed DNA repeat sequences, totaling 66 copies for 3.9 kb, i.e. 5.6 % of the genome size (Table 1). Repeats are defined here as sequences ≥30 nt in length found more than 4 times on the genome. All are located in intergenic regions. The average G + C content of the repeat families is similar to that of the whole mitochondrial genome (34–45 %). The dispersed repeat sequences are completely different from the repeats identified in the chloroplast genome of *L. incisa* [8] and are virtually absent from the *L. incisa* nuclear genome (at most two isolated hits for a given element, unpublished data). Comparison with the other mitochondrial genomes of Trebouxiophyceae reveals that all seven repeat families are unique to *L. incisa*. In addition, an array of 12 repetitions of the nonanucleotide GAGGGCTAC, also not found in other species, is located downstream of *rps12*.

A novel palindromic repeat, the LIMP

One of the repeats in the mitochondrial genome of *L. incisa* stands out as highly unusual, because it is both palindromic and internally repetitive. We termed it the LIMP (*L. Incisa M*itochondrial *P*alindrome). Nineteen copies are found in the genome (Table 2). Their structure can be summarized by the following pattern: 5'-$(AACCA)_m$[AATGAAA or TTTCATT] $(TGGTT)_n$-3',

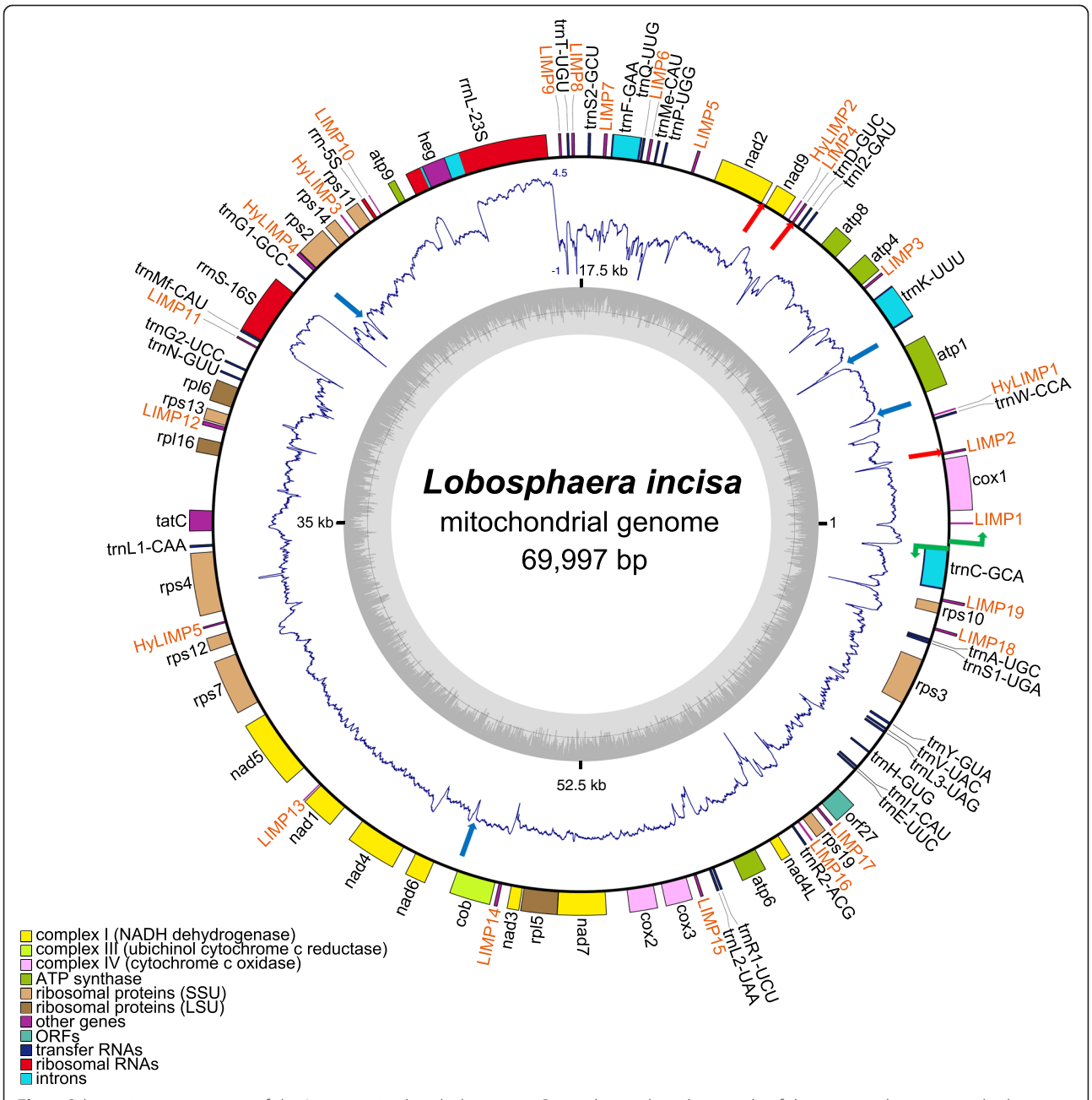

Fig. 1 Schematic representation of the *L. incisa* mitochondrial genome. Genes depicted on the outside of the outer circle are transcribed in counterclockwise orientation, on the inside in clockwise orientation. Putative promoters are shown as green bent arrows. Genes are color-coded according to functional classes, as listed in the bottom left corner. The LIMP and HyLIMP copies are shown in orange. The inner ring displays a graph of the G + C content (the circle marks the 50 % threshold), the middle one the RNA-Seq coverage ($\log_{10}$ scale, with 0 values brought to −1). Red arrows indicate the locations of the three short regions that are deleted in a fraction of the molecules, and blue arrows indicate the locations of the four palindromes (not LIMPs) of high predicted stability that are associated with a drop or loss of RNA-Seq coverage. The figure was generated using OrganellarGenomeDRAW [74]. The coverage plot was drawn using the "polar.plot" function from the plotrix package in R

where $2 \le (m,n) \le 13$ (Table 2 and Fig. 2a). Each LIMP is by itself repetitive, as it comprises a pentanucleotide motif (AACCA) that is repeated a variable number of times, and palindromic as the complementary motif TGGTT is repeated downstream, after a short loop. Hereafter, AACCA and TGGTT pentanucleotides will be referred to as "A-units" and "T-units", respectively. The number of A- and T-units for a given LIMP can differ. The two branches of the stem are separated by a conserved 7-nt loop that can be found in either of two orientations: AATGAAA or

Table 1 Families of interspersed DNA repeats identified in the *L. incisa* mitochondrial genome

Repeat name	# of copies	Consensus sequence[a]	Notes
LIMP	19	$(\underline{AACCA})_m$[AATGAAA or TTTCATT]$(\underline{TGGTT})_n$	Repetitive palindrome; $2 \leq (m,n) \leq 13$
Repeat_2	16	GCCTGTACAAATCTCTGCCCAACCGTAATGAAATGGTTGGCAAAAGAA AAAGAAATGGTGAGAGTAATCAAATGGTTGGCTC	Three copies interrupted by a LIMP; four copies next to LIMPs
Repeat_3	6	CCAGTTAAAATGAATGGCAAAAAACAAATGGTTGG	
Repeat_4	8	CTCCACACCATTTCATTAATCTCTGATTTGTTCA	
Repeat_5	7	TTTGGTTTGGTTTGGTTACAAATCAGAGAAAAGCAGGGGCTC	Six copies overlapping LIMPs
Repeat_6	5	TTACGGTTGGGCAGAGAAAAAAGGCAACCGTGAAAAAAAAAGCTGCGGT	three copies overlapping LIMPs
Repeat_7	5	AATCTCTGAT$\underline{TTGTTCAGGAACAA}$CTGGTTGGG	All copies adjacent to LIMPs

[a]palindromic positions are underlined

TTTCATT (reading on the plus strand). The first orientation will hereafter be referred to as +, the second one as -. LIMP copies are always intergenic, and found in both the clockwise and the counterclockwise gene clusters, with + and - orientations about equally represented (Table 2, Fig.2a). On the RNA, the LIMP sequence will always read 5'-AACCA…UGGUU-3', while the loop will read either AAUGAAA or UUUCAUU. While the loop sequence is invariant, variants of the pentanucleotide can be found just next to some LIMPs (A*G*CCA and *CGGTG* flanking LIMP #2; *GT*CCA *GA*CCA left of LIMPs #5, 8, 9 and 15, which leads to an increase in the length of the palindrome and in some cases makes the LIMP almost symmetrical; see Fig. 2a). As another evidence for mutational decay of the LIMP ends, 79 LIMP remnants were also found in the genome, containing a perfect or near-perfect 11–20 nt

Table 2 Features of LIMP repeats identified in the *L. incisa* mitochondrial genome

LIMP #	Center[a]	Structure[b]	Total length (nt)	Stem length (bp)	Palindromic region[c] (nt)	Imperfect genomic reads[d]	ΔG RNA hairpin (kcal.mol^{-1})[e]	transcriptome coverage[f]
1	33	5;+;4	52	20	52	0;6	−30.60	low (19)
2	2152	8;+;6	77	30	77	18;5	−49.70	NO (2)
3	7742	7;+;8	82	35	82	0;11	−63.00	NO (1)
4	10697	8;-;7	82	35	82	0;6	−61.20	NO (0)
5	14093	8;-;9	92	40	96	9;7	−77.60	NO (1)
6	15481	7;+;10	92	35	92	1;2	−66.60	NO (1)
7	16799	10;-;9	102	45	102	4;12	−80.00	NO (1)
8	17732	8;+;9	92	40	96	20;5	−76.80	NO (0)
9	18115	7;-;8	82	35	86	12;0	−68.20	NO (0)
10	23885	3;-;2	32	10	74	1;0	−55.20	normal (145)
11	29424	3;-;9	67	15	67	1;6	−27.40	normal (47)
12	32075	11;+;13	127	55	127	27;12	−101.70	NO (1)
13	43731	3;+;3	37	15	37	4;0	−21.70	normal (83)
14	50059	12;-;8	107	40	107	1;7	−70.40	NO (5)
15	55976	4;+;13	92	20	96	8;10	−48.00	low (13)
16	59523	5;-;5	57	25	57	0;0	−41.30	low (18)
17	60233	8;-;7	82	35	82	0;1	−61.20	NO (1)
18	66788	7;-;9	87	35	87	4;9	−63.10	NO (1)
19	67694	9;+;8	92	40	92	14;6	−68.20	NO (1)
average:			80.7	31.8	83.7		−59.6	
total:			1533	605	1591	124;105		

[a]position of first nt of first T-unit
[b]number of A-units; type of loop; number of T-units
[c]including additional nucleotides extending the stem, see Fig. 2a
[d]reads showing a change in the number of A-units; of T-units
[e]including extension of palindrome for #5, 8, 9 and 10
[f]coverage level (number of reads at position of lowest coverage; 5 or below is considered an interruption)

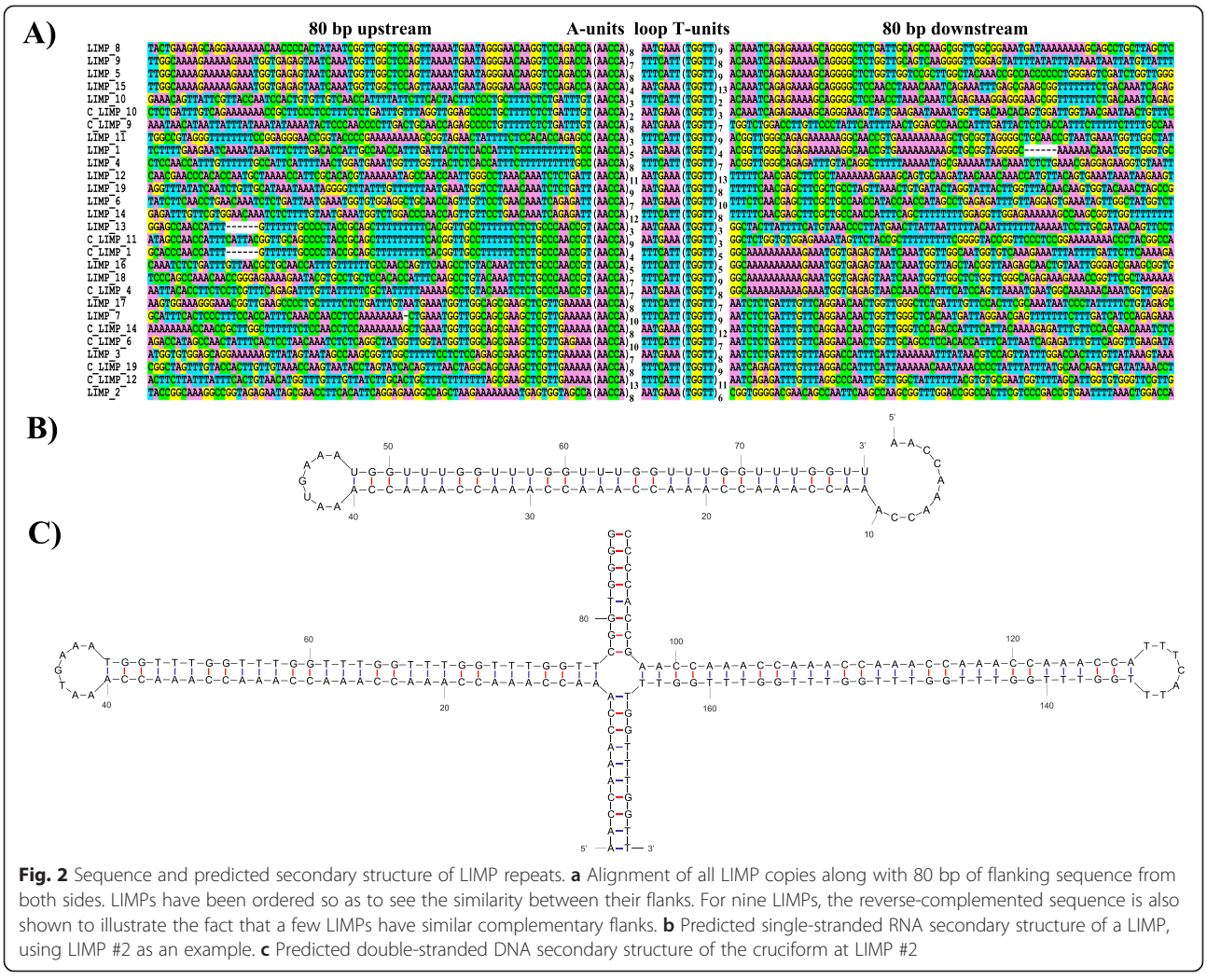

Fig. 2 Sequence and predicted secondary structure of LIMP repeats. **a** Alignment of all LIMP copies along with 80 bp of flanking sequence from both sides. LIMPs have been ordered so as to see the similarity between their flanks. For nine LIMPs, the reverse-complemented sequence is also shown to illustrate the fact that a few LIMPs have similar complementary flanks. **b** Predicted single-stranded RNA secondary structure of a LIMP, using LIMP #2 as an example. **c** Predicted double-stranded DNA secondary structure of the cruciform at LIMP #2

fragment of the LIMP (including the loop) but unable to form a stem-loop. These remnants often reside near *bona fide* LIMPs or are associated with other repeat sequences.

A striking feature of the LIMPs is the variable number of A- and T-units, which make the LIMP sequences vary from 32 to 127 bp in length and form asymmetric palindromes in 17 of the 19 cases (Table 2). In search for evidence of variability in the number of repeats, we extracted the genomic reads mapping to the LIMPs in an imperfect manner (i.e., with indels or with soft-clipping at one of the ends). When examined using the genome browser IGV [17, 18], only 229 reads were found to unambiguously indicate a number of repeat units different from that in the reference genome assembly (Table 2). Compared to the 62,478 reads that mapped to the LIMPs, this indicates a very low degree of variability within the population of DNA molecules sequenced. In particular, this analysis confirms that most LIMP stems in the genome are indeed asymmetrical, i.e., with different numbers of A- and T-units.

The propensity of the LIMPs to form hairpins as RNA and cruciforms as DNA, as depicted on Fig. 2b and c, respectively, was evaluated using mfold [19], a program that calculates the free energy (ΔG) of a DNA or RNA secondary structure. For the RNA hairpin, ΔG ranged between −101.7 and −21.7 kcal/mol (Table 2), compatible with the formation of at least some of these structures *in vivo*. To evaluate the propensity of the DNA to form cruciforms at the LIMP loci, we compared the free energy of the linear double-stranded B-DNA with that of the second-best structure, a cruciform where each strand folds onto itself according to the palindromic nature of the LIMP. As expected, the cruciform structure was markedly less stable, with an average difference 23.3 +/− 0.2 kcal/mol. This indicates that a DNA cruciform can form only if negative super-helicity is imposed to the molecule (see Discussion).

LIMPs can be sites of genome rearrangements

Remarkably, all LIMPs share (virtually) identical flanking sequences with at least one other LIMP (Fig. 2a). The

regions of identity are short, covering between 15 and 70 bp immediately next to the LIMPs, except for LIMPs #5 and 15 that share 104 bp of left flank and for LIMPs #10 and 15 that share 114 bp of right flank. Some LIMPs share only one flank (e.g., LIMPs #1 and 11), others show identity on both sides (e.g., LIMPs #16 and 18), and there are LIMPs sharing reverse-complemented flanks, where the left flank of a LIMP corresponds to the complementary sequence of the right flank of another LIMP or vice-versa (e.g., LIMPs #11 and 13 or LIMPs #7 and 14). For LIMP #10, the shortest of all LIMPs, the right flank is the reverse-complement of the left flank, so that the palindrome in this case is extended by 21 bp on each side. In a couple of instances the identical flanks actually correspond to copies of intergenic dispersed sequences that are found at other positions on the genome (Table 1). For example, LIMPs #4, 16 and 18 (Fig. 2a) are inserted within copies of repeat_2 that is found 16 times in the genome.

This similarity among LIMP flanks most likely reflects the history of their duplication. At the same time, it predictably increases the probability of recombination between LIMPs. Indeed, one of the minor variants observed in our genomic reads (see above) was a deletion of LIMP #4, due to recombination between its last T-units and a short LIMP remnant found 110 nt upstream. To further test the hypothesis that LIMPs can be sites of recombination, we mined the genomic sequencing reads for the occurrence of DNA fragments connecting the flanks of different LIMPs. To this end, we counted the number of fragments uniquely mapping to the 722 possible combinations of left and right flanks from the 19 LIMPs (using 400 bp of flanking sequences or up to the next LIMP; Fig. 3). Read pairs connecting the flanks of LIMPs that are distant in our assembly were extremely rare, at most 14, compared to 2,100-6,300 for the original LIMPs (cells on the diagonal). This suggests that LIMPs can indeed mediate mitochondrial genome rearrangements, but that the vast majority of the molecules in our culture conform to our genome map. Examination of total read coverage shows some variability among interLIMPs (i.e., regions in-between two LIMPs; Fig. 4), suggesting copy-number variation for interLIMPs. But because there is no correlation with the rare recombination events described above, we propose that this copy number variation is due to a bias in replication (see Discussion) and not to recombination between LIMPs.

LIMPs correspond to sites of low RNA-Seq coverage

Because of their palindromic nature, LIMPs have the potential to form stable stem-loop structures when transcribed as an RNA (see Table 2 and Fig. 2b). This raises the question of whether they could exert a function in transcript processing, i.e. help process the putative primary transcripts generated from the divergent promoters into shorter RNAs. We mapped the *L. incisa* paired-end and single-end transcriptomic reads to the mitochondrial genome assembly and found an average coverage of ~760X (~130X if the highly represented rRNAs are excluded). This indicates that, owing to their A/T-richness, the mitochondrial rRNAs and mRNAs were efficiently retained during poly-A RNA purification, the first step of the RNA-Seq protocol. Overall, coverage by RNA-Seq reads was relatively constant, with only a few regions showing an interruption of coverage. It therefore appears that many of the RNA molecules still comprise several cistrons, i.e. that maturation of the two precursor transcripts derived from the divergent promoters is not very efficient. Owing to their size, the tRNAs cannot be sequenced by our RNA-Seq method, yet all tRNA loci showed normal or slightly diminished coverage, indicating that they were only partially excised from the precursor transcripts. While interruption of coverage was sometimes seen near the 3' end of certain tRNAs, it is clear that the "tRNA punctuation" mode of transcript processing described in mitochondria of animals and to some extent yeast and red algae, relying on precise excision of the tRNAs from the primary transcript [20–22], is not predominant in the mitochondrion of *L. incisa*. The introns in tRNA genes and even that in *rrnL* were highly covered by RNA-Seq data, with many reads spanning the junctions, indicating that these introns are not spliced from the primary transcript but after excision of the tRNA and rRNA precursors.

Interestingly, LIMPs showed an overall lower coverage (~33X) compared to the whole mitochondrial genome (Fig. 1). In particular, we found that 13 of the LIMPs were associated with a complete or almost complete loss of RNA-Seq coverage (Table 2). Only LIMPs #10, 11 and 13 showed a coverage comparable to that of the surrounding regions, while LIMPs #1, 15 and 16 showed decreased but significant coverage. Interestingly, there was a strong correlation with the number of pairable A- and T-units, i.e. all LIMPs with stems shorter than 20 nt showed no coverage interruption, while those with stems longer than 25 nt showed complete interruption (Table 2). A partial effect was observed with stems of 20 or 25 nt. The short LIMP #10, in spite of the additional stability conferred by the extended palindrome, did not interrupt RNA-Seq coverage.

Another family of repetitive palindromes, the HyLIMP

We systematically searched the mitochondrial genome for palindromes using the program Palindrome from the EMBOSS 6.4.0 package [23], requesting a stem longer than 15 bp and a loop shorter than 21 nt. Among the 27 palindromes found (excluding the LIMPs), four, those with the highest predicted stability ($\Delta G < -65$ kcal.mol^{-1}), were associated with a drop or loss of RNA-Seq coverage (blue

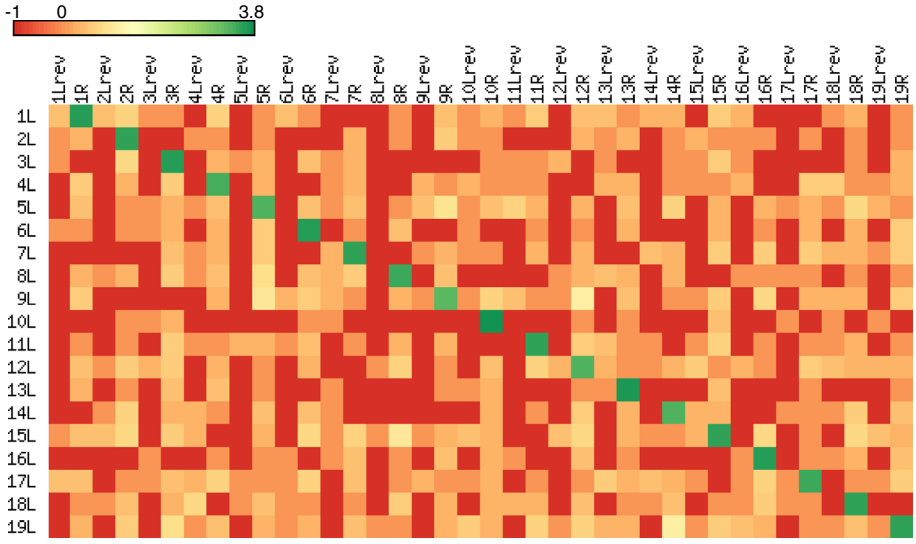

Fig. 3 Heat map of interLIMP joints, generated using matrix2png 1.2.2 [75]. The map gives a colored representation of the number of read pairs that could connect all possible combinations of flanks between the 19 LIMPs (L, left flank; Lrev, reverse-complemented left flank; R, right flank). Each cell in the map represents the $\log_{10}$-transformed number of read pairs for which one mate was located on each side of the LIMP flank combination considered. If there was no pair, the value was set to −1 (red)

arrows in Fig. 1). Three of them, unrelated to each other, were not internally repetitive and in two cases they obviously derive from the head-to-head assembly of two repeat elements (repeat_2 and repeat_4, respectively), probably brought together by recombination. But the fourth one (boxed sequence in Fig. 5) showed internal repetitivity in the stem, similar but different from that observed in the LIMPs. Its stem is made of three repeats of the 10-nt AACCA**GAGCC** sequence, i.e. a combination of the A-unit of a LIMP with a GAGCC pentanucleotide, prolonged by an unrelated decanucleotide. After a 3-nt loop, the complementary branch of the stem is found, with 5.5 repeats. This palindrome actually belongs to a small family of five related sequences (Fig. 5), all detected as palindromes by the criteria above. We call these repeats HyLIMPs, because of the hybrid nature of their repeat unit, partly similar to LIMPs. The other four HyLIMPs show the characteristic AACCAGAGCC-based palindrome, but they do not interrupt RNA-Seq coverage and their predicted stability is lower (-22 kcal.mol$^{-1} > \Delta G > -48$ kcal.mol^{-1}). Similar to LIMPs, HyLIMPs are often asymmetric, with one to three decanucleotide units in their stem. In HyLIMPs #1, 2 and 4, the non-repetitive inner part of the stem and the loop are conserved. Some sequence variability can be observed in the HyLIMP stem (italicized in Fig. 5), usually with complementary mutations in the other branch restoring Watson-Crick base-pairing. Like LIMPs, HyLIMPs show evidence for a dynamic evolution, with over 100 short sequences resembling HyLIMP #4 but not palindromic by our criteria, dispersed throughout the genome. As further proof of the relation between HyLIMPs and LIMPs, a HyLIMP decanucleotide-like sequence was found just left of LIMP #11 (see Fig. 2a), and the loop of HyLIMP #3 resembled that of the LIMPs (see Fig. 5). Fifty of these HyLIMP remnants were limited to the central part of the palindrome (in green in Fig. 5). Interestingly, this short palindrome is also found repeated 12 times in the chloroplast genome of *L. incisa* [8], and very similar repeats were found in the chloroplast genomes of three other Trebouxiophyceae of clade C, *Xylochloris irregularis*, *Leptosira terrestris* and *Dictyochloropsis reticulata*. No related sequence was found in the available Trebouxiophyceae mitochondrial genomes.

Discussion

Based on its relatively high percentage of coding regions (42 %), large gene repertoire (64 genes) and limited extent of repeats (5.6 %), the *L. incisa* mitochondrial genome can be classified as belonging to the "ancestral" type of algal mitochondrial DNA, as opposed to the "reduced-derived" pattern observed in *Pedinomonas minor* and Chlorophyceae [10]. Overall, the gene content is highly similar to that in the nine available mitochondrial genomes of Trebouxiophyceae [9–15]. In contrast to this conserved gene repertoire, highly diverged gene order and orientation have been observed among Trebouxiophyceae [10–16] including the present species. Even though there are remaining blocks of synteny shared between the mitochondrial genomes of Trebouxiophyceae, this diversity in organization highlights the numerous genome rearrangements that have occurred since the group originated. It also implies that the

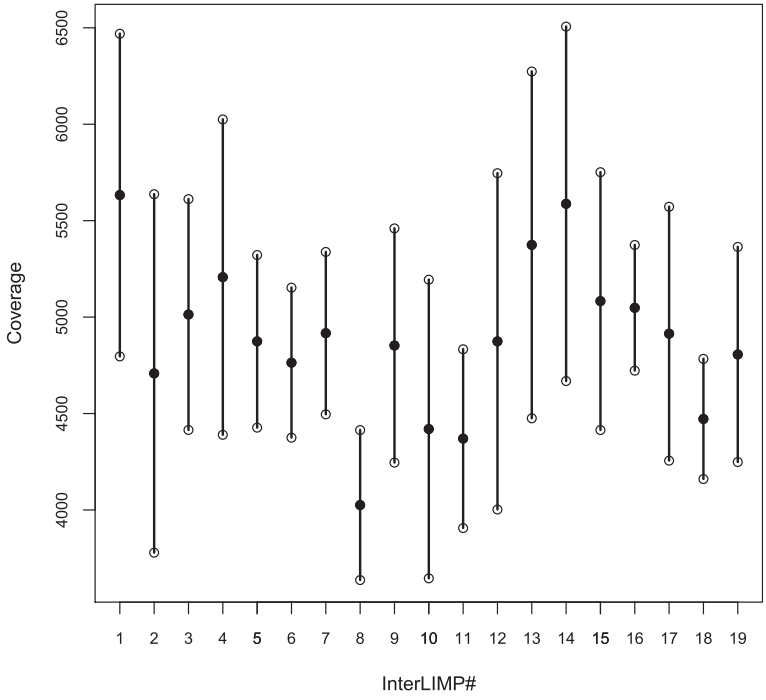

Fig. 4 Single read coverage of interLIMP regions. For each interLIMP, the filled dot indicates the average coverage, and the lower and upper open dots represent the values of average ± standard deviation. InterLIMP #*i* corresponds to the region in-between LIMPs #*i* and *i* + *1*, and interLIMP #19 is the region between LIMPs #19 and 1 (see Table 2)

mitochondrial genome is transcribed differently in the various species.

The most striking feature of the *L. incisa* mitochondrial genome is undoubtedly the presence of the LIMP. This repetitive palindromic repeat uses a fixed pentanucleotide ACCAA, repeated in direct and inverse orientation (A- and T-units) to form the two branches of a stem, with a fixed loop in-between. In the genome, LIMPs occur about equally in the two possible orientations, with no defined alternation pattern. No sequence similar to the LIMP can be found in any other organellar genome (chloroplast or mitochondrion) of Chlorophyta, and it therefore appears to have recently originated in the lineage leading to *L. incisa*. It is common for organellar genomes of algae to carry complex sets of short dispersed repeats (see e.g., [12, 24–31]), and each time they seem lineage- or even species-specific (in the case of *Volvox*). In some cases, repeat units can form extended tandem arrays [10, 32] or palindromes [26, 28, 33]. In *Volvox carteri*, both organelle genomes are bloated with repeats that form short (but non-repetitive) palindromes [26, 29]. Based on their large number and high sequence similarity, they have been described as selfish genetic elements, multiplying vigorously and spreading from the mitochondrial to the plastidial and even to the nuclear genome [26]. However, the stabilization of such structures in a polyploid genome where an efficient homologous recombination can rapidly correct the copy where the repeat would have duplicated, leads us to think that the repeats themselves confer a selective advantage to the cell, in other words that they have a function.

The two branches of the LIMP are rarely of equal length, implying that the stem formed by their annealing

#	START	ΔG (kcal.mol^{-1})	SEQUENCE
1	3316	-37.80	TTCCCAACCAGAGCCCCTTCAAATCAGAGATTTGAAGGGGCTCTGGTTGGCTTTTGAC
2	10816	-33.00	GTGCCAACCAGAGCCCCTGAAAATCAGAAATTTTCAGGGGCTCTGGTTGCTGT
3	24815	-15.30	GGGGTAAACCAACCATAATCAAACAGAGCCAACCATTTCATTATGGTTGGCTATGGTACGGGG
4	26423	-75.70	TTGCTGGTCCAACCATTCCTAACCAGAGCCAACCAGAGCCAACCAGAGCCGCTTGAAATCTCTGATTTCAAGCGGCTCTGGTTGGCTCTGGTTGGCTCTGGTTGGCTCTGGTTGGCTCTGGTTGACTGTGATC
5	37948	-39.00	GGTGCGACCAGAACCAACCAGAGCCAACCAGAGCCAACCAGAGTGGAATAGTTGGCTCTGGTTGGTTCTGGTTCAAGC

Fig. 5 HyLIMP sequences. HyLIMP #4 which interrupts RNA-Seq coverage is boxed. Palindromes are underlined, and the reported ΔG value is for the RNA form of the palindromic sequence. In the sequences, the repeat units common with LIMPs are written in red, the other pentanucleotide in blue and the conserved inner part of the palindrome in green. Residues differing from the consensus are italicized. The region in HyLIMP #3 that is identical to the LIMP loop is highlighted in yellow

needs not be extremely long to perform its function (it will only be as long as the shortest of the branches). Presumably, the repetitive nature of the palindrome allows expansion of the stem via DNA polymerase slippage, as is observed in microsatellites [34–36]. Indeed, detailed analysis of imperfectly matching reads shows that insertions and deletions of repeat units can occur, observable even within the culture we have analyzed. This could counterbalance stem shortening via mutational decay, attested by the imperfect A- and T-units next to LIMPs #2, 5, 8, 9 and 15 (Fig. 2a). The absolutely invariant sequence of the loop is another indication that LIMPs have a function. What then could be this molecular function? Two categories of hypotheses can be envisioned: LIMPs may act at the DNA level, imposing a local cruciform structure to the chromosome, or at the RNA level, by forming a stem-loop.

Palindromic sequences in DNA have the potential to form cruciform structures, with each strand folded as a stem-loop over the palindromic region. Because the energy of the linear form is lower for a molecule with unconstrained ends (in the case of the LIMPs, by about 23 kcal/mol), cruciforms are believed to form only if negative supercoiling is imposed on the molecule [37]. A higher degree of supercoiling favors the "folded" conformation, where the two stems stack onto the linear part of the chromosome, a conformation that is required for such processes as recombination and transposition. A folded cruciform indeed resembles a Holiday junction [38], and its "resolution" can lead to a double-stranded break and recombination [39, 40]. Another possible mechanism leading to double-stranded break is the formation of a single-stranded hairpin on the lagging strand during DNA replication [39]. Whatever the mechanism in the case of LIMPs, our finding that a small number of molecules in our preparation link the flank of different LIMPs (Fig. 3) indicates that recombination indeed occurs at these sites.

While this in principle could cause copy number variation among interLIMPs, by circularization and loss of the intervening sequence, we saw no correlation between the position of the main recombination events and the changes in sequence coverage along the genome (Fig 4). We therefore propose that these variations stem from another documented role of cruciforms, i.e. to serve as sites for initiation of DNA replication [41]. Such a role has been described for the rolling-circle replication of single-stranded viruses and plasmids [42–44], but also for eukaryotic nuclear DNA, where a protein of the 14-3-3 family stabilizes the cruciform at the replication origin [45]. Replication mechanisms in algal mitochondria have thus far received very little attention, and little can be inferred from comparison with other systems. In animal mitochondria, the heavy and light strands are replicated from distinct origins [46]. In land plants, mitochondrial DNA replication is probably very different because the DNA mostly occurs as linear branched molecules [47]. In Cryptophyte algae, it has been postulated that a palindrome found in a repetitive region is part of the replication origin [48], but no experimental evidence was provided. For *L. incisa* as well, we are tempted to speculate that DNA cruciforms formed at the LIMPs by negative supercoiling serve as origins of replication. This would explain, if these origins fire at different rates, the unequal representation of interLIMP regions in the population of DNA molecules sequenced. If replication is unidirectional as at animal mitochondrial origins, this direction could be determined by the orientation of the loop. During rolling circle replication, the unpaired displaced strand is expected to fold as a hairpin at the downstream LIMPs, which then could serve as a replication origin for the complementary strand.

Transcription is a source of local supercoiling *in vivo*, hence might contribute to the formation of the DNA cruciform. This could provide a link between transcriptional activity and DNA replication in the *L. incisa* mitochondrion. But when the LIMPs themselves are transcribed, the longest of them will form RNA stem-loops of decent stability (Table 2). These may be targets for post-transcriptional cleavage, which could explain why RNA-Seq coverage is interrupted. In *Prototheca*, processing sites on the two polycistronic pre-mRNAs have been mapped to the loop of short palindromic sequences [16]. Palindromic repeats have also been proposed to direct processing in *Chlamydomonas* mitochondria [27], as have the Repetitive Extragenic Palindromes (REP) elements found in many bacterial genomes [49]. RNase III-dependent RNA processing has been observed at the 26-27-nt long stem-loop structures formed by repetitive palindromic *nemis* elements in *Neisseria meningitidis* [50]. In the case of *L. incisa*, the fact that LIMPs with stems shorter than 21 nt seem to have little or no effect on RNA-Seq coverage, while longer ones efficiently interrupt it (Table 2), is compatible with an implication of RNase III: the dimeric bacterial enzyme is known to cleave only stems with more than two helical turns, i.e. longer than 20 nt [50, 51]. All 5 ribonuclease III homologs encoded in the nuclear genome of *L. incisa* contain N-terminal extensions compared to their bacterial homologs and are predicted to be directed to organelles (data not shown). Other endoribonucleases that are known to act at defined stem-loop or secondary structures [52, 53] could be involved as well in the processing of the precursor RNAs, as well as in their degradation. Incidentally, the fact that we observed unambiguous evidence for the addition of a 3' poly-A tail at a few specific locations (including rRNAs

and intergenic regions, data not shown) suggests that degradation of the *L. incisa* mitochondrial RNAs uses a system based on Poly(A)-Polymerase or PolyNucleotide Phosphorylase. Poly-A tailing, known to occur in chloroplast and mitochondria [54], has been proposed to operate in mitochondrial RNA degradation, in particular in *Chlamydomonas* [55].

With minor variations, most of the discussion above can also be applied to the HyLIMP. For example, only the HyLIMP with the highest stability is correlated with a drop in RNA-Seq coverage. The HyLIMP repeat is less well-defined and far less abundant than the LIMP, yet it shares with it a palindromic nature and internal repetitivity (at least for the longer of the HyLIMPs). In fact, the presence of the AACCA pentanucleotide in both is a strong indication for a partially common origin, as is the fact that HyLIMP #3 shows the same loop sequence as the LIMPs. The existence of many independent copies of the non-repetitive core of the HyLIMP (common to HyLIMPs #1, 2 and 4, see Fig. 5) in the *L. incisa* mitochondrial genome suggests that it preexisted and recruited the repeat units to extend its stem. Indeed, this core sequence is also present in the chloroplast genomes of *L. incisa* and two other clade C Trebouxiophyceae. Obviously, this central short palindromic repeat spread from the chloroplast to the mitochondrial genome, or vice-versa: in the absence of sequence for mitochondrial genomes of basal Clade C Trebouxiophyceae, it is not possible to ascertain the direction of the transfer. In the mitochondrion of *L. incisa*, HyLIMPs probably originated when the repeat co-opted for its elongation the repetitive unit of the LIMP along with another pentanucleotide. Remarkably, this recruitment respected the orientation found in the LIMP (A-units first). Whatever the phylogenetic path, it is probably not by chance that this core HyLIMP stem happens to be 10 nt long, and that it was extended by incorporating a 10-nt repeat. The pitch of a double-stranded RNA or DNA helix is about 10 nt, so that the repetitive part of LIMPs and HyLIMPs is predicted to show identical residues along a given generator of the helix. This property might be used by these repeats to perform their elusive molecular function, be it related to DNA or to RNA metabolism.

Conclusion

The mitochondrial genome of *L. incisa* encodes a unique type of repetitive palindromic DNA repeat sequence, the LIMP, and a related repeat, the HyLIMP. RNA sequencing suggests that the longest LIMPs and HyLIMPs are sites of transcript processing. Experimental studies are needed to confirm the functional role(s) of the LIMP and characterize the molecular mechanisms, and to identify the protein(s) that might interact with the LIMP.

Methods

Algal strain

The strain studied was an original isolate of *L. incisa* obtained from snow water patches in the alpine environment of Mt Tateyama, Japan [3] and deposited in the Göttingen University culture collection as SAG 2468. Algae were grown in modified BG11 medium at 25 °C and a light intensity of 130 μmol photons $m^{-2} \cdot s^{-1}$.

Genome sequencing and assembly

DNA was extracted following the CTAB DNA extraction protocol [56] with modifications. Whole-genome shotgun sequencing was performed using the Illumina short-read technology (HiSeq 2000 instrument) with PE (2 ×100 nt, insert size ~300 bp, performed by Tufts University core facility) and long jumping distance MP (2 × 100 nt, insert size ~8 kb, performed by MWG Eurofins) libraries. Reads were adapter- and quality-trimmed using cutadapt 1.1 (http://code.google.com/p/cutadapt/) and prinseq-lite 0.15 [57] with the following thresholds: min. read length 30 nt; min. base quality 20, min. average read quality 28. After preprocessing, a total of 189,672,359 reads remained for assembly. Several genome assemblies were computed using the de Bruijn graph-based assemblers SOAPdenovo 1.05 [58], Velvet 1.2.08 [59], and CLC Genomics Workbench 5.1.64 (CLC bio, Denmark; http://www.clcbio.com/) with k-mer values of 23, 35, 51, or 63 for each assembler. An assembly was also made with the super-read-based assembler MSR-CA (MaSuRCA) 1.9.3 [60] with an automatically set k-mer value of 31. In order to assemble specifically the mitochondrial genome, potentially mitochondrial sequences were identified by searching for similarity between the scaffolds and contigs of all assemblies and the complete sequences of 41 mitochondrial genomes from Chlorophyta available at the NCBI GenBank database. Searches were done both at the nucleotide level using BLASTN 2.2.25 [61] and the aminoacid level using BLASTX. A total of 1828 scaffolds/contigs were retained. All reads mapping to these selected scaffolds/contigs using SOAP 2.21 [62] were extracted, along with their mates, and then assembled using CLC with k-mer values of 23, 35, 51, or 63. Thirty-nine scaffolds/contigs generated in these four assemblies that were of length ≥100 nt and with a read coverage >1500, or that were of length ≥500 nt, were selected, after eliminating those coming from the *L. incisa* chloroplast genome [8]. A superassembly was then computed using the overlap-layout-consensus assembler in Geneious 5.5.2 (Biomatters Ltd., New Zealand; http://www.geneious.com/). To resolve the genome sequence, the superassembly and the original 39 scaffolds/contigs were compared using BLAT 35x1 [63] and manually joined based on the pattern of overlap, until a single continuous sequence could be reconstructed. This sequence

was iteratively refined by examining the mapping of genomic read pairs using BWA 0.6.2 [64] (options "-n 0.04 -o 0.02 -l 20"). In all analyses described above, alignment files in SAM and BAM format were sorted and converted using utilities from the samtools 0.1.18 [65] and bamtools 2.1.0 [66] packages.

Protein-coding genes were predicted using GeneMarkS 4.28 [67], rRNA genes were identified with RNAmmer 1.2 [68] and tRNA genes were predicted using tRNAscan-SE 1.21 [69]. RNAweasel [70] was used to find group I and group II introns, and repeated DNA sequences were identified using RepeatScout 1.0.5 [71] and RepeatMasker 3.3.0 [72]. The putative function of protein-coding genes was assigned based on aminoacid sequence similarity with homologs from proteins encoded by complete mitochondrial genomes of Chlorophyta using BLASTP searches. All annotations were manually reviewed. Open reading frames (ORFs) with no similarity to any sequence in NCBI GenBank were removed. The annotated sequence of the *L. incisa* mitochondrial genome has been deposited in the GenBank database under accession number [GenBank:KP902678] (BioProject PRJNA283614).

Transcriptome sequencing and analysis

The transcriptome of *L. incisa* was analyzed by high-throughput sequencing of cDNAs (RNA-Seq method) using the Illumina technology (HiSeq 1000 instrument). cDNAs were generated from mRNAs isolated under four growth conditions: exponential growth, 72 h nitrogen starvation under normal light (75 μmol photons $m^{-2} \cdot .s^{-1}$), as well as 12 h and 72 h nitrogen starvation under high light (150 μmol photons $m^{-2} \cdot .s^{-1}$). RNA was isolated from frozen samples using SV Total RNA isolation kit (Promega) after breaking the material with iron beads in liquid nitrogen. RNA quality was examined on a 2100 Electrophoresis Bioanalyzer. Illumina Truseq sequencing (50-nt single-end and 100-nt paired-end reads, insert size ~150 bp) was performed by the transcriptomics platform of the Institut de Biologie de l'École Normale Supérieure. For the purpose of this study, reads from all growth conditions were pooled, adapter-trimmed (using cutadapt 1.1), and mapped onto the mitochondrial genome sequence using GSNAP 2011-12-28 [73] (option "–max-mismatches = 0.04"). In total, 716,251 reads mapped to the mitochondrial assembly, of which 604,629 mapped to the rRNA genes.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NJT and OV performed the genome assembly, annotation and analysis and wrote the manuscript. NS and IKG carried out algal culture and DNA/RNA preparations. OV conceived the genome sequencing project. SB participated in study coordination. All authors read and approved the manuscript.

Acknowledgments

This work was supported by the European Commission's Seventh Framework Program for Research and Technology Development (FP7), project GIAVAP (Grant #266401), and by the French State "Initiative d'Excellence" program (Grant "DYNAMO", ANR-11-LABX-0011-01).

Author details

[1]Institut de Biologie Physico-Chimique, UMR CNRS 7141 - Université Pierre et Marie Curie, Paris, France. [2]Institut de Biologie Physico-Chimique, FRC CNRS 550, Université Pierre et Marie Curie, Paris, France. [3]Microalgal Biotechnology Laboratory, French Associates Institute for Agriculture and Biotechnology of Drylands, J. Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion 84990, Israel. [4]ARNA Laboratory, INSERM UMR 869, Université Bordeaux 2, Bordeaux, France.

Received: 26 February 2015 Accepted: 22 July 2015
Published online: 05 August 2015

References

1. Friedl T. Inferring taxonomic positions and testing genus level assignments in coccoid green lichen algae: a phylogenetic analysis of 18S ribosomal sequences from *Dictyochloropsis reticulata* and from members of the genus *Myrmecia* (Chlorophyta, Trebouxiophyceae cl. nov.). J Phycol. 1995;31:632–9.
2. Karsten U, Friedl T, Schumann R, Hoyer K, Lembcke S. Mycosporine-like amino acids and phylogenies in green algae: *Prasiola* and its relatives from the Trebouxiophyceae (Chlorophyta). J Phycol. 2005;41:557–66.
3. Watanabe S, Hirabayashi S, Boussiba S, Cohen Z, Vonshak A, Richmond A. *Parietochloris incisa* comb. nov. (Trebouxiophyceae, Chlorophyta). Phycol Res. 1996;44:107–8.
4. Lemieux C, Otis C, Turmel M. Chloroplast phylogenomic analysis resolves deep-level relationships within the green algal class Trebouxiophyceae. BMC Evol Biol. 2014;14:211.
5. Bigogno C, Khozin-Goldberg I, Boussiba S, Vonshak A, Cohen Z. Lipid and fatty acid composition of the green oleaginous alga *Parietochloris incisa*, the richest plant source of arachidonic acid. Phytochemistry. 2002;60(5):497–503.
6. Khozin-Goldberg I, Bigogno C, Shrestha P, Cohen Z. Nitrogen starvation induces the accumulation of arachidonic acid in the freshwater green alga *Parietochloris incisa* (Trebouxiophyceae). J Phycol. 2002;38:991–4.
7. Zorin B, Grundman O, Khozin-Goldberg I, Leu S, Shapira M, Kaye Y, et al. Development of a Nuclear Transformation System for Oleaginous Green Alga *Lobosphaera* (*Parietochloris*) *incisa* and Genetic Complementation of a Mutant Strain, Deficient in Arachidonic Acid Biosynthesis. PLoS One. 2014;9(8), e105223.
8. Tourasse NJ, Barbi T, Waterhouse JC, Shtaida N, Leu S, Boussiba S, Purton S, Vallon O. The complete sequence of the chloroplast genome of the green microalga *Lobosphaera* (*Parietochloris*) *incisa*. Mitochondrial DNA 2014, in press.
9. Wolff G, Plante I, Lang BF, Kuck U, Burger G. Complete sequence of the mitochondrial DNA of the chlorophyte alga *Prototheca wickerhamii*. Gene content and genome organization. J Mol Biol. 1994;237(1):75–86.
10. Turmel M, Lemieux C, Burger G, Lang BF, Otis C, Plante I, et al. The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor*. Two radically different evolutionary patterns within green algae. Plant Cell. 1999;11(9):1717–30.
11. Pombert JF, Keeling PJ. The mitochondrial genome of the entomoparasitic green alga *helicosporidium*. PLoS One. 2010;5(1), e8954.
12. Smith DR, Burki F, Yamada T, Grimwood J, Grigoriev IV, Van Etten JL, et al. The GC-rich mitochondrial and plastid genomes of the green alga *Coccomyxa* give insight into the evolution of organelle DNA nucleotide landscape. PLoS One. 2011;6(8), e23624.
13. Servin-Garcidueñas LE, Martinez-Romero E. Complete mitochondrial and plastid genomes of the green microalga *Trebouxiophyceae sp.* strain MX-AZ01 isolated from a highly acidic geothermal lake. Eukaryot Cell. 2012;11(11):1417–8.
14. Jeong H, Lim JM, Park J, Sim YM, Choi HG, Lee J, et al. Plastid and mitochondrion genomic sequences from Arctic *Chlorella sp.* ArM0029B. BMC Genomics. 2014;15:286.
15. Orsini M, Costelli C, Malavasi V, Cusano R, Concas A, Angius A, et al. Complete genome sequence of mitochondrial DNA (mtDNA) of *Chlorella sorokiniana*. Mitochondrial DNA. 2014;1–3.

16. Wolff G, Kuck U. Transcript mapping and processing of mitochondrial RNA in the chlorophyte alga *Prototheca wickerhamii*. Plant Mol Biol. 1996;30(3):577–95.
17. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.
18. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178–92.
19. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31(13):3406–15.
20. Ojala D, Montoya J, Attardi G. tRNA punctuation model of RNA processing in human mitochondria. Nature. 1981;290(5806):470–4.
21. Schafer B. RNA maturation in mitochondria of *S. cerevisiae* and *S. pombe*. Gene. 2005;354:80–5.
22. Richard O, Kloareg B, Boyen C. mRNA expression in mitochondria of the red alga *chondrus crispus* requires a unique RNA-processing mechanism, internal cleavage of upstream tRNAs at pyrimidine 48. J Mol Biol. 1999;288(4):579–84.
23. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends in Genet: TIG. 2000;16(6):276–7.
24. Maul JE, Lilly JW, Cui L, de Pamphilis CW, Miller W, Harris EH, et al. The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. Plant Cell. 2002;14(11):2659–79.
25. Pombert JF, Beauchamp P, Otis C, Lemieux C, Turmel M. The complete mitochondrial DNA sequence of the green alga *Oltmannsiellopsis viridis*: evolutionary trends of the mitochondrial genome in the Ulvophyceae. Curr Genet. 2006;50(2):137–47.
26. Smith DR, Lee RW. The mitochondrial and plastid genomes of *Volvox carteri*: bloated molecules rich in repetitive DNA. BMC Genomics. 2009;10:132.
27. Boer PH, Gray MW. Short dispersed repeats localized in spacer regions of *Chlamydomonas reinhardtii* mitochondrial DNA. Curr Genet. 1991;19(4):309–12.
28. Nedelcu AM, Lee RW. Short repetitive sequences in green algal mitochondrial genomes: potential roles in mitochondrial genome evolution. Mol Biol Evol. 1998;15(6):690–701.
29. Aono N, Shimizu T, Inoue T, Shiraishi H. Palindromic repetitive elements in the mitochondrial genome of *Volvox*. FEBS Lett. 2002;521(1–3):95–9.
30. Pombert JF, Lemieux C, Turmel M. The complete chloroplast DNA sequence of the green alga *Oltmannsiellopsis viridis* reveals a distinctive quadripartite architecture in the chloroplast genome of early diverging ulvophytes. BMC Biol. 2006;4:3.
31. Smith DR, Lee RW, Cushman JC, Magnuson JK, Tran D, Polle JE. The *Dunaliella salina* organelle genomes: large sequences, inflated with intronic and intergenic DNA. BMC Plant Biol. 2010;10:83.
32. Pombert JF, Otis C, Lemieux C, Turmel M. The complete mitochondrial DNA sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) highlights distinctive evolutionary trends in the chlorophyta and suggests a sister-group relationship between the Ulvophyceae and Chlorophyceae. Mol Biol Evol. 2004;21(5):922–35.
33. de Cambiaire JC, Otis C, Turmel M, Lemieux C. The chloroplast genome sequence of the green alga *Leptosira terrestris*: multiple losses of the inverted repeat and extensive genome rearrangements within the Trebouxiophyceae. BMC Genomics. 2007;8:213.
34. Eckert KA, Mowery A, Hile SE. Misalignment-mediated DNA polymerase β mutations: comparison of microsatellite and frame-shift error rates using a forward mutation assay. Biochemistry. 2002;41(33):10490–8.
35. Garcia-Diaz M, Kunkel TA. Mechanism of a genetic glissando: structural biology of indel mutations. Trends Biochem Sci. 2006;31(4):206–14.
36. Kunkel TA, Bebenek K. DNA replication fidelity. Annu Rev Biochem. 2000;69:497–529.
37. Mikheikin AL, Lushnikov AY, Lyubchenko YL. Effect of DNA supercoiling on the geometry of holliday junctions. Biochemistry. 2006;45(43): 12998–3006.
38. Shlyakhtenko LS, Potaman VN, Sinden RR, Lyubchenko YL. Structure and dynamics of supercoil-stabilized DNA cruciforms. J Mol Biol. 1998;280(1):61–72.
39. Casper AM, Greenwell PW, Tang W, Petes TD. Chromosome aberrations resulting from double-strand DNA breaks at a naturally occurring yeast fragile site composed of inverted ty elements are independent of Mre11p and Sae2p. Genetics. 2009;183(2):423–39. 421SI-426SI.
40. Kato T, Kurahashi H, Emanuel BS. Chromosomal translocations and palindromic AT-rich repeats. Curr Opin Genet Dev. 2012;22(3):221–8.
41. Brazda V, Laister RC, Jagelska EB, Arrowsmith C. Cruciform structures are a common DNA feature important for regulating biological processes. BMC Mol Biol. 2011;12:33.
42. Orozco BM, Hanley-Bowdoin L. A DNA structure is required for geminivirus replication origin function. J Virol. 1996;70(1):148–58.
43. Jin R, Novick RP. Role of the double-strand origin cruciform in pT181 replication. Plasmid. 2001;46(2):95–105.
44. Cheung AK. Palindrome regeneration by template strand-switching mechanism at the origin of DNA replication of porcine circovirus via the rolling-circle melting-pot replication model. J Virol. 2004;78(17):9016–29.
45. Zannis-Hadjopoulos M, Yahyaoui W, Callejo M. 14-3-3 cruciform-binding proteins as regulators of eukaryotic DNA replication. Trends Biochem Sci. 2008;33(1):44–50.
46. McKinney EA, Oliveira MT. Replicating animal mitochondrial DNA. Genet Molecular Biology. 2013;36(3):308–15.
47. Cupp JD, Nielsen BL. Minireview: DNA replication in plant mitochondria. Mitochondrion. 2014;19PB:231–7.
48. Kim E, Lane CE, Curtis BA, Kozera C, Bowman S, Archibald JM. Complete sequence and analysis of the mitochondrial genome of *Hemiselmis andersenii* CCMP644 (Cryptophyceae). BMC Genomics. 2008;9:215.
49. Di Nocera PP, De Gregorio E, Rocco F. GTAG- and CGTC-tagged palindromic DNA repeats in prokaryotes. BMC Genomics. 2013;14:522.
50. De Gregorio E, Abrescia C, Carlomagno MS, Di Nocera PP. Ribonuclease III-mediated processing of specific *Neisseria meningitidis* mRNAs. Biochem J. 2003;374(Pt 3):799–805.
51. Court DL, Gan J, Liang YH, Shaw GX, Tropea JE, Costantino N, et al. RNase III: Genetics and function; structure and mechanism. Annu Rev Genet. 2013;47:405–31.
52. Nicholson AW. Function, mechanism and regulation of bacterial ribonucleases. FEMS Microbiol Rev. 1999;23(3):371–90.
53. Tomecki R, Dziembowski A. Novel endoribonucleases as central players in various pathways of eukaryotic RNA metabolism. RNA. 2010;16(9):1692–724.
54. Schuster G, Stern D. RNA polyadenylation and decay in mitochondria and chloroplasts. Prog Mol Biol Transl Sci. 2009;85:393–422.
55. Zimmer SL, Schein A, Zipor G, Stern DB, Schuster G. Polyadenylation in *Arabidopsis* and *Chlamydomonas* organelles: the input of nucleotidyltransferases, poly(A) polymerases and polynucleotide phosphorylase. Plant J. 2009;59(1):88–99.
56. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 1987;19:11–5.
57. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27(6):863–4.
58. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010;20(2):265–72.
59. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821–9.
60. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29(21):2669–77.
61. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
62. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25(15):1966–7.
63. Kent WJ. BLAT–the BLAST-like alignment tool. Genome Res. 2002;12(4):656–64.
64. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
65. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
66. Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011;27(12):1691–2.
67. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 2001;29(12):2607–18.
68. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35(9):3100–8.

69. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5):955–64.
70. Lang BF, Laforest MJ, Burger G. Mitochondrial introns: a critical view. Trends Genet: TIG. 2007;23(3):119–25.
71. Price AL, Jones NC, Pevzner PA. *De novo* identification of repeat families in large genomes. Bioinformatics. 2005;21 Suppl 1:i351–8.
72. Tempel S. Using and understanding RepeatMasker. Methods Mol Biol. 2012;859:29–51.
73. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26(7):873–81.
74. Lohse M, Drechsel O, Kahlau S, Bock R. OrganellarGenomeDRAW–a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. 2013;41(Web Server issue):W575–81.
75. Pavlidis P, Noble WS. Matrix2png: a utility for visualizing matrix data. Bioinformatics. 2003;19(2):295–6.