A new lineage of non-photosynthetic green algae with extreme organellar genomes

Tomáš Pánek[1,2], Dovilė Barcytė[1], Sebastian C. Treitli[3], Kristína Záhonová[1], Martin Sokol[1], Tereza Ševčíková[1], Eliška Zadrobílková[2], Karin Jaške[1], Naoji Yubuki[2], Ivan Čepička[2], Marek Eliáš[1,*]

[1]Department of Biology and Ecology, Faculty of Science, University of Ostrava, 701 00 Ostrava, Czech Republic

[2]Department of Zoology, Faculty of Science, Charles University, 128 43 Prague, Czech Republic

[3]Department of Parasitology, Faculty of Science, Charles University, BIOCEV, 252 42 Vestec, Czech Republic

Correspondence: marek.elias@osu.cz

Abstract

Background: The plastid genomes of the green algal order Chlamydomonadales tend to expand their non-coding regions, but this phenomenon is poorly understood. Here we shed new light on organellar genome evolution in Chlamydomonadales by studying a previously unknown non-photosynthetic lineage. We established cultures of two new *Polytoma*-like flagellates, defined their basic characteristics and phylogenetic position, and obtained complete organellar genome sequences and a transcriptome assembly for one of them.

Results: We discovered a novel deeply diverged chlamydomonadalean lineage that has no close photosynthetic relatives and represents an independent case of photosynthesis loss. To accommodate these organisms, we establish a new genus, *Leontynka*, with two species *L. pallida* and *L. elongata* distinguished by both morphological and molecular characteristics. Notable features of the colourless plastid of *L. pallida* deduced from the plastid genome (plastome) sequence and transcriptome assembly include the retention of ATP synthase, thylakoid-associated proteins, carotenoid biosynthesis pathway, and plastoquinone-based electron transport chain, the latter two modules having an obvious functional link to the eyespot present in *Leontynka*. Most strikingly, the *L. pallida* plastome with its ~362 kbp is by far the largest among non-photosynthetic eukaryotes investigated to date. Instead of a high gene content, its size reflects extreme proliferation of sequence repeats. These are present also in coding sequences, with one repeat type found in exons of 11 out of 34 protein-coding genes and up to 36 copies per gene, affecting thus the encoded proteins. The mitochondrial genome of *L. pallida* is likewise exceptionally large, with its >104 kbp surpassed only by the mitogenome of *Haematococcus lacustris* among all members of Chlamydomonadales studied so far. It is also bloated with repeats, yet completely different from those in the *L. pallida* plastome, which contrasts with the situation in *H. lacustris* where both

organellar genomes have accumulated related repeats. Furthermore, the *L. pallida* mitogenome exhibits an extremely high GC content in both coding and non-coding regions and, strikingly, a high number of predicted G-quadruplexes.

Conclusions: With the unprecedented combination of plastid and mitochondrial genome characteristics, *Leontynka* pushes the frontiers of organellar genome diversity and becomes an interesting model for studying organellar genome evolution.

Keywords: Chlamydomonadales; GC content; G-quadruplex; green algae; mitochondrial genome; non-photosynthetic algae; plastid genome; repeat expansion

Background

Secondary loss of photosynthesis has occurred numerous times across the diversity of plastid-bearing eukaryotes, including land plants (Hadariová *et al.*, 2018; Sibbald & Archibald, 2020). Among algae, photosynthesis loss has been most common among groups characterised by secondary or higher-order plastids, with chrysophytes and myzozoans (including apicomplexans as the best-studied non-photosynthetic "algae") being the most prominent examples. In green algae, loss of photosynthesis is restricted to several lineages within two classes, Trebouxiophyceae and Chlorophyceae (Figueroa-Martinez *et al.*, 2015). Colourless trebouxiophytes are formally classified in two genera, *Helicosporidium* and the polyphyletic *Prototheca*, collectively representing three independent photosynthesis loss events (Suzuki *et al.*, 2018). While these organisms live as facultative or obligate parasites of metazoans (including humans), non-photosynthetic members of Chlorophyceae are all free-living osmotrophic flagellates. Two genera of such colourless flagellates have been more extensively studied and are represented in DNA sequence databases: the biflagellate *Polytoma* and the tetraflagellate *Polytomella*. They both fall

within the order Chlamydomonadales (Volvocales *sensu lato*), but are not closely related to each other. Furthermore, *Polytoma* as presently circumscribed is polyphyletic, since *P. oviforme* does not group with its congeners, including the type species *P. uvella* (Figueroa-Martinez *et al.*, 2015). Hence, photosynthesis was lost at least three times in Chlamydomonadales, but the real number is probably higher, since several other genera of colourless flagellates morphologically falling within this group were historically described (Ettl, 1983), but remain to be studied by modern methods. Indeed, a taxonomically unidentified non-photosynthetic chlamydomonadalean (strain NrCl902), not related to any of the three known lineages, was reported recently (Kayama *et al.*, 2020); whether it corresponds to any of the previously formally described taxa is yet to be investigated.

The non-photosynthetic chlamydomonadaleans are not only diverse phylogenetically, but they also exhibit diversity in the features of their residual plastids. Most notably, *Polytomella* represents one of the few known cases of a complete loss of the plastid genome (plastome) in a plastid-bearing eukaryote (Smith & Lee, 2014). In contrast, *Polytoma uvella* harbours the largest plastome amongst all non-photosynthetic eukaryotes studied to date (≥230 kbp). This is not due to preserving a large number of genes, but because of the massive accumulation of long arrays of short repeats in intergenic regions (Figueroa-Martinez *et al.*, 2017). The unusual architecture of the *P. uvella* plastome seems to reflect a more general trend of plastome evolution in Chlamydomonadales, i.e. a tendency to increase in size by the expansion of repetitive sequences. An extreme manifestation of this trend was recently unveiled by sequencing the 1.35-Mbp plastome of the photosynthetic species *Haematococcus lacustris*, a record-holder amongst all fully sequenced plastomes to date (Bauman *et al.*, 2018; Smith, 2018). Interestingly, *H. lacustris* also harbours the by far largest known mitochondrial genome (mitogenome) amongst all Chlamydomonadales, which has expanded to 126.4 kbp by the accumulation of repeats highly

similar to those found in the plastome, suggesting an inter-organellar transfer of the repeats (Zhang *et al.*, 2019). The mechanistic underpinnings of the repeat accumulation in chlamydomonadalean organellar genomes are still not clear.

When studying protists living in hypoxic sediments, we obtained cultures of two colourless flagellates that turned out to represent a novel, deeply separated lineage of Chlamydomonadales. We here describe them formally as two species in a new genus. Using a combination of different DNA sequencing technologies, we determined sequences of organellar genomes of one of the isolates, which turned out to exhibit extreme features concerning the size and/or composition. Our analysis of these genomes provides important new insights into the evolution of organelle genomes in general.

Results

A new lineage of non-photosynthetic Chlamydomonadales with two species

Based on their 18S rRNA gene sequences, the two new isolates – AMAZONIE and MBURUCU – constitute a clade (with full bootstrap support) that is nested within Chlamydomonadales, but separate from all the principal chlamydomonadalean clades as demarcated by Nakada *et al.* (2008) (Fig. 1a). Notably, this new lineage is clearly unrelated to all previously studied non-photosynthetic chlamydomonadaleans, including *Polytomella* (branching off within the clade *Reinhardtinia*), both lineages representing the polyphyletic genus *Polytoma* (*P. uvella* plus several other species in the clade *Caudivolvoxa* and *P. oviforme* in the clade *Xenovolvoxa*), and the strain NrCl902 (also in *Caudivolvoxa*; Additional file 1: Fig. S1). Our two strains were mutually separated in the 18S rRNA gene tree as deeply as other chlamydomonadalean pairs classified as separate species or even genera, and their 18S rRNA gene sequences differed in 13 positions (out

of 1703 available for comparison). In addition, the ITS1-5.8S-ITS2 rDNA regions of the two strains exhibited only 88% identity and the differences included several compensatory base changes (CBCs) in the helix II of the characteristic secondary structure of the ITS2 region (Additional file 1: Fig. S2). This and morphology-based evidence presented below led us to conclude that the two strains represent two different species of a new genus of chlamydomonadalean algae, which we propose be called *Leontynka pallida* (strain AMAZONIE) and *Leontynka elongata* (strain MBURUCU). Formal descriptions of the new taxa are provided in Additional file 2: Note S1.

The phylogenetic position of *L. pallida* was also studied by using protein sequences encoded by its plastome (see below). Phylogenomic analysis of a concatenated dataset of 24 conserved proteins encoded by plastomes of diverse members of Chlorophyceae, including a comprehensive sample of available data from Chlamydomonadales, revealed *L. pallida* as a separate lineage potentially sister to a fully supported broader clade comprising representatives of the clades *Caudivolvoxa* and *Xenovolvoxa* (*sensu* Nakada *et al.*, 2008; Fig. 1b). This position of *L. pallida* received moderate support in the maximum likelihood analysis (nonparametric bootstrap value of 78%), but inconclusive support from the PhyloBayes analysis (posterior probability of 0.68). Importantly, *L. pallida* was unrelated to *Polytoma uvella* (nested within *Caudivolvoxa* with full support). *Polytoma oviforme* and the genus *Polytomella* were missing from the analysis due to lack of plastome data or the complete absence of the plastome, respectively.

Both *Leontynk*a species lacked a green plastid (chloroplast). Instead, their cells were occupied by a colourless leucoplast containing starch grains, typically filling most of its volume (Fig. 2, Additional file 1: Figs S4 and S5). Two anterior, isokont flagella approximately as long as the cell body emerged from a keel-shaped papilla (Additional file 1: Fig. S3c, d). Cells of both

139 species also contained two apical contractile vacuoles (Fig. 2c, f, Additional file 1: Fig. S3a, c,
140 Fig. S4a, d, h), a central or slightly posterior nucleus (Additional file 1: Fig. S3c, Fig. S4h),
141 inclusions of yellowish lipid droplets (Fig. 2h, Additional file 1: Fig. S4h), and one or occasionally
142 two eyespots (Fig. 2a, b, and f–j, Additional file 1: Fig. S3a, b, e, g, h, Fig. S4a, c–f, h).
143 Reproduction occurred asexually through zoospore formation, typically with up to four zoospores
144 formed per the mother cell (Additional file 1: Fig. S3g, h). The two species differed in the cell
145 shape and position of the eyespot, as described in more detail in Additional file 2: Note S2.

146 The plastid of both species was bounded by a double membrane and composed of
147 numerous separate compartments connected by narrow "bridges" (Fig. 2d, e, i, j, Additional file
148 1: Fig. S5a-c, h). Each compartment contained either a single large or two smaller starch grains,
149 leaving essentially no room for the stroma or thylakoids. Rarely, starch-free compartments
150 containing membranous inclusions were present (Fig. 2e). The eyespot globules were inside the
151 plastid and were associated with structures that we interpret as thylakoids (Fig. 2j). Mitochondria
152 were highly abundant and contained numerous cristae (Additional file 1: Fig. S5e, f, i). It was
153 impossible to unambiguously determine the crista morphology in *L. pallida* (Additional file 1: Fig.
154 S5i), but in *L. elongata*, the cristae were of the discoidal morphotype (Additional file 1: Fig. S5e).
155 Further details on the ultrastructure of *Leontynka* spp. are presented in Additional file 2: Note S2.
156

157 **The extremely bloated plastome of *Leontynka pallida***

158 A complete plastome sequence was assembled for *L. pallida* using a combination of Oxford
159 Nanopore and Illumina reads. It corresponds to a circular-mapping molecule comprising 362307
160 bp (Fig. 3a). Thirty-four protein-coding genes (including two intronic ORFs), 26 tRNA genes (a
161 standard set presumably allowing for translation of all sense codons), and genes for the three

standard rRNAs were identified and annotated in the genome. Three genes are interrupted by introns: *atpA* with one group I intron that contains an ORF encoding a LAGLIDADG homing endonuclease, *tufA* with one group II intron that contains an ORF encoding a reverse transcriptase/maturase protein, and *rnl* with one group I and one group II intron, neither containing an ORF. No putative pseudogenes or apparent gene remnants were identified in the *L. pallida* plastome.

No genes that encode proteins directly associated with photosynthetic electron transport components and CO_2 fixation were identified in the *L. pallida* plastome. The genes retained encode proteins involved in transcription (RNA polymerase subunits), translation (*tufA* and ribosomal subunit genes), protein turnover (*clpP*, *ftsH*), and a protein of an unclear function (*ycf1*). Nearly all these genes have been preserved also in the plastome of *P. uvella* (Figueroa-Martinez *et al.*, 2017), except for *rps2*. The two non-photosynthetic *Leontynka* species share the absence of genes for two ribosomal proteins: *rpl32*, which is, however, also missing from a subset of photosynthetic representatives of Chlamydomonadales, and *rpl23* conserved in plastomes of all photosynthetic chlorophytes investigated to date (Turmel & Lemieux, 2018). Whereas an Rpl32 protein with a predicted plastid-targeting presequence is encoded by the *L. pallida* nuclear genome (Additional file 3: Table S1), the loss of *rpl23* does not seem to be compensated in a similar way. Interestingly, *rpl23* has been independently lost also from the plastome of *Helicosporidium* sp. (Figueroa-Martinez *et al.*, 2017), which suggests that this ribosomal subunit may become dispensable upon the loss of photosynthesis. In contrast to *P. uvella*, the *L. pallida* plastome has kept the same set of genes encoding ATP synthase subunits as typical for photosynthetic green algae, i.e., *atpA*, *atpB*, *atpE*, *atpF*, *atpH*, and *atpI*. As evidenced by the transcriptome data, the three missing ATP

184 synthase subunits (AtpC, AtpD, and AtpG) are encoded by the *L. pallida* nuclear genome and bear
185 predicted plastid-targeting signals (Additional file 3: Table S1).

186 What makes the *L. pallida* plastome truly peculiar are the intergenic regions. Their average
187 length is 4.7 kbp, which is 1.5 and five times more than the average length of the intergenic regions
188 in the plastomes of *P. uvella* and *C. reinhardtii*, respectively (Table 1). Furthermore, while the GC
189 content of the *P. uvella* intergenic regions is vastly different from that of coding regions (19%
190 versus 40%), the GC content of these two plastome partitions are highly similar in *L. pallida* (as
191 well as in *C. reinhardtii*; Table 1). A self-similarity plot generated for the *L. pallida* plastome
192 revealed a massive repetitiveness of the DNA sequence, with only short islands of unique
193 sequences scattered in the sea of repeats (Fig. 3b). The repeats are highly organised and occur in
194 various arrangements: tandem repeats, interspersed repeats, inverted repeats (palindromes), and
195 other higher-order composite repeated units. As an example, let us take the most abundant repeat,
196 the imperfect palindrome (IP) CAAACCAGT|NN|ACTGGTTAG. It is present in more than 1300
197 copies, with the dinucleotide AA as a predominant form of the internal spacer. This repeat is mostly
198 localised in clusters (>1200 cases) where its copies are interleaved by a repeat with the conserved
199 sequence TAACTAAACTTC, so together they constitute a composite extremely abundant tandem
200 repeat. In a single region, the palindromic repeat combines with a different interspersed repeat
201 (TAACTACTT), together forming a small cluster of composite tandem repeats in 14 copies.
202 Besides, the same palindromic repeat is also part of another, 146 bp-long repeat present in 27
203 copies across the plastome (for details see Additional file 2: Note S3).

204 Apart from intergenic regions, sequence repeats are found also in the four introns present
205 in the *L. pallida* plastome. The most prominent is a cluster of 20 copies of the motif
206 TGGTTAGTAACTAAACTTCCAAACCAGTAAAC in the intron inside the *atpA* gene that is

abundant also in intergenic regions (more than 1,000 copies typically located in huge clusters). Strikingly, when analysing the distribution of the most abundant IPs, we noticed that the motif AAGCCAGC|NNN|GCTGACTT and its variants are present also in coding regions, namely in exons of 11 out of 34 protein-coding genes of *L. pallida* plastome (Fig. 4a-c). They can be present in up to 36 copies per gene ("variant 8" in exons of *rpoC2*; Fig. 4b). In most cases the IP motif is part of a longer repeat unit including extra nucleotides at both ends ("variant 4" to "variant 8" in Fig. 4c). The most complex repeat unit variant is the following one (the IP core in round brackets; square brackets indicate alternative nucleotides occurring at the same position): AAAGAT-(AAGTCAGC|AGA|GCTGAC[AT]T)-CCAGACCACTAAAGTGGTCAGTAACTAAAAGTTAT. It is restricted to coding sequences (i.e., is absent from intergenic regions and introns) and occurs in eight copies inside three genes (*rpoC2*, *rpoC1*, *ftsH*), resulting in an insertion of a stretch of 20 amino acid residues in the encoded proteins. Other repeat variants (listed in Fig. 4c) have proliferated in exons as well as intergenic regions and introns. However, none of the aforementioned nucleotide motifs were found in plastomes of other chlamydomonadalean algae, indicating they have originated and diverged only in the *Leontynka* lineage.

Manual inspection of protein sequence alignments including chlamydomonadalean orthologs of the *L. pallida* proteins revealed that the intraexonic repeat insertions are located mostly in poorly conserved regions (see Fig. 4d for an example). Preferential proliferation in variable parts of coding regions is consistent with a high abundance of these motifs in proteins that exhibit a general tendency for including rapidly evolving and poorly conserved regions, namely FtsH (Additional file 1: Fig. S6), Ycf1, RpoC1, RpoC2, and RpoBb. Interestingly, the phase and orientation of the intraexonic repeats with respect to the reading frame and the direction of

transcription is not random and is potentially biased such that not only termination codons, but also codons generally rare in *L. pallida* plastid coding sequences are avoided from the actual frame in which the insertion is read during translation (for details see Fig. 4c, e, Additional file 2: Note S4). This bias does not merely reflect a possible bias in the orientation of the repeats relative to the DNA strand of the genome, as the repeats are distributed roughly equally in both strands when counted at the whole-genome level (Fig. 4c).

A high number of potential quadruplex-forming sequences in the GC-rich mitogenome of *Leontynka pallida*

The mitogenome sequence was assembled from Nanopore and Illumina reads as a linear molecule of 110515 bp with long (~5770 bp) nearly perfect (97.7% identity) direct terminal repeats differing primarily by the presence/absence of two short repetitive regions (13 and 70 bp) (Fig. 3c, Additional file 1: Fig. S7). This possibly indicates that the *L. pallida* mitogenome is in fact circular, with the slight differences in the terminal direct repeats of the assembled linear contigs reflecting sequence variability of a particular genomic region between the different genome copies in *L. pallida* or possibly sequencing or assembly artefacts. If circular, the mitogenome would then have a length of ~104812 bp. The suspected circularity of the mitogenome is also compatible with the absence of the *rtl* gene, which is present in all linear mitogenomes of Chlamydomonadales characterised to date and encodes a reverse transcriptase-like protein implicated in the replication of the mitogenome termini (Smith & Craig, 2021). Apart from *rtl*, the gene content of the *L. pallida* mitogenome is essentially the same as in other chlamydomonadalean mitogenomes sequenced before and includes seven protein-coding genes (with *cox1* interrupted by an ORF-free group II intron), only three tRNA genes, and regions corresponding to the 16S and 23S rRNA genes. As in

other chlamydomonadaleans studied in this regard (Boer & Gray, 1988; Denovan-Wright *et al.*, 1994; Fan *et al.*, 2003), the mitochondrial 16S and 23S rRNA genes in *L. pallida* are fragmented, consisting of multiple separately transcribed pieces. Four fragments, together constituting a presumably complete 16S rRNA, were annotated by considering the sequence and secondary structure conservation of the molecule. The number of the 16S rRNA fragments is thus the same as in *Chlamydomonas reinhardtii*, but the breakpoints are not completely identical. Due to a lower conservation of the 23S rRNA molecule, we could identify only a few of the presumed gene fragments in the *L. pallida* mitogenome.

The large size and the low density of coding sequences of the *L. pallida* mitogenome (~84.7% of its complete sequence is represented by intergenic regions) are atypical for Chlamydomonadales, including the other non-photosynthetic species: the mitogenome of *P. uvella* is 17.4 kbp long (Del Vasto *et al.*, 2015), and in *Polytomella* spp. the mitogenome size ranges from ~13 to 24.4 kbp (Smith *et al.*, 2010; Smith *et al.*, 2013). In fact, the *L. pallida* mitogenome can be compared only to the recently characterised mitogenome of *Haematococcus lacustris*, which with the same gene content is even larger (126.4 kbp) yet with a similar representation of intergenic regions (83.2%). A self-similarity plot generated for the *L. pallida* mitogenome revealed a highly repetitive nature of the genome sequence (Fig. 3d), similar to the plastome. However, the repeats are distributed less evenly than in the plastome, being present particularly in the terminal regions of the assembled linear sequence and in several internal hotspots.

With the GC content 62.6% (as counted for the circularised version of the genome), the mitogenome of *L. pallida* has the third highest documented mitochondrial GC content out of 11,077 examined mitogenomes available in GenBank, being surpassed only by the lycophyte *Selaginella moellendorffii* (68.2%; Hecht *et al.*, 2011) and the green alga *Picocystis salinarum*

(67.7%). These values contrast sharply with the median GC content value for the whole set of the mitogenomes examined, i.e. 38%. We also encountered an exceptionally high GC content (63.4%) and a strong bias towards using GC-rich codons in all protein-coding genes in the *L. pallida* mitogenome (see Additional file 3: Tables S2 and S3). Only two organisms are presently known to have an even higher GC content of mitochondrial protein-coding genes: the sponge *Leucosolenia complicata* (71.2%; Lavrov *et al.*, 2016) and *P. salinarum* (67.9%). Some *L. pallida* mitogenome-encoded proteins, namely Nad2 and Nad5, also exhibit a higher relative content of amino acids with GC-rich codons (G, A, R, and P) compared to most of their orthologs in other species (Additional file 3: Table S4). Thus, not only the expanded GC-rich intergenic regions, but also coding regions of the *L. pallida* mitogenome contribute to its extremely high GC content.

The repeats in the plastome and mitogenome of *H. lacustris* are nearly identical (Zhang *et al.*, 2019), so it was interesting to compare the two *L. pallida* organellar genomes to find out whether they behave similarly. However, as follows from the respective similarity plot (Fig. 3e) and comparison of most abundant inverted repeats and palindromes (Fig. 4a), the repeats in the two genomes do not resemble each other. The proliferation of different repeats in the two organellar genomes of *L. pallida* at least partially accounts for their strikingly different GC content (62.6% vs 37%). Interestingly, the most abundant IP in the *L. pallida* mitogenome contains the GGGG motif (Fig. 4a), which prompted us to bioinformatically investigate the possible occurrence of G-quadruplexes, unusual secondary structures in nucleic acids formed by guanine-rich regions (Burge *et al.*, 2006). Indeed, the *L. pallida* mitogenome was suggested to include up to 14.7 potential quadruplex-forming sequences (PQS) per 1,000 bp. A similar value was inferred for the *S. moellendorffii* mitogenome (15.6 PQS per 1,000 bp), whereas the other mitochondrial and plastid genomes that we analysed for comparison (for technical reasons focusing on GC-rich

299 genomes only) exhibited a much lower values (0.0-6.9 PQS per 1,000 bp; see Additional file 3:
300 Table S2).
301
302 **Discussion**
303 Both 18S rRNA and plastid gene sequence data concur on the conclusion that the two strains
304 investigated in this study, AMAZONIE and MBURUCU, represent a phylogenetically novel
305 lineage within Chlamydomonadales that is unrelated to any of the previously known non-
306 photosynthetic lineages in this order, i.e. *Polytomella*, *Polytoma sensu stricto* (including the type
307 species *P. uvella*), *Polytoma oviforme*, and the recently reported strain NrCl902. However,
308 morphological features of AMAZONIE and MBURUCU, including the cell shape and the
309 presence of two flagella, papilla, eyespot, and starch granules, make our organisms highly
310 reminiscent of the genus *Polytoma* (Ettl, 1983). This is consistent with the previous insight that
311 the *Polytoma* morphotype does not define a coherent phylogenetic unit (Figueroa-Martinez *et al.*,
312 2015). All other historically described genera of colourless flagellates assigned formerly to
313 Chlamydomonadales are sufficiently different from our strains as to consider them a potential
314 taxonomic home for AMAZONIE and MBURUCU (see Additional file 2: Note S5), justifying the
315 erection of the new genus *Leontynka* to accommodate the two strains. Furthermore, these strains
316 clearly differ from each other in morphology (cell shape and size, position of the eyespot) and are
317 genetically differentiated, as apparent from the comparison of the 18S rRNA gene and ITS2 region
318 sequences. Indeed, given the presence of several CBCs in the helix II of the conserved ITS2
319 secondary structure, the two strains are predicted to be sexually incompatible and hence
320 representing separate "biological species" (Coleman, 2000; Wolf *et al.*, 2013). We considered a
321 possibility that AMAZONIE and MBURUCU may represent some of the previously described

Polytoma species, but as detailed in Additional file 2: Note S5, none seems to be close enough in morphology as reported in the original descriptions. Given the fact that the majority of *Polytoma* species have been isolated and described from central Europe whereas our strains both come from tropical regions of South America, it is not so surprising that we encountered organisms new to science.

Leontynka spp. exhibit a number of ultrastructural similarities to the previously studied *Polytoma* species (Lang, 1963; Siu *et al.,* 1976; Gaffal & Schneider, 1980). For example, although photosynthetic chlamydomonadalean flagellates usually contain only a few mitochondria squeezed between the nucleus and the plastid, the cells of non-photosynthetic taxa, including *Leontynka*, are mitochondria-rich. It is possible that the proliferation of mitochondria compensates for the loss of the energetic function of the plastid in the non-photosynthetic species. Previous ultrastructural studies of *Polytoma obtusum* (Siu *et al.*, 1976) and *Polytomella* sp. (Dudkina *et al.*, 2010) showed that their mitochondria possess lamellar or irregular tubulo-vesicular cristae, respectively. The cristae of *L. pallida* resemble the latter morphotype, whereas *L. elongata* most probably possesses discoidal cristae (Additional file 1: Fig. S5e, f). Discoidal cristae are a very rare morphotype within the supergroup Archaeplastida, although they apparently evolved several times independently during the eukaryote evolution (Pánek *et al.*, 2020) and were previously noticed in several other non-photosynthetic chlorophytes (*Polytoma uvella*, *Polytomella agilis*, and *Prototheca zopfii*; Webster *et al.*, 1967).

A particularly notable feature of *Leontynka* spp. is the presence of two eyespots. These were more frequent in *L. elongata* (about half of the cells had two eyespots), whereas in the *L. pallida* cultures, such cells were rather rare. Variation in the number of eyespots (from none to multiple) in *Chlamydomonas reinhardtii* was shown to be a result of genetic mutations (Lamb *et*

345 *al.*, 1999), but the factors behind the eyespot number variation observed in *Leontynka* spp. are
346 unknown. The reddish colour of the *Leontynka* eyespots suggests the presence of carotenoids
347 (similar to the eyespot of *C. reinhardtii*; Böhm & Kreimer, 2020). In addition, searches of the *L.*
348 *pallida* transcriptome assembly revealed the presence of a homolog of the *C. reinhardtii* eyespot-
349 associated photosensor channelrhodopsin 1 (ChR1) that is the requires a carotenoid derivative,
350 retinal, as a chromophore (Petroutsos, 2017; Additional file 3: Table S1). The preservation of the
351 plastid-localized carotenoid biosynthetic pathway in non-photosynthetic eyespot-bearing
352 chlamydomonadaleans, namely certain *Polytomella* species and the strain NrCl902, has been noted
353 before (Asmail & Smith, 2016; Kayama *et al.*, 2020), and the same holds true for *L. pallida* based
354 on our analysis of its transcriptome assembly (Additional file 3: Table S1). Notably, like the
355 *Polytomella* species and the strain NrCl902 (Kayama *et al.*, 2020), *L. pallida* has also retained
356 enzymes for the synthesis of plastoquinone, which serves as an electron acceptor in two reactions
357 of carotenoid biosynthesis, and the plastid terminal oxidase (PTOX), which recycles plastoquinone
358 (from its reduced form plastoquinole) by passing the electrons further to molecular oxygen
359 (Additional file 3: Table S1). *Leontynka* thus represents an independent case supporting the notion
360 that retention of the eyespot constraints the reductive evolution of a non-photosynthetic plastid.

361 *Leontynka* is significant not only as a novel non-photosynthetic group *per se*, but also as
362 an independent lineage within Chlamydomonadales lacking any close photosynthetic relatives.
363 Specifically, based on the phylogenetic analysis of plastome-encoded proteins, *Leontynka*
364 branches off between two large assemblages, each comprised of several major
365 chlamydomonadalean clades defined by Nakada *et al.* (2008). One of these assemblages
366 (potentially sister to *Leontynka*) is comprised of the *Caudivolvoxa* and *Xenovolvoxa* clades, the
367 other includes *Reinhardtinia*, *Oogamochlamydinia*, and the genus *Desmotetra* (Fig. 1a). The

radiation of the *Reinhardtinia* clade itself was dated to ~300 MYA (Herron *et al.*, 2009), so the last common ancestor of *Leontynka* and any of its presently known closest photosynthetic relative must have existed even earlier. In other words, it is possible that *Leontynka* has been living without photosynthesis for hundreds of millions of years. The loss of photosynthesis in the four other known colourless chlamydomonadalean lineages certainly does not trace that far in the past. Specifically, the origin of *Polytomella* must postdate the radiation of *Reinhardtinia*, owing to the position of the genus with this clade, whereas *Polytoma sensu stricto* (*P. uvella* and relatives) has close photosynthetic relatives (*Chlamydomonas leiostraca*, *C. applanata* etc.) within the clade *Polytominia* in *Caudivolvoxa* (Fig. 1, Additional file 1: Fig. S1). *Polytoma oviforme* is specifically related to the photosynthetic *Chlamydomonas chlamydogama*, together constituting a clade in *Xenovolvoxa* that has not been formally recognised before and which we here designate "*Oviforminia*" (Additional file 1: Fig. S1). Finally, the recently reported strain NrCl902 is closely related to the photosynthetic *Chlamydomonas pseudoplanoconvexa* (Fig. 1A; Additional file 1: Fig. S1). The independent phylogenetic position of *L. pallida* based on plastome-encoded proteins is unlikely an artefact stemming from increased substitution rate of *L. pallida* plastid genes manifested by the markedly longer branch of *L. pallida* in the tree compared to most other species included in the analysis. Indeed, the branches of *P. uvella* and the strain NrCl902 are even longer (Fig. 1B), yet both organisms are placed at positions consistent with the 18S rRNA gene tree (Fig. 1A; Additional file 1: Fig. S1). Nevertheless, whether *Leontynka* represents a truly ancient non-photosynthetic lineage or whether it diverged from a photosynthetic ancestor rather recently needs to be tested by further sampling of the chlamydomonadalean diversity, as we cannot rule out the possibility that photosynthetic organisms closely related to the genus *Leontynka* are eventually discovered.

The presented considerations about the different ages of the separately evolved non-photosynthetic chlamydomonadalean lineages are somewhat at odds with features of their plastomes. Despite the presumably more recent loss of photosynthesis compared to *Leontynka*, both *P. uvella* and the strain NrCl902 exhibit a more reduced set of plastid genes (Table 1), whereas in *Polytomella*, plastome reduction triggered by photosynthesis loss has reached its possible maximum, i.e., a complete disappearance of the genome. It is likely that factors other than evolutionary time are contributing to the different degrees of plastome reduction in different evolutionary lineages, although little is known in this regard. Compared to *P. uvella* and the strain NrCl902, *L. pallida* has preserved one gene for a plastidial ribosomal protein (*rps2*) and, intriguingly, all standard plastidial genes for ATP synthase subunits, complemented by three more subunits encoded by the nuclear genome to allow for the assembly of a complete and presumably functional complex. It was proposed that the retention of ATP synthase in certain non-photosynthetic plastids is functionally linked to the retention of the twin-arginine protein translocase (Tat; Kamikawa *et al.*, 2015). The function of the translocase depends on a transmembrane proton gradient, which in photosynthetic plastids is primarily generated by the photosynthetic electron transport chain, whereas in non-photosynthetic ones its build-up would depend solely on the function of ATP synthase working in the opposite direction, i.e. pumping protons against the gradient at the expense of ATP. Interestingly, we found homologs of all three Tat subunits (TatA, TatB, TatC) in the nuclear transcriptome of *L. pallida* (Additional file 3: Table S1), providing further support to the hypothesis by Kamikawa *et al.* (2015). However, it must be noted that certain members of the non-photosynthetic trebouxiophyte genus *Prototheca* possess the plastidial ATP synthase in the absence of the Tat translocase (Suzuki *et al.*, 2018), suggesting that ATP synthase may be retained by a non-photosynthetic plastid for roles other than just

supporting the function of the Tat system. Directly relevant for the retention of ATP synthase in *L. pallida* might be its role in the functioning of the eyespot hypothesized in *C. reinhardtii* (Schmidt *et al.*, 2007).

The Tat translocase and the ATP synthase are both normally localised to the thylakoid membrane. While thylakoids may seem dispensable in a non-photosynthetic plastid, it seems there are putative thylakoids present in *Leontynka*, associated with the eyespot (Fig. 2j). Indeed, in the well-studied cases of *C. reinhardtii* and some other chlamydomonadalean algae, the layers of pigment granules are organized on the surface of thylakoids closely apposed to the plastid envelope (Kreimer, 1994; Böhm & Kreimer, 2020). Interestingly, our searches of the *L. pallida* transcriptome assembly revealed the presence of homologs of additional proteins functionally associated with thylakoids. These include components of several additional thylakoid-associated protein targeting or translocation systems (Schünemann *et al.*, 2007; Skalitzky *et al.*, 2011; Ziehe *et al.*, 2017), namely the plastidial SRP pathway (cpSRP54 and cpFtsY), ALB3 protein insertase, and thylakoid-specific Sec translocase (Additional file 3: Table S1). Furthermore, we also found in *L. pallida* homologs of proteins implicated in thylakoid biogenesis, such as VIPP1, FZL, THF1, or SCO2 (Mechela *et al.*, 2019; Additional file 3: Table S1). Interestingly, some of the corresponding transcripts have very low read coverage or are even represented by incomplete sequences, suggesting a low level of gene expression and presumably low abundance of the respective proteins. These observations support the notion that the thylakoid system is preserved in *Leontynka* plastids, however inconspicuous and likely reduced. Nevertheless, we cannot rule out that at least some of these proteins or complexes may have relocalised to the inner bounding membrane of the *Leontynka* plastid, or even to a cellular compartment other than the plastid (as suggested for some of these proteins by the results of *in silico* targeting prediction (Additional file

437 3: Table S1). The exact localisation of these complexes, the actual substrates of the plastidial SRP
438 pathway, ALB3 insertase, and Tat and Sec translocases, and indeed, the physiological functions
439 of the *L. pallida* leucoplast as a whole remain subjects for future research.

440 The most intriguing feature of *L. pallida* is the extreme expansion of its organellar
441 genomes. Generally, organellar genomes show a remarkable variation in the gene content,
442 architecture, and nucleotide composition, with most of them being AT-rich. The *L. pallida*
443 plastome is no exception in this respect, since its GC content is only ~37%. As noted by Smith
444 (2018), 98 % of plastomes are under 200 kbp and harbour modest amounts (<50 %) of non-coding
445 DNA. The *L. pallida* plastome, reaching 362.3 kbp, may not seem that impressive in comparison
446 with the giant plastomes recently reported from some photosynthetic species, including a distantly
447 related chlamydomonadalean *Haematococcus lacustris* (1.35 Mbp; Bauman *et al.*, 2018) or certain
448 red algae (up to 1.13 Mbp; Muñoz-Gómez *et al.*, 2017). However, it by far dwarfs plastomes of
449 all non-photosynthetic eukaryotes studied to date. The previous record holder, the plastome of
450 *Polytoma uvella* with ~230 kbp (Figueroa-Martinez *et al.*, 2017), accounts for only two thirds of
451 the size of the *L. pallida* plastome. The difference is not only because of a higher number of genes
452 in the latter, but primarily because of a more extreme expansion of intergenic regions in *L. pallida*
453 (4.7 kbp on average) than in *P. uvella* (3.0 kbp on average; Table 1). The plastome of the strain
454 NrCl902 with its size of 176.4 kbp, while exhibiting the same gene content as the *P. uvella*
455 plastome, is much less extreme (Kayama *et al.*, 2020), although still with the intergenic regions
456 substantially expanded as compared to the plastomes of non-photosynthetic trebouxiophytes
457 (Table 1).

458 Thus, despite its uniqueness, the organisation of the *L. pallida* plastome fits into the general
459 pattern observed in chlamydomonadalean algae, where plastomes in different lineages tend to

increase in size by accumulating repetitive sequences (Gaouda *et al.*, 2018; Smith, 2018). It was suggested that the repeats are prone to double-strand breaks, which are then repaired by an error-prone mechanism favouring repeat expansion (Smith, 2020a). However, the plastome of *L. pallida* is bloated not only due to extreme proliferation of repetitive DNA in intergenic regions, but also due to the expansion of some of them into the intronic regions and, much more surprisingly, even into exons (Fig. 4). The biased orientation and phase of the insertions with respect to the coding sequence and the reading frame avoid introduction of termination codons as well as rare codons or codons for rare amino acids (C, W) into the coding sequences (Fig. 4 c, d, Additional file 2: Note S4), which suggests that purifying selection eliminates those insertions that would disrupt or reduce the efficiency of translation of the respective mRNAs. Still, exons provide an important niche for the repeats: for example, for the "variant 8" repeat, the exonic copies constitute ~12.2% of the whole repeat population (compared to protein-coding sequences constituting ~17.2% of the total plastome length)! Such a massive proliferation of repeats to coding regions is unprecedented to our knowledge, although a much less extensive invasion of a different repeat into coding sequences was recently noticed in the plastome of another chlamydomonadalean alga, *Chlorosarcinopsis eremi* (Smith 2020a). Here the repeats are found in small numbers in the genes *ftsH*, *rpoC2*, and *ycf1*, paralleling the situation in *L. pallida* and consistent with the notion that genes encoding proteins rich in poorly conserved regions are most likely to tolerate the invasion of the repeats.

Recent sequencing of the mitogenome of *H. lacustris*, which is inflated by the accumulation of repeats highly similar to those found in the plastome of the same species (Zhang *et al.*, 2019), provided the first evidence that error-prone repair of double-strand breaks leading to repeat proliferation may operate also in chlamydomonadalean mitochondria. Smith (2020b)

483 recently reported the presence of highly similar repeats in the mitogenome of another
484 chlamydomonadalean alga, *Stephanosphaera pluvialis*, and proposed horizontal gene transfer
485 between the *H. lacustris* and *S. pluvialis* lineages as a possible explanation for the sharing of
486 similar mitochondrial repeats by the two organisms. Our characterisation of the *L. pallida*
487 mitogenome, which is also repeat-rich and larger than any chlamydomonadalean algal
488 mitogenome sequenced to date except that from *H. lacustris*, revealed that mitogenome inflation
489 may be more common in Chlamydomonadales. However, in contrast to *H. lacustris*, the GC
490 content as well as the repeats in the two organellar genomes of *L. pallida* differ significantly
491 (Additional file 3: Table S2), so the evolutionary path leading to the parallel inflation of both
492 genomes in this lineage may have been completely different from the one manifested in *H.*
493 *lacustris*. Strikingly, the specific nature of the mitochondrial repeats in *L. pallida* entails the high
494 abundance of PQS in the mitogenome. G-quadruplexes are increasingly recognised as regulatory
495 structures (Hänsel-Hertsch *et al.*, 2016), and they can form also in the mitogenomes, although their
496 role in mtDNA still needs to be elucidated (Falabella *et al.*, 2019). However, the PQS abundance
497 in the *L. pallida* mitogenome is truly extreme and comparable only with the situation in the
498 mitogenome of the lycophyte *S. moellendorffii* (Additional file 3: Table S2). Both species are thus
499 interesting candidates for studying the role of G-quadruplexes in mitochondrial DNA.
500

501 **Conclusions and future directions**

502 Our study indicates that continued sampling of microbial eukaryotes is critical for further progress
503 in our knowledge of the phylogenetic diversity of life and for better understanding of the general
504 principles governing the evolution of organellar genomes. The specific factors contributing to the
505 propensity of chlamydomonadalean organellar genomes to accumulate repetitive sequences,

reaching one of its extremes in *L. pallida*, remain unknown and may not be easy to define. However, future research on *Leontynka*, including characterisation of organellar genomes of *L. elongata*, may bring additional insights into the molecular mechanisms and evolutionary forces shaping the organellar genomes in this group. It will also be important to perform a detailed comparative analysis of the molecular machinery responsible for genome replication and maintenance in Chlamydomonadales and other green algae. The transcriptome assembly reported here for *L. pallida* will be instrumental not only in this enterprise, but will also serve as a resource for exploring the full range of physiological roles of the plastid in the *Leontynka* lineage and may help to further clarify the phylogenetic position of *Leontynka* within Chlamydomonadales. We posit that *Leontynka* may become an important model system for analysing the evolutionary and functional aspects of photosynthesis loss in eukaryotes with primary plastids.

Methods

Isolation, cultivation, and basic characterisation of new protist strains

Two strains, AMAZONIE and MBURUCU, were obtained from freshwater hypoxic sediment samples collected in Peru and Argentina, respectively. The strains were cultivated and morphologically characterised by light and transmission electron microscopy, using routine methods. Basic molecular characterisation was achieved by determining partial sequences of the rDNA operon. Further details are provided in Additional file 2: Methods S1-S3.

Organellar genome and nuclear transcriptome sequencing

Bacterial contamination in the AMAZONIE culture was minimised by filtration, and DNA and RNA were extracted using standard protocols detailed in Additional file 2: Methods S4. Nanopore

sequencing was performed using 4 µg of genomic DNA. The DNA was sheared at 20 kbp using Covaris g-TUBE (Covaris) according to the manufacturer's protocol. After shearing, two libraries were prepared using Ligation Sequencing Kit from Oxford Nanopore Technologies (SQK-LSK108). The prepared library was loaded onto a R9.4.1 Spot-On Flow cell (FLO-MIN106). Sequencing was performed on a MinION Mk1B machine for 48 hours using the MinKNOW 2.0 software. Basecalling was performed using Guppy 3.0.3 with the Flip-flop algorithm. Illumina sequencing of the genomic DNA was performed using 1 µg of genomic DNA with the Illumina HiSeq 2000 (2x150bp) paired-end technology with libraries prepared using TruSeq DNA PCR-Free (Illumina, San Diego, CA) at Macrogen Inc. (Seoul, South Korea). The transcriptome was sequenced using the HiSeq 2000 (2x100bp) paired-end technology with libraries prepared using the TruSeq RNA sample prep kit v2 (Illumina, San Diego, CA) at Macrogen Inc. (Seoul, South Korea).

Organellar genome and nuclear transcriptome assembly

Raw Illumina sequencing reads were trimmed with Trimmomatic v0.32 (Bolger *et al.*, 2014). Initial assembly of the Oxford Nanopore data was performed using Canu v1.7 with the corMaxEvidenceErate set to 0.15 (Koren *et al.*, 2017). After assembly, the plastome-derived contigs were identified using BLAST (Altschul *et al.*, 1997) with the *Chlamydomonas reinhardtii* plastome as a query. Nine putative plastid genome sequences were selected and polished using the raw nanopore reads with Nanopolish (Loman *et al.*, 2015) followed by polishing with Illumina reads with Pilon v1.22 (Walker *et al.*, 2014). After polishing of the contigs, the Illumina reads were re-mapped onto them, and the mapped reads were extracted and used as an input in Unicycler v0.4.8 (Wick *et al.*, 2017) together with the nanopore reads. Unicycler generated a single circular

552 contig of 362,307 bp. For the mitogenome, a single linear contig was identified in the Canu
553 assembly with BLAST with standard mitochondrial genes as queries; the contig sequence was
554 polished using the same method as described above, but it remained linear after a subsequent
555 Unicycler run. However, direct inspection of the contig revealed highly similar regions (about
556 5,600 bp in length) at both termini. The terminal regions were further polished by mapping of
557 Illumina genomic reads using BWA (Li, 2013) and SAMtools (Li *et al.*, 2009) followed by manual
558 inspection in Tablet (Milne *et al.*, 2016), which increased the sequence similarity of the termini to
559 97.7% (along the region of 5,771 bp).

560 Illumina genomic reads were also assembled separately with the SPAdes Genome
561 assembler v3.10.1 (Bankevich *et al.*, 2012) and used for cleaning the transcriptomic data as
562 follows. Contaminant bacterial contigs > 400,000 bp that were identified with BLAST in the
563 SPAdes (16 contigs) and Canu assemblies (11 contigs), together with published genome
564 assemblies of close relatives of bacteria identified in the AMAZONIE culture (*Curvibacter*
565 *lanceolatus* ATCC 14669, *Bacteroides luti* strain DSM 26991, and *Paludibacter jiangxiensis*
566 strain NM7), were used for RNA-seq read mapping (Hisat2 2.1.0; Kim *et al.*, 2015) to identify and
567 remove bacterial transcriptomic reads that survived the filtration of the culture and polyA
568 selection. This procedure removed ~4 % of the reads. Cleaned reads were used for transcriptome
569 assembly with the rnaSPAdes v3.13.0 using k-mer size of 55bp (Bushmanova *et al.*, 2019).

570

571 **Annotation of organellar genomes and other sequence analyses**

572 Initial annotation of both the plastid and mitochondrial genomes of the strain AMAZONIE were
573 obtained using MFannot (http://megasun.bch.umontreal.ca/cgi-bin/dev_mfa/mfannotInterface.pl).
574 The program output was carefully checked manually, primarily by relying on BLAST searches, to

find possible missed genes, to validate or correct the assessment of the initiation codons, to fix the delimitation of introns, and to ensure that all genes were properly named. ORFs lacking discernible homologs (as assessed by HHpred; Zimmermann *et al.*, 2018) encoding proteins shorter than 150 amino acid residues and ORFs consisting mostly of sequence repeats were omitted from the annotation. Distribution of repeats within the organellar genomes and comparison of repeats between organellar genomes of *L. pallida* and other selected chlamydomonadaleans were analysed using the dottup programme from the EMBOSS package (http://www.bioinformatics.nl/cgi-bin/emboss/dottup). Detailed analyses of imperfect palindromes and G-quadruplexes were performed using the Palindrome analyzer (Brázda *et al.*, 2016) and the G4hunter web-based server (Brázda *et al.*, 2019). The Palindrome analyzer was used to search for motifs 8-100 bp in length with spacers 0-10 bp, and a maximum of one mismatch in the palindrome. The G4hunter web-based server was used under the default settings, i.e., window=25 and threshold=1.2.

To understand the position of amino acid stretches encoded by the characteristic repeats that have invaded the coding sequence of the *ftsH* gene, the tertiary structure of the encoded protein was predicted by homology modelling using the Phyre2 program (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index; Kelley *et al.*, 2015). The secondary structure of the ITS2 region was modelled manually according to the consensus secondary ITS2 structure of two green algae (Caisová *et al.*, 2013), visualised by VARNA software (Darty *et al.*, 2009), and manually edited in a graphical editor. Homologs of nucleus-encoded plastidial proteins of specific interest were searched in the *L. pallida* transcriptome assembly by using TBLASTN and the respective proteins sequences from *Arabidopsis thaliana* or *C. reinhardtii* (selected based on the information from the literature or keyword database searches). Significant hits (E-value ≤1e-5) were evaluated by BLASTX searches against the NCBI non-redundant protein sequence

598 database to filter out bacterial contaminants and sequences corresponding to non-orthologous
599 members of broader protein families. Subcellular localization (for complete sequences only) was
600 assessed by using TargetP-2.0 (https://services.healthtech.dtu.dk/service.php?TargetP-2.0;
601 Almagro Armenteros *et al.*, 2019) and PredAlgo (http://lobosphaera.ibpc.fr/cgi-
602 bin/predalgodb2.perl?page=main; Tadrif *et al.*, 2012).
603
604 **Phylogenetic analyses**
605 Multiple sequence alignment of 18S rRNA gene relying on a total of 201 chlorophyte OTUs was
606 computed using MAFFT v7 (Katoh *et al.*, 2019) and trimmed manually. The 18S rRNA sequence
607 from *Polytoma oviforme* available in GenBank (U22936.1) was proposed to be chimeric (Nakada
608 *et al.*, 2008), but given the relevance of this organism for our analysis, we included it, masking the
609 regions putatively derived from a different source by strings of N. Maximum likelihood tree
610 inference was performed using IQ-TREE multicore v1.6.12 (Nguyen *et al.*, 2015) under
611 TIM2+F+I+G4 model with 100 non-parametric bootstrap replicates. For multigene analysis,
612 alignments of conserved plastome-encoded proteins used previously (Fučíková *et al.*, 2019) were
613 updated by adding the respective homologs from *L. pallida* and thirteen additional relevant
614 chlorophycean taxa not represented in the initial dataset. On the other hand, sequences representing
615 the OCC clade of Chlorophyceae (evidently only distantly related to *L. pallida* based on the 18S
616 rRNA gene phylogeny and morphological features) to keep the size of the dataset easier to analyse
617 with a complex substitution model. For the final matrix, a subset of 24 proteins (all having their *L.*
618 *pallida* representative) were used. Multiple alignments of the homologous amino acid sequences
619 were built using MAFFT v7.407 with the L-INS-i algorithm (Katoh & Standley, 2013) and
620 manually trimmed to exclude unreliably aligned regions. The final concatenated matrix comprised

5,020 amino acid residues. The tree was built using PhyloBayes v4.1 (Lartillot *et al.*, 2013) under the CAT + GTR model of sequence evolution, with two independent chains that converged at 15,298 generations with the largest discrepancy in posterior probabilities (PPs) (maxdiff) of 0.0535238 (at burn-in of 20%). The maximum likelihood (ML) tree was inferred with IQ-TREE multicore v1.6.12 using the LG+C60+F+G4 substitution model. Statistical support was assessed with 100 IQ-TREE non-parametric bootstraps with correction and PhyloBayes posterior probabilities.

Supplementary information

Supplementary information accompanies this paper at XXXXXXX.

Additional file 1: Supplementary Figs S1-S9. Fig. S1. Maximum likelihood phylogenetic tree (RAxML, GTRGAMMA+I substitution model) of 18S rRNA gene sequences from Chlorophyceae. **Fig. S2.** Predicted secondary structure of the ITS2 region of *Leontynka pallida*, with differences in the corresponding region of *Leontynka elongata* mapped onto it. **Fig. S3.** *Leontynka pallida* under the light microscope. **Fig. S4.** *Leontynka elongata* under the light microscope. **Fig. S5.** Ultrastructure of *Leontynka elongata* (a–f) and *Leontynka pallida* (g–i). **Fig. S6.** Occurrence of the "variant 8" repeat in the FtsH protein of *Leontynka pallida* mapped on its predicted structure. **Fig. S7.** Alignment of the highly similar terminal regions of the originally assembled linear mitogenome contig. **Fig. S8.** Occurrence of the "variant 8" repeat (translated in reading frame +0 as KDKPANLTS and -0 as KEVSFAGLSL) in variable region of protein sequence of the ribosomal protein Rps8 from *Leontynka pallida* (full protein alignment together with representatives of other chlamydomonadalean algae).

Additional file 2: Supplementary Notes S1-S5 and supplementary Methods S1-S4. Note S1. Taxonomic descriptions. **Note S2.** Further details on the morphology and ultrastructure of *Leontynka* spp. **Note S3.** Further details on various kinds of repeats in the plastome of *L. pallida*. **Note S4.** Further details on the repeat insertions in *L. pallida* plastid coding sequences. **Note S5.** Differential diagnosis of *Leontynka* spp. with regard to previously described colourless chlamydomonadalean taxa. **Methods S1.** Isolation and cultivation of strains. **Methods S2.** Light and transmission electron microscopy. **Methods S3.** Amplification and sequencing of 18S and ITS rDNA regions. **Methods S4.** DNA and RNA isolation.

Additional file 3: Supplementary Tables S1-S7. Table S1. Nuclear transcripts from *Leontynka pallida* specifically discussed in the paper. **Table S2.** Comparison of GC content, number of imperfect palindromes, and potential quadruplex-forming sequences in selected organellar genomes. **Table S3.** Strong codon usage bias in the mitochondrial genome of *Leontynka pallida*. **Table S4.** Relative frequency of amino acids with GC-rich codons (G, A, R, P) in proteins encoded by different mitogenomes. **Table S5.** Relative frequency of codons in plastid genes of *Leontynka pallida*. **Table S6.** Relative frequency of amino acids in proteins encoded by the plastome of *Leontynka pallida*. **Table S7.** The most abundant imperfect palindrome in the *Leontynka pallida* plastome that is missing in exons.

Declarations

Ethics approval and consent to participate

Not applicable.

667

668 **Consent for publication**

669 Not applicable.

670

671 **Competing interests**

672 The authors declare no competing interests.

673

674 **Availability of data and materials**

675 Sequences determined in this study are available from GenBank with the following accession
676 numbers: ##### and ##### – partial nuclear rDNA sequences (18S rRNA-ITS1-5.8S rRNA-ITS2)
677 from *L. pallida* and *L. elongata*, respectively; ##### and #### – plastid and mitochondrial genome
678 sequence from *L. pallida*; ###### – transcriptome assembly from *L. pallida*. The cultures of
679 *Leontynka* spp. investigated in this study are available upon request.

680

681 **Funding**

682 This work was supported by the Czech Science Foundation project 17-21409S (to M.E.) and the
683 project "CePaViP", supported by the European Regional Development Fund, within the
684 Operational programme for Research, Development and Education
685 (CZ.02.1.01/0.0/0.0/16_019/0000759). TP was also supported by the Charles University (UNCE
686 204069).

687

688 **Authors' contribution**

TP, DB, IČ and ME conceived the original research plans; TP, SCT, KZ, EZ, and KJ obtained nucleic acids for sequencing; SCT and TP obtained Oxford Nanopore data and generated organellar genome assemblies; EZ and IČ isolated the strains; TP and DB carried out the morphological characterisation of the strains; DB and NY obtained the TEM data; MS assembled the transcriptome; TP, DB, and TŠ carried out phylogenetic analyses; TP, IČ and ME analysed and annotated the organellar genome sequences; IČ and ME supervised the work of junior researchers and obtained funding; TP, DB and ME drafted the manuscript; all authors contributed to the final version of the text; ME agreed to serve as the author responsible for contact and ensuring communication.

Acknowledgements

We are thankful to Karolina Fučíková for providing alignments of chlorophycaean plastid proteins and Petr Janšta for collecting the sample MBURUCU.

References

Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, Nielsen H. 2019. Detecting sequence signals in targeting peptides using deep learning. *Life Science Alliance* **2**: e201900429.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Asmail SR, Smith DR. 2016. Retention, erosion, and loss of the carotenoid biosynthetic pathway in the nonphotosynthetic green algal genus *Polytomella*. *New Phytologist* **209**: 899–903.

712 **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,**
713 **Nikolenko SI, Pham S, Prjibelski AD *et al.* 2012.** SPAdes: a new genome assembly
714 algorithm and its applications to single-cell sequencing. *Journal of Computational*
715 *Biology* **19**: 455–477.

716 **Bauman N, Akella S, Hann E, Morey R, Schwartz AS, Brown R, Richardson TH. 2018.** Next-
717 generation sequencing of *Haematococcus lacustris* reveals an extremely large 1.35-
718 megabase chloroplast genome. *Genome Announcements* **6**: e00181–e001818.

719 **Böhm M, Kreimer G. 2020.** Orient in the World with a Single Eye: The Green Algal Eyespot and
720 Phototaxis. In: Cánovas FM, Lüttge U, Risueño MC, Pretzsch H (eds) Progress in Botany
721 Vol. 82. Progress in Botany, vol 82. Springer, Cham.

722 **Bolger AM, Lohse M, Usadel B. 2014.** Trimmomatic: a flexible trimmer for Illumina sequence
723 data. *Bioinformatics* **30**: 2114-2120.

724 **Boer PH, Gray MW. 1988.** Scrambled ribosomal RNA gene pieces in *Chlamydomonas*
725 *reinhardtii* mitochondrial DNA. *Cell* **55**: 399-411.

726 **Brázda V, Kolomazník J, Lýsek J, Bartas M, Fojta M, Šťastný J, Mergny J. 2019.** G4Hunter
727 web application: a web server for G-quadruplex prediction. *Bioinformatics* **35**: 3493–3495.

728 **Brázda V, Kolomazník J, Lýsek J, Hároníková L, Coufal J, Št'astný J. 2016.** Palindrome
729 analyser - A new web-based server for predicting and evaluating inverted repeats in
730 nucleotide sequences. *Biochemical and Biophysical Research Communications* **478**: 1739–
731 1745.

732 **Burge S, Parkinson GN, Hazel P, Todd AK, Neidle S. 2006.** Quadruplex DNA: sequence,
733 topology and structure *Nucleic Acids Research* **34**: 5402–5415.

734 **Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. 2019.** rnaSPAdes: a de novo
735 transcriptome assembler and its application to RNA-Seq data. *GigaScience* **8**: giz100.
736 **Caisová L, Marin B, Melkonian M. 2013.** A consensus secondary structure of ITS2 in the
737 Chlorophyta identified by phylogenetic reconstruction. *Protist* **164**: 482–496.
738 **Coleman AW. 2000.** The significance of a coincidence between evolutionary landmarks found in
739 mating affinity and a DNA sequence. *Protist* **151**: 1–9.
740 **Darty K, Denise A, Ponty Y. 2009.** VARNA: Interactive drawing and editing of the RNA
741 secondary structure. *Bioinformatics* **25**: 1974–1975.
742 **Del Vasto M, Figueroa-Martinez F, Featherston J, González MA, Reyes-Prieto A, Durand**
743 **PM, Smith DR. 2015.** Massive and widespread organelle genomic expansion in the green
744 algal genus *Dunaliella*. *Genome Biology and Evolution* **7**: 656–663.
745 **Denovan-Wright EM, Lee RW. 1994.** Comparative structure and genomic organization of the
746 discontinuous mitochondrial ribosomal RNA genes of *Chlamydomonas eugametos* and
747 *Chlamydomonas reinhardtii*. *Journal of Molecular Biology* **241**: 298–311.
748 **Dudkina NV, Oostergetel GT, Lewejohann D, Braun HP, Boekema EJ. 2010.** Row-like
749 organization of ATP synthase in intact mitochondria determined by cryo-electron
750 tomography. *Biochimica et Biophysica Acta (BBA)-Bioenergetics* **1797**: 272–277.
751 **Ettl H. 1983.** Chlorophyta I: Phytomonadina. In: Ettl H, Gerloff J, Heynig H, Mollenhauer D
752 (eds.), Süßwasserflora von Mitteleuropa. Band 9. Stuttgart: Gustav Fischer Verlag. 807 p.
753 **Falabella M, Fernandez RJ, Johnson FB, Kaufman BA. 2019.** Potential roles for G-
754 quadruplexes in mitochondria. *Current Medicinal Chemistry* **26**: 2918–2932.
755 **Fan J, Schnare MN, Lee RW. 2003.** Characterization of fragmented mitochondrial ribosomal
756 RNAs of the colorless green alga *Polytomella parva*. *Nucleic Acids Research* **31**: 769–778.

Figueroa-Martinez F, Nedelcu AM, Smith DR, Reyes-Prieto A. 2015. When the lights go out: the evolutionary fate of free-living colorless green algae. *New Phytologist* **206**: 972–982.

Figueroa-Martinez F, Nedelcu AM, Smith DR, Reyes-Prieto A. 2017. The plastid genome of *Polytoma uvella* is the largest known among colorless algae and plants and reflects contrasting evolutionary paths to nonphotosynthetic lifestyles. *Plant Physiology* **173**: 932–943.

Fučíková K, Lewis PO, Neupane S, Karol KG, Lewis LA. 2019. Order, please! Uncertainty in the ordinal-level classification of Chlorophyceae. *PeerJ* **7**: e6899.

Gaffal KP, Schneider GJ. 1980. Numerical, morphological and topographical heterogeneity of the chondriome during the vegetative life cycle of *Polytoma papillatum*. *Journal of Cell Science* **46**: 299–312.

Gaouda H, Hamaji T, Yamamoto K, Kawai-Toyooka H, Suzuki M, Noguchi H, Minakuchi Y, Toyoda A, Fujiyama A, Nozaki H *et al.* 2018. Exploring the limits and causes of plastid genome expansion in volvocine green algae. *Genome Biology and Evolution* **10**: 2248–2254.

Greiner S, Lehwark P, Bock R. 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Research* **47**: W59–W64

Gruber HE, Rosario B. 1974. Variation in eyespot ultrastructure in *Chlamydomonas reinhardi* (ac-31). *Journal of Cell Science* **15**: 481–494.

Hadariová L, Vesteg M, Hampl V, Krajčovič J. 2018. Reductive evolution of chloroplasts in non-photosynthetic plants, algae and protists. *Current Genetics* **64**: 365–387.

Hänsel-Hertsch R, Beraldi D, Lensing SV, Marsico G, Zyner K, Parry A, Di Antonio M, Pike J, Kimura H, Narita M *et al.* 2016. G-quadruplex structures mark human regulatory chromatin. *Nature Genetics* **48**: 1267–1272.

Hecht J, Grewe F, Knoop V. 2011. Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of *Selaginella moellendorffii* mitochondria: The root of frequent plant mtDNA recombination in early tracheophytes. *Genome Biology and Evolution* **3**: 344–358.

Herron MD, Hackett JD, Aylward FO, Michod RE. 2009. Triassic origin and early radiation of multicellular volvocine algae. *Proceedings of the National Academy of Sciences, USA* **106**: 3254–3258.

Kamikawa R, Tanifuji G, Ishikawa SA, Ishii K, Matsuno Y, Onodera NT, Ishida K, Hashimoto T, Miyashita H, Mayama S *et al.* 2015. Proposal of a twin arginine translocator system-mediated constraint against loss of ATP synthase genes from nonphotosynthetic plastid genomes. [Corrected]. *Molecular Biology and Evolution* **32**: 2598–2604.

Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* **20**: 1160–1166.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

Kayama M, Chen JF, Nakada T, Nishimura Y, Shikanai T, Azuma T, Miyashita H, Takaichi S, Kashiyama Y, Kamikawa R. 2020. A non-photosynthetic green alga illuminates the reductive evolution of plastid electron transport systems. *BMC Biology* **18**: 126.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* **10**: 845–858.

Kim D, Langmead B, Salzberg S. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**: 357–360.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27**: 722–736.

Kreimer G. 1994. Cell biology of phototaxis in flagellate algae. *International Review of Cytology* **148**: 229–310.

Lamb MR, Dutcher SK, Worley CK, Dieckmann CL. 1999. Eyespot-assembly mutants in *Chlamydomonas reinhardtii*. *Genetics* **153**: 721–729.

Lang NJ. 1963. Electron microscopic demonstration of plastids in *Polytoma*. *Journal of Protozoology* **10**: 333–339.

Lavrov DV, Adamski, M, Chevaldonne, P, Adamska, M. 2016. Extensive mitochondrial mRNA editing and unusual mitochondrial genome organization in calcaronean sponges. *Current Biology* **26**: 86-92.

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology* **62**: 611–615.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods* **12**: 733–735.

Mechela A, Schwenkert S, Soll J. 2019. A brief history of thylakoid biogenesis. *Open Biology* **9**: 180237.

Milne I, Bayer M, Stephen G, Cardle L, Marshall D. 2016. Tablet: visualizing next-generation sequence assemblies and mappings. *Methods in Molecular Biology* **1374**: 253–268.

Muñoz-Gómez SA, Mejía-Franco FG, Durnin K, Colp M, Grisdale CJ, Archibald JM, Slamovits CH. 2017. The new red algal subphylum Proteorhodophytina comprises the largest and most divergent plastid genomes known. *Current Biology* **27**: 1677–1684.e4.

Nakada T, Misawa K, Nozaki H. 2008. Molecular systematics of Volvocales (Chlorophyceae, Chlorophyta) based on exhaustive 18S rRNA phylogenetic analyses. *Molecular Phylogenetics and Evolution* **48**: 281–291.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**: 268–274.

Petroutsos D. 2017. Chlamydomonas Photoreceptors: Cellular Functions and Impact on Physiology. In: Hippler M. (eds) Chlamydomonas: Biotechnology and Biomedicine. Microbiology Monographs, vol 31. Springer, Cham.

Schmidt M, Luff M, Mollwo A, Kaminski M, Mittag M, Kreimer G. 2007. Evidence for a specialized localization of the chloroplast ATP-synthase subunits α, β, and γ in the eyespot apparatus of *Chlamydomonas reinhardtii* (Chlorophyceae). *Journal of Phycology* **43**: 284–294.

Schünemann D. 2007. Mechanisms of protein import into thylakoids of chloroplasts. *Biological Chemistry* **388**: 907–915.

Sibbald SJ, Archibald JM. 2020. Genomic insights into plastid evolution. *Genome Biology and Evolution* **12**: 978–990.

Siu C, Swift H, Chiang K. 1976. Characterization of cytoplasmic and nuclear genomes in the colorless alga *Polytoma*. *Journal of Cell Biology* **69**: 352–370.

Skalitzky CA, Martin JR, Harwood JH, Beirne JJ, Adamczyk BJ, Heck GR, Cline K, Fernandez DE. 2011. Plastids contain a second sec translocase system with essential functions. *Plant Physiology* **155**: 354–369.

Smith, DR. 2012. Updating our view of organelle genome nucleotide landscape. *Frontiers in Genetics* **3**: 175.

Smith DR, Hua J, Lee RW. 2010. Evolution of linear mitochondrial DNA in three known lineages of *Polytomella*. *Current Genetics* **56**: 427–438.

Smith DR, Hua J, Archibald JM, Lee RW. 2013. Palindromic genes in the linear mitochondrial genome of the nonphotosynthetic green alga *Polytomella magna*. *Genome Biology and Evolution* **5**: 1661–1667.

Smith DR, Lee RW. 2014. A plastid without a genome: evidence from the nonphotosynthetic green algal genus *Polytomella*. *Plant Physiology* **164**: 1812–1819.

Smith DR. 2018. *Haematococcus lacustris*: the makings of a giant-sized chloroplast genome. *AoB Plants* **10**: ply058.

Smith DR. 2020a. Can green algal plastid genome size be explained by DNA repair mechanisms? *Genome Biology and Evolution* **12**: 3797–3802.

867 **Smith DR. 2020b.** Common repeat elements in the mitochondrial and plastid genomes of green
868 algae. *Frontiers in Genetics* **11**: 465.
869 **Smith DR, Craig, R. J. 2021.** Does mitochondrial DNA replication in *Chlamydomonas* require a
870 reverse transcriptase? *New Phytologist* **229**: 1192–1195.
871 **Suzuki S, Endoh R, Manabe RI, Ohkuma M, Hirakawa Y. 2018.** Multiple losses of
872 photosynthesis and convergent reductive genome evolution in the colourless green algae
873 *Prototheca*. *Scientific Reports* **8**: 940.
874 **Tardif M, Atteia A, Specht M, Cogne G, Rolland N, Brugière S, Hippler M, Ferro M, Bruley**
875 **C, Peltier G, Vallon O, Cournac L. 2012.** PredAlgo: a new subcellular localization
876 prediction tool dedicated to green algae. *Molecular Biology and Evolution* **29**: 3625–3639.
877 **Turmel M, Lemieux C. 2018.** Chapter six - Evolution of the plastid genome in green algae.
878 *Advances in Botanical Research* **85**: 157–193.
879 **Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q,**
880 **Wortman J, Young SK *et al.* 2014.** Pilon: an integrated tool for comprehensive microbial
881 variant detection and genome assembly improvement. *PLoS One* **9**: e112963.
882 **Webster DA, Hackett DP, Park RB. 1968.** The respiratory chain of colorless algae: III. Electron
883 microscopy. *Journal of Ultrastructural Research* **21**: 514–523.
884 **Wick RR, Judd LM, Gorrie CL, Holt KE. 2017.** Unicycler: resolving bacterial genome
885 assemblies from short and long sequencing reads. *PLoS Computational Biology* **13**:
886 e1005595.
887 **Wolf M, Chen S, Song J, Ankenbrand M, Müller T. 2013.** Compensatory base changes in ITS2
888 secondary structures correlate with the biological species concept despite intragenomic
889 variability in ITS2 sequences--a proof of concept. *PLoS One* **8**: e66726.

Zhang X, Bauman N, Brown R, Richardson TH, Akella S, Hann E, Morey R, Smith DR. 2019. The mitochondrial and chloroplast genomes of the green alga *Haematococcus* are made up of nearly identical repetitive sequences. *Current Biology* **29**: R736–R737.

Ziehe D, Dünschede B, Schünemann D. 2017. From bacteria to chloroplasts: evolution of the chloroplast SRP system. *Biological Chemistry* **398**: 653–661.

Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2018. A completely reimplemented MPI Bioinformatics Toolkit with a new HHpred server at its core. *Journal of Molecular Biology* **430**: 2237–2243.

Table 1. Basic characteristics of plastomes of *Leontynka pallida*, selected other non-photosynthetic chlorophytes and the photosynthetic *Chlamydomonas reinhardtii* and *Haematococcus lacustris* for comparison.

	plastomes		genes		coding DNA**		intergenic regions		
species	accession number	size (bp)	protein-coding*	tRNAs	GC (%)	total length (%)	GC (%)	average length (bp)	total length (%)
Leontynka pallida	######	362,307	32	26	35	18	37	4,726	80
Polytoma uvella	KX828177.1	230,207	25	27	40	20	19	2,998	68
Volvocales sp. NrCl902	LC516060.1	176,432	25	30	36.1	32.6	42.8	1,944	60.9
Helicosporidium sp.	DQ398104.1	37,454	26	25	27.6	95.4	13.4	33.8	4.4
Chlamydomonas reinhardtii	NC_005353.1	203,828	69	29	34.6	43	34	923	48.5
Haematococcus lacustris (†)	NC_037007.1	1,352,306	63	28	38.6	6.8	51	12,326	91.15

*does not include ORFs inside introns

**does not include introns and ORFs inside introns, but includes tRNAs and rRNAs

(†) since the annotation of the *H. lacustris* plastome available in the respective GenBank record is highly incomplete, the values presented are based on a reannotation obtained by using MFannot (with a genetic code translating UGA as Trp, based on the insights by Smith, 2018).

928 **Figure legends**

929 **Fig. 1** Phylogenetic position of the new genus of non-photosynthetic green algae, *Leontynka*. Non-
930 photosynthetic taxa in Chlamydomonadales are highlighted in violet. (a) Maximum likelihood
931 phylogenetic tree (IQ-TREE, TIM2+F+I+G4 substitution model) of 18S rRNA gene sequences
932 from Chlamydomonadales and related chlorophytes. Non-parametric bootstrap support values
933 calculated from 100 replicates are shown when ≥50. Previously demarcated main clades (Nakada
934 *et al.*, 2008) are collapsed and the outgroup taxa are not shown for simplicity. The full version of
935 the tree is provided as Additional file 1: Fig. S1. (b) Phylogenetic analysis of a concatenated dataset
936 of 24 conserved plastome-encoded proteins (5,020 amino acid positions) from
937 Chlamydomonadales, including *Leontynka pallida*, and the sister order Sphaeropleales (*sensu lato*;
938 see Fučíková *et al.*, 2019). The tree topology was inferred using PhyloBayes (CAT+GTR
939 substitution model), branch support values correspond to posterior probability (from PhyloBayes)
940 / maximum likelihood bootstrap analysis (IQ-TREE, LG+C60+F+G4 substitution model, 100 non-
941 parametric bootstrap replicates). Black dots represent full support obtained with both methods,
942 asterisks denote bootstrap support values <50.

943

944 **Fig. 2** Light and transmission electron microscopy of *Leontynka pallida* (a–e) and *Leontynka*
945 *elongata* (f–j). Note the difference in the cell shape between the two species and the presence of a
946 single (a, f, h) or two (b, d, g, i, j) eyespots. Lipid droplets were also detected within the cells (d,
947 h, j). Abbreviations: bb – basal body; cv – contractile vacuole; e – eyespot; L – lipid droplet; m –
948 mitochondrion; N – nucleus; n – nucleolus; s – starch. Arrows point to thylakoids; asterisks mark
949 plastid "bridges" connecting separate compartments; arrowheads show membranous inclusions.
950 Scale bars: a-c, f-g = 10 μm; d, i = 0.2 μm; e = 0.5 μm; j = 1 μm.

951

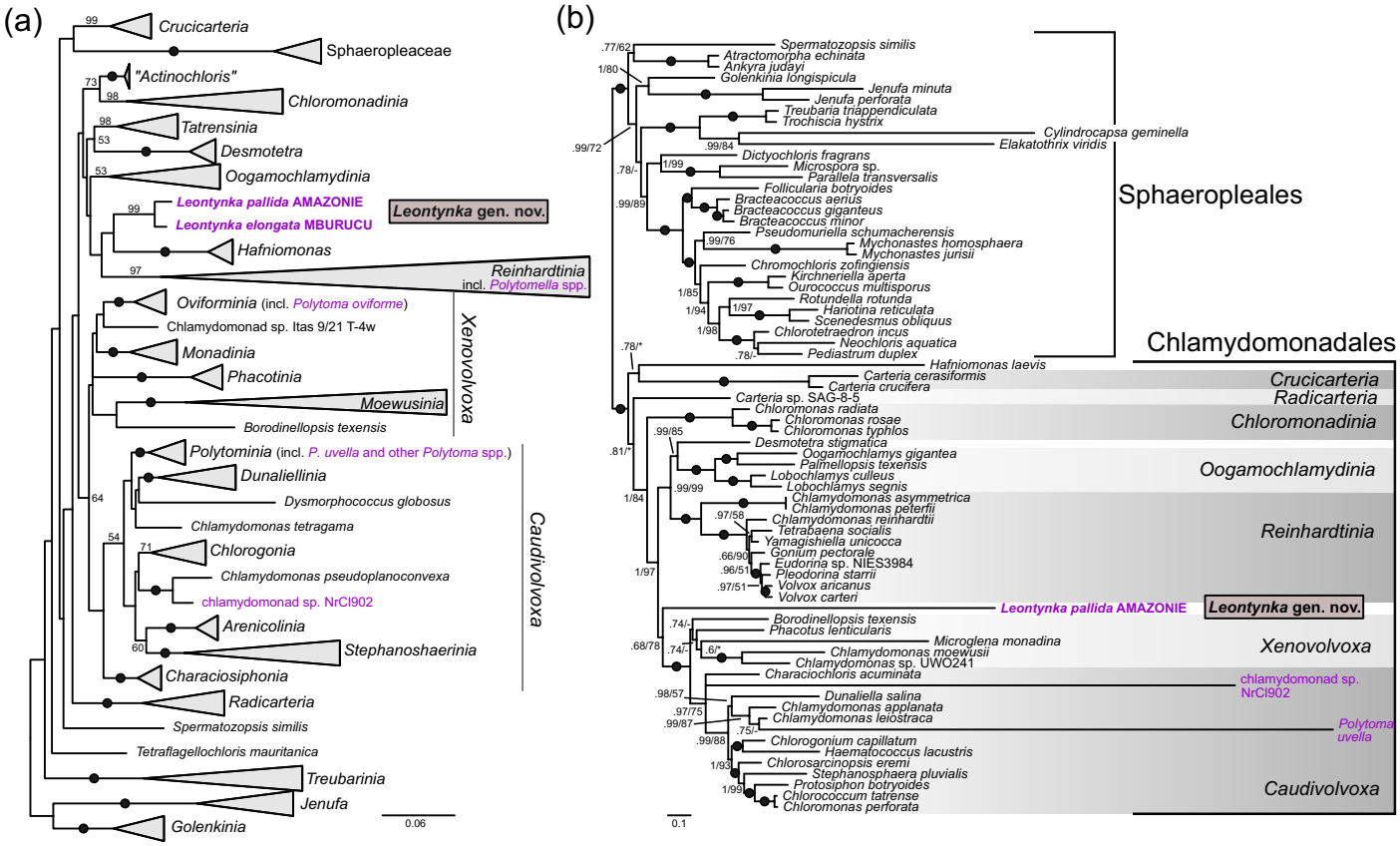
952 **Fig. 3** Characteristics of the organellar genomes of *Leontynka pallida*. (a) Gene map of the *L.*
953 *pallida* plastid genome. Genes are shown as squares (coloured according to the functional
954 category; see the graphical legend in the left bottom corner) on the inner or outer side of the outer
955 circle depending on their orientation (transcription in the clockwise or counter-clockwise direction,
956 respectively; see the grey arrows). Genes marked with an asterisk contain introns. The inner circle
957 represents a GC content plot. (b) Sequence self-similarity plot of the *L. pallida* plastome (ptDNA).
958 (c) Map of the *L. pallida* mitochondrial genome. The display convention is the same as for the
959 plastid genome. (d) Sequence self-similarity plot of the *L. pallida* mitochondrial genome
960 (mtDNA). (e) *L. pallida* plastome-mitogenome similarity plot. All similarity plots were generated
961 by using the word size of 15 bp and black dots represent the occurrence of the same word at the
962 places compared. The organellar genome maps (a, c) were visualised by using OGDRAW v1.3.1
963 (Greiner *et al.*, 2019).

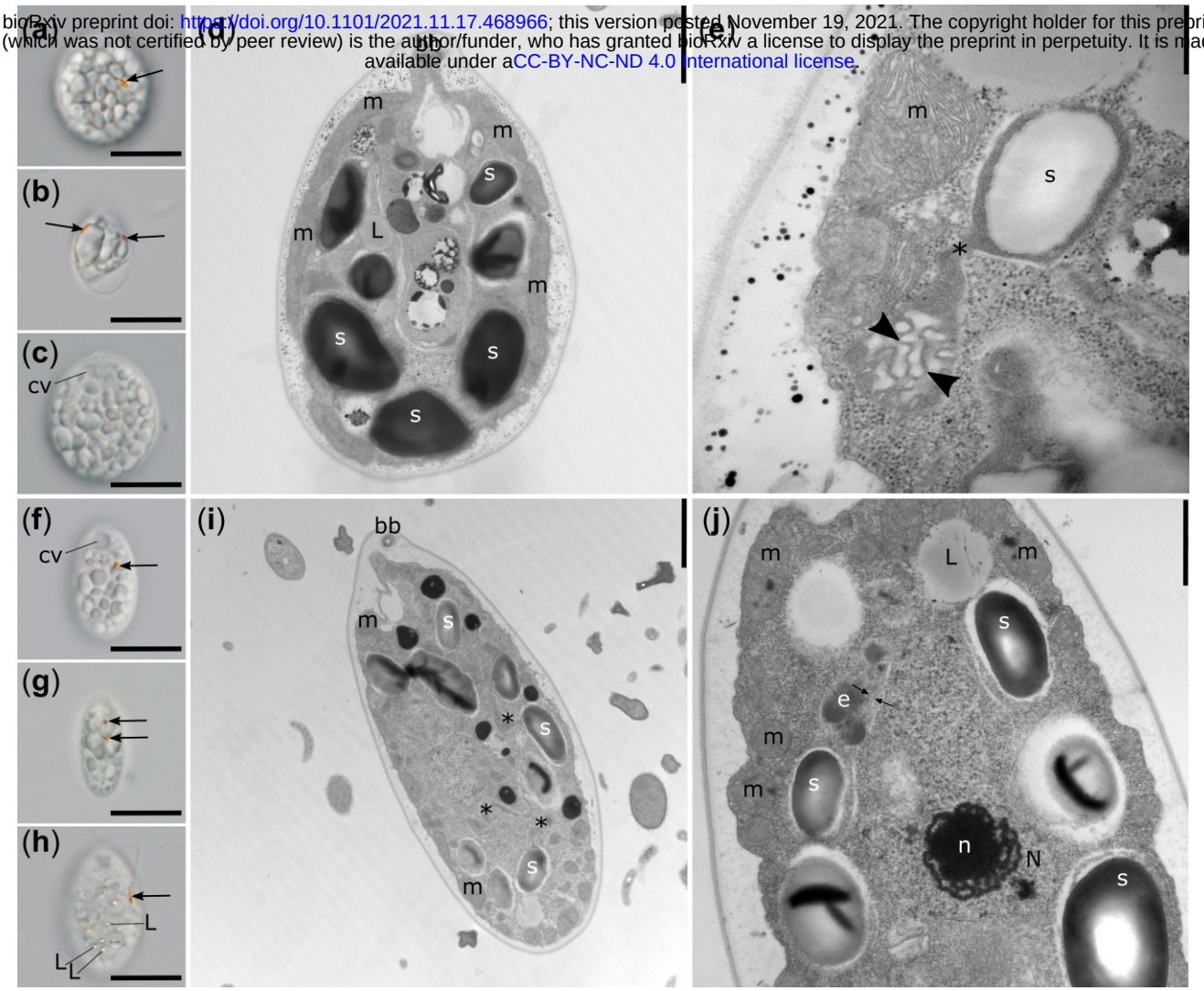
964

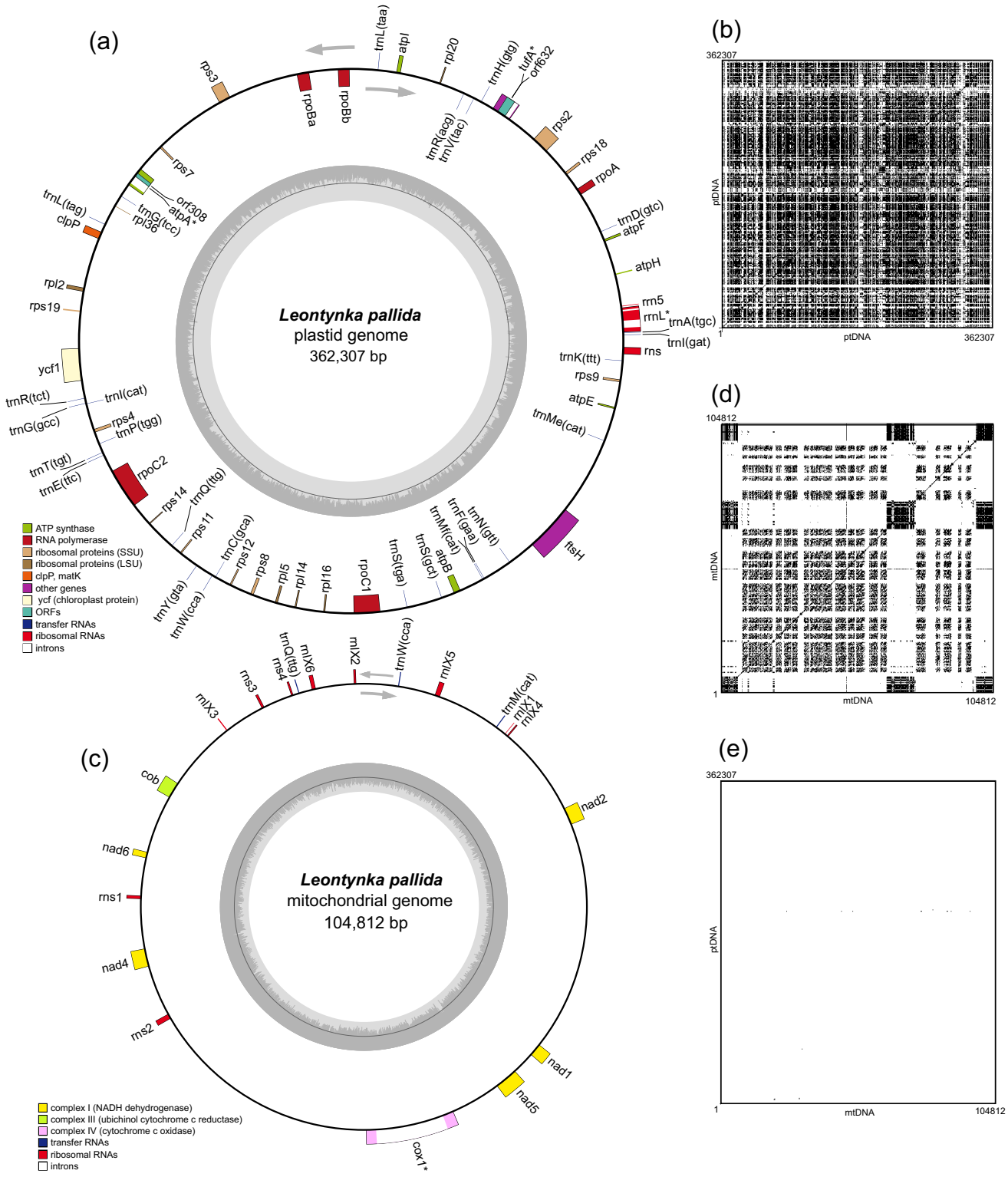
965 **Fig. 4** Distribution of repeats in organellar genomes of *Leontynka pallida*. (a) Most abundant
966 imperfect palindromes and their characteristics. The "Spacer" corresponds to the presumed loop
967 separating the palindromic regions presumably pairing to form a stem structure. The "Mismatch"
968 column indicates the number of positions that deviate from a perfect palindrome. The occurrence
969 of the repeats is given for the plastome (ptDNA) and mitogenome (mtDNA), with the number of
970 cases indicated for the whole organellar genome and separately for exons in protein-coding genes.
971 In two cases of mitogenome repeats, two variants – a shorter and a longer – are considered, with
972 the latter indicated in parentheses. (b) Distribution of the imperfect palindrome
973 AAGCCAGC|NNN|GCTGACTT and its most common variants within exons of the plastome. The

numbers show the abundance of the given repeat in direct / reverse complement orientation (relative to the coding sequence). In the case of "variant 1", the repeat has the same sequence in both directions, so only one number per gene is presented. Note that the variants considered are not mutually exclusive alternatives but correspond to nested categories with a different degree or relaxation of the sequence pattern. (c) Characterisation of repeats from (b) and their abundance in various regions of the plastome and the mitogenome of *L. pallida* as well as other plastomes of Chlamydomonadales deposited in NCBI databases. The numbers show the abundance of the given repeat in direct / reverse complement orientation (relative to the coding sequence in the case of exons, or relative to the DNA strand corresponding to the reference organellar sequence in case of the values for the whole organellar genome). (d) Occurrence of the "variant 8" repeat (translated in the reading frame +0 as KDKPANLTS) in a variable region of the ribosomal protein Rps8 (detail; the full alignment is available as Additional file 1: Fig. S8). (e) Occurrence of the "variant 4" repeat in protein-coding sequences and its translation in all six reading frames. The category of rare codons ("rare 2%") is defined as the sum of the least used codons together representing less than 2% of all codons in the plastome (100% = 19,899 codons); the categories of the 4%, 10%, and 20% rarest codons and more than 50% of most frequent codons are defined similarly (listed in Additional file 3: Table S5). The numbers indicated for the "codon usage" correspond to the minimal number of the codons of the respective category present in the respective reading frame, with the "max X" numbers indicating the maximal number of such codons depending on the actual nucleotide sequence of the degenerated "variant 4" repeat. Note that is some same the theoretical maximal number is not observed in the actual *L. pallida* plastid gene sequences (see the asterisks). The column "Rare AA" indicates the occurrence of amino acids belonging to the category of amino acids generally rarely used in plastome-encoded proteins in *L. pallida* (see Additional file 3: Table

S6). The occurrence of the repeat variants indicated for coding sequences (CDS) correspond to their occurrence as counted at the nucleotide level, whereas the occurrence in proteins is counted at the amino acid sequence level (and may be higher due to different nucleotide sequences encoding the same amino acid sequence). The analysis of intraexonic repeat insertions is commented in more detail in Additional file 2: Note S4.







(a)

Imperfect palindrome	Typical spacer	Spacer lenght	Mismatch	Occurence mtDNA	ptDNA	Exons
CCCGCCCC .... GGGGCGGG	TTTC	4-5	0	119	0	0
(G)AACCGTTC .... GAACCGTT(C)	CCCC	3-4	1	96 (80)	0	0
CCAGCCCCG ..... CGGGGCGGG	CCCCTTT	7	1	57	0	0
(AC)AACATGTT ..... AACATGTT(CT)	CTCTT	5	0 (1)	30 (9)	0	0
CAAACCAGT .. ACTGGTTAG	AA	2-3	1	0	1326	0
AAGCCAGC ... GCTGACTT	AAA	3	1	0	933	51
TATATTAGCT .. AGCTAATATA	AA	1-5	0	0	815	0
CTAGTGTTGCT . AGCAACACTAG	T	0-3	0	0	546	0
CTCGCTTGAACAT ATGTTCAAGCGAG	-	0	0	0	449	0

(b)

	atpI	*ftsH*	*rpoBa*	*rpoBb*	*rpoC1*	*rpoC2*	*rps2*	*rps3*	*rps8*	*rps9*	*ycf1*
IP (typical)	0/0	0/4	0/0	2/0	4/1	3/0	2/0	2/0	0/1	0/0	7/0
IP (general)	0/0	0/7	0/0	5/0	7/1	15/0	2/0	6/0	0/1	0/0	7/0
variant 1	1	15	2	1	9	22	4	1	0	0	4
variant 2	0/0	1/4	0/0	2/0	5/1	4/1	2/0	5/0	1/1	0/0	7/0
variant 3	0/0	4/1	0/0	0/1	4/0	1/0	1/1	1/0	0/0	1/0	1/0
variant 4	0/0	0/0	0/0	2/0	3/1	3/0	2/0	2/0	0/1	0/0	6/0
variant 5	1/0	10/0	0/2	0/0	3/0	14/1	1/0	0/0	0/0	0/0	1/0
variant 6	1/0	11/1	0/2	1/0	5/5	17/3	2/0	0/1	0/0	0/0	2/1
variant 7	1/0	12/4	0/3	5/0	9/2	31/1	4/1	7/0	0/1	0/0	10/0
variant 8	1/0	14/5	0/3	7/0	11/6	34/2	5/1	8/2	1/1	0/0	12/2

(c)

									Leontynka pallida AMAZONIE exons (+/-)	introns (+/-)	ptDNA (+/-)	mtDNA (+/-)	Other ptDNAs
IP (typical)			AAG	CCA	GCA	AAG	CTG	ACT	T	20/6	2/0	262/254	0/0	0/0
IP (general)			AAG	CCA	GCN	NNG	CTG	ACT	T	42/9	3/0	457/476	0/0	0/0
variant 1			AAG	TCA	GCN	NNG	CTG	ACT	T	59	1	142	0	0
variant 2			AA[AG]	CCA	GCA	AA[ACG]	C[CT][TG]	AC[AT]	T	27/7	2/0	270/261	0/0	0/0
variant 3			AA[AG]	TCA	GCA	AA[ACG]	C[CT][TG]	AC[AT]	T	13/3	0/0	15/13	0/0	0/0
variant 4	AAA	GAT	AAG	CCA	GCA	AAG	CTG	ACT	TC	18/2	2/0	187/196	0/0	0/0
variant 5	AAA	GAT	AAG	TCA	GCA	GAG	CTG	ACT	TC	30/3	0/1	40/52	0/0	0/0
variant 6	AAA	GAT	AAG	TCA	GCN	NNG	CTG	AC[AT]	TC	39/13	0/1	58/67	0/0	0/0
variant 7	AAA	GAT	AAG	[CT]CA	GCA	NNG	CT[TG]	ACT	TC	79/12	3/1	428/480	0/0	0/0
variant 8	AAA	GAT	AA[AG]	[CT][CT]A	GC[AGT]	NNN	[CT]T[TGA]	ACT	TC	93/22	3/1	444/502	0/0	0/0

(e)

RF	nucleotide sequence	AA sequence	codon usage rare 2%	rare 4%	rare 10%	rare 20%	frequent 50%	Rare AA	occurence CDS	proteins
+0	AAAGATAAGCCAGCAAAGCTGACTTCN	KDKPAKLTS	0 (max 1)*	0 (max 1)*	1 (max 2)	1 (max 2)	5 (max 6)	0	18	22
+1	NAAAGATAAGCCAGCAAAGCTGACTTC/N	XR*ASKADF	-	-	-	-	-	-	-	-
+2	NNAAAGATAAGCCAGCAAAGCTGACTT/CN	XKISQQS*L	-	-	-	-	-	-	-	-
-0	GAAGTCAGCTTTGCTGGCTTATCTTTN	EVSFAGLSX	0	0 (max 1)*	2 (max 3)	3 (max 4)	5 (max 6)	0	2	6
-1	NGAAGTCAGCTTTGCTGGCTTATCTTT/N	XSQLCWLIF	1 (max 2)	1 (max 2)	2 (max 3)	4 (max 5)	1	2	0	0
-2	NNGAAGTCAGCTTTGCTGGCTTATCTT/TN	XKSALLAYL	0 (max 1)	1 (max 2)	2 (max 3)	2 (max 3)	3 (max 4)	0	0	0

(*) these rare codons are missing in real data

(d)