

RESEARCH

Open Access



# The mitochondrial genome of the diploid oat *Avena longiglumis*

Qing Liu<sup>1,2,3\*</sup>, Hongyu Yuan<sup>1,2,4</sup>, Jiaxin Xu<sup>5</sup>, Dongli Cui<sup>1,2,4</sup>, Gui Xiong<sup>1,2,4</sup>, Trude Schwarzacher<sup>1,2,6</sup> and John Seymour Heslop-Harrison<sup>1,2,6\*</sup>

## Abstract

**Background** *Avena longiglumis* Durieu ( $2n = 2x = 14$ ) is a wild relative of cultivated oat (*Avena sativa*,  $2n = 6x = 42$ ) with good agronomic and nutritional traits. The plant mitochondrial genome has a complex organization and carries genetic traits of value in exploiting genetic resources, not least male sterility alleles used to generate F<sub>1</sub> hybrid seeds. Therefore, we aim to complement the chromosomal-level nuclear and chloroplast genome assemblies of *A. longiglumis* with the complete assembly of the mitochondrial genome (mitogenome) based on Illumina and ONT long reads, comparing its structure with Poaceae species.

**Results** The complete mitochondrial genome of *A. longiglumis* can be represented by one master circular genome being 548,445 bp long with a GC content of 44.05%. It can be represented by linear or circular DNA molecules (isoforms or contigs), with multiple alternative configurations mediated by long (4,100–31,235 bp) and medium (144–792 bp) size repeats. Thirty-five unique protein-coding genes, three unique rRNA genes, and 11 unique tRNA genes are identified. The mitogenome is rich in duplications (up to 233 kb long) and multiple tandem or simple sequence repeats, together accounting for more than 42.5% of the total length. We identify homologous sequences between the mitochondrial, plastid and nuclear genomes, including the exchange of eight plastid-derived tRNA genes, and nuclear-derived retroelement fragments. At least 85% of the mitogenome is duplicated in the *A. longiglumis* nuclear genome. We identify 269 RNA editing sites in mitochondrial protein-coding genes including stop codons truncating *ccmFC* transcripts.

**Conclusions** Comparative analysis with Poaceae species reveals the dynamic and ongoing evolutionary changes in mitochondrial genome structure and gene content. The complete mitochondrial genome of *A. longiglumis* completes the last link of the oat reference genome and lays the foundation for oat breeding and exploiting the biodiversity in the genus.

**Keywords** *Avena longiglumis*, Chloroplast, Genome recombination, Mitochondrial genome assembly, Mitogenome, Mitochondrial nuclear DNAs (MTNUs), Mitochondrial plastid DNAs (MTPTs), Organelle, Plants, RNA editing

\*Correspondence:

Qing Liu

liuqing@scib.ac.cn

John Seymour Heslop-Harrison

phh4@le.ac.uk; phh@molcyt.com

Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

As a widely grown temperate food crop, oats have long been favored by consumers because their grains are rich in nutrients such as protein, fat, vitamin B1, and  $\beta$ -glucan, as well as being a source of carbohydrates with low glycaemic index, dietary soluble fiber, different phenolic compounds, and minerals [1–4]. Oat is used as a high-quality feed for livestock with significant economic value [5].

Common oat (*Avena sativa* L.; abbreviated here ASA,  $2n=6x=42$ , genome composition AACCCD; Poaceae family) has a complex evolutionary origin, involving the separation of species from a common ancestor and then hybridization of an ancestral diploid and an ancestral tetraploid species, with further possible introgression from other ancestors [6]. *Avena longiglumis* Durieu (ALO,  $2n=2x=14$ , AA) is annual, grows primarily in the Mediterranean biome, and is considered a wild diploid ancestor to the common oat [7, 8]. As a crop wild relative, *A. longiglumis* has various agronomic traits, including stress resistance and nutritional traits, which are valuable to oat breeders.

Our extensive comparative analysis of the plastid (chloroplast) genomes of 13 *Avena* accessions identified the presence of rearrangement events in the plastomes [9], complementing other assemblies of oat plastomes [10, 11]. Due to the large genome size, presence of many repetitive DNA sequences, and complex ploidy in the oat nuclear genome, genomic studies lagged behind other gramineous crops, such as rice, maize and barley. It was not until 2022 that Peng et al. [12], Kamal et al. [13] and Liu et al. [14] published the reference nuclear genomes of cultivated oat and their diploid and tetraploid ancestors. Among them, the nuclear genome size of *A. longiglumis* was about 3.8 Gb.

In animals, mitochondrial genomes are mostly about 16 to 17 kb long, as a single circular assembly molecule with species-specific variation, making them extremely useful for phylogenetics [15, 16]. In contrast, angiosperm mitogenomes are much larger (typically 400 kb), with substantial variation in size, architecture, and extensive rearrangements [17] showing evidence of frequent sequence migration to plastid and nuclear genomes. Within species, there are with multiple complex configurations and this dynamic mitochondrial genome configurations has brought challenges to its assembly [18]. By October 2022, there were 525 plant mitogenomes in the NCBI organelle database (<https://www.ncbi.nlm.nih.gov/genome/organelle/>), fewer than nuclear genomes.

Grass (Poaceae) mitogenomes have been assembled from about 23 species including some cultivars from crops such as rice, barley and wheat, making Poaceae species an important group for comparative mitogenome

research [19–22]. The mitogenome can confer cytoplasmic male sterility [23, 24] with accessible restorer genes to enable seed multiplication: in rice [25] and maize [26], mitogenomes of several male sterile lines have been reported. The mitogenome is thus used for breeding and generating  $F_1$  hybrid seeds.

There is extensive sequence migration between the cellular genomes in higher plants during their evolution and divergence. Specifically, the mitogenome includes abundant foreign sequences, including mitochondrial plastid DNAs (MPTs) and mitochondrial nuclear DNAs (MTNUs) and retroelement fragments. We aim to generate an assembly of the *A. longiglumis* mitochondrial genome using a hybrid strategy with long, single-molecule sequencing to allow the resolution of alternative, large and complex structures [17, 27] and correction by deep sequencing and high-quality short reads. We then aim to characterize gene and repeat content, and the presence of sequence exchanges between plastomes and nuclear genomes, putting features into an evolutionary and phylogenetic context.

## Materials and methods

### Plant material and genome sequencing

Plants of *Avena longiglumis* Durieu (ALO) (accession PI 657,387; US Department of Agriculture, Beltsville, <https://www.ars-grin.gov/>), collected from the wild in Morocco in 1989 c. 500 m from the Mediterranean coast near Moulay Bouselham) was used for genome sequencing. After growing in South China Botanical Garden Greenhouse at 25 °C, 16 h light/8 h dark with 70% relative humidity for four weeks, plants were moved outside under natural daylight conditions (dry season in Guangzhou). Flowering specimens were identified by Qing Liu and vouchers were deposited in the South China Botanical Garden Herbarium (code IBSC).

Genomic DNA for Illumina mate-pair sequencing was extracted using the DNeasy Plant Mini Kit (Qiagen) from the 8-week-old leaves of ALO seedlings. A DNA library with an insert size of 350 bp was constructed following the manufacturer's protocol [28], and was sequenced using the Illumina NovaSeq 6000 platform, giving 201.9 Gb (gigabases) of raw data from 673,011,605 raw reads. Clean data were obtained by using Trimmomatic [29] removing low-quality sequences, defined by a quality value of  $Q < 19$  for more than 50% of the total bases or sequences in which more than 5% of bases were "N".

For ONT PromethION library construction and long read sequencing, genomic DNA was extracted from 3-week-old leaves of ALO seedlings using the Qiagen 695 Genomic DNA Extraction Kit. The DNA libraries were quantified using a Qubit 3.0 Fluorometer (Cat. E33216, Invitrogen, USA) and loaded into 12 lanes of a

PromethION, R9.4.1 flow cell (Oxford Nanopore Technologies, UK) for SMRT (single molecular real-time) sequencing. In total, 49.6 Gb of sequences in 2,131,575 reads were obtained.

### Plastid and mitochondrial genome assembly

For plastome assembly, the Illumina paired-end reads were assembled into contigs using the GetOrganelle v.1.7.3 pipeline [30] with the parameters of '-R 15 -k 21,45,65,85,105 -F embplant\_pt' recommended by the developer. We initially de novo assembled the mitogenome of *A. longiglumis* using the Nanopore raw long-reads with Flye v.2.9-b1774 [31] with the default settings. GFA format files from the assembly results of Flye were visualized using Bandage [32]. For all assembled units, we used makeblastdb to build the database of *A. longiglumis*. Subsequently, mitochondrial genes extracted from *Oryza sativa* indica Group (NC\_007886.1) mitogenome were used as query sequences to identify the unitig graph containing mitochondrial genes and sequences were compared with the nucleotide database [33] (see below for annotation details).

Considering that there are regions of mitogenome that are homologous to chloroplast sequences, it is likely that they are replaced by their chloroplast counterparts during polishing. Therefore, we mapped the Illumina short-reads to the mitogenome assembly by Nanopore long-reads using BWA and SAMtools v.0.1.19 [34, 35], and all unmapped reads were excluded. BEDTools [36] were used to transform BAM files containing mapped reads into FASTQ format files. We obtained about 52 Mb of mitochondrial reads from 201.9 Gb raw data, which also contained chloroplast homologous sequences. We used the same processing strategy of the chloroplast sequences to screen the Nanopore long-reads to obtain mitochondrial reads and exclude nuclear and plastid reads with Minimap2 [37].

Subsequently, all processed long-reads and short-reads were used as input files for hybrid assembly by using Unicycler [38] with default settings. Unicycler will call Spades [39] to assemble Illumina short-reads, and repetitive DNA sequences encountered during assembly will be resolved using long-reads. Then, four circular contigs (isoforms) were generated, representing the complete mitochondrial genome of *A. longiglumis* with a graph showing alternative configurations. We found that the four contigs have long repeated sequences within each other. Finally, a master circular genome was generated by merging them manually.

### Genome annotation

CPGAVAS2 [40] was used for the annotation of *A. longiglumis* plastome with the third option of a custom

reference, using the previously released GenBank format file of *A. longiglumis* plastome (NCBI accession number MK336391.1 [9]).

We first downloaded RNA-seq data from NCBI SRA database with the accession numbers PRJNA735431 [12] and PRJNA838431 [14]. For the annotation of protein-coding genes (PCGs) and rRNA genes in the mitogenome, we used GeSeq [41] to annotate the assembled mitogenome of *A. longiglumis* with two reference mitogenomes from GenBank: *Oryza sativa* (NC\_007886.1) and *Liriodendron tulipifera* (NC\_021152.1). Subsequently, the web-based tool public MITOFY [42] was used to identify genes. Annotation was manually edited by Apollo [43]. All transfer RNA genes were confirmed by tRNAscan-SE with default settings [44] and compared with studied tRNA genes previously [45]. Based on sequence similarity, we identified tRNA genes that migrated from plastome and tRNA genes that were native to mitochondria. We did not include those tRNA genes with no homology in previous studies. Finally, we drew the genome map using OGDRAW [46].

### Homologous sequences between mitochondrial, plastid and nuclear genomes

To identify DNA fragments that may be transferred to the mitochondria from the plastid and nuclear genomes, we compared the three genomes using makeblastdb and BLASTn program [33] with the following parameters: '-evalue 1e-5, -word\_size 9, -gapopen 5, -gapextend 2, -reward 2, -penalty -3, and -dust no'. The plastome assembled here (based on long-molecule sequences), and the nuclear genome of *A. longiglumis* downloaded from NCBI with the accession number of CM042677.1-CM042683.1 [12], OU342747.1-OU342753.1 [13] and SRR19279532-SRR19279533 [14] were used as reference in this study. For MTNUs, we only keep results with length longer than 1,000 bp.

### Analysis of repeat elements

Simple sequence repeats (SSRs) were identified using the web-application MISA [47] (<https://webblast.ipk-gatersleben.de/misa/>) for the assembled mitogenome, with parameters of the minimum numbers of mono-, di-, tri-, tetra-, penta-, and hexanucleotides set as 10, 5, 4, 3, 3, and 3, respectively. Additionally, forward, reverse, palindromic, and complementary repeat sequences were identified by REPuter (<https://bibiserv.cebitec.uni-bielefeld.de/reputer/>) with the following settings: hamming distance of three and minimal repeat size of 30 bp [48], and *e*-value is limited to less than 1e-05. Tandem Repeats Finder [49] (<https://tandem.bu>

[edu/trf/trf.html](http://edu/trf/trf.html)) is used to detect the tandem repeats of the mitogenome.

#### Identification of RNA editing sites in protein-coding region of mitochondrial genes

Two published RNA-seq datasets in the SRA database (<https://www.ncbi.nlm.nih.gov/sra/>; accession numbers SRR19216412 to SRR19216415 and SRR14760587 to SRR14760588) [12, 14] were combined to identify the RNA editing sites in mitochondrial PCGs. The RNA-seq data were mapped onto sequences of PCGs by using Bowtie2 [50] with the parameters: '-f -a -m 20 -al Reads\_aligned -un Reads\_unaligned'. We performed three runs with the mismatched numbers set to 3, 5, and 7, in order to ensure the excessive RNA editing events detected by enough mapping results; The anomalies results were checked by Tablet [51]. We used BCFtools [52] to call single nucleotide polymorphism sites (SNPs), which were considered as RNA editing sites. A standard that each RNA editing site was set to cover by at least 20 reads, and editing events had to occur in at least ten reads.

#### Phylogenetic inference

The mitogenomes of 28 Poaceae accessions of 23 species downloaded from NCBI were used to construct a phylogenetic tree, with two other monocotyledonous species, *Cyperus esculentus* (Cyperaceae) and *Phoenix dactylifera* (Arecaceae) as outgroups (Additional file 2: Table S1). To ensure comparisons were valid, these mitogenomes were re-annotated using former described tools. A total of 30 orthologous mitochondrial genes among the analyzed species were identified and extracted by PhyloSuite v.1.2.1 [53]. The corresponding nucleotide sequences were aligned by MAFFT v.7.450 [54]. Next, these aligned sequences were concatenated and used to construct the phylogenetic trees. The consensus sequences included 31,510 nucleotide sites. The maximum likelihood (ML) method implemented in RAxML v.8.2.4 [55]. The parameters were "raxmlHPC-PTHREADS-SSE3 -f a -N 1000 -m GTRGAMMA -x 551,314,260 -p 551,314,260". The bootstrap analysis was performed with 1,000 replicates. Bayesian inferences (BI) analysis was performed by MrBayes v.3.2.6 [56] with the Markov Chain Monte Carlo method for 200,000 generations and sampling trees every 100 generations. The first 20% of trees discarded as burn-in, the remaining trees were used to generate a consensus tree.

## Results

#### Mitochondrial genome structure of *Avena longiglumis*

Based on the hybrid assembly of Unicycler using Illumina paired short reads and ONT PromethION long reads, a graphical representation of the *A. longiglumis*

mitogenome is obtained. It contains 36 unitigs (high-confidence sub-contigs, Additional file 1: Figure S1), which show duplications and the alternative tiling paths. Length and sequencing depth of each unitig are given in Additional file 2: Table S2).

There are long repetitive sequences between the contigs which may potentially generate the multiple isoforms via genome recombination (Additional file 1: Figure S1, Additional file 2: Table S2). The alternative configurations (isoforms) are supported by Oxford Nanopore long reads. In order to describe the mitogenome succinctly, we simplify the four contigs into a master circular genome based on the shared repeats. The detailed solution and schematic diagram can be found in Additional file 1: Figure S2.

The total length of the mitogenome of *A. longiglumis* is 548,445 bp during the reconstruction of a single circle mediated by repeats. The GC content is 44.05%. Accuracy of the mitochondrial genome bases was confirmed by mapping the Illumina paired short read raw data (average 56-fold coverage) onto the assembled reference mitogenome (Additional file 1: Figure S3).

#### Gene content and repeat elements of mitochondria

We identify a total of 65 mitochondrial native genes (Table 1, Additional file 2: Table S3), including 39 mitochondrial native protein-coding genes (35 are unique), eight mitochondrial native rRNA genes (three are unique), and 18 mitochondrial native tRNA genes (11 are unique). Furthermore, there are ten complete plastid-derived genes, including one protein-coding gene (*rpl2*) and nine tRNA genes (eight are unique and there are two copies of *trnC-GCA*). The multiple dispersed and tandemly repetitive DNA elements are characterised (Additional file 2: Tables S4, S5 and S6).

The protein-coding genes (PCGs) (Table 1, Fig. 1) include 24 unique core mitochondrial genes: five subunits of ATP synthase, nine subunits of NADH dehydrogenase, four cytochrome C biogenesis genes, three cytochrome C oxidase genes, and one transport membrane protein, maturase, and ubiquinol cytochrome C reductase, respectively. Ten variable mitochondrial genes are identified, including two large subunits of ribosome proteins and eight small subunits of ribosome proteins. Among these, *atp1* has three copies, *atp4* and *rps4* have two copies of each one, and there are three rRNA genes (with 3 copies of *rrn5* and two copies *rrn18* and *rrn26* of each one). A total of 18 unique (total 27) tRNA genes are identified; six tRNA genes have multiple copies, e.g. *trnC-GCA* and *trnD-GUC* have two copies for each one (Table 1, Additional file 2: Table S3).

We find 103 simple sequence repeats (SSRs) in the mitogenome (Fig. 2A, Additional file 2: Table S4).

**Table 1** Gene composition of the mitogenome of *A. longiglumis*. For lengths and positions of genes see Additional file 2: Table S3

	Group of genes	Genes (Number of copies in parentheses if > 1)
Protein coding genes (PCGs)	ATP synthase	<i>atp1</i> (× 3), <i>atp4</i> (× 2), <i>atp6</i> , <i>atp8</i> , <i>atp9</i>
	NADH dehydrogenase	<i>nad1</i> , <i>nad2</i> , <i>nad3</i> , <i>nad4</i> , <i>nad4L</i> , <i>nad5</i> , <i>nad6</i> , <i>nad7</i> , <i>nad9</i>
	Cytochrome c biogenesis	<i>cob</i>
	Ubiquinol cytochrome c reductase	<i>ccmB</i> , <i>ccmC</i> , <i>ccmFC</i> , <i>ccmFN</i>
	Cytochrome c oxidase	<i>cox1</i> , <i>cox2</i> , <i>cox3</i>
	Maturases	<i>matR</i>
	Transport membrane protein	<i>mttB</i>
	Large subunit of ribosome	<i>rpl5</i> , <i>rpl16</i> , <i>rpl2</i> (cp)
	Small subunit of ribosome	<i>rps1</i> , <i>rps2</i> , <i>rps3</i> , <i>rps4</i> (× 2), <i>rps7</i> , <i>rps12</i> , <i>rps13</i> , <i>rps14</i>
	Ribosomal RNAs	<i>rrn5</i> (× 3), <i>rrn18</i> (× 3), <i>rrn26</i> (× 2)
Transfer RNAs	<i>trnC-GCA</i> (× 2, cp), <i>trnD-GUC</i> (× 2), <i>trnE-UUC</i> (× 3), <i>trnF-GAA</i> (cp), <i>trnFM-CAU</i> (× 4), <i>trnH-GUG</i> (cp), <i>trnI-CAU</i> , <i>trnK-UUU</i> , <i>trnM-CAU</i> (cp), <i>trnN-GUU</i> (cp), <i>trnP-UGG</i> (× 2), <i>trnP-UGG</i> (cp), <i>trnQ-UUG</i> , <i>trnS-GCU</i> , <i>trnS-GGA</i> (cp), <i>trnS-UGA</i> , <i>trnV-GAC</i> , <i>trnW-CCA</i> (cp), <i>trnY-GUA</i>	

The numbers in parentheses represent copy numbers of genes, and 'cp' represent the plastid-derived genes

Tetrameric repeats accounted for 38% of the SSRs (39 out of 103) compared to 20 or fewer for monomeric, dimeric and trimeric repeats; or 5 or fewer for hexamers and pentamers. For these SSRs, there are 18 stretches of the mononucleotide A/T, respectively. In addition, there are 5 dimers of TA and 8 tetramers of AATG (Fig. 2B).

We detect 19 long tandem repeats (> 11 bp repeat unit) (Fig. 2C, Additional file 2: Table S5). Duplicated non-coding sequences (≤ 30 bp) are more frequent than SSRs and tandem repeats: a total of 520 duplicated sequences are detected, among which 52.69% (274) are palindromic repeats, and 47.31% (246) are forward repeats (Additional file 2: Table S6), accounting for 42.5% (233,082 bp, elimination of overlapping region) of the 548 kb mitogenome of *A. longiglumis*. Most of these duplicated sequences (93%) are less than 100 bp in length, but three are much longer (R1 is 31,235 bp, R2 is 16,549 bp, and R3 is 14,514 bp), and five (R4-R7) between 5 and 10 kb long (Fig. 3, Additional file 2: Table S6).

#### Sequence exchange between genomes: Identification of MTPTs and MTNUs

We identify homologous sequences between the *A. longiglumis* mitogenome and the plastid and nuclear genomes. For MTPTs, the results show 22 homologous fragments between mitochondrial and plastid genomes (Fig. 4), the longest being 1,971 bp (MTPT5), followed by 1,125 bp (MTPT19). The total length of these fragments is 8,207 bp. These fragments contained some genes, including nine complete plastid-origin tRNA genes (*trnF-GAA*, *trnP-UGG*, *trnS-GGA*, *trnC-GCA*, *trnW-CCA*, *trnH-GUG*, *trnM-CAU*, *trnN-GUU* and *rpl2*), and nine plastid gene fragments, like *ndhA*, *ndhK*, *ndhJ*, *rpl14*, *atpA*,

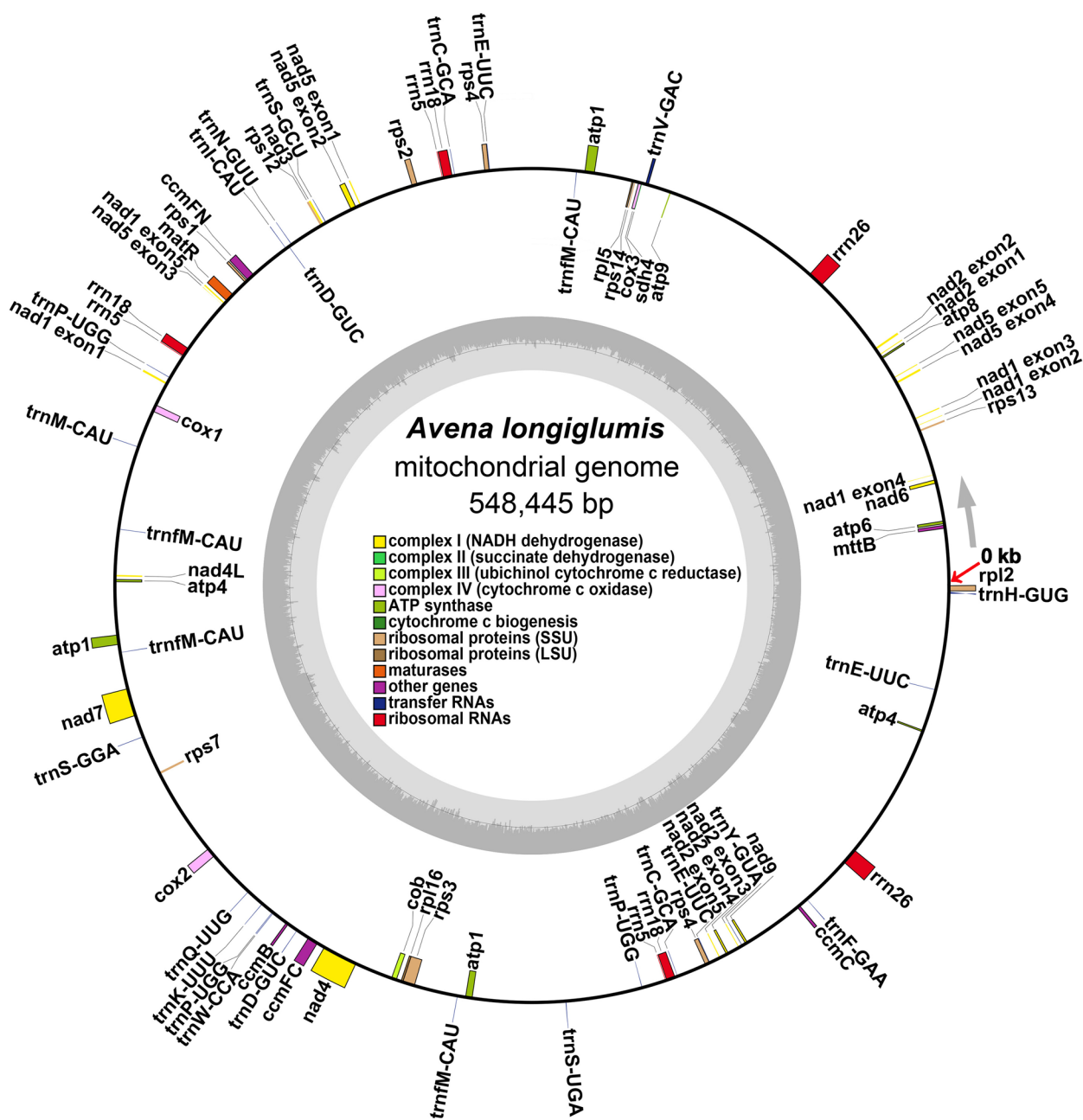
*rps19*, *rpl23*, *rps7* and *psbD* (Additional file 2: Table S7). Furthermore, a total of 16 sequence fragments of conserved protein domains (reverse transcriptase, RNaseH and integrase) of *copia* and *gypsy* nuclear retroelements are detected in the *A. longiglumis* mitogenome.

We detect a total of 502,359 bp (91.60%), 468,212 bp (85.37%) and 456,990 bp (85.32%) MTNUs in three nuclear genome assemblies of *A. longiglumis* [12–14], respectively (Additional file 2: Table S8). Most of the mitochondrial fragments are 1–2 kb long and only few more than 50 kb; they are dispersed throughout all seven chromosome assemblies of the nuclear genome.

#### RNA editing and gene structures

We identify the RNA editing sites of mitogenome-encoded protein coding genes. The results show that a total of 25 mRNAs of PCGs is identified RNA editing events, and all these edited sites are C to U editing. For these 25 genes, we identify 269 high-quality RNA editing sites (Fig. 5A, Additional file 2: Table S9; *nad2* has 26 editing sites, *nad1* and *nad4* have 23 of each one, *nad3* has 21 and *nad7* had 20).

The 269 editing sites involves 258 codons and resulted in 242 amino acid changes (Fig. 5B). Three edits alone accounted for more than 63% of the amino acid changes: serine replacement with leucine occurred 59 times, while substitution of proline with leucine and serine with phenylalanine are identified 51 and 44 times respectively (Fig. 5B). Though supported by relatively low coverage sequencing reads, we find RNA editing events for the *ccmFC* gene to create stop codon (Additional file 1: Figure S4A). By comparing it with the other two species, *Arabidopsis thaliana* and *Nicotiana tabacum*, we suggest

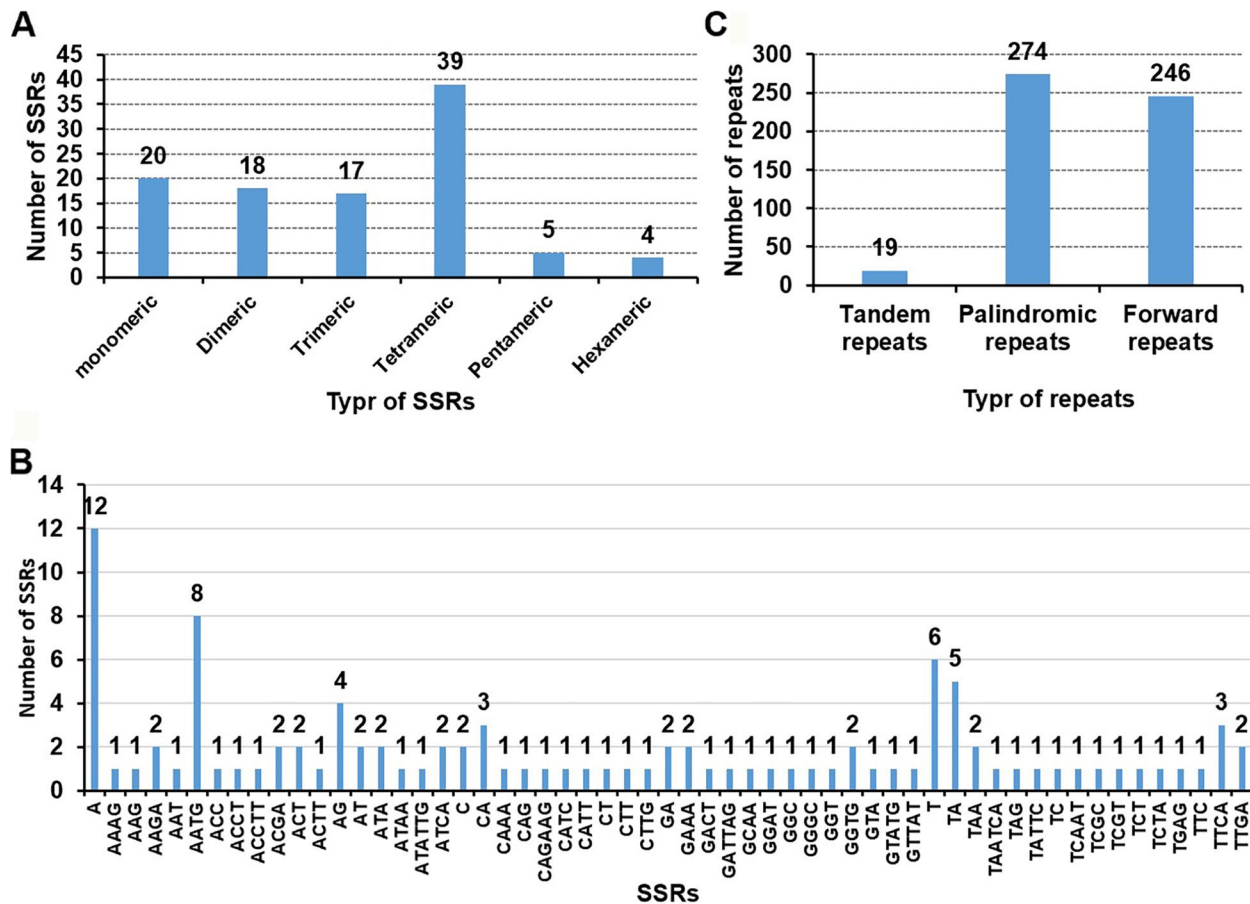


**Fig. 1** The circular mitogenome maps of *A. longiglumis*. The gene contents (see Table 1) of mitogenome with the starting point marked by red arrow and the direction of positive strand marked by the gray arrow. Genes transcribed clockwise and counter-clockwise are drawn on the inside and outside of the circles (location details see Additional file 2: Table S3). Genes belonging to different functional groups are color-coded

that this RNA editing might be meaningful to maintain the conserved length of the protein (Additional file 1: Figure S4B).

Eight of the 34 protein coding genes harbour introns: four genes contain one intron (two copies of *atp1* gene, *ccmFC* and *cox2*), one gene contains three introns (*nad4*), and four genes contain four introns (*nad1*, *nad2*, *nad5* and *nad7*). The *atp1* gene has two different gene

structures; one copy has no intron, while the other two copies contain one intron (Fig. 5C). However, the genuine identity of this intron needs to be treated cautiously since since introns have not been reported for the *atp1* gene in angiosperms. The insertion of this unknown intron sequence might be associated with a 39 bp small repeat sequence, perhaps introduced during genome repair.



**Fig. 2** The simple sequence repeats (SSRs), tandem repeats, and dispersed repeats identified in the mitogenomes of *A. longiglumis*. **A** The identified SSRs on the reference mitogenomes of *A. longiglumis*. Each column represents different nucleotide repeat units. The numbers of repeats in each category are shown on the top of corresponding columns (see Additional file 2: Table S4). **B** The type and number of SSRs. **C** Tandem repeats (with repeat unit  $\geq 12$  bp, see Additional file 2: Table S5) and dispersed repeats ( $\geq 30$  bp, see Additional file 2: Table S6) identified on four contigs

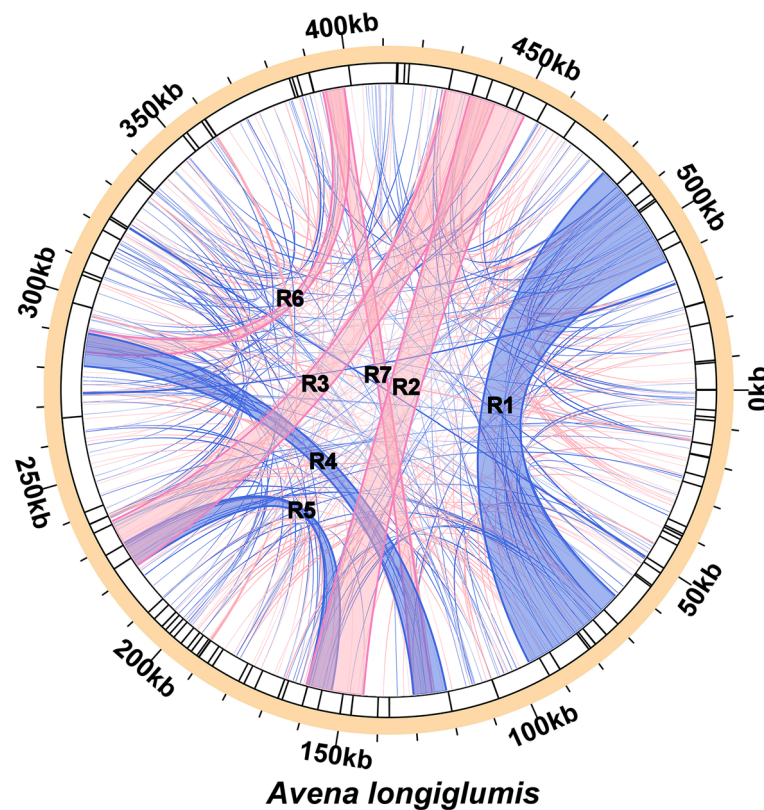
### Phylogenomic analysis

Phylogenetic analyses based on the mitochondrial genomes yield a phylogeny of Poaceae species with eight nodes having Maximum Likelihood (ML) bootstrap support values  $< 99\%$  or Bayesian Inference (BI) posterior probabilities  $< 1$  (Fig. 6A). Given the paucity of mitogenome data, the analysis presented here is based only on the available Poaceae mitogenomes (for just 24 species, Additional file 2: Table S1) of the more than 11,000 species in Poaceae family [57].

We compare gene content of mitogenomes of *A. longiglumis* with other 28 mitogenome accessions of 23 Poaceae species. Three unique rRNA genes and 24 core mitochondrial genes [58] are present in all the analyzed Poaceae species. For the tRNA genes, we examine the origin based on sequence similarity (Fig. 6B; see detail in Additional file 2: Table S10) between genomes and species [45]. Five of 17 tRNA genes originating from plastids (*trn\_cp*) presented in

most mitogenomes of Poaceae: *trnC-GCA\_cp*, *trnF-GAA\_cp*, *trnM-CAU\_cp*, *trnN-GUU\_cp* and *trnW-CCA\_cp* are identified in at least 27 out of 29 mitochondrial genomes. In contrast, plastid-derived *trnA-UGC\_cp*, *trnM-CAU\_cp*, *trnL-CAA\_cp*, *trnL-UAG\_cp* and *trnV-UAC\_cp* are found in four or fewer taxa. A total of eleven mitochondrial native tRNA genes are identified. Ten are present in most species, but *trnF-GAA\_mt* was detected in only *Sporobolus alterniflorus* (syn. *Spartina alterniflora*). Four mitochondrial native tRNA genes identified in other angiosperms are lost in Poaceae, *trnC-GCA*, *trnG-GCC*, *trnV-UAC* and *trnW-CCA*.

For the variable PCGs (large and small ribosomal proteins), eight are shared among 22 or more of the 24 species. Three frequent mitochondrial genes are unidentified in the analyzed species, including *rps10*, *sdh3* and *sdh4*. Genes have lost to varying degrees in the various species and many gene losses evidently occurred several times in the grass phylogeny.



**Fig. 3** The distribution of dispersed repeats among the mitogenome of *A. longiglumis*. Arcs connect similar repeats within the mitogenome; blue arcs represent 246 forward repeats and the red represents 274 palindromic repeats (see Additional file 2: Table S6). Seven long repeats over 5,000 bp in length are labelled

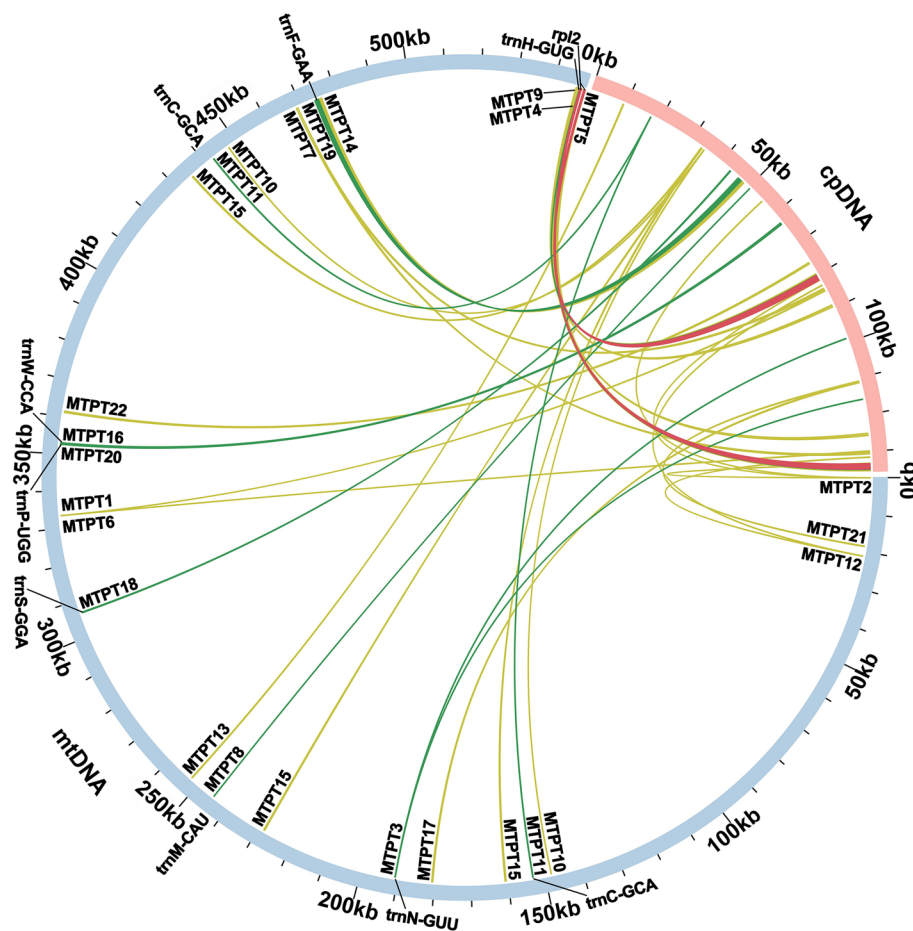
## Discussion

We present a complete sequence assembly, 548,445 bases long, of the *Avena longiglumis* mitochondrial genome (Fig. 1) and analyze repeats and the gene content in a phylogenetic context. However, the multi-isoform (multi-chromosomal) architecture is highly possible, because Unicycler obtained four independent and circular contigs based on resolving repeats using long-reads (Additional file 1: Figure S1). The results are consistent with our understanding of the mitogenomes of higher plants (especially angiosperms), showing possible rearrangements and genome recombination [59, 60]. The size is within the range reported in plant species, although longer than the average for land plants of 394,900 bases [61]. Kozik et al. [17] suggest that the plant mitogenomes should be presented as multiple sequence units showing their variable and dynamic connections rather than as circles. In our case, the complex assembly configurations, which are the result of recombination between repeats, are shown in Additional file 1: Figure S1A and S1B. For *A. longiglumis* mitogenome, we suggest that seven pairs of long dispersed repeats (R1 to R7, Fig. 3) are the key for

the mitogenome recombination. Dispersed repeats have been reported to mediate the recombination of the mitochondrial genome resulting in various alternative configurations (isomers) [62, 63], with data from various plants including *Nymphaea colorata* [64], *Silene latifolia* [41], *Mimulus guttatus* [65], and *Ginkgo biloba* [66]. In the *Arabidopsis thaliana* mitogenome, two large repeats (6.0 and 4.2 kb in size in the 368 kb genome) are constantly recombining and lead to multiple alternative structures [67].

In Poaceae species, there is only limited research with relatively few reports of mitochondrial genomes and their complexity [68]. Some assemblies have reported one, or several, circular contigs describing the whole mitogenome, enabling the data to be used to address key questions about gene content and gene polymorphisms, but not repeats nor genome structure. Sugarcane, for example, has two circular contigs as the complete mitogenome [68]. As we find in *A. longiglumis*, rearrangements involving inversion, translocation, fusion, and fission have also been shown in mitochondrial genomes of rice, with various frequencies of alternatives [69]. For





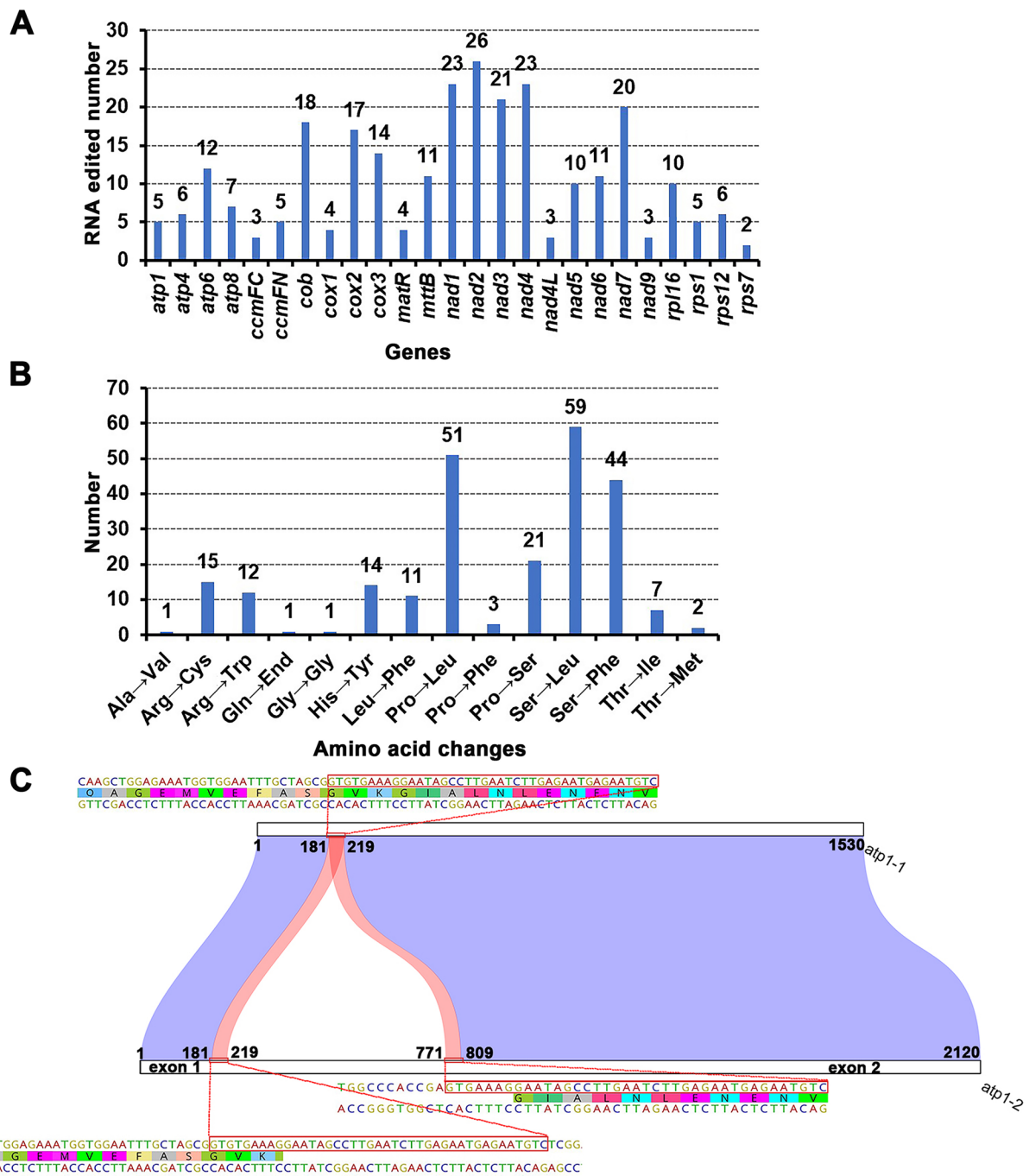
**Fig. 4** Schematic representation of the distribution of homologous sequences between mitogenome and plastome of *A. longiglumis*. Mitochondrial plastid DNAs (MTPTs) on the mitogenome are few and short in length. The outer circle of blue and pink section represents the mitogenome (mtDNA) and the plastome (cpDNA), respectively. The scale is 10 kb, marked on the genome. The yellow-green arcs in the middle represent the 22 MTPTs identified (please see Additional file 2: Table S7). We highlight the arcs in dark green and red if MTPT contained complete tRNA genes and protein-coding gene, respectively. They represent genes that migrated from the plastome to the mitogenome, including nine tRNA genes (eight are unique and there are two copies of *trnC-GCA*) and one protein-coding gene (*rpl2*). The homologous sequences located in the IR region of plastome was counted only once

most cultivated grasses, the complex and polyploid large nuclear genome, and low copy number of mitochondria, pose a challenge to the amount of data being able to be sequenced [19, 21]. Further long-molecule sequence coverage and PCR amplification across recombination breakpoints marking alternative configurations, possibly with enrichment of mitochondria based on differential centrifugation and subsequent DNA extraction for sequencing (as often used in human cell and tissue [70]), may be helpful.

In detailed analysis of sequence data, we find evidence for multiple RNA edits (Fig. 5). Notably, creation of stop codons via RNA editing events was in a few transcripts of *ccmFC* gene, with two of three reads supporting an edit to add a stop codon in the gene (Additional file 1: Figure

S3). This result is consistent with previous research with *ccmFC* [71], and editing maintaining the stability of the length of the protein product encoded by the *ccmFC* gene.

Exchange of sequence between mitochondrial genomes and those of the plastid—MTPTs— and nucleus—MTNUs—are ubiquitous in seed plants [72]. Studies of MTPTs in *A. longiglumis* here found relatively few plastid homologous sequences, representing only 8,207 bp or 1.5% of the mitogenome. This result shows sequence migration between these two organelle genomes. Large segments of MTPTs have been reported in many mitochondrial genomes. A total of 26.87 kb MTPTs were found in *Suaeda glauca* [73], accounting for 5.18% of its mitogenome. The large MTPTs are also found in *Salix*



**Fig. 5** Characteristics of the RNA editing sites identified in protein-coding genes (PCGs) and schematic structure of *atp1* gene of the *A. longiglumis* mitogenome. **A** Frequency of RNA editing sites identified in each PCGs. **B** Frequency of amino acid changes caused by RNA editing. **C** Schematic diagram of the *atp1* gene structures

*suchowensis* [74]. The MTPTs were thought to broadly impact eukaryotic evolution and promote genetic diversity [75]. Although the MTPTs were shorter in length in *A. longiglumis*, some plastid genes are still transferred

intact from the plastome to the mitogenome. There are eight tRNA genes. It is difficult to determine whether these plastid-derived tRNA genes have biological functions. Previous studies have shown that *trnH-GUG*

(cp) and *trnM-CAU* (cp) might still be functional in plant mitogenomes, and they migrated into the mitogenome early in the evolution of the species [41, 76]. In contrast, neither mitochondrial native *trnM-CAU* nor *trnH-GUG* has been reported in higher plants. Another three migrated tRNA genes, *trnN-GUU* (cp), *trnP-UUG* (cp), and *trnW-CCA* (cp), possibly also are functional, as reported by Richardson et al. [77]. The *trnF-GAA* (cp) and *trnC-GCA* (cp), although sporadically migrated to the mitogenome in other angiosperms, were detected in almost all Poaceae species. These results suggest that the migration of these two plastid-origin tRNA genes into mitogenome is universal in Poaceae, with only a few exceptions. Moreover, *trnS-GGA* (cp) was found only in *A. longiglumis* and some Poaceae plants. As for some plastid-derived tRNA genes found in other Poaceae species, such as *trnM-CAU* (cp) and *trnA-UGC* (cp) [78, 79]. Their migration is sporadic rather than universal, just like *trnS-GGA* (cp). They possibly have migrated along with the large plastid fragments recently, and this event is not universally shared among the family Poaceae. As for the variable PCGs, we found that only six of them are conserved in Poaceae. Besides this, two succinate dehydrogenases (*sdh3* and *sdh4*) are lost in all Poaceae species, which have been reported to transfer to the nuclear genome [80]. This result is a general characteristic of these taxa.

We identify large fragments of mitochondrial sequences in two sets of nuclear genomes, and find 85 to 92% of the mitochondrial sequence in the nuclear genome assemblies of *A. longiglumis* (Additional file 2: Table S8). The results are compatible with those from high-quality assemblies of *Arabidopsis thaliana* [67], although in *Arabidopsis*, the careful reassembly of the nuclear genome shows there is a single major insertion of a structurally complex MTNU representing more than one copy of most of the mitochondrial genome. Here, for *A. longiglumis*, we find many MTNUs of smaller size distributed over all nuclear chromosomes.

We detect multiple retroelement domains within the mitochondrial sequence of *A. longiglumis* (although no long stretches representing complete elements or even open reading frames, ORFs) showing the transfer

of fragments from the nucleus. We speculate that the transferred sequence fragments may originate from episomal DNA copies of retroelements that have been reverse transcribed from RNA transcripts present in the cell and integrated via recombination of party homologous sequences (see Richert-Pöeggeler et al. [81]) contrasting with the MTNUs originating directly from the nuclear genome. In many species including *Arabidopsis*, numerous fragments of retrotransposons have been found integrated into the mitochondrial genome [82].

Within the family Poaceae, phylogenetic inference based on mitochondrial sequence places *A. longiglumis* as a sister to *Lolium perenne* and the tribe Triticeae (Fig. 6A). However, subfamily Pooideae (including *A. longiglumis*) is placed only as weakly resolved sister to Bambusoideae (67% bootstrap support), and the subfamily Oryzoideae is a weakly resolved (85% bootstrap support) sister to Chloridoideae, and Panicoideae. Nuclear [83] and plastid [84] phylogenies, and indeed morphological studies, universally show strong support for two major grass clades, BOP (Bambusoideae, Oryzoideae, Pooideae) and PACMAD (Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae, and Danthonioideae; including the C<sub>4</sub> photosynthesis grasses) in the family Poaceae. Neither the mitochondrial sequence gene phylogeny (Fig. 6A), nor shared losses or gains of non-core genes (Fig. 6B), support separation of the clades, despite their being universally accepted as a natural division. Thus, our tree shows why, in contrast to animals, plant mitochondrial sequences are not usable in routine phylogenetic investigations. More mitochondrial assemblies are required for detailed analysis of the evolutionary and phylogenetic implications (including consideration of reasons for the low bootstrap or MI support for deep branches) for this organelle, including also basal grass genera such as *Anomochloa*.

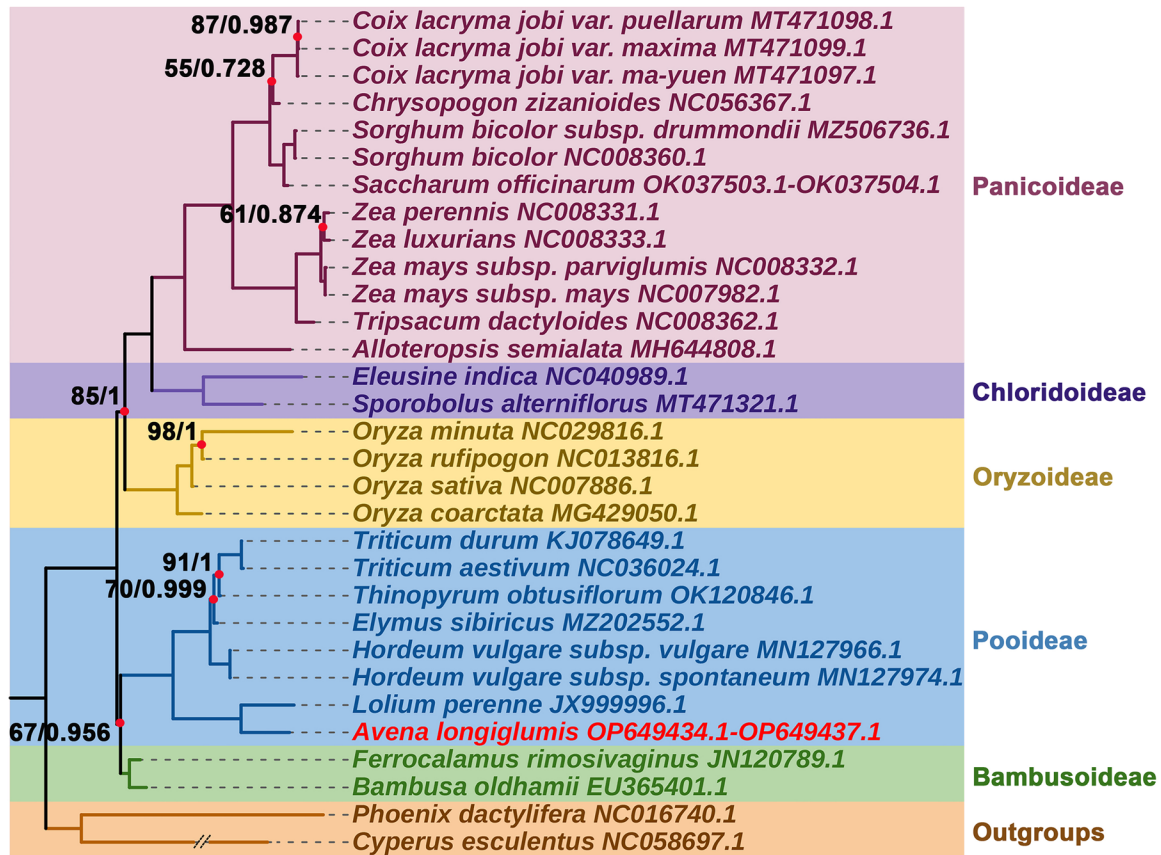
## Conclusions

Our reference mitogenome of *A. longiglumis* shows its complex structure and repeat-rich organization based on long-read sequencing, with important features of

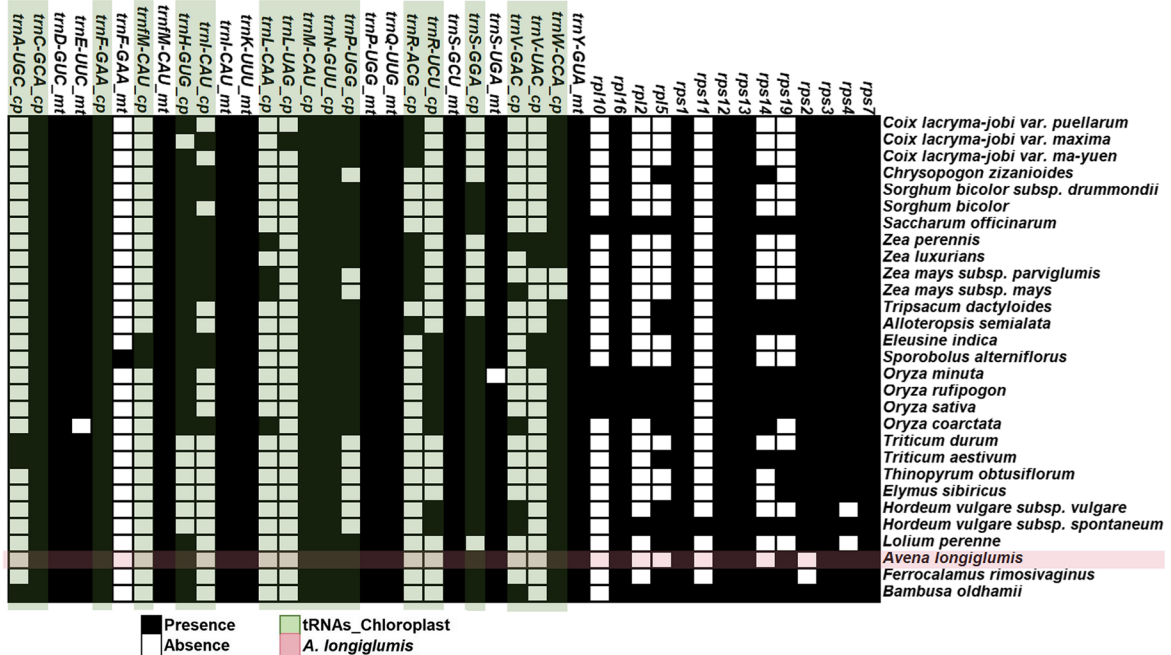
(See figure on next page.)

**Fig. 6** The phylogenetic relationships and gene distribution in mitogenomes of *A. longiglumis* compared with 23 Poaceae species. **A** Maximum Likelihood (ML) and Bayesian Inference (BI) tree reconstructing the phylogenetic relationships of mitochondrial genomes. The subfamilies of Poaceae (colored blocks) are only weakly resolved in red nodes. The tree is based on nucleotide sequences of 30 conserved mitochondrial PCGs (*atp1*, *atp4*, *atp6*, *atp8*, *atp9*, *nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7*, *nad9*, *ccmB*, *ccmC*, *ccmFC*, *ccmFN*, *cox1*, *cox2*, *cox3*, *cob*, *mttB*, *matR*, *rpl16*, *rps1*, *rps3*, *rps7*, *rps12* and *rps13*). The ML topology is indicated with ML bootstrap support values and BI posterior probabilities at each node; red nodes represent bootstrap support less than 100% (ML) or 1 (BI). Two other monocotyledonous species (*Cyperus esculentus* and *Phoenix dactylifera*) are used as outgroups. **B** Presence of non-core mitochondrial genes in Poaceae where filled squares represent presence of at least one complete copy. While some features are shared in subfamilies, there is substantial variation; *A. longiglumis* (red shading) shows characteristic differences. *\_cp* (green): tRNA genes that migrated from plastome; *\_mt* represents mitochondria-native tRNA genes

**A** Tree scale: 0.005



**B**



**Fig. 6** (See legend on previous page.)

recombination and RNA editing for the important cereal crop relatives. These results provide an important model for grass evolution and an essential reference for mitochondrial-associated characters, not least male sterility, in cereal crop breeding.

#### Abbreviations

BI	Bayesian Inference
BLAST	Basic local alignment search tool
cp	Plastome
GC	Guanine-cytosine
ML	Maximum Likelihood
mitogenome	Mitochondrial genome
mt	Mitochondria contig
MTPTs	Mitochondrial plastid DNAs
MTNUs	Mitochondrial nuclear DNAs
NCBI	National Center for Biotechnology Information
PCGs	Protein-coding genes
SNPs	Single nucleotide polymorphism sites
SSRs	Simple sequence repeats
tRNAs	Transfer RNAs

#### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12870-023-04217-8>.

**Additional file 1: Figure S1.** Graphical assembly display of the mitogenome based on Unicycler using Illumina short reads and ONT long reads. **Figure S2.** The master circular mitogenome generated by manually merging the four contigs based on long repetitive sequences. **Figure S3.** Sequencing coverage based on Illumina short-reads. **Figure S4.** RNA-editing to create a stop codon of gene *ccmFC*.

**Additional file 2: Table S1.** Species and GenBank accessions used for phylogenetic analysis. **Table S2.** Characteristics of unitigs assembled by Unicycler. **Table S3.** Annotated genes and their locations. **Table S4.** SSRs identified in the mitochondrial genome of *A. longiglumis*. **Table S5.** Tandem repeat sequences identified in the mitochondrial genome of *A. longiglumis*. **Table S6.** Duplicated repeats ( $\geq 30$  bp) identified in the mitochondrial genome of *A. longiglumis*. **Table S7.** The homologous DNA fragment identified among the mitochondrial genome and plastome of *A. longiglumis*. **Table S8.** The blastn results among the mitogenome (Query) and nuclear genome (Subject) of *A. longiglumis*. **Table S9.** RNA editing events identified in the mitochondrial protein coding gene of *A. longiglumis*. **Table S10.** tRNA genes identified in the mitochondrial genome of 29 Poaceae accessions.

#### Acknowledgements

We would like to thank National Plant Germplasm System (NGPS) of USDA-ARS providing seeds. We thank Grandomics Biosciences Co., Ltd. (Wuhan, China) for genome sequencing support, and Huawei Elastic Cloud Server (Jiangsu, China) for on-demand computational resources. We thank an anonymous reviewer who suggested the genome was presented as a single circle.

#### Authors' contributions

QL, TS and JSHH designed the research. HYY, JXX and GX conducted experiments. QL and HYY assembled and annotated the mitogenome. DLC performed bioinformatics analyses. TS and JXX prepared figures and tables. QL, DLC, TS and JSHH wrote the manuscript. All authors read and approved the final manuscript.

#### Funding

This work was supported by grants from the National Natural Science Foundation of China (32070359) to QL, the Guangdong Basic and Applied Basic Research Foundation (2021A1515012410) to QL, the Overseas Distinguished Scholar Project of South China Botanical Garden, Chinese Academy of

Sciences (Y861041001) to JSHH, Global Challenges Research Fund (GCRF) Foundation Awards for Global Agricultural and Food Systems Research (BB/P02307X/1), and University of Chinese Academy of Sciences (UCAS) Innovative Training Programs for Undergraduates to JXX (KCJH-80107–2020-047).

#### Availability of data and materials

The raw sequencing data for the Illumina short-reads and Nanopore long-reads and the mitogenome sequences have been deposited in NCBI (<https://www.ncbi.nlm.nih.gov/>) with accession number: PRJNA838431, SAMN28422612, SRR19279518, SRR19279531. The accession number of master circular mitogenome is OQ450323 (<https://www.ncbi.nlm.nih.gov/nuccore/OQ450323>).

#### Declarations

##### Ethics approval and consent to participate

We sampled the diploid oat *A. longiglumis* in this study from publicly available seed sources. The plant samples and experimental research comply with relevant institutional, national, and international guidelines and legislation.

##### Consent for publication

Not applicable.

##### Competing interests

The authors declare that they have no competing interests.

##### Author details

<sup>1</sup>Key Laboratory of Plant Resources Conservation and Sustainable Utilization / Guangdong Provincial Key Laboratory of Applied Botany, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China. <sup>2</sup>South China National Botanical Garden, Guangzhou 510650, China. <sup>3</sup>Center for Conservation Biology, Core Botanical Gardens, Chinese Academy of Sciences, Guangzhou 510650, China. <sup>4</sup>University of Chinese Academy of Sciences, Beijing 100049, China. <sup>5</sup>College of Plant Protection, South China Agricultural University, Guangzhou 510642, China. <sup>6</sup>Department of Genetics and Genome Biology, Institute for Environmental Futures, University of Leicester, Leicester LE1 7RH, UK.

Received: 6 December 2022 Accepted: 6 April 2023

Published online: 26 April 2023

#### References

- Kaur P, Kaur K, Basha SJ, Kennedy JF. Current trends in the preparation, characterization and applications of oat starch—A review. *Int J Biol Macromol.* 2022;212:172–81.
- Paudel D, Dhungana B, Caffè M, Krishnan P. A review of health-beneficial properties of oats. *Foods.* 2021;10:2591.
- Punia S, Sandhu KS, Dhull SB, Siroha AK, Purewal SS, Kaur M, et al. Oat starch: physico-chemical, morphological, rheological characteristics and its applications—A review. *Int J Biol Macromol.* 2020;154:493–8.
- Summo C, De Angelis D, Difonzo G, Caponio F, Pasqualone A. Effectiveness of oat-hull-based ingredient as fat replacer to produce low fat burger with high beta-glucans content. *Foods.* 2020;9(8):1057.
- Bochnia M, Boesel M, Bahrenthien L, Wensch-Dorendorf M, Zeyner A. Feed intake patterns of sport ponies and warmblood horses following iso-energetic intake of pelleted fibre-rich mixed feed, muesli feed and semicrushed oat grains. *J Anim Physiol Anim Nutr (Berl).* 2017; 101:Suppl 1:37–42.
- Chew P, Meade K, Hayes A, Harjes C, Bao Y, Beattie AD, et al. A study on the genetic relationships of *Avena* taxa and the origins of hexaploid oat. *Theor Appl Genet.* 2016;129(7):1405–15.
- Loskutov IG, Gnutikov AA, Blinova EV, Rodionov AV. The origin and resource potential of wild and cultivated species of the genus of oats (*Avena* L.). *Russ J Genet.* 2021;57(6):642–61.
- Cheng DW, Armstrong KC, Drouin G, McElroy A, Fedak G, Molnar SD. Isolation and identification of Triticeae chromosome 1 receptor-like kinase genes (*Lrk10*) from diploid, tetraploid, and hexaploid species of the genus *Avena*. *Genome.* 2003;46(1):119–27.

9. Liu Q, Li XY, Li MZ, Xu WK, Schwarzacher T, Heslop-Harrison JS. Comparative chloroplast genome analyses of *Avena*: insights into evolutionary dynamics and phylogeny. *BMC Plant Biol.* 2020;20(1):406.
10. Fu YB, Li P, Biligetü B. Developing chloroplast genomic resources from 25 *Avena* species for the characterization of oat wild relative germplasm. *Plants (Basel).* 2019;8(11):438.
11. Sebastin R, Lee KJ, Shin MJ, Cho GT, Ma KH, Lee JR, Lee GA, Chung JW. The complete chloroplast genome sequence of wild oat, *Avena sterilis* L. (Poaceae) and its phylogeny. *Mitochondrial DNA B Resour.* 2018;3(1):311–2.
12. Peng YY, Yan HH, Guo LC, Deng C, Wang CL, Wang YB, et al. Reference genome assemblies reveal the origin and evolution of allohexaploid oat. *Nat Genet.* 2022;54(8):1248–58.
13. Kamal N, Tsardakas Renhuldt N, Bentzer J, Gundlach H, Haberer G, Haberer G, et al. The mosaic oat genome gives insights into a uniquely healthy cereal crop. *Nature.* 2022;606(7912):113–9.
14. Liu Q, Yuan HY, Li MZ, Wang ZW, Cui DL, Ye YS, et al. Chromosome-scale genome assembly of the diploid oat *Avena longiglumis* reveals the landscape of repetitive sequences, genes and chromosome evolution in grasses. *bioRxiv.* 2022;479819.
15. Mustafa SI, Heslop-Harrison JS, Schwarzacher T. The complete mitochondrial genome from Iraqi Meriz goats and the maternal lineage using whole genome sequencing data. *Iran J Appl Anim Sci.* 2021;12(2):321–8.
16. Kavalco ML, Rocha-Reis DA, Heslop-Harrison JS, Schwarzacher T, Kavalco KF. Ten complete mitochondrial genomes of *Gymnocharacini* (Stethaprioninae, Characiformes): evolutionary relationships and a repetitive element in the control region (D-loop). *Front Ecol Evol.* 2021;9:650783.
17. Kozik A, Rowan BA, Lavelle D, Berke L, Schranz ME, Michelmore RW, et al. The alternative reality of plant mitochondrial DNA: one ring does not rule them all. *PLoS Genet.* 2019;15(8):e1008373.
18. Morley SA, Nielsen BL. Plant mitochondrial DNA. *Front Biosci (Landmark Ed).* 2017;22(6):1023–32.
19. Asaf S, Khan AL, Khan AR, Waqas M, Kang SM, Khan MA, et al. Mitochondrial genome analysis of wild rice (*Oryza minuta*) and its comparison with other related species. *PLoS One.* 2016;11(4):e0152937.
20. Fauron C, Casper M, Gao Y, Moore B. The maize mitochondrial genome: dynamic, yet functional. *Trends Genet.* 1995;11(6):228–35.
21. Hall ND, Zhang H, Mower JP, McElroy JS, Goertzen LR. The mitochondrial genome of *Eleusine indica* and characterization of gene content within Poaceae. *Genome Biol Evol.* 2020;12(1):3684–97.
22. Hisano H, Tsujimura M, Yoshida H, Terachi T, Sato K. Mitochondrial genome sequences from wild and cultivated barley (*Hordeum vulgare*). *BMC Genomics.* 2016;17(1):824.
23. Cheng Q, Wang P, Li T, Liu J, Zhang Y, Wang YH, et al. Complete mitochondrial genome sequence and identification of a candidate gene responsible for cytoplasmic male sterility in celery (*Apium graveolens* L.). *Int J Mol Sci.* 2021;22(16):8584.
24. Wang YP, Wang QB, Hao W, Li JX, Qi MX, Zhang L. Mitochondrial genome sequencing reveals *orf463a* may induce male sterility in NWB cytoplasm of radish. *Genes (Basel).* 2020;11(1):74.
25. Takatsuka A, Kazama T, Toriyama K. Cytoplasmic male sterility-associated mitochondrial gene *orf312* derived from rice (*Oryza sativa* L.) cultivar Tadukan. *Rice (NY).* 2021;14(1):46.
26. Xiao S, Xing J, Nie T, Su A, Zhang R, Zhao Y, et al. Comparative analysis of mitochondrial genomes of maize CMS-S subtypes provides new insights into male sterility stability. *BMC Plant Biol.* 2022;22(1):469.
27. Lu H, Giordano F, Ning Z. Oxford nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics.* 2016;14(5):265–79.
28. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods.* 2009;6(4):291–5.
29. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114–20.
30. Jin JJ, Yu WB, Yang JB, Song Y, dePamphilis CW, Yi TS, et al. GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* 2020;21(1):241.
31. Freire B, Ladra S, Parama JR. Memory-efficient assembly using Flye. *IEEE/ACM Trans Comput Biol Bioinform.* 2021;3108843.
32. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics.* 2015;31(20):3350–2.
33. Chen Y, Ye W, Zhang Y, Xu Y. High speed BLASTN: an accelerated MegaB-LAST search tool. *Nucleic Acids Res.* 2015;43(16):7762–8.
34. 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.
35. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754–60.
36. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26(6):841–2.
37. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094–100.
38. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol.* 2017;13(6):e1005595.
39. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19(5):455–77.
40. Shi L, Chen H, Jiang M, Wang L, Wu X, Huang L, et al. CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Res.* 2019;47(W1):W65–73.
41. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq- versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 2017;45(W1):W6–11.
42. Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, Palmer JD. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Mol Biol Evol.* 2010;27(6):1436–48.
43. Misra S, Harris N. Using Apollo to browse and edit genome annotations. *Current Protoc Bioinformatics.* 2005;12(1):9.5.1–9.5.28.
44. Chan PP, Lowe TM. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol Biol (Clifton, NJ).* 2019;1962:1–14.
45. Sloan DB, Alverson AJ, Chackalovcak JP, Wu M, McCauley DE, Palmer JD, et al. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biol.* 2012;10(1):e1001241.
46. Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 2019;47(W1):W59–64.
47. Beier S, Thiel T, Munch T, Scholz U, Mascher M. MISA-web: a web server for microsatellite prediction. *Bioinformatics.* 2017;33(16):2583–5.
48. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 2001;29(22):4633–42.
49. Benson G. Tandem Repeats Finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 1999;27(2):573–80.
50. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357–9.
51. Milne I, Bayer M, Stephen G, Cardle L, Marshall D. Tablet: visualizing next-generation sequence assemblies and mappings. *Methods Mol Biol.* 2016;1374:253–68.
52. Danecek P, McCarthy SA. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics.* 2017;33(13):2037–9.
53. Zhang D, Gao F, Jakovlic I, Zou H, Zhang J, Li WX, et al. PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol Ecol Resour.* 2020;20(1):348–55.
54. Rozewicki J, Li S, Amada KM, Standley DM, Katoh K. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res.* 2019;47(W1):W5–10.
55. Stamatakis A. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 2014;30(9):1312–3.
56. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* 2012;61(3):539–42.
57. Clayton WD, Vorontsova MS, Harman KT, Williamson H. GrassBase—The online world grass flora. 2006. Online: [http://www.kew.org/data/grass\\_base/index.html](http://www.kew.org/data/grass_base/index.html). Accessed 15 Feb 2023.
58. Mower J. Variation in protein gene and intron content among land plant mitogenomes. *Mitochondrion.* 2020;53:203–13.

59. Christensen AC. Plant mitochondrial genome evolution can be explained by DNA repair mechanisms. *Genome Biol Evol.* 2013;5(6):1079–86.
60. Fischer A, Dotzek J, Walther D, Greiner S. Graph-based models of the *Oenothera* mitochondrial genome capture the enormous complexity of higher plant mitochondrial DNA organization. *NAR Genom Bioinform.* 2022;4(2):lqac027.
61. Wu ZQ, Liao XZ, Zhang XN, Tembrock LR, Broz A. Genomic architectural variation of plant mitochondria—A review of multichromosomal structuring. *J Syst Evol.* 2020;60(1):160–8.
62. Sloan DB. One ring to rule them all? Genome sequencing provides new insights into the 'master circle' model of plant mitochondrial DNA structure. *New Phytol.* 2013;200(4):978–85.
63. Cheng N, Lo YS, Ansari MI, Ho KC, Jeng ST, Lin NS, et al. Correlation between mtDNA complexity and mtDNA replication mode in developing cotyledon mitochondria during mung bean seed germination. *New Phytol.* 2017;213(2):751–63.
64. Dong S, Zhao C, Chen F, Liu Y, Zhang S, Wu H, et al. The complete mitochondrial genome of the early flowering plant *Nymphaea colorata* is highly repetitive with low recombination. *BMC Genomics.* 2018;19(1):614.
65. Mower JP, Case AL, Floro ER, Willis JH. Evidence against equimolarity of large repeat arrangements and a predominant master circle structure of the mitochondrial genome from a monkeyflower (*Mimulus guttatus*) lineage with cryptic CMS. *Genome Biol Evol.* 2012;4(5):670–86.
66. Guo W, Grewe F, Fan W, Young GJ, Knoop V, Palmer JD, et al. *Ginkgo* and *Welwitschia* mitogenomes reveal extreme contrasts in Gymnosperm mitochondrial evolution. *Mol Biol Evol.* 2016;33(6):1448–60.
67. Field PD, Waneka G, Naish M, Schatz M, Henderson IR, Sloan DB. Complete sequence of a 641-kb insertion of mitochondrial DNA in the *Arabidopsis thaliana* nuclear genome. *Genome Biol Evol.* 2022;14(5):evac059.
68. Lloyd Evans D, Hlongwane TT, Joshi SV, Riaño Pachón DM. The sugarcane mitochondrial genome: assembly, phylogenetics and transcriptomics. *Peer J.* 2019;7: e7558.
69. Yang W, Zou J, Wang J, Li N, Luo X, Jiang X, et al. Wide crossing diversity mitogenomes of rice. *BMC Plant Biol.* 2020;20(1):159.
70. Liao PC, Bergamini C, Fato R, Pon LA, Pallotti F. Isolation of mitochondria from cells and tissues. *Methods Cell Biol.* 2020;155:3–31.
71. Tsujimura M, Kaneko T, Sakamoto T, Kimura S, Shigyo M, Yamagishi H, et al. Multichromosomal structure of the onion mitochondrial genome and a transcript analysis. *Mitochondrion.* 2019;46:179–86.
72. Wang XC, Chen H, Yang D, Liu C. Diversity of mitochondrial plastid DNAs (MTPDs) in seed plants. *Mitochondrial DNA A DNA Mapp Seq Anal.* 2018;29(4):635–42.
73. Cheng Y, He X, Priyadarshani S, Wang Y, Ye L, Shi C, et al. Assembly and comparative analysis of the complete mitochondrial genome of *Suaeda glauca*. *BMC Genomics.* 2021;22(1):167.
74. Ye N, Wang X, Li J, Bi C, Xu Y, Wu D, et al. Assembly and comparative analysis of complete mitochondrial genome sequence of an economic plant *Salix suchowensis*. *Peer J.* 2017;5: e3148.
75. Leister D, Kleine T. Role of intercompartmental DNA transfer in producing genetic diversity. *Int Rev Cell Mol Biol.* 2011;291:73–114.
76. Joyce PB, Gray MW. Chloroplast-like transfer RNA genes expressed in wheat mitochondria. *Nucleic Acids Res.* 1989;17(14):5461–76.
77. Richardson AO, Rice DW, Young GJ, Alverson AJ, Palmer JD. The "fossilized" mitochondrial genome of *Liriodendron tulipifera*: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. *BMC Biol.* 2013;11:29.
78. Wu X, Hu X, Chen X, Zhang J, Ren C, Song L, et al. Sequencing and characterization of the complete mitochondrial genome of *Thinopyrum obtusiflorum* (DC.) Banfi, 2018 (Poaceae). *Mitochondrial DNA B Resour.* 2022;7(3):539–40.
79. Xiong Y, Yu Q, Xiong Y, Zhao J, Lei X, Liu L, et al. The complete mitogenome of *Elymus sibiricus* and insights into its evolutionary pattern based on simple repeat sequences of seed plant mitogenomes. *Front Plant Sci.* 2021;12: 802321.
80. Adams KL, Palmer JD. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Mol Phylogenet Evol.* 2003;29(3):380–95.
81. Richert-Pöggeler KR, Vijverberg K, Alisawi O, Chofong GN, Heslop-Harrison JS, Schwarzacher T. Participation of multifunctional RNA in replication, recombination and regulation of endogenous plant pararetroviruses (EPRVs). *Front Plant Sci.* 2021;12: 689307.
82. Marienfeld J, Unsel M, Brennicke A. The mitochondrial genome of *Arabidopsis* is composed of both native and immigrant information. *Trends Plant Sci.* 1999;4(12):495–502.
83. Huang WC, Zhang L, Columbus JT, Hu Y, Zhao YY, Tang L, et al. A well-supported nuclear phylogeny of Poaceae and implications for the evolution of C<sub>4</sub> photosynthesis. *Mol Plant.* 2022;15(4):755–77.
84. Burke SV, Wysocki WP, Zuloaga FO, Craine JM, Pires JC, Edger PP, et al. Evolutionary relationships in Panicoid grasses based on plastome phylogenomics (Panicoidae; Poaceae). *BMC Plant Biol.* 2016;16(1):140.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

