



Article

Developing Chloroplast Genomic Resources from 25 *Avena* Species for the Characterization of Oat Wild Relative Germplasm

Yong-Bi Fu ^{1,*} , Pingchuan Li ² and Bill Biligetu ² 

¹ Plant Gene Resources of Canada, Saskatoon Research and Development Centre, Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK S7N 0X2, Canada

² Department of Plant Sciences, University of Saskatchewan, 51 Campus Drive, Saskatoon, SK S7N 5A8, Canada; lipingchuan@gmail.com (P.L.); bill.biligetu@usask.ca (B.B.)

* Correspondence: yong-bi.fu@canada.ca; +306-385-9298

Received: 4 September 2019; Accepted: 21 October 2019; Published: 23 October 2019



Abstract: Chloroplast (cp) genomics will play an important role in the characterization of crop wild relative germplasm conserved in worldwide gene banks, thanks to the advances in genome sequencing. We applied a multiplexed shotgun sequencing procedure to sequence the cp genomes of 25 *Avena* species with variable ploidy levels. Bioinformatics analysis of the acquired sequences generated 25 *de novo* genome assemblies ranging from 135,557 to 136,006 bp. The gene annotations revealed 130 genes and their duplications, along with four to six pseudogenes, for each genome. Little differences in genome structure and gene arrangement were observed across the 25 species. Polymorphism analyses identified 1313 polymorphic sites and revealed an average of 277 microsatellites per genome. Greater nucleotide diversity was observed in the short single-copy region. Genome-wide scanning of selection signals suggested that six cp genes were under positive selection on some amino acids. These research outputs allow for a better understanding of oat cp genomes and evolution, and they form an essential set of cp genomic resources for the studies of oat evolutionary biology and for oat wild relative germplasm characterization.

Keywords: Crop wild relative; *Avena*; chloroplast genome; chloroplast gene; positive selection

1. Introduction

Chloroplast (cp) genomics will play an important role in the characterization of crop wild relative (CWR) germplasm conserved in worldwide gene banks. Currently, many CWR collections are expanding to mitigate the threats of losing CWRs from climate change and other habitat disturbances and to conserve germplasm for plant breeding (e.g., see [1,2]). Thus, the need for CWR germplasm characterization is increasing. Most CWR germplasm has complex, polyploid, and/or uncharacterized genomes [3,4], and current tools based on nuclear genome sequences may not always be effective in identifying CWR germplasm and investigating its genetic variability [5,6]. Conserved CWRs are often misclassified or require taxonomic evaluation [7]. More informative barcoding [8] will be needed to distinguish among CWR accessions for specific traits or features. Acquired CWRs will be used to enhance the studies of plant biology and evolution (e.g., see [6,9]). The search for cp genes in conserved CWR will be increased for applications in genetic engineering [10]. Therefore, it is important to develop cp genomic resources for characterizing CWR germplasm.

Oat (*Avena* L.) is one of the most widely cultivated cereals and a valuable resource in many countries, both for human consumption and animal feed [11]. The genus *Avena* has up to 30 recognized species, including diploids, tetraploids, and hexaploids [7,12]. The cultivated hexaploid oat has 42

chromosomes, representing three different sets of nuclear genomes (A, C, and D) [6]. Currently, there are more than 31,000 accessions of oat wild relative germplasm conserved in more than 20 gene banks worldwide [13]. These genetic resources are known to harbor an important source of genetic variability [14] for oat genetic improvement through interspecific crossing and introgression [7,15,16]. Conserving and managing these wild relative accessions are challenging tasks [17]. However, there is no report so far on the research to develop oat cp genomic resources for characterizing these oat accessions [18], although several cp-based evolutionary studies of oat species were conducted [9,19–23]. This research has helped to provide insights into the maternal origins of oat genomes, confirming the general consensus in oat phylogeny [9,18,24].

The recent advances in next-generation sequencing have made the sequencing of organelle genomes more feasible than before [25–28]. We applied a multiplexed shotgun sequencing procedure to sequence the cp genomes of 25 *Avena* species with variable ploidy levels. In our companion paper [18], we extracted single nucleotide polymorphism (SNP), using the wheat cp genome as the reference and inferred maternal patterns of oat evolution through phylogenetic analyses. In this paper, we will perform various types of bioinformatics analysis to assemble all 25 cp genomes, annotate the genome assemblies, conduct comparative genomic analysis, and analyze sequence divergence and selective pressure for cp genes. It was our hope that these analyses would provide baseline information that could give us a better understanding of the oat cp genome and evolution and for oat wild relative germplasm characterization.

2. Results

2.1. Sequencing and Assembly

Four MiSeq runs generated a total of 95 million sequence reads for 25 *Avena* samples, each having 3.8 million sequence reads (Table 1). After removing sequence reads of poor quality ($Q < 15$ and read length < 150 bp), an average of 83% high-quality sequence reads were obtained for these samples. Thus, each sample still had a sequence length ranging from 1200 to 2500 Mbp, with an average of 1817 Mbp, corresponding to an approximately 8900× to 18,890× cp genome coverage. Such high genome coverages made the cp genome assembly simpler, with the smallest numbers of contigs and scaffolds under proper k-mer coverage and size setting.

De novo assembly with the pair-end sequence reads from each sample generated three major scaffolds, without any exception for all the 25 samples, as expected with two inverted repeats (IRs). As illustrated with 2x, 4x, and 6x oat species in Figure 1, all the circular *Avena* cp genomes consisted of four typical DNA fragment structures: one large single copy (LSC), one small single copy (SSC), and two inverted repeat regions (IRa and IRb). The cp genome sizes ranged from 135,557 to 136,006 bp, with an average of 135,878 bp (Table 1). The 25 complete cp genome sequences and the generic (or consensus) “*Avena*” cp genome sequence are available on Supplementary File S1 and File S2, respectively. Changes in cp genome size mainly were reflected in the LSC region. The average sizes for LSC, SSC, and IR among the 25 *Avena* species were 80k bp, 12.6k bp, and 21.6 kb, respectively. The average of GC content ranged from 38.41% to 38.53%, with an average of 38.49% (Table 1).

2.2. Chloroplast Genome Gene Annotation

The gene annotations revealed an identical set of 130 cp genes for these 25 oat species (Table 2; Supplementary File S3). Of all the 130 genes, 84 coding genes, eight rRNA, and 38 tRNA were identified as common genes across the oat species of variable ploidy (Table 2; Figure 1). There were 21 genes (13 coding genes and eight tRNA genes) that had at least one intron. Also, four pseudogenes were found in the two diploid species (*A. clauda* and *A. eriantha*), and six pseudogenes were identified for the other 23 oat species. All pseudogenes were distributed in the LSC region. Both IRa and IRb regions shared the same amount of 19 duplicated genes, but one marked difference is that the *ndhH* gene had different C terminals. As both 3′ ends of the *ndhH* genes were extended to either side of the SSC fragment,

it caused a different translation of peptide tails. Two genes *infA* and *rps16*, coding for a translation initiation factor 1 and a S16 ribosomal protein, respectively, were detected in these oat species.

Table 1. List of 25 studied *Avena* species from six botanical sections and their cp genome assemblies.

Section /Species	Pl	Raw Reads	CPG Size (bp)	CPG Region (bp)				GC%	NCBI Acc#	PGRC Acc#
				LSC	IRb	SSC	IRa			
Ventricosa										
<i>A. ventricosa</i>	2x	3,399,167	135,681	79,793	21,614	12,660	21,614	38.41	MG687301	CN21992
<i>A. clauda</i>	2x	3,037,449	135,557	79,667	21,619	12,652	21,619	38.43	MG687303	CN19205
<i>A. eriantha</i>	2x	3,932,778	135,560	79,669	21,619	12,653	21,619	38.43	MG687291	CN19256
Agraria										
<i>A. hispanica</i>	2x	3,138,274	135,935	80,099	21,605	12,626	21,605	38.49	MG687300	CN25788
<i>A. brevis</i>	2x	3,145,903	135,939	80,101	21,606	12,626	21,606	38.49	MG687310	CN3145
<i>A. muda</i>	2x	3,342,392	135,934	80,100	21,604	12,626	21,604	38.48	MG687306	CN79350
<i>A. strigosa</i>	2x	2,618,401	135,938	80,102	21,605	12,626	21,605	38.48	MG687309	CN22002
Tenuicarpa										
<i>A. canariensis</i>	2x	4,285,394	135,955	80,147	21,598	12,612	21,598	38.52	MG687297	CN25449
<i>A. damascena</i>	2x	2,906,067	135,925	80,101	21,602	12,620	21,602	38.51	MG687302	CN19458
<i>A. atlantica</i>	2x	4,092,127	136,006	80,168	21,606	12,626	21,606	38.48	MG687299	CN25859
<i>A. wiestii</i>	2x	4,372,976	135,944	80,109	21,605	12,625	21,605	38.49	MG687296	CN24315
<i>A. lusitanica</i>	2x	4,425,054	135,879	80,166	21,603	12,507	21,603	38.52	MG687295	CN25936
<i>A. longiglumis</i>	2x	4,288,420	135,728	79,881	21,605	12,637	21,605	38.53	MG687305	CN21407
<i>A. agadiriana</i>	4x	4,432,963	135,945	80,129	21,602	12,612	21,602	38.49	MG687294	CN25868
<i>A. barbata</i>	4x	4,362,663	135,946	80,111	21,605	12,625	21,605	38.49	MG687311	CN24462
Pachycarpa										
<i>A. insularis</i>	4x	4,001,844	135,967	80,130	21,603	12,631	21,603	38.5	MG674209	CN19178
<i>A. maroccana</i>	4x	4,160,684	135,887	80,102	21,604	12,577	21,604	38.51	MG687298	CN23057
<i>A. murphyi</i>	4x	5,199,242	135,892	80,108	21,604	12,576	21,604	38.51	MG687312	CN21989
Ethiopica										
<i>A. vaviloviana</i>	4x	3,870,454	135,946	80,111	21,605	12,625	21,605	38.49	MG687304	CN22413
<i>A. abyssinica</i>	4x	2,822,158	135,942	80,109	21,604	12,625	21,604	38.49	MG687293	CN22064
Avena										
<i>A. fatua</i>	6x	4,120,412	135,889	80,106	21,604	12,575	21,604	38.51	MG687307	CN21948
<i>A. hybrida</i>	6x	4,330,653	135,900	80,117	21,604	12,575	21,604	38.51	MG687292	CN24926
<i>A. occidentalis</i>	6x	2,947,248	135,893	80,114	21,602	12,575	21,602	38.51	MG687314	CN25946
<i>A. sterilis</i>	6x	2,467,386	135,888	80,107	21,603	12,575	21,603	38.51	MG687308	CN20625
<i>A. sativa</i>	6x	5,329,202	135,886	80,107	21,602	12,575	21,602	38.51	MG687313	CN24549

Note: Pl is for ploidy level. CPG is chloroplast genome. LSC, SSC, and IR are large single copy, small single copy and inverted repeat regions, respectively. NCBI Acc# are the accession numbers for the cp assemblies deposited in the National Center for Biotechnology Information (NCBI). PGRC Acc# are the accession numbers for the studied samples obtained from the oat collection at Plant Gene Resources of Canada (PGRC).

2.3. Comparison of Genomic Structures

Analyzing mVISTA percent identity plot revealed several major features of genomic variation, as illustrated in Figure 2 for nine oat species (and Supplementary File S4 for 25 oat species). First, no marked differences in genomic structure and gene arrangement were observed. Second, the degree of similarity between any two of 25 cp genome sequences ranged from 98.380% to 99.996%, with an average of 99.529%. Third, most of the nucleotide variations across the 25 cp genomes were located in intergenic regions. Fourth, sequence variation among the 25 genomes was also identified for the *ndhH* gene. Fifth, there were no specific variations in genomic structure and gene arrangement unique to each ploidy level.

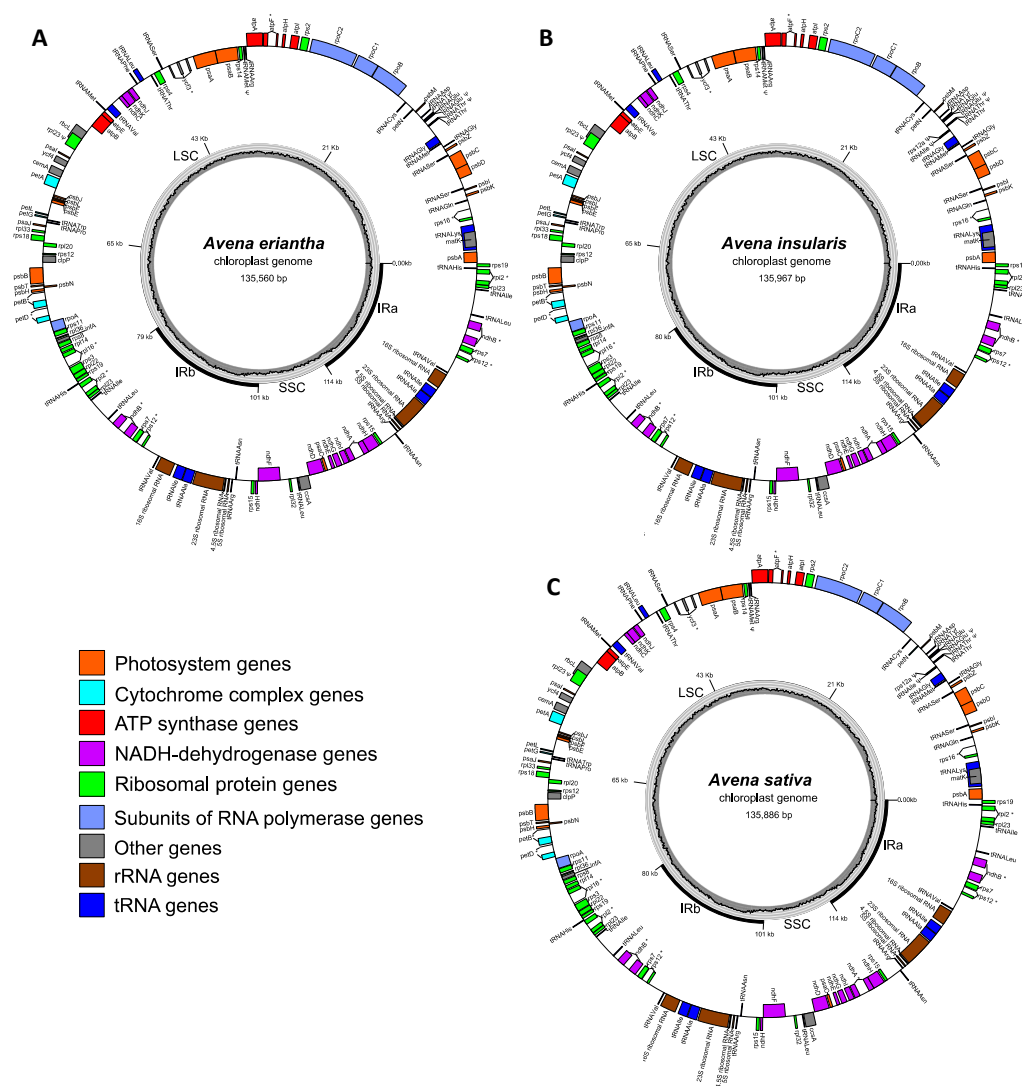


Figure 1. Gene maps of three selected *Avena* chloroplast genomes: (A) *A. eriantha* (2x), (B) *A. insularis* (4x), and (C) *A. sativa* (6x). Each map is represented in, moving counterclockwise from the right. The larger circle represents the layout of chloroplast genes distribution as per their transcription direction: outside boxes and inside boxes show the counterclockwise and clockwise transcription. The color of the gene box indicates the functional group that the gene belongs to. The smaller circle represents the CG content plot in the corresponding sample. LSC, large single copy region; SSC, small single copy region. IRa/b, inverted repeats. Intron-containing genes are marked by a ‘*’ symbol and pseudogenes are marked by a ‘Ψ’ symbol.

2.4. Sequence Variation and Divergence

The SNP calls based on the full length of 25 cp genome sequence alignments with the most stringent conditions revealed a total of 1313 SNPs for these 25 species (see Supplementary File S5). There were 583 SNPs (44.4%) located in the genic regions, 16 SNPs (1.2%) in the pseudogene, and 714 SNPs (54.4%) distributed in the intergenic regions.

The simple sequence repeat (SSR) analysis revealed considerable SSR polymorphism in these 25 oat cp genomes. A total of 6694 SSRs were identified for the 25 oat species. The SSR counts per species ranged from 256 (*A. clauda* and *A. eriantha*) to 280 (*A. atlantica*), with an average of 276.8. The SSR motifs were mainly poly-A, with eight to 18 repeats; poly-C, with eight to 14 repeats; poly-AT, with five to seven repeats; and poly-AG with five repeats (Table 3). Six abundant SSR motifs were poly-A, with 8,

9, 10, and 11 repeats, followed by poly-C, with eight repeats and poly-AT, with five repeats. However, no SSR motifs for tri-, tetra-, penta-, and hexa-nucleotides were found.

Table 2. List of 130 genes and their duplications found in the plastids of 25 *Avena* species.

Category	Gene*
Subunits of photosystem I	<i>psaA, psaB, psaC, psaI, psaJ</i>
Subunits of photosystem II	<i>psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ</i>
Subunits of cytochrome b/f complex	<i>petA, petB^a, petD^a, petG, petL, petN</i>
Subunits of ATP synthase	<i>atpA, atpB, atpE, atpF^a, atpH, atpI</i>
Large subunit of rubisco	<i>rbcL</i>
Subunits of NADH-dehydrogenase	<i>ndhA^a, ndhB(1)^a, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH(1), ndhI, ndhJ, ndhK</i>
Proteins of large ribosomal subunit	<i>rpl2(1)^a, rpl14, rpl16^a, rpl20, rpl22, rpl23(1), rpl32, rpl33, rpl36</i>
Proteins of small ribosomal subunit	<i>rps2, rps3, rps4, rps7(1), rps8, rps11, rps12(1)^a, rps14, rps15(1), rps16^a, rps18, rps19(1)</i>
Subunits of RNA polymerase	<i>rpoA, rpoB, rpoC1, rpoC2</i>
Cytochrome c biogenesis	<i>ccsA</i>
Transfer RNAs	<i>tRNA-Ala(2), tRNA-Arg(3), tRNA-Asn(2), tRNA-Asp, tRNA-Cys, tRNA-Gln, tRNA-Glu, tRNA-Gly(2), tRNA-His(2), tRNA-Ile(4), tRNA-Leu(4), tRNA-Lys, tRNA-Met(2), tRNA-Phe, tRNA-Pro, tRNA-Ser(3), tRNA-Thr(2), tRNA-Trp, tRNA-Tyr, tRNA-Val(3)</i>
Ribosomal RNAs	<i>16S rRNA(2), 23S rRNA(2), 4.5S rRNA(2), 5S rRNA(2)</i>
Maturase	<i>matK</i>
Protease	<i>clpP</i>
Conserved hypothetical genes	<i>ycf3^a, ycf4</i>
Envelope membrane protein	<i>cemA</i>
Translation initiation factor	<i>infA</i>

* The superscript ^a means the gene contains intron(s). The number in parentheses after a gene shows the number of duplications for the gene in the other genome regions.

Table 3. Simple sequence repeat (SSR) polymorphism found in the plastids of 25 *Avena* species.

SSR Type	A	A	A	A	A	A	A	A	A	A	A	C	C	C	C	C	AT	AT	AT	AG	
Repeat count	8	9	10	11	12	13	14	15	16	17	18	8	9	10	11	12	14	5	6	7	5
<i>A. ventricosa</i>	62	33	14	5	1	1	2	0	0	0	1	4	3	1	0	0	0	4	0	1	1
<i>A. clauda</i>	58	32	12	8	1	1	2	0	0	0	1	3	1	3	0	1	0	3	1	0	1
<i>A. eriantha</i>	57	30	15	8	1	1	2	0	0	0	1	3	1	3	1	0	0	3	1	0	1
<i>A. hispanica</i>	67	27	17	6	1	1	1	1	1	0	0	7	4	0	0	0	0	4	1	0	1
<i>A. brevis</i>	67	27	16	7	1	0	2	1	1	0	0	7	2	2	0	0	0	4	1	0	1
<i>A. nuda</i>	67	29	15	5	1	1	2	1	1	0	0	9	2	0	0	0	0	4	1	0	1
<i>A. strigosa</i>	67	29	15	2	4	2	1	1	1	0	0	7	4	0	0	0	0	4	1	0	1
<i>A. canariensis</i>	66	29	17	4	1	0	1	0	1	0	0	6	2	2	0	0	0	3	0	1	1
<i>A. damascena</i>	66	27	13	7	4	1	1	0	0	1	0	5	4	0	0	0	1	3	0	1	1
<i>A. atlantica</i>	68	30	15	2	2	2	3	1	0	0	0	7	2	2	0	0	0	4	1	0	1
<i>A. wiestii</i>	66	28	14	7	1	0	1	1	2	0	0	7	1	2	1	0	0	4	1	0	1
<i>A. lusitanica</i>	63	26	17	8	1	1	1	0	1	0	0	6	1	3	0	0	0	3	0	1	1
<i>A. longiglumis</i>	63	28	12	9	3	0	2	0	0	0	0	7	1	0	2	0	0	3	0	1	1
<i>A. agadiriana</i>	62	29	12	9	2	0	1	1	0	0	0	5	4	0	0	0	0	3	0	1	1
<i>A. barbata</i>	66	27	14	7	2	0	1	2	1	0	0	7	1	2	1	0	0	4	1	0	1
<i>A. insularis</i>	63	30	11	8	2	0	2	0	1	0	0	6	1	3	0	0	0	3	0	1	1
<i>A. maroccana</i>	66	29	11	8	1	0	2	0	0	0	0	5	3	0	2	0	0	3	1	0	1
<i>A. murphyi</i>	63	29	12	8	2	0	1	1	0	0	0	7	0	1	2	0	0	3	1	0	1
<i>A. vaaviloviana</i>	66	27	13	9	1	0	1	1	2	0	0	7	1	2	1	0	0	4	1	0	1
<i>A. abyssinica</i>	66	27	13	10	0	0	2	1	1	0	0	7	3	0	1	0	0	4	1	0	1
<i>A. fatua</i>	64	30	9	9	2	0	1	1	0	0	0	7	1	0	2	0	0	3	1	0	1
<i>A. hybrida</i>	64	30	11	8	1	0	2	0	0	0	0	7	0	1	2	0	0	3	1	0	1
<i>A. occidentalis</i>	64	32	10	7	1	0	2	0	0	0	0	7	2	1	0	0	0	3	1	0	1
<i>A. sterilis</i>	64	29	9	11	1	0	1	1	0	0	0	7	1	2	0	0	0	3	1	0	1
<i>A. sativa</i>	64	29	10	9	2	0	1	1	0	0	0	7	3	0	0	0	0	3	1	0	1
Mean	64.4	28.9	13.1	7.2	1.6	0.4	1.5	0.6	0.5	0.0	0.1	6.3	1.9	1.2	0.6	0.0	0.0	3.4	0.7	0.3	1.0

The sliding window analysis of nucleotide diversity showed that the genomic region with the highest nucleotide diversity was SSC, followed by LSC, while two repeat regions (IRa and IRb) had the lowest nucleotide diversity (Figure 3). Three specific genome positions with the highest diversity were the sliding windows near 108,349, 59,494, and 81,000 (Figure 3).

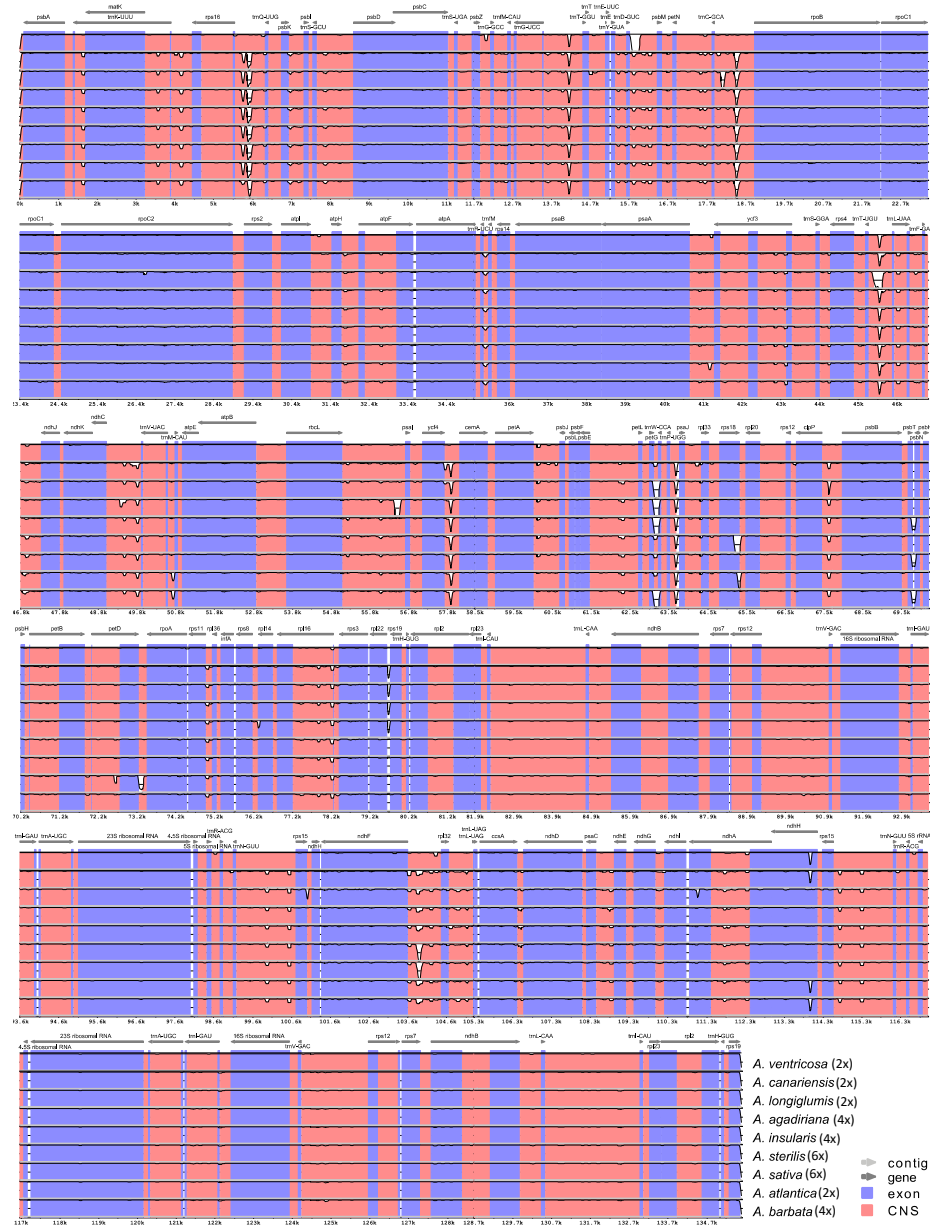


Figure 2. Percent identity plot by mVISTA of nine *Avena* chloroplast genome assemblies representing four diploid, three tetraploid, and two hexaploid *Avena* species, using *A. eriantha* as reference. Vertical scale indicates the percentage of identity ranging from 98.385% to 99.377%. Coding regions are in blue and non-coding regions are in orange.

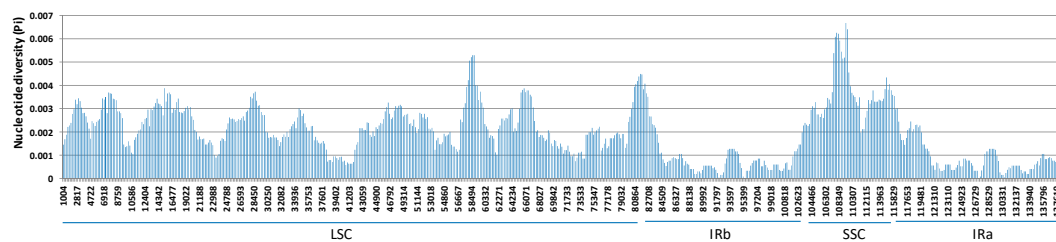


Figure 3. Nucleotide diversity (Π) from the sliding window analysis of 25 complete *Avena* chloroplast genome assemblies (window length: 2000 bp, step size 200 bp). X-axis: position of the window midpoint, Y-axis: nucleotide diversity within each window.

2.5. Selective Pressure Analysis

The positive selection analysis considered all 130 genes, with a total of 19,941 amino acids. The likelihood ratio tests for three models (M2a vs. M1a; M8 vs. M7; and M8 vs. M8a) appeared to suggest the presence of positive selections with p-values < 0.05 on many cp genes (Table 4). Based on the Naïve Empirical Bayes test, there were 114 codons and six codons showing positive selections with the posterior probability of 50% and 95%, respectively. However, if the Bayesian Empirical Bayes tests were used, there were only six codons showing positive selections with the posterior probability of 84% only. The six positively selected sites on the amino acids of six genes were 817 Gly (*matK*), 2150 Cys (*rpoB*), 4457 Val (*rpoC2*), 10,203 Glu (*rbcl*), 11,332 Leu (*rpl33*), and 16,682 Phe (*ccsA*). Additional maximum likelihood analysis of natural selection on individual codons identified the codon #17177 (CTA at the genome position of 55,930) for the amino acid leucine with the p-value of 0.0096. These results, together, indicate that positive selection was not strong and purifying selection was dominant, acting on these 25 oat cp genomes.

Table 4. Log-likelihood values (lnL) and parameter estimates under models of variable ω ratios among sites of 19,941 codons in 25 *Avena* chloroplast genomes.

Model Code	lnL	p-value for LRT ^a	Estimates of Parameters	Count of PSS ^b	
				NEB	BEB
M0 (One-Ratio)	−82795.020		$\omega = 0.11878$		
M3 (Discrete)	−82781.913	0.000013	$p_0 = 0.10199, p_1 = 0.85429, (p_2 = 0.04372), \omega_0 = 0.00000, \omega_1 = 0.00000, \omega_3 = 2.81808$	0(114)	
M1a (Nearly neutral)	−82785.669		$p_0 = 0.90153, (p_1 = 0.09847), \omega_0 = 0.00000, (\omega_1 = 1.00000)$		
M2a (Positive selection)	−82781.919	0.011758	$p_0 = 0.94578, p_1 = 0.02070, (p_2 = 0.03352), \omega_0 = 0.00110, (\omega_1 = 1.00000), \omega_2 = 2.98551$	114(0)	6(0)
M7 (Beta)	−82784.258		$p = 0.00500, q = 0.06086$		
M8 (Beta and ω)	−82781.940	0.049250	$p_0 = 0.97686, p = 0.05383, q = 1.11417, (p_1=0.02314), \omega=3.81079$	114(6)	6(0)
M8a (Beta and $\omega > 1$)	−82787.155		$p_0 = 0.95611, p = 0.07461, q = 0.93085, (p_1 = 0.04389), \omega = 1.00000$		
M8 (Beta and ω)	−82781.940	0.000671	$p_0 = 0.97686, p = 0.05383, q = 1.11417, (p_1 = 0.02314), \omega = 3.81079$	114(6)	6(0)

^a LRT = likelihood ratio test. ^b PSS = positively selected site; NEB = Naïve Empirical Bayes analysis; BEB = Bayesian Empirical Bayes analysis; and the first number is the count of PSS with posterior probabilities >50%, and the second number (in parenthesis) is the count of PSS with posterior probabilities >95%.

3. Discussion

Our cp genomic analysis here generated the first set of cp genome assemblies for these 25 oat species. The genomes typically had four regions (LSC, SSC, IRa, and IRb), with lengths of roughly 136,006 bp, and they carried 130 genes and four to six pseudogenes. Purifying selection was the dominant force acting on the cp genes. The genomes harbored 1313 SNPs and 277 SSRs per species. More nucleotide diversity was located in the SSR and LSC, rather than IRa and IRb, regions. These research outputs

allowed for a better understanding of oat cp genomes and evolution, and they formed an essential set of cp genomic resources for future oat studies and assessing oat genetic resources.

The analysis also revealed many comparative characteristics and some unique features of these oat cp genomes. First, the two genes *infA* and *rps16*, coding for a translation initiation factor 1 and a S16 ribosomal protein, respectively, were detected in these oat species, like other grass species, such as wheat and barley, but they were reported to be absent or nonfunctional in Malpighiales: *Passiflora edulis*, *Jatropha curcas*, and *Manihot esculenta* [29–31]. Second, a large sequence variation among 25 cp genomes was identified for the *ndhH* gene. Third, little differences in genomic structure and gene arrangement were identified across 25 species, and no marked genomic variations were unique to an oat ploidy level. Fourth, the positive selection was relatively weak acting on the 130 cp genes.

These comparative genomic features provided some empirical support for the previous inferences of oat evolution [9,18], as these genomes had the same gene count and arrangement, and purifying selection was the dominant force of selection acted on most of the 130 genes. Two C-genome species (*A. clauda* and *A. eriantha*) had only four pseudogenes, while the other 23 species had six. Such variation might be related to the major divergence of C-genome species 13–15 million years ago from A-genome species [18]. The indel variations for the *ndhH* gene seemed to be associated with the divergence of the As- and AB-genomes, but not with the divergence of hexaploid oat lineage. The cp genomic variations appeared to be not associated with the nuclear genome polyploidizations.

Managing more than 13,000 accessions of oat wild relatives is a challenging task, as many difficult issues must be properly addressed, such as taxonomic delimitation, duplication identification, viability monitoring, and field regeneration [17,32]. The cp genomic resources developed here can be utilized to develop taxonomic barcodes [8] for germplasm identification, e.g., from the region between the *psbZ* and *trnfM* (CAU) genes (see Supplementary File S4). Specific barcoding to identify wild relative germplasm of specific interests such as ecological types and geographic origin can also be developed. Molecular characterization of wild relative germplasm using cp markers, such as SNPs or SSRs, may be more informative than those using nuclear genomic tools, as cp markers are more conservative and nuclear genomic markers developed from polyploidy plants may not necessarily be informative. The inferred maternal phylogeny from these cp genomic resources (e.g., see [18]) can provide some guide for the search of evolutionarily related, economically important cp genes for gene introgression into oat breeding programs [15,16] through cp genome engineering [10]. The cp genome sequences are also essential for selecting intergenic spacer regions in oat cp genomes for transgene integration and assessing cp genome regulatory sequences in transgene expression [10]. We have initiated research to utilize these developed genomic resources to enhance the conservation, management and utilization of the oat collection in Plant Gene Resources of Canada.

4. Materials and Methods

4.1. Plant Material

We selected 25 *Avena* accessions of known species identity from the Plant Gene Resources of Canada (PGRC) oat collection, based on our previous oat research [3]. The selected accessions originated from various regions around the world and represent 25 species of the six botanical sections of the *Avena* genus, *Ventricosa*, *Agraria*, *Tenuicarpa*, *Pachycarpa*, *Ethiopica*, and *Avena*, and five distinct nuclear genomes organized in diploid (A or C), tetraploid (AB or AC), and hexaploid species (ACD). Table 1 shows the detailed information on the selected accessions with PGRC accession numbers, including botanical section and ploidy. About 300 seeds from each accession were planted in August 2013 in a 15 cm pot, grown for 8 to 10 days in the greenhouse at Saskatoon Research and Development Centre of Agriculture and Agri-Food Canada, and then incubated in the dark for 48 to 72 h. Up to 15 g from all 300 seedling leaves were collected and washed in cold water. Leaves were cut into 1 cm pieces with scissors, snap frozen with liquid nitrogen in a -20°C mortar, and then ground to a fine

powder. Ground samples, while still frozen, were transferred to 50 mL conical-bottom centrifuge tubes, cooled on dry ice, and then stored at -80°C , for up to one week.

4.2. DNA Extraction and MiSeq Sequencing

Plastid DNA isolation was performed following the method of Shi et al. [33] and optimized using the cp DNA extraction protocol developed by Diekmann et al. [34]. All the procedures were carried out on ice or at 4°C with buffers prechilled to 4°C . The enriched cp pellet was allowed to thaw at room temperature, and DNA was extracted using the Qiagen DNEasy Plant Mini kit standard method on a Qiacube robot (Qiagen, Mississauga, Canada) and eluted in 1/3x Qiagen AE buffer (3.33 mM Tris-Cl, 0.17 mM EDTA, pH9.0). DNA samples were quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies, Burlington, Canada). Final DNA yields ranged from 0.2 to $4.3\text{ ng}/\mu\text{l}$ and were diluted to $0.2\text{ ng}/\mu\text{l}$ with 10 mM Tris-HCl, with a pH of 8. The acquired cp DNAs were subjected to the genomic DNA library preparation with a Nextera XT DNA Library Preparation Kit (Illumina). Four MiSeq runs, each with six or seven libraries and pair-end of 250 bp, were performed to generate 25 forward and 25 reverse FASTQ files. All raw reads were deposited into the National Center for Biotechnology Information (NCBI), under the Bioproject PRJNA401438 (Table 1).

4.3. Chloroplast Genome Assembling and Annotation

All raw sequence reads were cleaned first with cutadapt [35] to remove sequence adapters and to perform quality trimming. Partial Nextera adapter sequence 'AGATGTGTATAAGAGACAG' was used to trim the raw sequence reads. All the sequence reads with both quality lower than 15 and shorter than 150 bp were discarded. SPAdes v3.11.1 [36] was used as the assembler for the circular cp genome assembly in the pair-end mode. Preliminary tests were performed to reach the least contigs number and the longest scaffold size by a series of combinations of different coverages and k-mer sizes. The k-mer size was eventually set to 127, and the coverage was set to 1000 folds, after a series of training analyses. The four major gaps located at the four junctions (LSC-IR, IR-SSC, SSC-IR, and IR-LSC) were filled in by the assistance of the four junction sequences. These junction sequences were obtained from the alignments of the scaffolds with their closely related species, including wheat (*Triticum aestivum*, NCBI Reference Sequence: AB042240.3; [37]), bent grass (*Agrostis stolonifera*; NCBI Reference Sequence: NC_008591.1), and ryegrass (*Lolium perenne*; NCBI Reference Sequence: NC_009950.1) cp genome. Each of the four junction sequences (ranging between 540 and 700 bp) containing both IR and another (either LSC or SSC) structure fragment was used as a bait to screen for reads for further gap sequence recovery. The selected reads from BLAST were also used to link adjacent structure fragments. The additional gaps located within the scaffolds of several *Avena* accessions were similarly filled with the assistance of the bait sequences acquired from either wheat or other *Avena* cp genomes with sequences at the same locations.

Gene annotations of 25 cp genomes were made using online DOGMA program [38], along with the cp genome annotations of wheat (NCBI Reference Sequence: AB042240), ryegrass (NCBI Reference Sequence: NC_009950), and bent grass (NCBI Reference Sequence: NC_008591.1). Manual curation was also made for the variations within coding genes, such as rRNA and tRNA, based on multiple sequence alignments with their closely related species in the Triticeae tribe. The circular maps of the *Avena* cp genomes were generated first with GenomeVx [39] and Circos [40] and finally merged in Inkscape (<https://inkscape.org>).

4.4. Comparative Genomic Analysis

To identify the genomic regions with substantial variability, the complete cp genomes of 25 *Avena* species were compared using mVISTA [41], with wheat cp genome as reference. For this comparison, the percent identity matrix among 25 cp genomes was also generated. To illustrate the genomic variations with respect to ploidy, further effort was also made, using *A. eriantha* cp genome as reference to compare nine cp genomes, representing diploid, tetraploid, and hexaploid species.

4.5. SNP, SSR, and Diversity Analysis

The SNP calling was performed on the basis of multiple sequence alignments (MSA) by SNP-sites with the default options [42]. MAFFT was used to generate MSA data with the FFT-NS-i×1000 alignment algorithm [43]. To identify SSRs, 25 cp genomes were analyzed using MISA [44], with the following setting of minimum numbers of repeats to 8, 5, 4, 3, 3, and 3 for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides, respectively. The sliding window diversity analysis was done using DnaSP v6 [45] to estimate nucleotide diversity across 25 oat species, with a sliding window of 2000 bp and step size of 200 bp.

4.6. Selective Pressure Analysis

We applied several site models (M0, M1, M2, M3, M7, and M8), implemented in codeml of PAML v 4.9i [46], to estimate the Ka/Ks and ω values, considering F3X4 codon frequencies. MAFFT, with the default options, was used to align the nucleotide sequences of all cp genes, and the phylogenetic tree of 25 *Avena* species, without tree-branch lengths, was obtained from the previous phylogenetic analysis [18]. Four nested site models (M3 vs. M0; M2 vs. M1; M8 vs. M7; and M8a vs. M8) were evaluated by log-likelihood ratio tests (LRT). The positively selected sites were analyzed by Naïve Empirical Bayes (NEB) analysis and Bayesian Empirical Bayes (BEB) analysis. Extra effort was also made to perform a maximum likelihood analysis of natural selection codon-by-codon, using MEGA 7 [47], following the method of Suzuki and Gojobori [48].

Supplementary Materials: The supplementary materials consist of five files as below. These files are accessible online on FigShare (DOI://10.6084/m9.figshare.9759449). File S1: A FASTA file for 25 complete *Avena* cp genome sequences; File S2: A FASTA file for a consensus genome sequence from 25 *Avena* cp genomes; File S3: An Excel file for gene annotation feature table; File S4: A pdf file for percent identity plot of 25 *Avena* cp genomes; File S5: An Excel file for 1313 cp SNPs across 25 cp genomes.

Author Contributions: Y.-B.F. conceived of the project, codesigned the research, conducted sequencing analysis, and cowrote the paper. P.L. performed sequence data analysis and cowrote the paper. B.B. codesigned the research and revised the paper. All authors read and approved the final manuscript.

Funding: This work was financially supported by an A-base Project of Agriculture and Agri-Food Canada (to Y.-B.F.) and the Beef Cattle Research Council of Canada (to B.B.).

Acknowledgments: We would like to thank Mr. Gregory Peterson and Ms. Carolee Horbach for their technical assistance in the research, as well as Dr. Isobel Parkin for the access to and use of the Illumina MiSeq instrument.

Conflicts of Interest: The authors declare no conflicts of interest.

Ethical Standards: The writing process of this manuscript complies with the current laws of Canada.

References

1. Castañeda-Álvarez, N.P.; Khoury, C.K.; Achicanoy, H.A.; Bernau, V.; Dempewolf, H.; Eastwood, R.J.; Guarino, L.; Harker, R.H.; Jarvis, A.; Maxted, N.; et al. Global conservation priorities for crop wild relatives. *Nat. Plants* **2016**, *2*, 16022. [[CrossRef](#)] [[PubMed](#)]
2. Greene, S.L.; Warburton, M.L. Wading into the gene pool: Progress and constraints using wild species. *Crop. Sci.* **2017**, *57*, 1039. [[CrossRef](#)]
3. Fu, Y.-B.; Williams, D.J. AFLP variation in 25 *Avena* species. *Theor. Appl. Genet.* **2008**, *117*, 333–342. [[CrossRef](#)] [[PubMed](#)]
4. Kole, C. *Wild Crop Relatives: Genomic and Breeding Resources*; Springer: Berlin/Heidelberg, Germany, 2011.
5. Brozynska, M.; Furtado, A.; Henry, R.J. Genomics of crop wild relatives: Expanding the gene pool for crop improvement. *Plant Biotechnol. J.* **2016**, *14*, 1070–1085. [[CrossRef](#)] [[PubMed](#)]
6. Yan, H.; Bekele, W.A.; Wight, C.P.; Peng, Y.; Langdon, T.; Latta, R.G.; Fu, Y.-B.; Diederichsen, A.; Howarth, C.J.; Jellen, E.N.; et al. High-density marker profiling confirms ancestral genomes of *Avena* species and identifies D-genome chromosomes of hexaploid oat. *Theor. Appl. Genet.* **2016**, *129*, 2133–2149. [[CrossRef](#)]
7. Loskutov, I.G.; Rines, H.W. *Avena*. In *Wild Crop Relatives: Genomic and Breeding Resources*; Kole, C., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 109–183.

8. Coissac, E.; Hollingsworth, P.M.; Lavergne, S.; Taberlet, P. From barcodes to genomes: Extending the concept of DNA barcoding. *Mol. Ecol.* **2016**, *25*, 1423–1428. [[CrossRef](#)]
9. Liu, Q.; Lin, L.; Zhou, X.; Peterson, P.M.; Wen, J. Unraveling the evolutionary dynamics of ancient and recent polyploidization events in *Avena* (Poaceae). *Sci. Rep.* **2017**, *7*, 44162. [[CrossRef](#)]
10. Daniell, H.; Lin, C.-S.; Yu, M.; Chang, W.-J. Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. *Genome Boil.* **2016**, *17*, 134. [[CrossRef](#)]
11. Strychar, R. World oat production, trade, and usage. In *Oats: Chemistry and Technology*; Webster, F., Wood, P., Eds.; AACC International, Inc.: Saint Paul, MN, USA, 2011; pp. 1–10.
12. Baum, B.R. *Oats: Wild and Cultivated. A monograph of the Genus Avena L. (Poaceae)*; Minister of Supply and Services: Ottawa, ON, Canada, 1977.
13. FAO. *The Second Report on the State of the World's Plant Genetic Resources*; FAO: Rome, Italy, 2010.
14. Dempewolf, H.; Baute, G.; Anderson, J.; Kilian, B.; Smith, C.; Guarino, L. Past and future use of wild relatives in crop breeding. *Crop. Sci.* **2017**, *57*, 1070. [[CrossRef](#)]
15. Rines, H.W.; Porter, H.L.; Carson, M.L.; Ochocki, G.E. Introgression of crown rust resistance from diploid oat *Avena strigosa* into hexaploid cultivated oat *A. sativa* by two methods: Direct crosses and through an initial 2 × 4 × synthetic hexaploid. *Euphytica* **2007**, *158*, 67–79. [[CrossRef](#)]
16. Aung, T.; Zwer, P.; Park, R.; Davies, P.; Sidhu, P.; Dundas, I. Hybrids of *Avena sativa* with two diploid wild oats (Clav6956) and (Clav7233) resistant to crown rust. *Euphytica* **2010**, *174*, 189–198. [[CrossRef](#)]
17. Fu, Y.-B. The vulnerability of plant genetic resources conserved *ex situ*. *Crop. Sci.* **2017**, *57*, 2314. [[CrossRef](#)]
18. Fu, Y.-B. Oat evolution revealed in the maternal lineages of 25 *Avena* species. *Sci. Rep.* **2018**, *8*, 4252. [[CrossRef](#)] [[PubMed](#)]
19. Steer, M.W.; Holden, J.H.W.; Gunning, B.E.S. *Avena* chloroplasts: Species relationships and the occurrence of stromacentres. *Can. J. Genet. Cytol.* **1970**, *12*, 21–27. [[CrossRef](#)]
20. Murai, K.; Tsunewaki, K. Chloroplast genome evolution in the genus *Avena*. *Genet.* **1987**, *116*, 613–621.
21. Gengenbach, B.G.; Boylan, K.L.; Storey, K.K.; Rines, H.W. Mitochondrial DNA diversity in oat cultivars and species. *Crop. Sci.* **1988**, *28*, 171–176.
22. Peng, Y.-Y.; Wei, Y.-M.; Baum, B.R.; Jiang, Q.-T.; Lan, X.-J.; Dai, S.-F.; Zheng, Y.-L. Phylogenetic investigation of *Avena* diploid species and the maternal genome donor of *Avena* polyploids. *Taxon* **2010**, *59*, 1472–1482. [[CrossRef](#)]
23. Yan, H.-H.; Baum, B.R.; Zhou, P.-P.; Zhao, J.; Wei, Y.-M.; Ren, C.-Z.; Xiong, F.-Q.; Liu, G.; Zhong, L.; Zhao, G.; et al. Phylogenetic analysis of the genus *Avena* based on chloroplast intergenic spacer psbA-trnH and single-copy nuclear gene *Acc1*. *Genome* **2014**, *57*, 267–277. [[CrossRef](#)]
24. Loskutov, I.G. On evolutionary pathways of *Avena* species. *Genet. Resour. Crop Evol.* **2008**, *55*, 211–220. [[CrossRef](#)]
25. Soltis, D.E.; Gitzendanner, M.A.; Stull, G.; Chester, M.; Chanderbali, A.; Chamala, S.; Jordon-Thaden, I.; Soltis, P.S.; Schnable, P.S.; Barbazuk, W.B. The potential of genomics in plant systematics. *Taxon* **2013**, *62*, 886–898. [[CrossRef](#)]
26. Fu, Y.-B.; Dong, Y.; Yang, M.-H. Multiplexed shotgun sequencing reveals congruent three-genome phylogenetic signals for four botanical sections of the flax genus *Linum*. *Mol. Phylogenet. Evol.* **2016**, *101*, 122–132. [[CrossRef](#)] [[PubMed](#)]
27. Twyford, A.D.; Ness, R.W. Strategies for complete plastid genome sequencing. *Mol. Ecol. Resour.* **2017**, *17*, 858–868. [[CrossRef](#)] [[PubMed](#)]
28. Tonti-Filippini, J.; Nevill, P.G.; Dixon, K.; Small, I. What can we do with 1000 plastid genomes? *Plant J.* **2017**, *90*, 808–818. [[CrossRef](#)] [[PubMed](#)]
29. Asif, M.H.; Mantri, S.S.; Sharma, A.; Srivastava, A.; Trivedi, I.; Gupta, P.; Mohanty, C.S.; Sawant, S.V.; Tuli, R. Complete sequence and organisation of the *Jatropha curcas* (Euphorbiaceae) chloroplast genome. *Tree Genet. Genomes* **2010**, *6*, 941–952. [[CrossRef](#)]
30. Cauz-Santos, L.A.; Munhoz, C.F.; Rodde, N.; Cauet, S.; Santos, A.A.; Penha, H.A.; Dornelas, M.C.; Varani, A.M.; Oliveira, G.C.X.; Bergès, H.; et al. The chloroplast genome of *Passiflora edulis* (Passifloraceae) assembled from long sequence reads: structural organization and phylogenomic studies in Malpighiales. *Front. Plant Sci.* **2017**, *8*, 941. [[CrossRef](#)]

31. Daniell, H.; Wurdack, K.J.; Kanagaraj, A.; Lee, S.-B.; Saski, C.; Jansen, R.K. The complete nucleotide sequence of the cassava (*Manihot esculenta*) chloroplast genome and the evolution of *atpF* in Malpighiales: RNA editing and multiple losses of a group II intron. *Theor. Appl. Genet.* **2008**, *116*, 723–737. [[CrossRef](#)]
32. Zhang, H.; Mittal, N.; Leamy, L.J.; Barazani, O.; Song, B.-H. Back into the wild—apply untapped genetic diversity of wild relatives for crop improvement. *Evol. Appl.* **2016**, *10*, 5–24. [[CrossRef](#)]
33. Shi, C.; Hu, N.; Huang, H.; Gao, J.; Zhao, Y.-J.; Gao, L.-Z. An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. *PLoS ONE* **2012**, *7*, e31468. [[CrossRef](#)]
34. Diekmann, K.; Hodkinson, T.R.; Fricke, E.; Barth, S. An optimized chloroplast DNA extraction protocol for grasses (Poaceae) proves suitable for whole plastid genome sequencing and SNP detection. *PLoS ONE* **2008**, *3*, e2813. [[CrossRef](#)]
35. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **2011**, *17*, 10. [[CrossRef](#)]
36. Nurk, S.; Bankevich, A.; Antipov, D.; Gurevich, A.; Korobeynikov, A.; Lapidus, A.; Prjibelsky, A.; Pyshkin, A.; Sirotkin, A.; Sirotkin, Y.; et al. Assembling genomes and mini-metagenomes from highly chimeric reads. In Proceedings of the Computer Vision—ECCV 2012, Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2013; Volume 7821, pp. 158–170.
37. Ogihara, Y.; Isono, K.; Kojima, T.; Endo, A.; Hanaoka, M.; Shiina, T.; Terachi, T.; Utsugi, S.; Murata, M.; Mori, N.; et al. Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. *Mol. Genet. Genom.* **2002**, *266*, 740–746.
38. Wyman, S.K.; Jansen, R.K.; Boore, J.L. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **2004**, *20*, 3252–3255. [[CrossRef](#)] [[PubMed](#)]
39. Conant, G.C.; Wolfe, K.H. GenomeVx: Simple web-based creation of editable circular chromosome maps. *Bioinformatics* **2008**, *24*, 861–862. [[CrossRef](#)] [[PubMed](#)]
40. Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; Jones, S.J.; Marra, M.A. Circos: An information aesthetic for comparative genomics. *Genome Res.* **2009**, *19*, 1639–1645. [[CrossRef](#)] [[PubMed](#)]
41. Frazer, K.A.; Pachter, L.; Poliakov, A.; Rubin, E.M.; Dubchak, I. VISTA: Computational tools for comparative genomics. *Nucleic Acids Res.* **2004**, *32*, W273–W279. [[CrossRef](#)] [[PubMed](#)]
42. Page, A.J.; Taylor, B.; Delaney, A.J.; Soares, J.; Seemann, T.; Keane, J.A.; Harris, S.R. SNP-sites: Rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genom.* **2016**, *2*, e000056. [[CrossRef](#)]
43. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [[CrossRef](#)]
44. Thiel, T.; Michalek, W.; Varshney, R.; Graner, A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* **2003**, *106*, 411–422. [[CrossRef](#)]
45. Rozas, J.; Ferrer-Mata, A.; Sánchez-DelBarrio, J.C.; Guirao-Rico, S.; Librado, P.; Ramos-Onsins, S.E.; Sánchez-Gracia, A. DnaSP 6: DNA sequence polymorphism analysis of large datasets. *Mol. Biol. Evol.* **2017**, *34*, 3299–3302. [[CrossRef](#)]
46. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **2007**, *24*, 1586–1591. [[CrossRef](#)]
47. Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **2016**, *33*, 1870–1874. [[CrossRef](#)] [[PubMed](#)]
48. Suzuki, Y.; Gojobori, T. A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* **1999**, *16*, 1315–1328. [[CrossRef](#)] [[PubMed](#)]

