TÜBİTAK

Turkish Journal of Agriculture and Forestry

http://journals.tubitak.gov.tr/agriculture/

Research Article

Turk J Agric For
(2022) 46: 525-535
© TÜBİTAK
doi:10.55730/1300-011X.3023

De novo assembly and characterisation of chloroplast genomes of broccoli cvs. Marathon and Green sprout using next generation sequencing

Ayisha ZIA[1,*], Misbah ZAHOOR[1,*], ABDULLAH[1], Asma NISAR[1], Neelam BATOOL[1], Amana BIBI[1], Kiran SABA[1,2], Ibrar AHMED[3], Arzu KARATAŞ[4], Ekrem GÜREL[5], Mohammad Tahir WAHEED[1,]**

[1]Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad, Pakistan
[2]Department of Biochemistry, Faculty of Life Sciences, Shaheed Benazir Bhutto Women University, Peshawar, Pakistan.
[3]Alpha Genomics Private Limited, Islamabad, Pakistan
[4]Department of Horticulture, Faculty of Agriculture, Recep Tayyip Erdoğan University, Rize, Turkey
[5]Department of Biology, Faculty of Science and Literature, Bolu Abant İzzet Baysal University, Bolu, Turkey

Received: 31.03.2022 • **Accepted/Published Online:** 21.07.2022 • **Final Version:** 09.08.2022

Abstract: The genus *Brassica* (family *Brassicaceae*) includes nutritionally and economically important species such as *Brassica napus*, *Brassica rapa*, and *Brassica oleracea*. Many varieties of *B. oleracea* are available in various morphological forms including nutritive vegetables such as cauliflower (var. *botrytis*), Brussels sprouts (var. *gemmifera*), kales and collards (var. *acephala*), kohlrabi (var. *gongylodes*), cabbage (var. *capitata*), and broccoli (var. *italica*). Objective of the present study was to sequence chloroplast genomes of two cultivars of broccoli: Marathon and Green sprout. The sequencing was done by next generation sequencing. The analysis was performed using Velvet, Geneious, GeSeq, tRNAscan-SE, ARAGORN, OrganellarGenomeDRAW, IRscope and REPuter. The genomes of both cultivars showed highly similar quadripartite structure of 153,365 bp. The LSC (83,136 bp) and SSC (17,835 bp) regions were separated by a pair of IR (26,197 bp) region. In total, 114 unique genes were present in both species, including 80 protein-coding, 30 tRNA and 4 rRNA genes, while 18 genes were duplicated in IRs. The highest amino acid encoding frequency was found for Leucine whereas cysteine was the least encoding amino acid. The codon usage analyses confirmed high encoding efficacy of codons that ended at 3′-end with A/T. Repeat analyses of these genomes revealed 415 microsatellites and 36 oligonucleotide repeats. Microsatellites motifs were mostly comprised of A/T instead of C/G. The comparative analyses confirmed the presence of 17 substitutions between both cultivars. Overall, this study will increase knowledge about the chloroplast genomes of broccoli and will provide a resource for chloroplast genetic engineering of this important edible plant.

Key words: Broccoli, next generation sequencing, chloroplast genome, oligonucleotide repeats, simple sequence repeats, codon usage

1. Introduction

Family Brassicaceae (Cruciferae) belongs to the order Brassicales. This family comprises 328 genera and 3628 species (Christenhusz and Byng, 2016). Species of the family *Brassicaceae* are distributed primarily in the temperate and alpine areas of all continents except Antarctica and used as food, condiments, oils and weed (Al-Shehbaz, 2001). The economically important plants belong to various genera, but most important genera are *Brassica*, *Raphanus*, *Nasturtium*, *Lepidium*, and *Eruca* (Al-Shehbaz, 2001).

The genus *Brassica* is very important due to economic importance and nutrient rich food, containing rutabaga and rape (*Brassica napus*), turnip (*Brassica rapa*), and *Brassica oleracea* (Anjum et al., 2012). Different varieties of *Brassica oleracea* exist in various morphological forms and include important vegetables such as: cauliflower (var. *botrytis*), Brussels sprouts (var. *gemmifera*), kales and collards (var. *acephala*), kohlrabi (var. *gongylodes*), cabbage (var. *capitata*), and broccoli (var. *italica*) with high nutritional value (Al-Shehbaz, 2001; Anjum et al., 2012; Aires, 2015).

Broccoli (var. *italica*) is considered very important due to its nutritional composition. Broccoli contains proteins, carbohydrates, lipids and fibres (Anjum et al., 2012). This plant also contains important minerals including calcium, iron, magnesium, phosphorus, potassium, sodium and zinc, whereas the important vitamins including

* These authors contributed equally to this manuscript.
** Correspondence: tahirwaheed@qau.edu.pk

Vitamins A, B_2, B_3, B_6, B_{12}, C, D, E, and K along with the important compounds glycosylates (Owis, 2015). Various important medicinal activities of Broccoli have also been reported i.e. anticancers, antioxidative, antimicrobial, antidiabetic, antiinflammatory, antiobesity along with hepatoprotective, immunomodulatory, cardioprotective, and gastroprotective properties (Owis, 2015).

The chloroplast is an important organelle which plays important role in photosynthesis and in the synthesis of amino acids and fatty acids (Daniell et al., 2016). The chloroplast genome contains its circular double stranded DNA and mostly exists in angiosperm as quadripartite structure in which single copy regions are separated by a pair of long inverted repeat regions (Palmer, 1985; Daniell et al., 2016; Abdullah et al., 2020) Chloroplast genome polymorphism is helpful for the study of phylogenetic (Shahzadi et al., 2019; Shahzadi et al., 2020; Henriquez et al., 2020a) to population genetics (Ahmed, 2014; Zhang et al., 2020). Different types of mutations events take place in the chloroplast genome, including substitutions, InDels, inversion, tandem repeats, and copy number variations even mutational events are also reported within the cultivars of same species (Ahmed et al., 2012; Ahmed et al., 2013; Xu et al., 2015; Iram et al., 2019; Mehmood et al., 2020; Waseem et al., 2020). Chloroplast genome is also used for the genetic transformation to get high expression of the desired gene with stable and high quality proteins (Ahmed et al., 2010; Lössl, 2011; Waheed et al., 2011a, 2011b; Khan et al., 2018).

In the current study, we de novo assembled chloroplast genomes of two cultivars of *Brassica oleracea* (var. *italica*): including Marathon and Green sprout to characterise chloroplast genome structure of these two important cultivars. The comparative analyses of amino acid frequency, codon usage, simple sequence repeats, and oligonucleotide repeats revealed high similarities and between these two cultivars only 17 substitutions were found. These genomic resources will be helpful for the identification of suitable regions for the purpose of chloroplast transformation.

2. Materials and methods

2.1. DNA extraction and sequencing

The whole genomic DNA two cultivars of broccoli, Marathon and Green sprout, were extracted from in vitro grown fresh leaves using DNeasy plant mini kit (Qiagen) according to manufacturer's protocol. The quality and quantity of DNA was confirmed by 1% agarose gel electrophoresis and nanodrop. High quality DNA was sent for sequencing to Novogene, Hong Kong. They sequenced DNA from pair end with short reads of 150 bp using next generation sequencing machine Hiseq 2500.

2.2. Chloroplast genome assembly and annotation

The quality of raw reads was confirmed by fastQC analyses (Andrews, 2018). The high-quality reads were de novo assembled by using Velvet 1.2.10 (Zerbino, 2008) following Abdullah et al. (2020) with various kmers values of 71, 91, 111, and 121. The generated contigs were combined by using de novo assembly option in Geneious R8.1 (Kearse et al., 2012). The boundaries of single copy regions (LSC: Large single copy and SSC: Small single copy) and inverted repeat regions were determined by manual inspection of scaffold regions. The validation of the assembled genome and coverage depth analyses were performed by mapping short reads to the respective genome using Burrow Wheel Aligner (BWA) (Li and Durbin, 2009) and visualisation in Tablet (Milne et al., 2010). We used GeSeq (Tillich et al., 2017) for the annotation of the genomes along with tRNAscan-SE v.2.0 (Lowe and Chan, 2016) and ARAGORN (Laslett and Canback, 2004). The circular map of the genome was drawn by using OrganellarGenomeDRAW v.1.3.1 (Greiner et al., 2019). These genomes were submitted to the National Center for Biotechnology Information (NCBI) under accession number MH388765 (*Brassica oleracea* cv. Marathon) and MH388764 (*Brassica oleracea* cv. Green Sprout).

2.3. Characterisation of chloroplast genome and comparative analyses

The characteristics of these genomes were analysed in Geneious R8.1 including amino acid frequency and codon usage (Kearse et al., 2012). The IRscope (Amiryousefi, 2018) was used to compare inverted repeat (IR) contraction and expansion among six species of *Brassica*.

The oligonucleotide repeat analysis was performed by REPuter (Kurtz et al., 2001) to detect forward, palindromic, reverse, and complementary repeat with minimum repeat size of 30 bp and 90% similarity. The simple sequence repeats (SSR) analysis was performed by using MISA software (Beier et al., 2007). One of the IR region was not included in the analysis to avoid the over representation of inverted repeat regions. The repeat unit was adjusted with a minimum value of seven nucleotides for mononucleotides, four for dinucleotides, and three for tri, tetra, penta, and hexa nucleotides. The number of repeats in LSC, SSC and IR regions was determined along with number of different types of repeats and motifs. We used MAFFT (Multiple Alignment using Fast Fourier Transform) for the determination of nucleotide differences between both cultivars.

3. Results

3.1. Features of the chloroplast genome of two cultivars of broccoli

The chloroplast genomes of both cultivars showed the same quadripartite structure of 153,165 bp. The LSC (83,136

bp) and SSC (17,835 bp) were separated by a pair of IRs (26,197 bp). The gene content was found to be identical in the chloroplast genomes of these two cultivars of the genus *Brassica*. The chloroplast genome of *Brassica* had 114 unique genes in which 80 were protein-coding genes, 30 were tRNA genes and 4 of them were rRNA genes. Among these, 18 genes were duplicated in IR regions. 18 genes contained introns in which 12 were protein-coding genes whereas 6 were tRNA genes. The *rps12* was a trans-spliced gene, having 5′ part in the LSC region and 3′ part in the IR regions, where rest of it was duplicated. Out of total 12 genes, 10 genes contained single intron whereas 2 genes including *ycf3* and *clpP* contained two introns. The comparative analyses of the complete chloroplast genome of both cultivars are given in Table 1. The circular map of both species is provided in Figure 1.

3.2. Amino acid frequency and codon usage

Detailed comparison of amino acid frequencies of genomes of these broccoli cultivars indicated that greater percentages of hydrophobic amino acids were encoded whereas uncharged polar amino acids were encoded in fewer amounts. Basic amino acids were also less prevalent i.e. Histidine (H). Amino acid frequencies of *Brassica* genomes indicated that Leucine (L) was the most abundant amino acid followed by the Isoleucine (Ile), whereas Tryptophan (W) and Cysteine (C) were found least prevalent. The graphical representation of amino acid frequency is shown in Figure 2. The RSCU (relative synonymous codon usage) value revealed that the codons which had A or T at the 3rd nucleotide positions were more frequently present in the genomes as compared to the codons having G or C at third nucleotide position. The RSCU values of both cultivars of broccoli are given in Table 2.

3.3. Boundary region of chloroplast genome

We also compared the boundary regions of the six species of genus *Brassica*. The analysis revealed the presence of *rps19* gene at LSC/IRb border. This gene originated from LSC and entered IRb to almost 113 bp. The *ycf1* gene was present at the border of SSC-IR region, leading to the formation of pseudogene *(ycf1$^{\psi}$)* of 1028 bp. The *ycf1* gene exceeded the border (IRb/SSC and SSC/IRa) from 1 to 1027 bp. The *ndhF* gene was also present at IRb/SSC border. This gene started from IRb region and crossed SSC with 2204 bp. These two genes, *ycf1* and *ndhF*, also overlapped at IRb/SSC border for about 39 bp. IR expansion led to

Table 1. Summary of the chloroplast genomes of both cultivars of *Brassica oleracea*.

Characteristics		***B. oleracea* var. *italica* cv. Marathon; cv. Green sprout**
Size (bp)		153,365
LSC length (bp)		83,136
SSC length (bp)		17,835
IR length (bp)		26,197
Number of genes		114
Protein-coding genes		80
tRNA genes		30
rRNA genes		4
Duplicate genes		18
GC content	Total (%)	36.4%
	LSC (%)	34.1%
	SSC (%)	29.1%
	IR (%)	42.3%
	CDS (%)	37.7%
	rRNA (%)	55.4%
	tRNA (%)	52.3%
	All genes (%)	37.3%
Protein coding part (CDS) (% bp)		49.0%
All gene (% bp)		68.90%
Noncoding region (% bp)		32.1%

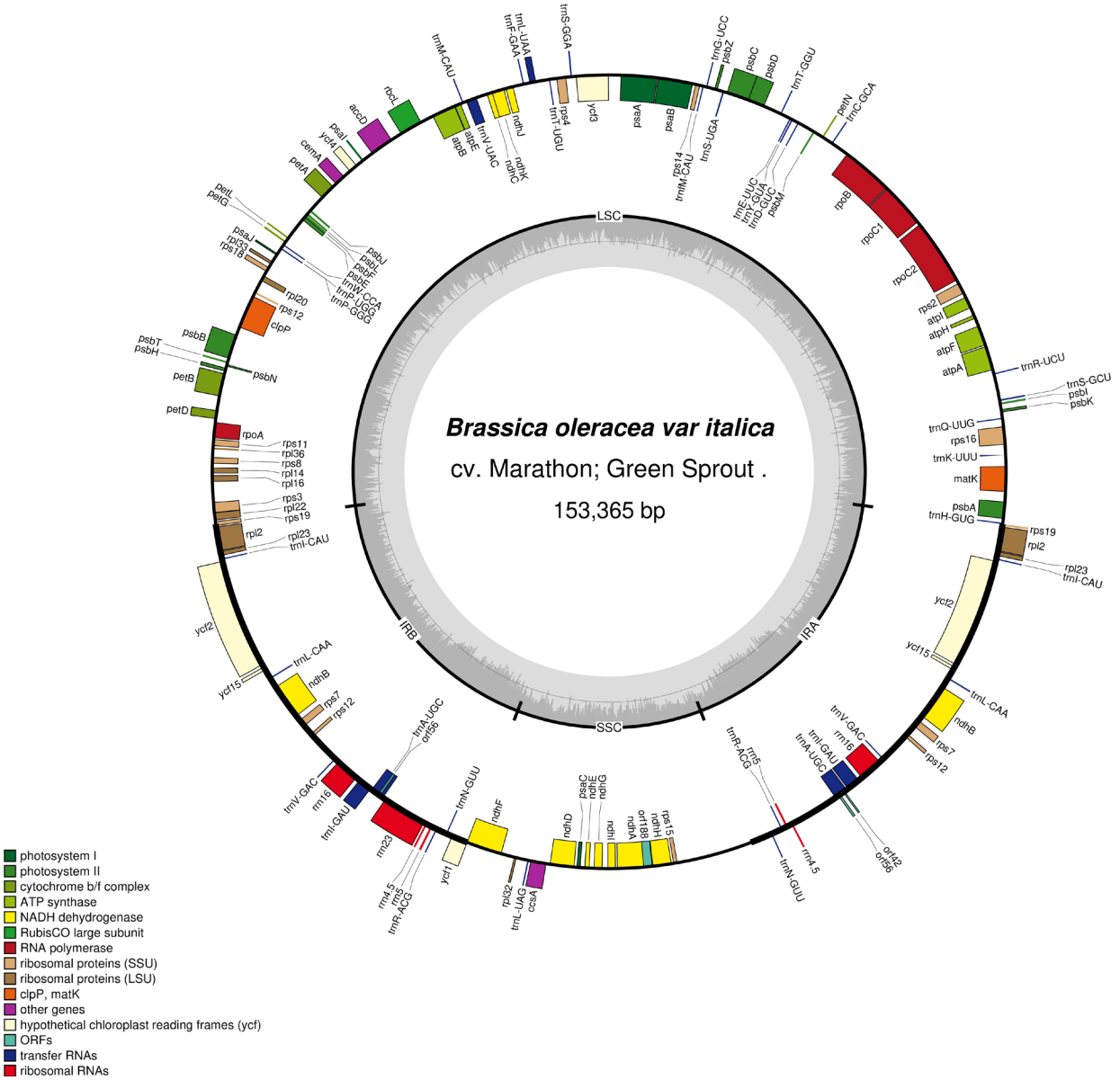

Figure 1. Circular map of chloroplast genome *Brassica oleracea* var. *italica*. Genes are coded based on their function. Genes present inside transcribe anticlockwise, and genes present outside transcribe clockwise.

the duplication of *rpI2* gene in all genomes. The gene *trnH* was present near IRa/LSC border. It was completely located in LSC region, 3 bp away from the IRa region in all the genomes. The complete analyses of IR expansion and contraction as well as the position of genes at junctions of the chloroplast genome region are shown in Figure 3.

3.4. Oligonucleotide repeats and simple sequence repeat analyses

We identified in *B. oleracea* species 36 repeats (F = 8, P = 23, R = 4, C = 1). The size of repeats varied from 30 bp to 47 bp. Among all the repeats, palindromic (P) repeats were most abundant (23) followed by the forward repeats (8) and then by reverse (4) and complementary (1) repeats. Most of the repeats lied in the intergenic spacer region as compared to coding and intronic regions (Table 3).

The SSR analysis revealed 415 SSRs in *B. oleracea*. Among mononucleotides, A/T motifs were most abundant, whereas in dinucleotide AT/TA SSR's were most abundant whereas AAT/ATT SSR's comprised most of trinucleotides. Microsatellite repeats were most abundant in LSC region followed by the SSC region and then by IR regions. The repeat unit of mononucleotide SSRs ranged from 7 to 15 units, dinucleotides ranged between 4 to 7

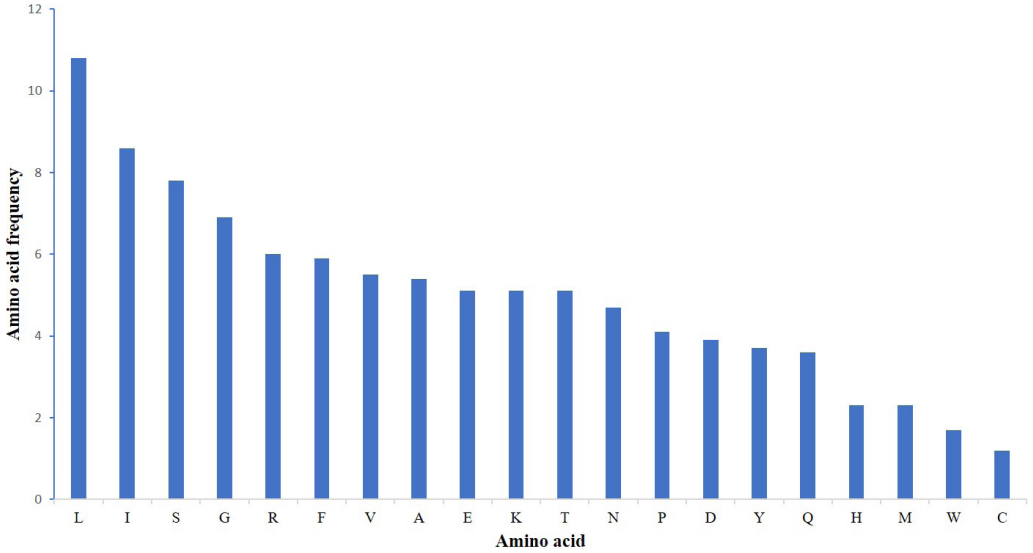

Figure 2. Amino acid frequency in chloroplast genomes of *Brassica oleracea* var. *italic.*

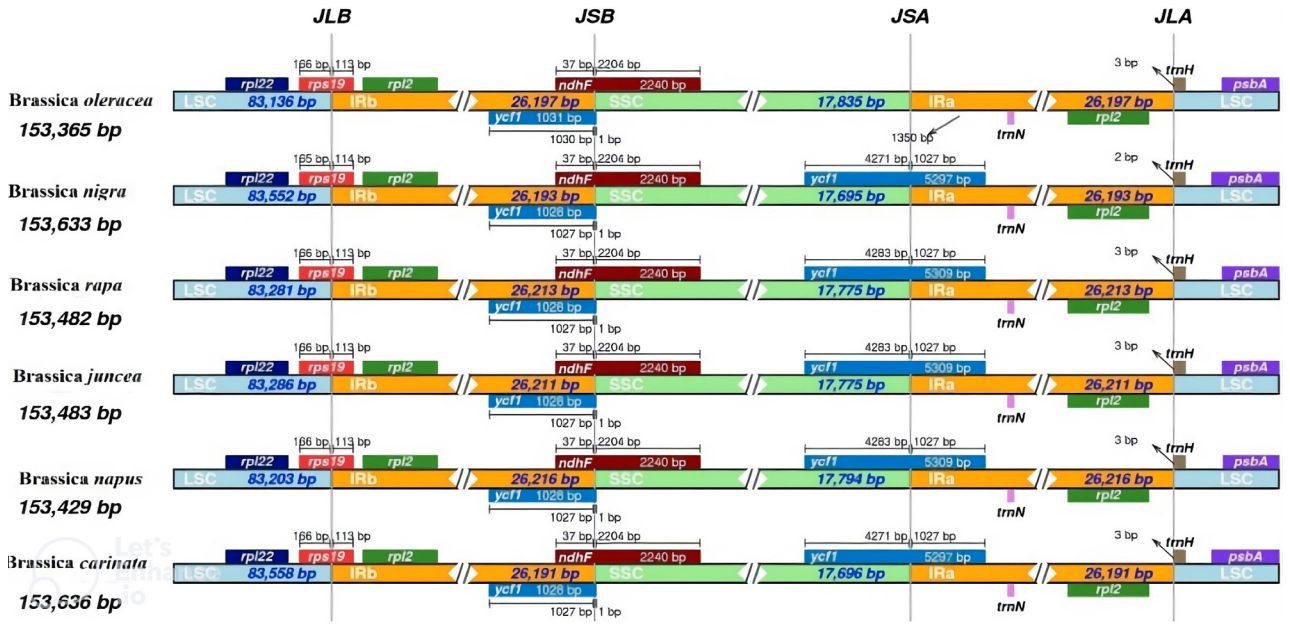

Figure 3. Comparative analysis of boundary regions: Large single copy (LSC), small single copy (SSC) and inverted repeated (IR) regions among six species of genus *Brassica* including *Brassica oleracea cv.* Marathon, *Brassica nigra*, *Brassica rapa*, *Brassica juncea*, *Brassica napus* and *Brassica carinata*. The sequences clearly depict that there are no differences between their boundaries. Lengths of arrows are illustrated. Base pairs indicate the distance of the gene from boundaries. The size of the complete chloroplast genome is given on left side.

units whereas the tri, tetra, and penta nucleotide SSRs repeat mostly existed in three repeat units (Table 4).

4. Discussion

The length of chloroplast genome sequences of both species was found equal in size. This shows that these cultivars have high resemblance regarding their chloroplast genome sequence length which showed substitutions of nucleotide at 17 positions. Therefore, similar features were observed for both species. The GC content fluctuated in the different regions of chloroplast genome and the IRs regions had high GC content due to the presence of ribosomal RNA genes and tRNA genes. Here, our results are similar to the previous studies of angiosperm chloroplast genomes

Table 2. Codon usage analyses in *Brassica oleracea.*

S. No	Codon	Amino acid	codon usage	S. No	Codon	Amino acid	codon usage
1	GCA	Alanine	1.10	33	CCA	Proline	1.14
2	GCC	Alanine	0.61	34	CCC	Proline	0.73
3	GCG	Alanine	0.42	35	CCG	Proline	0.53
4	GCT	Alanine	1.85	36	CCT	Proline	1.58
5	TGC	Cysteine	0.48	37	CAA	Glutamine	1.53
6	TGT	Cysteine	1.51	38	CAG	Glutamine	0.46
7	GAC	Aspartic acid	0.38	39	AGA	Arginine	1.75
8	GAT	Aspartic acid	1.61	40	AGG	Arginine	0.63
9	GAA	Glutamic acid	1.49	41	CGA	Arginine	1.36
10	GAG	Glutamic acid	0.50	42	CGC	Arginine	0.42
11	TTC	Phenylalanine	0.66	43	CGG	Arginine	0.49
12	TTT	Phenylalanine	1.33	44	CGT	Arginine	1.32
13	GGA	Glycine	1.65	45	AGC	Serine	0.38
14	GGC	Glycine	0.38	46	AGT	Serine	1.21
15	GGG	Glycine	0.65	47	TCA	Serine	1.18
16	GGT	Glycine	1.30	48	TCC	Serine	0.87
17	CAC	Histidine	0.50	49	TCG	Serine	0.58
18	CAT	Histidine	1.49	50	TCT	Serine	1.75
19	ATA	Isoleucine	1.74	51	ACA	Threonine	1.21
20	ATC	Isoleucine	1.09	52	ACC	Threonine	0.71
21	ATT	Isoleucine	1.49	53	ACG	Threonine	0.42
22	AAA	Lysine	1.49	54	ACT	Threonine	1.63
23	AAG	Lysine	0.50	55	GTA	Valine	1.44
24	CTA	Leucine	0.82	56	GTC	Valine	0.50
25	CTC	Leucine	0.40	57	GTG	Valine	0.56
26	CTG	Leucine	0.36	58	GTT	Valine	1.48
27	CTT	Leucine	1.24	59	TGG	Tryptophan	1
28	TTA	Leucine	2.01	60	TAC	Tyrosine	0.38
29	TTG	Leucine	1.13	61	TAT	Tyrosine	1.61
30	ATG	Methionine	1	62	TAA	Stop codon	1.80
31	AAC	Asparagine	0.46	63	TAG	Stop codon	0.71
32	AAT	Asparagine	1.53	64	TGA	Stop codon	0.47

Table 3. Oligonucleotide repeats in *Brassica oleracea*.

S. No	Type	Region	Functional regions	size	sequence
1	R	IR	*trnH-psbA/trnS*	31	TATTTTTTTCTATTTTATATATAATAGAAAA
2	P	LSC	*trnH-psbA*	40	ATAGTATTTTTTTCTATTTTATATATAATAGAAAAAAATA
3	R	LSC/SSC	*trnH-psbA/rpI32-ccsA*	30	TTTATATATAATAGAAAAAAATATATAAAA
4	P	LSC	*rps16-psbK*	32	AAATCTATATTTATATATATAAATATAGATTT
5	P	LSC	*rps16-psbK*	41	TTTGAATAGAAATCTATATTTATATATATAAATATAGATTT
6	R	LSC/SSC	*trnQ-psbK/ndhA*	32	TTTCTTTATTTATTTTTTTTTTTCCTTATATA
7	P	LSC	*psbI-trnS/trnS*	30	AGGGAAAGAGAGGGATTCGAACCCTCGGTA
8	F	LSC	*PsbI-trnS*	31	AAGGGAAAGAGAGGGATTCGAACCCTCGGTA
9	R	LSC	*trnS-trnR*	33	TTTTATTATATATATAAAATATATAATTATTTT
10	F	LSC	*trnS-trnR/trnG*	31	GCGGGTTCGATTCCCGCTACCCGCTCTAAAT
11	P	LSC	*trnS-trnR*	36	TAGCAATTGTGTAGTGAATTCACTACACAATTGCTA
12	P	LSC	*petN-psbM*	40	GCTAGTATGGTAGAAAGAGATCTCTTTCTACCATACTAGC
13	P	LSC	*psbC-trnS/trnS*	30	AAGGAGAGAGAGGGATTCGAACCCTCGATA
14	P	LSC	*trnS/ycf3-trnS*	30	GCCATCAACCACTCGGCCATCTCTCCAAAA
15	P	LSC	*rps14-psaB*	30	TTTTTATTATTTTAATAAAATGAATAAAAA
16	F	LSC	*PsaB*	43	CTATACATATGACCCGCAATGAGGAAAAGAATTGCGATAGCTA
17	F	LSC	*PsaB/PsaA*	46	AGGAAAAGAATTGCGATAGCTAGATGATGATGTGCCATATCG GTTA
18	P	LSC	*ycf3*	30	TGAGATTTTCATCTCATACGGCTCCTCCTT
19	P	LSC	*ycf3-trnS*	31	CTTTTCTTTTGTGAGAAAATTTTCTCACAAA
20	P	LSC	*ndhC-trnV*	31	TATTAATAATATAATATTAATATTATTAATA
21	P	LSC	*trnM-atpE*	39	AACTTATTAGACACCATGATCAATGGTGTCTAATAAGTT
22	P	LSC	*petA-psbJ*	30	ATTTTTCAATACAAATTTGTATTGAAAAAT
23	F	LSC	*petA-psbJ*	47	AATTGAAATTGATAGAATGTATCAATAATCAAGAGTTTTTTTCTA AT
24	P	LSC	*PsbE-petL/ccsA-ndhD*	31	TGAAGTTATATAATAGAGTTATTTTTTTTAT
25	P	LSC	*petL-petG*	30	ATGAATCTTTTTTGATCAAAAAAGATTTAT
26	P	LSC	*psaJ-rpI33*	30	CCCCCCTTTTTTTTTCTAATTCTTTTTTTA
27	P	LSC	*petB-petD*	45	TTATGTTTTTAGCTATTTTTTACTAAAAAATAGCTAAAAACATAA
28	C	LSC/IR	*rpI16-rps3/trnV-rps12*	30	CCTTATTTTATTTTTTTCATTGTTTTTTTC
29	P	SSC	*rpI32-trnL/ndhA*	33	AAAATAAAAAAAAAAAAGATAATATATATATATA
30	P	SSC	*ndhG/ndhG-ndhI*	30	GGCAAATCCATTATATTATTAAAAAAAGAA
31	P	SSC/IR	*ndhA/trnV-rps12*	37	AACCGTACATGAGGTTTTCGCCTCATACGGCTCCTCG
32	P	SSC	*rps15-trnN*	30	TAATTTTATAAAAAAAAAAGTTTAAATTTC
33	F	SSC	*rps15-trnN*	30	CTGTAGAATGAATAGATTTGTAGCAAACTG
34	P	SSC	*rps15-trnN*	30	CAAAAAAGATTATATATAGAATCTTTTTTG
35	F	IR	*rrn5-rrn4.5*	34	TGGTTTTTTCATGTTGTCAAAGAGTTGAACAATG
36	F	IR	*ycf2*	32	TTAGACAAAAAGAGAAGTAACTTGGACAAAAA

Table 4. Simple sequence repeats in *Brassica oleracea*.

Repeat	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Total
A/T	-	-	-	-	147	66	44	20	11	3	3	2	1	1	1	299
C/G	-	-	-	-	13	6	1		1							21
AC/GT	-	2														2
AG/CT	-	9														9
AT/AT	-	28	8	1	2											39
AAC/GTT	4															4
AAG/CTT	10															10
AAT/ATT	14	3														17
ACT/AGT	1															1
AGC/CTG	4															4
ATC/ATG	5															5
AAAC/GTTT	1															1
AAAG/CTTT	1															1
AAAT/ATTT	1															1
AGAT/ATCT	1															1
	Total															**415**

including genus *Brassica* in which the conserved chloroplast genome has been reported with high similarity in gene content, gene organisation, and intron content specifically at genus level (Xu et al., 2012; Yang et al., 2013; Cho et al., 2015; Li et al., 2017; Nguyen et al., 2017; Kim et al., 2018; Li et al., 2018; Shahzadi et al., 2019).

Location of boundaries is one of the most important factors to analyse the evolutionary patterns of chloroplast genome (Shahzadi et al., 2019; Abdullah et al., 2020; Henriquez et al., 2020a, 2020b). Here, we performed a detailed comparison of the boundary regions among chloroplast genomes of six species of genus *Brassica*, including the two sequenced in the present study. In our analyses, contraction and expansion of IR regions led to the origination of the pseudogene (*ycf1* $^{\Psi}$). The origination of pseudogene is also reported in previous studies (Shahzadi et al., 2019; Abdullah et al., 2020; Amiryousefi et al., 2020). Moreover, in the far divergent species such as at family level comparison the duplication of certain genes or deletion of a copy of gene has also been reported in chloroplast genome (Menezes et al., 2018; Abdullah et al., 2020). The phenomenon of IR contraction and expansion is also related to the phylogenetic studies in closely related species as suggested by many previous studies (Shahzadi et al., 2019; Shahzadi et al., 2020; Liu et al., 2018). Here, in the current study, the analyses of the *Brassica* chloroplast genome sequences revealed high resemblance in these species. However, a recent study does not agree with this finding (Henriquez et al., 2020c).

Codon usage analysis is important to study population pressure as well as phylogenetic analysis (Yang et al., 2014). Most protein coding genes employ ATG as the start codons. However, ATC, ATA, ATT, TTC can also be used as alternative codons (Henriquez et al., 2020b). Here, in our current study, we also found some other codons other than ATG codon as initiation codons. The high similarities in codon usage reveal that these species passed similar evolutionary conditions during the course of evolution (Menezes et al., 2018; Abdullah et al., 2020). We also found high similarities in amino acid frequencies among all studies species. We found Lysine as the most abundant amino acid whereas Cysteine as the rare encoded amino acid. This finding is similar to the previous reports (Iram et al., 2019; Abdullah et al., 2020; Henriquez et al., 2020a Waseem et al., 2020; Mehmood et al., 2020).

The relative synonymous codon usage (RSCU) value indicates the preference of a codon to code an amino acid. RSCU value greater than 1 indicates a higher frequency of codons whereas an RSCU value less than 1 indicates less prevalence of codons in the genome (Poczai and Hyvönen, 2017; Amiryousefi et al., 2018) . The codons having A or T at third nucleotide position showed higher RSCU values (>1) as compared to the codons having G or C at the third nucleotide position (<1). Such codon usage patterns have been observed in the chloroplast genomes of many other angiosperms (Henriquez et al., 2020c; Shahzadi et al., 2020).

Mononucleotide SSRs were more frequently present in the genome than di or tri nucleotide SSRs. Among

mononucleotide SSRs, A/T repeats were more abundant, as reported in previous studies of chloroplast genomes of angiosperms (Menezes et al., 2018; Mehmood et al., 2020). This might be due to the A/T rich composition of the chloroplast genome. Maximum repeats were found at the LSC region followed by the SSC region, as reported in some of the previous studies (Poczai and Hyvönen, 2017; Menezes et al., 2018).

Most of the repeats were found in intergenic spacer regions (IGS), followed by the coding region, whereas the least number of repeats were found in intronic regions. In several angiosperm lineages, numerous repeats have been identified in the IGS region (Poczai and Hyvönen, 2017; Abdullah et al., 2020; Mehmood et al., 2020). However, an abundance of repeats is also reported in the coding region of some angiosperms (Menezes et al., 2018). The moderate oligonucleotide range has been identified to induce origination of repeats and InDels (McDonald et al., 2011; Ahmed et al., 2012; Abdullah et al., 2020). Moreover, these repeats are also suggested as proxy for the identification of mutational hotspot regions (Ahmed et al., 2012; Abdullah et al., 2020). Hence, the repeats identified here could also be used for the identification of mutational hotspot regions.

5. Conclusion

In conclusion, our study provides insight into the evolutionary pattern of the chloroplast genome that exists in the two cultivars of broccoli, Marathon and Green sprout, in comparison to other species. These genomic resources will also be helpful for the development of vectors for chloroplast transformation of this important edible plants species.

Acknowledgements

This research work was funded by Higher Education Commission (HEC), Pakistan, under grant number: 7407/Federal/NRPU/R&D/HEC/2017. The authors acknowledge Mr. Irshad Khan for his technical support during the project.

References

Abdullah, Mehmood F, Shahzadi I, Waseem S, Mirza B et al. (2020). Chloroplast genome of Hibiscus rosa-sinensis (Malvaceae): Comparative analyses and identification of mutational hotspots. Genomics 112: 581–591.

Ahmad A, O Pereira E, J Conley A, S Richman A, Menassa R (2010). Green biofactories: recombinant protein production in plants. Recent Patents on Biotechnology 4 (3): 242-59.

Ahmed I, Biggs PJ, Matthews PJ, Collins LJ, Hendy MD et al. (2012). Mutational dynamics of aroid chloroplast genomes. Genome Biology and Evolution 4 (12): 1316-23.

Ahmed I, Matthews PJ, Biggs PJ, Naeem M, McLenachan PA et al. (2013). Identification of chloroplast genome loci suitable for high-resolution phylogeographic studies of C olocasia esculenta (L.) S chott (A raceae) and closely related taxa. Molecular Ecology Resources 13 (5): 929-37.

Ahmed I (2014). Evolutionary dynamics in taro [Internet]. Massey University, Palmerston North, New Zealand; Available from: https://mro.massey.ac.nz/handle/10179/5610.

Aires A (2015). Brassica composition and food processing. In Processing and impact on active components in food. Academic Press. 17-25.

Al-Shehbaz IA (2011). Brassicaceae (Mustard Family). eLS [Internet]. Chichester, UK: John Wiley & Sons, Ltd; Available from: http://doi.wiley.com/10.1002/9780470015902.a0003690.pub2.

Amiryousefi A, Hyvönen J, Poczai P (2018). IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics 34 (17): 3030-1.

Amiryousefi A, Hyvönen J, Poczai P (2018). The chloroplast genome sequence of bittersweet (Solanum dulcamara): Plastid genome structure evolution in Solanaceae. PLoS One 13: 1–23.

Andrews S (2018). Babraham Bioinformatics-FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Anjum NA, Gill SS, Ahmad I, Pacheco M, Duarte AC et al. (2012). The Plant Family Brassicaceae: An Introduction. In *The plant family Brassicaceae* (pp. 1-33). Springer, Dordrecht.

Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017). MISA-web: a web server for microsatellite prediction. Bioinformatics 33 (16): 2583-5.

Cho KS, Yun BK, Yoon YH, Hong SY, Mekapogu M et al. (2015). Complete chloroplast genome sequence of tartary buckwheat (*Fagopyrum tataricum*) and comparative analysis with common buckwheat (*F. esculentum*). PLoS One 10: 1–14.

Christenhusz MJM, Byng JW (2016). The number of known plants species in the world and its annual increase. Phytotaxa 261 (3): 201–217.

Daniell H, Lin CS, Yu M, Chang WJ (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biology 17 (1): 1-29.

Greiner S, Lehwark P, Bock R (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Research 47 (W1): W59-64.

Henriquez CL, Ahmed I, Carlsen MM, Zuluaga A, Croat TB et al. (2020). Molecular evolution of chloroplast genomes in Monsteroideae (Araceae). Planta 251 (3): 1-6.

Henriquez CL, Ahmed I, Carlsen MM, Zuluaga A, Croat TB et al. (2020). Evolutionary dynamics of chloroplast genomes in subfamily Aroideae (Araceae). Genomics 12 (3): 2349-2360.

Henriquez CL, Mehmood F, Shahzadi I, Ali Z, Waheed MT et al. (2020). Comparison among the first representative chloroplast genomes of Orontium, Lisa, Zamioculcas, and Stylochaeton of the plant family Araceae: inverted repeat dynamics are not linked to phylogenetic signaling. bioRxiv

Iram S, Hayat MQ, Tahir M, Gul A, Ahmed I (2019). Chloroplast genome sequence of Artemisia scoparia: comparative analyses and screening of mutational hotspots. Plants 8 (11): 476.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28 (12): 1647-9.

Khan MO, Mehmood MA, Mukhtar Z, Ahmad N (2018). Chloroplasts as cellular factories for the cost-effective production of cellulases. Protein and Peptide Letters 25 (2): 129-35.

Kim CK, Seol YJ, Perumal S, Lee j, Espinosa N et al. (2018). Re-exploration of U's Triangle Brassica Species Based on Chloroplast Genomes and 45S nrDNA Sequences. Scientific Reports 8: 1–11.

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J et al. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Research 29 (22): 4633-42.

Laslett D, Canback B (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Research 32 (1): 11-6.

Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li P, Zhang S, Li F, Zhang S, Zhang H et al. (2017). A phylogenetic analysis of chloroplast genomes elucidates the relationships of the six economically important Brassica species comprising the triangle of U. Frontiers in Plant Science 8: 111.

Li Y, Zhang Z, Yang J, Lv G (2018). Complete chloroplast genome of seven Fritillaria species, variable DNA markers identification and phylogenetic relationships within the genus. PLoS One 13 (3): e0194613.

Liu L, Wang Y, He P, Li P, Lee J et al. (2018). Chloroplast genome analyses and genomic resource development for epilithic sister genera Oresitrophe and Mukdenia (Saxifragaceae), using genome skimming data. BMC Genomics 19 (1): 1-7.

Loessl AG, Waheed MT (2011). Chloroplast-derived vaccines against human diseases: achievements, challenges and scopes. Plant Biotechnology Journal 9 (7): 527-539.

Lowe TM, Chan PP (2016). tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Research 44 (W1): W54-7.

McDonald MJ, Wang WC, Huang HD, Leu JY (2011). Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. PLoS Biology 9 (6): e1000622.

Menezes APA, Resende-Moreira LC, Buzatti RSO, Nazareno AG, Carlsen M et al. (2018). Chloroplast genomes of *Byrsonima* species (Malpighiaceae): Comparative analysis and screening of high divergence sequences. Scientific Reports 8: 1–12. doi: 10.1038/s41598-018-20189-4.

Mehmood F, Shahzadi I, Ali Z, Islam M, Naeem M et al. (2018). Correlations among oligonucleotide repeats, nucleotide substitutions, and insertion–deletion mutations in chloroplast genomes of plant family Malvaceae. Journal of Systematics and Evolution 59 (2): 388-402.

Mehmood F, Ubaid Z, Shahzadi I, Ahmed I, Waheed MT et al. (2020). Plastid genomics of *Nicotiana* (Solanaceae): insights into molecular evolution, positive selection and the origin of the maternal genome of Aztec tobacco (*Nicotiana rustica*). PeerJ 8: e9552.

Mehmood F, Abdullah, Shahzadi I, Ahmed I, Waheed MT et al. (2020) Characterization of *Withania somnifera* chloroplast genome and its comparison with other selected species of Solanaceae. Genomics 112: 1522–1530. doi: 10.1016/j.ygeno.2019.08.024.

Milne I, Bayer M, Cardle L, Shaw P, Stephen G et al. (2010) —next generation sequence assembly visualization. Bioinformatics 26 (3): 401-2.

Nguyen VB, Park HS, Lee SC, Lee J, Park JY et al. (2017). Authentication markers for five major Panax species developed via comparative analysis of complete chloroplast genome sequences. Journal of Agricultural and Food Chemistry 65 (30): 6298-306.

Owis AI (2015). Broccoli; the green beauty: a review. Journal of Pharmaceutical Sciences and Research 7 (9): 696.

Palmer JD (1985). Comparative organization of chloroplast genomes. Annual Review of Genetics 19 (1): 325-354.

Poczai P, Hyvönen J (2017). The complete chloroplast genome sequence of the CAM epiphyte Spanish moss (*Tillandsia usneoides*, Bromeliaceae) and its comparative analysis. PLoS One 12: 1–25.

Shahzadi I, Mehmood F, Ali Z, Malik MS, Waseem S et al. (2019) Comparative analyses of chloroplast genomes among three Firmiana species: Identification of mutational hotspots and phylogenetic relationship with other species of Malvaceae. Plant Gene 19: 100199.

Shahzadi I, Mehmood F, Ali Z, Ahmed I, Mirza B (2020). Chloroplast genome sequences of Artemisia maritima and Artemisia absinthium: Comparative analyses, mutational hotspots in genus Artemisia and phylogeny in family Asteraceae. Genomics 12 (2): 1454-63.

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A et al. (2017). GeSeq–versatile and accurate annotation of organelle genomes. Nucleic Acids Research 45 (W1): W6-11.

Waheed MT, Thönes N, Müller M, Hassan SW, Razavi NM et al. (2011). Transplastomic expression of a modified human papillomavirus L1 protein leading to the assembly of capsomeres in tobacco: a step towards cost-effective second-generation vaccines. Transgenic Research 20 (2): 271-82.

Waheed MT, Thönes N, Müller M, Hassan SW, Gottschamel J et al. (2011). Plastid expression of a double-pentameric vaccine candidate containing human papillomavirus-16 L1 antigen fused with LTB as adjuvant: transplastomic plants show pleiotropic phenotypes. Plant Biotechnology Journal 9 (6): 651-60.

Waseem S, Mirza B, Ahmed I, Waheed MT (2020). Comparative analyses of chloroplast genomes of *Theobroma cacao* and *Theobroma grandiflorum*. Biologia 75 (5): 761-71.

Xu Q, Xiong G, Li P, He F, Huang Y et al. (2012). Analysis of complete nucleotide sequences of 12 Gossypium chloroplast genomes: origin and evolution of allotetraploids. Plos One, e37128.

Xu JH, Liu Q, Hu W, Wang T, Xue Q et al. (2015). Dynamics of chloroplast genomes in green plants. Genomics 106 (4): 221-31.

Yang JB, Tang M, Li HT, Zhang ZR, Li DZ (2013). Complete chloroplast genome of the genus Cymbidium: lights into the species identification, phylogenetic implications and population genetic analyses. BMC Evolutionary Biology 13 (1): 1-2.

Yang X, Luo X, Cai X (2014). Analysis of codon usage pattern in *Taenia saginata* based on a transcriptome dataset. Parasites & Vectors 7 (1): 1-1.

Zerbino DR, Birney E (2018). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18: 821–829.

Zhang X, Lin N, Chen L, Zhang Z, Lei H et al. (2020). Genetic Diversity and Genetic Differentiation of Rheum palmatum by Chloroplast matK Sequences. The Natural Products Journal 10 (2): 96- 103.