Primer and barcode gap identification and DNA barcode generation for species discrimination in plants

Chapter · January 2015

CITATIONS
0

READS
78

1 author:

Deepu Mathew
Kerala Agricultural University
61 PUBLICATIONS **42** CITATIONS

SEE PROFILE

PRIMER AND BARCODE GAP IDENTIFICATION AND DNA BARCODE GENERATION FOR SPECIES DISCRIMINATION IN PLANTS

Dr.Deepu Mathew
Assistant Professor, Centre for Plant Biotechnology and Molecular Biology,
Kerala Agricultural University, KAU Post, Thrissur – 680656, India, E-mail: deepu.mathew@kau.in,
Phone: +914872438585 (office), +919446478503 (mobile)

I. INTRODUCTION

DNA barcoding isa methodology that offers assistance to the conventional taxonomy at species level, through analyzing the variation or barcode gaps those exist in the nucleotide sequences in various species of any genus in the specifically identified gene loci.This chapter details the various steps involved in the DNA barcoding of plants, with special reference to the genus *Momordica*. Different candidate loci for DNA barcoding in plants and animals, factors involved in the selection of candidate locus/ loci in any genus, design of genus specific primers through the use of universal primers and list of suggested universal primers, design of primers directly by mining the data from NCBI are detailed. Also, the use of the software Clustal Omega for multiple sequence alignment and to derive a base sequence of the candidate locus, for primer design is detailed. Barcode gap identification and species delineation, annotation of sequence using BLASTn, BioEdit and MEGA and finally, the steps involved in the submission of barcode data to BOLD systems, for the generation of DNA barcodes are also presented in this chapter.

Keywords: barcoding, BOLD, chloroplast DNA, evolution, *matK*, mitochondrial DNA, systematics; taxonomy

DNA barcoding is a methodology that offers assistance to the conventional taxonomy at species level. This technique uses the variations in the nucleotide sequences that may occur during the evolutionary process, commonly in all the members of a particular species, at the candidate gene loci. Mostly, the genes or the genomic loci employed in this strategy are lesser influenced in the structural changes during evolution. Thus, mitochondrial genes such as cytochrome oxydase I (*CoI*), chloroplast genes *ycf1, trnK, rbcL, accD, matK, ndhJ, , rpoB, rpoC1, ycf5, rps16* or chloroplast spacers such as *ycf1-rps15, rpl32-trnL, petA-psbJ, rps16-trnQ, ndhC-trnV, rpoB-trnC, psbE-petL, rbcL-accD, psbK-psbI, atpF-atpH* and *trnH-psbA* or the loci for internal transcribed spacers such as *ITS1* and *ITS2* on nuclear genome are widely used in the DNA barcoding works. Among these loci, *CoI* is found most suitable for the members of the animal kingdom whereas, the suitability of each locus or combination of various loci have to be evaluated for each member in the plant kingdom. DNA barcoding is not a method to substitute the classical taxonomy but can supplement the taxonomic process substantially by providing proof for the species delineation.

How DNA barcoding works?

As already mentioned, the characteristic variations occurring in all the members of a particular species, for any candidate locus/loci will be considered as the DNA barcode. Accordingly, all the DNA barcodes for a species for each locus will be identified and these definite sequence variations will be employed in species discrimination.

To identify the barcodes, the total genomic DNA that include the nuclear, chloroplast and mitochondrial DNAs will be isolated using any standard protocol. Specific protocols for the isolation of chloroplast (Palmer, 1986; Dally and Second, 1989; Mourad, 1998) and mitochondrial (White *et*

al., 1992) DNAs are also available. However, total genomic DNA itself will yield sufficient quantity of good quality material for the PCR amplification of the target sequence.

II. PREREQUISITE FOR DNA BARCODING

Apart from the routine prerequisites for a PCR amplification such as standardized thermal profile and chemicals, most vital components required are the identification of the most suitable locus/ combination of loci for the genus and identification of the specific forward and reverse primer sets for the loci.

a. Identification of the locus/ loci

For most of the members of the animal kingdom, *CoI* is identified as the most suitable locus. In all the genera, this locus is sufficiently long (~2.5kb) to identify the maximum number of barcodes. In usual barcoding programmes, the amplified locus will be eluted from the gel and sequenced directly using the NGS platforms. The inherent disadvantage of the NGS platforms is the shorter read length (less than 1 kb). This necessitates that the sequencing of the band has to be done in both forward and reverse directions so that the maximum sequence length could be assembled. Since it is more probable to get more barcodes with higher sequence lengths, it is always advised to choose the locus with the highest size.

For every member of the plant kingdom, the suitability of the particular locus has to be tested. Among the common loci, *ycf1-rps15* has the longest workable sequence length (~5.0 kb) followed by *rpoB* (~3.5 kb), *trnK* (~2.5 kb), matK (~1.5kb), *rbcL* (~1.5 kb), *rpoB-trnC*(1.3 kb), *psbE-petL*(1.3 kb), *rbcL-accD*(1.1 kb), *ndhJ* (0.9 kb), *rpl32-trnL* (0.85 kb), *rps16-trnQ* (0.8 kb), *petA-psbJ*(0.75 kb), *ndhC-trnV*(0.65 kb), *trnH-psbA* (0.55 kb), *psbK-psbI*(0.5 kb), *accD* (~0.5 kb), *atpF-atpH* (0.45 kb)etc. After sequencing a particular locus, the sequences of all the members from the different species of a genus has to be aligned to identify the base sequence for the genus. The sequences of the members of a particular species has to be studied in relation to the base sequence to identify the nucleotide variations common for all the member of the species from the base sequence. Such kind of a nucleotide variation is designated as the barcode gap.

If a particular locus fails to bring out any substantial or reliable uniform variations among all the members of any one of the species in the genus understudy, the locus is not suitable to be used for DNA barcoding. Thus the next candidate locus has to be checked for its capability to reveal the reliable barcode gaps for each species. Many a times, combination of various loci are found to be better in species delineation.

For barcoding the plants, universal barcode loci are recommended by many researchers and such recommendations are summarized in Table 1.

b. Identification of suitable primers

There are two strategies to identify the primer.

b.i. Use of universal primers

The universal primers for different loci are also described by many researchers. A consolidation on the recommended universal primers is given by Mathew (2013) and Girme (2014). It is unlikely that that these primers may amplify the locus in all the samples under study but most probably these primers may amplify the locus in at least few samples. The bands amplified using the universal

primers may be eluted from the gel, directly sequenced on NGS platforms and the sequences have to be aligned. From these sequences, the base sequence for the genus has to be worked out and using the base sequence, barcoding primers for the locus for the particular genus could be designed using any standard software such as Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/).

The identification of barcoding primers for the genus *Momordica* is detailed hereunder.

Primer sets for 3 common barcoding loci *matK* (Maturase K), *ITS2* (Internal transcribed spacer) and *trnH-psbA* intergenic spacer were used for the initial screening to assess their compatibility to generate the banding patterns in various*Momordica* accession. PCR was performed using the mix of genomic DNA (30 ng), 10X Taq assay buffer A - 2 µL, dNTP mix (10 mM each) - 1.5 µL, Taq DNA polymerase (3U) - 0.3 µL, primer (10 pM) - 0.75 µL each of F and W and autoclaved distilled water - 13.7 µL. The PCR protocol issued by CBOL (Ivanova *et al.*, 2006) was employed which constituted 40 cycles of initial denaturation at 94°C for 1 minute, denaturation at 94°C for 30 seconds, primer annealing (temperature as per GC composition of primers) for 40 seconds and primer extension at 72°C for 40 seconds. This was followed by the final extension at 72°C for 5 minutes and incubation at 4°C.

Table 1 Universal barcode loci recommended for plants

Sl. No.	Barcode loci recommended	Reference
1	*matK*	Lahaye *et al.* (2008)
2	*matK*, *trnH-psbA*	Newmaster *et al.* (2008), Seberg and Petersen (2009)
3	*ITS*	Chase *et al.* (2005), Kress *et al.* (2005)
4	*trnH-psbA*	Kress *et al.* (2005)
5	*trnL*	Taberlet *et al.* (2007)
6	*atpF-atpH*, *matK*, and *psbK-psbI*	Pennisi (2007)
7	*rpoC1*	Hollingsworth *et al.* (2009)
8	*rpoc1+rpoB+matK* and *rpoc1+matK+trnH-psbA*	Chase *et al.* (2007)
9	For bryophytes - *rbcL* and *rpoC1*	Liu *et al.* (2010)
10	trnH-psbA + rbcL	Kress and Erickson (2007)

Eight primer sets, of which 6 for *matK* and 1 for *ITS2* and 1 for *trnH-psbA* were initially used for amplifying the barcode loci in *Momordica* (Table 2).

Table 2: List and sequence of universal DNA barcoding primers used for barcoding the g *Momordica*

Sl No.	Primer	Nucleotide Sequences	Reference
1	*matK* F1	CCTATCCATCTGGAAATCTTAG	Jarret, 2008 ; Dunning and Savolainen, 2010
	matK R1	GTTCTAGCACAAGAAAGTCG	
2	*matK* F2	ATCCATCTGGAAATCTTAGTTC	
	matK R2	CTTCCTCTGTAAAGAATTC	
3	*matK* F3	CGATCTATTCATTCAATATTTC	
	matK R3	TCTAGCACACGAAAGTCGAAGT	Lahaye *et al*., 2007
4	*matK* F4	(T)AATTTACGATCAATTCATTC	Jarret, 2008 ; Dunning and Savolainen, 2010
	matK R4	TCTAGCACACGAGTCGAAGT	Lahaye *et al*., 2007
5	*matK* F5	CGATCTATTCATTCAATATTTC	Jing *et al*., 2011
	matK R5	GTTCTAGCACAAGAAAGTCG	
6	*matK* F6	ACCCAGTCCATCTGGAAATCTTGGT TC	Jarret, 2008 Dunning and Savolainen, 2010
	matK R6	CGTACAGTACTTTTGTGTTTACGAG	
7	*ITS2* F	GCGATACTTGGTGTGAAT	Jing *et al*., 2011
	ITS2 R	GACGCTTCTCCAGACTACAAT	
8	*trnH-psbA* F	CGCGCATGGTGGATTCACAATCC	Jarret, 2008
	trnH-psbA R	GTTATGCATGAACGTAATGCT C	

Based on the initial studies it was found that *matK* is giving sufficiently good amplification respo and the size of this locus remarkably larger than the other two loci. Hence it was decided to cho *matK* as the candidate locus for this study. Subsequently, 18 primer combinations using the forw and reverse primers for *matK* loci were attempted to amplify the loci, from 24 *Momordica* access (Table 3).

Table 3: List of combination of forward and reverse primer used for amplifying *matK* locus of *Momordica*

Sl. No.	Primer	Annealing Temp. (°C)	Combination
1	*matK* 1	51.7	*matK* F1
			matK R1
2	*matK* 2	45.2	*matK* F2
			matK R2
3	*matK* 3	49.5	*matK* F3
			matK R3
4	*matK* 4	45.2	*matK* F1
			matK R2
5	*matK* 5	53.2	*matK* F1
			matK R3
6	*matK* 6	51.7	*matK* F2
			matK R1
7	*matK* 7	52.7	*matK* F2
			matK R3
8	*matK* 8	49.5	*matK* F3
			matK R1
9	*matK* 9	45.2	*matK* F3
			matK R2
10	*matK* 10	48.4	*matK* F4
			matK R4
11	*matK* 11	49.5	*matK* F5
			matK R5
12	*matK* 12	55.6	*matK* F6
			matK R6
13	*matK* 13	48.4	*matK* F4
			matK R5
14	*matK* 14	48.4	*matK* F4
			matK R6
15	*matK* 15	49.5	*matK* F5
			matK R4
16	*matK* 16	49.5	*matK* F5
			matK R6
17	*matK* 17	57.9	*matK* 6
			matK R4
18	*matK* 18	51.7	*matK* F6
			matK R5

Various primer combinations have amplified various accessions and by using all the 18 combinations of the *matK* primers, the locus was successfully amplified in all the accessions under study.

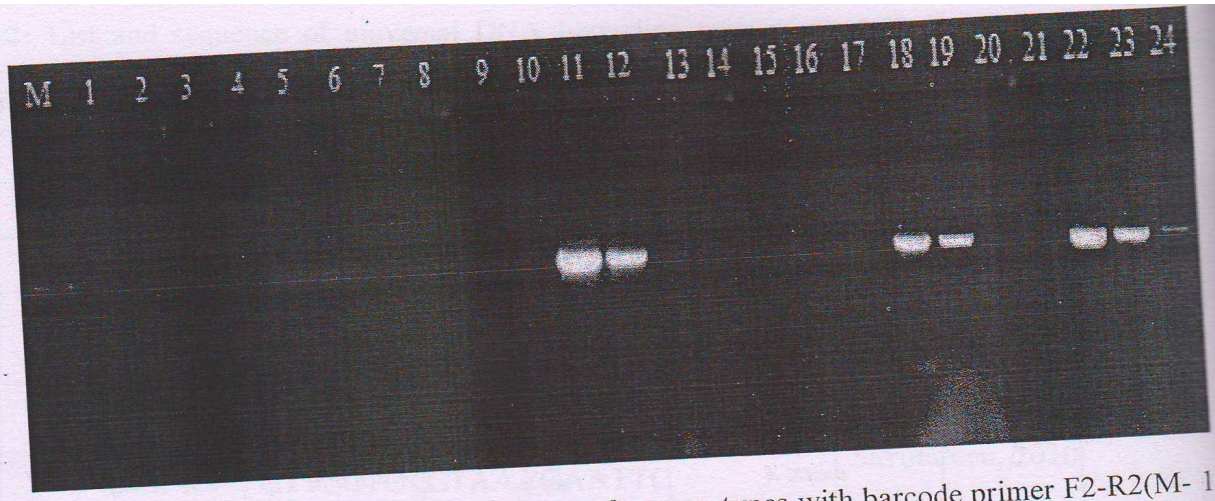
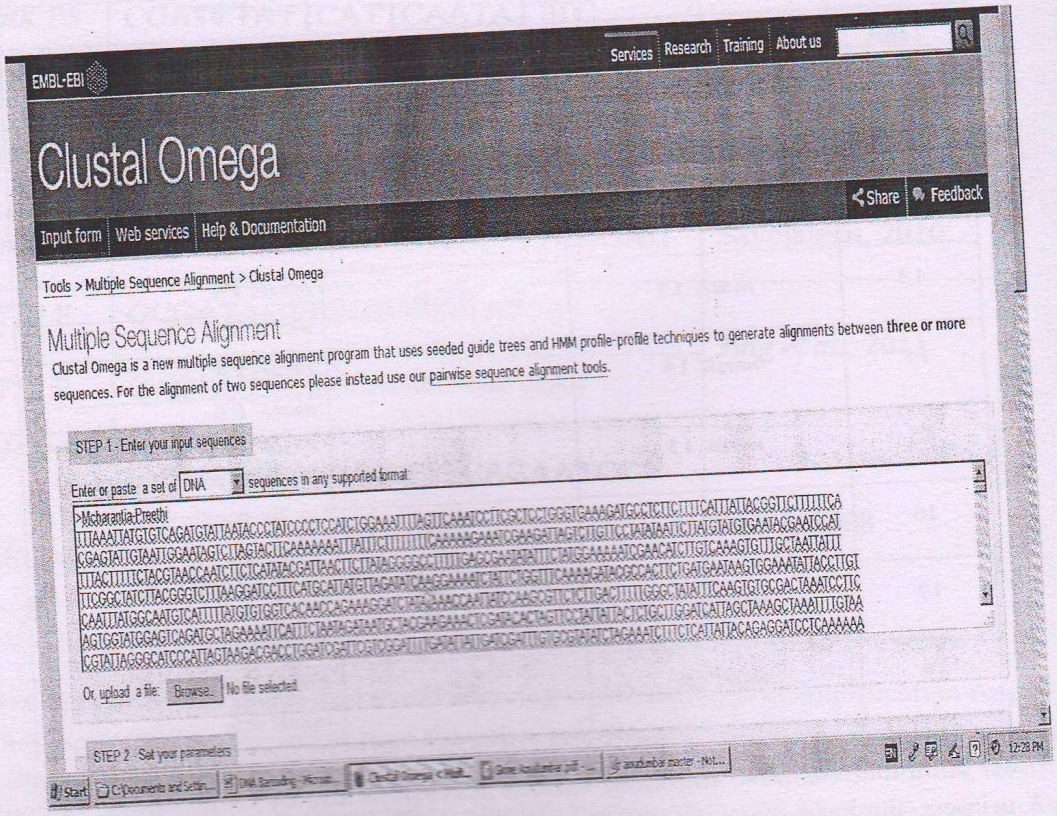


Fig. 1: DNA amplification pattern of the *Momordica* genotypes with barcode primer F2-R2(M- 1000 bp Ladder, 1 - Preethi , 2 - Kurupantara , 3 - JNM 7, 4 - Vadakara, 5 - V 53, 6 - Arka Gaurav, 7 - *M. charantia* cv. *muricata* , 8 - *M. dioica* Kerala 1, 9 - *M.dioica* Kerala 2, 10 - *M. dioica* Kerala 3, 11 - *M. dioica* Kerala 4, 12 - *M. dioica* Odisa,13 - Wild 2 (*M. charantia* cv. *muricata*),14 - Wild 1 *(M. sahyadrica)*, 15 - Wild 2 *(M. sahyadrica)*, 16 - *M. sahyadrica* Annamalai type, 17 - *M. balsamina*, 18 - *M. cochinchinensis* Andaman, 19 - *M. cochinchinensis* Northeast, 20 - *M.subangulata* 1, 21 - *M. subangulata* 2, 22 - *M. subangulata* 3, 23 - *M. subangulata* 4, 24 - *M. subangulata* 5).

The bands were eluted, sequenced and the sequences were analysed using BLASTn to confirm that it is the sequence of the locus of interest. The sequences were aligned using the software Clustal Omega (Fig. 2a-c).



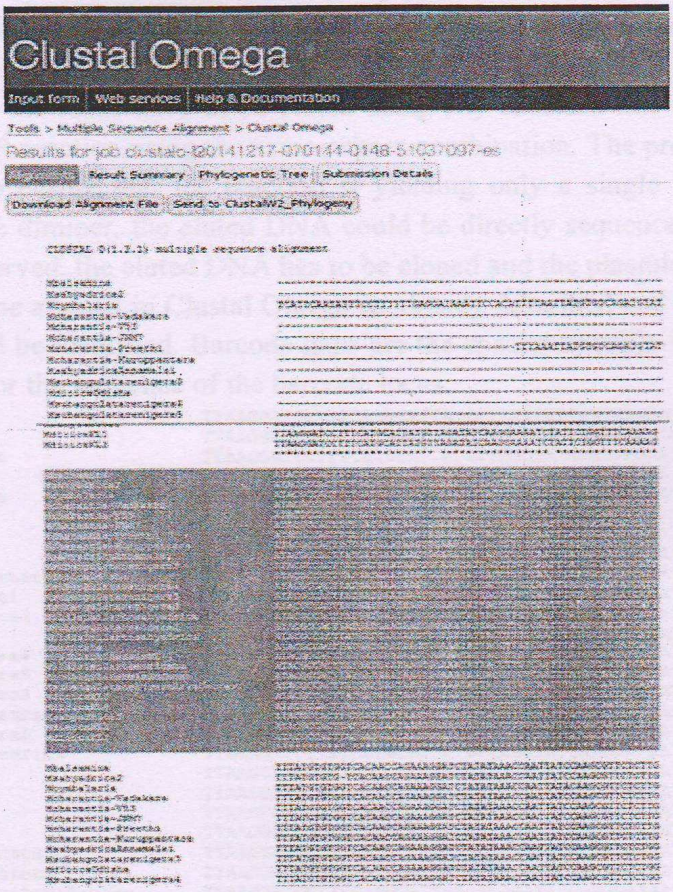


Fig. 2a-c: Steps in sequence alignment of the matK sequences for 25 Momordica accessions using the software Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/).

From the alignment, the base sequence of *matK* locus for the genus *Momordica* was arrived. This was subsequently used for designing primers for barcoding *Momordica* using the *matK* locus.

Forward - 5'AGGGTTTGGAGTCATTGTGG3' (Tm - 59.82°C)
Reverse – 5'GAATCGATCCAGGTCGTCTT3' (Tm - 59.09°C) Product size: 897

b.iii. Designing the primers using the sequences available in NCBI, for the loci for the same or related genus.

The sequences of the specific barcoding locus for the same or the closely related genus may be recovered from NCBI and they have to be aligned in the same way to design the primers for the genus.

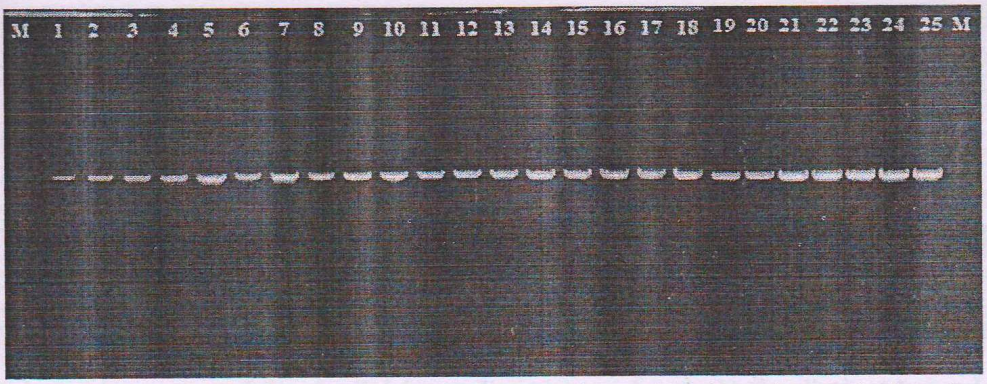


Fig. 3: Uniform amplification of *matK* locus in *Momordica* using the primers designed on the base sequences obtained by sequencing the bands generated by universal primers

III. Sequencing the amplified locus and identification of barcode gaps in each species

The bands representing the locus have to be eluted using any standard kits and the DNA has to be further amplified in PCR reaction using the same primer combination. The product has to be run on 2 per cent agarose gel to ensure that the reaction is yielding only a single band. If single band is amplified with no prime dimmer, the eluted DNA could be directly sequenced on NGS platforms. If multiple banding is observed, the eluted DNA has to be cloned and the plasmid has to be sequenced. The sequences have to be aligned in Clustal Omega as already described and based on the alignment, the barcode gaps should be identified. Barcode gaps are the species specific variations from the base sequence of the genus for the sequence of the barcode locus.

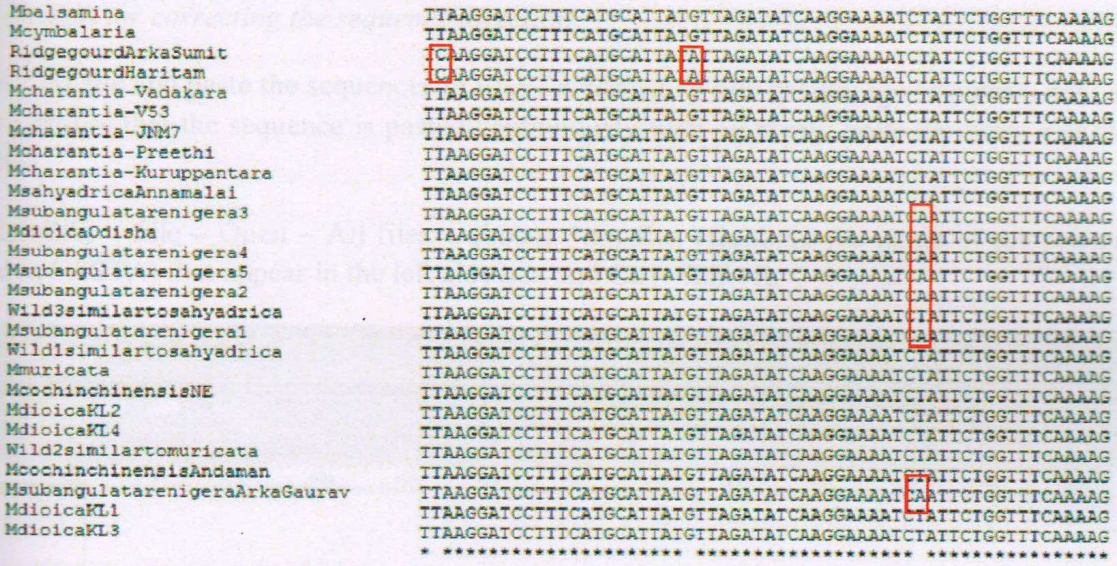


Fig. 4: The *matK* sequences of *Momordica* and *Luffa* accession aligned using ClustalW software to isolate the barcode gaps. Absence of * shows the presence of gaps.

IV. Manual generation of barcodes for various species in the genus.

The summation of the barcode gaps leads to the DNA barcodes for each species.

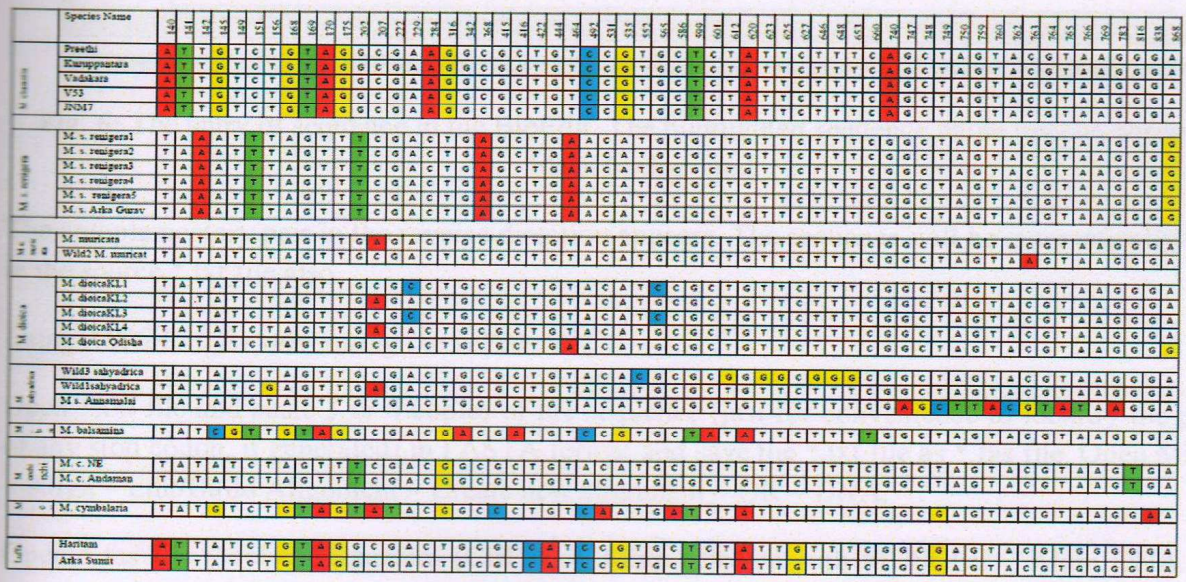


Fig. 5: DNA barcode generated for the 7 *Momordica* spp. under study the distinct barcode gap of *M. c.* ssp *charantia*, *M. subangulata*, *M. balsamina*, *M.cymbalaria* and *Luffa* may be noted.

V. Sequence annotation

Before being submitted to BOLD systems (The Barcode of Life Data Systems v3 http://www.boldsystems.org/), the sequences have to be annotated. Sequence annotation is a process through which the coding and non-coding regions and other specific locations that are of importance in a DNA sequence are identified.

First, using BLASTn, confirm that the sequence that we have obtained is in the 5'-3' forward direction. If the sequence is in reverse direction, the reverse complementary sequence has to be generated using the BioEdit software for submission to NCBI as well as BOLD.

Using BioEdit for correcting the sequence direction

Open a *.txt file and paste the sequences in FASTA format. While pasting the sequence, care must be taken to ensure that the sequence is pasted continuously with no break. Save the entry and close the *.txt file.

Open BioEdit – File – Open – All files – choose *.txt file. Highlight the name of the sequence (as entered in FASTA) that appear in the left side box and Ctrl+Shift+R.

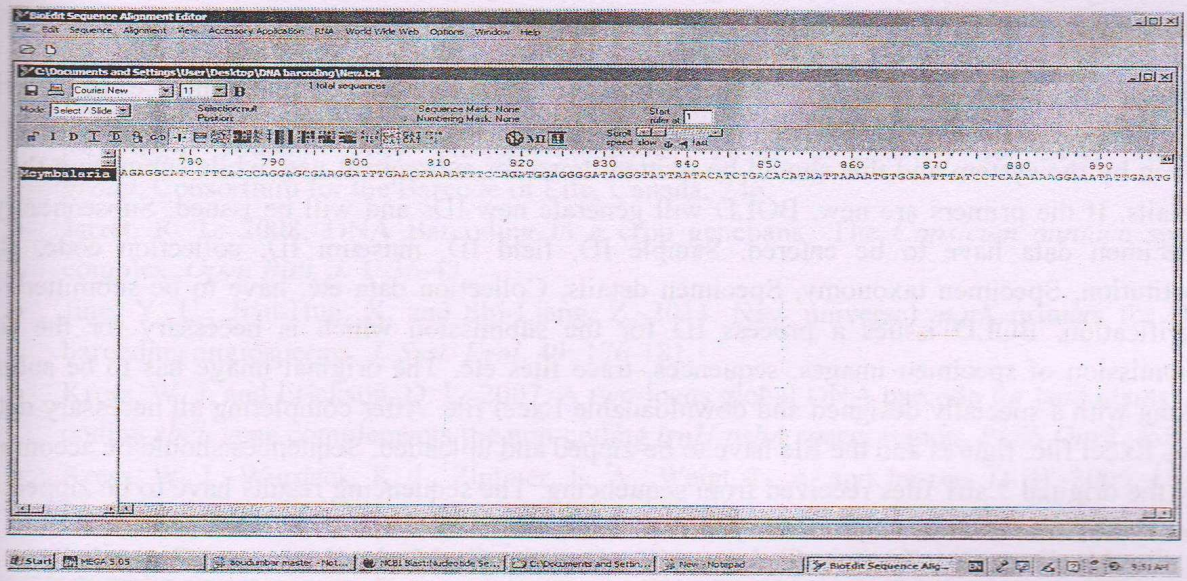


Fig. 6: The entry of sequence in the BioEdit. The highlighted sequence name may be observed in black background.

The complementary pair in the reverse direction appears. The sequence will be automatically changed in the source *.txt file also.

Using MEGA5.0 for annotating the sequence

Open *.txt file and paste 3 or 4 copies of the sequence (just for convenience to identify the presence of any stop codon, if generated) in FASTA format and save the *.txt file as *.fas file. Open MEGA5.0 – Align – Edit/Build Alignment – Create new alignment – OK – DNA.

Open *.fas file – select sequences (Ctrl A) – Align selected with MUSCLE – Align DNA. A new window opens and in that, Compute – Close – Sava as *.mas file.

Open MEGA5.0 and drag the *.mas file in to the work area. Then, Analyse – TA – Data. New box of code table appears and Select code table – preferably, plant plastid – OK. Click the Phen-UUC button to convert the nucleotide sequences to aminoacids sequences. This also shows the stop codons, if any, in between the amino acide sequences. There should be no stop codon in the sequences that have to be submitted and for this we have to check, if the start site in nucleotide sequence is changed to second or third position, whether the stop codon could be avoided. For this, select the icon 'Select and Edit Genes/Domains'. A new box appears, click 1 (present position choosen as the start point in the nucleotide sequence) and change to 2. If still the stop codon appears, try with 3 and still if the stop codon appears, we will have to trim the nucleotide sequence to avoid the stop codons in between the amino acid sequences.

VI. Submitting the details to BOLD

The DNA barcoding sequences and other details have to be submitted at BOLD systems for the generation of barcodes for every specimen. The details on submission are available in the BOLD handbook at http://www.boldsystems.org/libhtml_v3/static/BOLD_Handbook_Oct2013.pdf. Initially, an account may be started with a preferred username and password. After logging in, if you are new to the system, enter the workbench and start the new project. Project title, project code, project type, primary marker, supporting marker, involvement in any of the barcoding campaign, place in container, tags, project description etc. have to be entered here.

The primers that we have used in the barcoding have to be registered. If the sequences of the primers are already available in the BOLD systems, the unique ID of Forward and reverse primers will be given by the system and this should be written down for subsequent use, while submitting the other details. If the primers are new, BOLD will generate new IDs and will be issued. Subsequently, the specimen data have to be entered. Sample ID, field ID, museum ID, collection code, storing institution, Specimen taxonomy, Specimen details, Collection data etc. have to be submitted. After verification, BOLD issues a process ID for the submission which is necessary for the further submission of specimen images, sequences, trace files etc. The original image has to be submitted along with a specially designed and downloadable Excel file. After completing all necessary detail in this Excel file, figures and the file have to be zipped and uploaded. Sequences should be accompanied by the original *.ab1 files received from sequencing. The sequencing results have to be zipped along with *.ab1 files (forward and reverse) and uploaded.

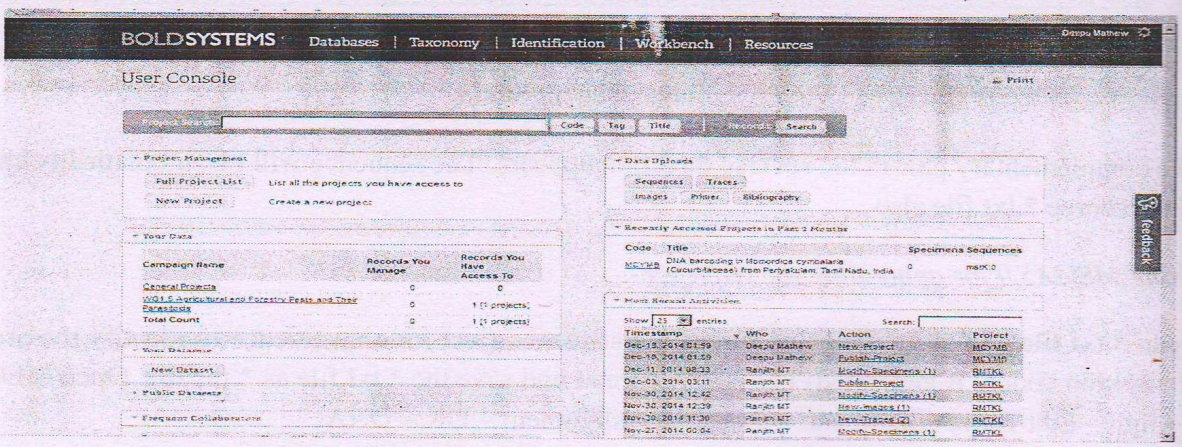


Fig. 7: The front page of the BOLD submission page, which shows the details to be submitted to generate the barcodes.

Once the submission of all the details are completed, BOLD will be generating the DNA barcode for the specimen.

VII. REFERENCES

- Chase, M. W., Cowan, R. S., Hollingsworth, P. M., van den Berg, C., Madrinan, S., Petersen, G., Seberg, O., Jorgsensen, T., Cameron, K. M., Carine, M., *et al.*, 2007. A proposal for a standardised protocol to barcode all land plants. *Taxon*, **56**: 295–299.
- Chase, M. W., Salamin, N., Wilkinson, M., *et al.*, 2005. Land plants and DNA barcodes: short-term and long-term goals. Philosophical Transactions of the Royal Society B: Biological Sciences, 360: 1889–1895.
- Dally, A. M. and Second, G. 1989. Chloroplast DNA isolation from higher plants: an improved non-aqueous method. *Plant Mol. Biol. Rep.***7**: 135-143.
- Dunning, L.T. and Savolainen, V. 2010. Broad-scale amplification of *matK* for DNA barcoding plants, a technical note. *Bot. J. Linnean Soc.* **164**: 1–9.
- Girme, A. R. (2014. DNA barcoding in *Momordica* spp. M. Sc. (Ag.) Plant Biotechnology thesis, Kerala Agricultural University, Thrissur, India, p. 83.
- Hollingsworth, M. L., Clark, A., Forrest, L. L., Richardson, J., Pennington, T. R., Long, D.G., Cowan, R., Chase, M. W., Gaudeul M. and Hollingsworth, P. M. 2009. Selecting barcoding loci for plants: evaluation of seven candidate loci with species level sampling in three divergent groups of land plants. *Mol. Ecol. Resour.***9**: 439–457.
- Ivanova, N. V., deWaard, J. R., Hajibabaei, M. and Hebert, P. D. N. 2006. *Protocols for high volume DNA barcode analysis, Draft submission to DNA working group*. Consortium for the Barcode of Life, Canada, 24p.
- Jarret, R. L. 2008. DNA Barcoding in a crop genebank: The *Capsicum annuum* species complex. *Open Biol. J.* **1**: 35-42.
- Jing, Y. U., Jian-Hua, X. and Shi-Liang, Z. 2011. New universal *matK* primers for DNA barcoding angiosperms. *J. Syst. Evol.* **49**: 176-181.
- Kress, W. J. and Erickson, D. L. 2007. A two-locus global DNA barcode for land plants: The coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PLoS One***2**: e508.
- Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A. and Janzen, D. H. 2005. Use of DNA barcodes to identify flowering plants. *Proc. Natl. Acad. Sci. USA***102**: 8369–8374.
- Lahaye, R., van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., Maurin, O., Duthoit, S., Barraclough, T. G. and Savolainen, V. 2008. DNA barcoding the floras of biodiversity hotspots. *Proc. Natl. Acad. Sci. USA***105**: 2923–2928.
- Liu, Y., Yan, H. F., Cao, T. and Ge, X. J. 2010. Evaluation of 10 plant barcodes in Bryophyta (Mosses). *J. Syst. Evol.***48**: 36–46.
- Mathew, D. 2013. DNA barcoding and its applications in horticultural crops. Biotechnology in Horticulture: Methods and Applications, Vol. 1 Peter, K.V. (Ed.), New India Publishing Agency (ISBN: 9789381450918), New Delhi, India, pp. 25-50;
- Mourad, G. S. 1998. Chloroplast DNA isolation. In: Martinez-Zapater, J. and Salinas, J. (Eds.), Methods in Molecular Biology, Vol. 82, *Arabidopsis Protocols*, Humana Press Inc., Totowa, NJ, pp. 71-77.
- Newmaster, S. G., Fazekas, A. J., Steeves, R. A. D. and Janovec, J. 2008. Testing candidate plant barcode regions in the Myristicaceae. *Mol. Ecol. Resour.***8**: 480–490.
- Palmer, J. D. 1986. Isolation and structural analysis of chloroplast DNA. *Method. Enzymol.***118**, 167-186.
- Pennisi, E. 2007. Taxonomy. Wanted: a barcode for plants. *Science***318**: 190-191.

- Seberg, O. and Petersen, G. 2009. How many loci does it take to DNA barcode a Crocus? *PLoS One***4**: e4598.
- Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermat, T., Corthier, G., Brochmann, C. and Willerslev, E. 2007. Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucl. Acids Res.***35**: e1–e8.
- White, P. S., Densmore III, L. D. and Hoelzel, A. R. 1992. Mitochondrial DNA isolation. In: Hoelzel, A.R. (Ed.), *Molecular genetic analysis of populations: A practical approach.*, IRL Press, Oxford, New York. 29-58.