American Journal of Molecular Biology, 2018, 8, 215-244
http://www.scirp.org/journal/ajmb
ISSN Online: 2161-6663
ISSN Print: 2161-6620

Molecular Footprint of Kenya's Gene Bank Repositories Based on the *cp*-Genome Signatures

Okoth Patrick[1,2], Muoma John[1], Omayio Dennis[1,2], Barasa Mustafa[3], Angienda Paul[2]

[1]Department of Biological Sciences, Masinde Muliro University of Science and Technology, Kakamega, Kenya
[2]Department of Zoology, School of Physical and Biological Sciences, Maseno University, Private Bag, Maseno, Kenya
[3]Department of Medical Laboratory Sciences (MLS), Masinde Muliro University of Science and Technology, Kakamega, Kenya
Email: okothpatrick@mmust.ac.ke

How to cite this paper: Patrick, O., John, M., Dennis, O., Mustafa, B. and Paul, A. (2018) Molecular Footprint of Kenya's Gene Bank Repositories Based on the *cp*-Genome Signatures. *American Journal of Molecular Biology*, **8**, 215-244.
https://doi.org/10.4236/ajmb.2018.84019

Received: June 14, 2018
Accepted: October 27, 2018
Published: October 30, 2018

Abstract

While the mutational processes that subsume biological diversity can be revealed in great detail through phylogenetic inferencing using plastid markers, few studies document their use. Accurate phylogenic inference can provide a framework for addressing a host of important evolutionary questions including a context to reconstruct molecular evolution of an organism. Despite the obvious utility of plastid markers in illuminating biological enquiry, many important questions still abound. The use of *cp*-DNA gene sequence data for phylogenetic inference can have an enormous impact on plant phylogenetics and systematics. The repertoire of genetic diversity of Kenya's Gene Bank repositories can be explored based on *cp*-genome signatures. This is because *cp*-DNA-based mutational changes are an important additional tool to the previous evidence available on plant evolution yet to be explored in biodiversity studies in Kenya. Taken together, these evolutionary changes can inspire development of realistic algorithms for phylogenetic inferencing based on molecular data. Phylogenetic reconstructions are at the very core of molecular evolution. Comparative sequence analyses of plastid markers can have utility beyond the study of phylogeny. The pattern of nucleotide substitution observed over evolutionary time can reflect functional constraints imposed due to natural selection. In line with this, it is possible to detect subtle anatomical variations associated with small fitness effects that can account for genetic diversity at varietal level. The lack of sequence information in Kenyan cowpea has limited the robust advancement of molecular markers use in dissecting diversity based on the putative plastid markers [1]. The present study sought to generate and upscale novel technologies such as genomics, DNA barcoding and bio-informatics in understanding molecular diversity of cowpea accessions from the Gene Bank of Kenya and ecotypes. A total of 298 sequences of cowpea germplasm conserved as *in situ* and *ex situ* in Kenya but sourced from phylogeographically diverse settings were examined and their genetic profiles were characterized and evaluated using molecular tools. The Gene Bank materials were purposefully sampled to develop subsets representative of the diversity in the genepool's collection. We present an extensive study on characterizing the genetic diversity of *cp*-DNA gene sequence data for the cowpea accessions from the Nation Gene Bank of Kenya. The comparative sequence analyses and phylogenetic clustering of seven plastid markers widely used in the DNA barcoding of land plants provide insights on the molecular evolution of this vascular plant. The detailed and in-depth genome characterization herein greatly enriches the genetic profile of this important crop, which can help in reconstructing realistic models of mutational process during plant evolutionary history. This study addressed this gap by employing a DNA barcode library for cowpea to determine the loci that yield the best species resolution. As well, this study examined the efficacy of custom DNA barcode loci for identification success, and compared phylogenetic diversity measures between sites and among variants.

Keywords

DNA Barcoding, Molecular Phylogeny, Molecular Diversity, Plastid Markers, Genome Signatures, Genotyping, NCBI

1. Introduction

Occurrence of *cp*-DNA microsatellites in the chloroplast genome has been widely utilized for delineating and reconstruction of phylogenetic relationships, taxonomic studies and the identification of maternal patterns in polyploids. The National Gene Bank of Kenya holds a repository of cowpea accessions from diverse phylogeographic backgrounds [1]. They carry a repertoire of genetic diversity, not adequately characterized. However, this immense potential needs to be unlocked using a suite of novel technologies, such as genomics. Vascular plants have a relatively slow rate of molecular evolution, and their frequent exposure to hybridization and introgression, often makes it difficult to discriminate them. Previous studies have examined these constraints in narrow geographic or taxonomic contexts, but the present investigation expanded analyses to consider the performance of seven plastid markers in molecular and phylogenetic clustering. With many genomic tools and resources becoming increasingly available, a more detailed and in-depth genome characterization of cowpea is crucial for their genetic improvement. The current genetic profile of cowpea displays an inadequate level of characterization. High variance in taxonomic scope, biogeographic focus, the number of DNA barcode markers employed and the methodologies used for making taxonomic assignments makes comparisons among past studies difficult. In fact, no prior study has involved a large-scale comparative analysis of the capacity of the seven standard barcode markers to deliver a varietal level identification for different biogeographic gene pools of cowpea using standard barcode library.

2. The *cp*-Genome

The *cp*-DNA architecture of plants is currently a focus of research in plant molecular evolution, phylogenetics and systematics. Several unique genome profile of *cp*-DNA confers plants excellent molecular evolutionary analyses. Foremost is the fact that, the *cp*-DNA genome is relatively small besides constituting an abundant component of cellular DNA. Further, it is not in doubt that the *cp*-DNA genome has been extensively characterized at the molecular level which has had the resultant effect of providing the basic information that support comparative evolutionary research and analyses. More importantly, the *cp*-DNA has relatively slow rates of nucleotide substitution which has the advantage of providing the requisite window of resolution for studying plant phylogeny and systematics at deep levels of molecular evolution. Despite a fairly conservative rate of genetic evolution and relatively stable gene content, comparative molecular analyses reveal complex patterns of mutational change. Noncoding regions of *cp*-DNA diverge through insertion/deletion changes that are site dependent. Rates of molecular change are often reported to vary among plant families and in a manner that violates the assumption of a simple molecular clock. Protein-coding genes exhibit patterns that are known to reveal subtle anatomical and functional relationships. Comparative studies of molecular sequences have the resolution to reveal this underlying complexity. To fully understand the mechanisms of evolutionary change and in formulating realistic models of mutational processes, a complete description of the complexity of molecular signature change is necessary. The conservation of *cp*-DNA gene content and a relatively slow rate of nucleotide substitution in coding genes have made the *cp*-genome an ideal focus for studies of plant evolutionary history [2]. The *cp*-DNA gene loci can enable the reconstruction of plant evolutionary history at a level of detail that is unprecedented in molecular systematics (**Figure 1**). Early work on *cp*-DNA sequences suggested that relative rates of nucleotide substitution do not follow a constant molecular clock but rather that substitution rates in chloroplast loci vary among evolutionary lineages.

Several DNA fingerprinting and genotyping assays based on molecular markers have been developed in the past and are still in use today [3]. Phylogenetic inference of a data set composed of *matK* and *rbcL*, sequences from basal angiosperms demonstrated parsimony informative traits and significantly more phylogenetic structure on average per parsimony informative site than the highly conserved chloroplast gene *rbcL* [4]. In the same study, sequence information from *matK* alone generated phylogenies as robust as those constructed from data sets comprised of 2 - 11 other genes combined [5]. The molecular information

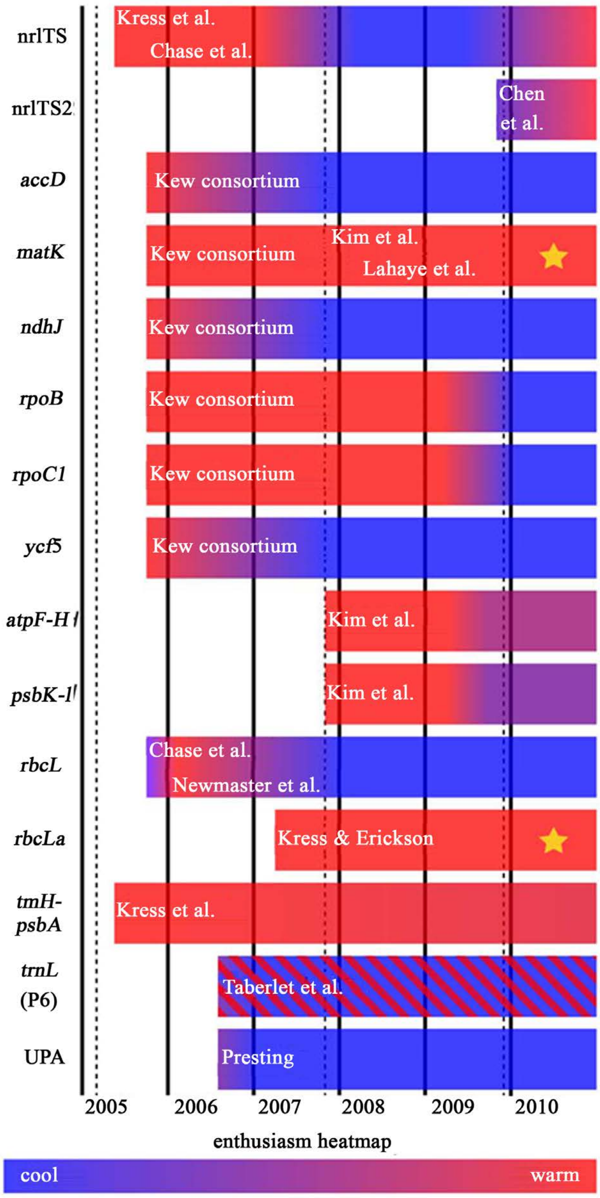


Figure 1. Schematic timeline of different DNA barcode markers. Colours (red = warm; blue = cool) represent an informal measure of enthusiasm among DNA barcoding researchers in the systematics community for CBOL and iBOL: Adapted from: PLoS ONE, http://www.plosone.org/.

generated from *matK* locus has been used to resolve many phylogenetic relationships from shallow to deep taxonomic levels among species [6] [7]. This locus distinguishes itself among plastid genes used in plant systematics due to its distinct mode and tempo of evolution. The *matK* rate of substitution is three times higher at the nucleotide level and is reported to be six times higher at the amino acid level than that of *rbcL*, denoting it as a rapidly evolving gene locus [8] [9]. The alignable and coding nature can facilitate character based analyses and allow inference of DNA barcode based diversity among phylogeographically diverse groups.

2.1. The *trnH-psbA* Intergenic Spacer

The *trnH-psbA* region is a straightforward region easily amplifiable across land plants, and is one of the most variable intergenic spacers [10]. It has been used successfully in a range of barcoding studies. [11] reports that the *trnH-psbA,* non-coding intergenic region exhibits significant sequence divergence with notable insertion/deletion rates. Studies by [12] indicate that this plastid region has highly conserved coding sequences that makes it an attractive marker. These attributes make *trnH-psbA* an important plant barcode for species discrimination. However, the complex molecular evolution and considerable length variation of *trnH-psbA* limits it as a barcode singly [13]. However, *trnH-psbA* is reported to suffer high rates of insertion or deletion in larger families of angiosperms. The *trnH-psbA* putative gene loci albeit a standard barcode region in most plants has been reported to suffer frequent inversions in some lineages of plants and singly as a barcode marker, may result to over estimation of genetic divergence and consequently inaccurate assignment of phylogenetic position. The *trnH-psbA* is the most widely used plastid marker with highly conserved coding sequences on both sides making the design of universal primers feasible. The non-coding intergenic region is reported to exhibit most sequence divergence with high rates of insertion/deletion [14]. These attributes make *trnH-psbA* highly suitable as a plant barcode for species discrimination. Alignment of the *trnH-psbA* spacer can be highly ambiguous because of its complex molecular evolution and considerable length variation [15], high rates of insertion or deletion in larger families of angiosperms [16].

2.2. The *matK* Gene Locus

The *matK* gene has been used in the reconstruction of grass phylogenies and to reveal polymorphisms. This gene sequence is one of the seven DNA candidate loci widely recommended for the DNA barcoding of plants [17]. Chloroplast genes and intergenic spacers have been amplified and amplicons used to reveal polymorphisms following direct sequencing [18]. The *matK* gene loci is the only putative group II intron maturase encoded in the chloroplast genome of plants and is the only plastid gene containing this putative maturase domain in higher plants [19]. The *matK* locus is maturase-kinase gene region, a plastid gene which is responsible for the chloroplast post-transcriptional processing in plants. It has an unusual evolutionary tempo, with relatively high substitution rates at both nucleotide and amino acid levels according to [20]. The strong phylogenetic signal from *matK* gene renders it invaluable gene loci in plant systematics and evolutionary studies at various evolutionary depths. This gene locus has 1500 base pair nested in the group II intron of the 50 and 30 exons of *trnK* in the large single copy region of the chloroplast genome of green plants. The *matK* gene sequence is one of the seven putative gene loci widely utilized in the DNA barcoding of land plants. Phylogenetic analysis based on *matK*, against other candidate genes has demonstrated excellent parsimony informative characters with significantly more phylogenetic structure per each parsimony-informative site contrary to the highly conserved chloroplast/plastid region. The *matK* sequence information has been reported to generate robust phylogenies and is considered to have a reliable evolutionary rate, suitable length and good interspecific divergence as well as a low transversion rate [21]. The *matK* locus is however difficult to amplify universally demonstrating that *matK* barcode albeit informative, may be inadequate and inconclusive when used in isolation as a universal barcode. This study therefore considered *matK* alongside six barcode markers.

2.3. The *rbcL* Locus

The *cp*-gene *rbc*L—RuBisCo large subunit is an ideal candidate barcode region and is considered the most abundant protein on earth, RuBisCo (Ribulose-1, 5-bisphosphate carboxylase oxygenase) catalyzes the first step of carbon fixation. *rbcL* is commonly used in phylogenetic investigations with over 50,000 sequences deposited in Genbank. It is easily amplifiable, sequenced and aligned. It is used in the catalysis of the first step of carbon fixation process and is a target region in many phylogenetic investigations due to its ease of amplification, sequencing and alignment. Many taxonomists consider *rbcL* gene an ideal DNA barcoding region at both family and generic level. The *rbcL* locus has the lowest divergence of plastid genes in flowering plants according to [22] [23] [24] report modest discriminatory power of this locus. Other studies however indicate that *rbcL* remains one of the best candidate barcodes based on the straightforward recovery of the gene sequence, easy accessibility and discriminatory power. The *rbcL* locus has by far the lowest divergence of plastid genes in flowering plants. While this region is considered versatile, no single study has explored its use in biodiversity studies in a Kenyan situation.

2.4. The *atpF-atpH* Intergenic Spacer

The second International Barcode of Life Conference proposes that *atpF-atpH* intergenic spacer is a potential plant barcode region. The fact that *atpF-atpH* marker has not been widely used in plant systematic and phylogeographics has led to paucity of data on its performance as barcode loci. However, the CBoL Plant Working Group indicate that *atpF-atpH* has relatively modest discriminatory power, intermediate sequence quality and universality and could be used as a plant DNA barcode. Recent studies document positive reports on the performance of *atpF-atpH* as a plant barcode region [25]. Studies on duckweeds [26] also demonstrated that *atpF-atpH;* a noncoding spacer could serve as a universal DNA barcoding marker for species-level identification. In the study on duckweeds, the utility of this non cording region in identification of new species by reason of its ease of amplification, straightforward sequence alignment and rates of DNA variation was reported [27]. In the same study, it's documented that DNA barcoding made significant contribution to the taxonomical structure in duckweeds as opposed to the less informative morphological classification and therefore recommends *atpF-atpH* as an important barcode region in biodiversity studies. The current study therefore sought to among others test the in formativeness of *atpF-atpH loci* in delineating cowpea diversity at sub-species level.

2.5. *psbK-psbI* Intergenic Spacer

The *psbK* and *psbI* locus encodes two low molecular weight polypeptides, *K* and *I* respectively, for the photosystem II and are conserved from algae to land plants [28]. The potential of the *psbK-psbI* intergenic spacer as a barcode for plants was tested in the flora of the Kruger National Park, South Africa. The results revealed its high PCR amplification and sequencing performances (98% of taxa) and ease in the alignment of sequences. Accordingly, *psbK-psbI* was proposed in conjunction with markers such as matK, *trnH-psbA* and *atpF-atpH* as appropriate for plant barcoding. The CBOL–Plant Working Group also observed that the species discriminatory power of this locus was better than that of *matK* and other loci, except *trnH-psbA*. However, due to the inconsistency in getting bidirectional unambiguous sequences, this has only been considered as a supplementary locus.

2.6. The *rpoB* and *rpoc*1 Gene Loci

The loci *rpoB* and *rpoc*1 are reported to encode three out of the four subunits of the *cp*-DNA. Genome wide substitution analyses in previous studies reveal that *rpoB*, *rpoC*1 and *rpoC2* are accumulating higher amount of nonsynonymous substitutions, which indicates either positive or relaxed selection. Their high substitution rate makes these genes highly suitable for phylogenetic inferencing. The locus *rpoC* has been reported as suitable for phylogenetic analysis. Currently, *rpoB* has been considered as the core gene for phylogenetic analyses and identification, especially when studying closely related isolates alongside 16S rRNA gene, the *rpoB* locus helps delineate new and refine bacterial flora [29]. These loci *rpoA*, *rpoB*, *rpoC*1 and *rpoC2* collectively, are considered ideal for phylogenetic studies. After extensive studies, these genes have been proposed for barcoding either individually or in combination by various research teams. The CBOL–Plant Working Group, on the other hand, observed that the species discrimination of *rpoC*1 was the least (43%) among the seven loci tested. Nonetheless, in recent years, *rpoC*1 has been found highly useful for barcoding the bryophytes [30]. In a nutshell, further research on these gene sequences is necessary in determining their suitability as an ideal barcode.

2.7. Molecular Phylogeny

Similarity of Biological functions and molecular mechanisms in living organisms often suggests an ancestral lineage of an organism. Molecular phylogeny explores the structure and function of molecules and how they change over time to help infer evolutionary relationships. Automated DNA based approaches can accelerate cataloguing of organisms. The primary objective of molecular phylogeny studies is reconstruction of the order of evolutionary events and proceed to represent such events in evolutionary trees which then graphically depict relationships among species or genes over time. Phylogenies are a fundamental tool for organizing knowledge of Biological diversity, for structuring classifications and for providing insight into events during evolutionary process. Furthermore, phylogenetic trees show descent from a common ancestor, thereby providing overwhelming evidence supporting the theory of evolution.

Molecular phylogenies can be generated from character datasets that provides evolutionary content and context of an organism. Bio molecular sequence alignments can also reveal character data of DNA. Molecular markers such as Single Nucleotide Polymorphisms (SNPs) can also inform genetic diversity. Evolution is modeled as a process that changes the state of a character, such as the type of nucleotide (AGTC) at a specific locus in a DNA sequence in which each character is a function that can map a set of taxa to distinct states [31] [32]. The idea dates back to Darwin, but the numerical calculation of trees using quantitative methods is relatively recent [33], and their application to molecular data even more so [34]. In the age of rapid and rampant gene sequencing, molecular phylogeny has come into its own, emerging as a major tool in Biodiversity studies. Genome-wide data can provide novel opportunities for resolving phylogenetic relationships. Multiple sequence alignment as one of the many heuristic methods for aligning sequences exist and improved algorithms continue to emerge.

Phylogenetic inference has recently gained popularity among molecular taxonomists and is used to describe relationships between paralogues within a gene family [35], histories of populations [36], the evolutionary and epidemiological dynamics of pathogens [37], the genealogical relationship of somatic cells during differentiation and cancer development [38] and the evolution of language [39]. In recent years, molecular phylogeny has become an indispensable tool for genome comparisons notably in classification of metagenomics [40]; to identify genes, regulatory elements and non-coding RNAs in newly sequenced genomes [41]; to interpretation of modern and ancient individual genomes [42] and reconstruction of ancestral genomes. Currently, phylogenetic inference using sequenced data has received great attention by molecular Biologists. The development of the coalescent theory [43] and the widespread availability of gene sequences for multiple individuals from the same species have prompted the development of genealogy-based inference methods. These methods have revolutionized Computational Biology. Knowledge about phylogenetic inference based on a suite of current methodologies for phylogenetic inference.

While molecular phylogeny is large and complex [44], it is nonetheless an important tool in taxonomy and systematics. Phylogeny reconstruction methods are either distance or character-based. In distance matrices, the distance between every pair of sequences is calculated, and the resulting matrix is used for tree reconstruction. Algorithms such as NJ [45] can apply a cluster algorithm to the distance matrix to resolve a phylogeny. Character-based matrices such as maximum parsimony, maximum likelihood is key in phylogenetic inference. In the current study, we explore a suite of current methodologies for phylogenetic inference based on sequence data of closely related cowpea variants. Genetic markers found on specific locations on a chromosome are considered landmarks for genome analyses [46]. However, molecular markers reveal greater polymorphisms at the protein or at DNA level as opposed to morphological traits.

3. Materials and Methods

3.1. DNA Amplification

Seven polymorphic *cp*-DNA genome signatures were used to screen and amplify three intergenic spacers' *atpF-atpH*, *psbK-psbL* and *trnH-psbA* and four genes: *matK*, *rbcL*, *rpoB*, *rpoC*1. PCR products of seven primer pairs with different dyes coloaded together in 96-well working plate vortexed and spined then an aliquot utilised. PCR products were resolved on a polyacrylamide gel (1%), using 0.5× TBE containing 1 mg/mL ethidium bromide with a vertical electrophoresis apparatus at 300 v. PCR amplification in a 0.2-mL PCR tube with a reaction volume of 25 µL, containing 2.5 µL 10× PCR buffer, 1 µM of each primer, 1 mM of each dNTPs, 0.5 U *Taq*DNA polymerase and 50 ng DNA. Tubes placed in an Eppendorf Master Cycler Gradient thermocycler programmed for initial denaturation at 94˚C for 1 min followed by 35 cycles of 30 s at 94˚C, 30 s at 55˚C, 1 min at 72˚C, and a final extension of 10 min at 72˚C. The PCR conditions were as follows: initial denaturation at 94.0˚C for 15 minutes; 30 cycles of denaturation at 95.0˚C for 1.0 minute, annealing temp of 50˚C for 40.0 seconds, extension at 72.0˚C for 1.0 min 30 sec and final extension at 72.0˚C for 5.0 minutes. The 30 µL total reaction mixture for the amplification of either the seven candidate *cp*-DNA contained 1.0 mL × PCR buffer, 1.67 mM $MgCl_2$, 0.3 mM dNTPs, 0.3 µM of each primer and 2.5 U *Taq*DNA together with 2.0 µL of DNA template. The PCR conditions were as follows: initial denaturation at 95.0˚C for 5 minutes; 35 cycles of denaturation at 95.0˚C for 1.0 minute, annealing temp at 56.0˚C for 30 seconds, extension at 72.0˚C for 1.0 min 15 sec; and final extension at 72.0˚C for 7 minutes.

3.2. Gel Electrophoresis

The amplicons were resolved on 1% agarose gel at 80 V for 48 minutes. The gels were observed for bands based on a UV trans-illuminator (FotoDyne model 3 3500 Foto-Prep). Photographs of the bands were taken using the software "Strata-gene Eagle View" that was integrated with the digital camera on the UV trans-illuminator. The gel was rinsed with distilled water and air dried. An aliquot of the PCR product was checked by agarose gel electrophoresis. The resolved products were extracted from the gel and purified using the Qiagen DNA purification kit according to the prescribed protocol. DNA quantification was done by using a DNA Nano Drop 2000/2000 c Spectrophotometer. The bands containing DNA of interest was excised and the DNA purified using the DNA purification kit from Qiagen plant DNA extraction kit following manufacturer's guidelines

3.3. DNA Sequencing

Cycle sequencing was done using Big Dye terminator v3.1 and sequencing on a 3130 xl genetic analyzer (Applied Biosystems, USA); electropherograms were edited using SEQUENCER 4.6 software (Genes Codes Corporation, USA) and DNA sequences aligned by BioEdit. Incomplete sequences at both ends were excluded from the analyses. Failed sequencing was exempted from the combined matrix in order to analyse complete matrices. The forward primer for each of the 7 markers was labelled at the 5' end of the oligonucleotide using fluorescent dyes for detection by the automated sequencer ABI 3730 genetic analyzers (Applied Biosystems). The 25 μL PCR reaction mixture for the amplification of the chloroplast gene contained 1.0× PCR buffer, 1.5 mM $MgCl_2$, 1.0 μL (99%) dimethyl sulphoxide, 0.4 mM dNTPs, 0.4 μM of each primer and 2.5 U TaqDNA polymerase (Super-Therm) (all supplies obtained from Bioneer Company South Korea together with 1.0 μL of DNA template. Sequencing primers are as indicated in Table 1.

3.4. Sequence Alignment

Sequence alignment was done by BioEdit. Similarity searches to find homologous sequences were conducted based on Basic Local Alignment Search Tool (BLAST) tool located at www.ncbi.nih.gov/blast; with the parameters set as follows: database-non redundant; search-mega blast and the expectant value set at E value ≤ 10^{-5}. The sequence that had the lowest expectant value (E-value) and

Table 1. List of *cp*-DNA genes/intergenic spacers amplified in the present study including primers and approximate amplicon lengths.

Genes/intergenic pacers	Primer Pair (5′-3′)	Amplicon length	T_a (˚C)	Source
atpF-atpH	ACTCGCACACACTCCCTTTCC GCTTTTATGGAAGCTTTAACAAT	621 bp	48˚C	Ki-Joong Kim; kimkj@KOREA.AC.KR)
***rpoc*1**	GGCAAAGAGGGAAGATTTCG CCATAAGCATATCTTGAGTTGG	490 bp	53˚C	http://www.kew.org/barcoding/protocols.html
rpoB	ATGCAACGTCAAGCAGTTCC CCGTATGTGAAAAGAAGTATA	490 bp	51˚C	http://www.kew.org/barcoding/protocols.html
matK	CGTACAGTACTTTTGTGTTTACGAG CCCAGTCCATCTGGAAATCTTGGTTC	892 bp	49.5˚C	Ki-Joong Kim; kimkj@KOREA.AC.KR)
psbK-psbI	TTAGCCTTTGTTTGGCAAG AGAGTTTGAGAGTAAGCAT	576 bp	60˚C	Ki-Joong Kim; kimkj@KOREA.AC.KR)
rbcL	GTAAAATCAAGTCCACCRCG ATGTCACCACAAACAGAGACTAAAGC	596 bp	50˚C	David Erickson; ERICKSOND@si.edu)
trnH-psbA	GTTATGCATGAACGTAATGCTC CGCGCATGGTGGATTCACAATCC	812 bp	50˚C	David Erickson; ERICKSOND@si.edu

Legend: The chloroplast markers *rbcL, rpoB, rpoC*1*, matK, atpF-atpH, trnH-psbA*, and *psbK-psbI*, proposed by the CBoL plant-working group, were amplified with a set of primers (**Table 1**). The amplicon sizes are shown. PCR reaction conditions followed guidelines from the CBOL plant-working group.

the highest Identity score was considered to be a similar sequence. BLAST used our sample sequence as the query sequence to search NCBI databases, to look for similar sequences, and then used a similarity matrix to measure the similarity between sequences and the possibility that the similarity could be due to chance based on the nucleotide sequence of the query versus its target. The higher the bit score, the more closely related that sequence was to the query sequence. The E value, on the far right, was the number of search matches to the current non-redundant sequence database expected by chance alone. The smaller the E value of the BLAST hit, the more likely that the similarity reflected a common descent and not that it occurred by chance. Sequences with an E value $\leq 10^{-5}$ were considered homologues. For sequence annotation and gene ontology, the contigs were analysed to predict Open Reading Frame (ORF). The NCBI taxonomy tool (https://www.ncbi.nih.gov/taxonomy) was used to infer the complete classification of the sequence which was confirmed in the International Plant Names Index website. The sequences were then named based on their molecular character. The sequences were then assembled into a single Fasta file format in BioEdit software version 7 (Thomas Hall & Abbott). This was followed by performing a local alignment using ClustalW in Mega 6 software with UPGMA as the clustering method.

Consensus sequences were generated and sequences of the candidate DNA barcodes aligned using ClustalW and verified by BioEdit. Sequence alignment was initially performed using ClustalW and manually adjusted using MEGA v.7.0. Phylogenetic analyses were performed using maximum-parsimony (MP) approaches. Maximum-parsimony (MP) analyses involved a heuristic search strategy based on 1000 replicates of random addition of sequences. Genetic distance matrices were calculated on the basis of Kimura 2-Parameter (K2P) substitution model for the seven chloroplast candidate DNA loci and the average values between subpopulations inferred. The distance matrices were inferred based on the Kimura 2 parameter substitution model for the seven chloroplast candidate DNA loci and the average values between subpopulations inferred.

3.5. Evolutionary Relationship

Analyses of DNA barcode sequences for cluster recognition provides an efficient approach for recognizing putative operational taxonomic units (OTUs). Sequence reads were edited by ChromasLite to collapse redundancies and unknown bases and to generate consensus sequences from sequence fragments with subsequent nucleotide alignments of partial *cp*-DNA loci and intergenic spacers to generate a consensus sequence using the multiple MAFFT online alignment. BLAST search on the Gene Bank database was performed to decipher similarities. The evolutionary history was inferred using the Neighbor-Joining algorithm [47]. The bootstrap consensus tree inferred from the 1000 replicates was considered to represent the evolutionary history of the sub-populations. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed [48]. The percentage of replicate trees in which the subpopulations clustered in the bootstrap test (1000 replicates) is indicated above the branches [49]. The evolutionary distances were computed using the Jukes-Cantor method [50] and are in the units of the number of base substitutions per site. The analysis involved 54 nucleotide sequences. Codon positions included were 1st + 2nd + 3rd + Noncoding. All positions containing gaps and missing data were collapsed. In estimating evolutionary divergence between sequences, the numbers of base substitutions per site are indicated. In order to confirm identities and show evolutionary relationships between operational taxonomic units (OTUs), the sequences were independently subjected to phylogenetic analyses. Initial alignment was done by BioEdit v. 7.0. and the rest of the analyses done in MEGA 6 software. Study sequences were subjected to estimation of the evolutionary divergence based on pairwise-distance tool in MEGA 6. The sequences were arranged in order of the plastid genes targeted in the study and were aligned, cut to size and concatenated using MAFFT v. 6.857b software. The concatenated sequences were transferred to MEGA 7.0 software and a combined phylogenetic tree constructed. To calculate the overall mean evolutionary divergence. Tajima's D test was conducted in MEGA to detect the nature of gene selection, nucleotide variation per site and nucleotide divergence per sequence.

3.6. Molecular Phylogenetic Analyses

MEGA 6 software was applied in phylogenetic analyses. Subsequent phylogenetic analyses and concatenation revealed between and within cowpea variants degree of diversity and similarity. Each of the raw sequence files were inspected and corrected prior to phylogenetic analysis using BioEdit software version 7.0. These sequences were compared to those in Gene Bank (National Centre for Biotechnology Information; http://www.ncbi.nlm.nih.gov/) using BLAST to verify identity and followed by phylogenetic data analyses. Consensus sequences from the forward and reverse primer pairs were generated in BioEdit ver. 7.0 [51] and subsequent nucleotide alignment generated using CLUSTAL W (Thompson *et al.,* 1994) and implemented in BioEdit *ver.* 7. The alignment file was subsequently loaded in MEGA 7 to infer evolutionary history using the algorithm of Neighbor-Joining method [52]. The bootstrap consensus tree was inferred based on 1000 replicates which were taken to represent the evolutionary divergence of the analyzed accessions. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed and not included in the final analyses. The evolutionary distances were computed based on the Jukes-Cantor method and were in the units of the number of base substitutions per site. The analyses involved various nucleotide sequences depending on each putative DNA loci/intergenic spacer. Codon positions included were 1st, 2nd, 3rd and noncoding. All positions containing gaps and missing data were collapsed.

DNA sequences were visualized and manually edited using BioEdit software to minimize sequencing errors and removal of gaps in the coding regions that could cause shifts in the open reading frames. The BLASTn algorithm (http://www.ncbi.nlm.nih.gov/BLAST) was used to perform sequence similarity searches against the non-redundant nucleotide databases of NCBI. The correspondence between the sequences of the PCR amplicons and the known sequences was inferred. Multiple sequence alignments were performed by BioEdit and the intra- and infraspecific genetic divergences were calculated using MEGA 6 beta software [53] according to the Kimura 2-parameter distance model [53]. Based on the pairwise nucleotide sequence divergences, the neighbor-joining (NJ) tree algorithm was used to infer diversity of all the accessions. A bootstrap analysis was conducted to infer the stability of the computed branches with 1000 resampling replicates. All nucleotide positions with gaps and missing bases were collapsed from the data set using complete deletion. Identification of accessions was based on generating a phylogenetic tree. The trees were built with MEGA 6 using the best algorithms methods of UPGMA and MP compared with other tree building techniques for DNA barcoding. UPGMA trees were made from K2P distances. The MP trees were constructed based on the close neighbour interchange (CNI) method with search level 1. Each tree contained the bootstrap value as generated by Mega 7.0 software. To confirm identities and evolutionary relationships between subpopulations of vigna, the sequences were independently subjected to phylogenetic analyses using MEGA 6 and the Neighbor-Joining method in a *p*-distance matrix model. To concatenate the plastid marker genes, sequences were arranged in a similar order based on on Fast Fourier Transform (MAFFT v. 7) and the genes combined manually prior to construction of a concatenated phylogenetic.

3.7. Phylogenetic Reconstruction

The use of a distance matrix of sequence similarity to produce a hierarchical clustering phenogram remains popular today [54]. The argument is that distance matrices are fast and easily computable when dealing with large data sets and have a reputation of resolving easily. In order to confirm identities and show evolutionary relationships between variants of vigna spp, seven chloroplast genes were independently subjected to phylogenetic analyses by employing MEGA 7.0 and the NJ algorithm in a *p*-distance model. To confirm the tree topologies, maximum-likelihood method was inferred [55]. MEGA 7.0 was used to estimate evolutionary divergence of the collections and concatenated chloroplast gene sequences were arranged in order using BioEdit v. 7.0. and alignment of the sequences done using multiple sequence alignment based on Fast Fourier Transform (MAFFT) and the genes manually combined to allow construction of a concatenated phylogenetic tree and subsequent estimation of mean evolutionary divergence Initial alignment was done by ClustalW in BioEdit v.7.0. and subsequently in MEGA 6 software. The study sequences were also subjected to estimation of the evolutionary divergence by using the “pairwise-distance” tool in MEGA 7.0. The sequences of the study accession samples were arranged in a similar order for each of the plastid genes targeted in the study. To employ a multi-tiered barcoding technique [56] [57] [58] [59], a subset of cowpea germplasms was tested at several genomic loci to determine polymorphic barcode markers at the intraspecific level. All the seven candidate loci chloroplast gene regions provided variable and informative in delineating the various cowpea germplasms at varietal level. The suitability of single locus ability to resolve phylogenetic relationships at the varietal level was clearly demonstrated by each marker but at varying degrees.

4. Results

4.1. DNA Extraction and PCR Amplification

Good quality DNA was obtained which resolved well on 1% agarose gel. Robust amplification products were obtained (Figures 2-4). Overall, PCR amplifications were largely successful with all the primer pairs designed for each DNA region exhibiting clear amplicons. Although double PCR products were usually not detectable in the gel, any problems that may have arisen from multiple comigrating amplicons of similar size were eliminated or collapsed. Figures 5-11 indicate aligned sequence contigs for various markers.

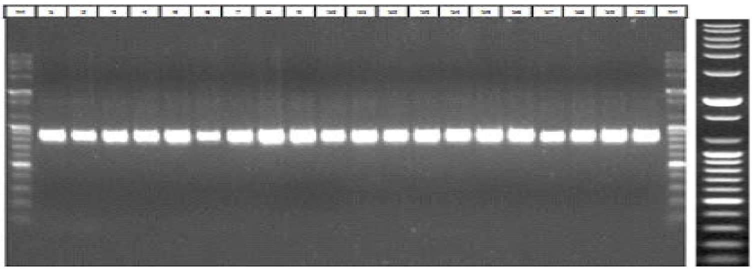


Figure 2. PCR profiles produced by *matK* loci [20 samples].

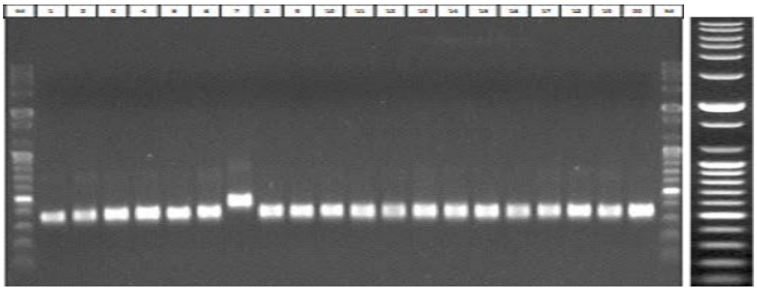


Figure 3. PCR profiles produced by *trnH_psbA* intergenic spacer [20 samples].

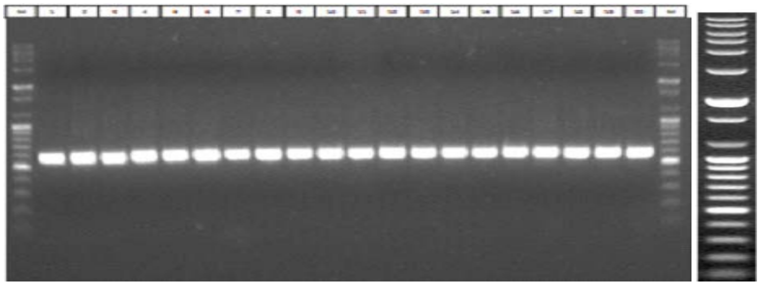


Figure 4. PCR profiles produced by *rbcL* gene loci [20 samples].

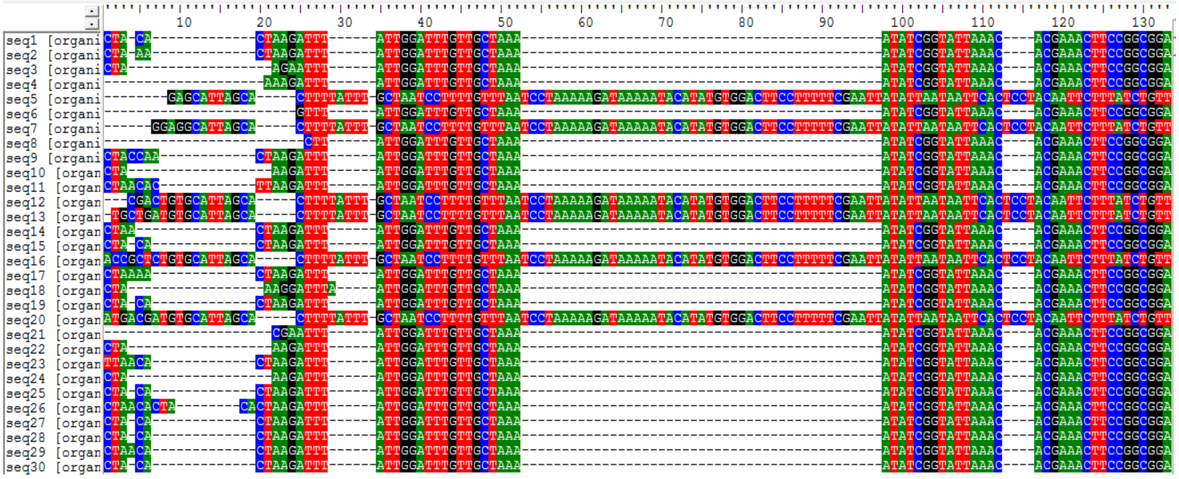


Figure 5. Sequence alignment [*atpF_atpH*].

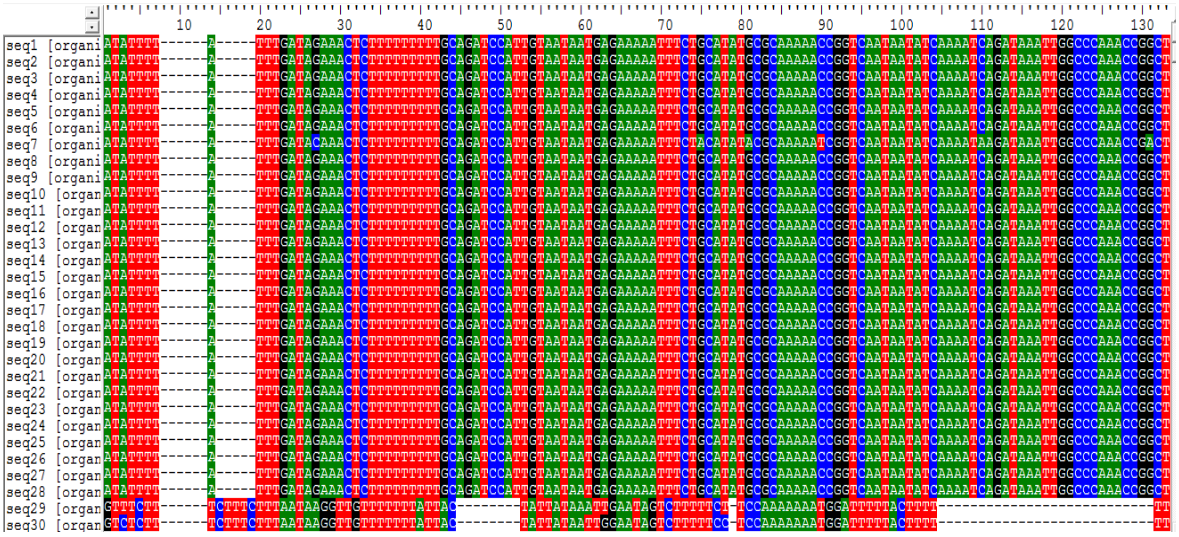


Figure 6. Sequence alignment [*matK*].

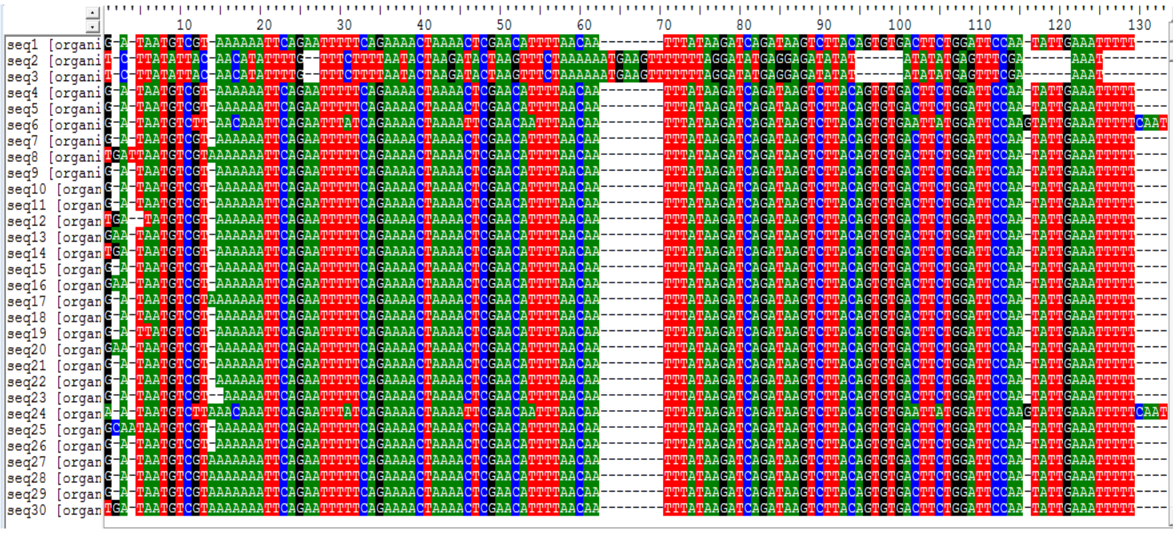


Figure 7. Sequence alignment [*psbK_psbL*].

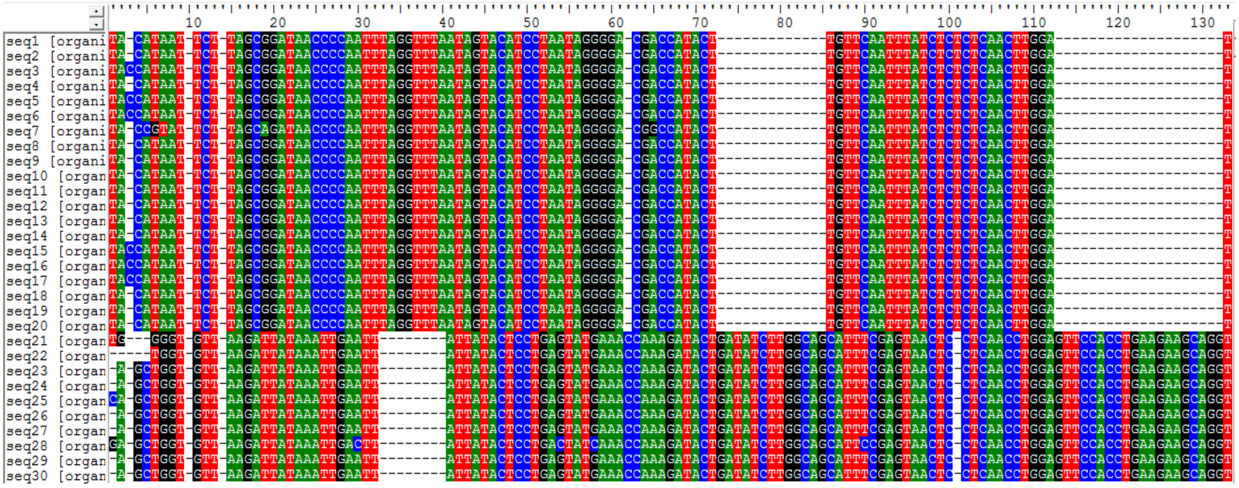


Figure 8. Sequence alignment [*rbcL*].

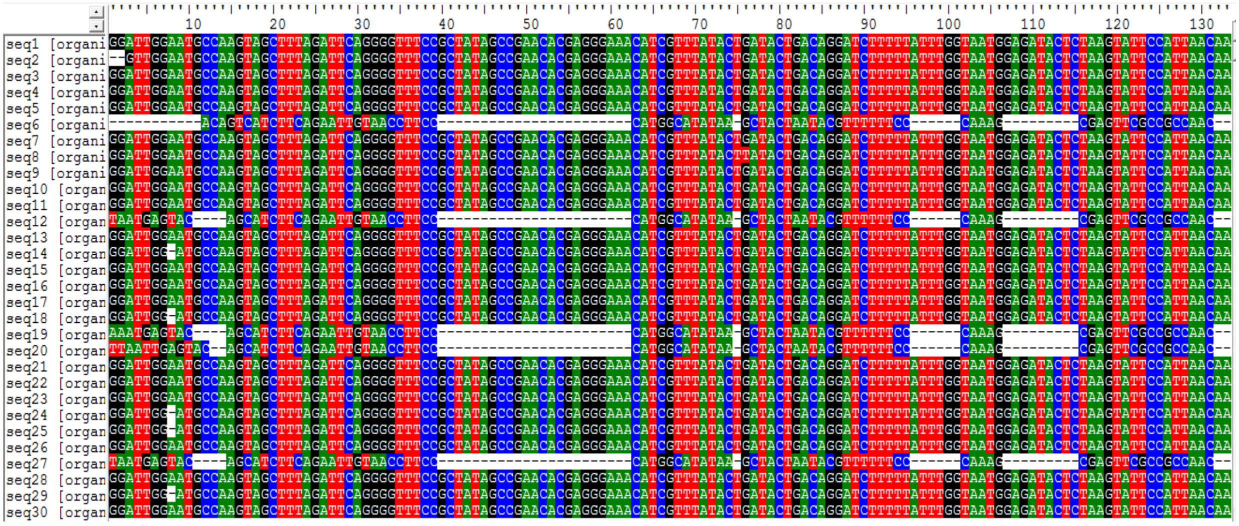


Figure 9. Sequence alignment [*rpoB*].

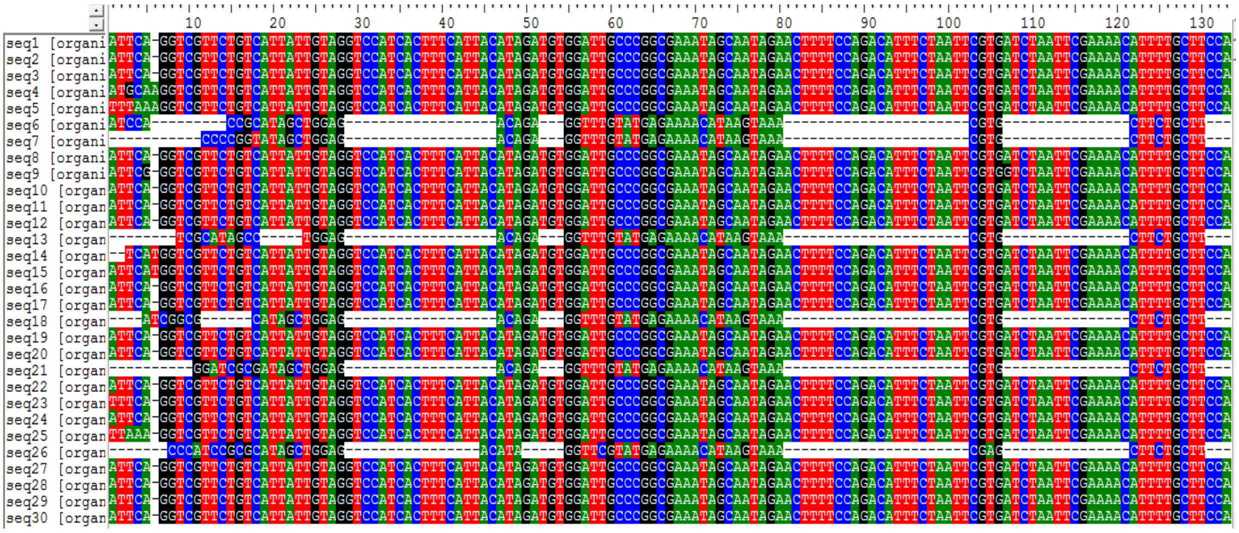
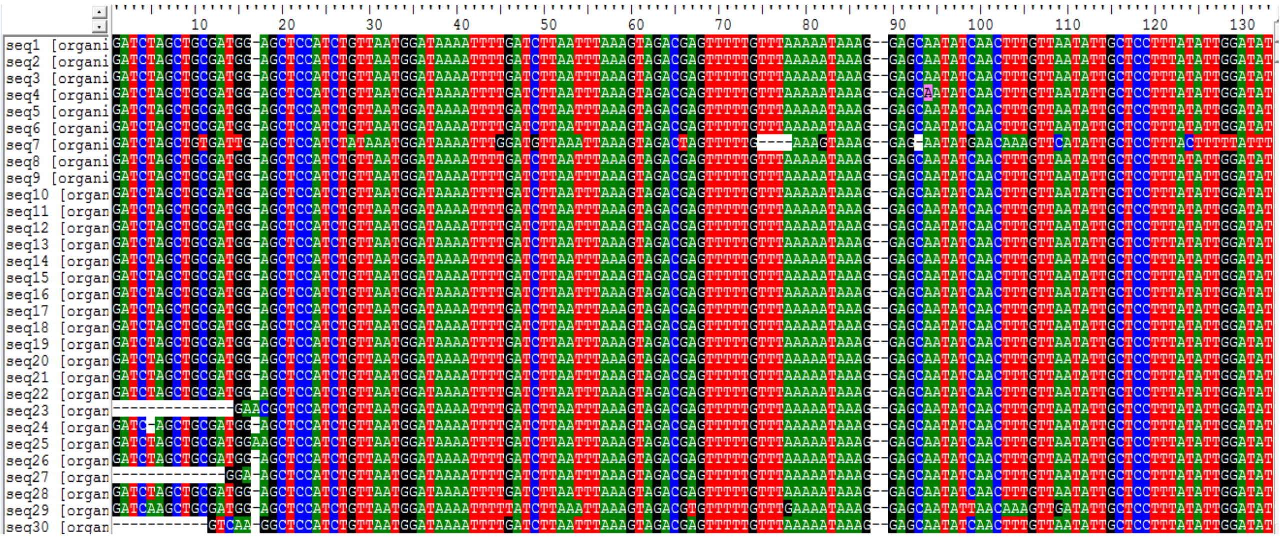


Figure 10. Sequence alighnment[*rpoc1*].



```
                     10        20        30        40        50        60        70        80        90       100       110       120       130
seq1 [organi GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq2 [organi GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq3 [organi GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq4 [organi GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq5 [organi GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq6 [organi GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq7 [organi GATCTAGCTGTGATTG-AGCTCCATCTATAAATGGATAAAATTTGGATGTTAAATTAAAGTAGACTAGTTTTTG----AAAGTAAAG--GAG-AATATGAACAAAGTTCATATTGCTCCTTTACTTTTTATTT
seq8 [organi GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq9 [organi GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq10 [organ GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq11 [organ GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq12 [organ GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq13 [organ GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq14 [organ GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq15 [organ GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq16 [organ GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq17 [organ GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq18 [organ GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq19 [organ GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq20 [organ GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq21 [organ GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq22 [organ GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq23 [organ --------------GAACGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq24 [organ GATC-AGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq25 [organ GATCTAGCTGCGATGGAAGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq26 [organ GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq27 [organ -------------GGA-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq28 [organ GATCTAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
seq29 [organ GATCAAGCTGCGATGG-AGCTCCATCTGTTAATGGATAAAATTTTTATCTTAAATTAAAGTAGACGTGTTTTTGTTTGAAAATAAAG--GAGCAATATTAACAAAGTTGATATTGCTCCTTTATATTGGATAT
seq30 [organ -----------GTCAA-GGCTCCATCTGTTAATGGATAAAATTTTGATCTTAATTTAAAGTAGACGAGTTTTTGTTTAAAAATAAAG--GAGCAATATCAACTTTGTTAATATTGCTCCTTTATATTGGATAT
```

Figure 11. Sequence alignment [*trnH-psbA*]. Legend: The conserved loci have same nucleotides [column]. Loci with variable nucleotides indicate divergence due to substitutions, deletions and or insertions.

4.2. Phylogenetic Analyses

The analyses involved a total of 298 sequences NCB published GenBank accessions [KX824129-KX824422]. Codon positions included were the 1st, 2nd, 3rd, and non-coding positions. Evolutionary analyses were conducted in MEGA software v7.0. Boot strapping with 1000 replicates used to measure branch confidence. Overall, the consensus trees topologies are presented here for each marker demonstrating the number of clades (clusters) resolved by each marker. Majority of the markers revealed between three to four clades represented by letters A, B, C, D and E. The most conserved marker was *rpoB* yielding only two clades while the most versatile and parsimonious region was *trnH-psbA* which revealed a total of five clades. In view of this, *rpoB* would not be a good marker for phylogenetic inferencing, because it is conserved. The *trnH-psbA* loci appear the most informative loci in inferring phylogenetic differences among the accessions (Figures 12-18). The marker *trnH-psbA* was the most parsimony informative in phylogenetic inferencing.

4.3. Discussion

4.3.1. Phylogenetic Analyses

Molecular phylogeny of cowpea variants was inferred based on the seven plastid markers singularly and collectively with a view to assessing the feasibility of these candidate loci in identification and intraspecific discrimination of phylogeographic groups into independent clades. The results are represented in Figures 12-18. An investigation of the relevance of DNA barcode loci to correctly delineate and cluster closely related cowpea variants into similar clades is presented and to evaluate the overall utility of chloroplast DNA barcode candidates in reconstructing phylogenetic relationships of cowpea at varietal level. The polymorphism in all the profiles was the result of nucleotide site mutations. Genetic

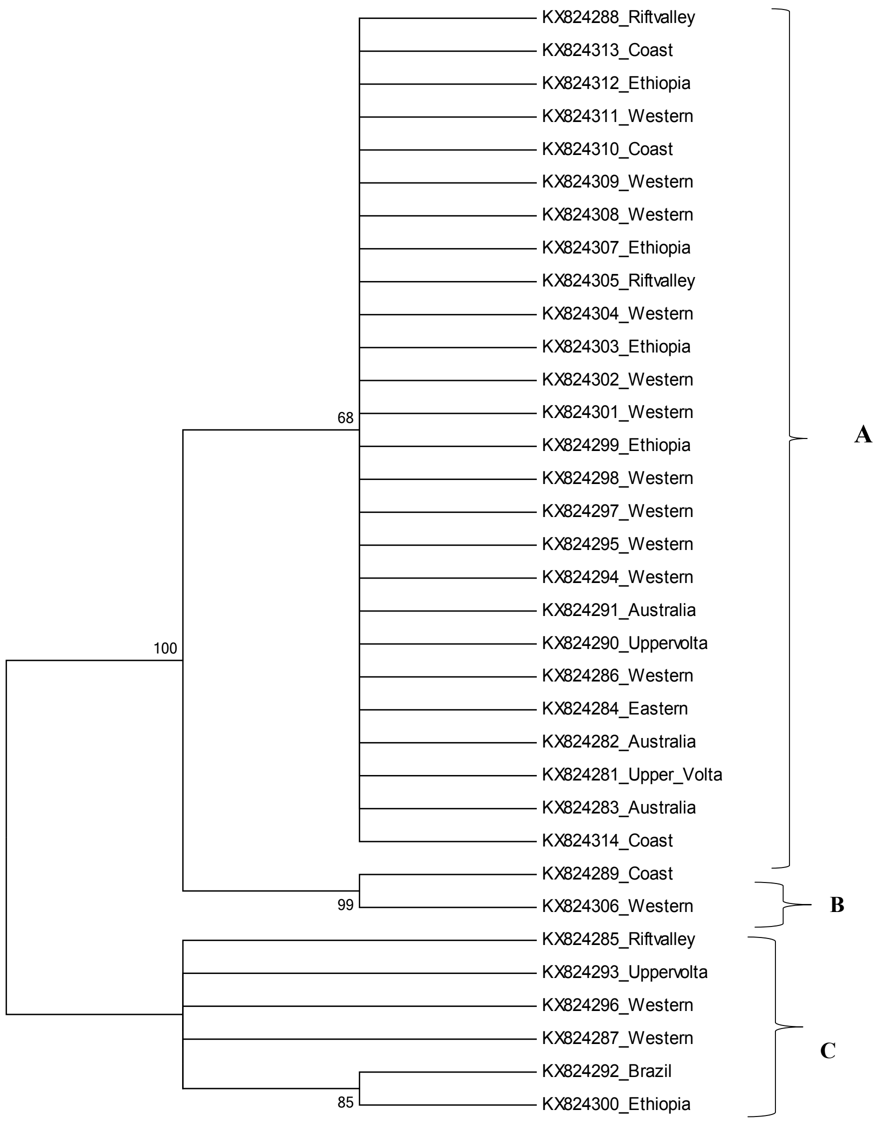


Figure 12. *atpF_atpH* loci: Phenogram of NJ cluster analyses. The phylogenetic tree based on *atpF_atpH* gene sequences, constructed using the neighbor-joining algorithm.

distance analyses based on the NJ algorithm generated dendograms with similar topologies resolving into 2, 3, 4 and 5 major clades as the case would apply. The marker *rpoB* was the most conservative generating only two clades (Figure 17). In line with this, *rpoB* gene locus owing to its conservative nature cannot be a good barcode locus. The markers *atpF-atpH, matK, psbK-psbL* and *rbcL* all generated three clusters each. ***trnH_ psbK*** was the most versatile generating five distinct clusters (Figure 18). The phylogeny reconstruction through the NJ method resulted in an optimal tree with a sum of branch length of 0.096 for *rpoB*, 3.11 for ***atpF_atpH***, 3.95 for *rbcL*, 0.81 for *matK*, and 1.15 for *psbK_psbL.* In the NJ bootstrap consensus tree, all the accessions were grouped into varying number of clades depending on each locus with distinct clades set at over 50% bootstrap support. The present study indicates that the dendograms based on MP analyses yielded similar topology as those of the NJ tree and gave parsimonious

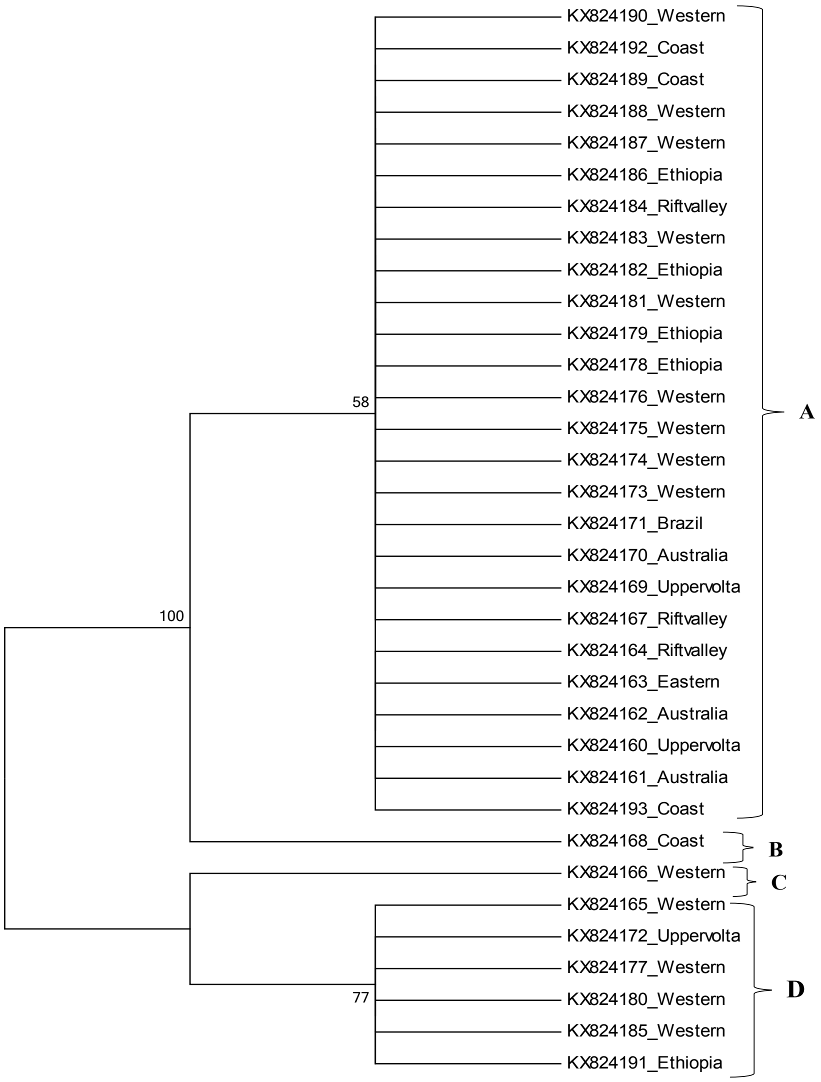


Figure 13. r*poC*$_1$loci: Phenogram of NJ cluster analyses. The phylogenetic tree based on rpoc1gene sequences, constructed using the neighbor-joining algorithm. Bootstrap values of more than 50 are shown at the internodes.

trees. The NJ tree allows conversion of sequence polymorphisms into genetic distance matrices based on nucleotide substitution models. Separate analyses for each barcode marker yielded NJ trees that correctly clustered together. In contrast, the NJ tree built for each barcode sequence did not show distinct uniqueness because tree phenograms generated were largely similar and this could be attributed to low divergence values among accessions perhaps because of the genetic homogeneity of cowpea as a crop. Overall, NJ phenograms constructed from the whole set of *cp*-DNA loci produced low discrimination among accessions owing perhaps to paucity of informative characters in some loci like *rpoB* that is highly conserved. This study documents 5 major clades (*trnH_psbA*) supported by high bootstrap values over 50%. However, each barcode loci identified several distinct haplotypes over all target regions corresponding to specific clades.

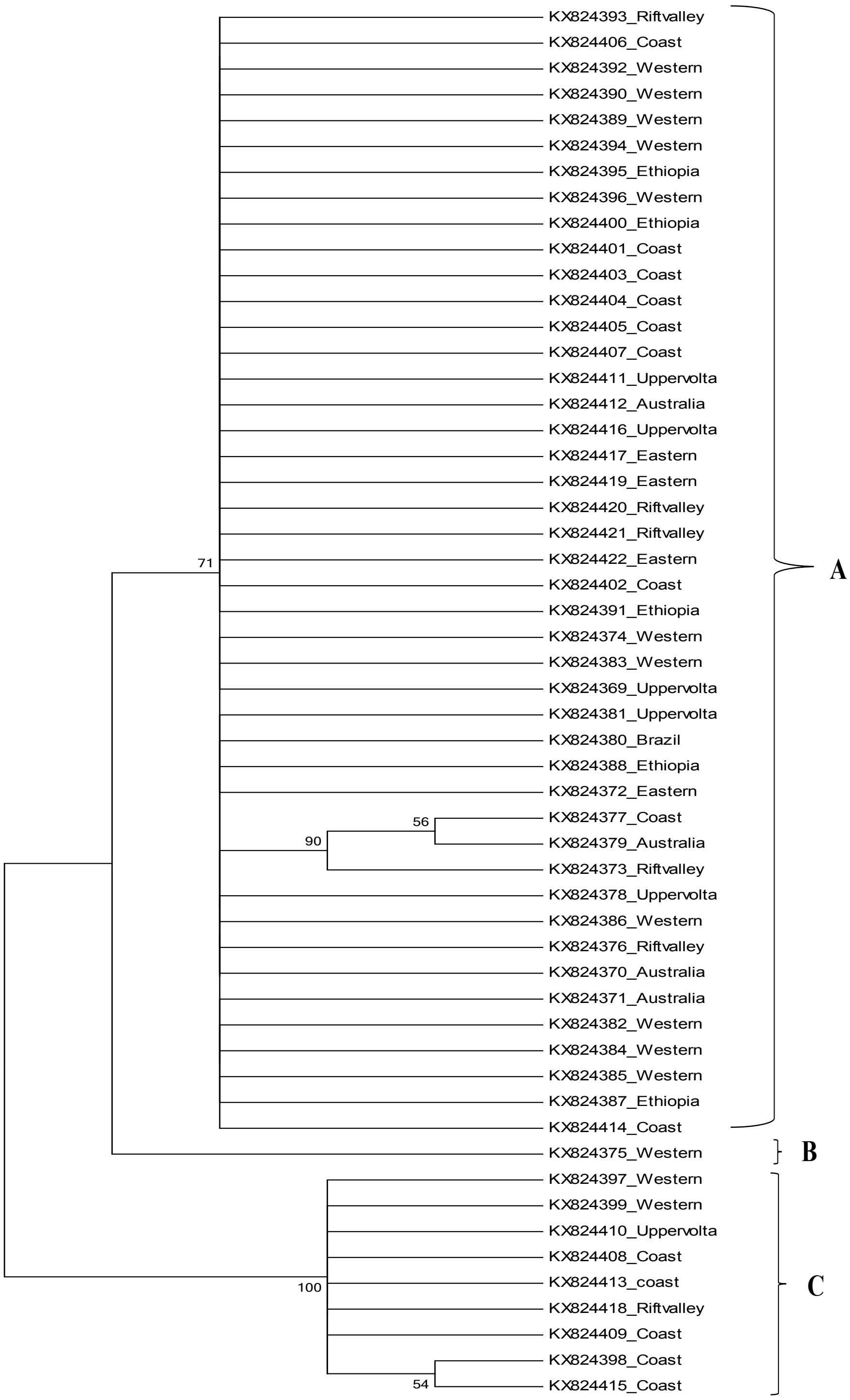


Figure 14. *matK* loci: Phenogram of NJ cluster analyses. The phylogenetic tree based on *matK* gene sequences, constructed using the neighbor-joining algorithm. Bootstrap values of more than 50 are shown at the internodes.

4.3.2. Phylogenetic Reconstruction Based on NJ Algorithm

In the current study, NJ algorithm was used to infer phylogenetic relationships. NJ algorithm has been documented empirically in phylogenetic analyses in many studies [60] and [61]. However, some studies report that NJ trees can be

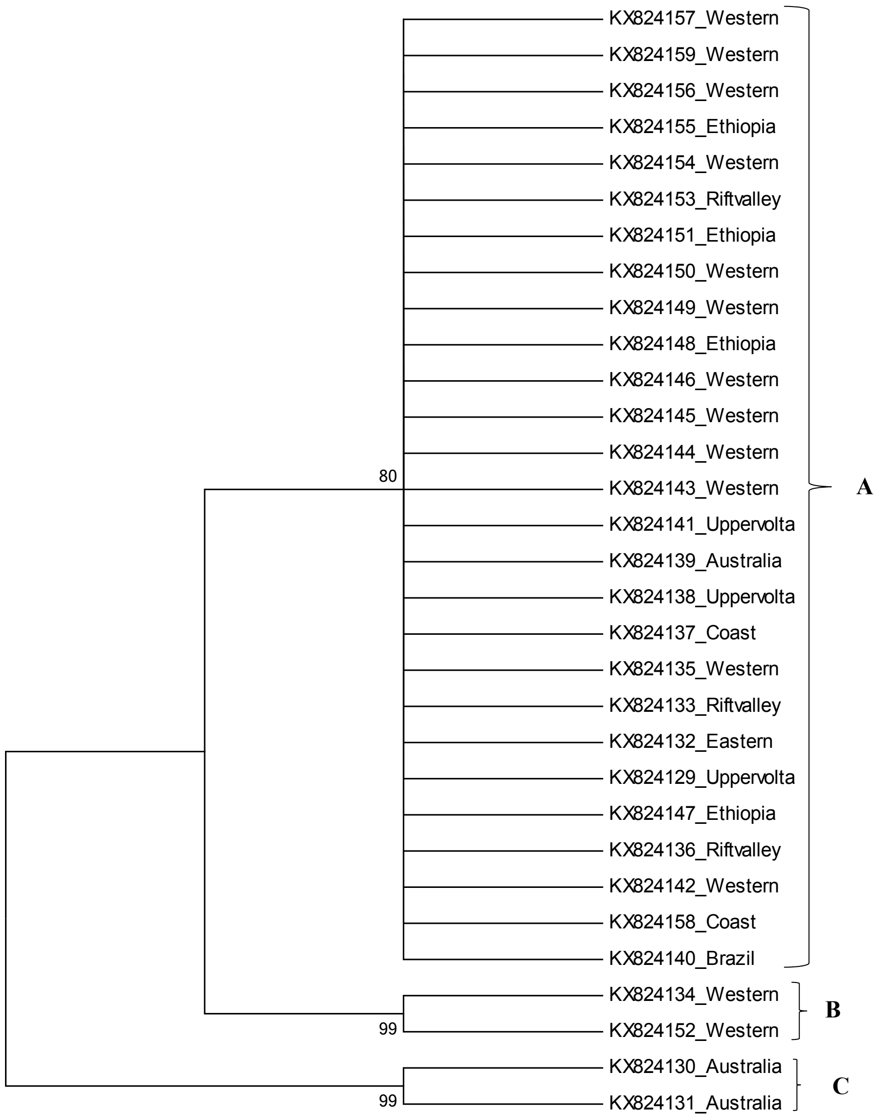


Figure 15. psbK_psbL loci: Phenogram of NJ cluster analyses. The phylogenetic tree based on pbsK_pbsL gene sequences, constructed using the NJ algorithm. Bootstrap values of more than 50 are shown at the internodes.

misleading to interpret in some circumstances, especially in cases of incompletely sampled reference library. Unless nested directly within a cluster, it's argued that the NJ tree may not discern if an unknown belongs to the closest topological cluster. In the current study, however, phylogenetic reconstruction NJ and MP methods placed the accessions into an average of 3 major clades consistent with several other studies. All sequences, whether analyzed separately or together, supported the distinctiveness of different varieties. In fact, nucleotide variability based on the occurrence of both SNPs and indels, clearly indicated genetic distinctiveness of these accessions. The chloroplast sequences contributed little or nothing toward resolving the genetic identities of landraces and varieties by resolving only a paltry 5 clusters. Although some concerns have arisen about the difficulties associated with the use of the *trnH-psbA* spacer [62], in the

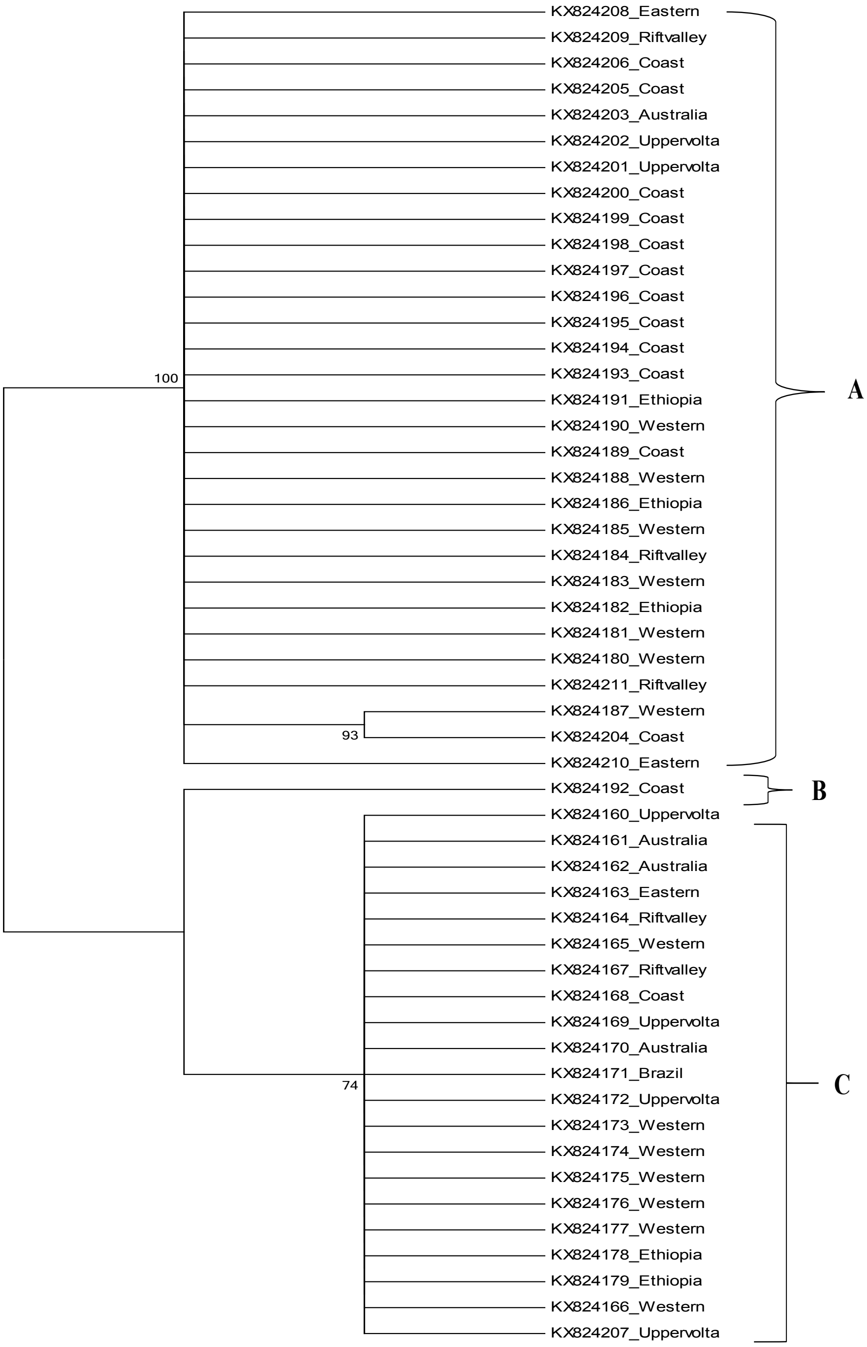


Figure 16. *rbcL* loci: Phenogram of NJ cluster analyses. The phylogenetic tree based on *rbcL* gene sequences, constructed using the neighbor-joining algorithm. Bootstrap values of more than 50 are shown at the internodes.

present study, the problems were not experienced with this marker and, on the contrary, it proved to be informative, followed by the *rbcL*. The NJ tree derived from the chloroplast combined data set as well did not exhibit a geographically related branching pattern, because the accessions from different regions clustered in similar clades. In this work, therefore, DNA barcoding loci provided a clear separation between geographically dissimilar gene pools by clustering them

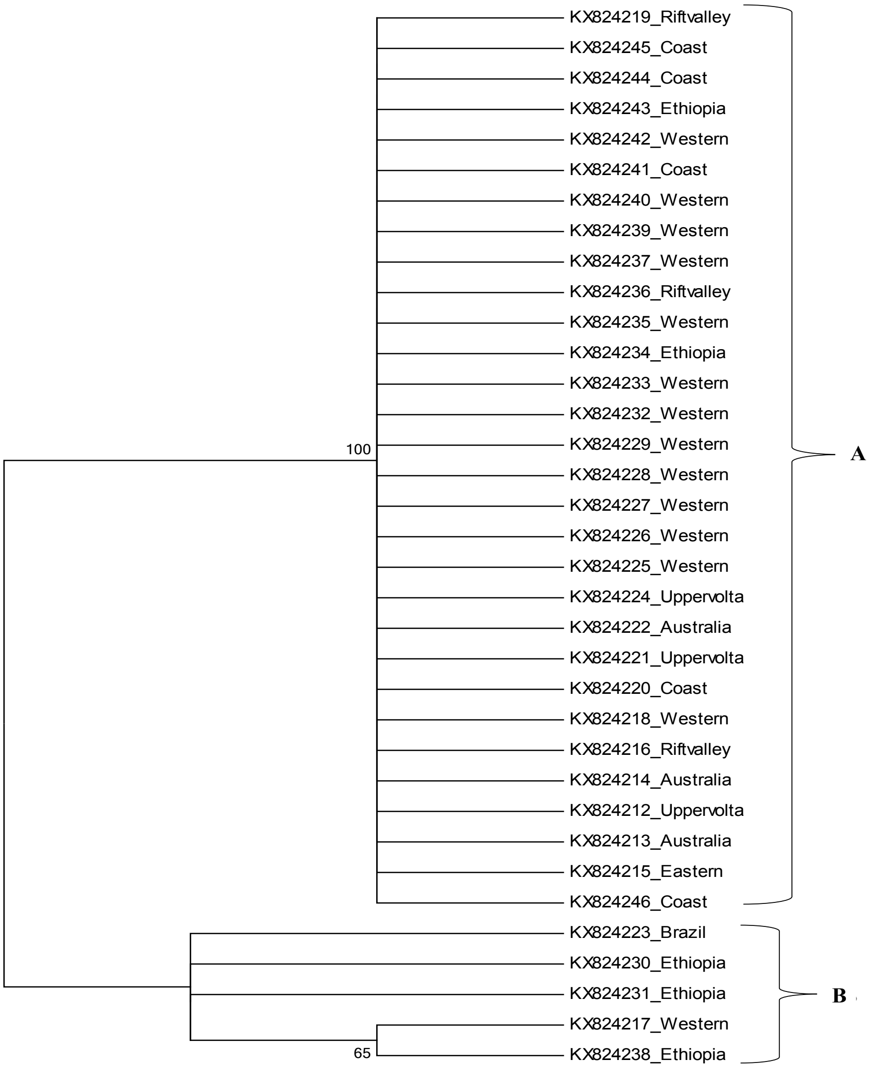


Figure 17. *rpoB* loci: Phenogram of NJ cluster analysis. The phylogenetic tree based on *rpoB* gene sequences, constructed using the neighbor-joining algorithm. Bootstrap values of more than 50 are shown at the internodes.

together consistent with several recent studies which successfully distinguished between these groups by using both chloroplast and nuclear SSR markers or genomic AFLP markers alone [63] [64] [65].

The analysis of DNA barcode sequences for cluster recognition was explored for recognizing putative cowpea sub variants based on operational taxonomic units (OTUs). This approach accelerates and improves taxonomic workflows. Taken together, the study further reveals the NJ for the reconstruction of phylogenetic trees from evolutionary distance data placing the accessions in distinct clades. This algorithm uses general data clustering techniques to sequence analysis using genetic distance as a clustering metric. The principle behind this method was to establish pairs of operational taxonomic units (OTUs) that minimize the total branch length at each clustering stage of OTUs. NJ was used to determine the branch lengths as well as the topology of parsimonious trees.

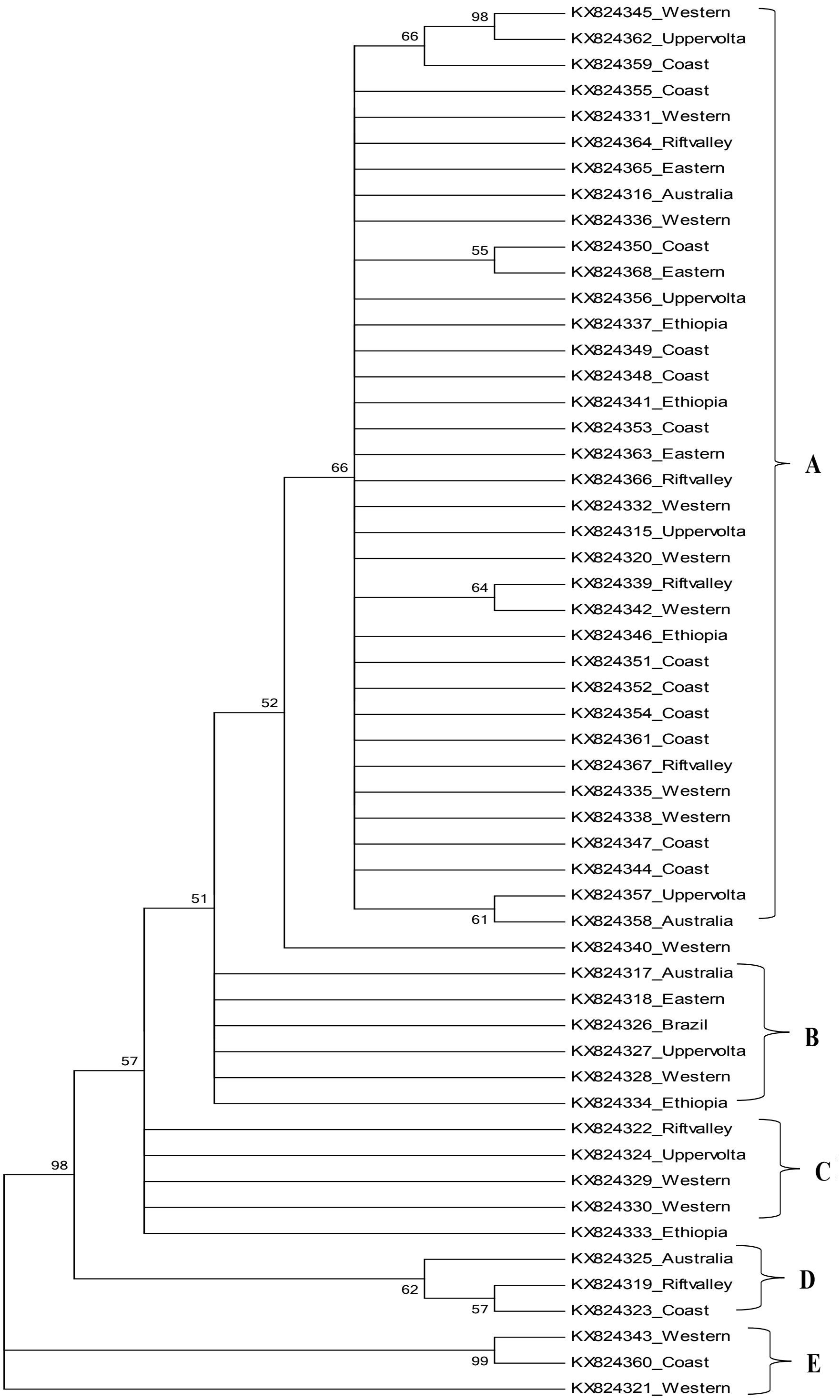


Figure 18. *trnH_psbA*: Phenogram of NJ cluster analyses. The phylogenetic tree based on *trnH_psbA*_gene sequences, constructed using the neighbor-joining algorithm. Bootstrap values of more than 50 are shown at the internodes.

The algorithm was used to infer the genetic distinctiveness of the subpopulations of cowpea based on *cp*-DNA markers and intergenic spacers. Preference of this algorithm was informed by its ability to convert polymorphic sequences into genetic distances based on nucleotide substitution models. Based on the coalescence of conspecific populations the NJ tree clusters closely related accessions into single but phylogenetically distinct clades.

4.3.3. Evolutionary Divergences

The evolutionary history was inferred using the Neighbor-Joining algorithm. The percentage of replicate trees in which the associated groups clustered together in the bootstrap test of 1000 replicates is shown above the branches. The evolutionary distances were computed using the Kimura 2-parameter method and are represented in the units of the number of base substitutions per site. The analysis involved a total of 54 nucleotide sequences depending on the marker. Codon positions included were 1st + 2nd + 3rd + non-coding. All ambiguous positions were removed for each sequence pair. Evolutionary analyses were conducted in MEGA 7.0. The phylogenetic tree topology generated using the NJ approach was consistent with the tree topology generated using the maximum likelihood method (ML). It was observed that there were a total of 5 clusters of the operational taxonomic units (OTUs) with bootstrap values greater than 50% supported at the internodes for *trnH_psbA*, apparently the most versatile and variable region. Phylogenetic trees indicate evolutionary relationships by means of bootstrap values.

However, it was regarded as critical to establish in numerical terms to what extent molecular markers differed in terms of their mean discriminatory power and nucleotide substitution between gene sequences. Hence, the gene sequences were subjected to mean divergence and "pairwise distance" estimation. The overall mean evolutionary divergence, computed using the distance menu of MEGA 6 was 0.19 for *matK* gene, 0.41 for *rbcL*, 0.24 for *rpoB*, and 0.64 for *atpF_atpH*. These findings are indicative of the superiority of concatenated trees in phylogenetic analysis. All the seven markers proved informative in discriminating between cowpea variants into distinct clusters albeit at varying degrees indicating the likelihood of *cp*-DNA and intergenic spacers of being used in delineating cowpea at varietal level. It would seem therefore that all the *cp*-DNA gene markers, which provided overall, mean evolutionary divergence with discriminatory power in delineating and clustering cowpea accessions into distinct clades. These findings are in agreement with Tajima's neutrality test findings reporting similar observations. Hence, the phylogeny based on the seven markers was complemented with phylogenetic analyses of DNA sequences of multiple *cp*-DNA genes and intergenic spacers in the multilocus sequence analyses. The MLSA approach is currently recommended for phylogenetic analysis of closely related species. The mutilocus sequence approach was tested in the current study to elucidate its ability in deciphering the phylogenetic relatedness among accessions. Overall, all the candidate loci were largely variable lead by *trnH_psbA* yielding five clades followed by *rpoc*1 gene yielding 4 clusters. The *matK* gene, *psbK_psbL* and *rbcL* gene revealed a total of 3 clusters each. However, *rpoB* was the most conservative gene loci revealing only 2 major clades. It would appear therefore that comparatively, *rpoB* would not be a good locus in phylogenetic analyses in the present study compared to *trnH_psbA*.

Acknowledgements

The author acknowledges support of the following who contributed in many ways to the success of this study: The National Gene Bank of Kenya, ASARECA, CIDA.

Conflicts of Interest

The authors declare no conflict of interest regarding the publication of this paper.

References

[1] Okoth, P., Muoma, J., Emmanuel, M., Clabe, W., Omayio, D.O. and Angienda, P.O. (2016) The Potential of DNA Barcode-Based Delineation Using Seven Putative Candidate Loci of the Plastid Region in Inferring Molecular Diversity of Cowpea at Sub-Species Level. http://www.scirp.org/journal/ajmb

[2] Chase, M.W., Cowan, R.S., Hollingsworth, P.M., van den Berg, C., Madrinan, S., Petersen, G. and Fay, M.F. (2009) Barcoding of Plants and Fungi. *Science*, **325**, 682-683. https://doi.org/10.1126/science.1176906

[3] Mohler, V. and Schwarz, G. (2008) Genotyping Tools in Plant Breeding: From Restriction Fragment Length Polymorphisms to Single Nucleotide Polymorphisms. *Molecular Marker Systems in Plant Breeding and Crop Improvement*, **55**, 23-38. https://doi.org/10.1007/3-540-26538-4_2

[4] Müller, K.F., Borsch, T. and Hilu, K.W. (2006) Phylogenetic Utility of Rapidly Evolving DNA at High Taxonomical Levels: Contrasting matK, trnT-F and rbcL in Basal Angiosperms. *Molecular Phylogenetics and Evolution*, **41**, 99-117. https://doi.org/10.1016/j.ympev.2006.06.017

[5] Hilu, K.W., Borsch, T., Müller, K., Soltis, D.E., Soltis, S.S., Savolainen, V., Chase, M.W., Powell, M.P., Alice, I.A., Evans, R., Sauquet, H., Neinhuis, C., Slotta, T.A.B., Jens, G.R., Campbell, C.S. and Chatrou, I.W. (2003) Angiosperm Phylogeny Based on matK Sequence Information. *American Journal of Botany*, **90**, 1758-1776. https://doi.org/10.3732/ajb.90.12.1758

[6] Johnson, L.A. and Soltis, D.E. (1994) *matK* DNA Sequences and Phylogenetic Reconstruction in Saxifragaceae s. str. *Systematic Botany*, **19**, 143-156. https://doi.org/10.2307/2419718

[7] Soltis, D.E. and Soltis, P.S. (2004) Amborella Not a “Basal Angiosperm”? Not so Fast. *American Journal of Botany*, **91**, 997-1001. https://doi.org/10.3732/ajb.91.6.997

[8] Shaw, J., Lickey, E.B., Schilling, E.E. and Small, R.L. (2007) Comparison of Whole Chloroplast Genome Sequences to Choose Noncoding Regions for Phylogenetic Studies in Angiosperms: The Tortoise and the Hare III. *American Journal of Botany*, **94**, 275-288. https://doi.org/10.3732/ajb.94.3.275

[9] 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. https://doi.org/10.1371/journal.pone.0000508

[10] Shaw, J. and Small, R.L. (2005) Chloroplast DNA Phylogeny and Phylogeography of

the North American Plums (*Prunus subgenus* Prunus Section Prunocerasus, Rosaceae). *American Journal of Botany*, **92**, 2011-2030. https://doi.org/10.3732/ajb.92.12.2011

[11] Chang (2006) The Chloroplast Genome of Phalaenopsis Aphrodite (Orchidaceae): Comparative Analysis of Evolutionary Rate with that of Grasses and Its Phylogenetic Implications. *Molecular Biology and Evolution*, **23**, 279-291. https://doi.org/10.1093/molbev/msj029

[12] Whitlock, B.A., Hale, A.M. and Groff, P.A. (2010) Intraspecific Inversions Pose a Challenge for the trnH-psbA Plant DNA Barcode. *PLoS ONE*, **5**, e11533. https://doi.org/10.1371/journal.pone.0011533

[13] 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. https://doi.org/10.1371/journal.pone.0000508

[14] 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., Pedersen, N., Hedderson, T.A.J., Conrad, F., Salazar, G.A., Richardson, J.E., Hollingsworth, M.L., Barraclough, T.G., Kelly, L. and Wilkinson, M. (2007) A Proposal for a Standardised Protocol to Barcode All Land Plants. *Taxon*, **56**, 295-299.

[15] Babbar, S.B., Raghuvanshi, S., Singh, H.K., Parveen, I. and Malik, S. (2012) An Overview of the DNA Barcoding of Plants. *Phytomorphology*, **62**, 69-99.

[16] Heinze, B. (2007) A Database of PCR Primers for the Chloroplast Genomes of Higher Plants. *Plant Methods*, **3**, 4-10. https://doi.org/10.1186/1746-4811-3-4

[17] Neuhaus, H. and Link, G. (1987) The Chloroplast tRNALys (UUU) Gene from Mustard (*Sinapsis alba*) Contains a Class II Intron Potentially Coding for a Maturase-Related Polypeptide. *Current Genetics*, **11**, 251-257. https://doi.org/10.1007/BF00355398

[18] Barthet, M.M. and Hilu, K.W. (2007) Expression of matK: Functional and Evolutionary Implications. *American Journal of Botany*, **94**, 1402-1412. https://doi.org/10.3732/ajb.94.8.1402

[19] Cameron, K.M. (2005) Leave It to the Leaves: A Molecular Phylogenetic Study of Malaxideae (Orchidaceae). *American Journal of Botany*, **92**, 1025-1032. https://doi.org/10.3732/ajb.92.6.1025

[20] Hickey, D.A. and Min, X.J. (2007) Assessing the Effect of Varying Sequence Length on DNA Barcoding of Fungi. *Molecular Ecology Notes*, **7**, 365-373. https://doi.org/10.1111/j.1471-8286.2007.01698.x

[21] 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. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 8369-8374. https://doi.org/10.1073/pnas.0503123102

[22] CBOL Plant Working Group (2009) A DNA Barcode for Land Plants. *PNAS*, **106**, 12794-12797. https://doi.org/10.1073/pnas.0905845106

[23] Nicolalde-Morejón, F., Vergara-Silva, F., González-Astorga, J., Stevenson, D.W. and Vovides, A.P. (2010) A Character-Based Approach in the Mexican Cycads Supports Diverse Multigene Combinations for DNA Barcoding. *Cladistics*, **26**, 1-15.

[24] Wang, W., Freeman, W.H., Wu, Y., Yan, Y., Ermakova, M., Kerstetter, R. and Messing, J. (2010) DNA Barcoding of the *Lemnaceae*, a Family of Aquatic Monocots. *BMC Plant Biology*, **10**, 205. https://doi.org/10.1186/1471-2229-10-205 http://www.biomedcentral.com/1471-2229/10/205

[25] McNeal, J.R., Kuehl, J.V., Boore, J.L. and de Pamphilis, C.W. (2007) Complete Plastid Genome Sequences Suggest Strong Selection for Retention of Photosynthetic Genes in the Parasitic Plant Genus *Cuscuta*. *BMC Plant Biology*, **7**, 57. https://doi.org/10.1186/1471-2229-7-57

[26] Lahaye, *et al*. (2008) DNA Barcoding the Floras of Biodiversity Hotspots. *Proceedings of the National Academy of Sciences*, **105**, 2923-2928. https://doi.org/10.1073/pnas.0709936105

[27] Guisinger, M.M., Kuehl, J.V., Boore, J.L. and Jansen, R.K. (2008) Genome-Wide Analyses of Geraniaceae Plastid DNA Reveal Unprecedented Patterns of Increased Nucleotide Substitutions. *Proceedings of the National Academy of Sciences*, **105**, 18424-18429. https://doi.org/10.1073/pnas.0806759105

[28] Adekambi, T., Drancourt, M. and Raoult, D. (2008) The *rpoB* Gene as a Tool for Clinical Microbiologists. *Trends in Microbiology*, **17**, 37-46. https://doi.org/10.1016/j.tim.2008.09.008

[29] Liu, Q., Triplett, J.K., Wen, J. and Peterson, M. (2011) Allotetraploid Origin and Divergence in Eleusine (Chloridoideae, Poaceae): Evidence from Low-Copy Nuclear Gene Phylogenies and a Plastid Gene Chronogram. *Annals of Botany*, **108**, 1287-1298. https://doi.org/10.1093/aob/mcr231

[30] Baldauf, S.L. (2003) Phylogeny for the Faint of Heart: A Tutorial. *Trends in Genetics*, **19**, 345-351. https://doi.org/10.1016/S0168-9525(03)00112-4

[31] Linder, C.R. and Warnow, T. (2005) An Overview of Phylogeny Reconstruction. In: Aluru, S., Ed., *the Handbook of Computational Molecular Biology*, Chapman and Hall/CRC, New York, 40.

[32] Saitou, N. and Nei, M. (1987) The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution*, **4**, 406-425.

[33] Zuckerkandl, E. and Pauling, L. (1965) Evolutionary Divergence and Convergence in Proteins. In: Bryson, V. and Vogel, H.J., Eds., *Evolving Genes and Proteins*, Academic Press, Cambridge, 97-166. https://doi.org/10.1016/B978-1-4832-2734-4.50017-6

[34] Edwards, S.V. (2009) Is a New and General Theory of Molecular Systematics Emerging? *Evolution*, **63**, 1-19. https://doi.org/10.1111/j.1558-5646.2008.00549.x

[35] Grenfell, B.T. (2004) Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science*, **303**, 327-332. https://doi.org/10.1126/science.1090727

[36] Salipante, S.J. and Horwitz, M.S. (2006) Phylogenetic Fate Mapping. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 5448-5453. https://doi.org/10.1073/pnas.0601265103

[37] Gregory, T.R. (2005) DNA Barcoding Does Not Compete with Taxonomy. *Nature*, **434**, 1067. https://doi.org/10.1038/4341067b

[38] Brady, A. and Salzberg, S. (2011) PhymmBL Expanded: Confidence Scores, Custom Databases, Parallelization and More. *Nature Methods*, **8**, 367. https://doi.org/10.1038/nmeth0511-367

[39] Pedersen, J.S. (2013) Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. *PLOS Computational Biology*, **2**, e33. https://doi.org/10.1371/journal.pcbi.0020033

[40] Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G. and Siepel, A. (2011) Bayesian Inference of Ancient Human Demography from Individual Genome Sequences. *Nature Genetics*, **43**, 1031-1034. https://doi.org/10.1038/ng.937

[41] Kingman, J.F.C. (1982) On the Genealogy of Large Populations. *Journal of Applied*

Probability, **19A**, 27-43. https://doi.org/10.2307/3213548

[42] Felsenstein, J. (2004) Inferring Phylogenies. Sinauer Associates, Sunderland.

[43] Yang, Z. (2006) Computational Molecular Evolution. Oxford Univ. Press, Oxford. https://doi.org/10.1093/acprof:oso/9780198567028.001.0001

[44] Kumar, L.S. (1999) DNA Markers in Plant Improvement: An Overview. *Biotechnology Advances*, **17**, 143-182. https://doi.org/10.1016/S0734-9750(98)00018-4

[45] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acid Research*, **22**, 4673-4680. https://doi.org/10.1093/nar/22.22.4673

[46] Tamura, K., Nei, M. and Kumar, S. (2004) Prospects for Inferring Very Large Phylogenies by Using the Neighbor-Joining Method. *Proceedings of the National Academy of Sciences*, **101**, 11030-11035. https://doi.org/10.1073/pnas.0404206101

[47] Felsenstein, J. (1985) Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, **39**, 783-791. https://doi.org/10.1111/j.1558-5646.1985.tb00420.x

[48] Jukes, T.H. and Cantor, C.R. (1969) Evolution of Protein Molecules. In: Munro, H.N., Ed., *Mammalian Protein Metabolism*, Academic Press, New York, 21-132. https://doi.org/10.1016/B978-1-4832-3211-9.50009-7

[49] Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Molecular Biology and Evolution*, **24**, 1596-1599.

[50] Tamura, K., Peterson, N., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011) MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance and Maximum Parsimony Methods. *Molecular Biology and Evolution*, **28**, 2731-2739.

[51] Hall, T.A. (1999) BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, **41**, 95-98.

[52] Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution*, **30**, 2725-2729. https://doi.org/10.1093/molbev/mst197

[53] Nei, M. and Kumar, S. (2000) Molecular Evolution and Phylogenetics. Oxford University Press, New York.

[54] Sass, C., Little, D.P., Stevenson, D.W. and Specht, C.D. (2007) DNA Barcoding in the Cycadales: Testing the Potential of Proposed Barcoding Markers for Species Identification of Cycads. *PLoS ONE*, **2**, e1154. https://doi.org/10.1371/journal.pone.0001154

[55] Guindon, S. and Gascuel, O. (2003) A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, **52**, 696-704. https://doi.org/10.1080/10635150390235520

[56] Newmaster, S.G., Fazekas, A.J. and Ragupathy, S. (2006) DNA Barcoding in Land Plants. Evaluation of rbcL in a Multigene Tiered Approach. *Canadian Journal of Botany*, **84**, 335-341. https://doi.org/10.1139/b06-047

[57] Wiemers, M. and Fiedler, K. (2007) Does the DNA Barcoding Gap Exist? A Case Study in Blue Butterflies (Lepidoptera: Lycanidae). *Frontiers in Zoology*, **4**, 8. https://doi.org/10.1186/1742-9994-4-8

[58] Meier, R., Shiyang, K., Vaidya, G. and Ng, P.K.L. (2006) DNA Barcoding and Tax-

onomy.

[59] Virgilio, M., Backeljau, T., Nevado, B. and De Meyer, M. (2010) Comparative Performances of DNA Barcoding across Insect Orders. *BMC Bioinformatics*, **1**, 209-235. https://doi.org/10.1186/1471-2105-11-206

[60] Collins, R.A., Boykin, L.M., Cruickshank, R.H. and Armstrong, K.F. (2012) Barcoding's Next Top Model: An Evaluation of Nucleotide Substitution Models for Specimen Identification. *Methods in Ecology and Evolution*, **3**, 457-465. https://doi.org/10.1111/j.2041-210X.2011.00176.x

[61] Kwak, M. and Gepts, P. (2009) Structure of Genetic Diversity in the Two Major Gene Pools of Common Bean (*Phaseolus vulgaris* L., Fabaceae). *Theoretical and Applied Genetics*, **118**, 979-992. https://doi.org/10.1007/s00122-008-0955-4

[62] Angioi, S.A., Desiderio, F., Rau, D., Bitocchi, E., Attene, G. and Papa, R. (2009) Development and Use of Chloroplast Microsatellites in Phaseolus spp. and Other Legumes. *Plant Biology*, **11**, 598-612. https://doi.org/10.1111/j.1438-8677.2008.00143.x

[63] Rossi, M., Bitocchi, E., Bellucci, E., Nanni, L., Rau, D., Attene, G. and Papa, R. (2009) Linkage Disequilibrium and Population Structure in Wild and Domesticated Populations of *Phaseolus vulgaris* L. *Evolutionary Applications*, **2**, 504-522. https://doi.org/10.1111/j.1752-4571.2009.00082.x

[64] Burle, M.L., Fonseca, J.R., Kami, J.A. and Gepts, P. (2010) Microsatellite Diversity and Genetic Structure among Common Bean (*Phaseolus vulgaris* L.) Landraces in Brazil, a Secondary Center of Diversity. *Theoretical and Applied Genetics*, **121**, 801-813. https://doi.org/10.1007/s00122-010-1350-5

[65] Hollingsworth, P.M., Forrest, L.L., Spouges, J.L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M.W., Cowan, R.S., Erickson, D.L., Fazekas, A.J., Graham, S.W., James, K.E., Kim, K.-J., Kress, W.J., Schneider, H., van AlphenStah, J., Barrett, S.C.H., van den Berg, C., Bogarin, D., Burgess, K.S., Cameron, K.M., Carine, M., Chacón, J., Clark, A., Clarkson, J.J., Conrad, F., Devey, D.S., Ford, C.S., Hedderson, T.A.J., Hollingsworth, M.L., *et al.* (2009) A DNA Barcode for Land Plants. *Proceedings of the National Academy of Sciences*, **106**, 12794-12797.