Association of \textit{nad7a} Gene with Cytoplasmic Male Sterility in Pigeonpea

Pallavi Sinha, K. B. Saxena, Rachit K. Saxena, Vikas K. Singh, V. Suryanarayana, C.V. Sameer Kumar, Mohan A.V.S. Katta, Aamir W. Khan, and Rajeev K. Varshney*

\textbf{Abstract}

Cytoplasmic male sterility (CMS) has been exploited in the commercial pigeonpea \textit{(Cajanus cajan (L.) Millsp.)} hybrid breeding system; however, the molecular mechanism behind this system is unknown. To understand the underlying molecular mechanism involved in \textit{A}$_4$ CMS system derived from \textit{C. cajanifolius} (Haines) Maesen, 34 mitochondrial genes were analyzed for expression profiling and structural variation analysis between CMS line (ICRI-SAT Pigeonpea A line, ICPA 2039) and its cognate maintainer (ICPB 2039). Expression profiling of 34 mitochondrial genes revealed nine genes with significant fold differential gene expression at \( P \leq 0.01 \), including one gene, \textit{nad4L}, with 1366-fold higher expression in CMS line as compared with the maintainer. Structural variation analysis of these mitochondrial genes identified length variation between ICPA 2039 and ICPB 2039 for \textit{nad7a} (subunit of \textit{nad7} gene). Sanger sequencing of \textit{nad4L} and \textit{nad7a} genes in the CMS and the maintainer lines identified two single nucleotide polymorphisms (SNPs) in upstream region of \textit{nad4L} and a deletion of 10 bp in \textit{nad7a} in the CMS line. Protein structure evaluation showed conformational changes in predicted protein structures for \textit{nad7a} between ICPA 2039 and ICPB 2039 lines. All above analyses indicate association of \textit{nad7a} gene with the CMS for \textit{A}$_4$ cytoplasm in pigeonpea. Additionally, one polymerase chain reaction (PCR) based Indel marker (\textit{nad7a}$_\text{del}$) has been developed and validated for testing genetic purity of \textit{A}$_4$ derived CMS lines to strengthen the commercial hybrid breeding program in pigeonpea.

\textbf{Cytoplasmic Male Sterility} is a unique, maternally inherited trait that has been identified in more than 150 flowering plant species. Formation of nonfunctional pollen occurs with CMS, though female fertility usually remains unaffected (see Schnable and Wise, 1998). Cytoplasmic male sterility is determined by recombination events in the mitochondrial genome that leads to formation of chimeric open reading frames (ORFs; Chase, 2007). These rearranged chimeric ORFs encode deleterious proteins resulting in reduced respiration and other associated mitochondrial defects that ultimately lead to pollen sterility (Igarashi et al., 2013; Luo et al., 2013). Effect of pollen sterility, however, can be suppressed by crossing CMS line with the line carrying nuclear fertility restorer (\textit{Rf}) genes (Wang et al., 2006; Eckardt, 2006; Tang et al., 2014). The \textit{Rf} genes encode pentatricopeptide repeat proteins that counteract the mitochondrial defects and reduce or remove the deleterious properties of the CMS-associated genes (Hu et al., 2012; Igarashi et al., 2013). Cytoplasmic male sterile lines, together with nuclear fertility restoration, are exploited for commercial production of hybrid seeds in various economically important crops such as rice \textit{(Oryza sativa L.); Cheng et al., 2004; Ellur et al., 2013}, wheat \textit{(Triticum aestivum L.; Longin...}
et al., 2013), barley (Hordeum vulgare L.; Mühleisen et al., 2013), maize (Zea mays L.; Crow, 1998; Duvick et al., 2004), sorghum [Sorghum bicolor (L.) Moench; Reddy et al., 2013], and pearl millet [Pennisetum glaucum (L.) R. Br.; Rai et al., 1999]. In the case of pigeonpea, significant efforts have been made to identify CMS lines using wide hybridization technology. For example, eight different cytoplasm systems (A1 to A8) have been identified (Saxena et al., 2010a; Saxena et al., 2013). Among these, the CMS line derived from Cajanus cajanifolius, designated as A4 cytoplasm has shown stability in variable environments and excellent performance in hybrid seed production (Saxena et al., 2005; Sawargaonkar, 2011). Using this system, the first CMS based pigeonpea hybrid (ICPH 2671) showing up to 47% higher yield over control varieties in multi-location trials has been released in India for commercial cultivation (Saxena et al., 2013). However, to further accelerate pigeonpea hybrid breeding programs in terms of yield and quality traits, understanding the molecular basis of CMS is important.

To understand the molecular basis of CMS in A4 cytoplasm, we sequenced mitochondrial genomes of four Cajanus genotypes: the CMS line ICPA 2039, its cognate maintainer line ICPB 2039, the hybrid line ICPH 2433, and the wild relative ICPW 29 (accession from Cajanus cajanifolius), source of A4 cytoplasm. Sequence annotation of mitochondrial genome of ICPA 2039 line predicted a total of 51 genes, including 34 protein coding, 14 tRNA, and 3 rRNA genes. Comparative analysis of four different Cajanus mitochondrial genomes revealed 31 chimeric ORFs, among which 13 were identified between CMS and maintainer lines (Tuteja et al., 2013). Chimeric ORFs found in the proximity of known mitochondrial protein coding genes were reported to be associated with the male sterility in a number of plant species (Schnable and Wise, 1998; Hanson and Bentolila, 2004; Igarashi et al., 2013). Considering the fact that chimeric ORFs are associated with CMS phenotype, the markers developed based on the structural differences between CMS and maintainer lines have been frequently utilized for genetic purity testing of CMS lines (Narayanan et al., 1996; Sane et al., 1997; Yashitola et al., 2004; Rajendrakumar et al., 2007; Saxena et al., 2010b). Mitochondrial DNA based markers can be used to differentiate CMS and maintainer lines at the early stage of plant growth. This can replace time consuming and labor-intensive traditional grow-out test that involves field trials, with accurate detection of contaminants in CMS.

In view of above, this study analyzes expression and structural variation of 34 protein coding mitochondrial genes between CMS (ICPA 2039) and its maintainer (ICPB 2039) lines to identify the CMS associated gene(s) for A4 cytoplasm derived from C. cajanifolius. Most importantly one mitochondrial gene based Indel marker that can differentiate the CMS line from its fertile counterpart has been developed in this study.

### Table 1. Plant materials used for validation gene based markers in cytoplasmic male sterile (CMS) and its maintainer lines.

<table>
<thead>
<tr>
<th>Serial No.</th>
<th>Source cytoplasm</th>
<th>CMS line</th>
<th>Maintainer line</th>
<th>Backcross generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cajanus cajanifolius</td>
<td>ICPA 2039 (A4)(^1)</td>
<td>ICPB 2039</td>
<td>BC(_{15})</td>
</tr>
<tr>
<td>2</td>
<td>Cajanus cajanifolius</td>
<td>ICPA 2199 (A4)</td>
<td>ICPB 2199</td>
<td>BC(_{1})</td>
</tr>
<tr>
<td>3</td>
<td>Cajanus cajanifolius</td>
<td>ICPA 2098 (A4)</td>
<td>ICPB 2098</td>
<td>BC(_{4})</td>
</tr>
<tr>
<td>4</td>
<td>Cajanus cajanifolius</td>
<td>ICPA 2092 (A4)</td>
<td>ICPB 2092</td>
<td>BC(_{6})</td>
</tr>
<tr>
<td>5</td>
<td>Cajanus cajanifolius</td>
<td>ICPA 2189 (A4)</td>
<td>ICPB 2189</td>
<td>BC(_{5})</td>
</tr>
<tr>
<td>6</td>
<td>Cajanus cajanifolius</td>
<td>ICPA 2166 (A4)</td>
<td>ICPB 2166</td>
<td>BC(_{5})</td>
</tr>
<tr>
<td>7</td>
<td>Cajanus cajanifolius</td>
<td>ICPA 2051 (A4)</td>
<td>ICPB 2051</td>
<td>BC(_{3})</td>
</tr>
<tr>
<td>8</td>
<td>Cajanus cajanifolius</td>
<td>ICPA 2048 (A4)</td>
<td>ICPB 2048</td>
<td>BC(_{3})</td>
</tr>
<tr>
<td>9</td>
<td>Cajanus cajanifolius</td>
<td>ICPA 2047 (A4)</td>
<td>ICPB 2047</td>
<td>BC(_{4})</td>
</tr>
<tr>
<td>10</td>
<td>Cajanus cajanifolius</td>
<td>ICPA 2043 (A4)</td>
<td>ICPB 2043</td>
<td>BC(_{10})</td>
</tr>
<tr>
<td>11</td>
<td>Cajanus sericus</td>
<td>ICPA 2061 (A1)</td>
<td>ICPB 2061</td>
<td>BC(_{1})</td>
</tr>
<tr>
<td>12</td>
<td>Cajanus scarabaeoides</td>
<td>ICPA 2052 (A2)</td>
<td>ICPB 2052</td>
<td>BC(_{5})</td>
</tr>
<tr>
<td>13</td>
<td>Cajanus lineatus</td>
<td>ICPA 2209 (A6)</td>
<td>ICPB 2209</td>
<td>BC(_{5})</td>
</tr>
<tr>
<td>14</td>
<td>Cajanus reticulatus</td>
<td>CMS reticulatus</td>
<td>Vyshali</td>
<td>BC(_{5})</td>
</tr>
</tbody>
</table>

\(^1\) Details in parentheses indicate the source of cytoplasm used to develop CMS lines.

### Materials and Methods

#### Plant Materials

A CMS line (ICPA 2039) and its cognate maintainer line (ICPB 2039) derived from C. cajanifolius (A4) were used for identification of mitochondrial gene variations through expression and structural variation analysis. A set of 10 plants each from ICPA 2039 and ICPB 2039 lines were used to confirm the structural variation associated with the gene(s) during the structural variation analysis. Nine different A4 derived CMS lines and their cognate maintainer lines were used for validation of the gene-based Indel marker (Table 1). To estimate the accuracy level of the gene-based Indel marker in detecting contamination of maintainer line, known levels of admixture of maintainer in CMS were formed through DNA admixture. To check the utility of this marker at the commercial scale, 100 seeds of ICPA 2043, female parent of world’s first CGMS based pigeonpea hybrid ICPH 2671 (Saxena et al., 2013) having A4 cytoplasm, were tested. The gene-based Indel marker was also amplified on CMS and maintainer lines of four different cytoplasm of pigeonpea, namely, A1 (C. sericus), A2 (C. scarabaeoides), A5 (C. lineatus), and A6 (C. reticulatus).

#### Ribonucleic Acid and Deoxyribonucleic Acid Extraction

Pool of 10 flower buds of 5 mm size from three randomly selected plants each from male sterile and male fertile plants as described in Dalvi et al. (2008) were harvested and used for RNA isolation. Further, the buds were fixed in Carnoy’s fluid to confirm the pollen fertility. After microscopic conformation of sterility or fertility, total RNA was extracted from the selected buds using TRIzol (Invitrogen, Grand Island, NY) and purified through RNeasy Plant Mini kit (Qiagen, GmbH, Hilden, Germany). Further, three
micograms of RNA was used to construct single strand cDNA using the SuperScript III RT enzyme (Invitrogen).

Genomic DNA was isolated from young leaves of single plant by following extraction method as described in Cuc et al. (2008). Quality and quantity of DNA was checked on 0.8% agarose gel and concentration was normalized to approximately 5ng/μL.

Primers were designed from the exonic regions which ranged from 1 to 5 exons per gene using the Primer3 software (Rozen and Skaletsky, 2000; Table S1). These primer pairs were used for both expressional and structural variation analysis. The criteria used for designing the primer pairs included annealing temperature range of 55 to 60°C with an average of 57°C, amplicon size 70 to 150 bp, primer length 20 ± 5 bp, and GC% 50 ± 5.

Quantitative Real-Time Polymerase Chain Reaction
Quantitative real-time PCR (qRT-PCR) was performed using ABI SYBR GREEN PCR reaction on an ABI Fast 7500 System (Applied Biosystems, Foster City, CA). To calculate mean relative expression levels, each reaction was performed in three biological and two technical replicates, along with a no-template control. For qRT-PCR analysis, 1 μL of cDNA was used as a template in a 10 μL PCR reaction mixture. Cycle initiated with preincubation at 50°C for 2 min, denaturation at 95°C for 10 min, followed by 40 cycles of denaturation at 95°C for 15 s and annealing and extension at 60°C for 1 min. The real-time PCR efficiencies for each reaction was >95%. Also, a melting curve analysis was performed to determine the specificity of the reaction. Actin was used as an internal control for normalization of expression level, and quantity of each mRNA was calculated from the threshold points located in the log-linear range. The data were analyzed using the 7500 Sequence Detection Software (Applied Biosystems, USA) with default baseline and threshold. Student’s t test was used for data analysis (at \( P \leq 0.01 \) and \( P \leq 0.05 \)) to declare statistically significant values. OrganellarGenomeDraw (OGDRAW, http://ogdraw.mpimp-golm.mpg.de/, verified 15 May 2015) was used to create high-quality visual representations of expressional differences.

DNA Amplification through Polymerase Chain Reaction
Polymerase chain reactions were performed in a 5 μL reaction volume [0.5 μL of 10× PCR buffer, 1.0 μL of 15 mM MgCl₂, 0.25 μL of 2 mM dNTPs, 0.50 μL of 2 PM/μL primer (MWG-Biotec AG, Bangalore, India), 0.1 U of Taq polymerase (Bioline, London, UK), and 1.0 μL (5 ng/μL) of template DNA in 96-well micro liter plate (ABGene, Rockford, IL) using thermal cycler GeneAmp PCR System 9700 (Applied Biosystems, Foster City, CA). The PCR program used to amplify the DNA fragments was: initial denaturation was for 5 min at 94°C, followed by 35 cycles of denaturation for 20 s at 94°C, annealing for 20 s at 60°C, and extension for 30 s at 72°C, and final extension for 7 min at 72°C. The amplified PCR products were separated on 3.5% Metaphor gels stained with ethidium bromide, visualized under UV light, and photographed on gel documentation system.

Sequence Analysis and Homology Search
Polymerase chain reaction amplicons showing expression difference and length variation were used for sequencing at both ends using corresponding primers by deploying Sanger sequencing methodology. Sequences obtained were merged and contigs were developed for each amplon using DNA baser software (http://www.dnabaser.com/, verified 15 May 2015). Contigs obtained from DNA baser software were subjected for homology search using BLASTn (http://blast.ncbi.nlm.nih.gov/Blast.cgi, verified 15 May 2015), and further multiple sequences were aligned by using Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/, verified 15 May 2015) software.

Protein Structure Analysis
To identify the conformational changes at protein level, protein structure evaluation of identified associated genes was performed between ICPA 2039 and ICPB 2039. Sequence data were subjected for ORF prediction using the getorf program of EMBOSS suite (Olson, 2002). One-dimensional property of the protein sequences were predicted using ProtParam (http://www.expasy.ch/tools/protparam.html, verified 15 May 2015). Secondary protein structures were predicted using Jpred3 (Cole et al., 2008). To predict protein disorders, DisEMBL (http://dis.embl.de, verified 15 May 2015) was used. Homology-based modeling was performed by submitting the sequences to automated homology prediction servers such as Phyre2 (Kelley and Sternberg, 2009), SWISS-MODEL (http://swissmodel.expasy.org, verified 15 May 2015), ModWeb (http://salilab.org/modweb, verified 15 May 2015), HHPredB (Söding et al., 2005), and M4T (Fernandez-Fuentes et al., 2007).

Results
Relative Expression Profiling of Mitochondrial Genes
Expression profiling of selected 34 protein coding genes revealed a range of expressional variation ranging from a minimum of 0.04-fold (nad1e and nad5d) to a maximum of 13.66-fold (nad4L) differences between the CMS (ICPA 2039) and the maintainer line (ICPB 2039). The differential gene expression values for each gene between the two lines are illustrated in Fig. 1. In summary, the genes of mitochondrial respiratory chain/electron transport chain of Complex I showed a wide range of differential expression between ICPA 2039 and ICPB 2039. The different exonic regions of the nad1 showed differential
expressions ranging from 0.04 to 2.22, for nad2 from 0.47 to 0.80. However, nad3 and nad4L possessing one exon each showed fold differences of 2.90 and 13.66, respectively. Similarly, the exonic region of nad4 ranged from 0.28 to 4.60, nad5 from 0.04 to 2.16, and nad7 from 0.21 to 1.43. The genes, nad6 and nad9 having one exon each, showed 1.22 and 0.26-fold differences, respectively.

sdh4 gene in the Complex II and cob gene in the Complex III showed expression differences of 0.56 and 0.95-fold, respectively, between the two lines. Further, in the Complex IV, which has two genes, cox1 and cox3 displayed expression differences of 0.44 and 6.21 between the two lines, respectively. The Complex V comprises of five genes namely atp1, atp4, atp6, atp8, and atp9. These
Table 2. Mitochondrial genes with significant fold expression difference between ICPA 2039 and ICPB 2039 lines of pigeonpea.

<table>
<thead>
<tr>
<th>Serial No.</th>
<th>Gene</th>
<th>Primer name</th>
<th>Fold difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>nad1</td>
<td>nad1a</td>
<td>0.20*</td>
</tr>
<tr>
<td>2</td>
<td>nad3</td>
<td>nad3</td>
<td>2.90*</td>
</tr>
<tr>
<td>3</td>
<td>nad4L</td>
<td>nad4a</td>
<td>13.66**</td>
</tr>
<tr>
<td>4</td>
<td>nad4</td>
<td>nad4a</td>
<td>4.60**</td>
</tr>
<tr>
<td>5</td>
<td>nad5</td>
<td>nad5a</td>
<td>0.22**</td>
</tr>
<tr>
<td>6</td>
<td>nad7</td>
<td>nad7e</td>
<td>0.21**</td>
</tr>
</tbody>
</table>

IV. Complex IV
5  cox3  cox3       6.21**

V. Complex V
8  atp6  atp6       2.20**

VI. Cytochrome c biogenesis
9  ccmFcb   ccmFcb   2.18*
10  ccmFn    ccmFn    3.71**

VII. Ribosomal protein
11  rps3  rps3b      0.11*
12  rps12  rps12     0.20*
13  rps14  rps14     2.18*
14  rps19  rps19a    0.36*  

VIII. Other protein coding
15  matR  matR       0.14**  

*P ≤ 0.05, unpaired student's t test.
**P ≤ 0.01, unpaired student's t test.

Structural Variations in Mitochondrial Genes
To further detect the possible structural variations in the mitochondrial genomes of ICPA 2039 and ICPB 2039, DNA polymorphism was analyzed for all 34 mitochondrial genes between A (ICPA 2039) and B (ICPB 2039) lines utilizing 55 exon based primers (Table S1). Only nad7a, fragment of nad7 gene, showed length variation between these lines (Fig. 2a). For the nad7a gene, the amplicons observed in ICPA 2039 and ICPB 2039 were of 150 and 160 bp sizes, respectively (Fig. 2b). Further, to confirm the identified structural variation, a set of 10 plants each for ICPA 2039 and ICPB 2039 were screened with primers for nad7a gene, which clearly differentiated the male sterile and fertile lines (Fig. 2c).

Based on expression profiling and length variation analysis between ICPA 2039 and ICPB 2039, the two genes, nad4L (13.66-fold gene expression in ICPA 2039 line as compared with ICPB 2039) and nad7a (10 bp deletion in ICPA 2039) were selected for Sanger sequencing from ICPA 2039 and ICPB 2039 lines. To amplify the full-length gene sequences, two primer pairs for nad4L and one primer pair for nad7a were designed. Polymerase chain reaction amplified products were sequenced, and the data generated were used for further analysis (Table 3). Comparative sequence analysis of amplicons for nad4L did not show any difference in the coding region of the gene between ICPA 2039 and ICPB 2039. However, two nucleotide substitutions were observed in the upstream region at 129 bp (A in ICPA 2039 to C in ICPB 2039) and 130 bp (C in ICPA 2039 to G in ICPB 2039) of the gene (Fig. 3a). Similarly, comparative sequence analysis of nad7a gene revealed 10 bp deletion in ICPA 2039 when compared with ICPB 2039 at 180 to 189 bp (Fig. 3b). Although, as per our earlier study (Tuteja et al., 2013), this deletion was present in the upstream of the coding region of nad7a gene, our annotation analysis of amplified sequence of nad7a gene in this study showed occurrence of the deletion in coding region.

Conformational Protein Changes for CMS Associated Genes
In addition to analyzing variation at DNA level, for understanding conformational protein changes, ORF prediction was performed for nad4L and nad7a gene sequences for ICPA 2039 and ICPB 2039 lines. While no change was observed in amino acids between ICPA 2039 and ICPB 2039 for nad4L gene, in the case of nad7a gene, 110 and 103 amino acids residues were predicted in ICPA 2039 and ICPB 2039 lines, respectively. In nad7a, because of the 10 bp deletion in ICPA 2039, the predicted ORF start site shifted 20 bases upstream of the deletion adding six amino acids. The seventh amino acid is formed from the triplet codon derived from the two residual bases upstream and one base downstream of the deletion. The deletion causes a frame-shift mutation thereby adding a seventh amino acid at N terminal region in ICPA 2039 resulting in a total of 110 amino acids. However in ICPB 2039 the ORF starts after the

Genes showed expression variation from 0.39 to 2.20 between ICPA 2039 and ICPB 2039. Similarly, five genes, ccmB, ccmC, ccmFCa, ccmFCb, and ccmFn, belonging to Cytochrome c biogenesis, showed differential gene expression of 0.88, 2.72, 0.61, 2.18, and 3.71 between these lines, respectively. In the ribosomal protein gene complex, which has 10 genes, the expression difference ranged from 0.11 for rps3 to 2.18 for rps14 between the lines. The other protein coding genes complex has two genes, matR and mttB, and for these genes, differential gene expression was observed as 0.14 and 0.72, respectively, between the A and B lines. Out of 34 protein coding genes, nine genes (subunits of nine genes: nad1e, nad4L, nad4a, nad4d, nad5a, nad5d, nad7e, cox3, atp6, ccmFn, and matR) showed significant fold difference at P ≤ 0.01 and 15 showed significant fold difference at P ≤ 0.05. However, out of nine genes showed significantly different at P ≤ 0.01 between the A and B lines, nad4L showed a very high level of significant fold difference between these lines (13.66-fold). The list of genes with significant fold difference between A and B lines are presented in Table 2.
same 10 bp region resulting in 103 amino acids (Fig. 3b). As a result, the molecular weight of NAD7 protein sequence was estimated as 12 kDa in ICPA 2039 as compared to 11.3 kDa in ICPB 2039. Additionally, changes in the number of negatively (Glu, Asp) and positively (Arg, Lys) charged residues were also observed in ICPA 2039 (8 negatively and 12 positively) and in ICPB 2039 (8 negatively and 11 positively). Further, the secondary protein structure analysis of ICPA 2039 and ICPB 2039 for nad7a showed significant differences in β sheet and helix with a predominant N-terminal loop regions (Supplemental Fig. S1). Preliminary analysis with predicted secondary protein structures of both the lines revealed significant disordered regions in NAD7A protein. Three different disordered loop regions in the predicted proteins were identified in ICPA 2039 (at 1 to 22, 42 to 74, and 83 to 110 amino acids) and ICPB 2039 (at 1 to 15, 35 to 67, and 76 to 103 amino acids). Tertiary protein structure of NAD7A from both the lines showed superimposed

![Figure 2](image_url)

Figure 2. Genomic organization of nad7 gene belonging to the respiratory chain Complex I. (a) Structure of the pigeonpea nad7 gene with lengths of exons (brown bars) and introns (lines) shown in base pairs. Chimeric genic regions associated with the nad7 region are presented in blue color bars. The position of introns and exons were determined by the published mitochondrial genome sequence. (b) Structural variation analysis of five subunits of nad7 showed presence of 10 bp length variation between ICPA 2039 and ICPB 2039 in the first subunit, nad7a. (c) Confirmation of identified 10 bp deletion in nad7a on a set of 10 plants each of ICPA 2039 and ICPB 2039.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Nucleotide position</th>
<th>SNPs/Indels</th>
<th>ICPA 2039</th>
<th>ICPB 2039</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>nad4L</td>
<td>129</td>
<td>A</td>
<td>C</td>
<td>Transversion</td>
<td></td>
</tr>
<tr>
<td></td>
<td>130</td>
<td>C</td>
<td>G</td>
<td>Transversion</td>
<td></td>
</tr>
<tr>
<td>nad7a</td>
<td>180–189</td>
<td>—</td>
<td>TTTTTAGTGT</td>
<td>Deletion</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Single nucleotide polymorphisms (SNPs) and Indels in nad4L and nad7a region of ICPA 2039 and ICPB 2039 by Sanger sequencing.
Figure 3. Sequence alignment of amplicons for nad4L and nad7a from cytoplasmic male sterile line ICPA 2039 and its cognate maintainer ICPB 2039. (a) Alignment of 360 bp nad4L amplified fragment that showed 13.66-fold differential expression between ICPA 2039 and ICPB 2039 provides two single nucleotide polymorphisms (SNPs) in the promoter region. (b) Sequence alignment of ICPA 2039 and ICPB 2039 for another gene nad7a of 619 bp shows a 10 bp deletion from position 180 to 189 in ICPA 2039 in comparison with its maintainer line.
structure onto each other with matching features and conformational differences between ICPA 2039 and ICPB 2039 that may account for the instability of the subunit (Fig. 4).

Gene Organization and Synteny of nad7 Gene
Based on Sanger sequencing analysis, the nad7 gene is comprised of five exons, which are ordered as nad7a (162 bp) along with unknown origin of sequence, nad7b (102 bp), nad7c (465 bp), nad7d (246 bp), and nad7e (327 bp), and separated by introns of 849, 1335, 1049, and 1808 bp, respectively (Fig. 2a). Further, to understand the synteny of nad7 gene of pigeonpea, mitochondrial genomes for select species such as rice, Arabidopsis [Arabidopsis thaliana (L.) Heynh.], lotus [Lotus japonicus (Regel) K. Larsen (= L. corniculatus var. japonicus Regel)], mung bean [Vigna radiata (L.) R. Wilczek], and soybean [Glycine max (L.) Merr.] were analyzed for homology search. Multiple sequence alignment showed that the nad7a subunit was highly conserved in the selected five plant species. For instance, the nad7a gene of pigeonpea showed 100% homology with the gene sequence of mung bean, soybean, and lotus. With the Arabidopsis sequence, 96.5% homology was observed, whereas least (48%) homology was observed with rice (Supplemental Fig. S2a). Subsequently, the upstream region of nad7a where the 10 bp deletion was observed in ICPA 2039, was searched for homology with other crops in NCBI database using BLASTN search. As a result, the sequence with 93% query coverage showed 88% nucleotide identity with orf124 of Beta vulgaris subsp. maritima (L.) Arcang. genotype male-sterile E mitochondrion (accession No. FQ014226.1; Supplemental Fig. S2b).

Table 4. Details for deletion based genic marker for seed purity testing analysis.

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Nature</th>
<th>Nucleotide sequence (5′–3′)</th>
<th>ICPA 2039</th>
<th>ICPB 2039</th>
</tr>
</thead>
<tbody>
<tr>
<td>nad7a_del</td>
<td>Codominant</td>
<td>F CTGACGGATGCTACTAT R CATCCTTGCAGCTTAG</td>
<td>150</td>
<td>160</td>
</tr>
</tbody>
</table>

Development and Validation of nad7a Gene-Based Marker
The marker analysis of nad7a between CMS and maintainer lines clearly differentiated the male sterile line (ICPA 2039) from the fertile line (ICPB 2039) and the marker was designated as nad7a_del (Table 4). Screening of a set of nine different A4 derived CMS and their maintainer lines along with ICPA 2039 and ICPB 2039 confirmed the marker ability to differentiate sterile and fertile lines (Fig. 5). The marker, nad7a_del was also used to detect the contaminants in ICPA 2039 with ICPB 2039 in different known admixtures (2, 5, 10, and 15%). The marker was able to detect as low as 2% admixtures level of ICPB 2039 specific fragment in DNA of ICPA 2039. The intensity of contaminant specific allele was more prominent and clear with increasing admixture level from 5 to 15% (Fig. 6). Further, to check the utility of the marker, DNA analysis of 100 individual seeds of ICPA 2043 revealed 97.7% purity (Supplemental Fig. S3).

Screening of the nad7a_del marker on the other four available pigeonpea cytoplasms also demonstrated its utility and specificity for A4 cytoplasm. All four different CMS lines, ICPA 2061 (A1), ICPA 2052 (A2), ICPA 2206 (A3) and CMS reticulatus (A4) and their corresponding maintainer lines (ICPB 2061, ICPB 2052, ICPB 2206 and Vyshali) did not show any structural variation with the marker nad7a_del. Due to amplification of a similar amplicon (150 bp) in both, CMS and its maintainer line, the marker could not differentiate the CMS line from its cognate maintainers for other four cytoplasms. Therefore, it was concluded from our study that the marker nad7a_del is specific for A4 derived CMS lines and can be utilized for purity testing of A4 derived CMS lines from their maintainers (Supplemental Fig. S4).

Discussion
Expression Variation of Mitochondrial Genes
To understand the molecular basis of A4 based CMS in pigeonpea, the CMS line, ICPA 2039, along with its maintainer, ICPB 2039, were compared for 34 protein coding genes, identified earlier through pigeonpea mitochondrial genome sequencing (Tuteja et al., 2013). Four genes, Cox3, nad7, mttB, and ccmFc, showed chimeric gene sequence between A and B lines during our previous studies (Tuteja et al., 2013). All except mttB showed significant fold differences between the two lines in expression profiling in the present study. Most of the
genes belonging to the Complex I of mitochondrial respiratory electron transport system (ETS) showed significant fold expression differences between ICPA 2039 and ICPB 2039 lines. The Complex I of respiratory ETS is one of the major pathways for the entry of electrons into cyanide-sensitive electron transfer chain (Marienfeld and Newton, 1994). Mutations in the mitochondrial DNA of this complex were found associated with CMS in many other plant species like Arabidopsis (Brandt et al., 1992), tobacco (*Nicotiana tabacum* L.; Gutierres et al., 1997), maize (Marienfeld and Newton, 1994), and wheat (Bonen et al., 1990). Interestingly, our study also showed a high level of expression difference between the CMS and its maintainer line in one of the component genes (*nad4L*) of Complex I. Furthermore, sequence analysis of *nad4L* in ICPA 2039 revealed two single nucleotide substitutions in comparison with maintainer lines in upstream regions of the gene. Altered transcription levels of mitochondrial genes was found to be associated with CMS in some plant species (Dieterich et al., 2003; Hanson and Bentolila, 2004; Wang et al., 2006). For instance, a comparative study of *atp9* gene between CMS and maintainer lines identified RNA editing with two nucleotide sites (C to U) in the conserved region of the gene associated with the male sterility (Jiang et al., 2011). However, the two nucleotide substitutions observed in ICPA 2039 at the upstream region of *nad4L* did not show any conformational changes at the protein structure level, which needs further functional confirmation.

**Structural Variation in Mitochondrial Genes**

The structural variation analysis of 34 mitochondrial protein coding genes at length level showed 10 bp deletion occurred at the upstream region of *nad7a* in ICPA 2039.
2039. The CMS-associated genes are often chimeric and are derived due to rearrangement in the mitochondrial genome from the time of evolution. These chimeric ORFs often include novel sequences of unknown origin, which lead to male sterility through altering the biological function of known candidate genes (see Schnable and Wise, 1998; and Chen and Liu, 2014). For instance, absence of the last two exons of nad7 in Nicotiana sylvestris Speg. & Comes CMS mutant, though not lethal to plant cells, was associated with the abnormal CMS phenotype due to deficiency of Complex I (Pla et al., 1995; Sabar et al., 2000). Classical studies on the presence of abnormality in nad7 gene in many crops revealed that it leads to respiratory deficiencies and causes abnormal respiration (Nehls et al., 1992; Connett and Hanson, 1990; Pla et al., 1995; Akçay et al., 2012). Recently, DNA methylation of sterile and fertile phenotypes of photoperiod-thermosensitive male sterile line of rice (PA64S) revealed nad7 as one of the candidate genes responsible for male sterile phenotype (Chen et al., 2014). However, to confirm the association of the above mentioned two genes (nad7 and nad4L) with CMS, the expression of the genes need to be further tested in the presence of fertility restorer genes in F1 plants.

**Implication of Structural Variation on NAD7A Protein**

To understand the conformational changes in protein structure of male sterile and fertile lines, prediction of longest ORFs between ICPA 2039 and ICPB 2039 was the initial and challenging step due to the presence of 10 bp deletion in the ICPA 2039 lines. This amino acids length variation between the two lines causes significant differences in protein molecular weights. Protein modeling of the nad7a gene between ICPA 2039 and ICPB 2039 lines confirmed the presence of high levels of disorder between these lines due to presence of predominant random coil (loop) regions. Protein modeling with various models revealed that a SWISS-MODEL-derived protein data bank was the most accurate structural alignment (root-mean square deviation near zero) with Chain 4 of the 4HEA crystal structure. The presence of 10 bp length variation was associated with the conformational changes in the secondary and tertiary protein structure of nad7 subunit (NAD7A) between ICPA 2039 and ICPB 2039. In earlier studies, aberration in nad7 gene or its associated subunits was not lethal for plants, but was associated with severe developmental defects including CMS due to reduced respiration rate (Sabar et al., 2000). Therefore, the conformational differences for NAD7A between the male sterile and fertile line might be associated with CMS in A4 cytoplasm of pigeonpea, which needs further functional confirmation. The 10 bp deletion observed in ICPA 2039 that caused changes in the ORFs in the CMS line might have affected the stability and functions of protein. It was evident from the multiple sequence alignment results that the nad7a region is highly conserved in legumes in comparison with other model crops (rice and Arabidopsis).

The upstream region having deletion showed 88% sequence similarity to the chimeric gene orf124 known candidate gene for CMS in Beta vulgaris (L.).

**Application of CMS Based Marker for Purity Testing**

To maintain the genetic purity of CMS lines, we have developed the mitochondrial gene based marker, nad7a_del, that can differentiate ICPA 2039 with its maintainer ICPB 2039. Further validation of the marker on different set of nine A4 cytoplasm derived CMS lines, along with their maintainers, clearly demonstrated that the marker can be successfully applied to distinguish and identify the off-type plants from A4 CMS seed lot. Contrary to the traditional, time consuming grow-out test for CMS seed purity testing, mitochondrial gene based markers provide breeders a quick and accurate tool for screening the admixtures (Yashitola et al., 2004; Ngangkham et al., 2010; Suzuki et al., 2013). Additionally, the efficiency of nad7a_del for testing the purity of ICPA 2039 up to 2% level of contamination demonstrated their immediate use in marker based seed purity of parental line. Earlier, marker based hybrid purity tests were also developed for genetic purity analysis of A4 based F1 hybrids (ICPH 2671 and ICPH 2438) in pigeonpea (Saxena et al., 2010b; Bohra et al., 2011). Our findings will further supplement the pigeonpea hybrid breeding program by selecting pure CMS seed lot for successful production of hybrid seeds.

To assist the hybrid pigeonpea breeding process, the marker developed for A4 cytoplasm (derived from C. cajanifolius) was also checked across other four promising cytoplasms. The finding revealed that the marker was not able to differentiate the CMS from their cognate maintainers. This clearly showed that the deletion in nad7 associated region was specific for A4 cytoplasm. Therefore, in the near future, such types of analyses will be required for other pigeonpea cytoplasms to understand the molecular mechanism involved in male sterility and to develop their associated markers for CMS seed purity.

**Conclusions**

Based on variation analysis of 34 proteins coding genes, two putative genes, nad4L and nad7a, were found associated with CMS in ICPA 2039. Further protein structure analysis predicted conformation changes in NAD7A between the CMS and maintainer lines due to 10 bp deletion in ICPA 2039, which needs further functional validation. To confirm the association of nad7a with A4 CMS, expression of the gene needs to be verified in the presence of a fertility restorer gene, in hybrids derived from male sterile and fertility restorer lines. The nad7a_del marker developed in this study is very effective to differentiate A4-derived CMS lines with their cognate maintainers. In summary, this study will assist with deciphering the molecular mechanism involved in male sterility for A4 cytoplasm, as well as for enhancing hybrid breeding in pigeonpea.
Acknowledgments

Authors thank United States Agency for International Development (USAID) for financial support for the research work. This work has been undertaken as part of the CGIAR Research Program on Grain Legumes. ICRISAT is a member of CGIAR Consortium.

References


