BMC Plant Biology

RESEARCH ARTICLE Open Access

The first set of EST resource for gene discovery and marker development in pigeonpea (*Cajanus cajan* L.)

Nikku L Raju[1], Belaghihalli N Gnanesh[1,2], Pazhamala Lekha[1], Balaji Jayashree[1], Suresh Pande[1], Pavana J Hiremath[1], Munishamappa Byregowda[2], Nagendra K Singh[3], Rajeev K Varshney[1,4*]

Abstract

Background: Pigeonpea (*Cajanus cajan* (L.) Millsp) is one of the major grain legume crops of the tropics and subtropics, but biotic stresses [*Fusarium* wilt (FW), sterility mosaic disease (SMD), etc.] are serious challenges for sustainable crop production. Modern genomic tools such as molecular markers and candidate genes associated with resistance to these stresses offer the possibility of facilitating pigeonpea breeding for improving biotic stress resistance. Availability of limited genomic resources, however, is a serious bottleneck to undertake molecular breeding in pigeonpea to develop superior genotypes with enhanced resistance to above mentioned biotic stresses. With an objective of enhancing genomic resources in pigeonpea, this study reports generation and analysis of comprehensive resource of FW- and SMD- responsive expressed sequence tags (ESTs).

Results: A total of 16 cDNA libraries were constructed from four pigeonpea genotypes that are resistant and susceptible to FW ('ICPL 20102' and 'ICP 2376') and SMD ('ICP 7035' and 'TTB 7') and a total of 9,888 (9,468 high quality) ESTs were generated and deposited in dbEST of GenBank under accession numbers GR463974 to GR473857 and GR958228 to GR958231. Clustering and assembly analyses of these ESTs resulted into 4,557 unique sequences (unigenes) including 697 contigs and 3,860 singletons. BLASTN analysis of 4,557 unigenes showed a significant identity with ESTs of different legumes (23.2-60.3%), rice (28.3%), *Arabidopsis* (33.7%) and poplar (35.4%). As expected, pigeonpea ESTs are more closely related to soybean (60.3%) and cowpea ESTs (43.6%) than other plant ESTs. Similarly, BLASTX similarity results showed that only 1,603 (35.1%) out of 4,557 total unigenes correspond to known proteins in the UniProt database (≤ 1E-08). Functional categorization of the annotated unigenes sequences showed that 153 (3.3%) genes were involved in cellular component category, 132 (2.8%) in biological process, and 132 (2.8%) in molecular function. Further, nineteen genes were identified differentially expressed between FW- responsive genotypes and 20 between SMD- responsive genotypes. Generated ESTs were compiled together with 908 ESTs available in public domain, at the time of analysis, and a set of 5,085 unigenes were defined that were used for identification of molecular markers in pigeonpea. For instance, 3,583 simple sequence repeat (SSR) motifs were identified in 1,365 unigenes and 383 primer pairs were designed. Assessment of a set of 84 primer pairs on 40 elite pigeonpea lines showed polymorphism with 15 (28.8%) markers with an average of four alleles per marker and an average polymorphic information content (PIC) value of 0.40. Similarly, *in silico* mining of 133 contigs with ≥ 5 sequences detected 102 single nucleotide polymorphisms (SNPs) in 37 contigs. As an example, a set of 10 contigs were used for confirming *in silico* predicted SNPs in a set of four genotypes using wet lab experiments. While occurrence of SNPs were confirmed for all the 6 contigs for which scorable and sequenceable amplicons were generated. PCR amplicons were not obtained in case of 4 contigs. Recognition sites for restriction enzymes were identified for 102 SNPs in 37 contigs that indicates possibility of assaying SNPs in 37 genes using cleaved amplified polymorphic sequences (CAPS) assay.

* Correspondence: r.k.varshney@cgiar.org
[1]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Greater Hyderabad 502 324, Andhra Pradesh, India

Conclusion: The pigeonpea EST dataset generated here provides a transcriptomic resource for gene discovery and development of functional markers associated with biotic stress resistance. Sequence analyses of this dataset have showed conservation of a considerable number of pigeonpea transcripts across legume and model plant species analysed as well as some putative pigeonpea specific genes. Validation of identified biotic stress responsive genes should provide candidate genes for allele mining as well as candidate markers for molecular breeding.

Background

Pigeonpea (*Cajanus cajan* (L.) Millsp) is one of the major grain legume crops of the tropical and subtropical regions of the world [1]. It is the only cultivated food crop of the *Cajaninae* sub-tribe and has a diploid genome with 11 pairs of chromosomes (2n = 2× = 22) and a genome size estimated to be 858 Mbp [2]. The genus *Cajanus* comprises 32 species most of which are found in India, Australia and one is native to West Africa. Pigeonpea is a major food legume crop in South Asia and East Africa with India as the largest producer (3.5 Mha) followed by Myanmar (0.54 Mha) and Kenya (0.20 Mha) [3]. It plays an important role in food security, balanced diet and alleviation of poverty because of its diverse usages as a food; fodder and fuel wood [4]. Several abiotic (e.g. drought, salinity and water-logging) and biotic (e.g. diseases like *Fusarium* wilt, sterility mosaic and pod borer insects) stresses, are serious challenges for sustainable pigeonpea production to meet the demands of the resource poor people of several African and Asian countries.

Fusarium wilt (FW) caused by *Fusarium udum* is an important biotic constraint in pigeonpea production in the Indian subcontinent, which results in 16-47% crop losses [5]. The fungus enters the host vascular system at the root tips through wounds or invasion made by nematodes, leading to progressive chlorosis of leaves, branches, wilting and collapse of the root system [6]. In India alone, the loss due to this disease is estimated to be US $71 million and the percentage of disease incidence varies from 5.3 to 22.6% [7].

Sterility mosaic disease (SMD) caused by pigeonpea sterility mosaic virus (PPSMV) is one of the wide spread diseases of pigeonpea, which is transmitted by an eriophyid mite (*Aceria cajani* Channabasavanna). The disease is characterized by the symptoms like bushy and pale green appearance of plants followed by reduction in size, increase in number of secondary and mosaic mottling of leaves and finally partial or complete cessation of reproductive structures. Some parts of the plant may show disease symptoms and other parts may remain unaffected [8].

Due to the above mentioned factors combined with limited water resources to the fields in the semi-arid tropic regions, where the crop is grown, the productivity has remained stagnant at around 0.7 t/ha during the past two decades [1]. With the advent of genomic tools such as molecular markers, genetic maps, etc., conventional plant breeding has been facilitated greatly and improved genotypes/varieties with enhanced resistance/tolerance to biotic/abiotic stresses have been developed in several crop species [9,10]. In case of pigeonpea, however, a very limited number of genomic tools are available so far [11,12]. For instance, 156 microsatellite or simple sequence repeat (SSR) markers [13-16], 908 expressed sequence tags (ESTs), at the time of undertaking the study, were available in pigeonpea. For enhancing the genomic resources in pigeonpea, transcriptome sequencing to generate ESTs should be a fast approach. ESTs, which are generated by large-scale single pass sequencing of randomly picked cDNA clones, have been cost - effective and valuable resource for efficient and rapid identification of novel genes and development of molecular markers [17]. Further, ESTs have been employed in bioinformatic analyses to identify the genes that are differentially expressed in various tissues, cell types, or developmental stages of the same or different genotypes [18,19].

In view of above facts, this study was undertaken to obtain a comprehensive resource of FW- and SMD-responsive ESTs in pigeonpea with the following objectives: (i) generation of FW- and SMD- responsive ESTs, (ii) functional annotation of assembled unigenes, (iii) *in silico* identification of putative FW- and SMD- responsive genes, and (iv) development of novel SSR and SNP markers in pigeonpea.

Results

Root tissue is the site for *Fusarium udum* infection, the causal fungal agent of *Fusarium* wilt in pigeonpea. With an objective to evaluate the transcriptional responses after infection of roots by *F. udum*, six unidirectional cDNA libraries were constructed. These are from each of FW- infected root tissues of resistant ('ICPL 20102') and susceptible ('ICP 2376') genotypes at different stages *viz.* 6, 10, 15, 20, 25, 30 days after inoculation (DAI). Infected roots were examined by light microscopy upon harvest at different stages. The severity of wilt disease in both susceptible and resistant genotype was observed in longitudinal sections of stem and root vascular region at 15 and 30 DAI (Figure 1). Likewise for SMD, leaf tissue is the specific site of infection and therefore leaf samples

Figure 1 *Fusarium* wilt (FW) challenged pigeonpea seedlings at 30 days after inoculation (DAI). a) *Fusarium* wilt challenged pigeonpea genotypes ('ICPL 20102') and ('ICP 2376') at 30 days after inoculation (30 DAI); b & c) Microscopic examination of FW-resistant pigeonpea genotype ('ICPL 20102') showing no disease symptoms on shoot and root vascular tissues; d & e) Microscopic examination of FW-susceptible pigeonpea genotype ('ICP 2376') showing severe wilt symptoms on shoot and root vascular tissues.

of SMD infected genotypes, 'ICP 7035' (SMD resistant) and 'TTB 7' (SMD susceptible) were harvested at 45 and 60 days after sowing (DAS). RNA was extracted and consequently unidirectional cDNA libraries were constructed (see Additional file 1).

Generation of FW- and SMD- responsive ESTs

A total of 16 unidirectional cDNA libraries were constructed from all the four genotypes i.e. 'ICPL 20102' and 'ICP 2376'; 'ICP 7035' and 'TTB 7' which represent parents of mapping population segregating for FW and SMD, respectively. Using Sanger sequencing approach 3,168 ESTs were generated from root cDNA libraries of 'ICPL 20102' and 2,880 from 'ICP 2376'. Similarly, 1,920 ESTs were generated from each leaf cDNA libraries of SMD- responsive genotypes, 'ICP 7035' and 'TTB 7'. Details of EST generation from different cDNA libraries are given in Figure 2. In brief, a total of 9,888 ESTs were generated and after stringent screening for shorter (<100 bp) and poorer quality sequences, 9,468 high quality ESTs were obtained, with an average varied-read length of 514 bp (Figure 2). All EST sequences were deposited in the dbEST of GenBank under accession numbers GR463974 to GR473857 and GR958228 to GR958231.

Pigeonpea EST assembly

With an objective to minimize redundancy, clustering and assembly was done for different EST datasets to define unigenes for (a) FW-responsive ESTs, (b) SMD-responsive ESTs, (c) FW- and SMD-responsive ESTs, and (d) the entire set of pigeonpea ESTs including those from the public domain. These unigene (UG) sets were referred to as UG-I, UG-II, UG-III and UG-IV, respectively. The UG-I comprised of 3,316 unigenes with 389 contigs and 2,927 singletons by clustering of 5,680 high quality ESTs. Similarly, for UG-II, clustering of 3,788 high quality sequences resulted in 1,308 unigenes (328 contigs and 980 singletons). Based on clustering of all the 9,468 high quality sequences generated in this study, the UG-III was defined with 4,557 unigenes (697 contigs and 3,860 singletons). The cluster analysis of 908 ESTs available in the public domain along with 9,468 pigeonpea ESTs resulted in UG-IV that included 5,085 unigenes with 871 contigs and 4,214 singletons. The number of ESTs in a contig ranged from 2 to 573, with an average of 7 ESTs per contig. As expected, contigs with two EST members exhibited a higher percentage (46.7%) than contigs with three or more EST members (Figure 3).

Comparison of pigeonpea unigenes with other plant EST databases

All the four sets of unigenes i.e. UG-I, UG-II, UG-III and UG-IV were analyzed for BLASTN similarity search against available EST datasets of legume species namely chickpea (*Cicer arietinum*), pigeonpea (*Cajanus cajan*), soybean (*Glycine max*), *Medicago* (*Medicago truncatula*), *Lotus* (*Lotus japonicus*), common bean (*Phaseolus vulgaris*) and three model plant species

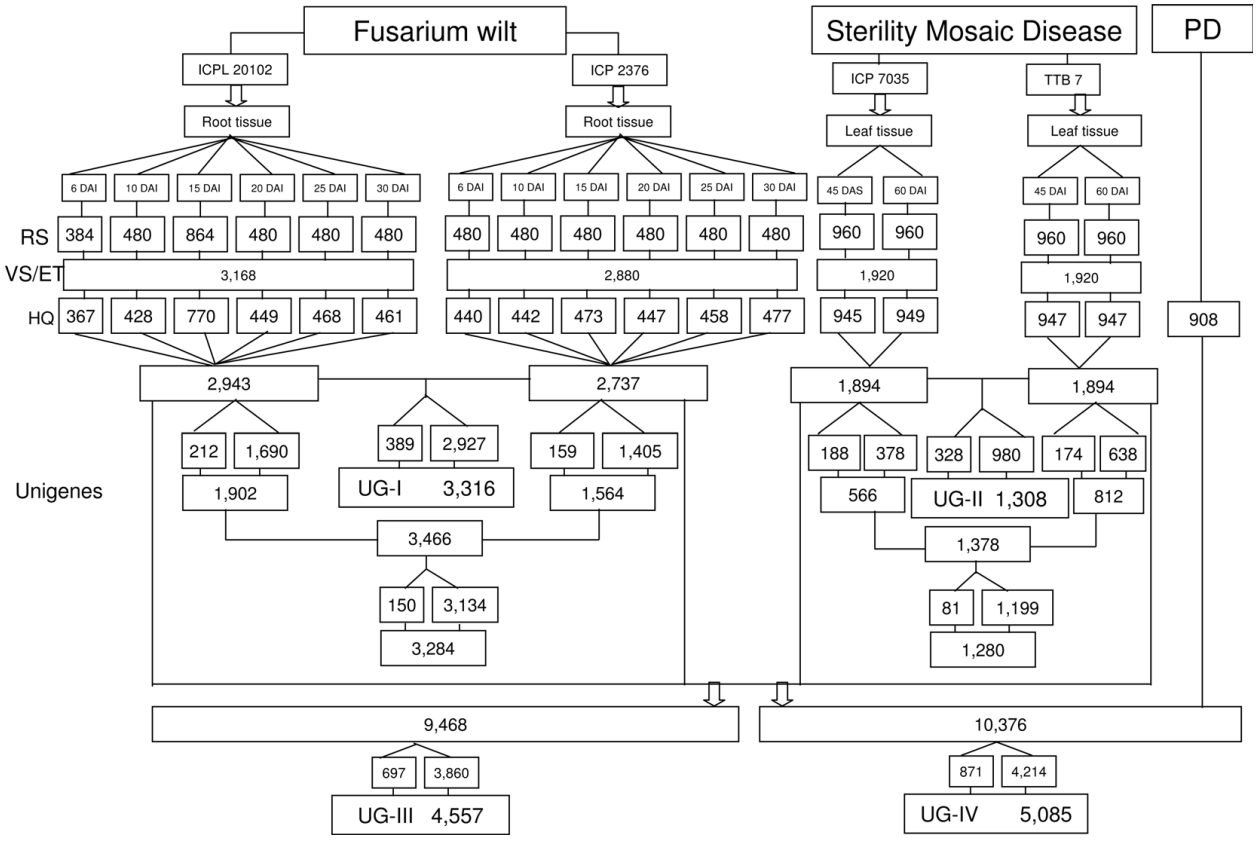

Figure 2 Summary of total ESTs generated from FW- and SMD- responsive pigeonpea genotypes. Generation and analysis of ESTs from 16 cDNA libraries of pigeonpea subjected to *Fusarium* wilt (FW) and Sterility mosaic disease (SMD) stresses; (A) Clustering and assembly of 2,943 and 2,737 HQS (High quality sequences) derived from FW-responsive cDNA libraries of pigeonpea genotypes 'ICPL 20102' and 'ICP 2376', respectively resulted in 3,316 unigenes (**UG-1)**; (B) Clustering and assembly of 1,894 HQS from each SMD-responsive pigeonpea genotypes 'ICP 7035' and 'TTB 7' resulted in 1,308 unigenes (**UG-II)**; (C) 9,468 HQS generated from all the four genotypes in the study as shown in (A) and (B) were analyzed together that provided a set of 4,557 unigenes (**UG-III)**; (D) Clustering and assembly of generated ESTs in this study along with 908 public domain pigeonpea ESTs, which resulted in 5,085 unigenes (**UG-IV)**, RS: Raw sequences; VS/ET: Vector trimmed/EST trimmed sequences; HQ: High quality sequences; PD: Public domain pigeonpea sequences from NCBI.

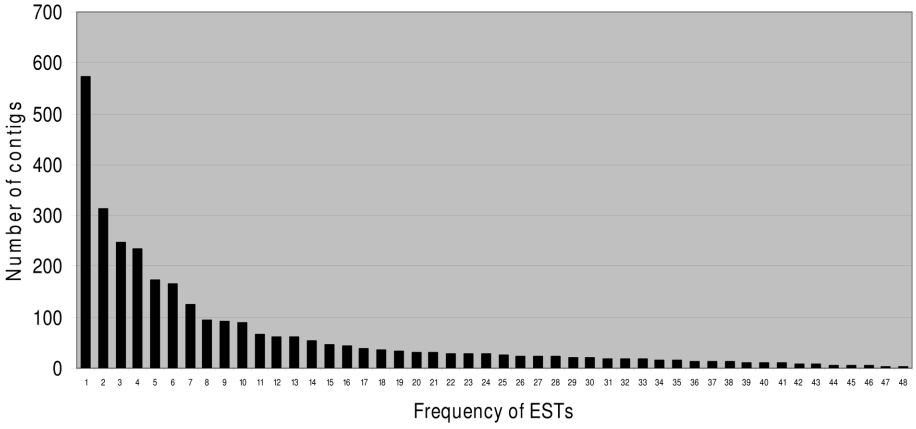

Figure 3 Frequency and distribution of pigeonpea ESTs among assembled contigs.

namely *Arabidopsis* (*Arabidopsis thaliana*), rice (*Oryza sativa*) and poplar (*Populus alba*). An E-value significant threshold of ≤ 1E-05 was used for defining a hit. Detailed results of BLASTN analyses for all the four unigenes sets are given in Table 1. For instance, analysis of UG-III found highest identity of 60.3% with soybean, followed by cowpea (43.6%), *Medicago* (43.0%), common bean (42.2%), *Lotus* (37.2%), and the least identity with chickpea (23.2%). Comparative BLASTN analysis of pigeonpea unigenes with EST databases of model plant species showed, high identity with poplar (35.4%), followed by *Arabidopsis* (33.7%) and the least similarity with rice (28.3%). Of 4,557 unigenes, 2,839 (62.2%) showed significant identity with ESTs of at least one plant species analysed, while 227 (4.9%) showed significant identity across all the plant EST databases in this study. It is also interesting to note that 39 unigenes did not show any homology with the legume species examined.

To identify the putative function of all the unigenes compiled in this study, the unigenes from all the four sets (UG-I, UG-II, UG-III and UG-IV) were compared against the non-redundant UniProt database, using the BLASTX algorithm. At a significant threshold of ≤ 1E-08, 1,005 (30.30%) of UG-I, 638 (48.77%) of UG-II, 1,603 (35.17%) of UG-III and 1,777 (34.94%) of UG-IV unigenes showed significant similarity with known proteins (Figure 4). Details of BLASTX and BLASTN analyses against UniProt database for all four unigene sets are provided in Additional files 2, 3, 4 and 5.

Table 1 BLASTN analyses of pigeonpea unigenes against legume and model plant ESTs

High quality ESTs generated Unigenes	UG-I 5,680 3,316	UG-II 3,788 1,308	UG-III 9,468 4,557	UG-IV 10,376 5,085
Legume ESTs				
Pigeonpea (*Cajanus cajan*) (908)	314 (9.4%)	224 (17.1%)	508 (11.1%)	1,052 (20.6%)
Chickpea (*Cicer arietinum*) (7,097)	585 (17.6%)	507 (38.7%)	1,059 (23.2%)	1,155 (22.7%)
Soybean (*Glycine max*) (880,561)	1,690 (50.9%)	946 (72.3%)	2,750 (60.3%)	2,865 (56.3%)
Cowpea (*Vigna unguiculata*) (183,757)	1,230 (37.0%)	817 (62.4%)	1,988 (43.6%)	2,215 (43.5%)
Medicago (*Medicago truncatula*) (249,625)	1,214 (36.6%)	803 (61.3%)	1,963 (43.0%)	2,153 (42.3%)
Lotus (*Lotus japonicus*) (183,153)	1,015 (30.6%)	738 (56.4%)	1,698 (37.2%)	1,861 (36.5%)
Common bean (*Phaseolus vulgaris*) (83,448)	1,202 (36.2%)	784 (59.9%)	1,927 (42.2%)	2,146 (42.2%)
Significant similarity with ESTs of at least one legume species	1,768 (53.3%)	1,001 (76.5%)	2,757 (60.5%)	3,201 (62.9%)
Significant similarity across legume ESTs	172 (5.1%)	156 (11.9%)	274 (6.0%)	383 (7.5%)
No similarity with legume species	39 (1.1%)	4 (0.3%)	39 (0.8%)	42 (0.8%)
Model plant ESTs				
Arabidopsis (*Arabidopsis thaliana*) (1,527,298)	913 (27.5%)	667 (50.9%)	1,536 (33.7%)	1,669 (32.8)
Rice (*Oryza sativa*) (1,240,613)	810 (24.4%)	520 (39.7%)	1,294 (28.3%)	1,389 (27.3%)
Poplar (*Poplus alba*) (418,223)	982 (29.6%)	678 (51.8%)	1,617 (35.4%)	1,753 (34.4%)
Significant similarity with ESTs of at least one Model plant species	1,161 (35.0%)	763 (58.3%)	1,872 (41.0%)	2,019 (39.7%)
Significant similarity across ESTs of all model plant species	635 (19.1%)	460 (35.1%)	1,066 (23.3%)	1,135 (22.3%)
Significant similarity with ESTs of at least one plant species analyzed	1,839 (55.4%)	1,015 (77.5%)	2,839 (62.2%)	3,280 (64.5%)
Significant similarity across ESTs of all plant species analyzed	150 (4.5%)	114 (8.7%)	227 (4.9%)	299 (5.8%)
No similarity with ESTs of any plant species	39 (1.1%)	4 (0.3%)	39 (0.8%)	41 (0.8%)

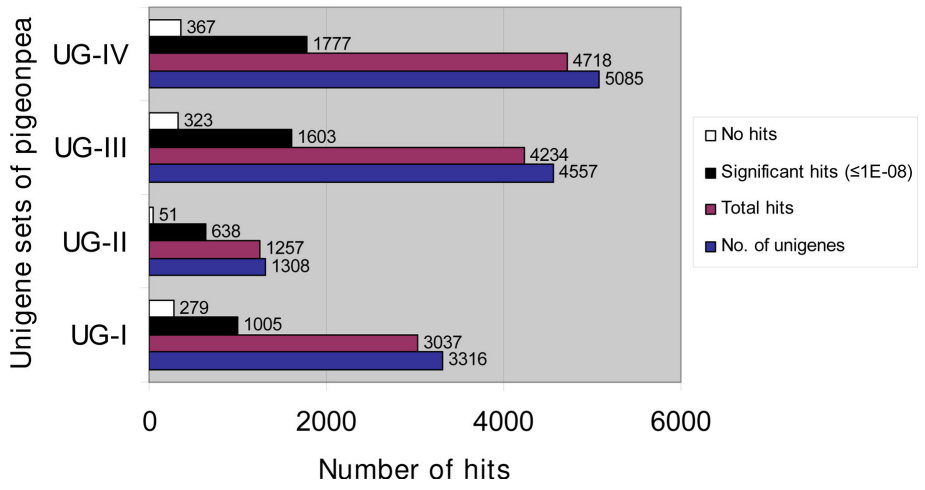

Figure 4 BLASTX analysis of pigeonpea unigenes against UniProt database. BLASTX homology search was performed for all the four unigene groups (UG-I, UG-II, UG-III and UG-IV) against the non-redundant UniProt database. The values against each bar represent total number of unigenes, total number of hits, significant hits at ≤ 1E-08 and no hits for each unigene set.

Functional categorization of pigeonpea unigenes

The unigenes from all the four sets that showed a significant hit (≤ 1E-08) against the UniProt database were further categorized into functional categories. As a result, 640 (63.6%) of UG-I, 448 (70.2%) of UG-II, 997 (62.1%) of UG-III and 1,119 (62.9%) of UG-IV unigenes were successfully annotated into three principal GO categories i.e. biological process, molecular function and cellular component. Like in earlier studies of this nature, it was observed that one gene could be assigned to more than one principal category, thus the total number of GO mappings from each category exceeded the number of unigenes analyzed. Details on full list of gene annotation for significant hits of four unigene sets are given in Additional file 6, 7, 8 and 9. For instance, of 1,603 (35.1%) unigenes of UG-III, only 997 (21.8%) were assigned to three principle categories. As a result, a total of 132 were grouped under biological process, 132 under molecular function and 153 under cellular component (Figure 5). Under the biological process category, cellular process accounted to 101, followed by metabolic process (82), biological regulation (32) and response to stimulus (21). In the cellular component category, 160 unigenes coded for cell part, 112 to organelle, and 70 to organelle part. In the last category of molecular function, majority of the unigenes were involved in binding (95) and catalytic activity (44). The remaining 606 unigenes which could not be classified into any of the three GO categories were grouped as "unclassified". The distribution of unigenes (UG-III) along with corresponding Gene Ontology (GO) categories are provided in Additional file 10. Based on GO annotation, enzyme commission IDs were also retrieved from the UniProt database to get an overview of unigenes (UG-III) putatively annotated to be enzymes. The major group of unigenes are included under oxidoreductases (107) followed by transferases (91), hydrolases (90), lyases (36), ligases (21) and isomerases (18). Similar patterns of distribution were observed in all the remaining Unigene sets.

In silico expression analysis

The identification of differentially expressed genes among specific cDNA libraries of FW- and SMD-responsive genotypes based on EST counts in each contig was done using a web statistical tool IDEG.6. As a result, 19 genes were identified to be differentially expressed between 'ICPL 20102' (FW- resistant) and 'ICP 2376' (FW-susceptible) genotypes, similarly, 20 genes were differentially expressed between 'ICP 7035' (SMD- resistant) and 'TTB 7' (SMD- susceptible) genotypes (Figure 6 and 7).

To assess the relatedness of each library and expressed genes in terms of expression pattern, a cluster analysis on the basis of EST abundance in each contig was performed [20]. Of the 697 contigs (UG-III), that were subjected to R-statistics [21] only 71 contigs were normalized with a true positive significance (R>8) and were eventually subjected to hierarchical clustering analysis (Additional file 11). The correlated gene expression pattern of all normalized 71 contigs/genes is displayed in Figure 8. All the 12 FW- derived libraries were grouped into a single cluster, while all the four SMD- challenged libraries were grouped into another cluster. About 49 genes were highly expressed in SMD- challenged libraries than in FW- challenged libraries and can be attributed to high accumulation of defence proteins

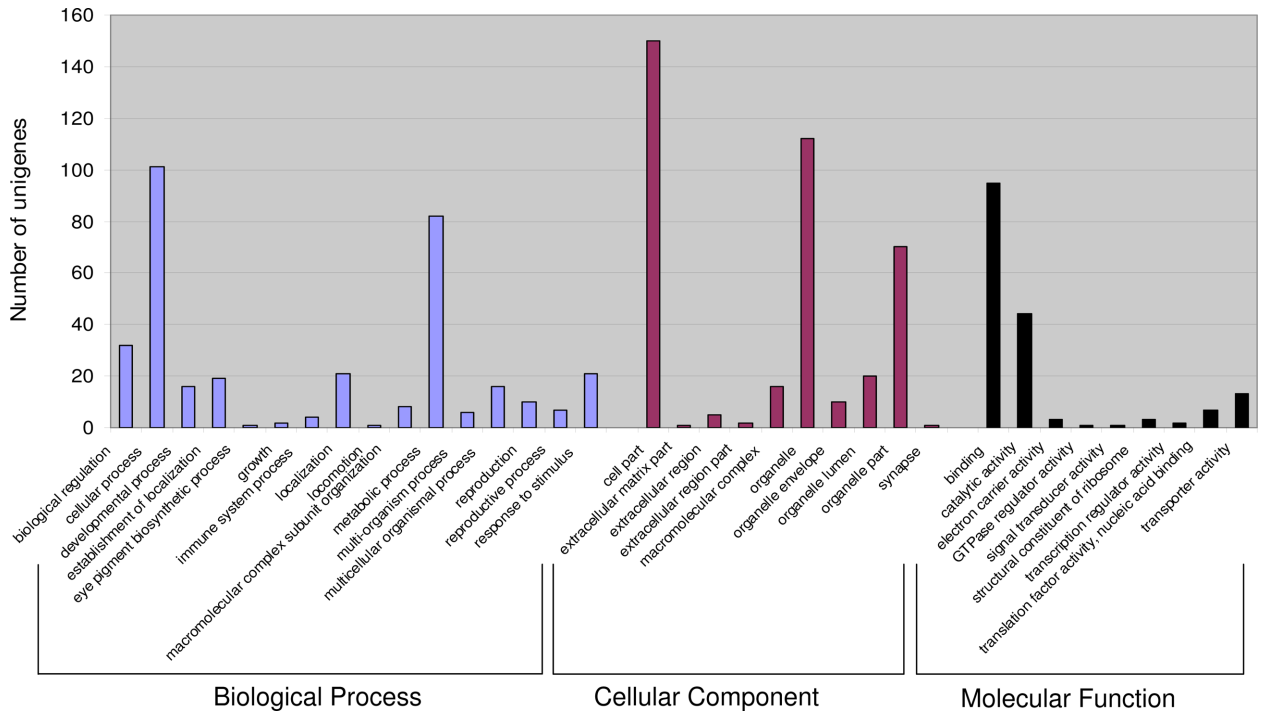

Figure 5 Gene Ontology (GO) assignment of pigeonpea unigenes (UG-III) by GO annotation. Functional categorization and distribution of 997 unigenes (UG-III) among three GO categories i.e biological process, cellular component and molecular function according to UniProt database.

during SMD infection. In the cluster of FW- challenged libraries, the 'ICPL 20102'-30 DAI library was distantly placed between FW- susceptible challenged libraries 'ICP 2376' - 6 DAI and 'ICP 2376' - 30 DAI. Each cluster represents a different pattern of gene expression as shown in Figure 8. Based on the clustering pattern and library specificity, Clusters I and IV were further divided into sub-clusters (represented in different colour bars). The above results indicated that the pattern and percentage of genes expression varied according to severity of the stress in specific library.

In Cluster I, 11.3% (8) of total genes were grouped and further sub divided into two groups with each sharing 2.8% (2) and 8.5% (6) genes, respectively. Similarly, Cluster II and Cluster III accounted for 4.2% (3) and 15.5% (11) genes and the largest Cluster IV, included 69.0% (49) of total genes with three sub groups IVa, IVb and IVc each sharing 14.0% (10), 10% (7) and 45% (32) of genes, respectively. Cluster analysis also showed high level expression of genes related to chloroplast/photosystem related proteins (22.5%), developmental proteins (19.7%), cellular proteins (15.4%), metabolic proteins (14.0%), defence/stimulus responsive proteins (4.3%), protein specific binding proteins (2.8%) and few uncharacterized proteins (19.8%).

Marker discovery

EST based markers can assay the functional genetic variation compared to other class of genetic markers and hence were targeted for marker development [22]. The unigene set based on generated ESTs in this study as well as the ones available in public domain was used for development of simple sequence repeats (SSR) and single nucleotide polymorphism (SNP) markers.

Identification and development of genic microsatellite markers

The entire set of 5,085 pigeonpea unigenes derived from UG-IV was used to identify the SSRs using *MISA* (*MI*cro*SA*tellite) tool [23]. As a result a total of 3,583 SSRs were identified at the frequency of 1/800 bp in coding regions (Table 2). 698 ESTs contained more than one SSR and 1,729 SSRs were found as compound SSRs. In terms of distribution of different classes of SSRs i.e. mono-, di-, tri-, tetra-, penta- and hexa-nucleotide repeats, mononucleotide SSRs contributed to the largest proportion (3,498, 97.6%). Only a limited number of SSRs of other classes were found. For instance, di- and tri- nucleotide SSRs accounted for 40 (1.1%) and 33 (0.9%) respectively. On the other hand, 9 tetrameric, 2 pentameric and 1 hexameric microsatellites were present (Figure 9). While using the criteria for Class I (> 20

UNIQID	Description	Lib1	Lib2	Lib1(norm)	Lib2(norm)
Contig18	>P51091|LDOX_MALDO Leucoanthocyanidin dioxygenase - Malus domestica	60	0	203.9	0
Contig21	>P41809|HKR1_YEAST Hansenula MRAKII killer toxin-resistant protein 1	1	33	3.4	120.6
Contig70	>Q6CQE5|TAR1_KLULA Protein TAR1 - Kluyveromyces lactis (Yeast)	13	71	44.2	259.4
Contig76	>Q8BLV3|SL9A7_MOUSE Sodium/hydrogen exchanger 7 - Mus musculus	5	38	16.1	138.8
Contig98	>P93276|M030_ARATH Uncharacterized mitochondrial protein AtMg00030 -	107	32	363.6	116.9
Contig124	>Q6CQE5|TAR1_KLULA Protein TAR1 - Kluyveromyces lactis (Yeast)	177	27	601.4	98.6
Contig180	>P72823|NU4C2_SYNY3 NAD(P)H-quinone oxidoreductase chain 4-2 -	6	39	20.4	142.5
Contig189	>Q8TGM7|ART2_YEAST Uncharacterized protein ART2 - Saccharomyces	138	220	468.9	803.8
Contig211	>P23472|CHLY_HEVBR Hevamine-A precursor [Includes: Chitinase - Hevea	21	1	71.4	3.7
Contig214	>Q99700|ATX2_HUMAN Ataxin-2 - Homo sapiens (Human)	39	97	132.5	354.4
Contig220	>Q0P591|MA6D1_BOVIN MAP6 domain-containing protein 1 - Bos taurus	0	92	0	336.1
Contig222	>Q99002|1433_TRIHA 14-3-3 protein homolog - Trichoderma harzianum	0	15	0	54.8
Contig224	>P26257|BGAL_THETU Beta-galactosidase - Thermoanaerobacter	0	32	0	116.9
Contig230	>Q1KMD3|HNRL2_HUMAN Heterogeneous nuclear ribonucleoprotein U-like	0	23	0	84
Contig231	>P30432|FUR2_DROME Furin-like protease 2 precursor - Drosophila	0	11	0	40.2
Contig240	>Q4U2V3|CBPC1_DANRE Cytosolic carboxypeptidase 1 - Danio rerio	0	25	0	91.3
Contig327	>Q3BAI2|YCX91_PHAAO Uncharacterized protein ORF91 - Phalaenopsis	2	19	6.8	69.4
Contig328	>Q05047|C72A1_CATRO Cytochrome P450 72A1 - Catharanthus roseus (Rosy	0	19	0	69.4
Contig377	>P0C5Q0|YL54F_YEAST Uncharacterized protein YLR154W-F - Saccharomyces	21	79	71.4	288.6

Figure 6 Differential gene expression between FW- responsive genotypes using IDEG.6 web tool. Differentially expressed genes between libraries of FW-resistant ('ICPL 20102') and susceptible ('ICP 2376') genotypes. Cells with different degrees of blue color represent extent of gene expression.

nucleotides in length) and Class II SSRs (< 20 nucleotides in length) as used by Temnykh and colleagues [24] and Kantety and colleagues [25], on all SSRs 641 SSRs represented Class I while 2,942 SSRs represented Class II (Table 2).

In general, mononucleotide SSRs are not included for primer designing and synthesis. However, as only a very limited number of SSR markers are currently available for pigeonpea in public domain and in a separate study some mononucleotide SSRs were found polymorphic [15], primer pairs were designed for 383 SSRs including mononucleotide SSRs. A total of 94 primer pairs were considered for validation after excluding the primers for monomeric SSR motifs and compound SSRs with mononucleotide repeats. However based on repeat number criteria, such as 5 minimum for di-, tri-, tetra-, pentanucleotides, primer pairs were synthesized for 84 SSRs. The details of newly developed pigeonpea EST-SSR primers along with corresponding SSR motif, primer sequence, annealing temperature and product size are provided in Additional file 12.

Newly synthesized 84 markers were analyzed on 40 elite pigeonpea genotypes (Additional file 13). As a result, 52 (61.9%) primer pairs provided scorable amplified products and 26 primer pairs produced a number of faint bands indicative of non-specific amplifications. A total of 15 (28.8%) markers showed polymorphism with 2-7 alleles with an average of 4 alleles per marker in genotypes examined. These markers showed a moderate PIC value ranging from 0.20 to 0.70 with an average of 0.40 (Table 3). To evaluate the genetic variability within a diverse collection of pigeonpea accessions which are parents of different mapping populations segregating for important agronomic traits and also to determine genetic relationship among them, phylogenetic analysis on the basis of dissimilarities was performed using NTSYS software package. The UPGMA cluster diagram showed clear segregation of wild and cultivated species (Figure 10).

SNP discovery and identification of CAPS markers

SNPs are an important class of molecular markers which are becoming more popular in recent times. To enhance the reliability of SNPs identification, the SNP which occurred in a contig ≥ 5 ESTs from more than one genotype was considered. *In silico* analysis showed a total of 102 SNPs in 37 (27,659 bp) contigs with a

UNIQID	Description	Lib1	Lib2	Lib1(norm)	Lib2(norm)
Contig1	>Q9FY64|RS154_ARATH 40S ribosomal protein S15-4 – Arabidopsis	13	0	68.6	0
Contig5	>P40620|HMGL_VICFA HMG1/2-like protein - Vicia faba (Broad bean)	19	0	100.3	0
Contig7	>Q6BK66|CCS1_DEBHA Superoxide dismutase 1 copper chaperone -	17	1	89.8	5.3
Contig9	>Q9XF89|CB26_ARATH Chlorophyll a-b binding protein CP26, chloroplast	21	2	110.9	10.6
Contig15	>Q43517|FER1_SOLLC Ferredoxin-1, chloroplast precursor - Solanum	43	8	227	42.2
Contig16	>P43399|MT1_TRIRP Metallothionein-like protein 1 - Trifolium repens	45	12	237.6	63.4
Contig20	>Q05502|HHEX_CHICK Homeobox protein PRH - Gallus gallus (Chicken)	40	5	211.2	26.4
Contig30	>P49107|PSAN_ARATH Photosystem I reaction center subunit N,	21	0	110.9	0
Contig44	>Q93VI8|TLP7_ARATH Tubby-like F-box protein 7 - Arabidopsis thaliana	15	0	79.2	0
Contig49	>Q06930|ABR18_PEA ABA-responsive protein ABR18 - Pisum sativum	13	0	68.6	0
Contig55	>P17067|CAHC_PEA Carbonic anhydrase, chloroplast precursor - Pisum	24	3	126.7	15.8
Contig57	>Q9XFB0|YAB2_ARATH Axial regulator YABBY 2 - Arabidopsis thaliana	13	0	68.6	0
Contig81	>Q5XJD3|FIP1_DANRE Pre-mRNA 3-end-processing factor FIP1 - Danio	22	1	116.2	5.3
Contig87	>Q9XEX2|PRX2B_ARATH Peroxiredoxin-2B - Arabidopsis thaliana	14	0	73.9	0
Contig177	>P93276|M030_ARATH Uncharacterized mitochondrial protein AtMg00030 -	0	26	0	137.3
Contig188	>Q9ULL4|PLXB3_HUMAN Plexin-B3 precursor - Homo sapiens (Human)	41	125	216.5	659.1
Contig198	>Q6CQE5|TAR1_KLULA Protein TAR1 - Kluyveromyces lactis (Yeast)	1	44	5.3	232.3
Contig203	>Q9MTN0|YCX6_OENHO Uncharacterized 6.9 kDa protein in psbD-trnT	4	32	21.1	168.1
Contig217	>Q8TGM7|ART2_YEAST Uncharacterized protein ART2 - Saccharomyces	6	57	31.7	300.1
Contig294	>Q59296|CATA_CAMJE Catalase - Campylobacter jejuni	188	94	992.6	496.3

Figure 7 Differential gene expression between SMD- responsive genotypes using IDEG.6 web tool. Differentially expressed genes between libraries of SMD resistant ('ICP 7035') and susceptible ('TTB 7') genotypes. Cells with different degrees of blue color represent extent of gene expression.

frequency of 1/271 bp (Table 4). With an objective of validating these *in silico* identified SNPs, as an example, 10 contigs were used to generate PCR amplicons and sequence four genotypes namely 'ICPL 20102', 'ICP 2376', 'ICP 7035' and 'TTB 7'. While a scorable and sequenceable amplicon was obtained in case of 6 contigs (contig 210, contig 433, contig 535, contig 555, contig 620 and contig 718), the scorable amplicons were not obtained in case of four contigs (contig 67, contig 330, contig 587 and contig 632). Sequencing of amplicons for all the four genotypes for all the six contigs showed occurrence of SNPs as predicted *in silico* (Additional file 14). For instance, for contig 433, a comparison of the amplified DNA sequences for four genotypes ('ICPL 20102', 'ICP 2376', 'ICP 7035' and 'TTB 7') with the 5 EST sequences coming from two genotypes ('ICP 7035' and 'TTB 7') showed the occurrence of the same SNP G to C between 'ICP 7035' and 'TTB 7' (Figure 11).

In order to perform cost-effective and robust genotyping assay for the detected 102 SNPs in 37 contigs, efforts were made to identify the restriction enzymes that can be used to assay SNPs via cleaved amplified polymorphic sequence (CAPS) assay. Results indicated that SNPs present in 37 contigs can be evaluated by using CAPS assay (Table 4).

Discussion

Plants are known to have developed integrated defence mechanisms against fungal and viral infections by altering spatial and temporal transcriptional changes. The EST approach was successfully utilized in identification of disease-responsive genes from various tissues and growth stages in chickpea [26], *Lathyrus* [27], soybean [28], rice [29] and ginseng [30]. Many earlier studies have shown that resistant genotypes have efficient mechanisms for stress perception and enhanced expression of defence-responsive genes, which maintain cellular survival and recovery [31]. Hence, the present study was undertaken to identify catalog of defence related genes in response to FW and SMD infection in pigeonpea by generating ESTs from different stress challenged tissues at various time intervals.

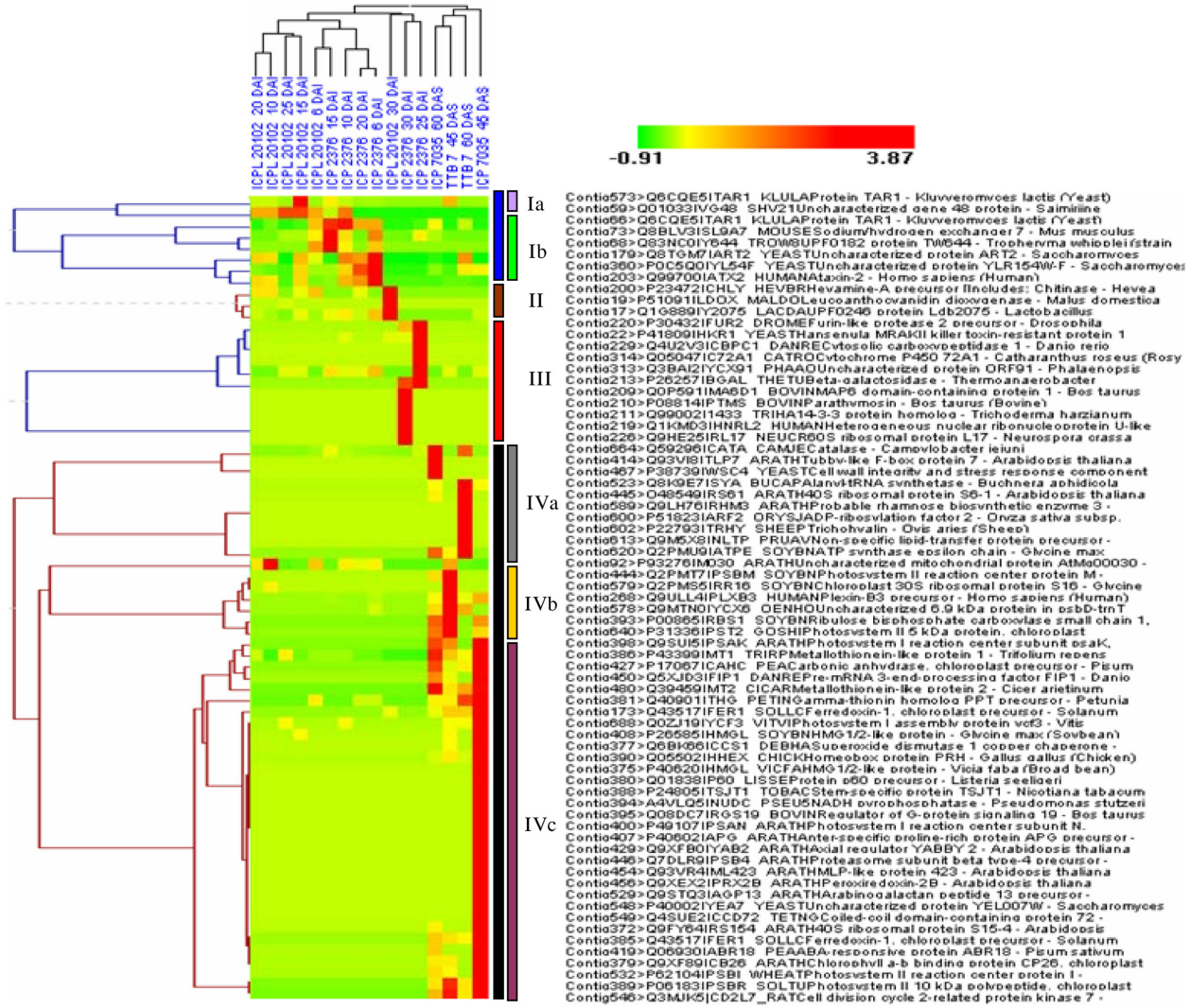

Figure 8 Hierarchical clustering analysis of differentially expressed genes from 16 libraries of pigeonpea using HCE version 2.0 beta web tool. Clusters of genes highly expressed in different libraries of pigeonpea genotypes subjected to FW and SMD stress. Columns represent different cDNA libraries and their relationship in a dendrogram. Clustering of highly expressed ESTs (normalized using R statistics, R>8) into four major clusters (indicated in vertical colour bars), and their cluster sub groups based on their library specificity. Colour scale represents the range of expression pattern by different genes with respect to libraries.

Generation of cDNA libraries and unigene assemblies

Roots provide a structural and physiological support for plant interactions with the soil environment by conducting transport of water, ions and nutrients. Plants are encountered with many biotic stress factors which includes bacterial, fungal and viral infection. Roots and leaves are the primary sites of infection by these organisms. Therefore, a total of 16 cDNA libraries were generated at different time intervals to specifically target the roots infected with *Fusarium udum* and leaves infected with SMD. In total 5,680 high quality ESTs were generated from FW- and similarly 3,788 high quality ESTs from SMD- challenged genotypes. Earlier, at the time of analysis in November 2008, the public domain consisted of only 908 ESTs for pigeonpea. Thus the present study contributes approximately 10-fold increase in the pigeonpea EST resource and an addition of 4,557 pigeonpea unigenes (UG-III).

Functional annotation of pigeonpea unigenes

Homology searches (BLASTN and BLASTX) against other plant ESTs and functional characterization was done for all the defined unigene datasets (UG-I, UG-II, UG-III and UG-IV). Of the 5,085 unigenes (UG-IV) assembled from all the pigeonpea ESTs, 3,280 (64.5%) had significant identity with ESTs of at least one plant

Table 2 Features of SSRs identified in ESTs

SSR database mining			
Total number of sequences examined			5,085
Total length of examined sequences (bp)			2,878,318
Number of ESTs containing SSRs			1,365 (26.8%)
Number of identified SSRs			3,583
Number of sequences containing more than 1 SSR			698
Number of SSRs present in compound formation			1,729
Frequency of SSR			1/0.8 kb
Distribution of SSRs			
Type	Class I	Class II	Total
Mono-nucleotides	607	2,891	3,498
Di-nucleotides	10	30	40
Tri-nucleotides	12	21	33
Tetra-nucleotides	9	0	9
Penta-nucleotides	2	0	2
Hexa-nucleotides	1	0	1
Total	641	2,942	3,583

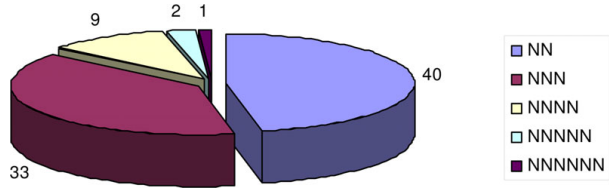

Figure 9 EST-SSR motifs derived from pigeonpea unigenes (UG-IV). Number of EST-SSR repeat motifs (excluding monomers) derived from unigenes (UG-IV) of pigeonpea cDNA libraries subjected to FW and SMD stress.

species analyzed, 299 (5.8%) unigenes showed significant identity with ESTs of all analyzed plant species in the study, while 41 (0.8%) were found to be novel to pigeonpea. A high significant identity was observed with soybean (56.3%), and the least percentage of similarity was observed with chickpea (22.7%) (Table 1). A similar BLASTN results were observed for the remaining three unigenes sets (UG-I, UG-II and UG-III) against the ESTs of plant species surveyed. Comparative analysis of newly defined UG-III dataset (4,557) with 908 public domain pigeonpea ESTs showed that only 508 (11.1%) shared identity and indicated that our EST sequencing study identified 4,049 (88.9%) new set of pigeonpea unigenes. Relatively, very low similarity of 36.5% with *Lotus* and 42.3% with *Medicago* was observed compared to soybean and cowpea than other legume species. These observations are in accordance with phylogenetic relationships of legumes [32].

The pigeonpea ESTs showed higher similarity to legume ESTs databases (22.7-56.3%) of the legume species than monocot species (27.3-33.4%). Comparative analysis of pigeonpea ESTs with monocot species like rice (27.3%) showed that the percentage of significance is much lower compared to any other legume species, inspite of larger EST repository. This is clearly attributed to phylogenetic divergence between dicots and monocots in course of evolution. These comparisons also indicate that several unigenes that were absent in analysed non-legumes but present in all legume species may be specifically confined to legumes.

BLASTX analyses indicated that those ESTs without significant identity to any other protein sequences in the existing database may be novel and involved in plant defence responses. Hence, this novel EST collection represented a significant addition to the existing pigeonpea EST resources and provides valuable information for further predictions/validation of gene functions in pigeonpea.

A comprehensive comparison of functionally categorized unigenes of all the four unigenes data sets (UG-I, UG-II, UG-III and UG-IV) showed a similar distribution. A large number of unigenes were involved in cell part, organelle, binding, organelle part, metabolic and cellular process among the significantly annotated ones. These observations are consistent with the earlier reported functional categorization studies in rice [29], soybean [33], barley [34] and tall fescue [35]. However, the sequences encoding activities related to categories such as biological regulation and response to stimulus are 28 and 20 incase of FW-responsive ESTs compared to 0 and 2 in case of SMD-responsive ESTs. This was possibly due to the fact that the ESTs generated from FW- challenged root libraries were most abundantly involved in stimulus to pathogenesis and ESTs derived from SMD stress are chloroplast binding proteins. Earlier studies such as Lee and colleagues [36], Ablett and colleagues [37], also reported that photosynthesis-related proteins were the most prevalent from aerial parts of the plant, which would help to make energy related activities such as cell division, growth, elongation and development. Similarly in this study, photosynthesis related genes were identified in larger proportion (30%) in SMD-responsive cDNA libraries derived from leaf tissues.

In silico differential gene expression

The invasion of pathogen not only results in expression of novel genes/transcripts, but also in altering the abundances of different ESTs resulting in induction or repression. This was evident from differential expression of 19 genes between FW-responsive genotypes and 20 genes between SMD-responsive genotypes. It is however, important to mention that *in silico* method of gene expression is not the ideal method to identity the

Table 3 Characteristics of pigeonpea EST-SSR markers

Primer ID	SSR motif	Tm (°C)	Product size (bp)	No. of alleles	PIC value
ICPeM0001	$(A)_{56}ttg(A)_{30}$	60	240	1	0.00
ICPeM0003	$(A)_{23}n(A)_{11}n(C)_{12}$	60	150	7	0.79
ICPeM0005	$(A)_{99}n(C)_{11}$	60	280	5	0.35
ICPeM0006	$(T)_{37}gg(T)_{56}$	60	246	1	0.00
ICPeM0009	$(A)_{85}g(A)_{27}$	60	208	1	0.00
ICPeM0010	$(T)_{11}n(T)_{20}$	60	266	1	0.00
ICPeM0011	$(A)_{58}gggg(A)_{24}$	60	280	1	0.00
ICPeM0013	$(A)_{87}g(A)_{29}n(C)_{11}$	62	150	4	0.66
ICPeM0017	$(AG)_8n(AT)_8$	59	240	1	0.00
ICPeM0018	$(A)_{10}taca(T)_{12}$	59	90	1	0.00
ICPeM0019	$(TTA)_7n(T)_{12}$	60	236	1	0.00
ICPeM0023	$(A)_{128}n(C)_{11}n(C)_{11}$	60	279	1	0.00
ICPeM0024	$(A)_{11}cccg(A)_{10}$	60	279	1	0.00
ICPeM0025	$(A)_{10}n(C)_{11}n(C)_{11}n(C)_{12}$	61	223	1	0.00
ICPeM0028	$(A)_{58}cc(A)_{27}$	61	239	1	0.00
ICPeM0029	$(A)_{57}t(A)_{30}$	60	184	1	0.00
ICPeM0030	$(A)_{52}tt(A)_{28}$	61	252	1	0.00
ICPeM0031	$(A)_{13}gn(A)_{67}n(A)_{19}$	60	279	1	0.00
ICPeM0033	$(A)_{12}tt(A)_{12}n(A)_{13}$	60	350	7	0.31
ICPeM0034	$(A)_{13}n(AT)_9$	60	236	1	0.00
ICPeM0035	$(T)_{21}n(A)_{11}$	60	218	1	0.00
ICPeM0038	$(C)_{15}acctcactaaccaaact(C)_{10}$	60	266	1	0.00
ICPeM0039	$(G)_{10}n(T)_{94}$	59	262	1	0.00
ICPeM0041	$(T)_{12}n(A)_{10}$	60	310	4	0.48
ICPeM0047	$(T)_{18}c(T)_{27}$	60	213	1	0.00
ICPeM0050	$(A)_{112}n(C)_{13}n(C)_{11}$	63	264	1	0.00
ICPeM0052	$(C)_{12}tccctcctctcgccca(C)_{12}$	60	233	1	0.00
ICPeM0053	$(C)_{24}t(C)_{26}$	60	136	1	0.00
ICPeM0054	$(G)_{10}agccc(G)_{10}$	60	<90	1	0.00
ICPeM0060	$(T)_{13}c(T)_{10}$	60	150	1	0.00
ICPeM0061	$(T)_{19}n(A)_{28}$	59	243	1	0.00
ICPeM0064	$(ATT)_7(T)_{10}$	59	300	3	0.49
ICPeM0065	$(A)_{10}(AT)_9$	60	90	1	0.00
ICPeM0066	$(AT)_9$	60	310	3	0.34
ICPeM0067	$(TA)_{11}$	60	200	3	0.29
ICPeM0068	$(GT)_{11}$	60	260	4	0.26
ICPeM0069	$(AT)_8$	60	<90	1	0.00
ICPeM0070	$(AT)_8$	61	310	1	0.00
ICPeM0071	$(GA)_9$	61	190	5	0.64
ICPeM0072	$(AT)_8$	60	<90	1	0.00
ICPeM0073	$(AG)_9$	60	<90	1	0.00
ICPeM0074	$(AGA)_6$	60	300	1	0.00
ICPeM0075	$(ACA)_6$	60	300	2	0.38
ICPeM0076	$(CTT)_6$	60	200	1	0.00
ICPeM0077	$(AAT)_7$	60	310	1	0.00
ICPeM0078	$(GCC)_6$	60	320	3	0.24
ICPeM0079	$(ATT)_6$	60	250	4	0.40
ICPeM0080	$(TGGAC)_5$	60	200	1	0.00
ICPeM0081	$(TAAT)_5$	60	300	1	0.00
ICPeM0082	$(AT)_9$	60	200	3	0.27
ICPeM0083	$(AG)_9$	60	190	1	0.00
ICPeM0084	$(TATG)_6$	60	240	3	0.59

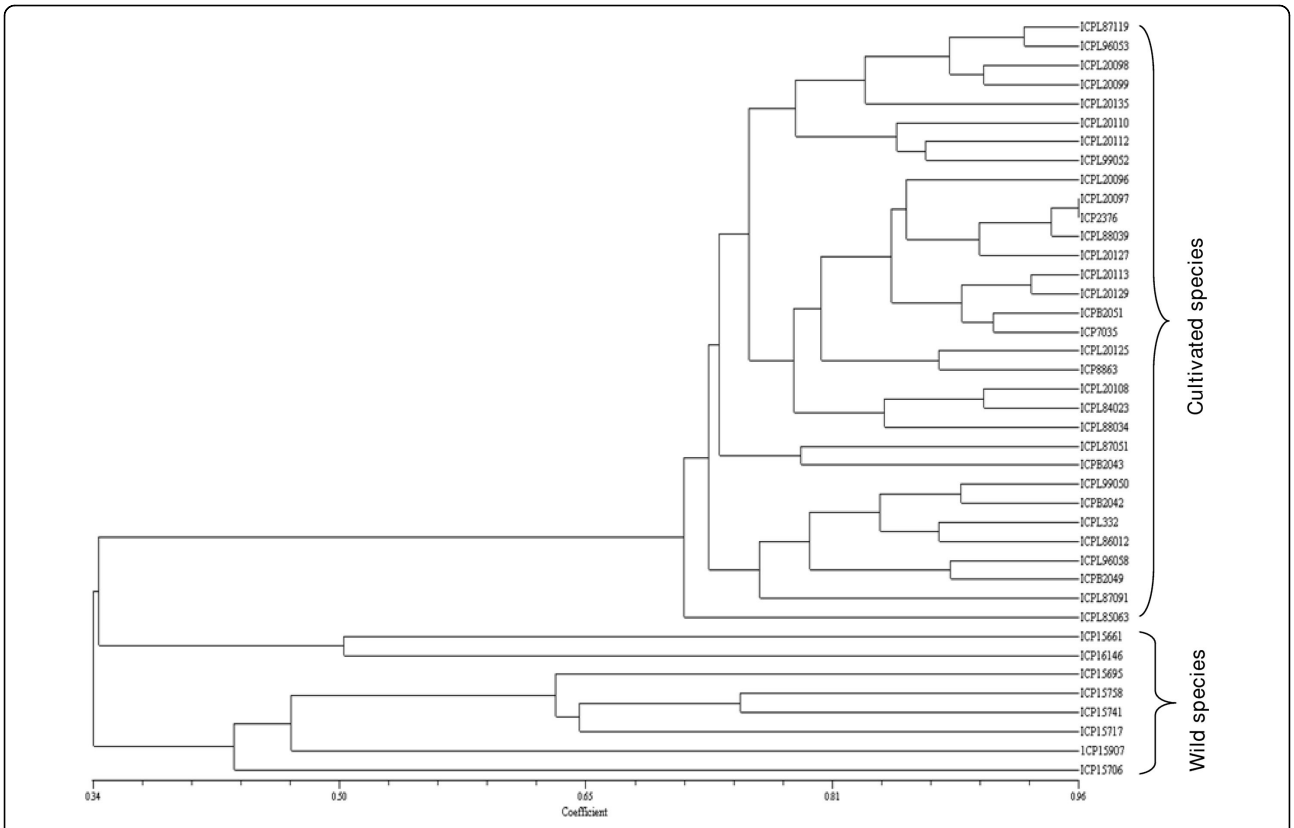

Figure 10 Dendrogram of elite pigeonpea accessions based on UPGMA analysis. Unweighted Pair Group Method using arithmetic average dendrogram showing relatedness among the forty elite pigeonpea genotypes representing 8 wild species and 32 cultivated genotypes. The scale at the bottom of the dendrogram indicates the level of similarity between the genotypes.

Table 4 Summary of SNPs and CAPS markers identified from pigeonpea ESTs

Total number of contigs examined (UG-IV)	**871**
Number of contigs containing ≥ 5 ESTs	133
Number of contigs containing SNPs	37
Total length of 37 contigs (bp)	27,659
Total number of identified SNPs in 37 contigs	102
Average SNP frequency	1/271 bp
Total number of contigs containing CAPS convertible SNPs	37

differentially expressed genes. Nevertheless, as large scale EST data were generated from FW- responsive and SMD- responsive genotypes, an effort like some earlier studies [18-21,34] was made to identify some putative genes differentially expressed in FW- and SMD- resistant and sensitive genotypes. Validation of these candidate genes by Northern analysis or real-time quantitative PCR analysis is essential before these candidate genes are deployed in some other studies.

Significant number of unigene sequences related to proteins like kinases, phosphatases, peroxidases, ribonucleases, endochitinases, glucanases and hormones like Abscisic acid responsive (ABA) genes were identified to be differentially expressed and are known to play a vital role in defence mechanism. For example, the cell wall degrading enzymes like endochitinases (EC: 3.2.1.14) implicate a major defence mechanism against pathogen [27]. Similarly, kinases play a major role in the plant's recognition to pathogen [38,39]. For instance, chitinase protein (UniProt ID: P23472), a class of pathogenesis related (PR) proteins with bi-functional role in lysozyme/chitinase activity involved in random hydrolysation of N-aetyl-beta-D-glucosaminide-beta linkages in chitin and chitodextrins during systemic acquired resistance (SAR), was expressed at higher concentrations in FW-responsive resistant genotype ('ICPL 20102') compared to susceptible genotype ('ICP 2376'). The high expression levels of chitinase in resistant genotype indicate the effectiveness within a narrow range of pathogenesis [40,41].

Similarly, the protein coding for ABA-responsive protein (ABR18) (UniProt ID: Q06930), which is involved in stimulus mechanism and cell localization etc. during plant development and one of the vital roles is in defence mechanism during biotic stress signaling. This gene was identified to be expressed relatively higher in

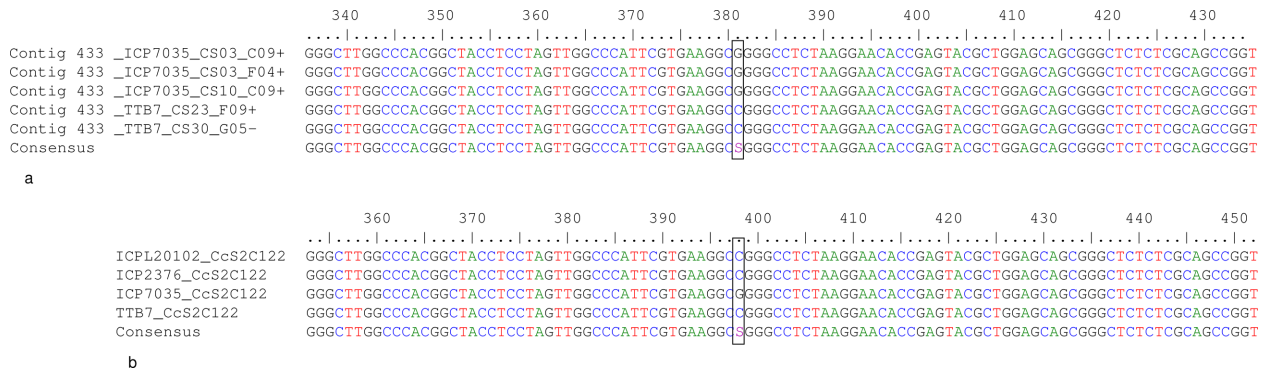

Figure 11 A snapshot of sequence alignment of EST sequences and amplicons for contig 433 validating the *in silico* predicted SNP. CAP3 alignment of ESTs (a) generated from 'ICP 7035' and 'TTB 7'in contig 433 showing a SNP between these genotypes and (b) Multiple sequence alignment of amplicon sequences generated from genomic DNA of four genotypes ('ICPL 20102', 'ICP 2376', 'ICP 7035' and 'TTB 7') with the primer pairs for the assembled contig 433. The *in silico* identified SNP in the EST contig 433 was confirmed in the amplicon sequences.

SMD-resistant pigeonpea genotype 'ICP 7035' compared to the susceptible genotype 'TTB 7'. During pathogen infection ABA inhibits the transcription of a basic β-1, 3-glucanase (EC: 3.2.1.39) that can degrade the β-1, 3-glucan callose, forming a physical barrier to viral spread through plasmodesmata. This down regulation of β-1, 3-glucanase by ABA can be termed as a resistance factor in plant pathogen interactions [42]. In our study, significant expression signals were observed in SMD resistant genotype 'ICP 7035' during viral infection. This positive correlation between the ABA levels and disease resistance was reported in plant species like common bean [43], rice [44] and tobacco [45]. Different enzymes like methyltransferases (HMT3) (UniProt ID: Q8LAX0) and dehydrogenases (G3PC) (UniProt ID: P34921) are putatively involved in synthesis of lignin in cell walls. These enzymes also play a major role in defence against pathogen interaction [46,47].

An *in silico* hierarchical clustering analysis of 71 differentially expressed and genes across 16 cDNA libraries using HCE V 2.0 was done to infer potential relation between the co-expressed genes. The profiles of some of the interesting gene families and genes that could play an important role in stress stimulus were explained.

In Cluster I, of the 8 contigs, 6 were identified to be highly expressed in FW- challenged libraries of susceptible genotype. The cluster includes genes encoding proteins involved in mitochondrial DNA (mtDNA) stability, Na^+/H^+ exchanger and a few uncharacterized proteins. The sub cluster Ia includes two ESTs, which are highly expressed in 'ICPL 20102' libraries (15 DAI and 25 DAI). The genes connected in sub cluster Ib are highly expressed in 'ICP 2376' libraries (6 DAI and 15 DAI). One of the putative proteins TAR1 (Transcript Antisense to Ribosomal RNA), a mitochondrial protein is known to be involved in regulation and respiratory metabolism. Over- expression of this protein suppresses the respiration-deficient petite phenotype of a point mutation in mitochondrial RNA polymerase that affects mitochondrial gene expression and mtDNA stability. This dysfunction of mitochondria might occur in response to biotic or abiotic stress [48]. The over-expression of these genes was observed only in 15 DAI libraries of both the FW- responsive genotypes. And their immediate disappearance in the later stages of infection in resistant genotype libraries and continued expression in susceptible genotypes supports a hypothesis that continual expression of this protein may lead to mitochondrial dysfunction and subsequent cell degeneracy.

Another protein Na^+/H^+ exchanger 7 (UniProt ID: Q8BLV3) is an ubiquitous ion transporter that serves multiple cell physiological processes such as intracellular pH homeostasis and electro neutral exchange of protons for Na^+ and K^+ across endomembranes. Biochemical studies suggest that Na^+/H^+ exchangers in the plasma membrane of plant cells contribute to cellular sodium homeostasis during salt stress, although the above protein is expressed in high salt stressed plants, it may also be expressed during biotic stress. Its high expression in susceptible genotype 'ICPL 2376' at 6 and 15 DAI libraries shows the severity of stress during fungal pathogenesis. And in the remaining FW- and SMD-challenged libraries this shows normal expression.

Cluster II genes include hevamine (EC: 3.2.1.14) and Leucoanthocyanidin dioxygenase (EC: 1.14.11.19) specifically expressed in 'ICPL 20102' 30 DAI library. The important protein hevamine represents a new class of polysaccharide-hydrolyzing $(\beta\alpha)_8$ barrel enzyme belonging to families of plant chitinases and lysozymes, which are vital for plant defence against pathogenic bacteria and fungi. Recent results indicate that these enzymes

may be involved not only in defence-related process or general stress response but also in growth and development processes [49]. The high expression of these proteins in the late phase of *Fusarium* infection indicates their prolonged defensive role against fungal pathogenesis.

The genes connected in Cluster III are highly expressed in FW-challenged 'ICP 2376' - 25 and 30 DAI libraries. The genes related to endoprotease activity, beta-glucan synthesis, carboxypeptidases, alkaloid biosynthesis, secologanin biosynthesis, beta-galactosidase activity, microtubule-stabilizing activity, nucleic acid binding protein, ribonucleoprotein and a few uncharacterized proteins were constituted in this cluster. For instance, antifungal class proteins such as beta-glucanases (EC: 3.2.1.39) and *Hansenula mrakii* killer toxin-resistant proteins (UniProt ID: P41809) located in epidermal leaf cells are believed to be involved in cell differentiation and defence against fungal pathogens [50]. Plants deficient in these enzymes generated by antisense transformation showed markedly reduced resistance to viral and fungal infection. Similarly, another class of proteins carboxypeptidases (EC: 3.4.16 - 3.4.18) with diverse functions ranging from catabolism to regulating biological processes, including function as a defence against pathogen attack [51].

The majority of genes (49) segregated in Cluster IV were highly expressed in SMD-responsive cDNA libraries derived from leaf tissues. In total Cluster IV genes showed high gene expression in SMD derived libraries. As expected, photosynthesis related transcripts were abundantly represented in sub clusters IVa, IVb and IVc, and these include putative transcripts like ribosomal proteins, mitochondrial proteins, chloroplast precursor proteins, photosystem I and II reaction centre proteins. The observed expression pattern of photosynthesis related proteins in this study is also consistent with experimental observations in barley [34].

Overall, the differentially expressed genes are involved in diverse pathways, displaying complex expression patterns. The different clusters based on monitoring of gene expression patterns propose that various pathways in response to biotic stress exist in pigeonpea and their interaction can lead to differential stress tolerance. The uncharacterized class of transcripts co-expressed could be repository of novel proteins and further characterization of these may reveal their significant role in plant stress responses [52].

Development of functional markers

One of our primary goals of our research programme is to develop molecular markers based on expressed sequences and screen them for polymorphism. During the last decade, microsatellites or SSRs have proven to be useful markers in plant genetic research and have been used for marker-assisted breeding purposes. The presence of SSRs in the coding region suggests their importance as functional or gene based markers [1,11,53]. Unfortunately, development of microsatellite markers is expensive, labor intensive, and time consuming if they are being developed from genomic libraries [54]. The data mining of microsatellites markers from EST data can be a cost effective option. The cost of mining EST libraries is far lower than other traditional methods, and SSR development from ESTs has been successful in EST data mining [22,23,53-56]. SSR motifs with repeats more than eight for di-nucleotides, six for tri-nucleotides, and five for tetra-nucleotides were considered. Dimeric repeat motifs (40) were relatively abundant than trimeric repeats (33). In addition to this, tetra-, penta- and hexameric repeat motifs were considerably less represented. A total of 94 SSR markers have been synthesized and characterized for polymorphism survey. However, there are some distant contrasts in frequency and distribution of SSRs in ESTs and in genomic survey sequences (GSSs). In general di-nucleotide SSRs of all repeat lengths are more common in GSSs and tri-nucleotide SSRs are common in the ESTs [22,23,56,57]. As against these reports, in our findings we observed that di-nucleotide repeats are more abundant than tri-nucleotide repeat motifs [58,59]. However this observation is not unexpected as the frequency and distribution of SSR depends on several factors such as size of dataset, tools and criteria used for SSR discovery [22].

In this study, a total of 15 polymorphic EST-SSRs primer pairs were validated and used for diversity study on forty pigeonpea genotypes representing 32 cultivated (*C. cajan*) and 8 wild species (six *C. scarabaeoides* and two *C. platycarpus*). All markers detected at least one allele in all genotypes tested, suggesting transferability of all markers across the *Cajanus* genus. In addition to high transferability, EST-SSRs are good candidates for the development of conserved orthologous sequence (COS) markers for genetic analysis and breeding of different species [10]. However, EST-SSRs were reported to be less polymorphic than genomic SSRs in crop plants due to greater DNA sequence conservation in transcribed regions [22,60]. For instance, the 15 SSR loci provided only 60 alleles with an average of 4 alleles per loci and an average 0.43 PIC value. Similar kind of diversity features were observed in earlier SSR based diversity studies in pigeonpea [14,15].

EST-SSR profiles obtained on 40 pigeonpea genotypes were used to compute pair-wise genetic distances among different genotypes to construct a dendrogram based UPGMA clustering. The neighbor joining tree grouped 40 pigeonpea genotypes into three major clusters (Figure 10). The Cluster I comprising 32 genotypes

(cultivated) is the largest cluster followed by Cluster III containing six wild genotypes (*C. scarabaeoides*) and Cluster II is the smallest cluster with two wild genotypes belonging to *C. platycarpus* species revealing clear segregation of the cultivated and the wild species. Less genetic variation was detected with in cultivated species, with only nine markers detecting polymorphism and a total of 35 alleles. The low genetic variability amongst cultivars when compared with the wild species genotypes suggests that natural and artificial selection has contributed to the selection of specific alleles and to changes of allelic frequencies at specific loci as reported by Odeny and colleagues [14]. The distinctness of *C. platycarpus* with *C. scarabaeoides* accessions observed in this study correlate well with earlier studies [61]. It is also important to note that 'ICPL 20097' and 'ICP 2376' genotypes were found closely related with high genetic similarity as both of these genotypes belong to the same geographic region. In conclusion, EST-SSR markers developed in this study complement the currently available or ongoing efforts on development of genomic SSRs that will be a valuable resource for linkage map development and marker assisted selection in pigeonpea [12].

SNPs and indels are an essentially inexhaustible resource of polymorphic markers for use in the high-resolution genetic map development of traits and for association studies. Although a variety of molecular markers are available SNPs are comparatively advantageous because of their abundance and amenability to high throughput approaches [62]. In addition, SNPs also offer several advantages like high-throughput and cost-effective genotyping [63] and identification of functional/gene-based markers for complex trait through linkage map development or association genetics [9,10,64,65]. Although SNP discovery was a cost effective task in past, advances in next generation sequencing technologies have made SNP discovery cheaper and faster [66]. However in case for a given species, ESTs are available from more than one genotype, *in silico* mining of ESTs is still a very inexpensive and fast approach for SNP discovery [17,64] and therefore we used this approach for mining SNPs in this study.

By using *in silico* mining approach in a total of 871 contigs coming from 10,376 ESTs (9,888 generated in this study and 908 available in public domain), a total of 102 potential SNPs were identified in 37 contigs that were consisted of ≥ 5 ESTs. Smaller contigs were not considered for SNP mining as these contigs are prone to errors due to lack of read depth as reported by Wang and colleagues [67]. Sequence analysis of PCR products for a subset of 6 out of 10 contigs confirmed the occurrence of SNPs in all the cases. As PCR products could not be generated for remaining four contigs, the presence of SNPs could not be confirmed in those cases. Furthermore, as SNP genotyping is another important criteria in breeding programmes, identification of CAPS markers for 37 contigs will facilitate SNP genotyping even in low tech laboratories [63].

Conclusion

This study has contributed a new and significant set of 9,888 ESTs that together with 908 public domain ESTs provides a unigene set of 5,085 sequences for pigeonpea. Detailed analysis of these datasets have provided several important features of pigeonpea transcriptome such as conserved genes (across legumes and model plant species) as well as possible pigeonpea specific genes, assignment of pigeonpea genes to different GO categories, identification of differentially expressed genes in response to FW- and SMD- stresses, etc. In terms of applied aspect of developed resource in breeding, this study has demonstrated development and application of gene-based molecular markers i.e SSRs, SNPs and CAPS. In summary, it is anticipated that this study is a significant contribution to enhance genomic resources in a so called orphan legume crop that will eventually impact pigeonpea breeding [11,12].

Methods

Plant material

Four pigeonpea genotypes namely 'ICPL 20102' (FW-resistant), 'ICP 2376' (FW- susceptible), 'ICP 7035' (resistant to SMD) and 'TTB 7' (highly susceptible to SMD) were used for constructing the cDNA libraries and generating the ESTs. Seeds of two genotypes ('ICPL 20102', 'ICP 2376') were procured from Legume Pathology section at ICRISAT and for the remaining two genotypes ('ICP 7035' and 'TTB 7') were obtained from Dr. M Byregowda, University of Agricultural Sciences, Bangalore, India.

A total of 40 genotypes including 32 genotypes from cultivated species (*C. cajan*) and 8 genotypes from 2 wild species (*C. platycarpus* and *C. scarabaeoides*) were used for validation and diversity analysis with new set of EST-SSR markers. These genotypes were obtained from Pigeonpea Breeding (Dr. KB Saxena) and Genebank (Dr. HD Upadhyaya) and have been listed in Additional file 13.

Inoculation treatment for FW and SMD

Seeds of FW-tolerant ('ICPL 20102') and FW-susceptible ('ICP 2376') were germinated in 15-inch deep polythene covers filled with sterile soil and sand (1:1) in a glass house at 23 ± 3°C under 80% relative humidity. The root, being the primary target of the pathogen *Fusarium udum* and the possible site of the initial defence response, was selected as the tissue of study. Ten days old seedlings were uprooted from pots and the root

system was thoroughly washed in running tap water and rinsed with distilled water. Seedlings of each genotype were inoculated by immersing the roots for 2 min in fungal inoculum (*Fusarium udum* culture). The spore suspension at 6×10^5 conidia/ml, was made by adding fungal spores from several culture plates (*Fusarium* was grown on potato dextrose media supplemented with 0.25 μg/ml tetracycline). Immediately following the inoculation, the seedlings were transplanted to sterilized sand and soil mixture (1:1) in pots and were transferred to glass house.

In order to capture the genes expressed in resistant and susceptible genotypes at different time periods after inoculation, six stages i.e. 6, 10, 15, 20, 25 and 30 days after inoculation (DAI) were selected arbitrarily to construct the cDNA libraries. From 6-15 DAI, chlorosis symptoms were observed on the leaves and aerial parts of plant material, indicated the severity of *Fusarium* wilt disease. Furthermore, from each stage of days after inoculation, shoot and root section cuttings were made and observed the fungal penetration into the vascular tissues. Based on microscopic observations at different stages, initial symptoms of fungal infection were noticed in root vascular tissue at 15 and 20 DAI stages. Beyond 20 DAI, though the fungus penetrates deeper into the vascular tissues of susceptible and resistant varieties to some extent, the susceptible variety shows complete *Fusarium* symptoms where as the resistant variety prevents most of the attacking fungus from reaching maturity and developing symptoms.

For SMD study, highly susceptible ('TTB 7') and resistant ('ICP 7035') pigeonpea genotypes that are parents of a mapping population segregating for resistance to SMD were chosen. Forty seeds from each accession were sown in plastic bags filled with sterilized soil and were maintained in a glass house under optimal physiological conditions as described above. Ten days after sowing, the aerial parts of the seedlings were stapled with mosaic virus infected leaves. The viral disease is caused by pigeonpea sterility mosaic virus (PPSMV) and transmitted by an eriophyid mite *Aceria cajani* Channabasavanna [7]. The disease slowly spreads into the vascular tissues from the aerial parts through mite population which is characterized by a bushy and pale green appearance of plants. Based on the severity of disease symptoms, leaves with visible SMD lesions were harvested at 45 and 60 days after sowing (DAS) stages for construction of cDNA libraries.

cDNA library construction

Root and leaf tissue samples were collected from FW- and SMD- responsive genotypes at different time-points till the infection stage reached stagnant phase. RNA was isolated from the above two tissue samples according to the protocol described by Schmitt and colleagues [68]. RNA quality was assessed using formamide gel electrophoresis and poly $(A)^+$ RNA was isolated with poly (A) tract mRNA isolation system IV (Promega, Madison, WI, USA) as described by the manufacturers. Double-strand cDNA was constructed using Super SMART™ PCR cDNA Synthesis kit (Clontech®, Mountain View, CA, USA) as described in the manufacturer's instructions. The resulting cDNA was size fractioned on 1.2% agarose gel. cDNA fractions containing fragments greater than 500 bp were selected for library construction. Subsequently, the cDNA was ligated into pGEM® Easy vector (Promega®, Madison, WI, USA) and ligation was allowed to proceed overnight at 14°C. The resulting plasmids were electroporated using One Shot® Top 10 Electrocomp™ cells (Invitrogen, Carlsbad, CA, USA). The transformants were spread on LB Agar plates containing 100 mg/ml ampicillin for direct picking. Based on blue/white screening, recombinant clones were picked into Nunc-Immuno™ 96 MicroWell™ Plates (Nunc™, Roskilde, Denmark) containing LB broth with 100 μg/ml ampicillin and grown for overnight at 37°C on a rotary shaker at 220 rpm. Glycerol stocks in 96-well format were prepared by combining 38 μl of 60% (v/v) glycerol with 150 μl of culture and frozen at -80°C.

EST sequencing, editing and assembly

Clones were randomly selected and on an average of 500 clones were sequenced per library in case of FW-response study and 1000 clones per library in case of SMD-responsive study. The plasmid DNA from these clones (i.e. colonies) was extracted using a 96-well alkaline lysis method prior to sequencing [69]. Plasmid DNA sequencing was performed by commercial DNA sequencing service provider (Macrogen Inc., Korea) using the standard M13 forward primer.

The FASTA files containing the raw sequences were edited by the software Sequencher™ 4.0 (Gene Codes Corporation, Ann Arbor. MI, USA) to remove the vector sequences. The vector screened sequences were subjected to EST trimmer [70], to trim poly-A ends and low quality sequences. High quality sequences of >100 bp were selected for further sequence analysis. ESTs were clustered and aligned into contigs and singletons using the CAP3 program [71].

In order to assess the number of unique and overlapping transcripts among the 16 libraries, four data sets were generated; those derived from libraries constructed from of FW-responsive genotypes (UG-I); those derived from libraries constructed from of SMD-responsive genotypes (UG-II); combined dataset of FW- and SMD-responsive ESTs (UG-III); and also from public domain sequences with total generated ESTs in this study (UG-IV). In addition to the above assembly of unigene sets,

CAP3 analysis was also performed to libraries derived from FW- resistant genotype, FW- susceptible genotype, SMD- resistant genotype and from SMD- susceptible genotype individually.

Homology search and functional annotation

The unigene sequences were also characterized for nucleotide homology search against the EST datasets of selected legume species [pigeonpea (*Cajanus cajan*)-908, chickpea (*Cicer arietinum*)-7,097, soybean (*Glycine max*)-880,561, *Medicago* (*Medicago truncatula*)-249,625, common bean (*Phaseolus vulgaris*)-83,448, cowpea (*Vigna unguiculata*)-183,757 and *Lotus* (*Lotus japonicus*)-183,153] and selected model plant species [rice (*Oryza sativa*)-1,240,613, *Arabidopsis* (*Arabidopsis thaliana*)-1,527,298 and poplar *(Populus alba*)-418,223] available at National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov) using BLASTN algorithm [72]. A match was considered significant at E-value ≤ 1E-05.

Each unigene dataset was subjected to BLASTX analysis against the non-redundant protein database of UniProt to deduce a putative function. Sequence similarity was considered as significant at E-value ≤ 1E-08. Each unigene was assigned a putative cellular function based on the significant database hit with lowest e-value. Subsequently, unigenes that showed a significant BLASTX hit were used for functional annotation based on Gene Ontology categories from UniProt database (UniProt-GO). This process allowed assignment of unigenes to the GO functional categories of biological process, cellular component and molecular function. Distribution of unigenes was further investigated in terms of their assignment to sub-categories of the main GO categories.

In silico expression and hierarchical clustering

In order to identify the differentially expressed genes in FW- and SMD- responsive genotypes, 389 contigs coming from FW-responsive genotypes and 328 contigs coming SMD-responsive genotypes were analyzed by using IDEG.6 web interface tool [73,74]. The IDEG.6 web tool allows running six different statistical analyses for the detection of differentially expressed genes in multiple tag experiments. For pair-wise comparisons, the Audic and Claverie test, Fisher exact test and chi-square tests (X^2) were used and in multiple comparisons R- statistics test, Greller and Tobin test and chi-square tests (X^2) were used [73,74].

Further, gene expression analysis was performed with hierarchical clustering expression (HCE) version 2.0 beta software [75] using transcript abundance data from UG-III set that includes 697 contigs derived from both the stress responsive libraries. As a pre-requisite for HCE analysis, all 697 contigs were subjected to R statistics (R>8) and only those contigs (71) were selected that have; (i) minimum 5 ESTs, and (ii) differential abundance of ESTs coming from different libraries. The matrix file developed based on the frequency of ESTs to each of 71 contigs was used as input file for above mentioned HCE tool.

Identification and development of SSR markers

A total of 5,085 unigenes (unigene set, UG-IV) developed based on 9,888 ESTs generated in this study and 908 public domain ESTs were searched with a Perl script program, *MISA* (*MI*cro*SA*tellite) [23,76] for identification and localization of SSRs. The SSR motifs, with repeat units more than five times in di-, tri-, tetra-, penta- and hexa- nucleotides were considered as SSR search criteria in *MISA* script. The Primer3 programme [77] was used for designing the primer pairs for SSRs and custom synthesized by MWG (MWG-Biotech AG, Bangalore, India).

The primer pairs for SSRs were tested for their utility as potential genetic markers on 40 elite genotypes of pigeonpea (Additional file 13). PCR amplifications were performed in 5 μl reactions containing 5 ng of genomic DNA, 1× SE-Taq DNA polymerase buffer (including 1.5 mM $MgCl_2$), 2 mM dNTPs, 10 pmol of each primer and 0.1 U *Taq* DNA polymerase (*SibEnzyme*, Novosibirsk, Russia) with the following touch down profile; 3 min at 95°C; 5 cycles of 20 sec at 94°C, 20 sec at 60°C minus 1° C/cycle, 30 sec at 72°C; 40 cycles of 20 sec at 94°C, 20 sec at 56°C, 30 sec at 72°C; and 20 min at 72°C for final extension. PCR products were separated on 6% non-denaturing polyacrylamide gels for 3 h at 600 V and visualized by silver staining. The polymorphism information content (PIC) of individual EST-SSR markers was calculated by using the standard formula [62]. Only data from polymorphic SSR loci were used for diversity analysis. Genetic similarities between any two genotypes were estimated according to Nei and Li [78]. All 40 genotypes were clustered with the Unweighted Pair Group Method using arithmetic average (UPGMA) in the SAHN procedure of the NTSYS-PC v2.10t [79].

SNP detection and their conversion into CAPS

All 871 contigs obtained from the collection of 5,085 unigenes (UG-IV) were searched for putative SNP/indels by using an integrated pipeline for large scale SNP discovery [80,81]. The pipeline utilized the CAP3 output files as input to detect SNPs/indels based on the nucleotide redundancy in the multiple sequence alignments. The auto SNP pipeline generated text file includes contig ID, number of sequences in the contig ID, consensus length, number of SNPs, mutation type and SNP frequency. The threshold for identification of SNPs was based on the number of sequences (≥ 5) in each

consensus sequence and two or more sequences from different genotype. In order to verify the SNPs at sequence level, the PCR amplicons of all four genotypes were sequenced using the corresponding forward and reverse primers for a set of 10 contigs (see Additional file 14). The amplicons were purified and further sequencing was done as described [80]. The sequenced data along with the sequences of ESTs (that provided the SNPs initially) were aligned and analyzed using BioEdit programme http://www.mbio.nesu.edu/BioEdit.html.

For converting SNPs into cleaved amplified polymorphic sequence (CAPS) markers, SNPs present in 37 contigs were analyzed to identify the recognition site for any of commercially available 725 restriction enzymes [82] by using integrated SNP2CAPS pipeline [80].

Additional file 1: Sterility mosaic disease (SMD) responsive pigeonpea seedlings. a) Sterility mosaic disease infected pigeonpea genotypes 'ICP 7035' and 'TTB 7' at 45 days after sowing (DAS); initiation of SMD infection to the aerial parts of susceptible genotype 'TTB 7'; b) Severe SMD infection observed in the susceptible genotype ('TTB 7') showing pale green and bushy aerial parts after 60 DAS as against resistant genotype (ICP 7035).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-45-S1.PPT]

Additional file 2: BLASTX and BLASTN result of UG-I dataset. Tables showing BLASTX and BLASTN results of unigene (UG-I) dataset with corresponding Genbank (GB) ID numbers, sequence name, length, score, E-value and identity.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-45-S2.XLS]

Additional file 3: BLASTX and BLASTN result of UG-II dataset. Tables showing BLASTX and BLASTN results of unigene (UG-II) dataset with corresponding GB ID numbers, sequence name, length, score, E-value and identity.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-45-S3.XLS]

Additional file 4: BLASTX and BLASTN result of UG-III dataset. Tables showing BLASTX and BLASTN results of unigene (UG-III) dataset with corresponding GB ID numbers, sequence name, length, score, E-value and identity.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-45-S4.XLS]

Additional file 5: BLASTX and BLASTN result of UG-IV dataset. Tables showing BLASTX and BLASTN results of unigene (UG-IV) dataset with corresponding GB ID numbers, sequence name, length, score, E-value and identity.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-45-S5.XLS]

Additional file 6: Gene Ontology categorization for UG-I dataset. Tables showing significant hits (≤ 1E-08) of unigenes from pigeonpea unigene (UG-I) dataset and its corresponding Gene Ontology categories according to UniProt database.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-45-S6.XLS]

Additional file 7: Gene Ontology categorization for UG-II dataset. Tables showing significant hits (≤ 1E-08) of unigenes from pigeonpea unigene (UG-II) dataset and its corresponding Gene Ontology categories according to UniProt database.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-45-S7.XLS]

Additional file 8: Gene Ontology categorization for UG-III dataset. Tables showing significant hits (≤ 1E-08) of unigenes from pigeonpea unigene (UG-III) dataset and its corresponding Gene Ontology categories according to UniProt database.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-45-S8.XLS]

Additional file 9: Gene Ontology categorization for UG-IV dataset. Tables showing significant hits (≤ 1E-08) of unigenes from pigeonpea unigene (UG-IV) dataset and its corresponding Gene Ontology categories according to UniProt database.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-45-S9.XLS]

Additional file 10: Gene Ontology categorization for UG-III dataset. Tables showing significant hits (≤ 1E-08) of unigenes from four pigeonpea unigene dataset (UG-III) and its corresponding Gene Ontology categories: a) Biological process b) Cellular component c) Molecular function.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-45-S10.PPT]

Additional file 11: Hierarchical clustering of UG-III contigs. Table showing data matrix of 71 contigs as four clusters, represented in Hierarchical clustering dendrogram with corresponding number of ESTs represented from each library.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-45-S11.XLS]

Additional file 12: List of newly developed pigeonpea EST-SSRs. List of newly developed pigeonpea EST-SSR markers with corresponding details of primer ID, SSR motif, primer sequence, melting temperature and product size.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-45-S12.XLS]

Additional file 13: List of pigeonpea elite genotypes used for diversity assessment. List of pigeonpea accessions used in assessment of newly synthesized EST-SSR markers with corresponding details of species name, geographical origin, type.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-45-S13.XLS]

Additional file 14: Validation of *in silico* identified SNPs in EST contigs through sequencing. Validation experiments of *in silico* identified SNPs have been shown in this file for 10 contigs.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-45-S14.XLS]

Acknowledgements
Authors are thankful to Indo-USAgricultural Knowledge Initiative (Indo-USAKI) supported by Indian Council of Agricultural Research (ICAR), Government of India and SP2-Leader Discretionary Grant from Generation Challenge Program http://www.generationcp.org for the financial support to undertake this study. Thanks are also due to Department of Biotechnology (DBT), Government of India for sponsoring a Post-Doctoral Fellowship to NLR. Authors are thankful to Dr. K.B. Saxena and Dr. H.D. Upadhayaya of ICRISAT

for providing the seeds/DNA of some genotypes used in this study. Thanks are also due to Mr. A. Bhanu Prakash and Ms. Spurthi Nayak for their help in data analysis and discussions. Authors are also thankful to three anonymous reviewers for their valuable suggestions on the first version of the MS that helped in improvement of the MS.

Author details
[1]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Greater Hyderabad 502 324, Andhra Pradesh, India. [2]University of Agricultural Sciences, Gandhi Krishi Vignyan Kendra (GKVK), Bangalore, 560 065, Karnataka, India. [3]National Research Centre on Plant Biotechnology (NRCPB), Indian Agricultural Research Institute, New Delhi 110 012, India. [4]Genomics towards Gene Discovery Sub Programme, Generation Challenge Programme (GCP) c/o CIMMYT, Int. Apartado Postal 6-641, 06600, Mexico, DF Mexico.

Authors' contributions
NLR, BNG and PL conducted experiments, NLR, BJ, PJH and RKV analyzed EST data, SP and MB contributed plant tissues to generate ESTs, BJ, NKS and RKV provided scientific inputs to analyze and interpret results, NLR, PJH and RKV wrote the manuscript in consultation with other co-authors, RKV conceived, planned coordinated and supervised the overall study and finalized the manuscript. All authors read and approved the final manuscript.

Received: 21 August 2009 Accepted: 11 March 2010
Published: 11 March 2010

References

1. Varshney RK, Hoisington DA, Upadhyaya HD, Gaur PM, Nigam SN, Saxena KB, Vadez V, Sethy NK, Bhatia S, Aruna R, Channabyre Gowda MV, Singh NK: **Molecular genetics and breeding of grain legume crops for the semi-arid tropics.** *Genomic-Assisted crop improvement* Springer Netherlands, The NetherlandsVarshney RK, Tuberosa R 2007, 207-241.
2. Greilhuber J, Obermayer R: **Genome size variation in *Cajanus cajan* (Fabaceae): a reconsideration.** *Plant Syst Evol* 1998, **212**:135-141.
3. **FAOSTAT 2006.** [http://faostat.fao.org].
4. Rao SC, Coleman SW, Mayeux HS: **Forage production and nutritive value of selected pigeonpea ecotypes in the southern great plains.** *Crop Sci* 2002, **42**:1259-1263.
5. Prasad P, Eswara Reddy NP, Anandam RJ, Lakshmikantha Reddy G: **Isozymes variability among *Fusarium udum* resistant cultivars of pigeonpea (*Cajanus cajan* (L.) (Millsp).** *Acta Physiol Plant* 2003, **25**:221-228.
6. Butler EJ: **The wilt disease of pigeonpea and pepper.** *Agricult J India* 1906, **1**:25-26.
7. Kannaiyan J, Nene YL, Reddy MV, Ryan JG, Raju TN: **Prevalence of pigeonpea disease and associated crop losses in Asia, Africa and America.** *Trop Pest Manage* 1984, **30**:62-71.
8. Kumar PL, Jones AT, Reddy DVR: **A novel mite transmitted virus with a divided RNA genome closely associated with Pigeonpea sterility mosaic disease.** *Phytopathol* 2003, **93**:81-91.
9. Varshney RK, Graner A, Sorrells E: **Genomics-assisted breeding for crop improvement.** *Trends Plant Sci* 2005, **10**:621-630.
10. Varshney RK, Hoisington DA, Tyagi AK: **Advances in cereal genomics and applications in crop breeding.** *Trends Biotechnol* 2006, **11**:490-499.
11. Varshney RK, Close TJ, Singh NK, Hoisington DA, Cook DR: **Orphan legume crops enter the genomics era.** *Curr Opin Plant Biol* 2009, **12**:202-210.
12. Varshney RK, Penmetsa RV, Dutta S, Kulwal PL, Saxena RK, Datta S, Sharma TR, Rosen B, Carrasquilla-Garcia N, Farmer AD, Dubey A, Saxena KB, Gao J, Fakrudin B, Singh MN, Singh BP, Wanjari KB, Yuan M, Srivastava RK, Kilian A, Upadhyaya HD, Mallikarjuna N, Town CD, Bruening GE, He G, May GD, McCombie R, Jackson SA, Singh NK, Cook DR: **Pigeonpea genomics initiative (PGI): an international effort to improve crop productivity of pigeonpea (*Cajanus cajan* L.).** *Mol Breed* 2009.
13. Burns MJ, Edwards KJ, Newbury HJ, Ford-Lloyd BV, Baggott CD: **Development of simple sequence repeat (SSR) markers for the assessment of gene flow and genetic diversity in pigeonpea (*Cajanus cajan*).** *Mol Ecol Notes* 2001, **1**:283-285.
14. Odeny DA, Jayashree B, Ferguson M, Hoisington D, Crouch J, Gebhardt C: **Development, characterization and utilization of microsatellite markers in pigeonpea.** *Plant Breed* 2007, **126**:130-136.
15. Saxena RK, Prathima C, Saxena K, Hoisington DA, Singh NK, Varshney RK: **Novel SSR markers for polymorphism detection in pigeonpea (*Cajanus* spp.).** *Plant Breed* 2009.
16. Odeny DA, Jayashree B, Gebhardt C, Crouch J: **New microsatellite markers for pigeonpea (*Cajanus cajan* (L.) millsp.).** *BMC Res Notes* 2009, **2**:35.
17. Sreenivasulu N, Kavi Kishor PB, Varshney RK, Altschmied L: **Mining functional information from cereal genomes -the utility of expressed sequence tags.** *Curr Sci* 2002, **83**:965-973.
18. Ogihara Y, Mochida K, Nemoto Y, Murai K, Yamazaki Y, Shin IT, Kohara Y: **Correlated clustering and virtual display of gene expression patterns in the wheat life cycle by large-scale statistical analyses of expressed sequence tags.** *Plant J* 2003, **33**:1001-1011.
19. Ronning CM, Stegalkina SS, Ascenzi RA, Bougri O, Hart AL, Utterbach TR, Vanaken SE, Riedmuller SB, White JA, Cho J: **Comparative analyses of potato expressed sequence tag libraries.** *Plant Physiol* 2003, **131**:419-429.
20. Ewing RM, Kahla AB, Poirot O, Lopez F, Audic S, Claverie JM: **Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression.** *Genome* 1999, **9**:950-959.
21. Stekel DJ, Git Y, Falciani F: **The comparison of gene expression from multiple cDNA libraries.** *Genome Res* 2000, **10**:2055-2061.
22. Varshney RK, Graner A, Sorrells E: **Genic microsatellite markers in plants: features and applications.** *Trends Biotechnol* 2005, **23**:48-55.
23. Thiel T, Michalek W, Varshney RK, Graner A: **Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.).** *Theor Appl Genet* 2003, **106**:411-422.
24. Temnykh S, DeClerck G, Lukashova A, Lipocic L, Cartinhour S, McCouch S: **Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential.** *Genome Res* 2001, **11**:1441-1452.
25. Kantety RV, La Rota M, Matthews DE, Sorrells ME: **Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat.** *Plant Mol Biol* 2002, **48**:501-510.
26. Coram TE, Pang ECK: **Isolation and analysis of candidate ascochyta blight defence genes in chickpea. Part I. Generation and analysis of an expressed sequence tag (EST) library.** *Physiol Mol Plant Path* 2005, **66**:192-200.
27. Skiba B, Ford R, Pang CK: **Construction of a cDNA library of *Lathyrus sativus* inoculated with *Mycosphaerella pinodes* and the expression of potential defence-related expressed sequence tags (ESTs).** *Physiol Mol Plant Path* 2005, **66**:55-67.
28. Iqbal MJ, Yaegashi S, Ahsan R, Shopinski KL, Lightfoot DA: **Root response to *Fusarium solani* f. sp. *glycines*: temporal accumulation of transcripts in partially resistant and susceptible soybean.** *Theor Appl Genet* 2005, **110**:1429-1438.
29. Jantasuriyarat C, Gowda M, Haller K, Hatfield J, Lu G, Stahlberg E, Zhou B, Li H, Kim H, Yu Y, Dean RA, Wing RA, Soderlund C, Wang GL: **Large-Scale identification of expressed sequence tags involved in rice and rice blast fungus interaction.** *Plant Physiol* 2005, **138**:105-115.
30. Goswami R, Punja ZK: **Molecular and biochemical characterization of defense responses in ginseng (*Panax quinquefolius*) roots challenged with *Fusarium equiseti*.** *Physiol Mol Plant Path* 2008, **72**:10-20.
31. Reddy PCO, Sairanganayakulu G, Thippeswamy M, Reddy PS, Reddy MK, Sudhakar C: **Identification of stress-induced genes from the drought tolerant semi-arid legume crop horsegram (*Macrotyloma uniflorum* (Lam.) Verdc.) through analysis of subtracted expressed sequence tags.** *Plant Sci* 2008, **175**:372-384.
32. Wojciechowski MF, Sanderson MJ, Steele KP, Liston A: **Molecular phylogeny of the "temperate herbaceous tribes" of papilionoid legumes: a supertree approach.** *Advances in Legume Systematics* Royal Botanic Gardens: KewHerendeen P, Bruneau A 2000, **9**:277-298.
33. Alkharouf N, Khan R, Matthews B: **Analysis of expressed sequence tags from roots of resistant soybean infected by the soybean cyst nematode.** *Genome* 2004, **47**:380-388.
34. Zhang H, Sreenivasulu N, Weschke W, Stein N, Rudd S, Radchuk V, Potokina E, Scholz U, Schweizer P, Zierold U, Langridge P, Varshney RK, Wobus U, Graner A: **Large-scale analysis of the barley transcriptome based on expressed sequence tags.** *Plant J* 2004, **40**:276-290.
35. Mian MAR, Zhang Y, Wang Z, Zhang J, Cheng X, Chen L, Chekhovskiy K, Dai X, Mao C, Cheng F, Zhao X, He J, Scott AD, Town CD, May GD: **Analysis**

of tall fescue ESTs representing different abiotic stresses, tissue types and developmental stages. *BMC Plant Biol* 2008, **8**:27.
36. Lee MC, Lee YJ, Lee MH, Nam HG, Cho TJ, Hahn TR, Cho MJ, Sohn U: **Large-scale analysis of expressed genes from the leaf of oilseed rape (*Brassica napus* L.).** *Plant Cell Rep* 1998, **17**:930-936.
37. Ablett E, Seaton G, Scott K, Shelton D, Graham MW, Baver-stock P, Lee LS, Henry R: **Analysis of grape ESTs: global gene expression patterns in leaf and berry.** *Plant Sci* 2000, **159**:87-95.
38. Yu I, Fengler KA, Clough SJ, Bent AF: **Identification of *Arabidopsis* mutants exhibiting an altered hypersensitive response in gene-for-gene disease resistance.** *Microbe Interact* 2000, **13**:277-286.
39. Piffanelli P, Zhou R, Casais C, Orme J, Jarosch B, Schaffrath U, Collins NC, Panstruga R, Schulze-Lefeert P: **The barley MLO modulator of defence and cell death is responsive to biotic and abiotic stress stimuli.** *Plant Physiol* 2002, **129**:1076-1085.
40. de las Mercedes Dana M, Pintor-Toro JA, Cubero B: **Transgenic tobacco plants over expressing chitinases of fungal origin show enhanced resistance to biotic and abiotic stress agents.** *Plant Physiol* 2006, **142**:722-730.
41. Ferreira RB, Monteiro S, Freitas R, Santos CN, Chen Z, Batista LM, Duarte J, Borges A, Teixeira AR: **The role of plant defence proteins in fungal pathogenesis.** *Mol Plant Pathol* 2007, **8**:677-700.
42. Mauch-Mani B, Mauch F: **The role of abscisic acid in plant-pathogen interactions.** *Curr Opin Biotech* 2005, **8**:409-414.
43. Mayek-Perez NO, Garcia-Espinosa R, Lopez-Castaneda C, Acosta-Gallegos JA, Simpson J: **Water relations, histopathology and growth of common bean (*Phaseolus vulgaris* L.) during pathogenesis of *macrophomina phaseolina* under drought stress.** *Physiol Mol Plant Pathol* 2002, **60**:185-195.
44. Koga H, Dohi K, Mori M: **Abscisic acid and low temperatures suppress the whole plant-specific resistance reaction of rice plants to the infection with *Magnaporthe grisea*.** *Physiol Mol Plant Pathol* 2004, **65**:3-9.
45. Whenham RJ, Fraser RSS, Brown LP, Payne JA: **Tobacco mosaic virus-induced increases in abscisic acid concentration in tobacco leaves: intracellular location in light and darkgreen areas, and relationship to symptom development.** *Planta* 1986, **168**:592-598.
46. Jennings DB, Ehrenshaft M, Pharr DM, Williamson JD: **Roles for mannitol and mannitol dehydrogenase in active oxygen-mediated plant defense.** *Proc Natl Acad Sci USA* 1998, **95**:15129-15133.
47. Liu CJ, Deavours BE, Richard SB, Ferrer JL, Blount JW, Huhman D, Dixon RA, Noel JP: **Structural basis for dual functionality of isoflavonoid O-methyltransferases in the evolution of plant defense responses.** *Plant Cell* 2006, **18**:3656-3669.
48. Day DA, Millar AH, Whelan J: **Plant mitochondria: from genome to function.** Springer publishers, The Netherlands 2004.
49. Scheltinga ATV, Kalk KH, Beintema JJ, Dijkstra B: **Crystal structures of Hevamine, a plant defence protein with chitinases and lysozymes activity, and its complex with an inhibitor.** *Structure* 1994, **2**:1181-1189.
50. Grenier J, Potvin C, Asselin A: **Barley pathogenesis-related proteins with fungal cell wall lytic activity inhibit the growth of yeasts.** *Plant Physiol* 1993, **103**:1277-1283.
51. De DN: **Plant cell vacuoles.** Collingwood, Australia: CSIRO Publishing 2000, 288.
52. Mehta PA, Sivaprakash K, Parani M, Venkataraman G, Paida AK: **Generation and analysis of expressed sequence tags from the salt-tolerant mangrove species *Avicennia marina* (Forsk) Vierh.** *Theor Appl Genet* 2005, **100**:416-424.
53. Kota R, Varshney RK, Thiel T, Dehmer KJ, Graner A: **Generation and comparison of EST-derived SSRs and SNPs in barley (*Hordeum vulagre* L.).** *Hereditas* 2001, **135**:145-151.
54. Gupta PK, Varshney RK: **The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat.** *Euphytica* 2000, **113**:163-185.
55. Shultz JL, Kazi S, Bashir R, Afzal JA, Lightfoot DA: **The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean.** *Theor Appl Genet* 2007, **114**:1081-1090.
56. Varshney RK, Thiel T, Stein N, Langridge P, Graner A: ***In silico* analysis of frequency and distribution of microsatellites in ESTs of some cereal species.** *Cell Mol Biol Lett* 2002, **7**:537-546.
57. Luo M, Dang P, Guo BZ, He G, Holbrook CC, Bausher MG, Lee RD: **Generation of expressed sequence tags (ESTs) for gene discovery and marker development in cultivated peanut.** *Crop Sci* 2005, **45**:346-353.
58. Yu JK, Sun Q, Rota ML, Edwards H, Tefera H, Sorrells ME: **Expressed sequence tag analysis in tef (*Eragrostis tef* (Zucc) Trotter).** *Genome* 2006, **49**:365-372.
59. Quilang J, Wang S, Li P, Abernathy J, Peatman E, Wang Y, Wang L, Shi Y, Wallace R, Guo X, Liu Z: **Generation and analysis of ESTs from the eastern oyster, *Crassostrea virginica* Gmelin and identification of microsatellite and SNP markers.** *BMC Genomics* 2007, **8**:157.
60. Scott KD, Eggler P, Seaton G, Rossetto M, Ablett EM, Lee LS, Henry RJ: **Analysis of SSRs derived from grape ESTs.** *Theor Appl Genet* 2000, **100**:723-726.
61. Sivaramakrishnan S, Seetha K, Reddy LJ: **Diversity in selected wild and cultivated species of pigeonpea using RFLP of mtDNA.** *Euphytica* 2002, **125**:21-28.
62. Kota R, Varshney RK, Prasad M, Zhang H, Stein N, Graner A: **EST-derived single nucleotide polymorphism markers for assembling genetic and physical maps of the barley genome.** *Funct Integr Genomic* 2008, **8**:223-233.
63. Varshney RK, Dubey A: **Novel genomic tools and modern genetic and breeding approaches for crop improvement.** *J Plant Biochem Biotechnol* 2009, **18**:127-138.
64. Rafalski A: **Applications of single nucleotide polymorphisms in crop genetics.** *Curr Opin Plant Biol* 2002, **5**:94-100.
65. Moreno-Vazquez S, Ochoa OE, Faber N, Chao S, Jacobs JM, Maison-Neuve B, Kesseli RV, Michelimore RW: **SNP-based codominant markers for a recessive gene conferring resistance to corky root rot (*Rhizomonas suberifaciens*) in lettuce (*Lactuca sativa*).** *Genome* 2003, **46**:1059-1069.
66. Varshney RK, Nayak SN, May GD, Jackson SA: **Next-generation sequencing technologies and their implications for crop genetics and breeding.** *Trends Biotechnol* 2009, **27**:522-530.
67. Wang S, Sha Z, Sonstegard TS, Liu H, Xu P, Somridhivej B, Peatman E, Kucuktas H, Liu Z: **Quality assessment parameters for EST-derived SNPs from catfish.** *BMC Genomics* 2008, **9**:450.
68. Schmitt ME, Brown TA, Trumpower BL: **A rapid and simple method for preparation of RNA from *Saccharomyces cerevisiae*.** *Nucl Acids Res* 1990, **18**:3091-3092.
69. Sambrook J, Fritsch EF, Maniatis T: **Molecular cloning.** Cold Spring Harbor Laboratory Press, Plainview 1989, **I-III.**
70. **EST trimmer.** [http://pgrc.ipk-gatersleben.de/misa/download/est_trimmer.pl].
71. Huang X, Madan A: **CAP3: a DNA sequence assembly program.** *Genome Res* 1999, **9**:868-877.
72. **NCBI EST database.** [http://www.ncbi.nlm.nih.gov/dbEST].
73. **IDEG.6 analysis tool.** [http://telethon.bio.unipd.it/bioinfo/IDEG6_form].
74. Romualdi C, Bortoluzzi S, D'alessi F, Danieli GA: **IDEG6: a web tool for detection of differentially expressed genes in multiple tag sampling experiments.** *Physiol Genomics* 2003, **12**:159-162.
75. **Hierarchical Clustering Explorer 2.0.** [http://www.cs.umd.edu/hcil/hce/hce2.html].
76. **SSR identification tool.** [http://pgrc.ipk-gatersleben.de/misa/].
77. Rozen S, Skaletsky HJ: **Primer3 on the WWW for general users and for biologist programmers.** *Bioinformatics Methods and Protocols: Methods in Molecular Biology* Humana Press, Totowa, NJKrawetz S, Misener S 2000, 365-386[http://fokker.wi.mit.edu/primer3/], Source code.
78. Nei M, Li WH: **Mathematical model for studying genetic variation in terms of restriction endonucleases.** *Proc Natl Acad Sci USA* 1979, **76**:5269-5273.
79. Rohlf FJ: **NTSYS-pc.** *Numerical taxonomy and multivariate analysis system, Version 2.10* Exeter Software, New York 2002.
80. Jayashree B, Hanspal MS, Srinivasan R, Vigneshwaran R, Varshney RK, Sputhi N, Eshwar K, Ramesh N, Chandra S, Hoisngton DA: **An integrated pipeline of open source software adapted for multi-CPU architectures: use in the large-scale identification of single nucleotide polymorphisms.** *Comp Funct Genomics* 2007, Article ID 35604.
81. Jayashree B, Bhanuprakash A, Jami A, Reddy SP, Nayak S, Varshney RK: **Perl module and PISE wrappers for the integrated analysis of sequence data and SNP features.** *BMC Res Notes* 2009, **2**:92.

82. **Restriction enzyme data base.** [http://rebase.neb.com/].

doi:10.1186/1471-2229-10-45
Cite this article as: Raju *et al.*: The first set of EST resource for gene discovery and marker development in pigeonpea (*Cajanus cajan* L.). *BMC Plant Biology* 2010 **10**:45.