nature biotechnology

ARTICLES

Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers

Rajeev K Varshney[1,2], Wenbin Chen[3], Yupeng Li[4], Arvind K Bharti[5], Rachit K Saxena[1], Jessica A Schlueter[6], Mark T A Donoghue[7], Sarwar Azam[1], Guangyi Fan[3], Adam M Whaley[6], Andrew D Farmer[5], Jaime Sheridan[6], Aiko Iwata[4], Reetu Tuteja[1,7], R Varma Penmetsa[8], Wei Wu[9], Hari D Upadhyaya[1], Shiaw-Pyng Yang[9], Trushar Shah[1], K B Saxena[1], Todd Michael[9], W Richard McCombie[10], Bicheng Yang[3], Gengyun Zhang[3], Huanming Yang[3], Jun Wang[3,11], Charles Spillane[7], Douglas R Cook[8], Gregory D May[5], Xun Xu[3,12] & Scott A Jackson[4]

Pigeonpea is an important legume food crop grown primarily by smallholder farmers in many semi-arid tropical regions of the world. We used the Illumina next-generation sequencing platform to generate 237.2 Gb of sequence, which along with Sanger-based bacterial artificial chromosome end sequences and a genetic map, we assembled into scaffolds representing 72.7% (605.78 Mb) of the 833.07 Mb pigeonpea genome. Genome analysis predicted 48,680 genes for pigeonpea and also showed the potential role that certain gene families, for example, drought tolerance–related genes, have played throughout the domestication of pigeonpea and the evolution of its ancestors. Although we found a few segmental duplication events, we did not observe the recent genome-wide duplication events observed in soybean. This reference genome sequence will facilitate the identification of the genetic basis of agronomically important traits, and accelerate the development of improved pigeonpea varieties that could improve food security in many developing countries.

Pigeonpea (*Cajanus cajan* L.), a diploid legume crop species ($2n = 2x = 22$), is a member of the tribe *Phaseoleae*. This tribe is located in the millettioid (tropical) clade within the subfamily *Papilionoideae*, which includes many important legume crop species such as soybean (*Glycine max*), cowpea (*Vigna ungiculata*), common bean (*Phaseolus vulgaris*) and mung bean (*Vigna radiata*). The sister galegoid (temperate) clade also contains many important legume crops such as alfalfa (*Medicago sativa*), chickpea (*Cicer arietinum*), clover (*Trifolium* spp.), pea (*Pisum sativum*), lentil (*Lens culinaris*) as well as barrel medic (*Medicago truncatula*) and lotus (*Lotus japonicus*). The last two have emerged as important model species for understanding legume genomics[1].

Pigeonpea is grown on ~5 million hectares (ha), making it the sixth most important legume food crop globally. Domesticated >3,500 years ago in India[2–4], it is the main protein source for more than a billion people in the developing world and a cash crop that supports the livelihoods of millions of resource-poor farmers in Asia, Africa, South America, Central America and the Caribbean[5]. In the developing world, protein is often only available at levels less than one-third of minimum dietary requirements[6], and without improvements in agricultural productivity, this challenge is likely to worsen due to increases in human population and crop yield stagnation. From a food security perspective, legumes provide a highly balanced and nutritious source of calories and protein that is not provided by cereals, especially those commonly grown in semi-arid regions.

Owing to biotic and abiotic stresses, and the fact that pigeonpea is grown in low-input and risk-prone marginal environments, there is a large gap between potential yield (2,500 Kg/ha) and yields obtained on farmer's fields (866.2 kg/ha in Asia and 736.2 kg/ha in Africa)[5]. Together, limited genomic resources and low levels of genetic diversity in the primary gene pool have constrained genetic improvement of pigeonpea[7]. It is one of a range of orphan (or neglected) crops that have not benefited from intensive scientific research despite their importance for regional food security in the world's poorest regions.

To accelerate the application of genomics to improve yield and quality, we generated and analyzed a draft genome sequence for the pigeonpea genotype ICPL 87119, popularly known as Asha (meaning hope in Hindi). This is an inbred line and a widely cultivated medium duration Indian variety resistant to several important diseases (*Fusarium* wilt (FW) and sterility mosaic disease (SMD)), for which a number of genetic and genomic resources have been recently developed[8]. This is the first draft genome sequence for a grain legume as well as the first for an orphan legume crop and probably the first for a nonindustrial crop. It will help to increase the

[1]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India. [2]CGIAR Generation Challenge Programme (GCP), c/o CIMMYT, Mexico DF, Mexico. [3]Beijing Genomics Institute (BGI)-Shenzhen, Shenzhen, China. [4]University of Georgia, Athens, Georgia, USA. [5]National Center for Genome Resources (NCGR), Santa Fe, New Mexico, USA. [6]University of North Carolina, Charlotte, North Carolina, USA. [7]National University of Ireland Galway (NUIG), Botany and Plant Science, Galway, Ireland. [8]University of California, Davis, California, USA. [9]Monsanto Company, Creve Coeur, Missouri, USA. [10]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [11]Department of Biology, University of Copenhagen, Denmark. [12]BGI-Americas, Cambridge, Massachusetts, USA. Correspondence should be addressed to R.K.V. (r.k.varshney@cgiar.org).

Received 19 July; accepted 3 October; published online 6 November 2011; doi:10.1038/nbt.2022

Table 1 Assembly and annotation statistics for the pigeonpea genome

	All scaffolds	Scaffolds longer than 2 kb
Number of scaffolds	137,542	6,534
Total span	605.78 Mb	578 Mb
N50 (scaffolds)	516.06 kb	585 kb
Longest scaffold (pseudomolecule)	48.97 Mb	48.97 Mb
Number of contigs	173,708	35,854
Longest contig	185.39 kb	185.39 kb
N50 (contigs)	21.95 kb	23.1 kb
GC content	32.8%	32.7%
Number of gene models	48,680	
Number of gene models (non-TE containing)	40,071	
Mean transcript length	2,348.70 bp	
Mean coding sequence length	959.35 bp	
Mean number of exons per gene	3.59	
Mean exon length	267.39 bp	
Mean intron length	536.89 bp	
Number of genes annotated	46,750 (96.04%)	
Number of genes unannotated	1,930 (3.96%)	
Number of miRNA genes	862	
Mean length of miRNA genes	106.92 bp	
miRNA genes share in genome	0.0152%	
Number of rRNA fragments	329	
Mean length of rRNA fragments	129.59 bp	
rRNA fragments share in genome	0.0070%	
Number of tRNA genes	763	
Mean length of tRNA genes	75.18 bp	
tRNA genes share in genome	0.0095%	
Number of snRNA genes	363	
Mean length of snRNA genes	114.02 bp	
snRNA genes share in genome	0.0068%	
Total size of transposable elements (TEs)	313,027,948 bp	
TEs share in genome	51.67%	

TE, transposable element.

efficiency of pigeonpea improvement by integrating biotechnological tools in conventional breeding and the use of genome information of pigeonpea in other legume species.

RESULTS

Sequencing and assembly

We used the Illumina GA and HiSeq 2000 Sequencing system to sequence 11 small-insert (180–800 bp) and 11 large-insert (2–20 kb) libraries. This generated a total of 237.2 Gb of paired-end reads, ranging from 50–100 bp (**Supplementary Table 1**). Filtering and correction of the sequence data for very small and/or bad-quality sequences yielded 130.7 Gb of high-quality sequence, ~163.4× coverage of the pigeonpea genome. Analysis of sequence data for GC content indicated a similar GC content distribution in the genomes of pigeonpea and soybean (**Supplementary Fig. 1**). Additionally, a set of 88,860 bacterial artificial chromosome (BAC) end sequences were generated using Sanger sequencing from two BAC libraries (69,120 clones) by using the HindIII (34,560 clones) and BamHI (34,560 clones) restriction enzymes.

We used the assembler SOAPdenovo[9] to assemble 605.78 Mb of the pigeonpea genome *de novo*, generating a sequence with a contig N50 of 21.95 kb, and longest contig length of 185.39 kb. We then improved the assembly by using both the paired-BAC end sequences (41,302) that passed after filtering through RepeatMasker, and a genetic map comprising 833 marker loci (**Supplementary Table 2**). This increased N50 to 516.06 kb (longest scaffold in chromosome level of 48.97 Mb) (**Table 1**). Our draft genome assembly has <5.69% (~34 Mb) unclosed gaps[10]. These analyses showed that mapped genetic loci provide additional information for assembling superscaffolds, especially in regions in which scaffolds were not large enough to cross the repeat rich regions (**Supplementary Fig. 2**). The generated chromosome-scale scaffolds can be considered as 'pseudomolecules' (**Supplementary Table 3**). We estimated the pigeonpea genome size, based on K-mer statistics, to be 833.07 Mb (**Supplementary Table 4** and **Supplementary Fig. 3**), suggesting that the assembly captures 72.7% of the genome in the genome scaffolds. If only the 6,534 scaffolds >2 kb are considered, the assembly spans 578 Mb with an N50 of 0.58 Mb (**Table 1**).

Analysis of our genome assembly against the database of bacterial genomes using Megablast showed no contamination of any bacterial contig in the genome assembly. This was expected, as the GC-depth graph and distribution analysis of the genome data sets used for assembly did not show any characteristic feature of microbial genomes. In terms of checking the assembly for organellar DNA contamination, analysis of the soybean chloroplast DNA against the genome assembly showed hits with only 36 of 6,534 scaffolds (>2 kb). The longest one is about 50 kb in the scaffold 000124. This observation is also not unexpected, as long stretches of chloroplast and mitochondrial DNA have been shown to be transferred into nuclear chromosomes of several plant and animal species during evolution[11].

The transcriptome assembly (CcTAv2.0, http://cajca.comparative-legumes.org/data/lista_cajca-201012.tgz) composed of 21,434 contigs, referred to as transcriptome assembly contigs (TACs), was mapped to the draft genome assembly. Of the 21,434 TACs, 97% of the total length of them could be mapped to the genome assembly with >90% sequence identity. We found 94.78% of TACs in the genome assembly at >90% identity and >50% coverage of query length. Using more stringent criteria (>90% identity and >90% of the coverage), 88.53% of TACs were captured (**Supplementary Table 5**). These results indicate an extremely low proportion of misassemblies, at least in the gene-rich regions. The mapping of the CcTAv2.0 was only for quality control purposes. The transcriptome assembly does not necessarily represent all pigeonpea genes; it is restricted to the tissue types used and genes expressed at low levels are likely to be under-represented.

Repetitive sequences

De novo repeat identification using RepeatModeler and homology analysis against the RepBase library identified repetitive DNA (excluding low-complexity sequences) in 51.67% of the genome, most of which could not be associated with known transposable element (TE) families. The fraction of repetitive sequences in the genome is comparable to other genomes, like those of soybean (59%)[12], castor bean (50%)[13] and grapevine (41%)[14], but less than that seen in the genomes of maize (85%)[15] and sorghum (62%)[16]. Classification of the observed transposable elements into known classes revealed that the majority of repetitive sequences were retrotransposons (37.12%), whereas 8.77% of the transposable elements were DNA transposons (**Table 2**). Like the soybean[12] and castor bean[13] genomes, the most abundant repeats identified are long-terminal repeat elements, of which 22.81% are *Gypsy*-type elements and 12.04% are *Copia*-type elements (**Table 2**).

Gene annotation

We used a combination of *de novo* gene prediction programs and homology-based methods to predict gene models in the pigeonpea genome. These were combined using the GLEAN algorithm[17], resulting in the identification of 48,680 genes with an average transcript length of 2,348.70 bp, coding sequence size of 959.35 bp and 3.59 exons per gene (**Supplementary Table 6**). The majority of these predicted genes (99.6%) were supported either by *de novo* gene prediction, expressed

Table 2 Repetitive sequences in the pigeonpea genome

	Length occupied (bp)	Total repeats (%)	Genome (%)
Retrotransposons	116,194,477	37.12	19.18
Gypsy	71,402,096	22.81	11.79
Copia	37,676,825	12.04	6.22
Line	6,717,918	2.15	1.11
Sine	375,342	0.12	0.06
Other	22,296	0.01	0.00
Unclassified elements	169,378,278	54.11	27.96
DNA transposons	27,455,193	8.77	4.53
Total transposable elements	313,027,948	–	51.67
Low complexity sequences	2,807,079	–	0.46

sequence tags (EST)/unigenes or homology-based searching, or a combination of these approaches (**Supplementary Fig. 4**). To further validate the gene predictions, we used the predicted pigeonpea gene set to search KOGs, the core genes from the core eukaryotic gene mapping approach (CEGMA) pipeline[18]. The presence of 453 out of 458 (98.9%) KOGs within the pigeonpea gene set (**Supplementary Table 7**) confirms that annotation of the pigeonpea genome is close to being complete, although the number of genes may be inflated owing to the breaking of genes onto separate contings during the assembly process.

When compared to other sequenced plant genomes, such as those from cucumber (26,682)[19], cacao (28,798)[20], grapevine (29,585)[14] and *L. japonicus* (38,483)[21], the number of predicted genes in the pigeonpea genome is higher, but comparable to poplar (45,555)[22], soybean (46,430)[12] and *M. truncatula* (47,529) (Nevin Young, University of Minnesota, personal communication). Comparison of the features of pigeonpea genes with those of other dicot genomes indicates that they have similar characteristics, e.g., the size of mRNAs, coding sequences, exons and introns. However, the average number of exons per gene in the pigeonpea (3.59) is less than for soybean (5.80), whereas average exon (267.39 bp) and intron (536.89 bp) lengths are longer than those for soybean (216.13 bp exons and 419.43 bp introns) (**Supplementary Table 8**).

All predicted genes were functionally annotated following a consensus approach of either known homologous or predictive sequence signatures using Swissprot, GO, TrEMBL[23], InterPro[24] and KEGG[25]. The largest number of genes showed homology with proteins in TrEMBL (95.77%) followed by those in the InterPro (70.01%) database. In total, 46,750 (96.04%) genes had sufficient similarity to entries in databases to tentatively assign gene functions. Only 1,930 (3.96%) genes remain unannotated (**Supplementary Table 9**). In addition to protein-coding genes, we have identified 862 microRNA (miRNA), 763 tRNA, 329 rRNA and 363 small nuclear (snRNA) genes in the pigeonpea genome set (**Supplementary Table 10**). It is important to mention that rRNA genes in pigeonpea genome were predicted by aligning the 5.8S, 18S and 25S rRNA of *Arabidopsis* and 28S rRNA of rice against the pigeonpea genome assembly using BLASTN. We also determined the rDNA loci cytogenetically in the pigeonpea genome using fluorescence *in situ* hybridization (FISH) (**Supplementary Fig. 5**).

Synteny with sequenced plant genomes

The *Papilionoideae* subfamily, which contains pigeonpea and several other crop legumes, diverged into two major subgroups, the millettioid and galegoid clades, ~54 million years (Myr) ago[26]. Within the millettioid clade, pigeonpea diverged from soybean ~20–30 Myr ago. In spite of this long period of divergence, high levels of synteny are observed between the millettioid species pigeonpea and soybean (**Fig. 1** and **Supplementary Fig. 6**) as well as between pigeonpea and the galegoid species *M. truncatula* (**Supplementary Fig. 7**) and *L. japonicus* (**Supplementary Fig. 8**). Details of these syntenic blocks are provided in **Supplementary Table 11.** Each pigeonpea chromosome shows extensive synteny with two or more than two chromosomes in soybean, likely due to the independent duplication event in soybean[12] following divergence from pigeonpea. Even with this duplication event, the level of synteny and the blocks themselves are prominent (**Supplementary Fig. 9**).

Absence of recent genome duplication

We analyzed gene content and gene order, using the i-ADHoRe[27] tool to find syntenic blocks by identifying successive pairs of duplicated genes. Using the same parameter set as used in the soybean genome[12], we identified a total of 28 duplicated syntenic blocks in the genome (**Supplementary Fig. 10**). The number of homologs (gene pairs) within a block averaged 6, with a range from 4 to 18. Interestingly, chromosome 11 seems to be highly fragmented with block matches to five other chromosomes (CcLG02, CcLG03, CcLG06, CcLG08 and CcLG10). We also found three intrachromosomal duplications, two of which were located in CcLG06, with the third present in the chromosome CcLG11. Although chromosome number can be reduced after polyploidization, the chromosome number of pigeonpea ($2n = 22$) relative to other diploid legumes, also supports the lack of a genome duplication event in pigeonpea (**Supplementary Fig. 11**). This contrasts with soybean, the only other member of the *Phaseoleae* for which a genome sequence is available[12]. Comparison of the pigeonpea and soybean genomes (**Fig. 1** and **Supplementary Fig. 6**) confirms the recent whole-genome duplication ~13 Myr ago in the soybean genome[12], which is missing in the pigeonpea genome.

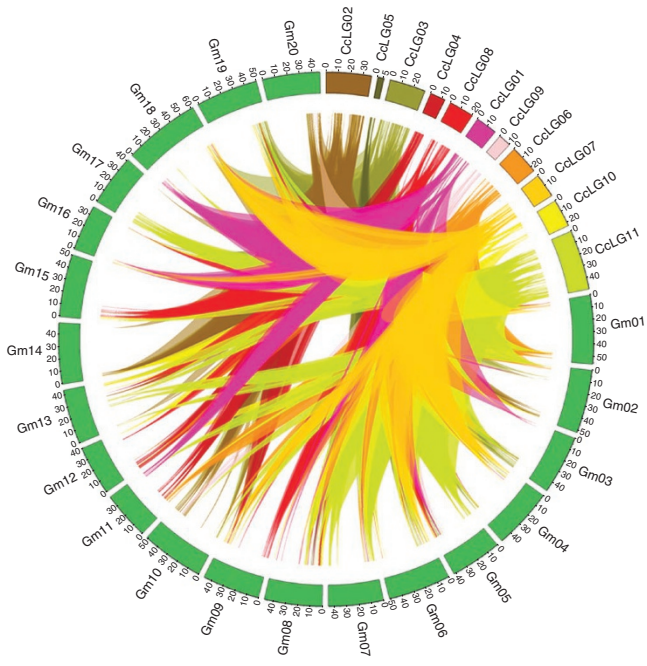

Figure 1 Extensive synteny between the pigeonpea and soybean genomes. Soybean pseudomolecules, labeled as Gm, are represented as green boxes. Numbers along each chromosome box are sequence length in megabases. Pigeonpea pseudomolecules, labeled as CcLG, are shown with each chromosome as a different color. Syntenic blocks were identified through reciprocal best matches between gene models and block identification using i-ADHoRe. Each line radiating from a pigeonpea pseudomolecule represents a gene match found in a block between soybean and pigeonpea.

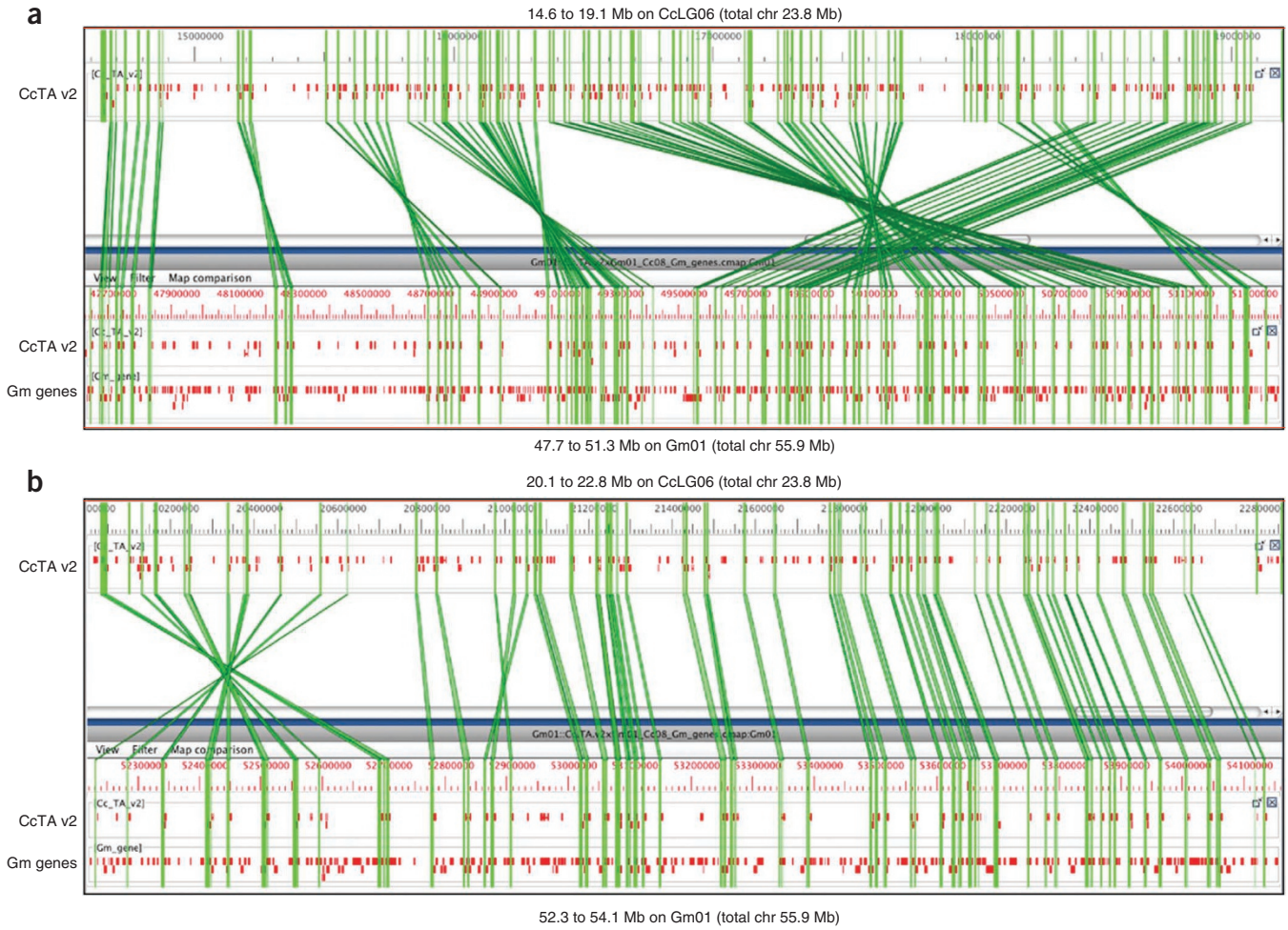

Figure 2 Microsynteny analysis between pigeonpea and soybean genomes. One chromosome arm of soybean chromosome 01S (south arm) and pigeonpea CcLG06 (indicated as a green circle in the whole-genome dot-plot in **Supplementary Fig. 6**) is shown here as a representation of microsynteny. Mapping of the pigeonpea transcriptome assembly contigs (TACs) of the pigeonpea transcriptome assembly (CcTA v2) onto both genomes (indicated by green lines) was used as a measure of conserved gene order. (**a**) The first part shows local rearrangements. (**b**) The later part indicates very good collinearity among genes in the two genomes.

Moreover, no major duplications (similar to the duplication of chromosomes 5–8) found in *M. truncatula* (Nevin Young, personal communication) were retained from the 58 Myr ago whole-genome duplication that occurred in all members of *Papilionoideae*. These observations indicate that there was most likely a period of extensive rearrangements after the whole-genome duplication 58 Myr ago, which stabilized before the split between the millettioid and galegoid clades 54 Myr ago, and that some of these rearrangements are lineage specific. However, as expected, local rearrangements have occurred during the course of evolution. For instance, this is seen in the microsynteny between chromosome CcLG06 of pigeonpea and chromosome 1 of soybean (**Fig. 2**).

Comparisons of gene families among eudicots

We used several strategies to identify pigeonpea-specific gene families. For instance, we used protein sequence similarities to cluster gene families from all of the sequenced legume genomes (*M. truncatula*, soybean, *L. japonicus* and pigeonpea), using grapevine as an out-group (nonlegume) species. This revealed 4,311 clusters of genes, containing 72,193 genes that are common to all five eudicot genomes, 903 clusters containing 7,513 genes found only in the four legume genomes, 1,024 clusters specific to soybean and pigeonpea, and 3,068 gene families with 15,076 genes specific to the pigeonpea genome (**Fig. 3**). About 31% of pigeonpea genes are specific to pigeonpea. However, because a majority of these encode hypothetical proteins supported only at the transcript level (**Supplementary Table 12**), this is probably an overestimate. In fact, InterPro annotation of the 15,076 genes revealed that 6,714 are of unknown function, whereas 4,788 are transposon related (reverse transcriptase, retrotransposon gag protein, transposon and transposase) and the remaining 3,574 have other functions. Analysis of the proteomes for the five eudicot species using Gene Ontology (GO) and InterPro terms revealed differences between pigeonpea and the other four species (**Supplementary Fig. 12**). Some of these differences may reflect the evolutionary history of pigeonpea or its adaptations to specific environments.

ORFan genes

The previous section identifies a large number of genes (15,076) that could not be grouped into gene families with genes from the other three sequenced legume genomes (or the grapevine genome) and therefore occupy pigeonpea-specific gene families (some of which could be ORFan genes). ORFan genes are protein-encoding genes

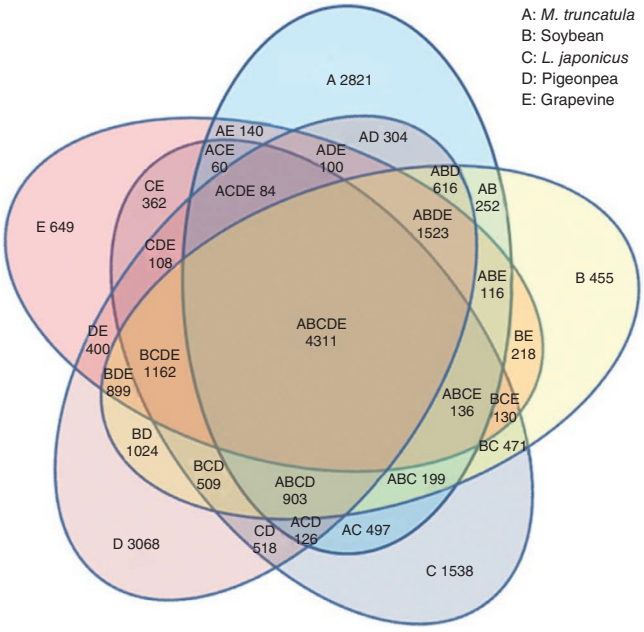

Figure 3 Distribution of gene families among five eudicot genomes (*M. truncatula*, soybean, *L. japonicus*, pigeonpea and grapevine). Homologous genes in pigeonpea, soybean, *M. truncatula*, *L. japonicus* and grapevine were clustered to gene families. The numbers of gene families are indicated for each species and species intersection.

that have no significant sequence similarity to any other proteins and/or peptides in the genome or protein databases outside the taxon of interest[28–31]. ORFan genes may represent lineage-specific adaptations (or innovations), such as in stress responses, as highlighted as a key feature of ORFans in both rice[32] and *Arabidopsis*[31]. To identify pigeonpea ORFan genes, we used a BLAST filtering approach involving sequence data from all available taxa against the 48,680 pigeonpea genes to identify 266 gene models restricted to the tribe *Phaseoleae*. Of the 266 *Phaseoleae*-restricted ORFans, 97 ORFans had significant sequence similarity ($e < 0.0001$) to peptides from soybean, leaving 169 putative pigeonpea-specific ORFan genes. The *Phaseoleae*-restricted ORFans in pigeonpea display many of the characteristics of ORFans identified in other species[33], namely short length, few introns and unusual GC content (**Supplementary Table 13**).

To determine the evolutionary origin of the pigeonpea ORFans, we identified significant hits to conserved (non-ORFan) pigeonpea genes by performing all-against-all BLASTP searches ($e < 0.0001$). Of the 266 ORFans, 202 have at least one significant hit to a non-ORFan gene with a mean percentage coverage of alignment of 84.38 ± 24.3541. To identify sequence matches to non-ORFans in different reading frames, we performed all-against-all BLASTN searches ($e < 0.0001$) on coding sequences. This identified an additional 11 ORFans with at least one significant sequence match, representing out-of-frame hits with mean percentage coverage of alignment of 87.6 ± 22.6854. These data indicate that the vast majority of ORFans (213) in the pigeonpea genome are duplicates, which have evolved by a duplication-divergence model[33]. Of the remaining 53 ORFans, 11 have evolved due to frameshifts generating novel open reading frames (as identified by genes with coding sequence hits to soybean) and two ORFans originated either *de novo* or as a result of gene loss (as identified by intergenic hits in soybean). The remaining 40 ORFans have no identifiable sequence similarity to any sequence tested, making it impossible to discern their evolutionary origins.

Large-scale identification of genetic markers

Narrow genetic diversity, coupled with limited genomic resources, has been a major bottleneck for applying molecular plant breeding for improvement of pigeonpea. Simple sequence repeat (SSR) and single-nucleotide polymorphism (SNP) markers are currently the markers of choice for plant breeding. Analysis of the pigeonpea genome provided a total of 309,052 SSRs (**Supplementary Table 14**). For designing SSR primers, 29,467 sequences containing tri-, tetra-, penta-, hexa- or compound repeat units, which are generally polymorphic, were considered and 23,410 primer pairs were successfully designed that can be converted into genetic markers (**Supplementary Table 15**).

The genome assembly was also used to align 128.9 million Illumina transcript reads from 12 different pigeonpea genotypes that are parents of 6 mapping populations for identification of SNPs[34]. By aligning the transcript reads for a given parental combination onto the pigeonpea genome assembly, we identified sequence variants that differ between the parental combinations. The number of SNPs between two parental genotypes ranged from 2,164 (BSMR 736 × TAT 10) to 16,651 (ICP 28 × ICPL 87091) (**Supplementary Table 16**). In total, we identified 28,104 novel SNPs across the 12 genotypes (**Supplementary Table 17**). These SNPs can be used for germplasm characterization and for genetic improvement. It is important to mention that SNP calling between the high-quality sequence reads and the draft genome assembly showed a heterozygosity rate of only 0.067% in the Asha genome (**Supplementary Table 18**), confirming the inbred nature of the Asha accession that was used for genome sequencing.

DISCUSSION

Recently, draft genome sequences have become available for two model legume species, *M. truncatula* (Nevin Young, personal communication) and *L. japonicus*[21], and one industrial legume crop, soybean[12]. This report presents the genome of the first orphan legume crop and the second food legume (after soybean). Pigeonpea plays a substantial role in the livelihood of resource-poor smallholder farmers in marginal environments. Fungal diseases (e.g., FW), viral diseases (e.g., SMD) and insect pests (e.g., *Helicoverpa armigera* (pod borer)), together with abiotic stresses such as salinity and water logging, have limited the yield of pigeonpea to about one-third of the potential yield. Moreover, pigeonpea crop productivity has remained stagnant for the last 50 years. Low genetic diversity, coupled with availability of only a few hundred useful markers, has hampered the development of intraspecific genetic maps for identification of markers associated with quantitative trait loci for resistance and/or tolerance to these yield drags. As a result, pigeonpea breeders have not been able to increase the varietal yields. However, pigeonpea is the first legume to have hybrid varieties released based on cytoplasmic-nuclear male sterility[35]. Improvement of parental lines for biotic and abiotic resistance stresses as well as maintaining the purity of hybrid seeds is critical for sustainable hybrid production.

The availability of a draft genome sequence opens new avenues for pigeonpea improvement. In the short-term, the genome sequence will usher the pigeonpea crop into the molecular breeding era by deploying the SSR and SNP markers identified for genetic mapping and trait identification. Breeding approaches such as marker-assisted recurrent selection and genomic selection will now be feasible for pigeonpea breeding, and may be even further advanced by genotyping by sequencing that can be done with the help of the draft genome sequence. In the long term, with the help of low-cost sequencing technologies or approaches, the draft genome sequence will facilitate understanding of the genetic basis of many traits at genome level and allow the undertaking of genome-wide association studies involving

thousands of pigeonpea accessions (13,632 accessions in ICRISAT genebank)[7]. It will lead to the identification and manipulation of candidate genes or genomic regions to enable breeding of varieties or hybrids resistant or tolerant to biotic and abiotic stresses as well as global climate fluctuations[36].

For example, one of the most attractive features of pigeonpea relative to other legume crops is its tolerance of drought stress. As a preliminary screen to begin to understand the genetic basis of this drought tolerance, we analyzed 511 universal drought-responsive protein sequences from the *Viridiplantae*[37] in the sequenced legume genomes. Of the 511 proteins, 427 had homologous sequences in the legumes. Pigeonpea had a higher number (111) than either *M. truncatula* (90; Nevin Young, personal communication) or *L. japonicus* (58) (data not shown). It had a similar number of drought-responsive genes to the number found in soybean (109), but soybean underwent a recent genome duplication event. These genes need to be confirmed experimentally but are, nonetheless, candidates that can be used to begin to gain insight into the genetic architecture of pigeonpea's drought tolerance and for screens to identify superior haplotypes for improvement (**Supplementary Table 19**).

In addition to bioinformatics-based comparative and predictive approaches at the genomic level, differential gene expression studies through RNA-Seq will facilitate the identification of candidate genes for biotic and abiotic stresses, for example, resistance to *Helicoverpa armigera*, SMD and FW. These candidate genes will be critical for the improvement of pigeonpea and perhaps other crops. In brief, the availability of a pigeonpea reference genome sequence will facilitate greater integration of biotechnological tools into pigeonpea breeding efforts to minimize the yield gap in farmer's fields in Asia and Africa. The availability of pigeonpea genome sequence will also facilitate assembly and alignment of genomes of other *Phaseoloid* species, such as cowpea and common bean. These comparative approaches may allow other legumes to leverage the unique characteristics found in pigeonpea. For example, candidate genes associated with drought tolerance and the cytoplasmic-nuclear male-sterility system that are unique to pigeonpea may be used for improving other legume crops such as soybean and common bean that are adversely affected by drought stress and possibly bring hybrid seed production in other legume species.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession code. Genome assembly is available at National Center for Biotechnology Information as BioProject ID PRJNA72815 (http://www.ncbi.nlm.nih.gov/bioproject?term=PRJNA72815). Genome assembly, annotation data and all supplementary figures and tables are available for viewing and/or downloading at http://www.icrisat.org/gt-bt/iipg/Genome_Manuscript.html.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We would like to thank CGIAR Generation Challenge Programme, US National Science Foundation (DBI 0605251, BIO 0822258), BGI-Shenzhen, China, and The International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), India, for their financial contribution to this study. Thanks are also due to Indian Council of Agricultural Research (ICAR), India for financial support to some earlier work that was used for analyzing genome sequence data. We would like to thank N.K. Singh, National Research Centre on Plant Biotechnology, Indian Agricultural Research Institute, New Delhi, India, for his immense help and support in various ways while carrying out this study. Our sincere appreciation to W.D. Dar, D.A. Hoisington, C.L.L. Gowda, O. Riera-Lizarazu from ICRISAT; J.-M. Ribaut from CGIAR Generation Challenge Programme; and M. dela Bastide from Cold Spring Harbor Laboratory for their helpful advice and assistance, wherever required, during the course of the study. We also thank D. Johnson from the University of Ottawa for providing 18S rRNA gene cloned from soybean (*G. max*).

AUTHOR CONTRIBUTIONS

R.K.V., W.C., R.K.S., G.F., R.V.P., H.D.U., K.B.S., W.R.McC., B.Y., G.Z., D.R.C., G.D.M., X.X., contributed to generation of genome sequence, transcriptome sequence and genetic mapping data; W.C., G.F., R.T., W.W., S.-P.Y., T.M., W.R.McC., G.Z., H.Y., J.W., X.X., worked on genome assembly; W.C., Y.L., A.K.B., R.K.S., S.A., A.D.F., H.Y., J.W., X.X., contributed to genome annotation and gene function; R.K.V., W.C., Y.L., A.K.B., R.K.S., J.A.S., J.S., A.I., M.T.A.D., A.M.W., A.D.F., J.S., R.T., T.S., C.S., D.R.C., G.D.M., X.X., S.A.J., worked on genome analysis and comparative genomics and R.K.V., together with S.A.J., D.R.C., C.S., W.C., A.K.B., R.K.S., S.A., J.A.S., wrote and finalized the manuscript. R.K.V. conceived and directed the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/nbt/index.html.

1. Cannon, S.B., May, G.D. & Jackson, S.A. Three sequenced legume genomes and many crop species: rich opportunities for translational genomics. *Plant Physiol.* **151**, 970–977 (2009).
2. Vavilov, N.I. The origin, variation, immunity, and breeding of cultivated plants. *Chron. Bot.* **13**, 1–366 (1951).
3. De, D.N. Pigeonpea. in *Evolutionary Studies in World Crops: Diversity and Change in the Indian Subcontinent* (ed., Hutchinson, J.). 79–87 (Cambridge University Press, London, 1974).
4. Royes, W.V. Pigeonpea. in *Evolution of Crop Plants* (ed., Sommonds, N.W.). 154–156 (Longmans, London and New York, 1976).
5. Mula, M.G. & Saxena, K.B. *Lifting the Level of Awareness on Pigeonpea—a Global Perspective* (International Crops Research Institute for the Semi-Arid Tropics, 2010).
6. Latham, M.C. Human nutrition in the developing world. FAO Food and Nutrition Series No. 29 (UN Food and Agriculture Organization, 1997) ⟨http://www.fao.org/DOCREP/W0073e/w0073e05.htm⟩.
7. Bohra, A. *et al.* Harnessing the potential of crop wild relatives through genomics tools for pigeonpea improvement. *J. Plant Biol.* **37**, 85–100 (2010).
8. Varshney, R.K. *et al.* Pigeonpea genomics initiative (PGI): an international effort to improve crop productivity of pigeonpea (*Cajanus cajan* L.). *Mol. Breed.* **26**, 393–408 (2010).
9. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
10. Li, R. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
11. Timmis, J.N. *et al.* Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* **5**, 123–135 (2004).
12. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
13. Chan, A.P. *et al.* Draft genome sequence of the oilseed species *Ricinus communis*. *Nat. Biotechnol.* **28**, 951–956 (2010).
14. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
15. Schnable, P.S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
16. Paterson, A.H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
17. Elsik, C.G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).
18. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
19. Huang, S. *et al.* The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* **41**, 1275–1281 (2009).
20. Argout, X. *et al.* The genome of *Theobroma cacao*. *Nat. Genet.* **43**, 101–108 (2011).
21. Sato, S. *et al.* Genome structure of the legume, *Lotus japonicus*. *DNA Res.* **15**, 227–239 (2008).
22. Tuskan, G.A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006).
23. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
24. Zdobnov, E.M. & Apweiler, R. InterProScan–an integration platform for the signaturerecognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).

25. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
26. Lavin, M., Herendeen, P.S. & Wojciechowski, M.F. Evolutionary rates analysis of *Leguminosae* implicates a rapid diversification of lineages during the tertiary. *Syst. Biol.* **54**, 575–594 (2005).
27. Simillion, C., Janssens, K., Sterck, L. & van de Peer, Y. i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics* **24**, 127–128 (2008).
28. Brcic-Kostic, K. Neutral mutation as the source of genetic variation in life history traits. *Genet. Res.* **86**, 53–63 (2005).
29. Wilson, G.A. *et al.* Orphans as taxonomically restricted and ecologically important genes. *Microbiology* **151**, 2499–2501 (2005).
30. Schmid, K. & Aquadro, C. The evolutionary analysis of "orphans" from the *Drosophila* genome identifies rapidly diverging and incorrectly annotated genes. *Genetics* **159**, 589–598 (2001).
31. Donoghue, M.T. *et al.* Evolutionary origins of *Brassicaceae* specific genes in *Arabidopsis thaliana*. *BMC Evol. Biol.* **11**, 47 (2011).
32. Guo, W.J., Li, P., Ling, J. & Ye, S.P. Significant comparative characteristics between orphan and nonorphan genes in the rice (*Oryza sativa* L.) genome. *Comp. Funct. Genomics* **2007**, 21676 (2007).
33. Taylor, J.S. & Raes, J. Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.* **38**, 615–643 (2004).
34. Dubey, A. *et al.* Defining the transcriptome assembly and its use for genome dynamics and transcriptome profiling studies in pigeonpea (*Cajanus cajan* L.). *DNA Res.* **18**, 153–164 (2011).
35. Stokstad, E. The plant breeder and the pea. *Science* **316**, 196–197 (2007).
36. Varshney, R. *et al.* Agricultural biotechnology for crop improvement in a variable climate: hope or hype? *Trends Plant Sci.* **16**, 363–371 (2011).
37. Isokpehi, D.R. *et al.* Identification of drought-responsive universal stress proteins in *Viridiplantae*. *Bioinform Biol. Insights* **5**, 41–58 (2011).

ONLINE METHODS

High-molecular-weight DNA preparation. High-quality genomic DNA was prepared from seeds of the Asha variety. Seeds were grown in a dark chamber for 2 weeks before harvest. A standard phenol/chloroform method of DNA extraction was performed. The extracted DNA was treated with RNase A and proteinase K, respectively, to prevent RNA and protein contamination, and further precipitated with ethanol.

Whole-genome shotgun sequencing. We used a whole-genome shotgun sequencing strategy with Illumina Genome Analyzer sequencing technology and Hiseq 2000 Sequencing System. To get enough DNA for the library construction and sequencing, we carried out whole-genome amplification. We constructed a total of 22 paired-end sequencing libraries with insert sizes of about 180 base pairs (bp), 250 bp, 350 bp, 500 bp, 800 bp, 2 kb, 5 kb, 10 kb and 20 kb. In total, we generated 237.2 Gb data of paired-ends ranging from 50 to 100 bp. To reduce the effect of sequencing error to the assembly, we have taken a series of checking and filtering steps on reads generated[10]. Using stringent criteria, only 130.7 Gb data were considered for *de novo* genome assembly.

The genome size was calculated using the total length of sequence reads divided by sequencing depth. To estimate the sequencing depth, we counted the copy number of a certain K-mer (e.g., 17-mer) present in sequence reads, and plotted the distribution of copy numbers. The peak value of the frequency curve represents the overall sequencing depth. We used the algorithm: $(N \times (L - K + 1) - B)/D = G$, where N is the total sequence read number, L is the average length of sequence reads and K is K-mer length, defined as 17 bp here. To minimize the influence of sequencing error, K-mers with low frequency (<4) are discarded. B is the total number of low frequency 17-mer. G denotes the genome size, and D is the overall depth estimated from K-mer distribution.

We carried out the whole-genome assembly using SOAPdenovo for the remaining reads after the above filtering and correction steps[9,10]. The contigs after SOAPdenovo corrections were formed without any gap. We realigned all the usable reads onto the contig sequences and obtained aligned paired ends. We then calculated the amount of shared paired-end relationships between each pair of contigs, weighted the rate of consistent and conflicting paired ends, and then constructed the scaffolds step by step, from short insert-sized paired ends, to long insert-sized paired ends. Subsequently, 88,860 BAC end sequences were generated from two BAC libraries of Asha. All the raw paired BAC end sequences (88,860) were filtered against RepeatMasker. Subsequently, 41,302 paired-BAC end sequences that passed after filtering were used for mapping to scaffolds to obtain the super scaffolds. Finally, we used the genetic map (*C. cajan* ICP 28 × *C. scaraboides* ICPW 94) comprising 833 marker loci including 209 BAC end sequence–derived SSR markers and 624 conserved orthologous sequence–based markers based on legume transcript sequence for developing the final scaffolds or pseudomolecules. To close the gaps inside the constructed scaffolds, which were mainly composed of repeats that were masked before scaffold construction, we used the paired-end information to retrieve the read pairs that had one end mapped to the unique contig and the other located in the gap region, then did a local assembly for these collected reads.

The quality of genome assembly for microbial contamination was checked by its analysis against a database of bacterial genomes using Megablast (E-value < 1e-5, > 90% identity, > 200 bp length mapped to scaffold sequence). For checking the contamination of assembly with organellar DNA, soybean chloroplast DNA (152,218 bp) downloaded from http://www.ncbi.nlm.nih.gov/nuccore/DQ317523 was screened against the pigeonpea genome assembly.

For checking the completeness of the assembly, a transcriptome assembly comprising 21,434 unigenes, referred to as transcriptome assembly contigs (TACs) and defined based on 10,817 Sanger ESTs, 2.19 million 454/FLX transcript reads[34] and 128.91 million Illumina transcript reads[34] was used for mapping the TACs to the assembly genome with the help of BLAT software. Analysis was done at different criteria of percent sequence homology and percent coverage (**Supplementary Table 5**).

Identification of repetitive elements. There are two main types of repeats in the genome (tandem repeats and interspersed repeats). We searched the genome for tandem repeats using Tandem Repeats Finder[38] and Repbase (composed of many transposable elements) to identify the interspersed repeats. Transposable elements in the genome assembly were identified both at the DNA and protein level. For identification of transposable elements at the DNA level, RepeatMasker was applied using a custom library comprising a combination of Repbase and the *de novo* transposable element library of the pigeonpea genome. At the protein level, RepeatProteinMask, updated software in the RepeatMasker package, was used to perform RM-BlastX against the transposable elements protein database[39]. In this context, we used the software RepeatModeler to build a new repeat library based on the genome. These results were used to construct a new library for RepeatMasker and RepeatMasker was run again to find homolog repeats in the genome. Identified repeats were classified into different known classes as per standard genome analysis[12,40].

Gene prediction and annotation. To predict genes, we used three main approaches: homology-based method (H), *de novo* method (D) and EST/unigenes-based method (C). Results of these three methods were integrated by the GLEAN program[17] and then filtered multiple times and checked manually.

Protein sequences from six sequenced eudicot species, namely *Arabidopsis thaliana*, *Cucumis sativus*, *Carica papaya*, *Vitis vinifera*, *Populus trichocarpa* and *Glycine max*, were used to perform prediction, taking one species each time. We mapped them to the genome assembly using TblastN with E-value -1e-5 (ref. 41). After this, homologous genome sequences were aligned against the matching proteins using GeneWise (version 2.0)[42] for accurate spliced alignments. Subsequently, we filtered pseudogenes from the homology-search results from six data sets.

For *de novo* prediction, Augustus[43], GENSCAN[44] and GlimmerHMM[45] were used to predict genes with parameters trained on *A. thaliana*. We merged three *de novo* predictions into a unigene set. *De novo* gene models that were supported by two or more *de novo* methods were retained. For overlapping gene models, the longest one was selected and finally, we got *de novo*-based gene models (43,647).

In the third approach, we used the transcribed sequences, that is, 10,376 Sanger ESTs to align against the genome assembly using BLAT[41] to generate spliced alignments, and then filtered the overlaps to link the spliced alignments using PASA (http://www.lerner.ccf.org/moleccard/qin/pasa/). As a result, 2,246 genes were defined.

Finally, using one *de novo* set (43,647) and six homolog-based results as gene models (33,360 to 39,749), together with an EST-based gene set (2,246), integration was done using the GLEAN program[17]. Finally, we got the GLEAN gene set (referred to as G-set, 48,369).

Sixty-nine genes, with high GeneWise scores, complete ORFs, from the homology-based prediction, which were not included in the G-set but were supported by ESTs/unigenes, were added to the GLEAN result (48,438 genes, termed the GH-set). Subsequently, three *de novo* gene sets based on homology-search analysis were compared with the GH-set. In cases where there was an overlap between two or more genes in these gene sets, the longest gene was selected. Subsequently, the gene set was translated using SwissProt and EST/unigene results translated into proteins and only those genes that got support as mentioned above and did not have any overlap with GH-set were selected. At this point also, we filtered 14 genes that have a coding sequence length of <150 bp or N content >50%, the genes that have internal stop codon or frameshift. And subsequently, we got a gene set, referred to as GHD-set, comprising 48,671 genes.

For checking the completeness of the gene set defined, the core eukaryotic gene-mapping approach (CEGMA)[18] that rapidly assess genome completeness and gene structure prediction was used with the gene set defined. CEGMA analysis includes a set of 453 core genes that are supposed to be highly conserved and single-copy genes present in all eukaryotes. Based on this analysis, a set of 9 genes that did not align with any gene defined was also included in the final gene set of 48,680 genes, referred to as Official Gene Set (OGSv1.0).

Gene functions were assigned according to the best match of the alignments using BLASTP (1e-5) to SwissProt[23] and TrEMBL databases. InterProScan[24] determined motifs and domains of genes against protein databases including Pfam, PRINTS, PROSITE, ProDom and SMART[46]. Gene Ontology IDs for each gene were obtained from the corresponding InterPro entry. All genes were aligned against KEGG proteins[25], and the pathway in which the gene might be involved was derived from the matching genes in KEGG.

doi:10.1038/nbt.2022

Identification of noncoding RNA genes. The tRNA genes were predicted by tRNAscan-SE[47] with eukaryote parameters. Aligning the rRNA template sequences from plants (e.g., *Arabidopsis thaliana* and rice) using BlastN with E-value 1e-5 identified the rRNA fragments. The miRNA and snRNA genes were predicted by INFERNAL software against the Rfam database (Release 9.1).

For determining the rDNA loci cytogenetically in the pigeonpea genome, 18S rRNA gene, cloned from soybean (*G. max*), provided by D. Johnson, and 5S rRNAgene cloned from common bean (*Phaseolus vulgaris*) were used for fluorescence *in situ* hybridization (FISH). Nick translation method was used to directly label 18S rRNA gene and 5S rRNA gene with Texas red-12-dUTP and Fluorescein-12-dUTP, respectively (Invitrogen). Chromosomes were counterstained with 4′,6-diamidino-2-phenylindoleand (DAPI). Images were taken with Zeiss Axio Imager M2 microscope, equipped with AxioCamMRm, controlled by Axio Vision software. The image was adjusted for publication using Adobe Photoshop CS3 (Adobe Systems).

Construction of syntenic blocks. Whole genome dot plots were generated with pigeonpea scaffolds representing the 11 chromosomes on the *x* axis against chromosome arms of *M. truncatula* (Mt), soybean (Gm) and *L. japonicus* (Lj) on the *y* axis. The three reference genome assemblies were downloaded from http://www.medicagohapmap.org/downloads.php (*M. truncatula*), ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v7.0/Gmax/assembly (soybean) and ftp://ftp.kazusa.or.jp/pub/lotus/lotus_r2.5/pseudomolecule (Lotus), respectively. The Mt, Gm and Lj chromosomes were broken into 'North' and 'South' arms based on the estimated position of the centromeric region. In the case of soybean, the large pericentromeric regions were removed for the synteny analysis. Amino-acid based sequence alignment was carried out with the Promer package of MUMmer 3.22 (ref. 48). It looks for Maximal Unique Matches (MUMs) in all six frames as anchors for amino-acid–based alignment. Mummerplot & gnuplot 4.4 patch level 2 were used to generate whole genome dot plots of MUMs. Vmatch[49] was used to identify reciprocal best matches between the soybean and pigeonpea genomes, thereby enabling definition of syntenic blocks. The parameter set chosen was query and subject coverage of 85 and 70 (respectively) with an exdrop of 100 and a minimum length of 100. Matches were then piped into i-ADHoRe[27] to identify syntenic blocks between the two genomes.

Estimation of genome duplication. We used the Vmatch software package[49] to generate clusters of similar genes based on sequence similarity using the predicted gene models. The resulting clusters, consisting of two to six genes each, were further analyzed using the yn00 program of PAML to determine gene duplicate pairs within each cluster. These duplicate pairs were then analyzed using i-ADHoRe[27], which finds syntenic blocks by identifying successive pairs of duplicated genes. The Circos image was generated by identifying the first and last gene of each block, and placing its position in the genome using the .GFF file. To show relative block size, the Ribbon option of Circos[50] was used to draw thick lines which, at the start and end points, have a thickness that directly corresponds to the size of the duplicated block.

Defining gene families. All the predicted protein sequences of four sequenced legume genomes, namely Mt, Lj, Gm and pigeonpea, together with an out-group species grape (Vv), were compared against each other by using BLASTP[23]. The BLASTP results were filtered if the E-value > e-15, or aligned region length < 60% of any one of the aligned two sequences. For defining gene families, Markov cluster (MCL) algorithm[51] was used to cluster the BLASTP[23] results into groups of homologous proteins at inflation (I) parameter as 6.0 and other default parameters. Protein family emergence and extinction within phylogenetically related organisms were detected by custom Perl scripts.

Identification of ORFan genes. The ORFans in the pigeonpea genome were identified using a BLAST filtering approach (BLASTP, e-value < 0.01). All predicted pigeonpea peptide sequences were searched against all available peptide sequences of fully sequenced *Viridiplantae* genomes outside of the *Phaseoleae* tribe, that is, all genomes represented at phytozome.org (v7) minus the soybean genome. All pigeonpea peptides with a significant hit to a non-*Phaseoleae* peptide were filtered out. The remaining pigeonpea ORFan candidates were then BLAST searched against the NCBI nonredundant (nr) protein and EST (expressed sequence tag) databases using BLASTP and t-BLASTN, respectively (e-value < 0.01). For the NCBI multi-species databases the species names of all significant hits were retrieved using the gene accession and blastdbcmd program. Again those pigeonpea peptides with hits to non-*Phaseoleae* peptides were filtered out. Further filtering was done using position-specific PSIBLAST on the NCBI nr protein database (e-value < 0.01) and InterProScan[24]. For InterProScan only hits of type family were considered; if the taxonomic coverage of the family extended past *Phaseoleae*, then those ORFan candidates matching that family were removed.

ORFans originated by duplication events were identified by all-against-all BLASTP and BLASTN searches for all ORFans versus non-ORFans within the pigeonpea genome. Orthologs containing frameshifts (therefore producing novel peptides) were identified using BLASTN against all sequenced plant genome coding sequences. *De novo* origination or gene loss events were identified using BLASTN against the genome assemblies of all sequenced plant genomes and compared to known open reading frames.

Identification of SSRs and SNPs. SSRs were mined in the genome sequence using the MIcroSAtellite (*MISA*)[52] program, with the following parameters: at least ten repeats for mono-, six repeats for di-, and five repeats for tri-, tetra-, penta- and hexa-nucleotide for simple SSRs. The Primer3 program[53] was used for designing the primer pairs for identified SSRs based on the following criteria: (i) annealing temperature (Tm) between 50–65 °C with 60 °C as optimum; (ii) product size ranging from 100 bp to 350 bp; (iii) primer length ranging from 18 bp to 24 bp with an optimum of 20 bp; (iv) GC % content in the range of 40–60%.

SNPs were identified on the basis of alignment of Illumina transcript reads generated from each of the genotypes against the genome assembly using SOAPdenovo (http://soap.genomics.org.cn/), allowing not more than two mismatches. Based on the alignment results, with consideration and analysis of data characters, sequence quality and other influences of experiments, the Bayesian model was applied to calculate the probability of genotypes with the actual data. The genotype with the highest probability was selected as the genotype of the sequenced individual at the specific locus and a quality value is designated accordingly to reflect the accuracy of the genotype. Using the consensus sequence, polymorphic loci against the reference sequence are selected and then filtered under certain requirements (such as the quality value must be greater than 20 and result must be supported by at least three reads). Two additional filter steps were used to remove unreliable portions of the consensus sequence: (i) the average copy times of all the reads mapped to this position would be less than twice. (ii) The SNPs had to be at least 3 bp away from each other.

Estimation of heterozygosity. Heterozygosity in Asha genotype was estimated in the following four steps. (i) All the high-quality reads from the genomic DNA of Asha genotypes were mapped to the genome assembly using the software SOAP2 (http://soap.genomics.org.cn/soapaligner.html) with the cutoff less than five mismatches. (ii) The reads alignment results were analyzed for SNP mining using SOAPsnp (http://soap.genomics.org.cn/soapsnp.html). (iii) The sites that met the following criteria were searched and named "criterion effective sites": (a) quality score of consensus genotype in the SNP mining result is greater than 20; (b) count of all the mapped best and second-best bases are supported by at least four unique reads; (c) sequencing depth is more than 10×; and (d) SNPs are at least 5 bp away from each other. (iv) In addition to the parameters for the criterion effective sites, the sites whose number of reads supported by the best base calling is less than four times the number of reads supported by the second-best base calling (best base calling reads count <4 second-best base calling reads count) were identified as heterozygosis sites. Finally, the rate of the heterozygosity was estimated as the number of heterozygosis sites divided by the number of criterion effective sites.

38. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
39. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
40. Price, A.L., Jones, N.C. & Pevzner, P.A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).

41. Kent, W.J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
42. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
43. Stanke, M. *et al.* Augustus: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* **34**, 435–439 (2006).
44. Salamov, A.A. & Solovyev, V.V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
45. Majoros, W.H., Pertea, M. & Salzberg, S.L. TigrScan and GlimmerHMM: two opensource *ab initio* eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
46. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
47. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
48. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
49. Beckstette, M., Homann, R., Giegerich, R. & Kurtz, S. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics* **7**, 389 (2006).
50. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
51. Enright, A.J., van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575 (2002).
52. Thiel, T., Michalek, W., Varshney, R.K. & Graner, A. Exploiting EST databases for the development and characterization of gene derived SSR-markers in barley (*Hordeumvulgare* L.). *Theor. Appl. Genet.* **106**, 411–422 (2003).
53. Rozen, S. & Skaletsky, H.J. Primer3 on the WWW for general users and for biologist programmers. in *Bioinformatics Methods and Protocols: Methods in Molecular Biology* (eds., Krawetz, S. & Misener, S.) 365–386 (Humana, Totowa, 2000).

doi:10.1038/nbt.2022