the plant journal

SEB Society for Experimental Biology

The Plant Journal (2013) **74,** 715–729 doi: 10.1111/tpj.12173

FEATURED ARTICLE

A draft genome sequence of the pulse crop chickpea (*Cicer arietinum* L.)

Mukesh Jain, Gopal Misra†, Ravi K. Patel†, Pushp Priya, Shalu Jhanwar, Aamir W. Khan, Niraj Shah, Vikas K. Singh, Rohini Garg, Ganga Jeena, Manju Yadav, Chandra Kant, Priyanka Sharma, Gitanjali Yadav, Sabhyata Bhatia, Akhilesh K. Tyagi* and Debasis Chattopadhyay*

National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi 110067, India

Received 24 January 2013; revised 27 February 2013; accepted 4 March 2013; published online 12 March 2013.
*For correspondence (e-mails akhilesh@genomeindia.org, debasis@nipgr.ac.in).
†Equal contribution.

SUMMARY

***Cicer arietinum* L. (chickpea) is the third most important food legume crop. We have generated the draft sequence of a desi-type chickpea genome using next-generation sequencing platforms, bacterial artificial chromosome end sequences and a genetic map. The 520-Mb assembly covers 70% of the predicted 740-Mb genome length, and more than 80% of the gene space. Genome analysis predicts the presence of 27 571 genes and 210 Mb as repeat elements. The gene expression analysis performed using 274 million RNA-Seq reads identified several tissue-specific and stress-responsive genes. Although segmental duplicated blocks are observed, the chickpea genome does not exhibit any indication of recent whole-genome duplication. Nucleotide diversity analysis provides an assessment of a narrow genetic base within the chickpea cultivars. We have developed a resource for genetic markers by comparing the genome sequences of one wild and three cultivated chickpea genotypes. The draft genome sequence is expected to facilitate genetic enhancement and breeding to develop improved chickpea varieties.**

Keywords: chickpea, *Cicer arietinum*, ICC4958, desi-type, genome sequence, gene expression, marker resource, technical advance.

INTRODUCTION

Legumes are the second most important crop for humans. Grain and forage legumes are grown in about 15% of the world's cultivated land, account for 27% of world's primary crop production and provide 33% of dietary nitrogen requirement, apart from being a natural fertilizer (Graham and Vance, 2003). *Cicer arietinum* (chickpea) ranks third in food legume crop production in the world, with 96% of its cultivation occurring in developing countries. The Indian subcontinent is the principal chickpea-producing and -consuming region, contributing almost 70% of the world's total production (FAOSTAT, 2009). In addition, chickpea is also produced in eastern Africa, western Asia, the Mediterranean Basin, Australia, and in some parts of the European and American continents. More than 90% of chickpea production is consumed locally (FAOSTAT, 2009), demonstrating its importance to the community where it is grown. It is a rich source of protein and starch for the predominantly vegetarian population in the growing countries. Moreover, being a grain legume it plays an integral part in diversifying the cereal-based cropping system because of its ability to fix atmospheric nitrogen and break disease cycles (Arnon, 1972). According to seed morphology, cultivated chickpeas are of two types: kabuli and desi. Kabuli chickpeas have large seeds (100-seed mass $\geq$ 50 g) with thin seed coats of white–cream colour, and are usually grown in regions of temperate climate. Desi chickpea seeds are relatively smaller (100-seed mass $\leq$ 30 g) with thick seed coats of dark-brown colour, and are grown in semi-arid tropical regions.

Chickpea, a diploid ($2n = 16$) annual legume crop species of the family *Leguminosae* and subfamily *Faboideae*, is a member of the West Asian Neolithic crop assemblage. Its origin of cultivation can be traced back to 7500 years ago in Turkey, and extends back to Central Asia (Zohary and Hopf, 2000). Unlike other crops of the same origin, the wild progenitor of cultivated chickpea, *Cicer reticulatum*, is

narrowly distributed. During domestication, a series of evolutionary bottlenecks resulted in a narrow genetic base among the cultivated chickpea varieties (Abbo *et al.*, 2003a,b). Chickpea cultivation requires low external inputs, and it is generally grown in marginal lands with residual moisture. Globally chickpea is grown on 11.5 million hectares (ha) to produce 10.4 million tons, with an average yield of about 0.9 ton ha^{-1}, which is far below its yield potential of 6 ton ha^{-1} under optimum growing conditions (FAOSTAT, 2009). The narrow genetic base of cultivated chickpea varieties attenuates the efforts of marker-assisted crop improvement and production of elite cultivars with durable stress-resistance by conventional breeding, which has so far been compounded by limited genomic resources, lack of comprehensive intergenic and intragenic molecular marker maps, and a lack of the genome sequence (Young and Bharti, 2012; Gaur *et al.*, 2012b). Therefore, a concerted genomic approach has been undertaken encompassing whole-genome sequencing to have a deep insight into the gene content and organization of the chickpea genome. The chickpea cultivar ICC4958, a desi drought-tolerant genetic stock and a popular breeding parent isolated from India has been used for this purpose. The chickpea draft genome sequence, diversity analysis of chickpea cultivars and the marker resource generated are expected to help elucidate the molecular basis of agronomically important traits, and to facilitate the genetic improvement of a nutritionally and economically important crop.

RESULTS

Sequencing, assembly and coverage

We generated 13.354 Gb of high-quality sequence data for chickpea ICC4958 by sequencing whole-genome shotgun (WGS) libraries and mate-pair (MP) libraries of 3–20-kb insert sizes using a 454/Roche GS FLX Titanium platform. The Illumina GA IIx sequencing platform was used to sequence two WGS libraries to produce an additional 43.7 Gb of quality-filtered paired-end (PE) sequence data (Table S1a). The Illumina data set was assembled using ABySS (Simpson *et al.*, 2009) to produce 304 948 126 bases of assembled sequences. These contigs and all the quality-filtered reads generated by GS FLX platform were assembled together using NEWBLER v2.5.3 to obtain the primary assembly. Further scaffolding of the assembled fragments was performed using publicly available bacterial artificial chromosome (BAC)-end sequences (Thudi *et al.*, 2011). The sequence scaffolds were assigned to their corresponding linkage groups by mapping the marker sequences used to make an integrated genetic linkage map of the *C. arietinum* ICC4958 × *C. reticulatum* PI489777 mapping population (Gaur *et al.*, 2012a). The assembled sequence spanned 519 846 222 bases, with an N50 length of 77 313 bases and the N50 index (rank of the N50-scaffold according to size) of 931 (Table 1). About 84% of the assembly consists of scaffolds with 2 kb or more in size (Figure S1; Table S2). Altogether, 532 scaffolds spanning 124.4 Mb were assigned to eight linkage groups. The largest pseudomolecule was linkage group 3, with a length of 23.4 Mb (Table S3). The draft assembly covered 70% of the 740 Mb predicted genome length (Arumuganathan and Earle, 1991) at an average of 15X GS FLX read coverage (Figure S2). The total genome size based on read alignment was estimated to be 740.52 Mb (Table S4). Overall, 98.66% of the assembled bases had a quality of Q40 or more. Independent assessment of heterozygosity by mapping reads generated by the SOLiD platform (Table S1a.iii) revealed an average heterozygosity of 0.049% in the linkage groups (Table S5) and an average of 0.052% in the whole assembly, which is comparable with the hetrozygosity of another closed flower-pollinating *Fabaceae* family crop, *Cajanus cajan* (pigeonpea, 0.067%; Varshney *et al.*, 2012). The mapping of high-quality Illumina reads to the assembly revealed a mismatch error frequency of 1.2 bases/10 kb and an indel frequency of 1.6/10 kb. Although the predicted heterozygosity is extremely low, it may contribute a small fraction to the mismatch error. The assembly was aligned to the working draft sequences of 12 BACs of chickpea available from public databases. The unordered pieces of the BAC contigs aligned to the scaffolds with 95–100% identity over the entire stretches.

Transcriptome coverage in the assembled chickpea genome was calculated using three data sets, 34 760 chickpea transcripts, 1 931 224 high-quality Roche 454 RNA-seq reads generated in a previous study (Garg *et al.*, 2011) and 41 045 expressed sequence tags (ESTs), available at NCBI

Table 1 Assembly and annotation statistics of the chickpea genome

Total size (bp)	519 846 222
Number of scaffolds	181 462
Minimum scaffold length (bp)	200
Maximum scaffold length (bp)	23 376 002
Average scaffold length (bp)	2865
N50 length (bp)	77 313
N50 index	931
GC content (%)	26.93
Size of repetitive content (bp)	210 201 779 (40.4%)
Protein coding genes	27 571
Average gene length (bp)	3122
Average coding sequence length (bp)	962
Average number of exons	4.23
Average exon length (bp)	270
Average intron length (bp)	606
tRNA loci	627
rRNA loci	249
miRNA loci	60
snoRNA loci	278

after cleaning. A BLAT search of these sequences in the chickpea genome revealed that about 84% of the transcripts were covered in the genome with at least 90% identity and 80% coverage, and about 82% transcripts showed at least 90% identity and 90% coverage in the genome (Table S6). Furthermore, about 77% of the large transcripts (>2 kb) overlapped significantly with the chickpea genome (with at least 90% identity and 80% coverage). Likewise, more than 69% of the ESTs and about 81% of the Roche 454 RNA-seq reads were present in the genome, with at least 90% identity and 80% coverage. Overall, these results indicated that more than 80% of the expressed sequences are covered in the assembled chickpea genome, which is comparable with transcript coverage of the draft genome sequence of date palm (Al-Dous *et al.*, 2011).

Identification of repetitive elements and annotation of protein coding genes

A total of aproximately 210 Mb, representing about 40.4% of the draft genome sequence, was identified as interspersed repeat sequences, in which 27.31% constitutes retrotransposons with more than 0.3 million copies. About 96% of these retrotransposons are long terminal repeats, and 8.6% of repeats could not be classified in any known category (Table S7). Another 4.55% of the repetitive sequence represented DNA transposons. Simple sequence repeats (SSRs) constitute 0.329% of the genome assembly, and this fraction is similar to *Medicago truncatula* (*Medicago*; 0.203%) and *Glycine max* (soybean; 0.328%). However, fractions of trimers and tetramers are much higher in chickpea in comparison with those in *Medicago* and soybean (Table S8).

Using the repeat-masked genome sequence, we analysed the protein-coding genes based on various gene prediction approaches [*ab initio*, homology-based, EST and Core Eukaryotic Genes Mapping Approach (CEGMA) analyses], followed by generating a non-redundant consensus gene set by merging their prediction results. We predicted a total of 27 571 non-redundant consensus genes (a total of 86.1 Mb, 5.3 genes/100 kb), with an average gene length of 3122 bp, coding sequence length of 962 bp (total of 26.5 Mb) and an average of 4.23 exons per gene (Tables 1 and S9). At least 3109 (11.3%) chickpea genes had an average of 2.37 transcript isoforms, representing a total of 31 844 transcripts. The overall GC content within the coding sequence was higher (43.6%) than that of the whole genome (26.93%). The total N content in the coding sequence of the predicted gene set was only 0.05% (overall only 29 genes contained N, and only 15 genes had $\geq$10% of N in the coding sequence). An analysis by CEGMA pipeline revealed that about 78% eukaryotic orthologous groups (KOGs) with $\geq$80% coverage, and more than 86% KOGs with $\geq$50% coverage, were included in our predicted gene set (Table S10). Overall, >73% of predicted genes were supported by transcripts, ESTs and/or RNA-seq data (Table S11).

Comparative genome features of chickpea

The GC content distribution in the chickpea genome (bin window size of 500 bp) was comparable with other dicot genomes, including legumes (soybean, *Medicago*, pigeonpea, *Lotus japonicus*, *Arabidopsis thaliana* and poplar), whereas the GC distribution in the *Oryza sativa* (rice) genome was much higher (Figure S3). We did a comparative analysis of different features of the predicted genes in other sequenced legumes (soybean, *Medicago*, pigeonpea, *L. japonicus*) and a few dicots [Arabidopsis, *Cucumis sativus* (cucumber), *Theobroma cacao* (cocoa) and *Vitis vinifera* (grapevine)], as presented in Table S12. Although the total number of predicted genes in chickpea (27 571) was less than in other legumes (ranging from 38 482 genes in *L. japonicus* to 62 379 genes in *Medicago*), the number was comparable with other dicots (ranging from 21 503 in cucumber to 28 798 in cocoa) (The Arabidopsis Genome Initiative, 2000; Jaillon *et al.*, 2007; Huang *et al.*, 2009; Argout *et al.*, 2011). The possibility of the presence of a few genes in the unsequenced portion of the chickpea genome is not ruled out. The average transcript length in chickpea (3122 bp) was larger than those in legumes *Medicago* (2028 bp), pigeonpea (2520 bp) and *L. japonicus* (1494 bp), but was close to that of soybean (3693 bp). The average exon length of chickpea genes (270 bp) was comparable with other legumes, but the average intron length was higher (606 bp) than in the other legumes and dicots analysed, excepting grapevine.

Multi-species comparison and identification of lineage-specific gene families

For the delineation of unique and shared gene families, the predicted proteomes of chickpea, soybean, pigeonpea, *Medicago* and grapevine were analysed. A total of 19 218 chickpea genes clustered with 145 801 genes from the other three legumes (soybean, *Medicago* and pigeonpea) and grapevine (non-legume out-group) in 22 281 gene families of two or more members (Figure 1a). Among these, 6474 gene families containing 94 503 genes were conserved in all five species, whereas 1327 families containing 9990 genes were restricted to four legume species. A total of 626 families representing 1491 genes were unique to chickpea. Chickpea shared the largest number (864) of gene families with *Medicago*, whereas only 118, 219 and 131 gene families were shared with soybean, pigeonpea and grapevine, respectively. This indicates a close relationship of chickpea with *Medicago* as compared with other legumes and grapevine. The number of gene families unique to chickpea was fewer (626) than those of *Medicago* (3799) and pigeonpea (1,389), but greater than that of soybean (363).

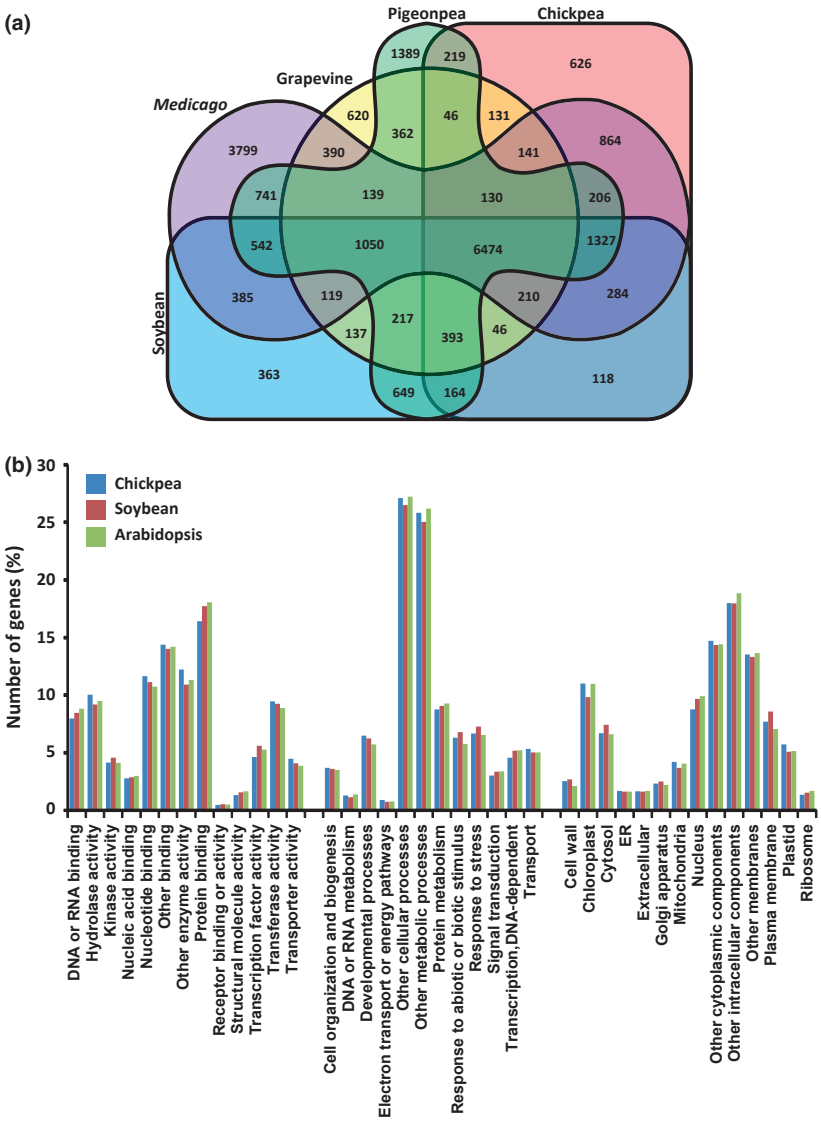


Figure 1. Comparative features of chickpea protein-coding genes.
(a) Venn diagram showing distribution of gene families among *Cicer arietinum* (chickpea), *Glycine max* (soybean), *Medicago*, *Cajanus cajan* (pigeonpea) and *Vitis vinifera* (grapevine). Comparative analysis revealed that a total of 165 019 genes from five species are clustered into 22 281 gene families. The number of gene families that are unique and are shared among different species is indicated. (b) Distribution of various GOSlim categories in chickpea, soybean and Arabidopsis genes. GOSlim categories for Arabidopsis genes were extracted from the TAIR database. GOSlim categories were assigned to all the chickpea and soybean genes, based on their corresponding best hit gene from Arabidopsis.

Based on the BLAST searches, more than 89% of the predicted genes showed similarity to at least one of the public protein databases analysed, and 80.3% of genes were assigned a gene ontology (GO) term using the Blast2GO pipeline (Table S13). The transferase activity (14.3%), metabolic processes (12.2%) and membrane (22.7%) were the most abundant GO categories represented (Figure S4). A similar distribution of GOSlim terms was found among the chickpea, soybean and Arabidopsis genes (Figure 1b). The genes encoding kinase domain proteins were the most abundant (2.5%) in chickpea (Figure S5).

Using various BLAST searches, we further identified 2751 (aproximately 10%) genes as putative chickpea-specific orphan genes, which did not show detectable similarity to any sequence analysed (Figure S6). This was similar to the earlier prediction based on the chickpea transcriptome (Garg *et al.*, 2011) and less than rice (17.4%) (Campbell *et al.*, 2007), but higher than Arabidopsis (4.9%) (Lin *et al.*, 2010). The other 954 genes represented putative legume-specific genes (Figure S6). The GO analysis revealed that receptor and transporter activity among the molecular function, metabolic process and transport among the biological process, and integral to membrane and outer membrane-bound periplasmic space among the cellular component terms, were more abundant in chickpea-specific gene families (Figure S7). A comparison of features of lineage-specific genes with that of other genes was performed (Table S14). The average GC content of lineage-specific genes was much higher than that of other genes. About 73% of legume-specific and 78% of chickpea-specific orphan genes were single exonic, which is much higher than in other genes (aproximately 32%). The average coding region and protein length of lineage-specific genes were much shorter than those of other genes. The

evidence for expression of about 51% of legume-specific and only 22% of chickpea-specific genes was available, as compared with 88% of the other genes. The difference in the structures of genes from different lineages and poor quality of gene calls for lineage-specific genes can be further minimized by experimental evidence of their structure and biological relevance. Considering the fact that the genome sequence presented here is the first-draft genome sequence, the exact picture will improve once updated versions of the genome sequences and an annotated gene set of chickpea and other legumes become available.

Gene families

Based on the hidden Markov model (HMM) profile search, we identified a total of 1748 (6.3%) chickpea genes belonging to 84 transcription factor families (Table S15). This fraction is much less than that in soybean, but is comparable with that in other legumes and in Arabidopsis (Libault *et al.*, 2009; Schmutz *et al.*, 2010; Figure S8). The MYB domain-containing proteins were the most abundant among the transcription factor families, followed by basic helix-loop-helix (bHLH) and Apetela2 (AP2)-domain-containing proteins encoded by 110, 103 and 101 genes, respectively. The high representation of MYB/MYB-related transcription factors in the chickpea draft assembly is in accordance with other plants; however, unlike a high abundance of CCHC domain-containing transcription factors in pigeonpea, the fraction of this group of genes in the chickpea draft genome is relatively low (totalling 21 in number), similar to other plants. The abundance of auxin response factors (ARF) in the chickpea draft genome (21 only) is similar to that in Arabidopsis, but is less than one-third of those present in *L. japonicus* and soybean.

To delineate resistance-related genes, previously described *R*-genes were selected from the PRG database (Sanseverino *et al.*, 2010) and NCBI protein database using BLASTP. The initial screening of the chickpea unigene set revealed the presence of 729 putative resistance-related genes, including receptor-like kinases. The presence of specific R-protein domains was inspected using InterProScan and the InterPro database (Hunter *et al.*, 2009) to select 119 high-probability *R*-gene candidates. Five hundred and thirty-five proteins with receptor-like kinase domains were also mentioned to include the proteins putatively involved in resistance processes. The same *R*-gene prediction pipeline with the same parameters was applied to the annotated protein databases for *Medicago*, soybean, Arabidopsis and grapevine to retrieve the predicted *R*-genes from those plants. Each sequence was assigned to an appropriate class of *R*-genes based on their conserved domains and compared among the plant species mentioned above (Table S16). The chickpea draft assembly possesses many fewer *R*-genes compared with the two other legumes analysed; however, the number of receptor-like kinases is comparable with those in other plants, except soybean. Given the complexity of *R*-genes, it is also possible that the present draft assembly could not capture the full complement of *R*-genes. Phylogenetic analysis of the CC-NBS-LRR genes of chickpea and *Medicago* shows that the chickpea genes form separate clusters, suggesting divergence of this family from the *Medicago* genes (Figure S9a). A list of resistance-associated genes is available at http://nipgr.res.in/CGAP/home.php.

Genes involved in various events associated with nodule development were identified by comparison with orthologous nodulation-related genes of other legumes. For this, 76 sequences of key regulatory genes involved in nodulation (largely including the NOD factor receptors, receptor-like kinases, calcium/calmodulin-dependent kinases, transcription factors and ion-channel transporters) and 56 sequences of nodulin genes from *Medicago*, soybean, *Lotus* and *Pisum sativum* were downloaded from the NCBI database. A unique set of genes was identified to generate the database, which consisted of 26 nodulation regulatory genes and 44 nodulin genes. These were used to search for sequence similarity with protein sequences of chickpea, *Medicago*, soybean, pigeonpea and Arabidopsis. In the present chickpea genome assembly, 89 nodulation-related genes (53 nodulation regulatory and 36 nodulin genes) were identified that were lower than the nodulation genes present in *Medicago* (total 166), soybean (total 256) and pigeonpea (total 156), but as expected, were much higher than the 65 genes (39 nodulation regulatory genes and 26 nodulins) in the non-leguminous Arabidopsis (Table S17). The nodulation-related genes in chickpea predominantly consisted of transporters (sugar, water and nuclear), transcription factors, i.e. GRAS, ERF and bZIP, receptor-like kinases and other signalling components involved in cell signalling during the nodulation process. (A list of nodulation-associated genes is available at http://nipgr.res.in/CGAP/home.php). Among 89 nodulation genes in chickpea, 21 had at least one homologue, and some genes like the ERF transcription factor and pectate lyase had several homologues. Furthermore, phylogenetic analysis of the leghaemoglobin genes of chickpea, *Medicago* and soybean showed that three of the chickpea leghaemoglobin genes clustered distinctly, whereas one clustered with the *Medicago* homolog (Figure S9b). An analysis using BLASTP and HMM profile searches at a cut-off of $1e^{-10}$ revealed 170 and 643 unique genes associated with the metabolism of carotenoids and flavonoids, respectively. Both the gene families have many fewer genes as compared with those in Arabidopsis and *Medicago* (Table S18). The comparison of gene families presented here is based on the draft genome assembly and its annotation, and therefore might have bias towards certain families, which will be clear with the availability of the complete annotated gene set.

Annotation of non-coding RNA genes

The tRNA genes were predicted by scanning the sequence using tRNAscanSE using eukaryotic parameters (Lowe and Eddy, 1997; Table S19). A total of 627 tRNA loci, excluding three pseudo tRNA, spanning 46 343 bp with an average length of 74 bp/locus, were predicted. Interestingly, only tRNA genes coding for methionine and tyrosine showed the presence of introns. The average intron lengths of the tRNA genes for methionine and tyrosine are 9.7 and 18.2 bp, respectively. Genes encoding ribosomal RNAs were predicted using BLASTN by aligning the rRNA sequences of Arabidopsis (GenBank accession: AJ307399.2), rice (M82426.1) and seven species of the genus *Medicago* (*Medicago polyceratia*, AJ28842; *Medicago medicaginoides*, AJ288239.1; *Medicago brachycarpa*, AJ288234.1; *Medicago monantha*, AJ288266.1; *Medicago fischeriana*, AJ288214.1; *Medicago saxatilis*, AJ288270.1; *Medicago aurantiaca*, AJ288260.1 and *Medicago papilosa*, AJ288224.1) with a cut-off of $1e^{-10}$. A total of 249 rDNA loci spanning 40 856 bp were predicted. For the prediction of miRNA and snoRNA, a combination of BLASTN and INFERNAL 1.0.2 was used (Nawrocki *et al.*, 2009). INFERNAL identified 1455 non-coding RNAs of 121 distinct Rfam families. Among these, 60 miRNA loci, with an average precursor miRNA length of 120.23 bp/locus, and spanning 7214 bp, were identified, which represented 0.0014% of the genome. The miRNAs represented 20 unique families, of which MIR169_2 (15%) and MIR159 (11.67%) were most abundant. Similarly, 278 snoRNA loci with average length of 96.13 bp/locus and spanning 26 724 bp were identified, which represented 0.0051% of the genome. The snoRNAs represented 75 unique families, of which snoZ159 (5.04%) and snoZ278 (4.68%) were present in maximum abundance.

Gene expression analysis

RNA-seq data was generated from libraries prepared with RNA isolated from different tissues/organs using the Illumina GA-IIx platform. A total of 274 million filtered reads from six tissue samples representing different tissues/organs and three tissue samples representing stressed/non-stressed conditions were analysed (Table S20). Data analysis showed the expression of about 72% of genes ($\geq$1 read per million) in at least one tissue/condition (Figures 2 and 3). In a tissue-by-tissue comparison, the largest number of genes was differentially expressed between shoot and root, followed by flower and root, and then leaf and root. A smaller number of genes were differentially expressed between the green (shoot and leaf, shoot and stem, and leaf and stem) tissues, as expected. In general, the expression profile of chickpea genes in root tissue was the most distinct from other tissues analysed (Figure 2). A large number (5843) of genes exhibited preferential expression in a particular tissue. A maximum number of genes showed preferential or specific expression in root and flower tissues, followed by pod (Figure 2; Table S21). Of the 1680 transcription factor genes with RNA-seq data, at least 526 (31.3%) genes exhibited tissue-preferential expression. A wide variety of GO terms were enriched within tissue-preferential/tissue-specific and stress-responsive chickpea genes (Figure 3; Table S22). In shoot and leaf tissues, the genes involved in photosynthesis-related processes were significantly enriched. In root, the genes involved in response to stimulus/stress and those having heme/iron binding activity were over-represented. In flower, the genes involved in enzyme regulator activity were significantly represented. The nutrient reservoir activity, endopeptidase activity and proteolysis-related genes were significantly enriched in pod (Table S22).

A total of 2000 chickpea genes exhibited differential expression under drought and/or salt stress conditions. Among the 1278 and 1163 genes regulated under drought and salt stress, respectively, 441 genes were in common (Figure 3). More than 13% of the transcription factor genes were differentially expressed under drought and/or salt stress conditions (Table S21). The genes with oxidoreductase activity, transcription factor activity and protein kinase activity were significantly enriched in the stress-responsive genes. About 26 and 10% of the legume-specific and chickpea-specific genes, respectively, exhibited tissue-preferential and/or stress-responsive expression (Table S21).

Synteny and genome duplication

At least 32 duplicated blocks (30 interchromosomal and two intrachromosomal) of genes were identified in the chickpea genome (Figure 4a). Notably, linkage group 3 harboured 14 duplicated blocks shared with five other linkage groups (CaLG2, CaLG4, CaLG6, CaLG7 and CaLG8), the largest (five) being with each of linkage groups 4 and 7. The distribution analysis of synonymous substitution rate (K_s) within the paralogous gene pairs (Figure S10) confirmed the absence of the recent whole genome duplication (WGD) event, as predicted in soybean (Schmutz *et al.*, 2010). Considering the old legume genome duplication event to have occurred about 58 Mya, the rate of synonymous substitution per site per year is calculated to be 6.05×10^{-9} in chickpea, which is 12% faster than that predicted in the model legume *Medicago* (Young *et al.*, 2011).

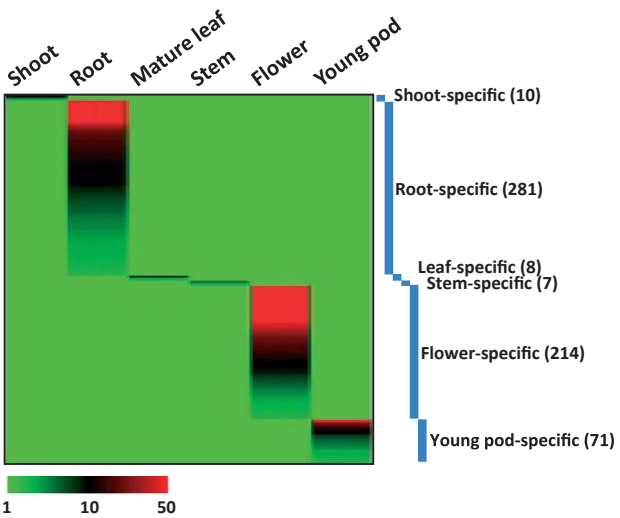
To analyse genome-wide synteny within legumes, a whole-genome dot plot was generated with eight chickpea linkage groups on the *x*-axis against chromosome arms (north and south) of soybean and *Medicago* on the *y*-axis. For the soybean genome, large pericentromeric regions were removed. Even after divergence of *Medicago* and chickpea by about 25 Myr (Choi *et al.*, 2004), extensive synteny and conservation of gene order were observed between the chickpea and *Medicago* genomes, particularly

Figure 2. Tissue-specific differential expression of *Cicer arietinum* (chickpea) genes.
(a) Genes preferentially expressed in each tissue sample, as compared with others, in a tissue-by-tissue comparison. The genes showing at least twofold change (upregulated above the blank cells and downregulated below the blank cells), as compared with other tissue samples, are given. (b) Heatmap showing genes preferentially/specifically expressed in various tissue samples. The number of genes and tissue specificity is noted on the right side. The colour scale (1–50) represents reads per million (RPM) value. The transcripts with >50 RPM are not distinguished by the colour scale (they are represented by the same colour as that of genes with 50 RPM), so as to differentiate the genes with low–high expression.

(a)

	Shoot	Root	Mature leaf	Stem	Young pod	Flower
Shoot		661	410	366	478	725
Root	1171		1011	880	833	1118
Mature leaf	816	674		867	796	854
Stem	705	741	381		541	897
Young pod	694	731	556	646		831
Flower	447	578	389	405	361	

(b)



between chickpea LG5 and *Medicago* LG3, chickpea LG3 and *Medicago* LG5 (Figures 4b, S11 and S12). The absence of duplicated syntenic blocks between these two genomes within the galegoid clade of the *Papilinoideae* subfamily suggested the absence of recent whole genome duplication in chickpea and/or *Medicago*. In contrast, the chickpea genome showed less synteny with soybean, a member of the millettoid clade. However, the presence of duplicated syntenic blocks of chickpea genes in the soybean genome (Figure S13) reflects recent whole genome duplication in soybean.

Genome evolution

Genome-wide analyses have provided evidence of whole genome duplication in plants as a mechanism of diversification and adaptation to an altered environment (Otto and Whitton, 2000; Taylor and Raes, 2004; Comal, 2005). One of the methods used for the dating of large-scale duplication or divergence is to plot synonymous substitution per synonymous site (K_s) between the paralogous or orthologous genes present in the syntenic blocks within or between the genomes, respectively (Lynch and Conery, 2000; Paterson *et al.*, 2004; Lavin *et al.*, 2005).

Distribution of K_s within a number of orthologous gene pairs between grapevine and *Medicago*, grapevine and chickpea, *Medicago* and *Lotus*, soybean and *Medicago* and paralogs within chickpea were plotted (Figure 5). *Medicago*–grapevine and chickpea–grapevine orthologs share peaks at K_s 1.5. Incidentally, chickpea paralogs also show a peak at K_s 1.5. Assuming a synonymous substitution rate per synonymous substitution of 6.1×10^{-9} per year (Lynch and Conery, 2000) for eudicots, this peak is most probably attributed to 'gamma triplication', common for all eudicots (Jaillon *et al.*, 2007), and coincides with the period of divergence of the Eurosids from grapevine (Tang *et al.*, 2008; Fawcett *et al.*, 2009) about 120 Mya. The major peak of the chickpea paralogs at K_s 0.7 (Figures 5 and S10) is shared by several other legumes (Pfeil *et al.*, 2005; Cannon *et al.*, 2006; Schmutz *et al.*, 2010; Young *et al.*, 2011), and is thought to be a signature of common WGD aproximately 58–60 Mya, marking the origin of legumes coinciding with the Cretaceous–Tertiary (K–T) boundary, which experienced similar polyploidization in other plant lineages, possibly because of sudden environmental changes (Wilf and Johnson, 2004; Fawcett *et al.*, 2009). A comparison of overlapping peaks of the *Medicago*–soybean with *Medicago*–*Lotus* orthologs suggests that the diversification of the milletoid (soybean) and galegoid clades (*Medicago*) (41–46 Mya), within the *Papilionoideae* subfamily, slightly precedes the diversification of the *Trifoleae* tribe (*Medicago*) from *Loteae* (*Lotus*) (38–40 Mya) within the galegoids. Previous reports following independent methods also suggested the separation of milletoids and galegoids immediately before the separation of these two tribes within the galegoid clade (Choi *et al.*,

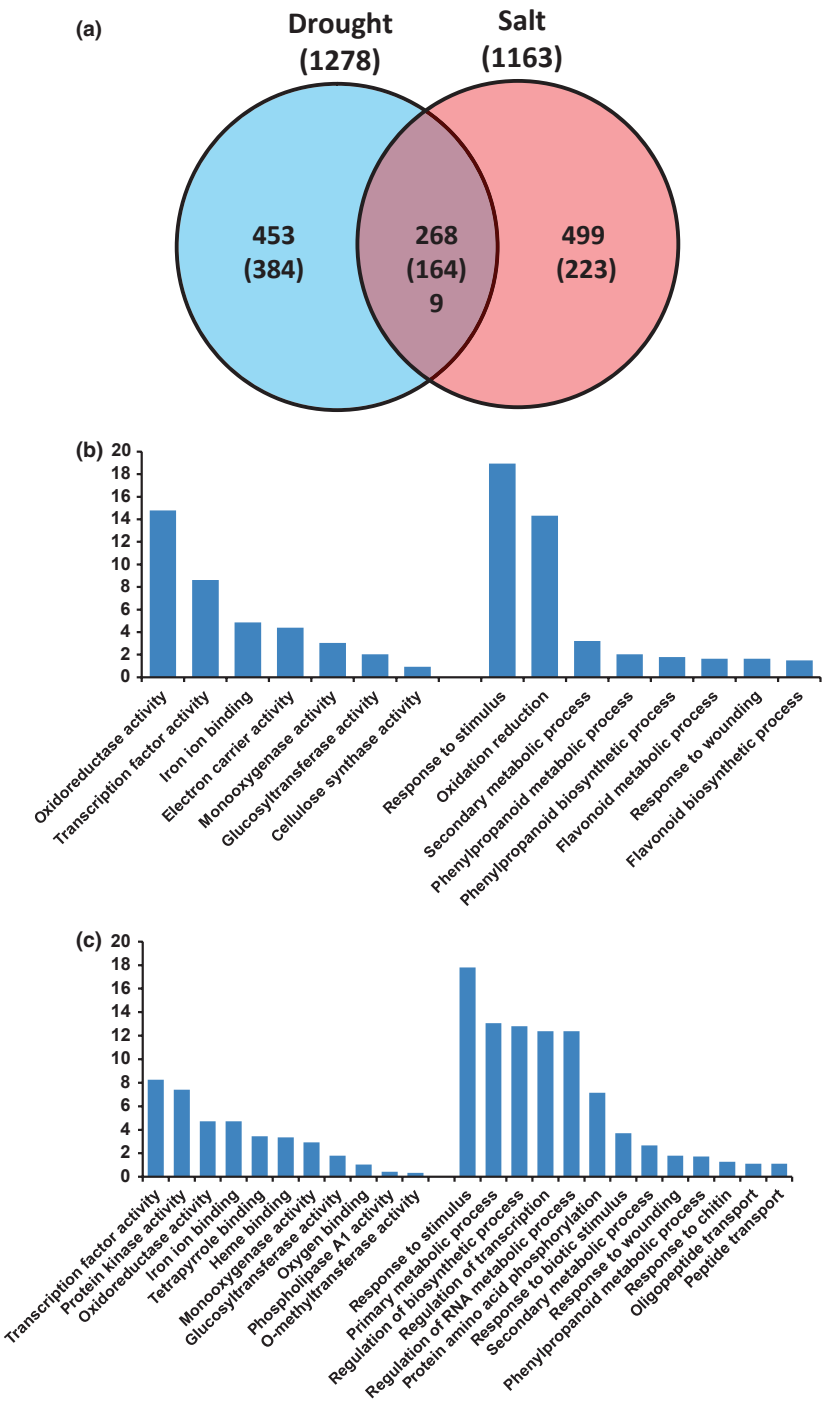


Figure 3. Differential expression of *Cicer arietinum* (chickpea) genes in response to drought and salt stress.
(a) Venn diagram showing the number of genes up- and downregulated (parentheses) under drought, salt and both stress conditions, as compared with the control. Nine genes were upregulated under one but downregulated under the other stress conditions. The total number of genes regulated under drought and salt stress conditions are shown outside the Venn diagram. (b, c) Enrichment of molecular function and biological process gene ontology terms in chickpea genes differentially expressed under drought (b) and salt (c) stress with $P \leq 0.01$.

2004; Pfeil *et al.*, 2005). K_s dating suggested the divergence of chickpea (*Cicereae* tribe) and *Medicago* by 25–30 Myr, well after the separation of *Medicago* and *Lotus*. Accordingly, we found considerable synteny between the chickpea and *Medicago* genomes. In the absence of recent WGD in *Medicago* (Young *et al.*, 2011), chickpea and *Lotus* (Sato *et al.*, 2008), one of the reasons for the diversification of these three plants belonging to three different tribes might be small chromosomal rearrangements and lineage-specific gene gain/loss and evolution. It is reported that *Medicago* has undergone extensive local gene duplications (Young *et al.*, 2011). Our previous study based on transcriptome data suggested a speciation within the *Cicereae* tribe separating chickpea from its wild progenitor *Cicer reticulatum* aproximately 0.6 Mya (Jhanwar *et al.*, 2012).

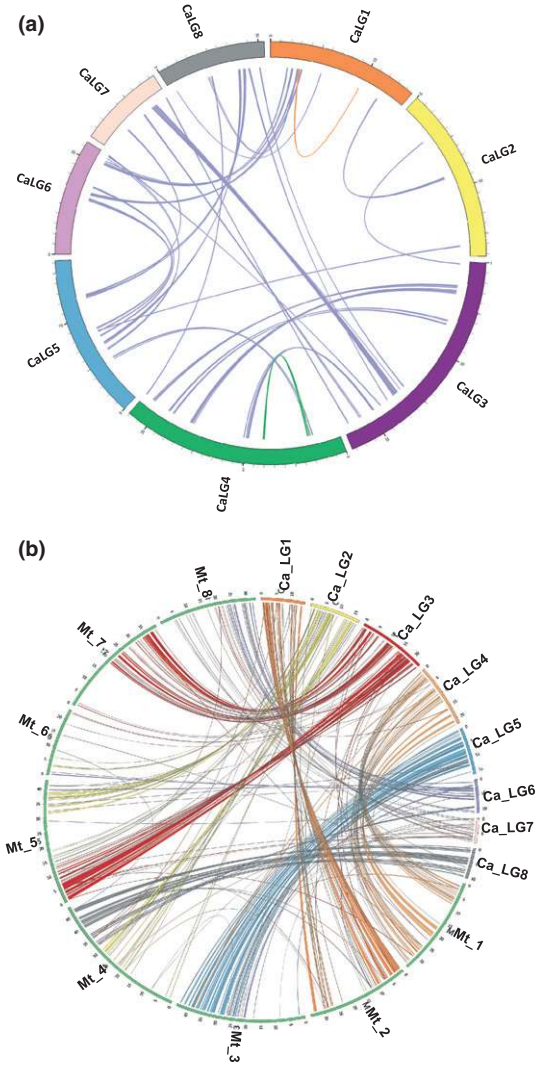


Figure 4. Whole genome duplication and synteny analysis.
(a) Duplicated blocks in the *Cicer arietinum* (chickpea) genome. All eight linkage groups are shown in coloured blocks, arranged in a circle. The interlinkage group duplicated blocks are marked by blue connecting lines and the intralinkage group duplicated blocks are marked by connecting lines of the same color as that the linkage groups. (b) Circos diagram presenting the syntenic relationship between chickpea and *Medicago truncatula* (Mt) pseudomolecules. Mt pseudomolecules were shown in green and labelled as Mt_1–8. Chickpea pseudomolecules are shown in different colours and labelled as Ca_LG1–8.

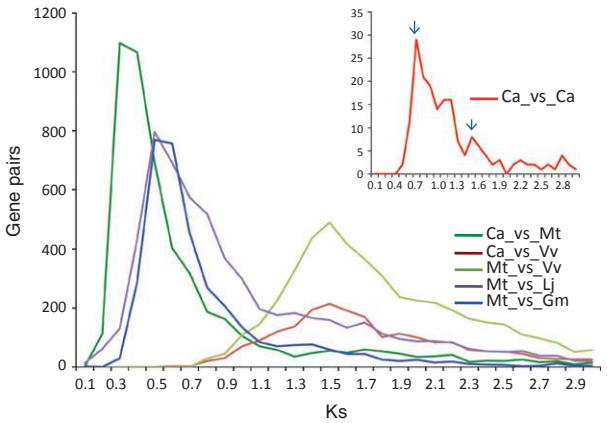


Figure 5. Synonymous substitution rate (K_s) dating of duplication blocks in *Cicer arietinum* (chickpea) and different combinations of orthologs of chickpea, *Vitis vinifera* (grapevine), *Glycine max* (soybean), *Medicago* and *Lotus japonicus*.
Distribution of K_s of chickpea paralogs and orthologs of chickpea (Ca) and *Medicago* (Mt), chickpea and grapevine (Vv), *Medicago* and grapevine, *Medicago* and *L. japonicus* (Lj), and *Medicago* and soybean (Gm) plotted against number of gene pairs are shown in different coloured lines. Inset shows distribution of K_s between the gene pairs present in the duplicated blocks within the chickpea genome. Blue arrows indicate major peaks that suggest duplication in the chickpea genome.

The distribution of non-synonymous substitution rate (K_a)/K_s (ω) and K_s between the orthologous gene pairs of *Medicago* and chickpea formed three clusters according to the K_s values. These gene clusters are centred around K_s values of 0.3, 1.5 and beyond (Figure S14). The average ω values of the genes decreased with the average K_s of the clusters. Clusters with $K_s \geq 1.5$ were attributed to pan-eudicot palaeoploidization. The low ω of these clusters indicates that these genes are under neutral selection. The cluster with $K_s < 1$ refers to neopolyploidization, and its higher ω suggests that the genes in this cluster are under purifying selection. A cluster with $K_s > 2$ probably represents genes originated during polyploidization, before gamma triplication.

The distribution of K_a/K_s resulted in the identification of 555 chickpea genes with pair-wise $K_a/K_s > 1$ (Figure S15). These are probably newly evolving genes under positive selection pressure. More than 90% of the valid GOSlim matches of this gene set encode catalytic and binding activities, and 21 genes encode transporter activity (Figure S16).

Nucleotide diversity

For rapid assessment of nucleotide diversity at a genomic scale, sequence reads were generated by sequencing WGS libraries (Table S1b) of three other chickpea genotypes (ICCV2/IC12968, kabuli; JG62/ICC4951, desi; and PI489777, wild). Alignment of the reads to the draft genome sequence of ICC4958 and to the corresponding map-based assemblies of other genotypes showed that the nucleotide diversity based on genome-wide single nucleotide polymorphism (SNP) between each pair of the cultivated varieties varied from 0.00125 to 0.000915 bp^{-1}, which was less than the values for other *Fabaceae* family members (soybean, 0.00189 bp^{-1}; *M. truncatula*, 0.0043 bp^{-1}; Lam *et al.*, 2010; Branca *et al.*, 2011) and much lower than those for cereals (rice, 0.00229 bp^{-1}; maize, 0.0066 bp^{-1}; Caicedo *et al.*, 2007; Gore *et al.*, 2009). In comparison, higher average nucleotide diversity (0.0031 bp^{-1}) was observed

Table 2 Nucleotide diversity (per base) within cultivated and wild chickpea genotypes. Values represent the genome-wide (in bold) and transcriptome-based assessments, respectively

ICC4958	ICCV2	JG62	PI489777	Genotypes
	0.00125	**0.000958**	**0.00308**	ICC4958
0.00005		**0.000915**	**0.00400**	ICCV2
0.00003	0.00004		**0.00232**	JG62
0.00103	0.00082	0.00075		PI489777

between the cultivated and wild chickpea genotypes. Likewise, at the transcriptome level, very low nucleotide diversity was observed within the cultivated genotypes (Table 2). Overall, the average nucleotide diversity among various genotypes was less in the transcriptome as compared with the genome, indicating higher sequence conservation in the transcripts. In addition to closed flower pollination, a characteristic of the *Fabaceae* family members, low nucleotide diversity within the cultivated chickpea genotypes may have resulted from the evolutionary bottlenecks it suffered during domestication. Higher nucleotide diversities in *Medicago* and soybean genomes may also result from the higher frequency of cross-pollination in *Medicago* (Siol *et al.*, 2008; Bonnin *et al.*, 2010) and the higher rate of expected heterozygosity in soybean genome as a result of recent polyploidization (Schmutz *et al.*, 2010).

Development of genetic marker resources

Sequence assemblies of the cultivated and wild chickpea genotypes were compared and analysed to develop genetic marker resources for breeding programmes. Primer sequences for 30 000 non-monomeric simple sequence repeats (SSRs) from ICC4958 genome sequence, and primers for several hundreds of polymorphic non-monomeric SSRs and flanking bases of several thousands of SNPs between different combinations of sequenced chickpea genotypes were catalogued (Appendices S1–S13) to convert them into genetic markers. A high amplification efficiency and polymorphism potential for SNPs (90.75% for both) and a set of SSRs (98% and 94.7%, respectively) between ICC4958 and PI489777 were observed in experimental validations of the sequence-based predictions (Gaur *et al.*, 2012a; Jhanwar *et al.*, 2012).

DISCUSSION

We have presented the draft genome assembly of an economically important legume crop, chickpea (desi-type), based on next-generation sequencing data. This assembly is expected to provide the majority of the information required for a wide variety of fundamental and applied research, although a comprehensive approach of clone-by-clone sequencing, along with physical, genetic and cytogenetic mapping, is required to gain a deeper insight into the genome structure. After soybean and pigeonpea, this study reports a genome assembly of the third food legume crop. While this article was under review, a draft genome assembly of a kabuli-type chickpea was published (Varshney *et al.*, 2013). Chickpea is the only domesticated species of the *Cicerae* tribe, and its genome assembly is the second (after *Medicago*) of a member of the inverted-repeat-lacking clade (IRLC), members of which share a number of morphological features, including predominantly herbaceous habit, epulvinate compound leaves and base chromosome number 7–8 (Wojciechowski *et al.*, 2004). Nucleotide diversity analysis between the cultivated (desi and kabuli types) and wild genotypes has provided an assessment of a narrow genetic base in this crop, a limiting factor in the application of molecular breeding approaches. The wild progenitor of domesticated chickpea *C. reticulatum* grows in its native area at a high altitude, where it is exposed to sub-zero winter temperatures. The domestication of chickpea made it an autumn-sown crop in India and east Africa, and a spring-sown crop in the Mediterranean basin, probably to avoid *Ascochyta* blight (Abbo *et al.*, 2003a,b). During this process, winter hardiness and vernalization requirement alleles were lost (Summerfield *et al.*, 1989; Singh *et al.*, 1997). A comparison of sequences generated for the wild and the domesticated desi and kabuli genotypes is expected to be useful in identifying alleles for low-temperature tolerance and vernalization. Nucleotide diversity analysis between the cultivated and wild genotypes has provided a rich resource of polymorphic SSRs and SNPs between the chickpea genotypes, which is useful to establish the marker–trait relationship, especially for quantitative trait loci and molecular breeding.

More than 40% of the assembled draft genome represents interspersed repeats, including transposons and retrotransposons. This is considerably lower than soybean (59%) and pigeonpea (52%), but higher than *Medicago* (27%) and *Lotus* (34%) (Sato *et al.*, 2008; Schmutz *et al.*, 2010; Varshney *et al.* 2012; Young *et al.*, 2011). Although a lower proportion of repeats may reflect a lower level of pericentromeric sequencing, or the method used for sequencing and assembly, it may also be indicative of genomic differences between the milletoid (soybean and pigeonpea) and galegoid (*Medicago*, *Lotus* and chickpea) clades or genome size. The soybean genome is reported to have only 283 legume-specific gene families containing 448 genes (Schmutz *et al.*, 2010). We predicted 954 legume-specific genes in the present chickpea genome assembly. This number is expected to further increase with the availability of more legume genome sequences, because a number of genes previously found to be species-specific would show homology to the newly sequenced genes from other legumes. Features of the legume-specific and chickpea-specific genes are different from that of the other genes. Similar features of lineage-

specific genes, including short length, fewer number of introns and unusual GC content have also been observed in previous studies (Campbell *et al.*, 2007; Lin *et al.*, 2010). Although the origin of lineage-specific genes is not exactly clear, lateral gene transfer, gene duplication, followed by rapid sequence divergence, and *de novo* emergence from non-genic sequences may be attributed to the prediction of lineage-specific genes. Some of the lineage-specific genes predicted might result from genome assembly and annotation artifacts as well.

The quantification of gene expression levels provides clues about gene function and the molecular mechanisms underlying biological processes. A gene expression atlas covering various tissues has been reported for the model legumes of soybean and *Medicago* (Benedito *et al.*, 2008; Libault *et al.*, 2010). To present a global view of transcriptome activity of all the predicted genes in chickpea, we generated RNA-seq data from various tissues representing different tissues/organs and stress treatment. Based on RNA-seq data analysis, we could detect the expression of about 72% of the predicted genes. The comparison of gene expression profiles of all the genes within six different tissues/organs identified differentially expressed and tissue-specific genes, which presumably orchestrate their differentiation and development. The largest differences were observed in root (underground tissue) transcriptome, compared with the other (above ground) tissues analysed here, as was expected, and has been reported in other studies (Benedito *et al.*, 2008; Libault *et al.*, 2010; Garg *et al.*, 2011). Likewise, a maximum number of genes exhibited specific expression in root and flower tissues, which indicates major differences in the transcriptional programmes of these tissues, as compared with others. The genes involved in various biological processes were found to be specifically expressed in different tissues, which might play a crucial role in the biology of a particular tissue/organ. We also identified several genes responsive to drought and salt stresses related to various metabolic processes, regulation of transcription and transport. Several lineage-specific genes were also found to display tissue-specific and stress-responsive expression; many of these genes might presumably be involved in lineage-specific biological processes and adaptation. Overall, the tissue-specific and stress-responsive genes identified here will be very important in selecting the target genes for functional analysis.

Although the number of genes in the present chickpea genome assembly is equivalent to those in other dicots, such as Arabidopsis and grapevine it is much less than those in other sequenced legume genomes (Sato *et al.*, 2008; Schmutz *et al.*, 2010; Young *et al.*, 2011; Varshney *et al.*, 2012). However, when compared with soybean and Arabidopsis unigene data sets, chickpea genes exhibited a similar distribution of proportions of various functions. Out of aproximately 46 000 genes of soybean, aproximately 31 000 exist as paralogs. It was inferred that the pre-duplication proto-soybean possessed aproximately 30 000 genes, and expansion within soybean gene families has occurred because of recent genome duplication (Schmutz *et al.*, 2010). Considering a coverage of about 85% of genes, chickpea is also expected to have aproximately 32 000 genes. Therefore, the occurrence of fewer genes in this chickpea genome assembly, in comparison with other sequenced legume genomes, might not be the result of a contraction of genes in chickpea, but rather the result of expansions in gene complements in other legumes. It is evident that the high gene count in *Medicago* is the result of genome-wide extensive local gene duplication (Young *et al.*, 2011), which seems to be absent in chickpea. Apart from soybean, which has undergone a recent WGD, the average CDS lengths of four other sequenced legumes are less than those of the non-legume plants. As the average exon lengths of all the plants compared are similar, this may reflect a faster rate of gene evolution dynamics in legumes. Interestingly, the chickpea-specific and soybean-specific gene families (626 and 363, respectively) are fewer in number than the *Medicago*-specific or pigeonpea-specific gene families (3799 and 1389, respectively), indicating that a larger lineage-specific gain of genes occurred in *Medicago* and pigeonpea than in chickpea and soybean.

The draft genome sequence of chickpea and its analysis has provided rich information on the similarity and diversity of structural and organizational components in relation to other sequenced legume genomes. The comparison of cultivated and wild chickpea genomes and transcriptomes, along with gene expression analysis and certain evolutionary aspects, would help accelerate research efforts. A coordinated effort to use the resources made available through this study along with global germplasm collections is required to usher in an exciting era of genetic enhancement and breeding in chickpea for traits like stress tolerance, improved yield and symbiotic nitrogen fixation.

EXPERIMENTAL PROCEDURES

Sequencing and assembly

For the draft assembly, sequence data were generated primarily by the 454/Roche GS FLX Titanium platform (454, http://www.454.com, a Roche company) using pyrosequencing technology. The construction of the WGS (insert size of 300–900 bp) and matepair (MP; insert size of 3, 15 and 20 kb) libraries were performed as described by the manufacturer (Margulies *et al.*, 2005). The reads were filtered according to the method followed for *Solanum lycopersicum* (tomato) genome sequence (The Tomato Genome Consortium, 2012). The Illumina GA-IIx (Illumina Inc., San Diego, CA, USA) short-read sequencing platform was used to sequence two small-insert libraries (average insert size of 520 and 620 bp) to produce 43.7 Gb (aproximately 59X) paired-end (PE) high-quality sequence data of 100-base read lengths after quality filtering (Patel

and Jain, 2012). The short-read data set was assembled using ABYSS 1.2.6 with a K-mer length of 47 to produce 304 948 126 bases of assembled sequences. These contigs and all the filtered reads generated by the GS FLX platform were assembled by the *de novo* assembly tool NEWBLER 2.5.3 (GS *de novo* assembler; Roche Applied Sciences, http://www.roche-applied-science.com) to obtain the primary assembly. Further scaffolding using publicly available BAC end sequences (GenBank gi numbers 14 645 554–270 242 271) and genetic markers (Gaur *et al.*, 2012a) was performed. Sequencing and assembly methods are described in the Methods S1 in detail.

Repetitive element and gene prediction

Transposable elements in the chickpea genome were identified at both the DNA and protein levels using REPEATMASKER and REPEATPROTEINMASK (Chen, 2004). In addition, four software packages, REPEATMODELER, PILER, REPEATSCOUT and LTR_FINDER (Edgar and Myers, 2005; Price *et al.*, 2005; Xu and Wang, 2007), were used for *de novo* identification of repeat sequences in the chickpea genome. The prediction of protein-coding genes from the repeat masked genome involved three approaches – *ab initio*, and homology- and EST-based approaches – and finally integrating them using GLEAN (Elsik *et al.*, 2007) and EVIDENCEMODELER (Haas *et al.*, 2008). For *ab initio* prediction, we used AUGUSTUS (Stanke and Waack, 2003) and GENSCAN (Salamov and Solovyev, 2000) with parameters trained on Arabidopsis, and FGENESH++ trained on *Medicago*, to generate gene models. For homology-based prediction, we first aligned the protein sequences from nine sequenced plants [Arabidopsis, *Carica papaya* (papaya), grapevine, *Populus* (poplar), cucumber, rice, *Zea mays* (maize), soybean and pigeonpea] onto the chickpea genome using TBLASTN with a cut-off of $1e^{-5}$, followed by alignment of homologous chickpea genomic sequences with matching proteins using GENEWISE (Birney *et al.*, 2004) for the prediction of accurate spliced alignment. In the EST-based approach, we aligned all the ESTs available at NCBI after cleaning (41 045) and non-redundant Roche 454 reads ($\geq$200 bp; 811 558) from our previous study (Garg *et al.*, 2011) for the prediction of spliced alignments using PASA (Campbell *et al.*, 2006). The outputs from all three approaches were integrated by GLEAN and EVM to generate consensus gene sets. Furthermore, the outputs of both GLEAN (20 791) and EVM (27 203) were combined to filter the non-redundant set of protein-coding genes. PASA was run on this gene set for the identification of spliced variants and addition of the untranslated regions (UTRs). This resulted in a total of 31 862 transcripts (Table S10). Furthermore, three genes predicted by CEGMA analysis (Parra *et al.*, 2007) in the genome but not predicted by the above strategies were added to the predicted gene set. We filtered out 21 genes with coding sequences of <90 bp and those predicted across the sequencing gaps (coding sequence with N content of $\geq$50%). Finally, a total of 31 844 transcript sequences were obtained, representing the final set of 27 571 protein-coding genes (CGAP v1.0). All the protein-coding genes were assigned a unique identifier number (from Ca_00001 to Ca_27571).

Functional annotation

Putative gene function was assigned to the chickpea genes based on the best alignment to the protein sequences in SwissProt, TrEMBL and TAIR10 databases using BLASTP with a cut-off of $1e^{-5}$. Gene ontology terms were assigned to the genes using the Blast2GO pipeline (Conesa *et al.*, 2005). PFAM domains in the chickpea genes were identified using the AutoFACT pipeline (Koski *et al.*, 2005). The proteins encoding transcription factors were identified using HMMs downloaded from PFAM and Plant Transcription Factor (PlnTFDB) databases, as previously described (Garg *et al.*, 2011).

Multi-species gene family analyses and identification of lineage-specific genes

The proteomes of chickpea, soybean, pigeonpea, *Medicago* and grapevine were used for multi-species gene family analyses. TRIBE-MCL (Enright *et al.*, 2002) was used to generate clusters from all-against-all BLASTP search results using $I = 6$ and scheme = 4, keeping other parameters at their default values. To reduce noise, genes with <25% of median similarity hits within the clusters were marked as outliers. The identification of lineage-specific genes in chickpea was performed using the strategy as described by Garg *et al.* (2011). The chickpea genes showing significant hits (cut-off $\leq 1e^{-5}$) with genome/proteome, plant transcript assemblies and EST/unigene sequences of *Fabaceae* plant species only, were identified as legume-specific candidates, and those not showing significant similarity with any of the plant species were identified as candidate chickpea-specific genes.

Gene expression analysis

RNA-seq was performed with total RNA isolated from different chickpea tissues/organs (root, shoot, mature leaves, stem, flowers and young pod) and roots of seedlings subjected to control, drought and salt stress conditions. 51- or 54-bp-long single-end read sequencing was performed using the Illumina GA-II platform. High-quality reads filtered with NGS QC TOOLKIT (Patel and Jain, 2012) were mapped using CLC GENOMICS WORKBENCH to the mRNA sequences of predicted chickpea genes, for quantification of gene expression, allowing two mismatches. Only the uniquely mapped reads were considered for gene expression analysis. Differential gene expression analysis was performed using DESEQ (Anders and Huber, 2010). The genes showing a fold-change of at least two-fold with $P \leq 0.05$ were regarded as differentially expressed. GO-term enrichment analysis was performed using BINGO (Maere *et al.* 2005). The detailed method is described in Methods S1.

Synteny and genome duplication

Duplicated blocks within the chickpea genome were identified using I-ADHORE 3.0 (Proost *et al.*, 2012), with various parameters, including alignment method gg2, gap_size 30, cluster_gap 35, q_value 0.85, probability_cut-off 0.01, anchor_points 4, table_type family and level_2_only false. Protein sequences corresponding to detected anchor points or collinear regions were aligned using CLUSTALW, followed by an estimation of the synonymous substitution rate (K_s) using CODEML in PAML 4.5 (Yang, 2007). Collinear blocks were positioned and visualized on the genome using Circos (Krzywinski *et al.*, 2009). The whole genome dot plot was generated with chickpea scaffolds representing eight linkage groups on the *x*-axis against chromosome arms (north and south) of soybean and *M. truncatula*. The PROMER package of MUMMER 3.22 (Delcher *et al.*, 2002) was used to align annotated genes based on amino acid sequence. Whole genome dot plots were generated using MUMMERPLOT (Delcher *et al.*, 2002) and GNUPLOT 4.4 (www.gnuplot.info/) patch level 2.

Development of a genetic marker resource

The coordinates of the SNPs between two genotypes obtained from the annotation files generated using GS REFERENCE MAPPER and the flanking bases were extracted by custom-made PERL scripts. The transcriptome sequencing of three genotypes

(ICC4958, PI489777 and ICCV2) has been reported earlier (Garg *et al.*, 2011; Agarwal *et al.*, 2012; Jhanwar *et al.*, 2012). The transcriptome of genotype JG62 was sequenced using the Roche 454 platform. SNP detection among the transcriptomes of different chickpea genotypes was performed using GIGABAYES, as previously described (Jhanwar *et al.*, 2012). SSRs in each genotype were detected using the MISA tool. Identification of polymorphic SSRs between two genotypes was performed using the coordinates and MISA. The flanking bases were extracted by custom-made PERL scripts. Primer sequences were determined in batch mode by the PRIMER 3 tool (Rozen and Skaletsky, 2000). The flanking bases of SNPs, and the primers for the SSRs are listed in Appendices S1–S13 at http://nipgr.res.in/CGAP/home.php.

ACKNOWLEDGEMENTS

We acknowledge funding from the Department of Biotechnology, Government of India, under the Next Generation Challenge Programme in Chickpea Genomics (grant no. BT/PR12919/AGR/02/676/2009 from 2009-14). Authors declare no conflict of interest.

ACCESSION CODES

Genome assembly is available at National Centre for Biotechnology Information (NCBI) as Bioproject ID PRJNA78951 (GenBank accession no. AHII00000000) (http://www.ncbi.nlm.nih.gov/bioproject?term=PRJNA78951). The sequence data are available in Short Read Archive under accession numbers SRA053228, SRA053197, SRA053141 and SRA053687. Genome assembly, annotation data and all the supplementary figures, tables, data are available for viewing and downloading at http://nipgr.res.in/CGAP/home.php.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Fragment distribution of the *de novo* assembly of ICC4958.

Figure S2. Read depth at assembled bases of chickpea ICC4958 (based on 454/Roche read alignment).

Figure S3. GC content distribution in the genome sequence of chickpea, as compared with other plant species.

Figure S4. Top-20 GO terms represented in the chickpea gene set.

Figure S5. Top-20 PFAM domains represented in the chickpea gene set.

Figure S6. Strategy for the identification of lineage-specific genes in the chickpea genome.

Figure S7. Top-10 GO terms represented in the genes included in chickpea-specific gene families.

Figure S8. Gene distribution in different transcription factor families in chickpea, other sequenced legumes and Arabidopsis.

Figure S9. Phylogenetic analysis of chickpea and *Medicago* genes belonging to CC-NBS-LRR (a) and Leghaemoglobin (b) families.

Figure S10. K_s distribution analysis of paralogous chickpea gene pairs to determine the genome duplication event.

Figure S11. The whole genome dot plot was generated between chickpea linkage groups (*x*-axis) and *Medicago truncatula* chromosome arms (*y*-axis).

Figure S12. Microsynteny of chickpea (Ca) LG 5 with *Medicago truncatula* (Mt) chromosome 3.

Figure S13. The whole genome dot plot was generated between chickpea linkage groups (*x*-axis) and *Glycine max* chromosome arms (*y*-axis).

Figure S14. Scatter plot showing the distribution of K_a/K_s (ω) with respect to K_s between gene pairs present in the collinear blocks of chickpea and *Medicago*.

Figure S15. K_a/K_s distribution analysis of chickpea gene pairs.

Figure S16. Distribution of various GOSlim categories (level 2) in chickpea gene pairs with $K_a/K_s > 1$.

Table S1. Sequencing data generated for chickpea genotypes.

Table S2. Statistics of draft assembly.

Table S3. Anchoring of scaffolds to linkage groups.

Table S4. Estimation of chickpea genome length based on read alignment.

Table S5. Estimated heterozygosity in ICC4958 draft genome.

Table S6. Transcriptome coverage in the assembled chickpea genome.

Table S7. Repeat content in the assembled chickpea draft genome.

Table S8. Comparative analysis of microsatellite sequences in chickpea draft genome with those in other legumes.

Table S9. Statistics of protein-coding gene prediction.

Table S10. Assessment of gene prediction using the CEGMA pipeline.

Table S11. Experimental evidence for the predicted protein-coding genes.

Table S12. Statistics of protein-coding genes from different plant species.

Table S13. Functional annotation of the predicted protein-coding genes.

Table S14. Features of lineage-specific genes in chickpea.

Table S15. Transcription factor/regulator families in the chickpea draft genome.

Table S16. Comparison of *R*-gene family in chickpea draft genome with other sequenced plant genomes.

Table S17. Comparison of nodulation-associated gene families in chickpea draft genome with other sequenced plant genomes.

Table S18. Comparison of number of genes associated with carotenoid and flavonoid metabolism in chickpea draft genome with other sequenced plant genomes.

Table S19. Non-coding RNA genes in the chickpea draft genome.

Table S20. Summary of RNA-seq data generated from different tissues/treatments to study gene expression.

Table S21. Summary of tissue-preferential and stress-responsive gene expression results based on RNA-seq data.

Table S22. GO terms enriched in the chickpea genes expressed in tissue-specific manner.

Appendices S1–13. SNP and SSR marker resources.

Methods S1. Experimental methods and URLs used.

REFERENCES

Abbo, S., Berger, J. and Turner, N.C. (2003a) Evolution of cultivated chickpea: four bottlenecks limit diversity and constrain adaptation. *Funct. Plant Biol.* **30**, 1081–1087.

Abbo, S., Shtienberg, D., Lichtenzveig, J., Lev-Yadun, S. and Gopher, A. (2003b) The chickpea, summer cropping, and a new model for pulse domestication in the ancient near East. *Quant. Rev. Biol.* **78**, 435–448.

Agarwal, G., Jhanwar, S., Priya, P., Singh, V.K., Saxena, M.S., Parida, S.K., Garg, R., Tyagi, A.K. and Jain, M. (2012) Comparative analysis of kabuli chickpea transcriptome with desi and wild chickpea provides a rich resource for development of functional markers. *PLoS ONE*, **7**, e52443.

Al-Dous, E.K., George, B., Al-Mahmoud, M.E. *et al.* (2011) *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nat. Biotechnol.* **29**, 521–527.

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106.

Argout, X., Salse., J., Aury, J.M. *et al.* (2011) The genome of *Theobroma cacao*. *Nat. Genet.* **43**, 101–108.

Arnon, I. (1972) *Crop Production in Dry Regions*, Vol. I. London: Leonard Hill, pp. 460–462.

Arumuganathan, K. and Earle, E.D. (1991) Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9**, 208–218.

Benedito, V.A., Torres-Jerez, I., Murray, J.D. *et al.* (2008) A gene expression atlas of the model legume *Medicago truncatula*. *Plant J.* **55**, 504–513.

Birney, E., Clamp, M. and Durbin, R. (2004) Genewise and genomewise. *Genome Res.* **14**, 988–995.

Bonnin, I., Ronfort, J., Wozniak, F. and Olivieri, I. (2010) Spatial effects and rare outcrossing events in *Medicago truncatula* (Fabaceae). *Mol. Ecol.* **10**, 1371–1383.

Branca, A., Paape, T.D., Zhou, P. *et al.* (2011) Whole-genome nucleotide diversity, recombination, and linkage disquilibrium in the model legume *Medicago truncatula*. *Proc. Natl Acad. Sci. USA*, **108**, E864–E870.

Caicedo, A.L., Williamson, S.H., Hernandez, R.D. *et al.* (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet.* **3**, e163.

Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M. and Buell, C.R. (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC Genomics*, **7**, 327.

Campbell, M.A., Zhu, W., Jiang, N., Lin, H., Ouyang, S., Childs, K.L., Haas, B.J., Hamilton, J.P. and Buell, C.R. (2007) Identification and characterization of lineage-specific genes within the Poaceae. *Plant Physiol.* **145**, 1311–1322.

Cannon, S.B., Sterck, L., Rombauts, S. *et al.* (2006) Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *Proc. Natl Acad. Sci. USA*, **103**, 14959–14964.

Chen, N. (2004) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics*, **25**, 4.10.1–4.10.14.

Choi, H.K., Mun, J.H., Kim, D.J. *et al.* (2004) Estimating genome conservation between crop and model legume species. *Proc. Natl Acad. Sci. USA*, **101**, 15289–15294.

Comal, L. (2005) The advantages and disadvantages of being polyploidy. *Nat. Rev. Genet.* **6**, 836–846.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Delcher, A.L., Phillippy, A., Carlton, J. and Salzberg, S.L. (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478–2483.

Edgar, R.C. and Myers, E.W. (2005) PILER: identification and classification of genomic repeats. *Bioinformatics*, **21**(Suppl), i152–i158.

Elsik, C.G., Mackey, A.J., Reese, J.T., Milshina, N.V., Roos, D.S. and Weinstock, G.M. (2007) Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13.

Enright, A.J., Van Dongen, S. and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584.

FAOSTAT (2009) Faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567 anchor.

Fawcett, J.A., Maere, S. and van de Peer, Y. (2009) Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc. Natl Acad. Sci. USA*, **106**, 5737–5742.

Garg, R., Patel, R.K., Jhanwar, S., Priya, P., Bhattacharjee, A., Yadav, G., Bhatia, S., Chattopadhyay, D., Tyagi, A.K. and Jain, M. (2011) Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. *Plant Physiol.* **156**, 1661–1678.

Gaur, R., Azam, S., Jeena, G., Khan, A.W., Choudhary, S., Jain, M., Yadav, G., Tyagi, A.K., Chattopadhyay, D. and Bhatia, S. (2012a) High-throughput SNP discovery and genotyping for constructing a saturated linkage map of chickpea (*Cicer arietinum* L.). *DNA Res.* **19**, 357–373.

Gaur, P.M., Jukanti, A.K. and Varshney, R.K. (2012b) Impact of genomic technologies on chickpea breeding strategies. *Agronomy*, **2**, 199–221.

Gore, M.A., Chia, J.M., Elshire, R.J. *et al.* (2009) A first-generation haplotype map of maize. *Science*, **326**, 1115–1117.

Graham, P.H. and Vance, C.P. (2003) Legumes: importance and constraints to greater use. *Plant Physiol.* **131**, 872–877.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R. and Wortman, J.R. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7.

Huang, S., Li, R., Zhang, Z. *et al.* (2009) The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* **41**, 1275–1281.

Hunter, S., Apweiler, R., Attwood, T.K. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215.

Jaillon, O., Aury, J.M., Noel, B. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.

Jhanwar, S., Priya, P., Garg, R., Parida, S.K., Tyagi, A.K. and Jain, M. (2012) Transcriptome sequencing of wild chickpea as a rich resource for marker development. *Plant Biotechnol. J.* **10**, 690–702.

Koski, L.B., Gray, M.W., Lang, B.F. and Burger, G. (2005) AutoFACT: an automatic functional annotation and classification tool. *BMC Bioinformatics*, **6**, 151.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645.

Lam, H.M., Xun, X., Liu, X. *et al.* (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* **42**, 1053–1059.

Lavin, M., Herendeen, P.S. and Wojciechowski, M.F. (2005) Evolutionary rates analysis of leguminosae implicates a rapid diversification of lineages during the tertiary. *Syst. Biol.* **54**, 575–594.

Libault, M., Joshi, T., Benedito, V.A., Xu, D., Udvardi, M.K. and Stacey, G. (2009) Legume transcription factor genes: what makes legumes so special? *Plant Physiol.* **151**, 991–1001.

Libault, M., Farmer, A., Joshi, T., Takahashi, K., Langley, R.J., Franklin, L.D., He, J., Xu, D., May, G. and Stacey, G. (2010) An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. *Plant J.* **63**, 86–99.

Lin, H., Moghe, G., Ouyang, S., Iezzoni, A., Shiu, S.H., Gu, X. and Buell, C.R. (2010) Comparative analyses reveal distinct sets of lineage-specific genes within *Arabidopsis thaliana*. *BMC Evol. Biol.* **10**, 41.

Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964.

Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Maere, S., Heymans, K. and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

Margulies, M., Egholm, M., Altman, W.E. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Nawrocki, E.P., Kolbe, D.L. and Eddy, S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.

Otto, S.P. and Whitton, J. (2000) Polyploid incidence and evolution. *Annu. Rev. Genet.* **34**, 401–437.

Parra, G., Bradnam, K. and Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Patel, R.K. and Jain, M. (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE*, **7**, e30619.

Paterson, A.H., Bowers, J.E. and Chapman, B.A. (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl Acad. Sci. USA*, **101**, 9903–9908.

Pfeil, B.E., Schlueter, J.A., Shoemaker, R.C. and Doyle, J.J. (2005) Placing paleoploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. *Syst. Biol.* **54**, 441–454.

Price, A.L., Jones, N.C. and Pevzner, P.A. (2005) *De novo* identification of repeat families in large genomes. *Bioinformatics*, **21**(Suppl), i351–i358.

Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van de Peer, Y. and Vandepoele, K. (2012) i-Adhore 3.0-fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.* **40**, e11.

Rozen, S. and Skaletsky, H. (2000) Primer 3 on WWW for general users and for biologists programmers. *Methods Mol. Biol.* **132**, 365–386.

Salamov, A.A. and Solovyev, V.V. (2000) *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522.

Sanseverino, W., Roma, G., De Simone, M., Faino, L., Melito, S., Stupka, E., Frusciante, L. and Ercolano, M.R. (2010) PRGdb: a bioinformatics platform for plant resistance gene analysis. *Nucleic Acids Res.* **38**, D814–D821.

Sato, S., Nakamura, Y., Kaneko, T. *et al.* (2008) Genome structure of the legume, *Lotus japonicus*. *DNA Res.* **15**, 227–239.

Schmutz, J., Cannon, S.B., Schlueter, J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J. and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123.

Singh, K.B., Malhotra, R.S., Saxena, M.C. and Bejiga, G. (1997) Superiority of winter sowing over traditional spring sowing of chickpea in the Mediterranean region. *Agron. J.* **89**, 112–118.

Siol, M., Prosperi, J.M., Bonnin, I. and Ronfort, J. (2008) How multilocus genotypic pattern helps to understand the history of selfing populations: a case study in *Medicago truncatula*. *Heredity*, **100**, 517–525.

Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**(Suppl), ii215–ii225.

Summerfield, R.J., Ellis, R.H. and Roberts, E.H. (1989) Vernalization in chickpea (*Cicer arietinum*): fact or artefact? *Ann. Bot.* **64**, 599–604.

Tang, H., Wang, X., Bowers, J.E., Ming, R., Alam, M. and Paterson, A.H. (2008) Unraveling ancient hexaploidity through multiply-aligned angiosperm gene maps. *Genome Res.* **18**, 1944–1954.

Taylor, J.S. and Raes, J. (2004) Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.* **38**, 615–643.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

Thudi, M., Bohra, A., Nayak, S.N. *et al.* (2011) Novel SSR markers from BAC-end sequences, DArT arrays and a comprehensive genetic map with 1291 marker loci for chickpea (*Cicer arietinum* L.). *PLoS ONE*, **6**, e27275.

Varshney, R.K., Chen, W., Li, Y. *et al.* (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**, 83–89.

Varshney, R.K., Song, C., Saxena, R.K. *et al.* (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.* **31**, 240–246.

Wilf, P. and Johnson, K.R. (2004) Land plant extinction at the end of the Cretaceous: a quantitative analysis of the North Dakota megafloral record. *Paleobiology*, **30**, 347–368.

Wojciechowski, M.F., Lavin, M. and Sanderson, M.J. (2004) A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. *Am. J. Bot.* **91**, 1846–1862.

Xu, Z. and Wang, H. (2007) LTR_FINDER: an efficient tool for the prediction of full- length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.

Young, N. and Bharti, A. (2012) Genome-enabled insights into legume biology. *Annu. Rev. Plant Biol.* **63**, 283–305.

Young, N.D., Debellé, F., Oldroyd, G.E. *et al.* (2011) The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature*, **480**, 520–524.

Zohary, D. and Hopf, M. (2000) *Pulse. Domestication of plants in the old world*. New York: Oxford University Press, pp. 108–111.

