9 Sequencing Pigeonpea Genome

Vikas K. Singh, Rachit K. Saxena and Rajeev K. Varshney

Abstract

Availability of draft genome has brought quantum jump in pigeonpea status and facilitated to move it to the league of genomic resource rich crops. It is important to mention that pigeonpea became the first orphan and non-industrial grain legume in 2012 to have the draft genome sequence. An elite pigeonpea genotype Asha (ICPL 87119) was used to develop the draft genome in two different sequencing efforts. The pigeonpea genome sequence effort led by International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) used Illumina Genome Analyzer and HiSeq 2000 NGS platform, and a total of 237.2 Gb of sequence was generated. *De-novo* genome assembly combined with Sanger-based bacterial artificial chromosome end sequences and a genetic map was used to assemble raw reads into scaffolds representing 72.7% (605.78 Mb) of the 833.07 Mb pigeonpea genome. Genome analysis predicted 48,680 genes with an average transcript length of 2348 bp, coding-sequence size of 959.35 bp and 3.59 exons per gene. Analysis of genome assembly for repetitive DNA showed presence of transposable elements (TEs) in 49.61% of assembled genome. The pigeonpea genome sequencing led by National Research Centre on Plant Biotechnology (NRCPB) used 454 GS-FLX sequencing chemistry, with mean read lengths of >550 bp and >10-fold genome coverage, was used to assemble ~511 Mb sequence data. In this study, 47,004 protein-coding genes were predicted. This study also reported 1213 disease resistance/defense

V.K. Singh · R.K. Saxena · R.K. Varshney (✉)
International Crops Research Institute for the
Semi-Arid Tropics, Patancheru 502324, India
e-mail: r.k.varshney@cgiar.org

R.K. Varshney et al. (eds.), *The Pigeonpea Genome*,
Compendium of Plant Genomes, DOI 10.1007/978-3-319-63797-6_9

response genes and 152 abiotic stress tolerance genes. The available pigeonpea draft genome information is expected to facilitate genomics-assisted breeding for the targeted traits that could improve food security in many developing countries.

9.1 Introduction

In continuation to the previous chapter in the book entitled "Whole-genome sequencing of pigeonpea: requirement, background history, current status and future prospects for crop improvement," we would like to present a focused chapter on de novo sequencing the pigeonpea genome. In the previous chapter, we have discussed about the background history of two genome sequencing efforts by Varshney et al. (2012) and Singh et al. (2012). However, in this short chapter, we present detailed comparisons between above-mentioned de novo sequencing projects in terms of (i) sequencing data, (ii) draft genome assemblies statistics, (iii) repetitive sequences in genome, (iv) gene annotation, (v) genome duplication and synteny with sequenced plant genomes, and (vi) novel marker and genes repertoire. Further, we have also presented utility of pigeonpea genome sequence by providing one example in soybean for crop improvement.

9.2 Sequencing Data

Genome sequencing of pigeonpea was undertaken in two different studies. Whole-genome shotgun sequencing strategy was used in both the studies. Illumina Genome Analyzer and Hiseq 2000 Sequencing System was used by Varshney et al. (2012), and 454 GS-FLX Phase D platform was used by Singh et al. (2012). Both of these next generation sequencing platforms have their advantages and disadvantages (Luo et al. 2012). In the case of Varshney et al. (2012), 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 were used for sequencing on Illumina platforms. In the case of Singh et al. (2012), GS-FLX sequencing was undertaken on 20-kb-long fragments sequencing library. In these studies, 237.2 Gb and 10.1 Gb sequencing data were generated on Illumina sequencing and GS-FLX sequencing platform, respectively. Further, to reduce the effect of sequencing errors to the assemblies, a series of checking and filtering steps on reads generated were performed. After applying stringent criteria, only 130.7 Gb and 9.48 Gb data were used for developing draft genome assemblies by Varshney et al. (2012) and Singh et al. (2012), respectively.

9.3 Draft Genome Assemblies Statistics

In general, both studies had taken a series of steps to assemble the filtered/corrected sequencing reads. As the first step, raw sequencing reads were aligned to form contigs and then calculated the amount of shared PE relationships between each pair of contigs, and then constructed the scaffolds. Subsequently in Varshney et al. (2012), BAC-ends sequences (BESs) were used for mapping of scaffolds to obtain the super scaffolds, and a genetic map (*Cajanus cajan* ICP 28 × *C. scaraboides* ICPW 94) was used in developing the final scaffolds or pseudomolecules. A total of 137,542 and 59,681 scaffolds spanning 60,578 Mb and 5108 Mb genome assemblies were developed in Varshney et al. (2012) and Singh et al. (2012), respectively. The N50 of the assembly was 51,606 kb (scaffolds) in Varshney et al. 2012 and 4522 bp (contig) in Singh et al. (2012).

9.4 Repetitive Sequences in the Genome

In Varshney et al. (2012), repetitive DNA (excluding low-complexity sequences) was identified in 51.67% of the genome, most of which could not be associated with known transposable element (TE) families by the *de novo* repeat identification using RepeatModeler and homology analysis against the RepBase library. Majority of repetitive sequences were classified as retro-transposons (37.12%), whereas 8.77% of the transposable elements were DNA transposons. Long-terminal repeat elements, of which 22.81% are *Gypsy*-type elements and 12.04% are *Copia*-type element, were the most abundant. *De novo* analysis of RE were conducted using RepeatModeler software in the other study (Singh et al. 2012). As a result, a total of 1,127,729 REs in the were identified (63.95%) in the genome which covers a total of 326,671,068 bp sequences. Majority of the RE was retro-transposons (23.6%) (including Line: 1.03%; Copia 6.1%; and Gypsy 16.02%) and 2.99% was DNA transposons, whereas 66.2% was unclassified. Simple direct repeats and low-complexity repeats represented only 2.57% and 4.63% of the total RE, respectively.

9.5 Genome Annotation

Genome analysis combined with *de novo* gene prediction programs identified 48,680 pigeonpea genes (Varshney et al. 2012) (Table 9.1). The average transcript length was found to be of 2,348.70 bp, coding-sequence size of 959.35 bp and 3.59 exons per gene. *De novo* gene prediction supported majority of these predicted genes (99.6%). The annotation of the pigeonpea genome was found completed by observing 453 out of 458 (98.9%) KOGs within the pigeonpea gene set. The genes that have been predicted in pigeonpea genome are comparable to poplar (*Populus trichocarpa*), soybean (*Glycine max*), and *Medicago truncatula.* The average length of exon and intron in pigeonpea genome was found to be 267.39 bp and 536.89 bp, respectively, whereas the average number of exons per gene is 3.59. A total of 46,750 (96.04%) genes were found to be similar to entries in databases to tentatively assign gene functions and 1930 (3.96%) genes remain unannotated. Further, 862 microRNA (miRNA), 763 tRNA, 329 rRNA, and 363 small nuclear (snRNA) genes were identified in the pigeonpea genome set in addition to the protein-coding genes.

In the case of Singh et al. (2012), FGENESH software was used for gene prediction using 454 GS-FLX large sequence contigs containing ~511 Mb of high-quality sequence. A total of 59,515 genes were predicted with average gene size of 1170 bp. The gene with largest size was of 11,523 bp and the gene with smallest size of 501 bp. The average exon and intron sizes were 268 bp and 288 bp, respectively. The predicted 99.9% of the genes showed significant matches within the pigeonpea transcriptome database. A total of 47,004 protein-coding genes and 12,511 transposable elements related genes were reported. Additionally, 1213 disease resistance/defense response genes and 152 abiotic stress tolerance genes in the pigeonpea genome were also reported (Table 9.1).

After going in detail on these studies for last 2–3 years, we understand the number of genes predicted in both the genome assemblies mentioned above is an overestimate. This may be due to the quality of final genome assemblies which seemed to be fragmented. Improved version of assembly in near future may provide us the accurate number of genes in pigeonpea.

9.6 Genome Duplication and Synteny with Sequenced Plant Genomes

The synteny analysis in Varshney et al. (2012) revealed that pigeonpea diverged from soybean ~20–30 Myrs ago. In spite of this long period of divergence, high levels of synteny were observed between pigeonpea and soybean as well as between pigeonpea and the galegoid species *M. truncatula* and *Lotus japonicus*. Each pigeonpea chromosome showed extensive

Table 9.1 Comparative account on gene annotation in pigeonpea in two studies

Features	Varshney et al. (2012)	Singh et al. (2012)
Number of protein-coding genes	48,680	47,004
Number of gene models (non-TE containing)	40,071	34,493
Mean transcript length	2,348.70 bp	–
Mean coding-sequence length	959.35 bp	1170 bp
Mean number of exons per gene	3.59	4.90
Mean exon length	267.39 bp	268 bp
Mean intron length	536.89 bp	288 bp
Number of genes annotated	46,750 (96.04%)	–
Number of genes unannotated	1930 (3.96%)	–
Number of miRNA genes	862	100
Number of rRNA genes	329	448
Number of tRNA genes	763	671
Number of snRNA genes	363	226

synteny with two or more than two chromosomes in soybean, likely due to the independent duplication event in soybean following divergence from pigeonpea. The close relationships between pigeonpea and soybean genomes were also detected in Singh et al. (2012). In this study, a total of 31,937 (67.94%) of the pigeonpea genes showed synteny with soybean genes, whereas 9067 genes were unique to pigeonpea.

9.7 Novel Marker and Genes Repertoire

The completion of the pigeonpea genome has made a significant contribution to the genomic resources available for pigeonpea through sequencing of the pigeonpea genome. In Varshney et al. (2012), a total of 309,052 simple sequence repeats (SSRs) and 28,104 novel single nucleotide polymorphisms (SNPs) were identified. Further, a detailed comparative analysis has identified 111 drought-responsive genes for drought tolerance, an important trait that can be transferred to other legume crops, whereas in Singh et al. (2012) study, 1,89,895 SSRs comprising of 100,373 mono-nucleotide, 49,325 di-nucleotide, 18,505 tri-nucleotide, 2217 tetra-nucleotide, 512 penta-nucleotides, 815 hexa-nucleotide, and 18,148 compound repeats were reported. A total of 437 SSRs were experimentally validated for PCR amplification and high rate of polymorphism among pigeonpea varieties were reported.

9.8 Application of Pigeonpea Genome Sequence

The genome sequence provided hope to the pigeonpea community to use the genome sequence to harness pigeonpea's genetic diversity at genome level and to identify the molecular markers and genes for targeted traits. Such information will allow researchers to develop superior varieties and parental lines of hybrids in pigeonpea. The genome sequence will also be useful in identifying germplasm lines or advanced breeding lines with a broader genetic base for future breeding programs. Modern genetics and breeding approaches such as genotyping by sequencing, marker-assisted recurrent selection, and genomic selection will now be possible in this crop to improve the efficiency of pigeonpea breeding. Genome sequence will be useful in utilizing gene sequences in genetic engineering approaches also. Several projects are underway at present to harness the full potential of pigeonpea genome for crop improvement (Varshney 2016).

It is also important to mention here that the pigeonpea genome sequence information has also been used for crop improvement in other species like soybean. A very first example has come recently where pigeonpea genome has been used to bring rust resistance in soybean (Kawashima et al. 2016). In soybean, Asian soybean rust (ASR) is one of the most economically important diseases. This disease can only treatable with use of fungicides. However, due to the emergence of fungicide resistance in pest, it becomes less effective. Moreover, there are no commercial soybean cultivars with durable resistance. Interestingly, a gene *CcRpp1* (*Cajanus cajan* Resistance against *Phakopsora pachyrhizi 1*) from pigeonpea has been found useful to confer resistance to ASR in soybean. By analyzing the pigeonpea genome sequence, an intracellular immune receptor from pigeonpea was identified and transferred into soybean shows that *CcRpp1* confers full resistance to ASR in soybean. This will be helpful in achieving a higher level of resistance, which might provide commercial control superior to current strategies. This study has clearly shown the importance of pigeonpea genome and opened new avenues for its use not only in pigeonpea but also in other crops species (primarily closely related legumes) for crop improvement.

Acknowledgements Authors are thankful to the United States Agency for International Development (USAID), Biotechnology Industry Partnership Programme (BIPP) and Department of Biotechnology of Government of India for supporting various activities related to pigeonpea genomics and molecular breeding projects at ICRISAT. This work has been undertaken as part of the CGIAR Research Program on Grain Legumes. ICRISAT is a member of CGIAR Consortium.

References

Kawashima CG, Guimarães GA, Nogueira SR, MacLean D, Cook DR, Steuernagel B, Baek J, Bouyioukos C, do VA Melo B, Tristão G, de Oliveira JC, Rauscher G, Mittal S, Panichelli L, Bacot K, Johnson E, Iyer G, Tabor G, Wulff BHB, Ward E, Rairdan GJ, Broglie KE, Wu G, van Esse HP, Jones JDG, Brommonschenke SH (2016) A pigeonpea gene confers resistance to Asian soybean rust in soybean. Nat Biotechnol 34:661–665

Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT (2012) Correction: direct comparisons of illumina vs. roche 454 sequencing technologies on the same microbial community dna sample. PLOS ONE 7 (3):10–1371

Singh NK, Gupta DK, Jayaswal PK, Mahato AK, Dutta S, Singh S, Bhutani S, Dogra V, Singh BP, Kumawat G, Pal JK, Pandit A, Singh A, Rawal H, Kumar A, Prashat RG, Khare A, Yadav R, Raje RS, Singh MN, Datta S, Fakrudin B, Wanjari KB, Kansal R, Dash PK, Jain PK, Bhattacharya R, Gaikwad K, Mohapatra T, Srinivasan R, Sharma TR (2012) The first draft of the pigeonpea genome sequence. J Plant Biochem Biotechnol 21:98–112

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MTA, Azam S, Fan G, Whaley AM, Farmer AD, Sheridan J, Iwata A, Tuteja R, Penmetsa RV, Wu W, Upadhyaya HD, Yang SP, Shah T, Saxena KB, Michael T, McCombie WR, Yang B, Zhang G, Yang H, Wang J, Spillane C, Cook DR, May GD, Xu X, Jackson SA (2012) Draft genome sequence of pigeonpea *(Cajanus cajan)*, an orphan legume crop of resource-poor farmers. Nat Biotechnol 30:83–89

Varshney RK (2016) Exciting journey of 10 years from genomes to fields and markets: some success stories of genomics-assisted breeding in chickpea, pigeonpea and groundnut. Plant Sci 242:98–107