

1 **Refinement of Draft Genome Assemblies of Pigeonpea (*Cajanus cajan*)**

2 Soma Marla<sup>a\*</sup>, Pallavi Mishra<sup>a</sup>, Ranjeet Maurya<sup>a</sup>, Mohar Singh<sup>a</sup>, D. P. Wankhede<sup>a</sup>, Anil. K.  
3 Gupta<sup>b</sup>, N. S. Rao<sup>c</sup>, S. K. Singh<sup>a</sup>, Rajesh Kumar<sup>a</sup>

4 **AUTHOR AFFILIATIONS**

5 <sup>a</sup>ICAR-National Bureau of Plant Genetic Resources, New Delhi, India. 

6 <sup>b</sup>Rani Lakshmi Bai Central Agricultural University, Jhansi, India. 

7 <sup>c</sup>Jawaharlal Nehru University, New Delhi, India 

8 \*Address for correspondence: [soma.marla@icar.gov.in](mailto:soma.marla@icar.gov.in) 

9 \*Address for correspondence: [Soma.Marla@icar.gov.in](mailto:Soma.Marla@icar.gov.in)

10 **Abstract**

11 Genome assembly of short reads from large plant genomes remains a challenge in computational  
12 biology despite major developments in Next Generation sequencing. Of late multiple draft  
13 assemblies of plant genomes are reported in many organisms. The draft assemblies of *Cajanus*  
14 *cajan* are with different levels of genome completeness; contain large number of repeats, gaps  
15 and segmental duplications. Draft assemblies with portions of genome missing, are shorter than  
16 the referenced original genome. These assemblies come with low map accuracy affecting further  
17 functional annotation and prediction of gene component as desired by crop researchers. Genome  
18 coverage *i.e.* number of sequenced raw reads mapped on to certain locations of the genome is an  
19 important quality indicator of completeness and assembly quality in draft assemblies. Present  
20 work was aimed at improvement of coverage in reported *de novo* sequenced draft genomes  
21 (GCA\_000340665.1 and GCA\_000230855.2) of Pigeonpea, a legume widely cultivated in India.  
22 The two assemblies comprised 72% and 75% of estimated coverage of genome respectively. We  
23 employed assembly reconciliation approach to compare draft assemblies and merged them to  
24 generate a high quality near complete assembly with enhanced contiguity. Finished assembly has  
25 reduced number of gaps than reported in draft assemblies and improved genome coverage of  
26 82.4%. Quality of the finished assembly was evaluated using various quality metrics and for  
27 presence of specific trait related functional genes. Employed pair-end and mate-pair local library  
28 data sets enabled to resolve gaps, repeats and other sequence errors yielding lengthier scaffolds

29 compared to two draft assemblies. We report prediction of putative host resistance genes from  
30 improved sequence against *Fusarium* wilt disease and evaluated them in both wet laboratory and  
31 field phenotypic conditions.

32 **Introduction**

33 Rapid developments in sequencing technologies facilitated generation of several draft assemblies  
34 in plants. These are valuable resources for elucidating genetic information and understanding  
35 biology of the crop. However, each of these draft assemblies have strengths and weaknesses as  
36 were sequenced and assembled based on different algorithms [1,2]. Draft assemblies differ on  
37 the sequencing technology and also the assembly software employed. One assembly may be  
38 conservative in selection of reads resulting in low genome coverage but with many gaps.  
39 Another assembler is vigorous, yielding more contigs but with many errors. Draft genomes are  
40 typically sets of large contingent of assembled contigs and scaffolds that are often fragmented  
41 due to presence several gaps interlaced by repetitive regions. In a misassembly different contigs  
42 are improperly joined. Mis-joins problem arises due to inversions, relocation or a translocation.  
43 Gaps arise also due to incorrect insertion or deletion of a sequenced read in a misassembly. An  
44 inversion or a translocation alters placement of a contig on to scaffold belonging to a different  
45 chromosome. Hence, annotation of unfinished and partially assembled genomes creates  
46 ambiguities while accessing complete genetic information as desired by biologists.

47 In misassemblies some of the reasons for incompleteness include 1. Gaps appearing due to  
48 polymorphisms in complex genomes where reads on either side of a gap representing two  
49 haplotypes belonging to two separate chromosomes, 2. Abundance of repeat elements, multiple  
50 ways to fill the gaps and confusing the assembler thus leaving a gap unfiled, 3. Lack of more  
51 reads to cover that part of the genome, requiring additional library of reads to fill the gaps.  
52 Besides, in draft genome assembly base call errors, variations in read depth coverage also cause  
53 gaps and pose serious computational challenges while connecting nodes in a *De Bruijn* graph [3].

54 Complex eukaryotic genomes are known to contain large volume of near identical copies of  
55 DNA repeats and fragments. Various types of repeats present in genomes of wheat, pigeonpea,  
56 maize or potato include transposable elements, highly conserved gene clusters and segmental  
57 duplications. Presence of identical (or near identical) DNA fragments further complicate

58 computational assembly. During pre-assembly, short reads of equal size tend to be masked  
59 together and complicate construction of *De Bruijn* graphs [4]. Recently introduced third  
60 generation single molecule real time technologies [5] and Oxford Nano pore technologies [6]  
61 generate large sized reads which can readily be inserted for filling gaps caused by repetitive  
62 elements. However, due to low level of sensitivity, high sequencing error rates and expensive  
63 technologies many plant researchers are opting to short read sequencing technologies. Two draft  
64 *de novo* genomes compared in the present study are short read assemblies generated from second  
65 generation sequencing technologies. Abundance of repeats obviates gap closing and responsible  
66 for low levels of genome coverage reported in draft assemblies. Along with reads, modern  
67 sequencing platforms generate paired end reads or mate- pairs. The mate pair libraries are  
68 generated in different sizes (ranging from 3bp to 5bp) and orientations. Hence they could serve  
69 as potential inserts while filling gaps. Mate pair libraries are recommended as a potential  
70 approach to mitigate repeats in computational assembly. In the present work we demonstrated  
71 incorporation of suitable mate pairs to metasassembly for gap closing, which in turn yielded  
72 significant improvement of both genome coverage and quality of the finished Pigeonpea  
73 assembly.

74 Major techniques suggested for gaining contiguity and higher coverage in draft genomes broadly  
75 include, use of long inserts for gap filling [7] assembly reconciliation, hybrid assembly [8],  
76 filtering repeats [9] and iterative mapping using short reads to close the remaining gaps [10].  
77 Use of paired end or mate pairs for filling the gaps is a robust computational approach [11].  
78 Reconciliation approach for closing gaps and correcting misassemblies involves comparing  
79 available data sets from different draft genomes of same or related species, mapping their reads  
80 and finally merge them together to gain improved scaffold lengths with higher contiguity [12].

81 Pigeonpea (*Cajanus cajan* (L) Millsp. cv. Asha) is a major food legume grown in India is diploid  
82 (2n = 22) with genome size of 833.07 Mbp [2]. Widely cultivated and is a major source of  
83 dietary proteins in India with annual production of 2.31mt and productivity of 678 kg/ha [13].  
84 Prevailing low crop productivity may be attributed to absence of high yielding cultivated  
85 varieties possessing resistance to various pests and diseases. In plants, resistance genes (R genes)  
86 play important roles in recognition and protection from invading pests and pathogens. A few  
87 sources of resistance to biotic stresses can be found in available germplasm collections.  
88 Resistance genes are identified and found primarily organized in individual clusters that are

89 strictly linked across the genome [14]. Modern plant breeding techniques such as Marker assisted  
90 and Genomic selections develop superior crop varieties making use of genomic resources and  
91 genetic information emitting from sequenced genome projects. Pigeonpea genome was *de novo*  
92 sequenced independently by [1,2]. These draft assemblies, available in public domain  
93 (GCA\_000340665.1 and GCA\_000230855.2) are valuable resources for breeders. However, both  
94 the assemblies are incomplete with sizable number of fragmented contigs and gaps. Lack of  
95 accurate genetic information is a major limitation towards prediction of gene compliment  
96 associated with desirable traits. Hence our primary objective in the present work is to generate a  
97 more contagious finished assembly with improved genome coverage. We report a finished  
98 assembly based on genome reconciliation approach that first compares the two available draft  
99 assemblies, scoring matching blocks at each location followed by their merger. Metassembler  
100 tool employed in the present study detected gaps and filled them iteratively using right sized  
101 inserts from local pair-end and mate-pair libraries. Completeness and map accuracy of the  
102 reconstructed assembly was verified for the presence of conserved plant resistance genes (R  
103 genes). Here we report prediction of putative R genes, their isolation and PCR screening of a set  
104 of known cultivars against *Fusarium* wilt disease in both laboratory and field conditions.

## 105 **Results**

### 106 **Improvement of the draft genome assemblies employing reconciliation algorithm**

107 Reconciliation assembly approach was employed in the present work to refine the fragmented  
108 draft genome assemblies A1 and A2. For selection of optimum K-mers, hybridSPAdes [15], was  
109 employed and combinations ranging from 21 to 55 were evaluated. We observed with k-mer  
110 sizes 21, 33, 55 and 77 yielded few fragmented sequences, less number contigs with high N50,  
111 mean and median scaffold lengths in superior assemblies. The Illumina HiSeq sequence reads  
112 resulted in 46,979 reads with the N50 length of 24,087. Metassembler was employed for merging  
113 of two assemblies. Metassembler implements reconciliation algorithm to refine and obtain  
114 reconstructed genome. In order to capture the suitable reference assembly set for alignment  
115 during merger process we examined the required order in which assemblies A1 and A2 are to be  
116 chosen as master set (GCA\_000230855.2) and slave sets for aligning with the former,  
117 (GCA\_000340665.1). We observed that choice of A1 as master set with and A2 as slave set  
118 resulted in a highly contiguous superior assembly. Superiority of resulting merged metasassembly

119 was systematically evaluated with the compression-expansion (CE) statistics. Gaps present in the  
120 scaffolds were closed using mate pairs. The remaining gaps were filled by searching unique  
121 contig end sequences against unused reads. We observed that repeat structure analysis and  
122 resulted significant reduction of gaps and contributed to prediction of specific genes. The  
123 improved assembly had 46,979 contigs with total size of 548.2 Mb and covers 82.4% of the  
124 genome with high contiguity (**Table 1**).

125 **Table 1: Genome assembly statistics of draft assemblies A1, A2 and finished A3 assembly.**

Parameter	A1, Assembly GenBank accession: GCA_000340665.1	A2, Assembly GenBank accession: GCA_000230855.2	A3, Finished Assembly GenBank accession: WWND00000000
Number of Contigs	360,028	72,923	46,979
Contig N50	5,341	22,480	24,087
Contig L50	30,054	7,524	6,925
Number of scaffold	NA	36,536	13,101
Scaffold N50	NA	555,764	574,622
Scaffold L50	NA	72	57
Total scaffold Length	NA	592,970,700	548,600,000
Number of Gaps	NA	72,774	36,561
Number of Ns	NA*	34,435,295	34,188,871
Genome Coverage	199x	160x	174x
Percentage mapping	75.6%	72.7%	82.4%
GC content	37.2%	32.8%	45.5%
File size (Mb)	648 Mb	605 Mb	548 Mb

126 **(Data Source: <https://www.ncbi.nlm.nih.gov>)**

127 \*'N's masked.

128 **Read alignment/mapping of pigeonpea**

129 Read mapping increased from 75.6% and 72.7% in two compared misassemblies to 82.4% in  
130 finished metasassembly. More reads were found to be mapped to merged assembly than in A1 and  
131 A2 misassemblies. Mapping depth is a measure of number of reads used for aligning the finished  
132 genome. It also helps to estimate the extent of similarity between final finished assembly and the  
133 compared misassemblies. Among the two draft assemblies, A2 is superior to A1 in depth of read  
134 coverage. Relatively higher read depth in A2 misassembly can be attributed to the high-identity  
135 Illumina reads used both in initial assembly and later polishing steps. Our finished final assembly  
136 in terms of depth of coverage is superior to A2, with more gaps filled. In addition, refined  
137 assembly has more GC-rich regions (**Table 1**) and improved gene component predicted. The  
138 total GC content in A1 and A2 assemblies had GC content i.e. 37.2% and 32.8% respectively and  
139 enhanced to 45.5% in metasassembly reported in the present work. The improvement in GC rich  
140 fraction and of N50 values in both contigs and scaffolds in the finished genome was achieved  
141 largely due to gap filling. High GC content is known to be associated with concentration of  
142 coded genes in certain regions of genome [16]. In the present study high GC content obtained in  
143 refined assembly A3 has contributed to increased number of predicted genes in the finished  
144 genome of pigeonpea.

#### 145 **Metasassembly, annotation and quality assessment**

146 Two draft assemblies were merged and reassembled employing two approaches as described  
147 above. We wanted to ascertain which type of mate-pair libraries effectively resolve repeat  
148 problem. In assembly employing Meta assembler tool we used in one experiment only 648 Mb  
149 library and in the second 605 Mb and 548 Mb libraries taken together. Initially we used all the  
150 single paired read data sets available (minus two mate-pair data sets) of A1 along with all data  
151 sets from A2. In the second treatment included the two mate-pair data sets from A1 along with  
152 all full data available from A2. At the end of analysis, all the output values and statistical  
153 metrics were collected for comparative performance analysis. We observed that all the available  
154 Pigeonpea mate pair libraries taken together resulted improvement in genome coverage. It is  
155 presumed that incorporation of variable size mate pair inserts helped in gap closing during  
156 assembly.

157 In our final assembly the contig N50 is increased by 24,087, and scaffold N50 increased by  
158 574,622. Total number of gaps decreased across the genome by 50.23%, comprising from A2

159 (Table 1). It is observed that the order in which the input draft assemblies are inputted to  
160 Metassembler drastically influences the alignment quality and the resulting read coverage. In  
161 primary assembly we treated Assembly A1 as master and aligned it with Assembly A2. In other  
162 variant we used Assembly A2 as master and aligned against its counterpart Assembly A1. Output  
163 of resulting primary assembly yielded us a scaffold length 548,600,000. We initially used  
164 unpaired reads for assembly adopting overlapping read approach. As no significant improvement  
165 was observed in both read mapping depth and eventual coverage we resorted to available mate  
166 paired libraries to close gaps. We used mate pairs during different alignment steps during  
167 metassembly and succeeded in resolving repeat problems.

168 **Closure of repeat-derived gaps**

169 For each round of alignments undertaken between A1 and A2 misassemblies, metassembler  
170 builds a graph, with vertices being the above alignments and edges joining two alignments. If  
171 both have the same direction, they are readily rearranged in to a single block thus providing  
172 contiguity. In case, where the examined genomic segments from two misassemblies do not share  
173 same direction, indicates the existing distance from each other and need to fill the prevailing  
174 gaps. In such cases, variable sized local pair-end and mate-pair libraries could offer right inserts  
175 to fill these gaps. While building the graph, metassembler searches the mate-pair library for right  
176 sized inserts to complete the shortest path between any of these contigs, to fill a gap.

177 We evaluated the closure performances of the Gapcloser and Gapfiller tools on the repeat  
178 derived gaps using the raw mate pair reads. We first tested the performance of each tool using  
179 the raw mate pair reads. Both the above tools used first raw pair end and Mate pair libraries. We  
180 monitored the gap closure efficiency by evaluating number of gaps closed. In improved  
181 pigeonpea assembly, we estimated 37,145 repeat-derived gaps of which 584 gaps and 322,780  
182 nucleotides out of total 34511651 were closed. The gap sizes ranged from 200 bp to 15,510 bp.  
183 Gap closer was more efficient by filling most of the gaps with 82.4% and with low error rates.  
184 We achieved improved contiguity by using long mate-pairs to fill gaps in assembly and thereby  
185 achieving higher coverage in the finished assembly. Two-draft genome assemblies A1  
186 (GCA\_000340665.1) and A2 (GCA\_000230855.2) are used in the present study to improve  
187 scaffold contiguity and achieve read coverage completeness of Pigeonpea genome. Draft  
188 assembly A1 had 360,028 contigs with N50 and L50 of 5,341 and 30,054 respectively. Reported

189 genome coverage was 199x with a similarity of 75.6 %. Draft Assembly A2 had 72,923 contigs  
190 with N50 and L50 of 22,480 and 7,254 respectively. A2 had 592,970,700 scaffolds with reported  
191 genome coverage of 160x with a similarity of 72.7 %. We present an improved reference  
192 assembly of pigeonpea genome.

193 **Completeness of the merged assembly**

194 The BUSCO [17] evaluation of completeness of the conserved proteins in all three assemblies of  
195 the pigeonpea genome sequence predicted that it was 94.02% complete in A3 assembly, where a  
196 proportion of total 1,440 BUSCO groups were searched, the genome assembly found to contain  
197 1,321 complete single-copy (S) BUSCOs, 33 complete duplicated (D) BUSCOs, 57 fragmented  
198 (F) BUSCOs, and 29 missing (M) BUSCOs. Whereas comparatively in A1 and A2 assemblies it  
199 was found 85.27% (S:76.87%, D:8.40%, F:5.62%, M:9.09%) and 87.9% (S:80.9%, D:7%, F:5%,  
200 M:7.1%) complete respectively (**Supplementary Table 1**). The gene completeness as measured  
201 by BUSCO is increased in improved assembly, while the numbers of fragmented and missing  
202 BUSCO genes are reduced. This genome comparison can be used to help such draft assemblies  
203 towards becoming finished genome.

204 **Functional annotation of predicted gene content**

205 FGENESH module of the Molquest v4.5 software package (<http://www.softberry.com>) and  
206 Augustus was employed and 51,737 genes are predicted in the finished metasassembly.  
207 Predicted numbers of genes are less compared to A1 and higher than A2.

208 **Table 2: Results of Gene finding.**

Parameter	A1 Assembly	A2 Assembly	A3 Assembly
No. of genes predicted	56,888	48,680	51,737

209 In the total gene component predicted we found 1,303 disease resistance related genes in  
210 pigeonpea. Finished assembly, A3 yielded 51,737 total genes which are less to A1 but more than  
211 reported in A2 assembly, improvement in read mapping depth results in reduction of number of  
212 earlier reported incomplete genes and yields a complete gene set. Of the predicted gene set 54-

213 resistance single copy gene analogues containing known conserved domain NBS LRR were  
214 selected and *in silico* mapped on to the corresponding chromosomes (**Supplementary Table 2**).

215 **Identification of repetitive sequences and transposable elements in the improved assembly**

216 Repeat elements are some extra copies of DNA sequences generated and planted at various  
217 locations in the genome to meet certain challenges and improve fitness during the course of  
218 evolution. Repetitive elements in Pigeonpea occupy nearly half of the genome of *Cajanus cajan*  
219 [18]. Repeats pose many computational challenges in read alignment and assembly [19] such as  
220 creation of gaps, overlaps and leading to many mapping inaccuracies in misassemblies [20]. One  
221 can always filter and exclude the reads but it is essential to map them on to chromosomal  
222 locations where gaps exist. Mate-pair libraries were used for resolving repeat problems and  
223 obtaining contiguous scaffolds in both prokaryotic [21] and eukaryotic organisms [22].  
224 Metassembler searches for contigs that can be placed in the gap using mate pairs, and then again  
225 looks to see if there is a recorded shortest path exists between any of these contigs. In assembly,  
226 over lapping reads are used as edges to connect reads belonging to same region of genome.  
227 However in complex genomes like Pigeonpea abundance of repeats cause coverage gaps and  
228 read errors thus leaving numerous gaps to fill between contigs while scaffolding. Filling of gaps  
229 requires adoption of robust computational approaches to effectively address repeats problem.  
230 Sequenced pair-end and mate-pair reads can potentially bridge over gaps efficiently to order and  
231 orientate contigs by estimating the gap lengths to the edges while filling the scaffolding graph  
232 [23].

233 High level of assembly was achieved using mate-pair reads in wheat, a genome ridden with large  
234 content of repeats [24]. We analyzed the repeat content in comparison to A1 and A2 assemblies  
235 and classified them in to various classes (**Table 3**). In course of iterative use of reads during  
236 assembly we observed transposon derived repeats collapse against identical reads resulting in  
237 closure of a significant portion of gaps. Similar observations were reported on gap filling using  
238 retro transposon related repeats in human genome assembly [25].

239 **Table 3: Repetitive sequences of draft assemblies A1, A2 and finished A3 assembly.**

Transposable Elements	A1 Assembly	A2 Assembly	A3 Assembly
Retrotransposons	77,096,057	116,194,477	89,089,240

Gypsy	52,354,920	71,402,096	59,247,991
Copia	19,937,308	37,676,825	24,339,237
Line	5,261,337	6,717,918	5,914,324
Unclassified elements	216,262,607	169,378,278	158,228,382
DNA transposons	9,772,250	27,455,193	19,826,943
Total transposable elements	303,130,914	313,027,948	267,144,565

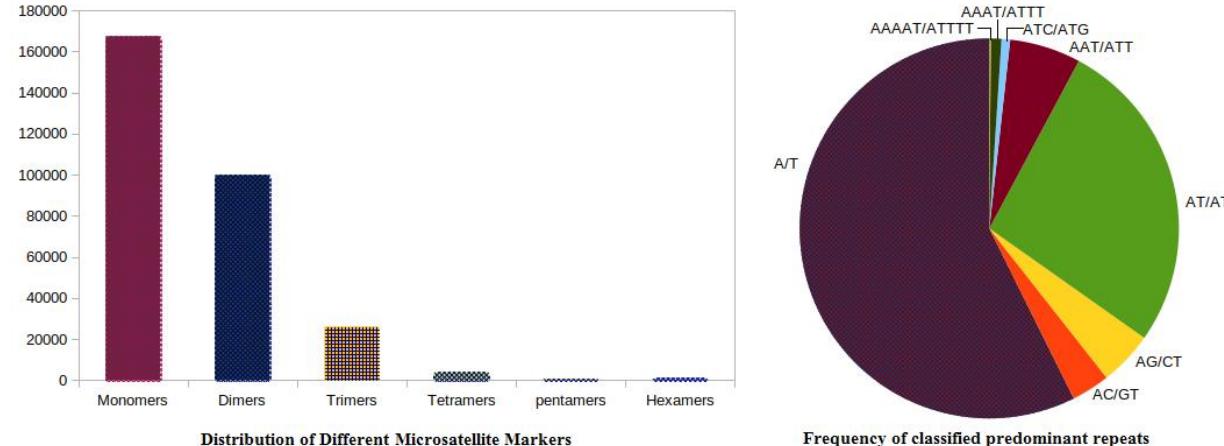
240 **Identification of microsatellites**

241 Improved Pigeonpea assembly was mined for single sequence repeats and out of 2,98,732,  
242 2,97,294 were simple and the remaining 1,438 of complex types. Mononucleotide repeats were  
243 the abundant with 56.05% of total SSRs, mined. Dinucleotides occupying 33.45% dinucleotides  
244 (99949), 8.72% (26069), trinucleotides and 1.27% tetranucleotides (3811) repeats. The  
245 remaining SSRs were a complex type, with 0.25% of hexa nucleotides and 0.22% of penta.  
246 Among 167,465 mononucleotide repeats, the mononucleotide motifs were in majority with A/T  
247 repeats of 98.25% and of with the rest of. 1.74% occupied by C/G types. Among 99,949  
248 dinucleotides microsatellites, AT/AT type (77.34%) of microsatellites were most common type  
249 in the genome followed by AG/CT type (13.21%), and AC/GT type (9.40%). The CG/CG type  
250 dinucleotides microsatellites were present at a very low proportion (0.03%). In trinucleotide  
251 SSRs repeats (26,069), around 66.71%, 12.31%, 8.07%, 5.98% of SSRs were of AAT/AAT,  
252 AAG/CTT, ATC/ATG and AAC/GTT types, and were most abundant respectively. Among the  
253 other types of repeats, the ACG/CGT type was lowest (0.36%) in the genome of Pigeonpea. The  
254 highest distribution (68.06%) of tetra nucleotides microsatellites was present in the genome of  
255 Pigeonpea. Maximum numbers of predominant SSRs repeats were of A/T type followed by  
256 AT/AT, AG/CT, AAG/CTT, AAT/ATT and AAAT/ATT (Supplementary Table 3). The  
257 overall analysis showed that the relative abundance of tetra, penta and hexa SSRs types were low  
258 as compared to mono, di and tri SSR types in Pigeonpea genome sequences (Figure 1).

259 **Table 4: Results of microsatellite search in the improved pigeonpea assembly A3.**

Total number of sequences examined	13,102
Total size of examined sequences (bp)	584435790
Total number of identified SSRs	298732

Number of SSR containing sequences	6494
Number of sequences containing more than 1 SSR	4603
Number of SSRs present in compound formation	41002



260

Distribution of Different Microsatellite Markers

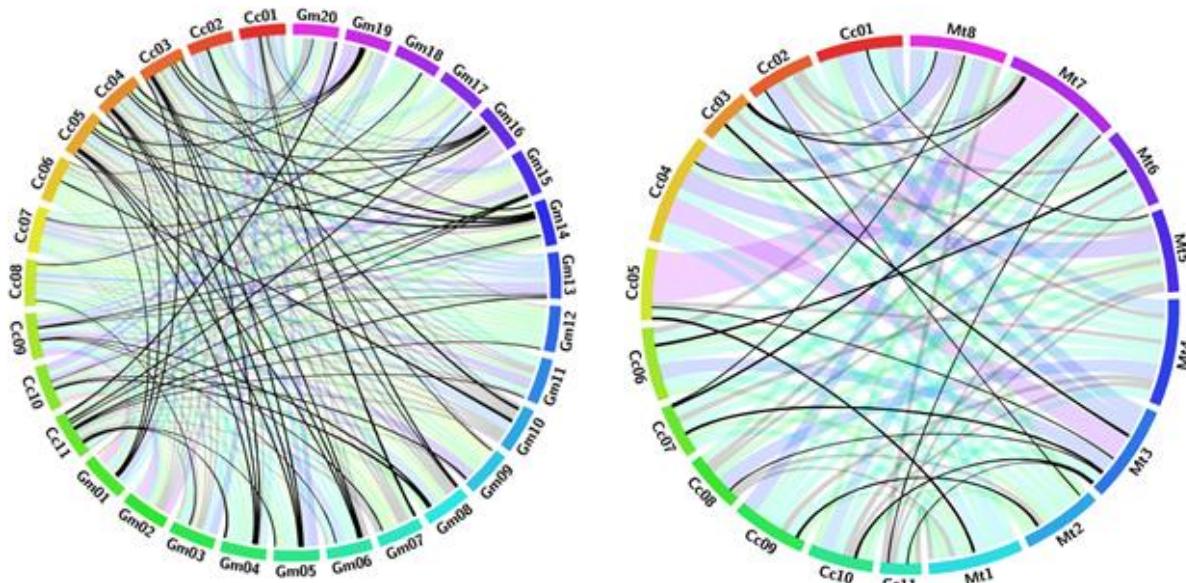
261 **Figure 1: SSR distribution frequency** (A) Distribution of different repeats type classes (B)  
262 Frequency of classified predominant repeats.

263 **Characterization and syntenic analysis of pigeonpea NBS-LRR like resistance gene analogs**

264 We verified the presence of already known conserved disease resistance genes in the refined  
265 metasassembly. The nucleotide-binding site (NBS)-leucine rich repeat (LRR) protein sequences for  
266 other genomes were downloaded from Phytozome [26]. Comparison of predicted coding  
267 sequences against bean (*Phaseolus vulgaris*) cluster resulted in more than 100 resistance gene  
268 analogues (RGA). The predicted gene annotation revealed presence of known disease resistance  
269 domains such as ARC-NBS-LRR, Transmembrane and Kinases. Nucleotide-binding site (NBS)  
270 disease resistance genes play an important role in defending plants from a variety of pathogens  
271 and insect pests. Many R-genes have been identified in various plant species. However, little is  
272 known about presence of NBS-encoding genes in pigeonpea genome. In this study, using  
273 computational analysis of the refined genome, we identified 54 NBS-encoding single copy genes  
274 and characterized them on the basis of structural diversity and conserved protein motifs. The  
275 RGAs had high amino acid identity (77-98%) with putative disease resistance proteins in *Glycine*  
276 *max* several sequences with high similarity to NBS-LRR resistance (R) proteins were identified.  
277 We mined 1,301 resistance gene analogues sharing up to 78% of homology with Soybean,  
278 Chickpea, barrel clover, field bean and other species (**Supplementary Table 4**). Of them 251

279 NBS-LRR domain containing resistance gene analogues to pigeonpea were found  
280 (**Supplementary Table 5**).

281 Syntenic relationship with selected legume genomes *Glycin max* and *Medicago truncatula*  
282 revealed extensive conservation among pigeonpea and other legume plants, with 89–91 per cent  
283 of the pigeonpea assembly showing signs of RGA conservation. 41 NBS-LRR orthologs *Glycin*  
284 *max*, 73 NBS-LRR orthologs *Medicago truncatula*, for some 57 per cent NBS-LRR pigeonpea  
285 genes, were identified for the closely related organisms. *Glycin max* was found to have the  
286 largest number of extended conserved synteny blocks indicating its recent ancestry followed by  
287 *Medicago truncatula*. The genome assembly of pigeonpea comprises 251 homologs of the  
288 disease resistance gene, of which 229 are anchored in pseudomolecules. The number of 41  
289 pigeonpea genes had significant sequence homology with *Glycin* genes and 73 with *Medicago*  
290 genes. Homologous blocks containing more than 4 R genes in *C. cajan* with *G. max* and *M. truncatulum*  
291 are noted. Of these, there are 23 genes between the pigeonpea and *Glycin* genome assemblies with 57  
292 collinear blocks (**Figure 2**). Overall all pigeonpea RGAs displayed extensive collinearity with different  
293 chromosomes of *Glycin* and *Medicago*. Homologous blocks connecting chr4 in *C. cajan* with chr4 of *G.*  
294 *max*; chr11 of *C. cajan* with chr20 and chr17 of *G. max*; Chr3 of *C. cajan* with chr19 in *G. max*. Similarly  
295 comparative analysis of draft assembly A2 [2] reported homologous blocks connecting chr3 in *C. cajan*  
296 with chr19 of *G. max*.



297  
298 **Figure 2:** Circos diagram presenting synteny relationship between NBS-LRR proteins from

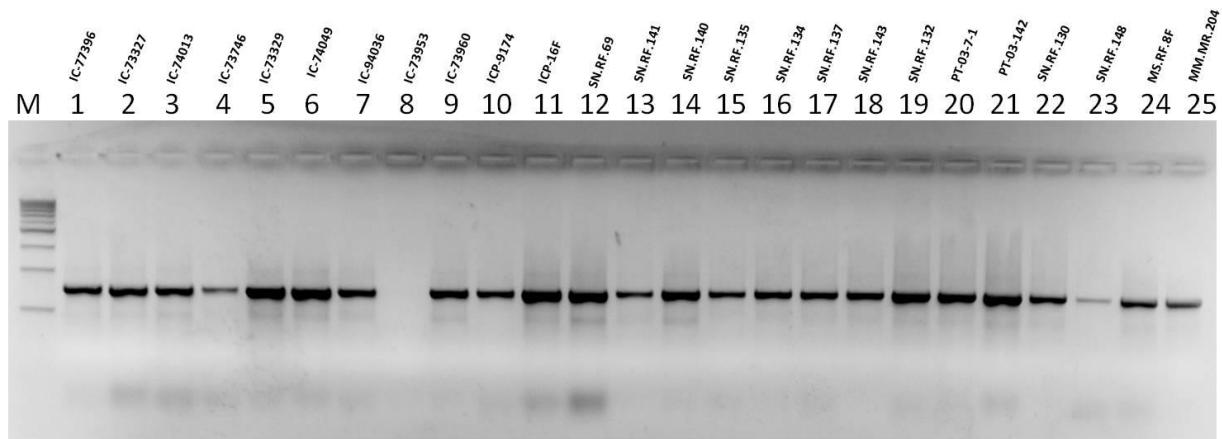
299 pigeonpea (Cc), *Glycin max* (Gm) and *Medicago truncatula* (Mt) pseudomolecules.  
300 Pseudomolecules of the two target species were labeled as Gm01-20 and Mt1–8. Pigeonpea  
301 pseudomolecules are labeled in different colours and labeled as Cc01-11. Colinear blocks are  
302 coloured according to the colour of the corresponding Pigeonpea pseudomolecule. Each ribbon  
303 radiating black from a pigeonpea pseudomolecule represents a NBS-LRR similarity block  
304 between pigeonpea and other legumes.

305 **Cloning, isolation and PCR amplification of identified putative R gene analogs (RGAs)**

306 For designing primer sets for PCR amplification of predicted resistance gene (R) orthologs  
307 BLASTN was employed in comparison with Soybean genome. Primer sets for PCR  
308 amplification were designed using EPrimer tool [27]. PCR amplicons were eluted and sequenced  
309 by Sanger sequencing method. Isolated Pigeonpea resistance gene analogues were deposited to  
310 NCBI (**Supplementary Table 6**). List of primer sequences used in PCR amplification are given in  
311 (**Supplementary Table 7**).

312 Genomic DNA from 15 day old seedlings of 34 Pigeonpea cultivars was extracted employing  
313 CTAB method [28]. Purity and concentration of DNA was estimated with Nanodrops ND-1000.  
314 Nine primers were selected for polymorphism study (**Supplementary Table 7**). Polymerase  
315 chain reaction (PCR) was performed in a total volume of 20  $\mu$ l containing 60 ng of template  
316 DNA, 200  $\mu$ M of dNTPs, 2.5 mM MgCl<sub>2</sub>, 1x PCR buffer, 0.4  $\mu$ M of each primer, 0.75 U Taq  
317 DNA polymerase and water to make the final volume up to 20  $\mu$ l.

318 Amplification were carried out using thermocycler Bioer Gene Pro and PCR conditions was set  
319 as initial denaturation at 94°C for 5 minutes, 30 cycles of denaturation at 94°C for 30 seconds,  
320 primer annealing at 50°C for 30 seconds, primer extension at 72°C for 2 minutes and final  
321 extension step at 72°C for 7 minutes. The amplified products were visualized by ethidium  
322 bromide stained 1.5 % agarose gels in SYNGENE G-Box gel documentation unit (**Figure 3**).



1-11: F.wilt Resistant; 12-18: F.wilt tolerant; 19-23: F.wilt susceptible; 24-25: yellow mosaic resistant

M: 500 bp ladder, Primer ID: 06, 1-25 Pigeon Pea genotypes, Product size 0.7 kb

323

324 **Figure 3: PCR amplification of *Fusarium* wilts resistant RGA among Pigeonpea genotypes.**

## 325 Discussion

326 In the present work we chose two available incomplete draft assemblies and employed  
327 reconciliation algorithm to correct errors. Two compared draft assemblies A1 and A2 had low  
328 genome coverage and several repeats and gaps resulting disjoin between contigs to yield lengthy  
329 scaffolds with correct contiguity. Assembly tool, Metassembler employed in the present work is  
330 based on genome reconciliation algorithm. The computational framework includes merger of two  
331 draft assemblies, A1 and A2, align them by selecting matches and mis-matches present in both,  
332 resolving gaps and other sequence errors to obtain a consensus and complete assembly.

333 To begin with we wanted to select the order in which the input draft assemblies are to be merged  
334 to gain subsequent superior alignment and read mapping. After several permutations, we  
335 observed that treating assembly A1 as master and aligning it with assembly A2 yielded better  
336 read mapping and lengthier scaffolds of 592,970,700 mb. Merging the two draft assemblies, in  
337 course of alignment, Meta assembler yielded matched and mismatched portions in the merged  
338 assembly by identifying homologous genomic regions with shared set of reads. Mis matches  
339 include gaps that are to be filled with right sized read sequences.

340 Metassembler initially utilized all available raw reads from both draft assemblies using  
341 conventional read overlapping technique to fill the existing gaps and join the contigs. However,  
342 no notable success was observed in gap filling and repeat resolution. Alternatively, we employed  
343 local pigeonpea pair-end and mate pair libraries to fill the gaps. Metassembler generated

344 statistics, compared the distances between the mapped mates and the required sizes of insert  
345 reads to fill a gap. For example, gaps measuring < 500 mb were filled by pair-end reads while  
346 mate-pair reads were maximum utilized for filling larger gaps measuring 3 to 5 KB. Similar  
347 reports using large sized mate-pairs for filling bigger gaps in assembly of large genomes were  
348 reported [29]. In the present study employed pair-end and mate-pair reads contributed  
349 significantly to fill the gaps and thereby in joining the contigs in to full length scaffolds. Further,  
350 iterative use of pair-end and mate-pair libraries during successive alignments resulted in  
351 identification of maximal portions shared by same library of reads. This in turn has contributed  
352 to dramatic improvement of genome coverage in the resultant assembly A3. Resulting A3  
353 assembly quality was judged using metrics- contig number, scaffold lengths, N50 and L50,  
354 genome coverage of 160x with a similarity of 72.7 %. Genome similarity score can also be  
355 useful in estimating extent of redundancy present in both genomes [30-31].  
356 Draft assembly A1 had 360,028 contigs with N50 and L50 of 5,341 and 30,054 respectively.  
357 Reported genome coverage was 199x with a similarity of 75.6 %. Draft Assembly A2 had 72,923  
358 contigs with N50 and L50 of 22,480 and 7,254 respectively. A2 had 592,970,700 scaffolds with  
359 reported genome coverage of 160x with a similarity of 72.7 %.  
360 FGENESH predicted 51,737 genes using the finished metasassembly, A3. Predicted number of  
361 genes are less in our finished assembly, A3 are less compared to A1 but higher than A2 (**Table**  
362 **2**). Annotation of improved assembly yielded 51,737 genes predicted. Wet lab PCR  
363 amplification is the Gold standard for verification of predicted gene presence and their  
364 functionality. For PCR based gene amplification 23 primer sets were designed to screen 34  
365 pigeon cultivars. Out of the 34 genotypes screened 14 were found to be *Fusarium* wilt resistant  
366 (**Supplementary Table 8**), 7, *F.* wilt tolerant, 5 *F.* wilt susceptible, 5 yellow mosaic susceptible  
367 genotypes (**Figure 3**). Data on yellow mosaic disease reaction is not presented here. PCR  
368 amplified genes were isolated, cloned and submitted to NCBI (**Supplementary Table 6**).  
369 Genotype, environment interaction in the field determines the phenotypic performance of  
370 isolated plant genes [32]. Phenotypic evaluation of predicted resistance genes in field trials is  
371 also required for transfer of the obtained results to pigeonpea downstream breeding programs for  
372 development of disease resistant cultivars. Field experiments were conducted to assess the  
373 disease reaction of predicted R genes to *Fusarium* wilt taking cv. Asha ( object of present  
374 study) as control with 34 Pigeonpea cultivars. The replicated field experiments were conducted

375 at Ranchi (Jharkhand state) and Rahuri (Maharastra), India during 2011, 2012 and 2012, 2013  
376 rainy seasons. Of the 26 screened cultivars against check cv. Asha, 14 resistant and 6 tolerant at  
377 Ranchi farm and at Rahouhuri farm 8 resistant, one tolerant and 6 susceptible disease reaction  
378 was observed to the *F.* wilt disease of Pigeonpea. Observed variation in disease incidence reflect  
379 the natural agro climatic conditions prevailing at the individual trail site.

### 380 **Conclusion**

381 In the present work genome reconciliation algorithm was adopted to reconstruct draft assemblies  
382 to produce an accurate and near complete genome assembly of pigeonpea. We demonstrated  
383 successful implementation of our reassembly frame work by merging two chosen draft  
384 assemblies employing pair-end and mate-pair libraries to correct gaps and other sequencing  
385 errors. Resulting reconstructed metasassembly was superior to compared two draft assemblies in  
386 terms of measured assembly quality statistics *viz.* N50 and scaffold lengths. Quality of finished  
387 assembly was assessed for presence known conserved resistance gene loci (imparting resistance  
388 to *Fusarium* wilt disease in Pigeonpea). Annotation of improved assembly yielded prediction of  
389 1303 resistance genes (including six extra genes gained from metasassembly). PCR screens and  
390 field experiments validated the resistance reaction of isolated genes against *Fusarium* wilt thus  
391 making the results available to Pigeonpea breeders.

### 392 **Methods**

393 We developed a workflow model (**Figure 4**) based on reconciliation algorithm, that includes 1.  
394 Merging two mis-assemblies, 2. Finding matches and mismatches and other sequencing errors, 3.  
395 Closing gaps using pair-ends, mate -pair libraries, 4. Assessment of finished assembly quality,  
396 prediction of disease resistance gene families, their isolation and characterization.

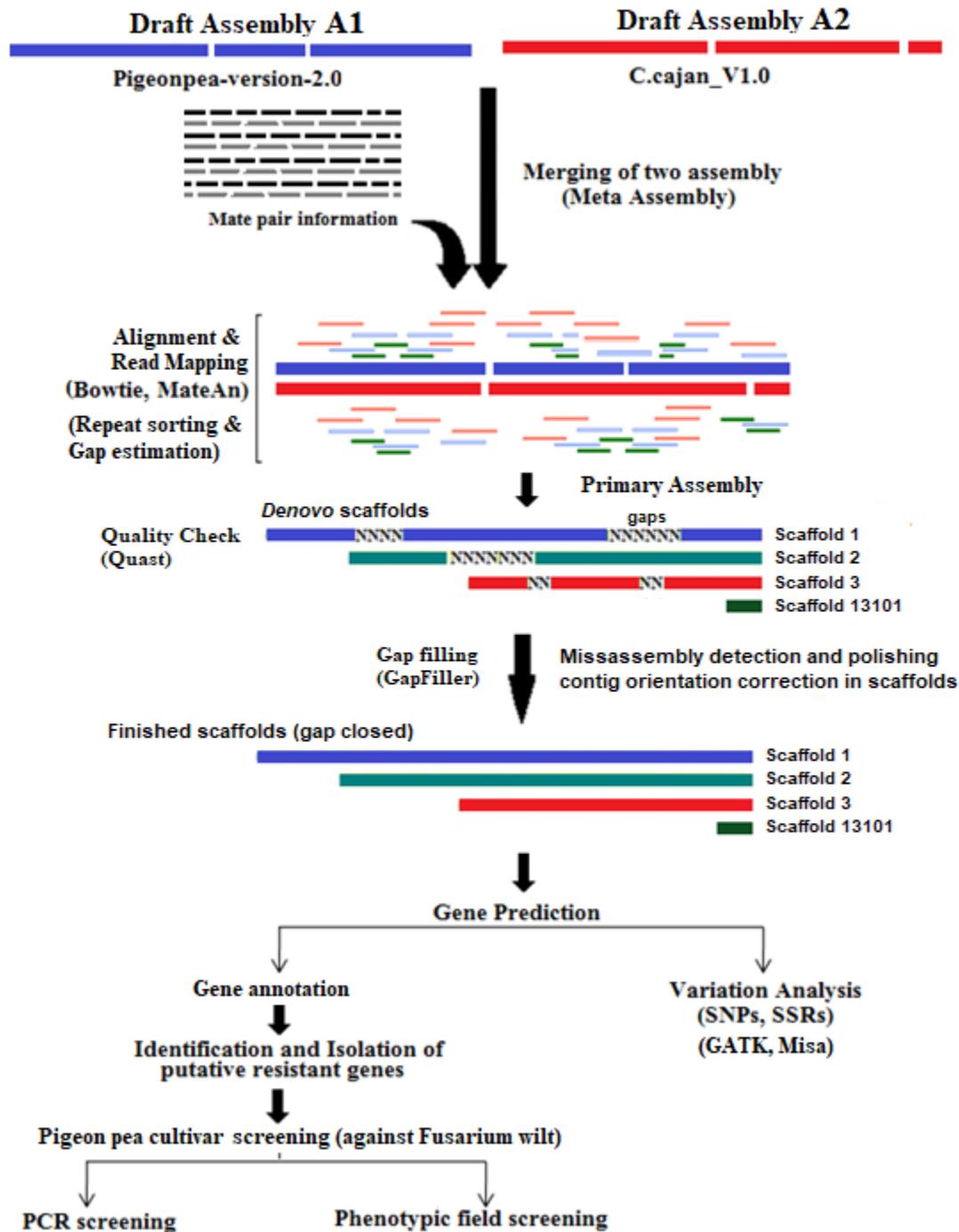
### 397 **Retrieval of pigeonpea genome datasets**

398 Complete data sets belonging to two whole genome sequences of Pigeonpea and associated 23  
399 SRA reads were downloaded from the National Center for Biotechnology Information (NCBI)  
400 (<https://www.ncbi.nlm.nih.gov>) to the local storage- GCA\_000340665.1 (SRA accessions  
401 SRR5922904-SRR5922907) and GCA\_000230855.2 (SRA accessions SRR6189003-  
402 SRR6189021) for the *cv* Asha.

### 403 **Genome reconstruction and quality assessment**

404 Illumina pair-end and mate-pair library sequence reads of Pigeonpea, *cv* Asha were quality  
405 checked using FASTQC v0.11.8 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>).  
406 Contaminated reads were removed to get error corrected reads. Reads with sequence quality of  
407 Phred scores of less than Q30 (base calling accuracy with less than 99.99%) were removed using  
408 PRINSEQ v0.20.4 (<https://sourceforge.net/projects/prinseq>) and reads were repaired using  
409 BBmap v37.66 (<https://sourceforge.net/projects/bbmap>).  
410 Reported Pigeonpea draft assemblies A1 [1] and A2 [2] were both sequenced employing  
411 Illumina technology and assembled with SoapDenovo v2.3.1 assembler. In present work, data  
412 sets A1 (GCA\_000340665.1 consisting 4 SRA read sets) and A2 (GCA\_000230855.2 of 19 SRA  
413 read sets) were analyzed employing reconciliation algorithm [11]. The work flow includes the  
414 steps: 1) Merger of two mis-assemblies, 2) Finding matches, mismatches and other structural  
415 errors, 3) Closing gaps using pair-end, mate pair libraries, 4) Assessment of finished assembly  
416 quality 5) Prediction of disease resistance gene families, their isolation and characterization. A1  
417 consisted of 360,028 initial contigs (N50 5341, 648 Mb) with 30% of gaps within contigs. A2  
418 contained 72,923 scaffolds (N50 22480, 605 Mb) with 20% of intra scaffold gaps. We used all  
419 the read datasets available belonging to A1 and A2 with NCBI. All the computations including  
420 read pre-processing, quality control, comparison of two draft assemblies, their alignment, gap  
421 filling, assembly merger, map accuracy, quality assessment, putative gene prediction were  
422 performed on HPC server employing Meta Assembler [33].  
423 GapFiller [34] was employed to find the existing gaps (A1 30%; A2 20%). Initially short reads  
424 were used for filling gaps, resulting A1 genome size of 648 Mb and 605 MB of A2 draft  
425 assembly.  
426 We initially employed overlap approach with available read sequences followed by used pair-  
427 end as well as mate-pair library data sets for resolving repeat redundancy, gap filling and other  
428 structural errors. Firstly, we used entire single paired read data sets available (minus two mate-  
429 pair data sets) of A1 along with all data sets from A2. Alternatively, second treatment included  
430 two mate-pair data sets from A1 and all full data available from A2. At the end of analysis, all  
431 the output values and statistical metric data were collected for comparative performance analysis.  
432 Draft assembly A1 was sequenced in 2011 and had genome coverage of 199% [1]. However,  
433 using the same raw read data, the authors had again reassembled employing A3 assembler and  
434 reported gain of coverage, i.e. an increase of ~15% (from 60.0% to 75.6%,) and resubmitted to

435 NCBI . In our present work we used this recent assembly set, A1 along with A2 assembly data  
436 [2] for reassembly and improvement (Figure 4).



437

438 **Figure 4:** Experimental Frame work depicting reconstruction steps of Pigeonpea genome.

439 We observed that in our reassembly pair-end insert read sizes below 500 bp in our library were  
440 maximum utilized for filling smaller gaps. Mate-pair sizes up to 5.0 kilo base pairs are available  
441 in our library. In our metassembly these mate pairs were employed effectively used for closing  
442 medium and long distanced gaps (even up to 20-25 kb). Similar results on use of large sized  
443 mate-pairs for filling bigger gaps was reported in assembly of large genomes [29].

444 **Merging misassemblies and gap closure**

445 Draft assembly sequences A1 and A2 were merged in to a single sequence. Alignment and  
446 merger of A1 and A2 assemblies resulted in a total scaffold length of 548 Mb. Resulting merged  
447 assembly is compared to A1 and A2 draft assemblies (75.6% and 72.7% respectively) has an  
448 improved genome coverage of 82.4 %. Yet the Merged sequence contained 10% of gaps.

449 To improve further contiguity and accuracy of merged sequence existing intra scaffold gaps were  
450 filled. Repeat content and existing gaps were estimated employing Gapcloser and Gapfiller tools  
451 [34]. In the second round of gap filling various computational approaches such as paired end,  
452 mate pair libraries and remaining unused short reads were used. Gap content, estimation of  
453 repeats (**Table1**). Iterative use of left over short reads (300bp) contributed to filling nearly 20%  
454 of gaps. After polishing and another round of reassembly yielded a scaffold length 13,348  
455 (scaffolds of N50 574,622) with a coverage of 174x %.

456 **Finished genome assembly and quality assessment**

457 Increased N50, maximum scaffold length and minimum number of contigs, increased N50 values  
458 together with longer scaffolds contribute to improved genome coverages. In mis-assemblies the  
459 number of gaps and 'N's caused due to repeats were measured. In course of metassembly we  
460 strived to minimize gaps and other sequencing errors. We employed Quast v4.5 [35] to gather  
461 extensive assembly statistics. BUSCO v3.2 [17] was employed for assessing the genome  
462 completeness, annotation and sets of predicted genes. Mapping accuracy and identification of  
463 resistant gene analogue loci were assessed. In addition 75% of unigenes were aligned to the  
464 reassembled genome.

465 **Gene prediction and function annotation**

466 Metasassembly was first repeat-masked using RepeatModler and Repeat Masker tools [36],  
467 followed by *ab initio* gene prediction using the FGENESH module of the Molquest v4.5  
468 software package (<http://www.softberry.com>). The predicted genes were annotated using  
469 BLASTX (E<10<sup>6</sup>) search against the NCBI non-redundant (nr) protein database using Blast2GO  
470 software [37]. Synteny blocks between the genomes of pigeonpea and other legumes were  
471 computed by blastp combined with the Circos [38] to understand homology to the NBS-LRR  
472 gene from *Glycin max* (Gm) and *Medicago truncatula* (Mt) pseudomolecules.

473

#### 474 **Identification of genome wide SSR**

475 Refined genome sequence of Pigeonpea was analyzed identify to various Single Sequence  
476 Repeat markers (SSRs) types using Microsatellite Identification tool (MISA)  
477 (<http://pgrc.ipkgatersleben.de/misa/>). Minimum length for SSR motifs per unit size was set to 10  
478 for mono, 6 for di and 5 for a tri, tetra, penta, hexa motifs. We calculated the total lengths of all  
479 mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide repeats in terms of base pairs of SSR per  
480 mega base pair (Mb) of DNA.

#### 481 **Gene validation**

482 Genome similarity score recorded set of sequenced reads originating from one draft genome  
483 correctly mapped on to a second genome. To check the accuracy in finished Pigeonpea genome  
484 we wanted to verify the location of certain genomic regions or loci present in the inputted two  
485 assemblies. A set of genes imparting resistance against various pests and diseases are located in  
486 B4 cluster on chromosomes in two examined draft assemblies of pigeonpea, (*Cajanus cajan*)  
487 Asha. As a test case location of B4 gene cluster syntenic regions was verified in the present study  
488 to estimate the accuracy of read mapping achieved in the finished assembly.

#### 489 **Computational resources**

490 We run all reassembly and merging using HPC Cluster having CentOS-Linux version 7.2.93  
491 GHz 2x Intel Xeon 8 core processors and 2 TB of RAM. Majority of the running time is spent on  
492 assembly process and about 1/4 on graph construction and analysis. However, Reconciliator uses  
493 more than 1.5 TB of RAM to merge the Asha isolates, Pigeonpea assemblies.

#### 494 **Data availability**

495 The improved draft genome assembly of Pigeonpea is available at NCBI/ENA/GenBank, under  
496 the Accession Number WWND00000000.

497 **Acknowledgments**

498 Authors thank ICAR-National Bureau of Plant Genetic Resources, New Delhi for providing  
499 research facilities, and Centre for Agricultural Bioinformatics (CABIN, ICAR-IASRI), New  
500 Delhi India for providing high performance computing (HPC) facility.

501 **Author information**

502 Soma Marla, Pallavi Mishra, Ranjeet Maurya contributed equally.

503 **Affiliations**

504 1) ICAR-National Bureau of Plant Genetic Resources, New Delhi, India.

505 Soma Marla, Pallavi Mishra, Ranjeet Maurya, Mohar Singh, D. P. Wankhede, Anil K. Gupta,  
506 S. K. Singh, Rajesh Kumar

507 2) Rani Lakshmi Bai Central Agricultural University, Jhansi, India.

508 Anil K. Gupta

509 3) Jawaharlal Nehru University, New Delhi, India.

510 N. S. Rao

511 **Contributions**

512 SM conceptualized and supervised all the experiments, interpreted the results and wrote the  
513 manuscript text. PM and RM performed all the high-throughput bioinformatics analysis,  
514 software and formulated the manuscript. MS conducted field phenotypic experiments, SNR and  
515 RK updated the final manuscript for publication. DPW, AKG and SKS performed all the wet lab  
516 experiments.

517 **Corresponding author**

518 Correspondence to Soma Marla; **Soma.Marla@icar.gov.in**

519 **Ethics declarations**

520 **Competing Interests**

521 The authors declare that they have no conflict of interest. We have also followed the accepted  
522 principles of ethical and professional conduct and no animals or humans are involved in this  
523 research.

524 **References**

- 525 1. Nagendra K. Singh, Deepak K. Guptam Pawan K Jayaswal, Ajay K Mahato et al., *et. al.* The  
526 first draft of the Pigeonpea genome sequence. *Journal of plant biochemistry and*  
527 *biotechnology*. **21**, 98-112 (2012).
- 528 2. Rajeev K Varshney, Wenbin Chen, Yupeng Li, Arvind K Bharti, Rachit K Saxena *et. al.*  
529 Draft genome sequence of Pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-  
530 poor farmers. *Nature biotechnology*. **30**, 83 (2012).
- 531 3. Dieval Guizelini, Roberto T. Raitt, Leonardo M. Cruz, Emanuel M. Souza, Maria B. R.  
532 Steffens & Fabio O. Pedrosa. GFinisher: a new strategy to refine and finish bacterial  
533 genome assemblies. *Scientific reports*. **6**, 34963 (2016).
- 534 4. Compeau PE, Pevzner PA, Tesler G. How to apply *de Bruijn* graphs to genome assembly.  
535 *Nat Biotechnol*. **29**, 987–991 (2011).
- 536 5. Ardui, S., Ameur, A., Vermeesch, J. R., & Hestand, M. S. Single molecule real-time  
537 (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic*  
538 *acids research*. **46**, 2159-2168 (2018).
- 539 6. Brown, C. G., & Clarke, J. Nanopore development at Oxford nanopore. *Nature*  
540 *biotechnology*. **34**, 810-811 (2016).
- 541 7. Tsai, I. J., Otto, T. D., & Berriman, M. Improving draft assemblies by iterative mapping and  
542 assembly of short reads to eliminate gaps. *Genome biology*. **11**, R41 (2010).
- 543 8. Wang, Y., *et al.* Optimizing hybrid assembly of next-generation sequence data from  
544 *Enterococcus faecium*: a microbe with highly divergent genome. *BMC systems biology*. **6**,  
545 S21 (2012).

546 9. Tarailo-Graovac, M. & Chen, N. Using Repeat Masker to Identify Repetitive Elements in  
547 Genomic Sequences. *Current protocol in Bioinformatics* **25**, 1–14 (2009).

548 10. Alhakami, H., Mirebrahim, H., & Lonardi, S.A. Comparative evaluation of genome assembly  
549 reconciliation tools. *Genome biology*. **18**, 93 (2017).

550 11. Pallavi Mishra, Ranjeet Maurya, Vijai K Gupta, Pramod W Ramteke, Soma S  
551 Marla, Anil Kumar. Comparative genomic analysis of monosporidial and  
552 monoteliosporic cultures for unraveling the complexity of molecular pathogenesis of  
553 *Tilletia indica* pathogen of wheat. *Scientific reports*. **9**, 8185 (2019).

554 12. Anil Kumar, Pallavi Mishra, Ranjeet Maurya, A K Mishra, Vijai K Gupta, Pramod W  
555 Ramteke, Soma S Marla, Improved draft genome sequence of a monoteliosporic culture of  
556 the Karnal bunt (*Tilletia indica*) pathogen of wheat. *Genome Announc*. **6**, e00015-18 (2018).

557 13. Pande, S., Sharma, M., & Guvvala, G. An updated review of biology, pathogenicity,  
558 epidemiology and management of wilt disease of pigeonpea (*Cajanus cajan* (L.) Millsp.).  
559 *Journal of Food Legumes*. **26**, 1-14 (2013).

560 14. Perrine David, Nicolas W.G. Chen, Andrea Pedrosa-Harand, Vincent Thareau., *et al.*  
561 A nomadic subtelomeric disease resistance gene cluster in common bean. *Plant  
562 Physiology*. **151**, 1048-1065 (2009).

563 15. Antipov, D., Korobeynikov, A., McLean, J. S., & Pevzner, P. A. hybridSPAdes: an  
564 algorithm for hybrid assembly of short and long reads. *Bioinformatics*. **32**, 1009-1015  
565 (2016).

566 16. Rao, Y.S., Chai, X.W., Wang, Z.F., Nie, Q.H. & Zhang, X.Q. Impact of GC content on gene  
567 expression pattern in chicken. *Genetics Selection Evolution*. **45**, 9 (2013).

568 17. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M.  
569 BUSCO: assessing genome assembly and annotation completeness with single-copy  
570 orthologs. *Bioinformatics*. **31**, 3210-3212 (2015).

571 18. Lazarow K, Doll M-L, Kunze R. Molecular Biology of Maize Ac/Ds elements: an overview.  
572 In: Peterson T, editor. *Plant transposable elements*. Totowa: Humana Press, 59–82 (2013).

573 19. Treangen, T. J., & Salzberg, S. L. Repetitive DNA and next-generation sequencing:  
574 computational challenges and solutions. *Nature Reviews Genetics*. **13**, 36 (2012).

575 20. Ranjeet Maurya, Yeshveer Singh, Manisha Sinha, Kunal Singh, Pallavi  
576 Mishra, Shreenivas Kumar Singh., *et. al.* Transcript profiling reveals potential

577 regulators for oxidative stress response of a necrotrophic chickpea pathogen  
578 *Ascochyta rabiei*. *3 Biotech*. **10**, 1-14 (2020).

579 21. Wetzel, J., Kingsford, A., & Pop, M. Assessing benefits of using mate-pairs to resolve  
580 repeats in *de novo* short read prokaryotic assemblies. *BMC Bioinformatics*. **12**, 95 (2011).

581 22. Jiří Macas, Petr Novák, Jaume Pellicer., *et. al.* In depth characterization of repetitive DNA  
582 in 23 plant genomes reveals sources of genome size variation in the legume tribe  
583 Fabeae. *PLoS One*, **10** (2015).

584 23. Ghurye, J., & Pop, M. Modern technologies and algorithms for scaffolding assembled  
585 genomes. *PLoS computational biology*. 15(6). (2019).

586 24. Bernardo J. Clavijo, Luca Venturini, Christian Schudoma, Gonzalo Garcia Accinelli., *et. al.*  
587 An improved assembly and annotation of the allohexaploid wheat genome identifies  
588 complete families of agronomic genes and provides genomic evidence for chromosomal  
589 translocations. *Genome research*. **27**, 885-896 (2017).

590 25. Marin, D., *et. al.* Validation of a targeted next generation sequencing-based comprehensive  
591 chromosome., *et. al.* Validation of a targeted next generation sequencing-based  
592 comprehensive chromosome screening platform for detection of triploidy in human  
593 blastocysts. *Reproductive biomedicine online*. **36**, 388-395 (2018).

594 26. David M Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane. *et. al.* Phytozome:  
595 a comparative platform for green plant genomics. *Nucleic acids research*. **40**, D1178-D1186  
596 (2012).

597 27. Spapé, M., Verdonschot, R., Dantzig, S. V., & Steenbergen, H. V. The E-Primer: An  
598 introduction to creating psychological experiments in E-Prime®. Leiden University Press.  
599 (p. 208) (2014).

600 28. Zhen Wang, Tao Chen, Weixiang Chen, Kun Chang, Lin Ma., *et. al.* CTAB-assisted  
601 synthesis of single-layer MoS 2-graphene composites as anode materials of Li-ion  
602 batteries. *Journal of Materials Chemistry A*. **1**, 2202-2210 (2013).

603 29. Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop  
604 potato. *Nature*. **475**, 189 (2011).

605 30. Lindner, M. S., Kollock, M., Zickmann, F., & Renard, B. Y. Analyzing genome coverage  
606 profiles with applications to quality control in metagenomics. *Bioinformatics*. **29**, 1260-1267  
607 (2013).

608 31. Pozzi, C., & Salamini, F. Genomics of wheat domestication. In Genomics-assisted crop  
609 improvement. *Springer, Dordrecht*. 453-481 (2007).

610 32. Pandey, I.B. and Tiwari, S. & Singh, S.K. Integrated nutrient management for sustaining the  
611 productivity of Pigeonpea (*Cajanus cajan*) based intercropping systems under rainfed  
612 condition. *Indian Journal of Agronomy*. **58**, 192-197 (2013).

613 33. Wences, A. H., & Schatz, M. C. Metassembler: merging and optimizing de novo genome  
614 assemblies. *Genome biology*. **16**, 207 (2015).

615 34. Boetzer, M., & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome  
616 biology*. **13**, R56 (2012).

617 35. Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. QUAST: quality assessment tool for  
618 genome assemblies. *Bioinformatics*, **29**, 1072-1075 (2013).

619 36. Bailly-Bechet, M., Haudry, A., & Lerat, E. “One code to find them all”: a perl tool to  
620 conveniently parse RepeatMasker output files. *Mobile DNA*. **5**, 13 (2014).

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

624 38. Martin Krzywinski, Jacqueline Schein, İnanç Birol, Joseph Connors. *et al.* Circos: an  
625 information aesthetic for comparative genomics. *Genome research*. **19**, 1639-1645 (2009).

626

627

628

629

630

631

632

633

634

635

636

637

638

639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654

655 **Figure legends**

656 **Figure 1:** SSR distribution frequency (A) Distribution of different repeats type classes (B)  
657 Frequency of classified predominant repeats.

658 **Figure 2:** Circos diagram presenting syntenic relationship between NBS-LRR proteins from  
659 pigeonpea (Cc), *Glycin max* (Gm) and *Medicago truncatula* (Mt) pseudomolecules.  
660 Pseudomolecules of the two target species were labeled as Gm1-20 and Mt1-8. Pigeonpea  
661 pseudomolecules are labeled in different colours and labeled as Cc1-11. Colinear blocks are  
662 coloured according to the colour of the corresponding Pigeonpea pseudomolecule. Each ribbon  
663 radiating from a pigeonpea pseudomolecule represents a NBS-LRR similarity block between  
664 pigeonpea and other legumes.

665 **Figure 3:** PCR amplification of *Fusarium* wilts resistant RGA among Pigeonpea genotypes.

666 **Figure 4:** Experimental Frame work depicting reconstruction steps of Pigeonpea genome.

667

668 **Table legends**

669 **Table 1:** Genome assembly statistics of draft assemblies A1, A2 and finished A3 assembly.

670 **Table 2:** Results of Gene finding.

671 **Table 3:** Repetitive sequences of draft assemblies A1, A2 and finished A3 assembly.

672 **Table 4:** Results of microsatellite search in the improved pigeonpea assembly A3.

673 **Supplementary Information legends**

674 **Supplementary Table 1:** BUSCO (Benchmarking Universal Single-Copy Orthologs) genes  
675 distribution.

676 **Supplementary Table 2:** Putative Disease resistance genes predicted from improved reassembly  
677 of Pigeonpea.

678 **Supplementary Table 3:** Numbers of predominant SSRs repeats.

679 **Supplementary Table 4:** Resistance gene analogues sharing homology with species.

680 **Supplementary Table 5:** NBS-LRR domain containing resistance gene analogues to Pigeonpea  
681 [*Cajanus cajan* (L.) Millsp.].

682 **Supplementary Table 6:** List of Pigeonpea disease resistance genes submitted to NCBI.

683 **Supplementary File 7:** List of primer sequences used in PCR amplication.

684 **Supplementary Table 8:** Pigeonpea phenotypic field scores for *Fusarium* wilt disease reaction.

685

686

687

688

689

690

691

692

693

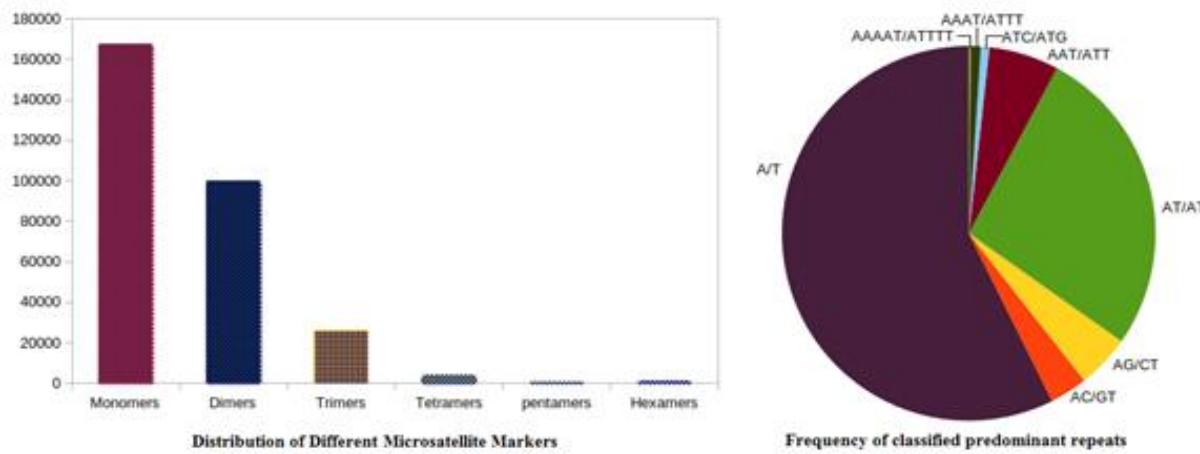
694

695

696

697

698  
699  
700  
701  
702  
703  
704  
705  
706  
707



708  
709

710 **Figure 1: SSR distribution frequency** (A) Distribution of different repeats type classes (B)  
711 Frequency of classified predominant repeats.

712

713

714

715

716

717

718

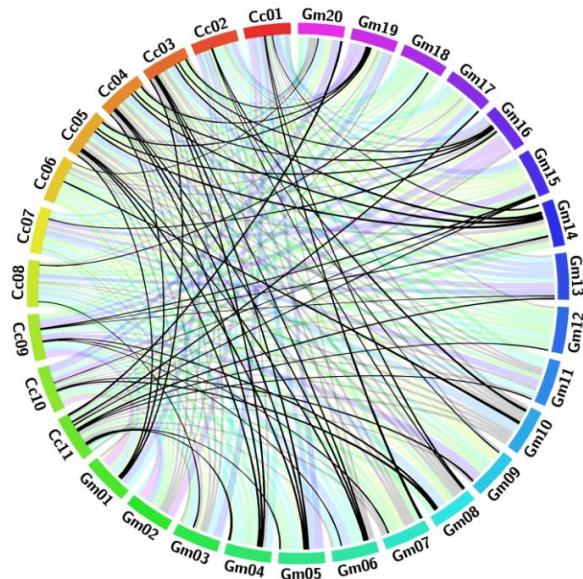
719

720

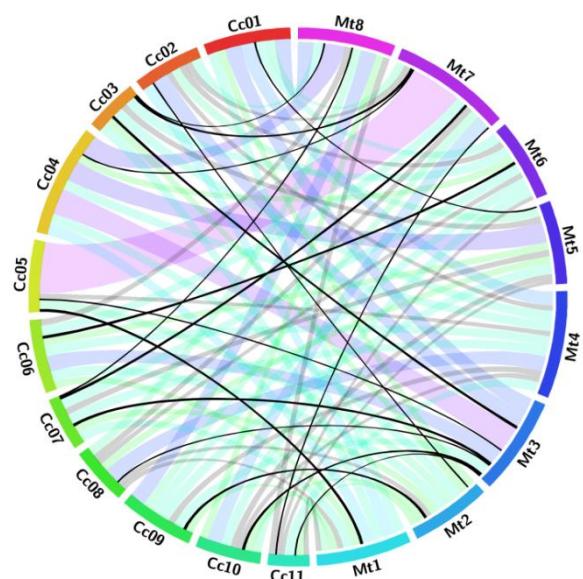
721

722

723



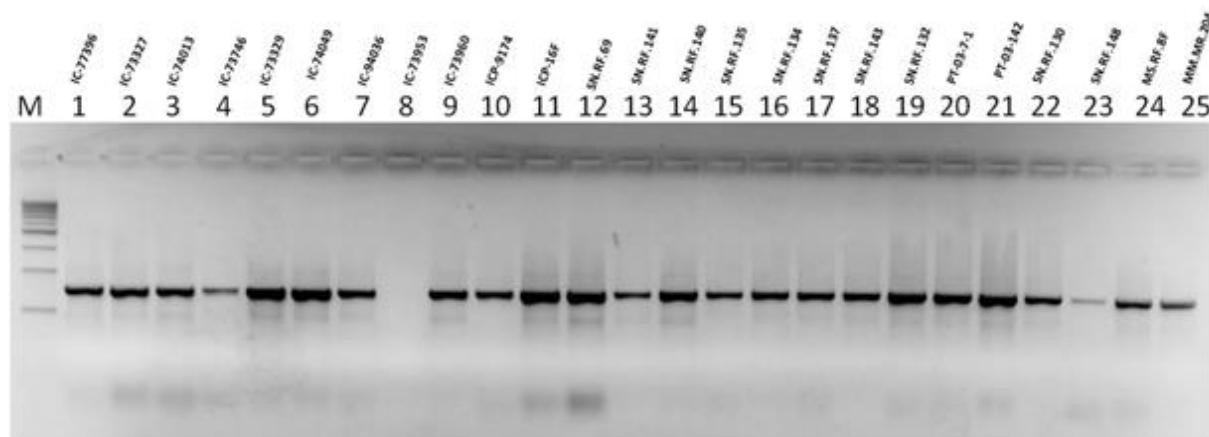
724



725

726 **Figure 2:** Circos diagram presenting syntenic relationship between NBS-LRR proteins from  
727 pigeonpea (Cc), *Glycin max* (Gm) and *Medicago truncatula* (Mt) pseudomolecules.  
728 Pseudomolecules of the two target species were labeled as Gm1-20 and Mt1–8. Pigeonpea  
729 pseudomolecules are labeled in different colours and labeled as Cc1-11. Colinear blocks are  
730 coloured according to the colour of the corresponding Pigeonpea pseudomolecule. Each ribbon  
731 radiating from a pigeonpea pseudomolecule represents a NBS-LRR similarity block between  
732 pigeonpea and other legumes.

733

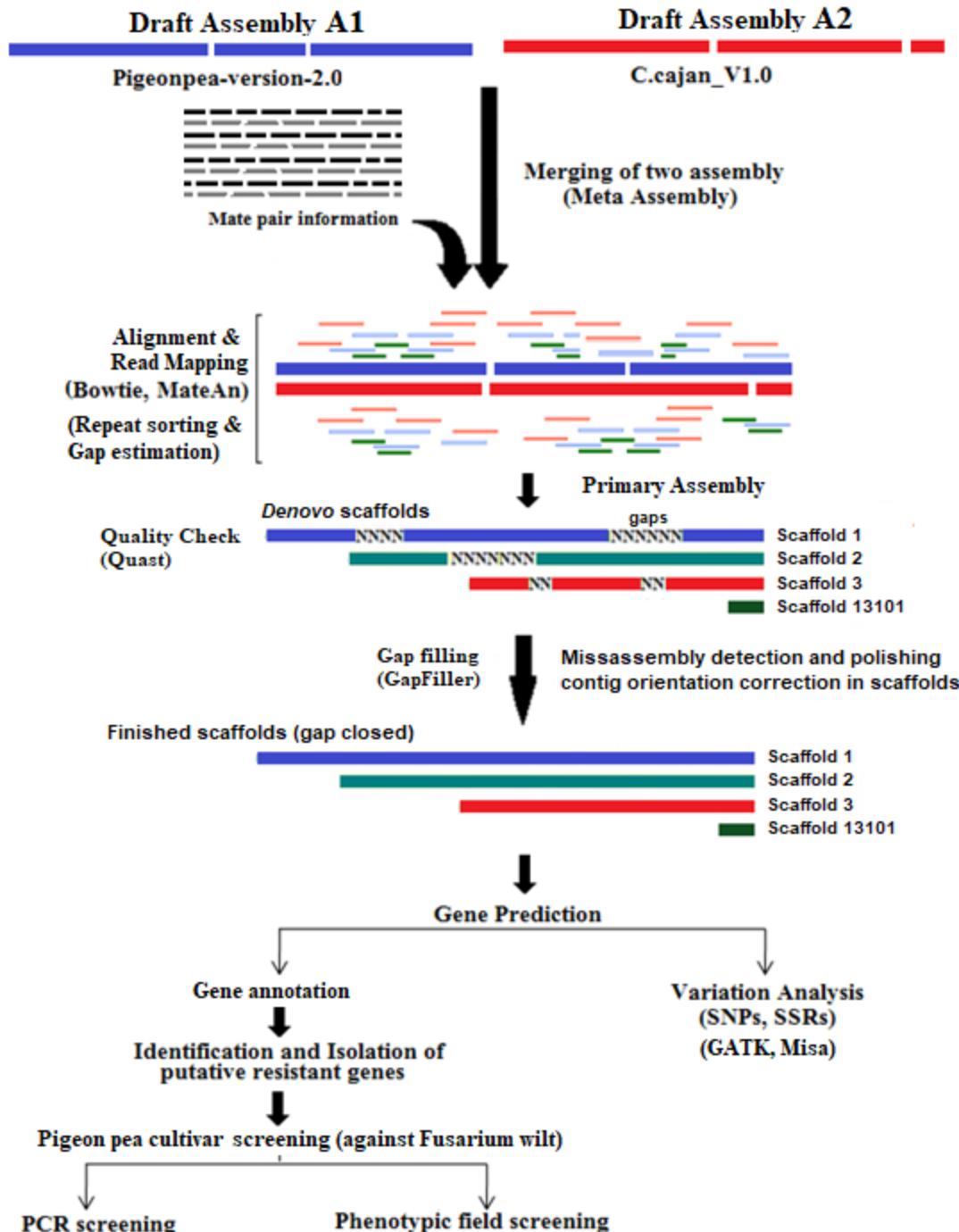


1-11: Fwilt Resistant; 12-18: Fwilt tolerant; 19-23: Fwilt susceptible; 24-25: yellow mosaic resistant  
M: 500 bp ladder, Primer ID: 06, 1-25 Pigeon Pea genotypes, Product size 0.7 kb

734  
735

**Figure 3:** PCR amplification of *Fusarium* wilts resistant RGA among Pigeonpea genotypes.

736



737

738

**Figure 4:** Experimental Frame work depicting reconstruction steps of Pigeonpea genome.

## Figure legends

**Figure 1: SSR distribution frequency** (A) Distribution of different repeats type classes (B) Frequency of classified predominant repeats.

**Figure 2:** Circos diagram presenting syntenic relationship between NBS-LRR proteins from pigeonpea (Cc), *Glycin max* (Gm) and *Medicago truncatula* (Mt) pseudomolecules. Pseudomolecules of the two target species were labeled as Gm1-20 and Mt1–8. Pigeonpea pseudomolecules are labeled in different colours and labeled as Cc1-11. Colinear blocks are coloured according to the colour of the corresponding Pigeonpea pseudomolecule. Each ribbon radiating from a pigeonpea pseudomolecule represents a NBS-LRR similarity block between pigeonpea and other legumes.

bioRxiv preprint doi: <https://doi.org/10.1101/2020.08.10.243949>; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

**Figure 3: PCR amplification of *Fusarium* wilts resistant RGA among Pigeonpea genotypes.**

**Figure 4:** Experimental Frame work depicting reconstruction steps of Pigeonpea genome.

## Table legends

**Table 1:** Genome assembly statistics of draft assemblies A1, A2 and finished A3 assembly.

**Table 2:** Results of Gene finding.

**Table 3:** Repetitive sequences of draft assemblies A1, A2 and finished A3 assembly.

**Table 4:** Results of microsatellite search in the improved pigeonpea assembly A3.

## Supplementary Information legends

**Supplementary Table 1:** BUSCO (Benchmarking Universal Single-Copy Orthologs) genes distribution.

**Supplementary Table 2:** Putative Disease resistance genes predicted from improved reassembly of Pigeonpea.

**Supplementary Table 3:** Numbers of predominant SSRs repeats.

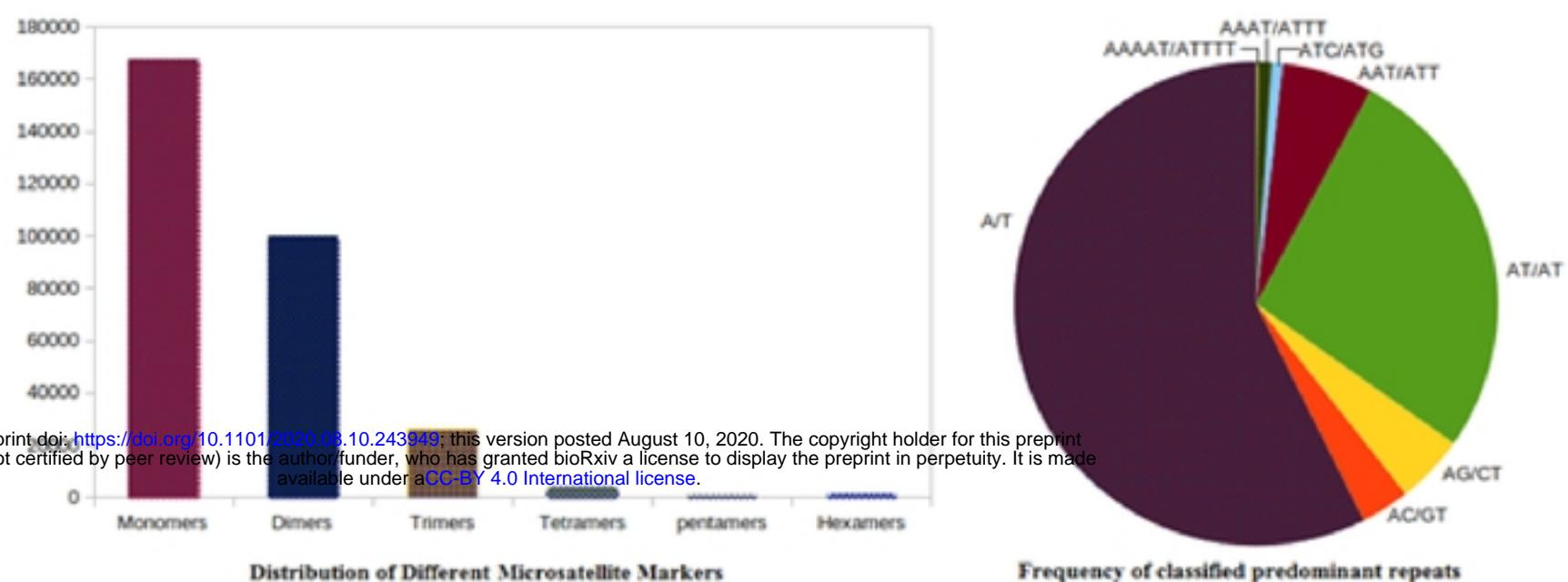
**Supplementary Table 4:** Resistance gene analogues sharing homology with species.

**Supplementary Table 5:** NBS-LRR domain containing resistance gene analogues to Pigeonpea [*Cajanus cajan* (L.) Millsp.].

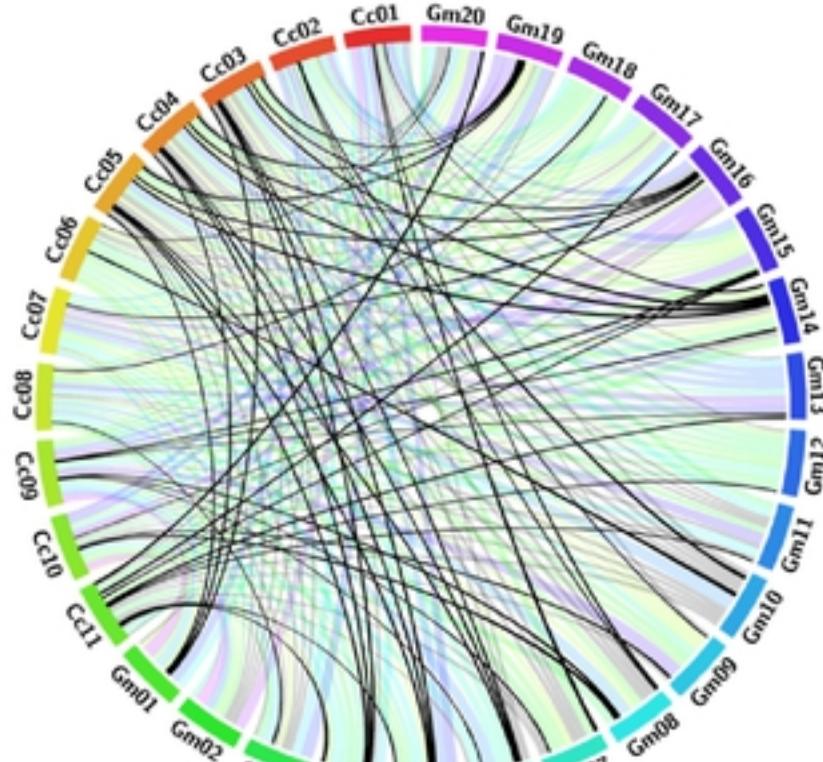
**Supplementary Table 6:** List of Pigeonpea disease resistance genes submitted to NCBI.

**Supplementary File 7:** List of primer sequences used in PCR amplification.

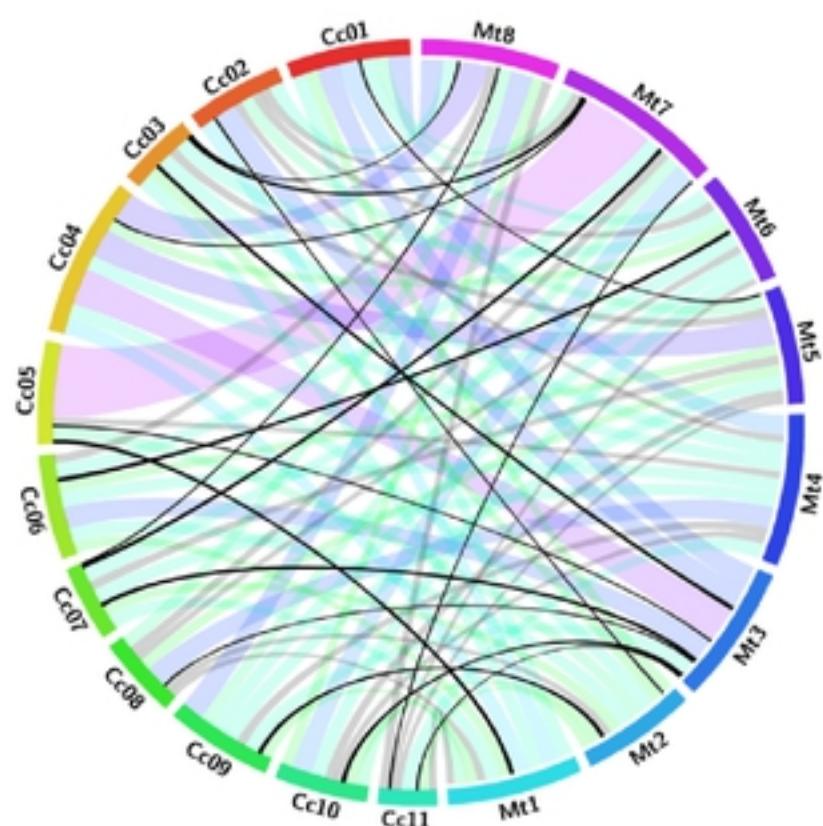
**Supplementary Table 8:** Pigeonpea phenotypic field scores for *Fusarium* wilt disease reaction.



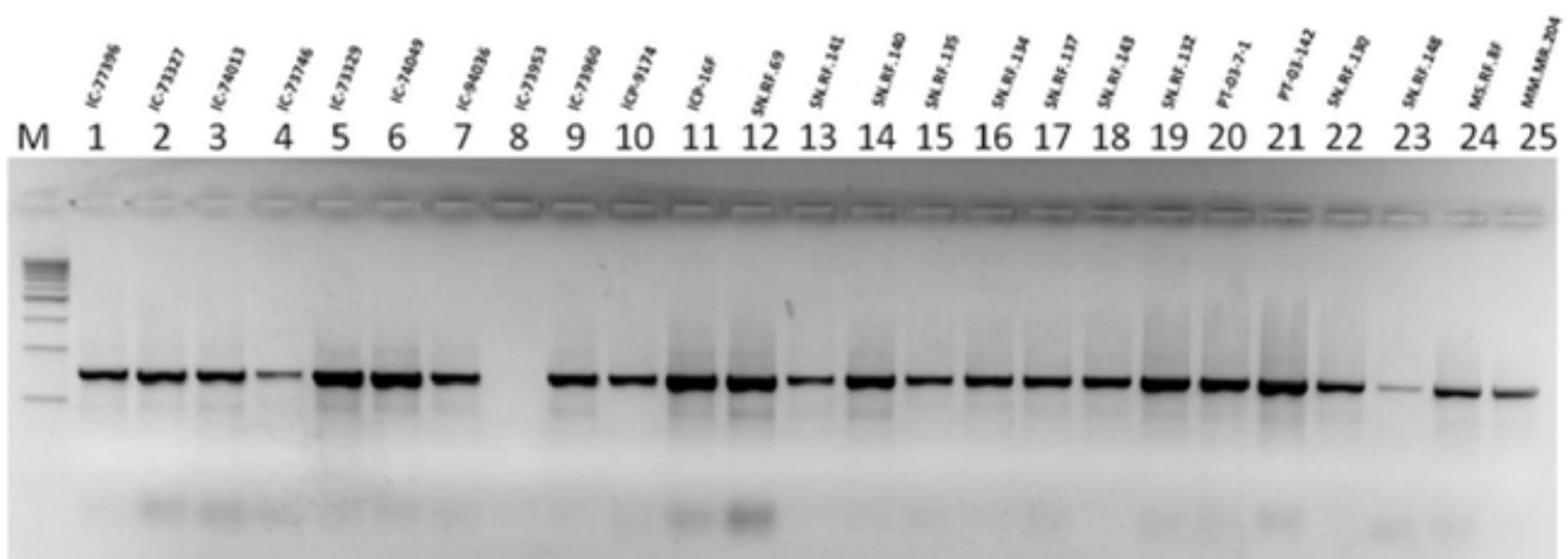
**Figure 1: SSR distribution frequency (A) Distribution of different repeats types (B) Frequency occurrence of repeats.**



bioRxiv preprint doi: <https://doi.org/10.1101/2020.08.10.243949>; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

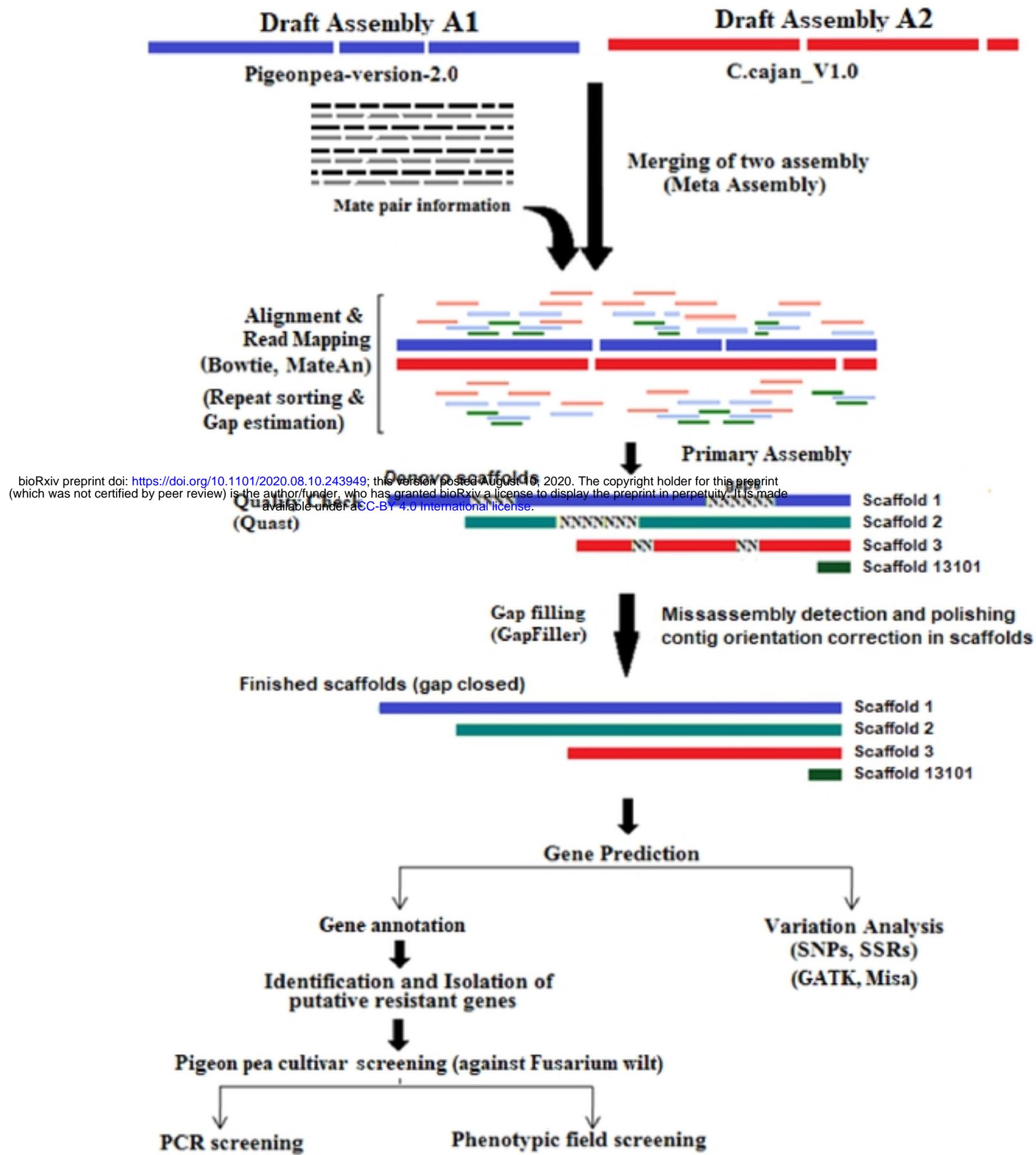


**Figure 2:** Circos diagram presenting synteny relationship between NBS-LRR proteins from pigeonpea (Cc), *Glycin max* (Gm) and *Medicago truncatula* (Mt) pseudomolecules. Pseudomolecules of the two target species were labeled as Gm1-20 and Mt1-8. Pigeonpea pseudomolecules are labeled in different colours and labeled as Cc1-11. Colinear blocks are coloured according to the colour of the corresponding Pigeonpea pseudomolecule. Each ribbon radiating from a pigeonpea pseudomolecule represents a NBS-LRR similarity block between pigeonpea and other legumes.



1-11: Fwilt Resistant; 12-18: Fwilt tolerant; 19-23: Fwilt susceptible; 24-25: yellow mosaic resistant  
 M: 500 bp ladder, Primer ID: 06, 1-25 Pigeon Pea genotypes, Product size 0.7 kb

bioRxiv preprint doi: <https://doi.org/10.1101/2020.08.10.243940>; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



**Figure 4:** Experimental Frame work depicting reconstruction