American Journal of Botany

American Journal of Botany 99(2): 000–000. 2012.

Coverage-based consensus calling (CbCC) of short sequence reads and comparison of CbCC results to identify SNPs in chickpea (*Cicer arietinum*; Fabaceae), a crop species without a reference genome[1]

Sarwar Azam[2], Vivek Thakur[2,3], Pradeep Ruperao[2,4], Trushar Shah[2], Jayashree Balaji[2], BhanuPrakash Amindala[2], Andrew D. Farmer[5], David J. Studholme[6,7], Gregory D. May[5], David Edwards[4], Jonathan D. G. Jones[6], and Rajeev K. Varshney[2,8,9]

[2]Centre of Excellence in Genomics (CEG), International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502324, Andhra Pradesh, India; [3]C4 Center, International Rice Research Institute (IRRI), DAPO 7777, Metro Manila, Philippines; [4]Australian Centre for Plant Functional Genomics, School of Agriculture and Food Sciences, University of Queensland, Brisbane, QLD 4072, Australia; [5]National Center for Genome Resources (NCGR), New Mexico, Santa Fe 87505 USA; [6]The Sainsbury Laboratory (TSL), JIC, Norwich Research Park, Norwich NR4 7UH, UK; [7]School of Biosciences, Geoffrey Pope Building, University of Exeter, Stocker Road, Exeter, EX4 4QD, UK; and [8]Theme-Comparative and Applied Genomics, CGIAR Generation Challenge Program (GCP), c/o CIMMYT, Mexico DF, Mexico

- *Premise of the study:* Next-generation sequencing (NGS) technologies are frequently used for resequencing and mining of single nucleotide polymorphisms (SNPs) by comparison to a reference genome. In crop species such as chickpea (*Cicer arietinum*) that lack a reference genome sequence, NGS-based SNP discovery is a challenge. Therefore, unlike probability-based statistical approaches for consensus calling and by comparison with a reference sequence, a coverage-based consensus calling (CbCC) approach was applied and two genotypes were compared for SNP identification.
- *Methods:* A CbCC approach is used in this study with four commonly used short read alignment tools (Maq, Bowtie, Novoalign, and SOAP2) and 15.7 and 22.1 million Illumina reads for chickpea genotypes ICC4958 and ICC1882, together with the chickpea trancriptome assembly (CaTA).
- *Key results:* A nonredundant set of 4543 SNPs was identified between two chickpea genotypes. Experimental validation of 224 randomly selected SNPs showed superiority of Maq among individual tools, as 50.0% of SNPs predicted by Maq were true SNPs. For combinations of two tools, greatest accuracy (55.7%) was reported for Maq and Bowtie, with a combination of Bowtie, Maq, and Novoalign identifying 61.5% true SNPs. SNP prediction accuracy generally increased with increasing reads depth.
- *Conclusions:* This study provides a benchmark comparison of tools as well as read depths for four commonly used tools for NGS SNP discovery in a crop species without a reference genome sequence. In addition, a large number of SNPs have been identified in chickpea that would be useful for molecular breeding.

Key words: chickpea; *Cicer arietinum*; Fabaceae; legumes; molecular breeding; next-generation sequencing; SNP discovery.

The importance of genomics for human health and agricultural applications led to the development of high-throughput sequencing methods that can generate data rapidly at a relatively low cost. Several such technologies, commonly known as next-generation sequencing (NGS), have been developed over the last decade (Mardis, 2008; MacLean et al., 2009; Metzker, 2010). Among them, Illumina (http://www.illumina.com), Roche 454/FLX (http://www.454.com), and ABI-SOLiD (http://www.appliedbiosystems.com) are already in commercial use, and others such as Heliscope (http://www.helicosbio.com) and Pacific Biosciences (http://www.pacificbiosciences.com) are recent entrants. These technologies differ not only in their sequencing chemistry but also in terms of rate of data generation, average length of sequence reads, accuracy, and costs (Varshney et al., 2009).

NGS is being increasingly applied for crop genome sequencing (Imelfort and Edwards, 2009; Marshall et al., 2010; Varshney et al., 2011), and the availability of reference genome sequences for human and several animal and plant species has triggered whole genome resequencing of genotypes of such species by using NGS technologies (Bentley, 2006; Hillier et al., 2008; Hudson, 2008; Wang et al., 2008; Wheeler et al., 2008; Ahn et al., 2009). The alignment of resequence data with a reference genome facilitates the identification of sequence variants such as single nucleotide polymorphisms (SNPs) and insertion–deletions (indels) between individuals as well as with the

[1] Manuscript received 27 August 2011; revision accepted 2 January 2012.

The authors are grateful to the CGIAR Generation Challenge Program for financial support. Thanks are due to several colleagues, in particular Chris Town and Yongli Xiao from The J. Craig Venter Institute (USA); Jimmy Woodward from NCGR; Doug Cook from University of California-Davis; Dan MacLean from TSL; Richard Bruskiewich from IRRI; and Abhishek Rathore, Spurthi Nayak, Pavana Hiremath, Roma Rani Das, Neha Gujaria, Ashish Kumar, and Shamshad Alam from ICRISAT for their cooperation and help in generation of some data, analysis, and interpretation of results presented in this study. S.A. and V.T. contributed equally to this study.

[9] Author for correspondence (e-mail: r.k.varshney@cgiar.org)

doi:10.3732/ajb.1100419

reference genome sequence. High-throughput transcriptome sequencing, also known as RNA-seq (Wang et al., 2009), using NGS technologies can be used for variant detection (Imelfort et al., 2009), splice-site detection (Sultan et al., 2008) and gene expression profiling (Marioni et al., 2008; Morin et al., 2008). Plant researchers have started to use NGS approaches routinely, especially with the decreasing costs associated with NGS technologies (Varshney et al., 2009; Edwards and Batley, 2010). While NGS is being used for allele discovery and gene expression analysis in model or major crop species (Deschamps et al., 2010; Duran et al., 2010; Hyten et al., 2010a; Lu et al., 2010), SNP marker discovery is the major growth area of NGS applications in plant species that do not have a reference genome sequence data (Duran et al., 2009a; Trick et al., 2009; Parchman et al., 2010; Dubey et al., 2011; Garget al., 2011; Hiremath et al., 2011).

Several bioinformatics tools are available to detect SNPs from sequence data generated using traditional Sanger sequencing approach (Barker et al., 2003; Zhang et al., 2005; Stephens et al., 2006; Duran et al., 2009b, c; Jayashree et al., 2009). These are generally unsuitable for the detection of SNPs from short sequence reads such as those generated by Illumina/Solexa or ABI-SOLiD technologies. In the last few years, several tools have been developed for SNP discovery from short read data. These tools generally employ three steps for SNP identification: (1) alignment of reads from different genotypes onto a reference sequence (i.e., mapping), (2) the generation of a consensus sequence for individual genotypes on the basis of posterior probability (i.e., consensus calling), and (3) the identification of SNPs by comparison with the reference sequence (i.e., SNP calling).

A number of open source tools are currently available that can execute one, two, or all three steps mentioned. For instance, Maq (Li et al., 2008), SOAP2/SOAPsnp (Li et al., 2009), Mosaik/PolyBayes (Marth et al., 1999), and Bowtie (Langmead et al., 2009) are tools that can perform mapping, consensus calling, and SNP calling, while several others such as Novoalign (Hercus, 2009) and RMAP (Smith et al., 2008) can only be used for mapping. In case of mapping, all tools use heuristics techniques to align reads to the reference because running accurate alignment algorithms to identify all possible places where reads may map to the reference sequence is computationally infeasible (Flicek and Birney, 2009). Therefore, these tools use various approaches to identify a subset of places in the reference sequence where the best mapping is most likely to be found. Then, more accurate alignment algorithms such as Smith-Waterman are run on the limited subset (Batzoglou, 2005). Algorithms used to search the small set of potential alignments in the reference sequence can be broadly classified into two main categories: hash based, hashing either reference sequence (Novoalign) or reads (Maq); and Burrows–Wheeler transformation (Flicek and Birney, 2009; Li and Homer, 2010) (e.g., Bowtie and SOAP2). For consensus calling, in general, the posterior probability of all bases from different reads is often computed. Subsequent comparison with the reference sequence identifies SNPs.

Apart from the open source tools mentioned, some proprietary tools are also available as a part of workbench or integrated analyses solutions, e.g., NextGENe (http://www.softgenetics.com/NextGENe.html), CLCBio Genomics workbench (http://www.clcbio.com), ELANDv1 or ELANDv2 from Illumina (http://www.illumina.com/support/sequencing/sequencing_software/casava/downloads.ilmn), and Alpheus (Miller et al., 2008). NGS technologies are being used in our research to identify SNPs in crop species such as chickpea *(Cicer arietinum* L.; Fabaceae) and pearl millet (*Pennisetum glaucum* L.; Poaceae) that do not have a reference genome sequence. In addition to several other studies, Varshney et al. (2009) proposed using a transcript assembly (TA) based on Sanger ESTs or 454/FLX transcript reads as a reference and then mapping reads, generated from transcriptomes of parental genotypes of mapping/segregating populations. The bioinformatics analysis tools currently available, as mentioned already, have limited utility in this regard because they identify SNPs between the genotypes and the reference genotype whose genome/transcriptome is used for mapping the reads.

When calling variants based on short reads (especially those generated by the first version of NGS machines like Illumina GA I), there are at least two major challenges: the first is the poor read quality, especially toward the 3′ end of sequence reads. This is likely to affect the reliability of consensus calling, particularly at positions with lower coverage. The second major challenge is accuracy of mapping of reads especially when these are single-ended and short (≤36 bp). To address these issues, this study explores the possibility of using coverage-based consensus calling (CbCC) and comparison of CbCC results for SNP calling without considering the reference sequence. In this context, the current study was undertaken with the following objectives: (1) CbCC for reads from two chickpea genotypes, (2) comparison of four open source tools (Bowtie, Maq, Novoalign, and SOAP2) to identify SNPs, and (3) optimization of the read depth criteria. In addition, this analysis has also provided SNPs to develop genetic markers for use in genetics research and breeding applications in chickpea.

MATERIALS AND METHODS

***Short sequence read and transcript assembly (TA) data sets*—**Illumina GA I sequencing of transcripts from drought-stress-challenged tissues of chickpea genotypes ICC4958 and ICC1882 generated 15.7 and 22.1 million reads of 36-bp length, respectively (SRA030700.1, Hiremath et al., 2011). These genotypes are the parents of a mapping population that segregate for drought tolerance. For mapping reads, the chickpea transcriptome assembly (CaTA), comprising 98 534 tentative unique sequences (TUSs) including 46 740 contigs and 51 794 singletons, was used as a reference (Hiremath et al., 2011).

***Mapping tools for aligning reads to the chickpea transcriptome assembly*—**Four tools, Maq (Li et al., 2008) (version 0.7.0; http://maq.sourceforge.net/), SOAP2 (Li et al., 2009) (version 2.18; http://soap.genomics.org.cn/soapaligner.html), Bowtie (Langmead et al., 2009) (version 0.12.7; http://bowtie-bio.sourceforge.net/index.shtml), and Novoalign (Hercus, 2009) (version 2.03.12; http://www.novocraft.com/main/index.php) were used for mapping the reads onto the CaTA. In case of Maq, Bowtie, and SOAP2, mapping was performed using following criteria: (1) for each read, a maximum of seven mismatches i.e., two in seed region (first 24 bases) and five in nonseed region were allowed, and (2) the sum of quality of mismatch bases should not exceed 70. Because Novoalign does not have the option to specify the above parameters, default parameters are used for mapping. With all the tools, reads were allowed to map randomly if multiple best alignments were found.

***CbCC-based SNP calling*—**Aligned reads from all four tools were used for defining the major base across the alignment for each genotype using an in-house developed Perl script (ConsensusCallingSNP.pl, http://www.icrisat.org/azam_et_al_2012/ConsensusCallingSNP.pl). Subsequently, the major base (which is a base in majority or consensus base on a locus) at each position in aligned reads was compared between both genotypes. If a variation is found between the major bases of the two genotypes, the variation was reported as candidate SNP.

***Experimental validation of SNPs*—**For validating the predicted SNPs, two approaches, allele-specific sequencing (Nayak et al., 2010) and KASPar SNP genotyping (KBioscience, England, http://www.kbioscience.co.uk/) assays were applied. In the case of allele-specific sequencing, primer pairs were designed using the program Primer3-v.0.4.0 (Rozen and Skaletsky, 2000)/GENETOOL (version 1.0; http://www.doubletwist.com), based on the corresponding TUS so that the selected SNP is present in the target sequencing region. Subsequently, these primer pairs were used to generate amplicons for ICC4958 and ICC1882, and these amplicons were sequenced following the protocol of Nayak et al. (2010). Sequence data generated for both genotypes were compared with the sequence of the corresponding TUS at the targeted SNP position using DNA Baser (DNA Baser Sequence Assembler v3.0, 2011; Heracle BioSoft, http://www.DnaBaser.com). In some cases, SNP genotyping was conducted using KASPar assays at KBioscience UK. Complete details on the principle and procedure of the assay are available at http://www.kbioscience.co.uk/reagents/KASP_manual.pdf. The called alleles were compared with those of the predicted SNPs in the corresponding TUS.

All the raw data as well as the processed data related to the manuscript can be accessed through the URL: http://www.icrisat.org/azam_et_al_2012/index.html .

RESULTS AND DISCUSSION

Mapping of reads onto CaTA—Illumina GA I sequencing on RNA samples isolated from drought-stress-challenged root tissues of two genotypes namely ICC4958 (drought tolerant) and ICC1882 (drought susceptible) generated 15.7 and 22.1 million reads of 36-bp length, respectively (SRA030700.1, Hiremath et al., 2011). Four tools, Bowtie, Maq, Novoalign, and SOAP2 were used for mapping 37.8 million reads of these genotypes onto the chickpea transcriptome assembly (CaTA), developed in an earlier study (Hiremath et al., 2011) (Table 1). Among these tools, Bowtie could align 73.6% and 63.5% of reads from ICC4958 and ICC1882, respectively, while SOAP2 aligned 68.5% and 58.7% of reads in these genotypes. These observations were not unexpected because the tools employ different algorithms and parameters. For instance, Maq and Novoalign use a hash-based algorithm, while Bowtie and SOAP2 are based on Burrows–Wheeler transformation (Flicek and Birney, 2009; Li and Durbin, 2009; Li and Homer, 2010). However, the hashing strategies used by Maq and Novoalign are different (Flicek and Birney, 2009; Li and Homer, 2010); Maq makes hashes of reads and then searches for set of potential good alignments, while Novoalign makes a hash of the reference sequence.

When aligning reads onto the CaTA, 47–60% of reads identified more than one location. This is not uncommon (Pasaniuc et al., 2010), and similar results have been observed in other organisms including animal (17–24%) (Mortazavi et al., 2008; Costa et al., 2010; Li et al., 2010) and plant species (52%) (Li et al., 2010). Duplicate mapping could be due to the following reasons: (1) low quality of the reference sequence (CaTA) or sequencing errors in the reads, (2) occurrence of paralogous gene families, and (3) high similarity between alternatively spliced isoforms of the given gene (Li et al., 2010). Because the proportion of such reads was quite large, discarding these reads was not affordable. Therefore, a unique location, selected randomly was assigned to such reads. This strategy has been previously applied with human data set (Li et al., 2008).

Coverage-based consensus calling—Aligned reads for each genotype from all four tools were used for coverage-based consensus calling (CbCC). A Perl script was written (ConsensusCallingSNP.pl) that defines the major base at every position of the aligned sequences for each genotype using the following criteria: (1) the base quality is >20, (2) the minimum read depth is 2 in both genotypes, and (3) the frequency of the major base (f_{major}) in each of the two genotypes is >0.66.

The current data sets comprise single short reads (≤36 bp), which contain sequence errors, especially toward the 3′ end of reads. In addition, because multiple best hits were identified for 47–60% of reads, a unique location was selected randomly for such reads. In this case, a probability-based statistical approach for consensus calling, as performed in general cases (Marth et al., 1999; Li et al., 2008), is not appropriate. Considering the multiple read mapping and the lack of a genomic reference, the CbCC approach is a more appropriate method for consensus calling, especially when using relatively small and low-quality data sets.

SNP discovery and validation—On the basis of a comparison of CbCC results from four alignment tools, a total of 7015 SNPs was predicted. Novoalign predicted the least number of SNPs (1272), with the maximum number identified by Bowtie (2454). Maq and SOAP2 predicted 1329 and 1960 SNPs, respectively (Table 1). A significant portion of SNPs predicted by each tool was unique to the tool (Fig. 1). For example, Novoalign had the maximum proportion of unique SNPs (58.3%), while Maq had the fewest unique SNPs (33.6%). Bowtie and SOAP2 predicted 43.9% and 35.6% unique SNPs. Unique SNP identification can be attributed to the different sensitivity of read alignments using different tools. This may be due to the intrinsic alignment policy or different algorithms employed to identify the best alignment or may be due to random placement of multiple best hit (Souche et al., 2007). After considering redundancy, a total of 4543 nonredundant SNPs were predicted (Fig. 1).

To estimate the accuracy of SNP prediction using each of the four alignment tools, a set of 224 randomly selected SNPs was validated. A total of 190 SNPs were selected for validation using allele-specific sequencing of PCR amplicons in both directions using Sanger sequencing. In addition, KASPar assays were developed to validate 34 SNPs. Only 79 of the predicted SNPs (35.3%) were found to be correct (60 SNPs via allele-specific sequencing and 19 SNPs via KASPar assays). In a comparison of these results with other studies including *Brassica napus*

Table 1. Mapping of short sequence reads and SNP prediction between genotypes ICC 4958 and 1882 of chickpea.

	Mapping on to CaTA					
	ICC 4958		ICC 1882			
Tool	Reads aligned	% of alignment	Reads aligned	% of alignment	No. of TUSs having alignment in both genotypes	Predicted SNPs between ICC4958 and ICC1882
Bowtie	11 530 446	73.6	14 036 566	63.5	46 170	2454
Maq	11 446 328	73.1	13 959 395	63.1	45 525	1329
Novoalign	11 481 173	73.3	13 869 083	62.7	46 029	1272
SOAP2	10 731 068	68.5	12 957 191	58.7	44 304	1960

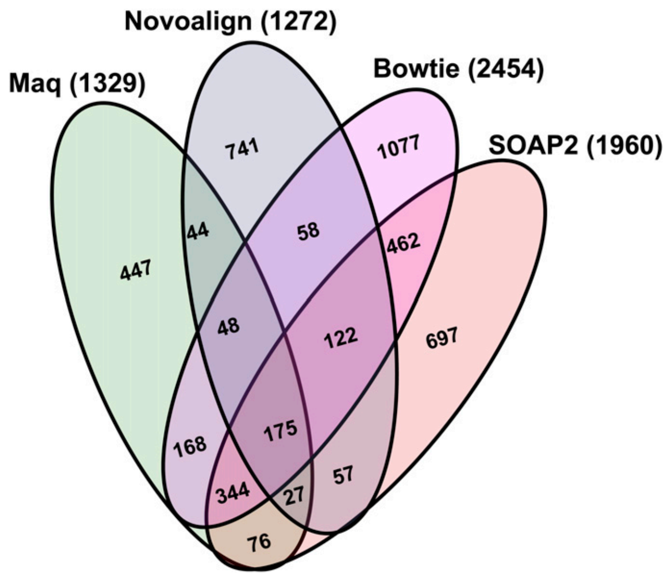

Fig. 1. A venn diagram showing distribution of predicted SNPs by different/ combination of tools. Numbers of SNPs predicted by the four tools have been shown in four differently colored ellipses. The numbers of SNPs present in intersections of two or more ellipses represent the SNPs detected by combination of two or more tools. Sum of the number of SNPs present in different intersections in a given ellipse represents the number of redundant SNPs for the respective tool.

(84–91%) (Trick et al., 2009), *Eucalyptus grandis* (83%) (Novaes et al., 2008), maize (overall 85%, while at lower depth 64%) (Barbazuk et al., 2007), rice (96.4%) (Deschamps et al., 2010), soybean (79–92%) (Hyten et al., 2010a) and common bean (86%) (Hyten et al., 2010b), our SNP discovery rate as well as prediction accuracy is low. Lower SNP frequency is a result of the narrow genetic base of the gene pool of cultivated chickpea as well as the use of transcript reads for SNP discovery (Gujaria et al., 2011; Hiremath et al., 2011). Differences in prediction accuracy in chickpea as compared to other species as mentioned earlier can be attributed to the larger data sets, use of paired reads as well as the availability of reference genome sequences in several of the aforementioned studies. Considering a 35.3% success rate across the total nonredundant (4543) SNPs, a set of 1602 valid SNPs can be anticipated, and this set of SNPs will be helpful to integrate in the transcript genetic map (Gujaria et al., 2011). These markers will be useful to accelerate research and breeding applications in chickpea by deploying markers for the molecular characterization of germplasm collections, genetic diversity studies, trait mapping, and marker-assisted selection studies (Upadhyaya et al., 2011).

Comparison of four tools—To compare the use of different combinations of alignment tools, we categorized all predicted SNPs into 15 classes (Table 2). Each class represents SNPs either unique to each tool or an intersection of two or more tools. From this set, 224 SNPs were selected for validation, but only 79 of these SNPs were found as true positives. The prediction and validation results for these SNPs were compared across all 15 classes where stringent classes are subsets of larger classes. For example, class XII (Bowtie, Maq, SOAP2), is the subset of class V (Bowtie, Maq), class VII (Bowtie, SOAP2), and class IX (Maq, SOAP2), and class XV (Bowtie, Maq, Novoalign, SOAP2) is the subset of all classes. Detailed comparative analysis across individual tools highlighted the superiority of Maq, as 50.0% of Maq predicted SNPs were found to be correct in the validation study. In contrast, only 30.2% of SNPs identified by Novoalign were correct. When comparing combinations of two tools, 55.7% accuracy was reported for SNPs predicted by both Maq and Bowtie (class V), while SNPs predicted by both Novoalign and SOAP2 (class X) were the least accurate (38.1%). Among combinations of three tools, class XIV (Bowtie, Maq, and Novoalign) provided greatest accuracy (61.5%). It was observed that inclusion of Maq in a combination of two or three tools provided greatest accuracy. As expected, class XV containing SNPs predicted by all four tools provided the greatest accuracy (62.5%).

The confidence of a predicted SNP being correct increases if it is predicted by more than one tool (Souche et al., 2007). For instance, in the current study, 2962 (65.2%) SNPs were unique to a particular tool, and 1581 (34.8%) SNPs were identified by more than one tool. When these results were correlated with true positives, SNPs unique to one tool showed a poor rate (mean 31.2%) of accuracy compared to those identified by two (mean 46.6%), three (mean 53.9%), or all four tools (62.5%). An attempt was made to estimate the accuracy rate of predicted SNPs that are unique to individual tools. The greatest accuracy (47%) was observed for SOAP2, which compared to 9% for Novoalign (Appendix S1, see online Supplemental Data for this article). These results highlight the use of SOAP2 for SNP discovery because it identified a greater proportion of unique SNPs compared to the other tools. The present analysis suggests Novoalign has a low accuracy rate (30%), findings which support earlier studies (Xu et al., 2011; Yu, 2011).

Comparison of SNP prediction and validation for different read depths—To identify the optimal read depth for SNP prediction, we analyzed SNPs that were predicted at different thresholds of read depth in the context of the validation results. Four different read depth thresholds 2, 3, 4–10, and >10 were assessed. The number of SNPs predicted by each of the tools at different read depths has been summarized in Table 3. Most of the predicted SNPs(40.0–44.3%) are from the category with read depth 2. Between 33.6 and 33.7% of SNPs were predicted at a read depth of 3, 16.7–17.6% at a depth between 4 and 10, while only 5.5–9.2% of SNPs were predicted at a read depth >10.

As the read depth increases, the rate of SNP prediction accuracy also increases (online Appendix S2). The rate of SNP prediction accuracy was poor (20.0–26.7%) at read depth of 2. As the depth increased to 3, Maq had the greatest accuracy 51.6%, much higher than the other tools (20.7–38.7%). Similarly, at a depth of 4–10, Maq demonstrated the greatest accuracy (58.6%) compared to 44.4% for Novoalign and 47.6% and 47.2% for Bowtie and SOAP2, respectively. At read depths of >10, Novoalign showed the least accuracy (60.0%), Maq and SOAP2 showed 75.0% each, and Bowtie showed 82.4%, though these figures may not be precise due to the relatively small number of SNPs predicted at this read depth. The greater SNP prediction accuracy at higher read depth is similar to observations in other species such as *Eucalyptus grandis* (83%) (Novaes et al., 2008), maize (85%) (Barbazuk et al., 2007), soybean (79–92%) (Hyten et al., 2010a), and common bean (86%) (Hyten et al., 2010b). However, it is important to note that our results are mainly based on short sequence reads and low coverage data as compared to data sets in the other studies mentioned.

Table 2. Summary of predictions, wet laboratory experiments, and validation of SNPs classified as per tool or combination of tools.

		Prediction		Wet laboratory experiments			Validation			
Class	Tools or combination of tools	No. of unique SNPs	No. of genes for unique SNPs	No. of SNPs targeted in allele sequencing	No. of SNPs targeted in KASPar assays	Total	True positive SNPs in allele sequencing	True positive SNPs in KASPar assay	Total true positives	% of True positives
I	Bowtie	2454	2135	110	25	135	36	14	50	37.0
II	Maq	1329	1177	90	18	108	40	14	54	50.0
III	Novoalign	1272	1132	78	8	86	21	5	26	30.2
IV	SOAP2	1960	1725	101	28	129	36	15	51	39.5
V	Bowtie, Maq	735	674	54	16	70	27	12	39	55.7
VI	Bowtie, Novoalign	403	385	38	6	44	13	4	17	38.6
VII	Bowtie, SOAP2	1103	1010	65	21	86	22	12	34	39.5
VIII	Maq, Novoalign	294	278	32	6	38	16	5	21	55.3
IX	Maq, SOAP2	622	568	49	16	65	22	12	34	52.3
X	Novoalign, SOAP2	381	363	35	7	42	12	4	16	38.1
XI	Bowtie, Maq, Novoalign	223	212	21	5	26	12	4	16	61.5
XII	Bowtie, Maq, SOAP2	519	480	35	16	51	17	12	29	56.9
XIII	Bowtie, Novoalign, SOAP2	297	284	21	6	27	7	4	11	40.7
XIV	Maq, Novoalign, SOAP2	202	193	18	5	23	9	4	13	56.5
XV	Bowtie, Maq, Novoalign, SOAP2	175	166	11	5	16	6	4	10	62.5

In summary, this study provides a critical appraisal of four commonly used short read alignment tools, along with the read-depth criteria for the identification of SNPs in a crop species without a reference genome sequence. Our results, though based on 36-bp sequence reads generated by Illumina GA I, are expected to be applicable with higher confidence to the data sets with longer sequence reads (than 36 bp) being generated by modern sequencing machines such as GA II, HiSeq. 1000, HiSeq 2000, 454/Titanium, and Ion Torrent. We have predicted a large number of SNPs using the CbCC approach, and these will be useful for molecular genetics and breeding for chickpea improvement. Of the four tools used for SNP discovery (Bowtie, Maq, Novoalign, SOAP2), Maq was found to be most accurate and sensitive, at even low read depth. All four tools demonstrated greater accuracy at higher read depth. Furthermore, SNPs predicted by three or four tools were more likely to be correct.

Table 3. Read depth wise categorization of predicted and validated SNPs

Tools	Read depth category	SNPs predicted	% of predicted SNPs	SNPs validated	True positive (%)
Bowtie	2	1088	44.3	52	23.1
	3	826	33.7	29	20.7
	4–10	409	16.7	42	47.6
	>10	131	5.3	17	82.3
	Total	2454			
Maq	2	531	40.0	30	26.7
	3	448	33.7	31	51.6
	4–10	230	17.3	29	58.6
	>10	120	9.0	20	75.0
	Total	1329			
Novoalign	2	551	44.3	33	21.2
	3	427	33.57	27	25.9
	4–10	224	17.6	18	44.4
	>10	70	5.5	10	60.0
	Total	1272			
SOAP2	2	852	43.5	45	20.0
	3	661	33.7	31	38.7
	4–10	334	17.0	36	47.2
	>10	113	5.8	20	75.0
	Total	1960			

In terms of the future implications of this study in the crop species without a reference genome, SNP discovery using the optimized tools and approaches and their genotyping using GoldenGate or KASPar assays is expected to be in greater use for accelerating genetics research and breeding applications. In the long term, due to decreasing costs in NGS technologies, SNP discovery and imputation of allele, in case of missing data, across the germplasm collection or mapping populations, popularly called genotyping-by-sequencing (GBS), is going to be the approach of future. Recommendations made in this study will be very useful while analyzing GBS data especially in the crop species without a gold-standard reference genome.

LITERATURE CITED

Ahn, S. M., T. H. Kim, S. Lee, D. Kim, H. Ghang, D. Kim, B. C. Kim, et al. 2009. The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. *Genome Research* 19: 1622–1629.

Barbazuk, W. B., S. J. Emrich, H. D. Chen, L. Li, and P. S. Schnable. 2007. SNP discovery *via* 454 transcriptome sequencing. *Plant Journal* 51: 910–918.

Barker, G., J. Batley, H. O'Sullivan, K. J. Edwards, and D. Edwards. 2003. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics* 19: 421–422.

Batzoglou, S. 2005. The many faces of sequence alignment. *Briefings in Bioinformatics* 6: 6–22.

Bentley, D. R. 2006. Whole-genome re-sequencing. *Current Opinion in Genetics & Development* 16: 545–552.

Costa, V., C. Angelini, I. deFeis, and A. Ciccodicola. 2010. Uncovering the complexity of transcriptomes with RNA-Seq. *Journal of Biomedicine & Biotechnology* 2010: article ID 853916, 1–19.

Deschamps, S., M. L. Rota, J. P. Ratashak, and P. Biddle. 2010. Rapid genome-wide single nucleotide polymorphism discovery in soybean and rice via deep resequencingof reduced representation libraries with the Illumina genome analyzer. *Plant Genome* 3: 53–68.

Dubey, A., A. Farmer, J. Schlueter, S. B. Cannon, B. Abernathy, R. Tuteja, J. Woodward, et al. 2011. Defining the transcriptome assembly and its use for genome dynamics and transcriptome profiling studies in pigeonpea (*Cajanus cajan* L.). *DNA Research* 18: 153–164.

Duran, C., N. Appleby, T. Clark, D. Wood, M. Imelfort, J. Batley, and D. Edwards. 2009c. *AutoSNPdb*: An annotated single nucleotide polymorphism database for crop plants. *Nucleic Acids Research* 37 (supplement 1): D951–D953.

Duran, C., N. Appleby, D. Edwards, and J. Batley. 2009a. Molecular genetic markers: Discovery, applications, data storage and visualisation. *Current Bioinformatics* 4: 16–27.

Duran, C., N. Appleby, M. Vardy, M. Imelfort, D. Edwards, and J. Batley. 2009b. Single nucleotide polymorphism discovery in barley using AutoSNPdb. *Plant Biotechnology Journal* 7: 326–333.

Duran, C., D. Eales, D. Marshall, M. Imelfort, J. Stiller, P. J. Berkman, T. Clark, et al. 2010. Future tools for association mapping in crop plants. *Genome* 53: 1017–1023.

Edwards, D., and J. Batley. 2010. Plant genome sequencing: Applications for crop improvement. *Plant Biotechnology Journal* 8: 2–9.

Flicek, P., and E. Birney. 2009. Sense from sequence reads: Methods for alignment and assembly. *Nature Methods* 6: S6–S12.

Garg, R., R. K. Patel, A. K. Tyagi, and M. Jain. 2011. *De novo a*ssembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Research* 18: 53–63.

Gujaria, N., A. Kumar, P. Dauthal, A. Dubey, P. Hiremath, A. BhanuPrakash, A. Farmer, et al. 2011. Development and use of genic molecular markers (GMMs) for construction of a transcript map of chickpea (*Cicer arietinum* L.). *Theoretical and Applied Genetics* 122: 1577–1589.

Hercus, C. 2009. Novocraft short read alignment package. Website http://www.novocraft.com/.

Hillier, L. W., G. T. Marth, A. R. Quinlan, D. Dooling, G. Fewell, D. Barnett, P. Fox, et al. 2008. Whole-genome sequencing and variant discovery in *C. elegans. Nature Methods* 5: 183–188.

Hiremath, P. J., A. Farmer, S. B. Cannon, J. Woodward, H. Kudapa, R. Tuteja, A. Kumar, et al. 2011. Large-scale transcriptome analysis in chickpea (*Cicer arietinum* L.), an orphan legume crop of the semi-arid tropics of Asia and Africa. *Plant Biotechnology Journal* 9: 922–931.

Hudson, M. 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources* 8: 3–17.

Hyten, D. L., S. B. Cannon, Q. Song, and N. Weeks. 2010a. High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics* 11: 38.

Hyten, D. L., Q. Song, E. W. Fickus, and C. V. Quigley. 2010b. High-throughput SNP discovery and assay development in common bean. *BMC Genomics* 11: 475.

Imelfort, M., C. Duran, J. Batley, and D. Edwards. 2009. Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnology Journal* 7: 312–317.

Imelfort, M., and D. Edwards. 2009. *De novo* sequencing of plant genomes using second-generation technologies. *Briefings in Bioinformatics* 10: 609–618.

Jayashree, B., A. Bhanuprakash, A. Jami, P. S. Reddy, S. Nayak, and R. K. Varshney. 2009. Perl module and PISE wrappers for the integrated analysis of sequence data and SNP features. *BMC Research Notes* 2: 92.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10: R25.

Li, B., V. Ruotti, R. Stewart, J. Thomson, and C. Dewey. 2010. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26: 493–500.

Li, H., and H. Durbin. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25: 1754–1760.

Li, H., and N. Homer. 2010. A survey of sequence alignment algorithms for next generation sequencing. *Briefings in Bioinformatics* 11: 473–483.

Li, H., J. Ruan, and R. Durbin. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* 18: 1851–1858.

Li, R., C. Yu, Y. Li, T. W. Lam, S. M. Yiu, K. Kristiansen, and J. Wang. 2009. SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* 25: 1966–1967.

Lu, T., G. Lu, D. Fan, C. Zhu, W. Li, Q. Zhao, Q. Feng, et al. 2010. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Research* 20: 1238–1249.

MacLean, D., J. D. G. Jones, and D. J. Studholme. 2009. Application of 'next-generation' sequencing technologies to microbial genetics. *Nature Reviews Microbiology* 7: 287–296.

Mardis, E. R. 2008. The impact of next-generation sequencing technology on genetics. *Trends in Genetics* 24: 133–141.

Marioni, J., C. E. Mason, S. M. Mane, M. Stephens, and Y. Giladi. 2008. RNA-Seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* 18: 1509–1517.

Marshall, D. J., A. Hayward, D. Eales, M. Imelfort, J. Stiller, P. J. Berkman, T. Clark, et al. 2010. Targeted identification of genomic regions using TAGdb. *Plant Methods* 6: 19.

Marth, G. T., I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu, H. Zakeri, N. O. Stitziel, et al. 1999. A general approach to single-nucleotide polymorphism discovery. *Nature Genetics* 23: 452–456.

Metzker, M. L. 2010. Sequencing technologies—The next generation. *Nature Reviews Genetics* 11: 31–46.

Miller, N. A., S. F. Kingsmore, and A. D. Farmer. 2008. Management of high-throughput DNA sequencing projects: Alpheus. *Journal of Computer Science & Systems Biology* 1: 132.

Morin, R., M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. Pugh, H. McDonald, et al. 2008. Profiling the *HeLa S3* transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques* 45: 81–94.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 5: 621–628.

Nayak, S. N., H. Zhu, N. Varghese, S. Datta, H. K. Choi, R. Horres, R. Jüngling, et al. 2010. Integration of novel SSR and gene-based SNP marker loci in the chickpea genetic map and establishment of new anchor points with *Medicago truncatula* genome. *Theoretical and Applied Genetics* 120: 1415–1441.

Novaes, E., D. R. Drost, W. G. Farmerie, G. J. Jr. Pappas, D. Grattapaglia, R. R. Sederoff, and M. Kirst. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 9: 312.

Parchman, T. L., K. S. Geist, J. A. Grahnen, C. W. Benkman, and C. A. Buerkle. 2010. Transcriptome sequencing in an ecologically important tree species: Assembly, annotation, and marker discovery. *BMC Genomics* 11: 180.

Pasaniuc, B., N. Zaitlen, and E. Halperin. 2010. Accurate estimation of expression levels of homologous genes in RNA-seq experiments. *In* Proceedings of 14th Annual International Conference on Research in Computational Biology (RECOMB 2010), Lisbon, Portugal, 397–407. International Computer Science Institute, Berkeley, California, USA.

Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology* 132: 365–386.

Smith, A. D., Z. Xuan, and M. Q. Zhang. 2008. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* 9: 128.

Souche, E. L., B. Hellemans, H. J. K. J. Van, and A. Canario. 2007. Mining for single nucleotide polymorphisms in expressed sequence tags of European sea bass. *Journal of Integrative Bioinformatics* 4: 73.

Stephens, M., J. S. Sloan, P. D. Robertson, P. Scheet, and D. A. Nickerson. 2006. Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nature Genetics* 38: 375–381.

Sultan, M., M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321: 956–960.

Trick, M., Y. Long, J. Meng, and I. Bancroft. 2009. Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnology Journal* 7: 334–346.

Upadhyaya, H. D., M. Thudi, N. Dronavalli, N. Gujaria, S. Singh, S. Sharma, and R. K. Varshney. 2011. Genomic tools and germplasm diversity for chickpea improvement. *Plant Genetic Resources* 9: 45–58.

Varshney, R. K., W. Chen, Y. Li, A. K. Bharti, R. K. Saxena, J. A. Schlueter, M. A. Donoghue, et al. 2011. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nature Biotechnology* 30: 83–89.

Varshney, R. K., S. N. Nayak, G. D. May, and S. A. Jackson. 2009. Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology* 27: 522–530.

Wang, J., W. Wang, R. Li, Y. Li, G. Tian, L. Goodman, W. Fan, et al. 2008. The diploid genome sequence of an Asian individual. *Nature* 456: 60–65.

Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics* 10: 57–63.

Wheeler, D. A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452: 872–876.

Xu, G., N. Deng, Z. Zhao, T. Judeh, E. Flemington, and D. Zhu. 2011. SAMMate: A GUI tool for processing short read alignments in SAM/BAM format. *Source Code for Biology and Medicine* 6: 2.

Yu, X. 2011. How well do alignment programs perform on sequencing data with varying qualities and obtained from repeat regions? *In* Meeting of Sequencing Data Interest Group (SDIG), Cleveland, Ohio, USA; available as pdf at website http://cancer.case.edu/workinggroups/SDIG/index.html; cancer.case.edu/workinggroups/SDIG/2011/yu2-28-11.pdf.

Zhang, J., D. A. Wheeler, I. Yakub, S. Wei, R. Sood, W. Rowe, P. P. Liu, et al. 2005. SNPdetector: A software tool for sensitive and accurate SNP detection. *PLoS Computational Biology* 1: e53.