Atlas Journal of Biology 3 (1): 206–211, 2014
doi: 10.5147/ajb.2014.0134

SNP-E: A New Method For Multiple Sequence Alignments Analysis And Accurate Single Nucleotide Polymorphism Evaluation

Melody N. Hemmati-Sholeh[1], Larry A. Sholeh[2], and David A. Lightfoot[1,*]

[1] Genomics Core Facility; Department of Plant Soil and Agricultural Systems, and the Illinois Soybean Center, Southern Illinois University at Carbondale, Carbondale, IL 62901, USA; [2] Department of Electrical and Computer Engineering, Southern Illinois University at Carbondale, Carbondale, IL 62901, USA

Received: June 19, 2014 / Accepted: September 5, 2014

Abstract

Identification of single nucleotide polymorphisms (SNPs) and insertion-deletion mutations are important for discovering the connection between the genetic mutations and complex diseases. The objective of this study was to develop a sensitive and accurate computational method for SNP detection among Multiple Sequence Alignments (MSAs) to be run on Microsoft Office Suite™ and Windows™. The SNP-Evaluator, was designed to simulate the process of human eye visual change-identification. Analysis of three 82-Kbp genomic loci derived from Sanger sequencing and the corresponding SNPs from 31 genomes from Illumina™ sequencing of soybean (*Glycine max* L. Merr.) demonstrated that the SNP-E was an effective method for medium-scale genomic research.

Keywords: Single Nucleotide Polymorphism (SNP); Sanger; Illumina™; Cultivar; Variation.

* Corresponding author: ga4082@siu.edu

Introuction

Identification of single nucleotide polymorphisms (SNPs) in multiple sequence alignments (MSA) involves looking across MSA and identifying base discrepancies (Wegrzyn et al., 2009). SNPs and mutations among highly polymorphic regions are often associated with useful traits such as resistance to disease (Ruben et al., 2006; Srour et al., 2012) and identification of them is important to discover the link between the genetic mutations and complex diseases (Zhang et al., 2005). The genomes of crop plants, animals and humans contain regions of diversity much less than 98% identity among individuals interspersed in more conserved region. One such region is the *Rhg1/Rfs2* locus of soybean (*Glycine max* L. Merr.) which appeared to span over 150kbp and encompassed more than 20 genes. Various methods have been developed (Chang, 2009) and already available for SNP detection and calling single-nucleotide polymorphisms from next generation sequencing data (DePristo et al., 2011; Kobold et al., 2009; Li et al., 2009; Nijveen et al., 2013) however, most of these methods require expensive high-depth sequencing to perform satisfactory (Xu et al., 2012) or they are Java-based programs such as Seq-SNPing (Chang et al., 2009) or they require Linux command line skills to run and a separate program to visualize the results (Nijveen et al., 2013) or requires Phred commands in case of PineSAP. So, there was a need for a method to efficiently and accurately call, analyze and identify SNPs in MSAs for medium scale loci, ~ 100 Kbp. The method needed to be run on a popular and user friendly system such as Microsoft Office Suite™ to prevent the need for complex and costly operating system such as UNIX. SNP-Evaluator was created and demonstrated that can accurately align relatively diverse sequences and allow researchers to identify

SNPs of interest for further analyses. SNP-E, has the ability to search among MSA and recognize base polymorphisms. It is capable of making a confident identification (base-call) because visual confirmation is not a reliable option for sequences from next generation sequencer with higher number of sequences at the cost of higher error rates and patchy sequence coverage. Also, this method needed to give the young researchers who are still enhancing their skills in different programming languages and would like to work with medium-scale genomic sequences, more flexibility for next generation data analysis. Furthermore, by applying this method, researchers have the ability to associate a specific alignment nucleotide identification number to each nucleotide (nucleotide's ID), customize data analysis and apply functions and formulas on data (nucleotides). Moreover, this method functions as a visible tool due to its table format that can provide SNP-E the capability to simulate the process of human eye visual change-identification.

Materials and Methods

The genomic DNA region used here was the 82Kbp isolated and embedded in a BAC, B73P06, which encompassed 10 genes and one highly polymorphic region of about 59 Kbp (743 SNPs from 1,500 bp to 60,500 bp). A multiple sequence alignment between *Rhg1/Rfs2* locus on chromosome 18, of three sequences from soybean cultivars; 'Forrest' BAC B73P06 (Hemmati and Lightfoot, 2011), Asgrow 3244 and Williams 82 (Srour et al., 2012) was performed through NCBI-MSA. The out-put of NCBI-MSA in FASTA format by flat query anchored with dots for identities was used as raw data.

Excel™ applications was used as a sufficient interface to communicate data, SNPs, and SNPs-detection.

SNP-E Design Phases

Designing SNP-E consisted of 3 major phases. First phase was data migration of MSA-FASTA-Format to Column-Alignment-Format (CAF), second phase was data conversion of Column-Alignment-Format (CAF) to Vertical-Alignment-Format (VAF) and the third phase was Single Nucleotide Polymorphism Evaluation, SNP-E.

Data Migration of MSA-FASTA-Format to Column Alignment Format (CAF)

To convert this raw data to excel applicable format, data migration of FASTA format sequence of multiple sequence alignments from a horizontal alignment text format to excel vertical alignment format was needed. A cascade conversion process was applied to convert entire MSA-FASTA text format to Excel format, column. Text to column function was applied to entire MSA-FASTA. Up to this stage a 3×4 of MSA-Excel format had been created. Each NCBI-MSA-FASTA converted to MSA-Excel format consisted of column A indicating sequence ID, column B indicating sequence starting numbers, column C containing 60 nucleotides of each sequence from each row with no assigned identification number to each nucleotide and finally, column D indicating sequence ending numbers (Figure 1).

Data Conversion of Column Alignment Format (CAF) to Vertical Alignment Format (VAF)

The greatest challenge at this stage was extraction of each sequence form the multiple sequence alignments while conserving alignments in excel format, VAF. In order to overcome this challenge, data filtering was applied to call for already-aligned-sequences; Forrest, Asgrow and Williams 82 (W82) by calling their sequence ID, Query (Forrest), 31218 (Asgrow 3244) and 31219 (W82) accordingly. The purpose of filtering was to isolate each sequence from the FASTA format alignment. Then, each isolated sequence (Asgrow 3244, W82 and Forrest) was transferred to EmEditor to be converted to a single vertical sequence in a column by applying a separation method such as line break or End-Of-Line (EOL). Next, each Vertical Isolated Sequence (VIS) was imported from EmEditor to excel in a single column. And the final outcome of this process was a Vertical Alignment Format (VAF).

Sequence Numbering Method (SNM) and Challenges

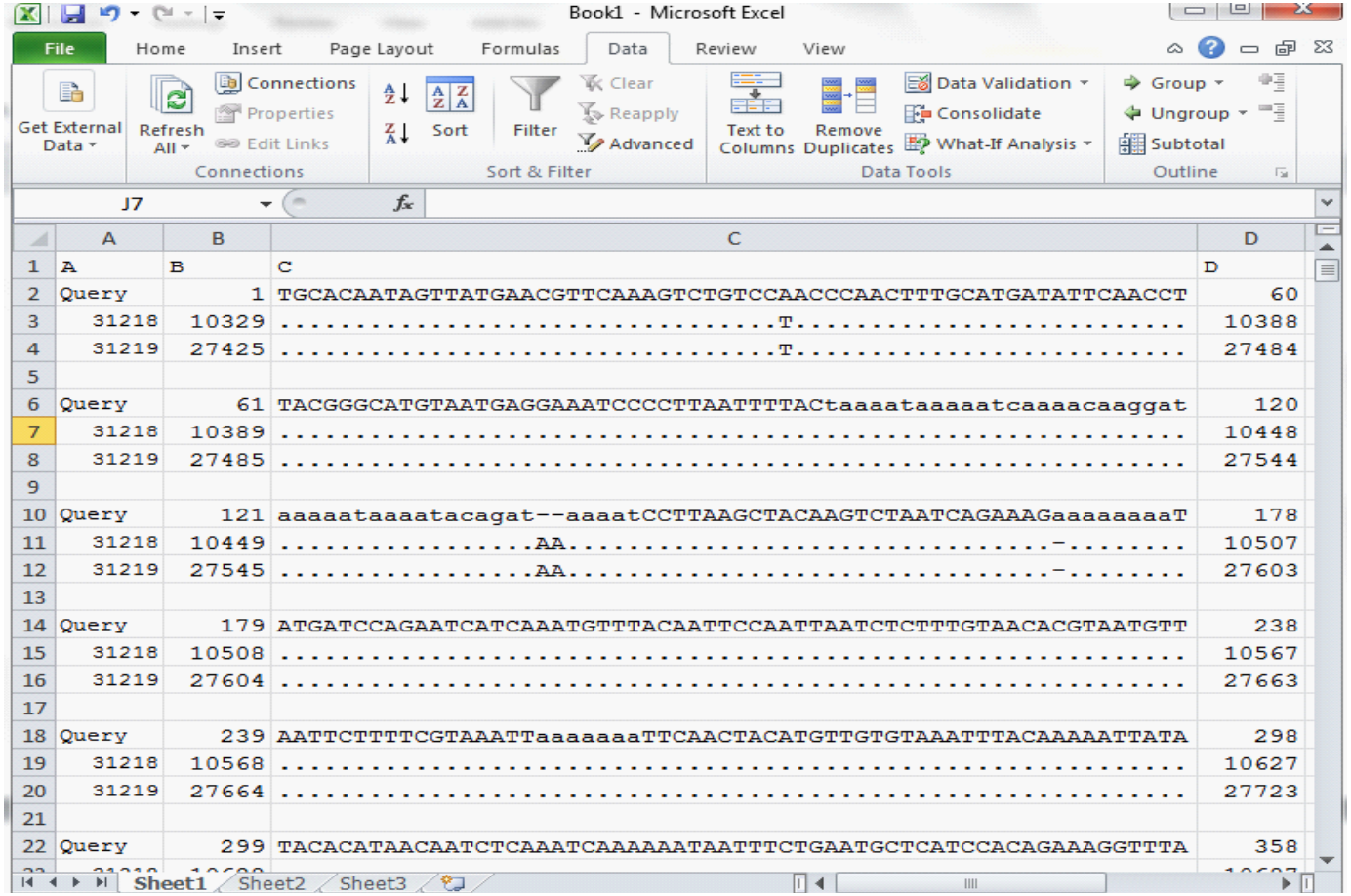
The purpose of this step was to assign an ascending number (ID) to each nucleotide from each sequence located at separate columns in excel. Gaps were regions where the greatest challenge in SNM occurred. To synchronize SNM with Sequence Alignment Gaps SAG, a combination of logic functions i.e. "IF", were applied, in order to stop counting nucleotide numbering at gap regions and release counting at nucleotides regions.

Single Nucleotide Polymorphism Evaluation, SNP-E

In order to identify SNPs among aligned sequences, a combination of logic functions i.e. "IF" and "OR" were applied. All SNPs were flagged in a separate column as the result of logic function, "0" for "FALSE" and "1" for "TRUE". Filtering on TRUEs revealed SNPs for further analysis. The Algorithm Flowchart of SNP-E was pictured in Figure 2. The step by step guideline is also accessible through http://pbgc.siu.edu/docs/illustration.docx link.

Integration of 31 Genomes SNPs

The available SNP motifs of 31 genomes by Lam et al. (2010) which identified over 400 more potential SNP positions in this region (Lam et al., 2010) was compared to the results of SNP-E from three analyzed sequences mentioned above (MASTER-SNPs). The comparison demonstrated that the SNP motif s could be added by reference to the bp position of Williams 82. For those regions that misalignments had occurred, they were clear from the SNP motifs disagreement and could be logically nudged to the nearest likely correct positions. Once aligned the SNPs were de-convoluted to single columns each corresponding to a single genotype (Hauge et al., 2006; Lam et al., 2010).



	A	B	C	D
1	A	B	C	D
2	Query	1	TGCACAATAGTTATGAACGTTCAAAGTCTGTCCAACCCAACTTTGCATGATATTCAACCT	60
3	31218	10329	.....T.....	10388
4	31219	27425	.....T.....	27484
5				
6	Query	61	TACGGGCATGTAATGAGGAAATCCCCTTAATTTTACtaaaataaaaatcaaaacaaggat	120
7	31218	10389	.....	10448
8	31219	27485	.....	27544
9				
10	Query	121	aaaaataaaatacagat--aaaatCCTTAAGCTACAAGTCTAATCAGAAAGaaaaaaaaT	178
11	31218	10449	.....AA.....-	10507
12	31219	27545	.....AA.....-	27603
13				
14	Query	179	ATGATCCAGAATCATCAAATGTTTACAATTCCAATTAATCTCTTTGTAACACGTAATGTT	238
15	31218	10508	.....	10567
16	31219	27604	.....	27663
17				
18	Query	239	AATTCTTTTCGTAAATTaaaaaaaTTCAACTACATGTTGTGTAAATTTACAAAAATTATA	298
19	31218	10568	.....	10627
20	31219	27664	.....	27723
21				
22	Query	299	TACACATAACAATCTCAAATCAAAAAATAATTTCTGAATGCTCATCCACAGAAAGGTTTA	358

Figure 1. The first step of converting MSA-FASTA text fromat to excel format.

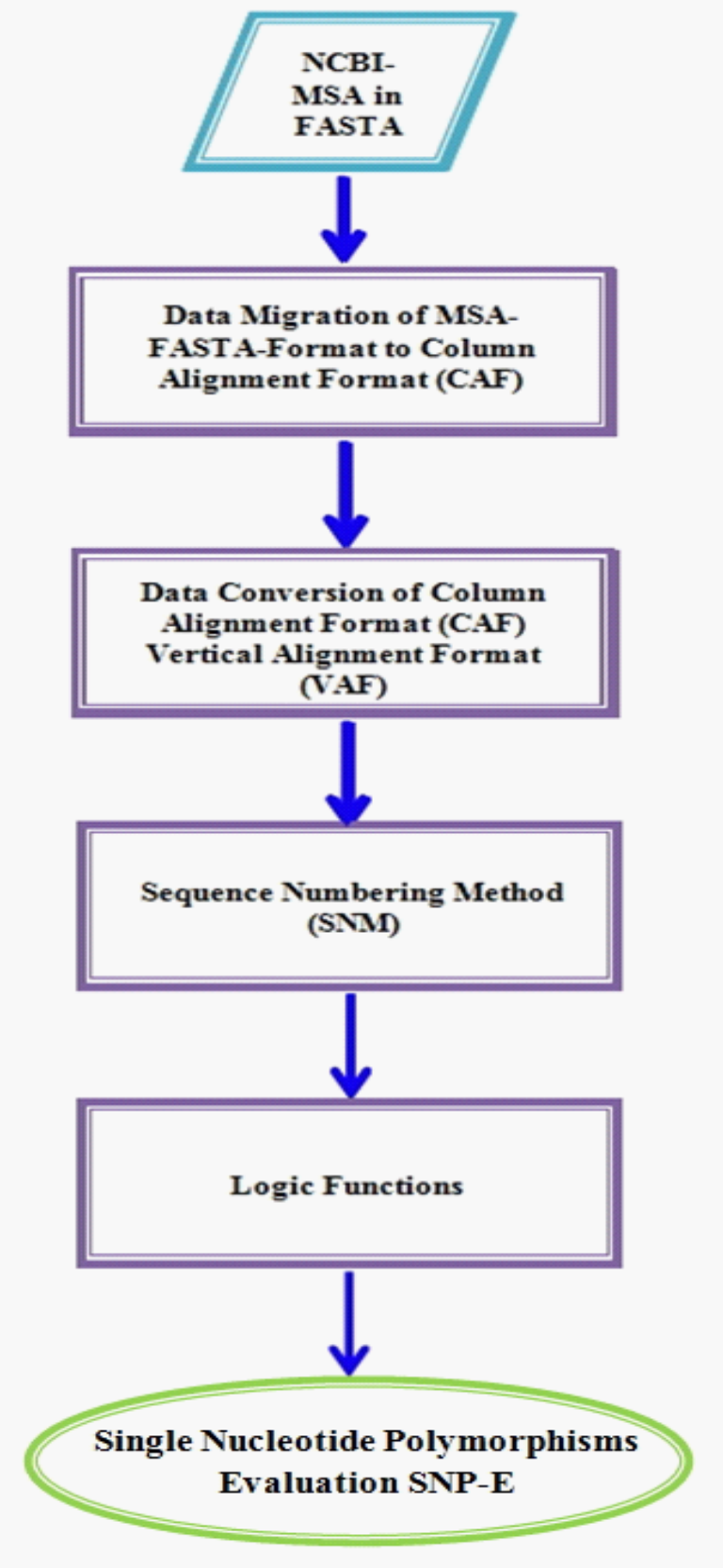


Figure 2. SNP-E algorithm flow chart.

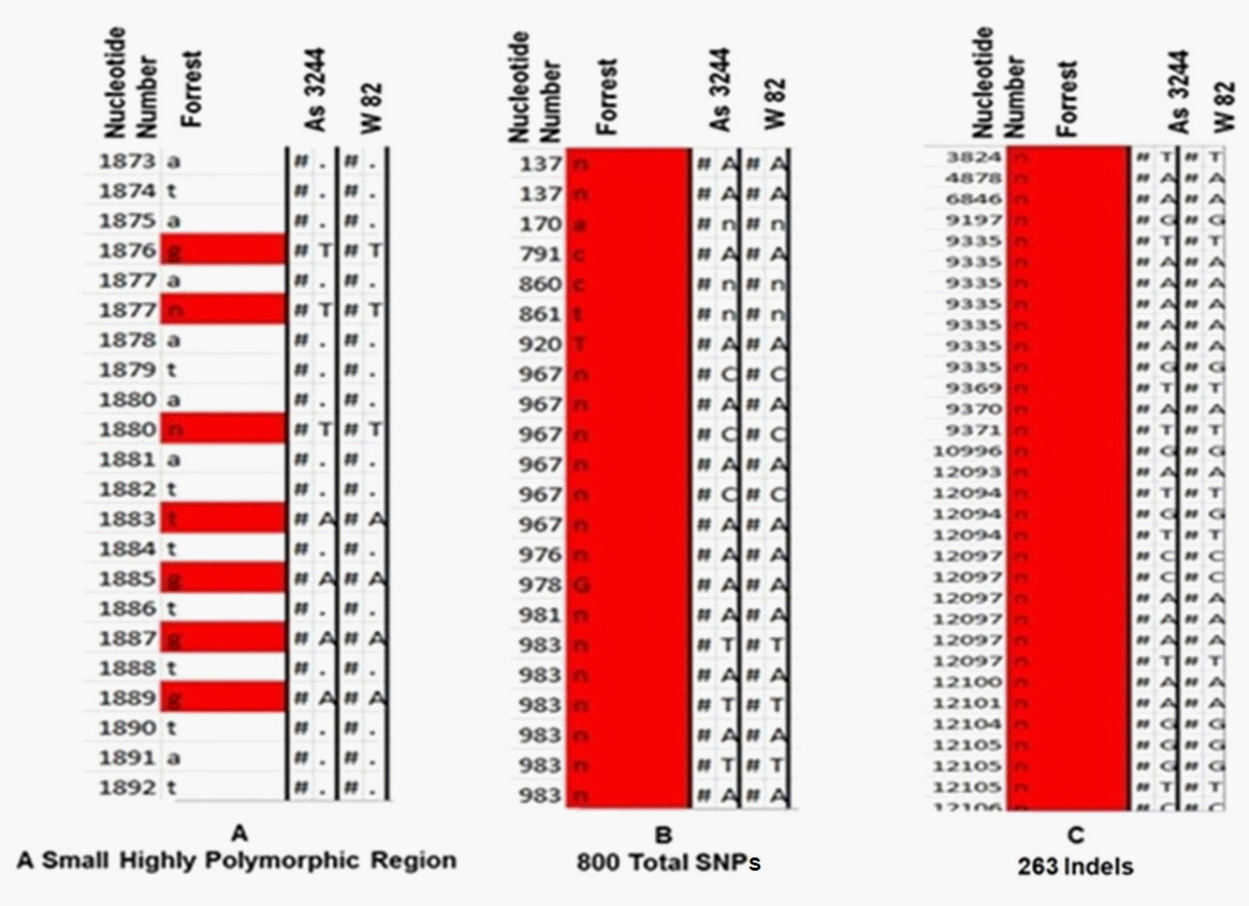


Figure 3. Examples of alignments in SNP-E. Panel A presents a small highly polymorphic region of the automated final alignment of Forrest, Asgrow 3244 and Williams 82 genomic sequences. Panel B presents a view of a typical region among the total number of SNPs. Panel C presents a view of a typical region among the total number of Indels.

Results and Discussion

SNP-E Results

Automated SNP-E recorded 800 SNPs from polymorphic region (Figure 3-panel A) of the alignment among Forrest, Asgrow 3244 and Williams 82 genomic sequences (Figure 3-panel B). Of those, 263 were total nucleotide insertions. No, manual annotation was needed for SNP evaluation at any stage of process. The results have been stored in SNP-E-MASTER-Excel-File.

Results of Integration of 31 Genomes SNPs

The results indicated that the Master-SNPs file (SNP-E-MASTER-Excel-File) was homogenous ascending sequentially with the 31-SNPs-Genome-consensus alignments. The MSA between SNPs results from SNP-E-MASTER-Excel-File, and 31-column-SNPs-Sequence resulted in one aligned-MASTER-SNPs and 31-column-SNPs-consensus-sequences. The final results showed that 78% SNPs aligned automatically for each of the 31-SNPs-Genomes and no manual annotation was involved. Therefore, Sanger sequence of the BACs and Next Gen sequence of genome inferred that there were more than 1,400 SNPs in the 80 kbp region introgressed from Peking. 27.97 % of the indels aligned automatically for each of the 31-SNPs-Genomes and no manual annotation was needed. Sanger sequence has more strength in detecting indels than Nextgen sequencing, so this was to be expected (Chang et al., 2009; DePristo et al., 2011; Kobold et al., 2009; Li et al., 2009; Nijveen et al., 2013). The methods to compare sequences have to account for the sequence quality of the different available methods.

Abbreviations

SNP-E, Single Nucleotide Polymorphism Evaluation
MSAs, Multiple Sequence Alignments
NCBI, National Center for Biotechnology Information
CAF, Column Alignment Format
VAF, Vertical Alignment Format
EOL, End Of Line
VIS, Vertical Isolated Sequence
SNM, Sequence Numbering Method
SAG, Sequence Alignment Gap

Acknowledgements

The physical map location of B73P06 was supported by the NSF under Grant No. 9872635 and 0487654. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

Chang H, L Chuang, Y Cheng, C Ho, C Wen, C Yang (2009) Seq-SNPing: multiple-alignment tool for SNP discovery, SNP ID identification, and RFLP genotyping. Omics 13:253-260.

DePristo M, E Banks, KV Garimella, J Maguire, C Hartl, A Philippakis, G del Angel, M Rivas, Mea Hanna (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43:491–498.

Hauge B, M Wang, J Parsons, L Parnell (2006) Methods of introgressing nucleic acid molecules associated with soybean cyst nematode resistance into soybean. US Patent 7:154,021.

Hemmati M, D Lightfoot (2011) *Glycine max* cultivar Forrest clone BAC 73P06 genomic sequence. GenBank: HQ0089381.

Kobold tD, K Chen, T Wylie, D Larson, MD McLellan, E Mardis, G Weinstock, R Wilson, L Ding (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25:2283–2285.

Lam H, X Xu, X Liu, W Chen, G Yang, F Wong, M Li, W He, N Qin, Bea Wang (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nature genetics 42:1053-1059.

Li H, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078 – 2079.

Nijveen H, M van Kaauwen, D Esselink, B Hoegen, B Vosman (2013) QualitySNPng: a user-friendly SNP detection and visualization tool. Nucleic Acids Research Advance Access:1-4.

Ruben E, A Jamai, J Afzal, V Njiti, K Triwitayakorn, M Iqbal, S Yaegashi, R Bashir, S Kazi, P Arelli , C Town, H Ishihara, K Meksem, D Lightfoot (2006) Genomic analysis of the *Rhg1/Rfs2* locus: candidate genes that underlie soybean resistance to the cyst nematode. Mol Genet Genom 276:503-516.

Srour A, A Afzal, N Saini, L Blahut-Beatty, N Hemmati, D Simmonds, H El Shemy, C Town, H Sharma, D Lightfoot (2012) The receptor like kinase transgene from the *Rhg1/Rfs2* locus caused pleiotropic resistances to soybean cyst nematode and sudden death syndrome. BMC Genomics 13:368.

Wegrzyn J, J Lee, J Liechty, D Neale (2009) Sequence analysis, PineSAP—sequence alignment and SNP identification pipeline. Bioinformatics Application Note 25:2609-2610.

Xu F, W Wang, P Wang, M Jun Li, P Chung Sham, J Wang (2012) A fast and accurate SNP detection algorithm for next-generation sequencing data. Nature Communications 3:1258.

Zhang J, D Wheeler, I Yakub, S Wei, R Sood, W Rowe, P Liu, R Gibbs, K Buetow (2005) SNPdetector: A Software Tool for Sensitive and Accurate SNP Detection. PLoS Computational Biology 1:395-404.

Atlas Journal of Biology - ISSN 2158-9151. Published By Atlas Publishing, LP (www.atlas-publishing.org)