Journal of
Proteomics & Bioinformatics

Okii et al., J Proteomics Bioinform 2014, 7:1
DOI: 10.4172/jpb.1000296

Research Article Open Access

Application of Bioinformatics in Crop Improvement: Annotating the Putative Soybean Rust resistance gene *Rpp3* for Enhancing Marker Assisted Selection

Okii D[1,3]*, Chilagane LA[2], Tukamuhabwa P[1] and Maphosa M[1]

[1]*Department of Agricultural Production, Makerere University, P.O. Box 7062, Kampala, Uganda*
[2]*Department of Crop Science and Production, Sokoine University of Agriculture, P.O. Box 3005, Morogoro, Tanzania*
[3]*National Crops Resources Research Institute, Namulonge, P.O. Box 7084, Kampala, Uganda*

Abstract

Despite the wide availability of DNA sequence information freely available online, the challenge is to convert this mass of data into knowledge that can be readily applied in crop improvement programs. The main objective of this study was to annotate the *Rpp3* locus in soybean for enhancing the crop's marker assisted selection (MAS). The specific objectives were: (i) to do structural and functional annotation of the *Rpp3* locus genetically mapped on Linkage groups (LG) - C2 and physically located on chromosome 6 and (ii) to generate novel markers linked to the rust resistance for MAS in soybean. The soybean query sequence of interest was downloaded from NCBI (www.ncbi.nlm.nih.gov/nuccore/NW_003722736.1) and subsequently analysed with an array of bioinformatics tools to capture information on the characteristics of the *Rpp3* gene. The study found DNA transposons as the predominant repeats in the soybean genomic region analysed. 16 non-overlapping genes were predicted to be tightly linked to marker Satt460 and code for various functions from BLASTx analyses. Gene 1 and 12, both code for structural and enzymatic roles, while gene 13 suggests storage proteins mobilization in seeds. Genes 6, 7 and 8 codes for transcription activation, while gene 10 is a transcription deactivator. There was homology to model organisms; *Arabidopsis thaliana* (dicots) Chromosome 5 as best hit, with expected-value (E-value) of 3e-128 and 76% sequence identity to *Oryza sativa japonica* Chromosome 2, *Oryza sativa*, with E-value of 2e-21 and 84% sequence identity. 15 short random primer sequences with 18-24 base pairs were designed to amplify the *Rpp3* predicted genes and introns in soybean chromosome 6 though not validated in the study due to economic reasons. Similar studies are recommended on other genes conferring resistance to rust disease for effective gene pyramiding and shortening the soybean breeding cycle.

Keywords: Marker assisted selection; Bioinformatics; Annotation; Satt460; *Rpp3* gene

Introduction

Genomic information available online is key to understanding plant development and associated traits, for crop improvement [1]. The National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/) and soybean breeder's (http://soybeanbreederstoolbox.org/) databases among the many have useful information for enhancing soybean improvement through Marker Assisted Breeding. The wide array of bioinformatics tools freely available online can enable information capture and management from repositories of genomic data in attempts to understanding and modelling living systems [2]. Bioinformatics refers to the new field in biology that merges, computer science and information technology with wide applications such as; genome sequencing [2,3], molecular marker discovery [4,5], transcriptomics [6], candidate gene identification [7] and taxonomy [8]. Despite its wide application, the challenge however, remains to convert this mass of data into knowledge that can be readily applied in crop improvement programs [1]. Although a glimpse of the distribution of genic and repetitive sequences in soybean has been seen [9], a detailed analysis is lacking. There is no report on structural and functional annotation of the *Rpp3* locus in soybean, despite its effectiveness in contributing to rust disease resistance in the crop. Few annotation studies have however recently been undertaken given availability of the whole crop's draft genome sequences deposited in the NCBI database for utilisation. Soybean rust (*Phakopsora pachyrhizi*) is one of the most serious foliar diseases of soybean worldwide (Yang et al. [10]; Sinclair and Hartman [11], Monteros et al. [12]). The rust disease spreads rapidly causing yield losses of up to 80%, making it a very important disease in soybean production [13]. The *Rpp3* gene was chosen in this annotations study due to its greater effectiveness in a pairwise gene pyramiding combination study for resistance to soybean rust populations in Uganda [14]. Various studies, such as, *Rpp3* locus genetic mapping [15]; pair wise gene pyramiding studies involving the *Rpp3 locus* [14]; have provided detailed explanations on the locus in conferring resistance to soybean rust disease worldwide. In addition, Ray et al. [16] reported recessive resistance at or near the R*pp3* locus using phenotypes of 24 F1 plants in Paraguay. There are six major rust genes mapped to different linkage groups (LG) as follows: R*pp1* to linkage group G [17], *Rpp3 to* C2 [15] and http://soybeanbreederstoolbox.org), R*pp2* and R*pp4 to* J and G, respectively [18,19], R*pp Hyuuga to* C2 [12] and R*pp5 to* N [20]. The *Rpp3* locus is tightly linked to marker Satt460 (a Simple Sequence Repeat) mapped to linkage group C2 at 106.991 cM (centimorgans) [15]. Kendrick et

***Corresponding author:** Okii D, National Crops Resources Research Institute, Namulonge, P.O. Box 7084, Kampala, Uganda, Tel: +256782177552; Fax: +256414531641; E-mail: dennis.okii@gmail.com

Received October 23, 2013; **Accepted** January 08, 2014; **Published** January 10, 2014

Citation: Okii D, Chilagane LA, Tukamuhabwa P, Maphosa M (2014) Application of Bioinformatics in Crop Improvement: Annotating the Putative Soybean Rust resistance gene *Rpp3* for Enhancing Marker Assisted Selection. J Proteomics Bioinform 7: 001-009. doi:10.4172/jpb.1000296

al. [21], mapped two genes, *Rpp3* and *Rpp?*(Hyuuga) to chromosome 6 (LG-C2) in cultivar *Hyuuga,* revealing a natural case of gene pyramiding for Asian soybean rust (ASR) resistance and underscored the importance of utilizing multiple isolates of *P. pachyrhizi* when screening for ASR resistance. The identification of genes and molecular markers underlying important agronomic traits enhances breeding processes, and leads to varieties with improved yield and quality, tolerance to unfavourable environmental conditions and resistance to disease [2]. Structutural and functional annotations of the *Rpp3* locus could provide an insight on gene function and discovery of more polymorphic alternative primers to Satt460 for selecting *Rpp3* bearing rust resistant progenies in different soybean populations. The main objective of this study was thus to characterise the *Rpp3* locus for enhancing the crop's marker assisted selection (MAS). The specific objectives were: (i) to do structural and functional annotation of the *Rpp3* locus genetically mapped on Linkage groups (LG) - C2 and physically located on chromosome 6 and (ii) to generate novel markers linked to the rust resistance for MAS in soybean.

Material and Methods

Materials

To do structural and functional annotation of the *Rpp3* locus and to design new primers for marker based selection of soybean rust disease, the sequence of interest was downloaded from NCBI (www.ncbi.nlm.nih.gov/nuccore/NW_003722736.1). The programs or algorithms used in the study (Table 1) were accessed via the internet.

Methods

The identification of repeats and gene predictions in the soybean *Rpp3* genomic region employed at least two different online algorithms or programs (Table 2) to allow for comparison and validation of results. The soybean chromosome 6 genomic scaffold (>gi|353336025) spanning region 43291218 - 44292393 was then analysed as mentioned below. The steps in the method were divided into five broad categories: Identification of repetitive elements and putative genes, annotation of *Rpp3* locus alleles and functional protein predictions, comparative genomics and primer design using the query soybean sequence from the NCBI as follows.

In the first step repetitive elements surrounding the putative soybean *Rpp3* locus genome were identification using two programs as follows;

i) The CENSOR repeat finder

The query soybean FASTA format sequence was uploaded to the CENSOR program [22] available online to find repetitive elements. Data on the alignments, names or type of repeats, and repeat class was specifically collected.

Rpp3 Locus	Description	Reference
Marker Satt460 position (bp)	>gi|353336025:43291318-43291393 Glycine max chromosome 6 genomic scaffold.	http://soybeanbreederstoolbox.org/cmap/cgi-bin/cmap/feature?feature_acc=GmConsensus40_C2_Satt460
Satt460 position in linkage group (LG)C2 incentiMorgans.	GmConsensus40_C2 (106.991cM)	http://soybeanbreederstoolbox.org/cmap/cgibin/cmap/feature?feature_acc=GmConsensus40_C2_Satt460 and Hyten et al. [15]
Chromosome	6	http://soybeanbreederstoolbox.org
Positive DNA strand (bp) download in 5' to 3' direction.	43291218(bp) to 44292393 (bp) ~ 1.01 (cM) (Size : 100bp+Satt460+1.01x10^6 (bp))	www.ncbi.nlm.nih.gov/nuccore/NW_003722736.1

The positions, description and references of the *Rpp3* locus and the linked marker Satt460 in to soybean linkage group (LG) - C2. The query DNA sequence downloaded was used in subsequently analysed in the methods section

Table 1: The putative *Rpp3* locus genomic sequences with fragment size, position, relations to linked Satt460 marker on reference soybean linkage map.

	Steps	Programs or algorithms	Roles	Websites
1	Repeats Identification	Censor	Finds repeats and masks homologous portions.	http://www.girinst.org/censor/index.php
		TREP and BLASTn	Discovery and annotation of repetitive elements.	http://wheat.pw.usda.gov/ggpages/Repeats/blastrepeats3.html
		DNA Subway	Repeat sequences masking and gene predictions.	http://dnasubway.iplantcollaborative.org/
2	Gene Identification and functional annotations	FGENESH	Polish annotations	http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind.
		DNA Subway and Augustus	Gene Prediction, Annotation of Genomic Sequences.	http://dnasubway.iplantcollaborative.org/ andStanke et al. [48].
		BLASTn	Finds ESTs that match proteins.	blast.ncbi.nlm.nih.gov
		BLASTp	Finds possible proteins.	
		TREP and BLASTp	Predicts proteins.	http://wheat.pw.usda.gov/ggpages/Repeats/blastrepeats3.html
3	Primer designs	Primer 3.	Primer design	http://www.simgene.com/Primer3 and Rozen and Skaletsky [24].

Programs used to annotate the putative soybean rust resistance gene *Rpp3* tightly linked to marker Satt460. Some programs were multi-purpose i.e DNA subway and BLAST algorithms used for identification of repeat and putative genes. BLASTn find ESTs (expressed sequence tags)

Table 2: The algorithms used to analyse *Rpp3* gene in Soybean (*Glycine max*) query sequence from DNA base 43291218 – 44292393 in the NCBI database and their roles.

ii) The TREP and BLASTn sequence repeat finders

The query soybean FASTA sequence was uploaded to the TREP tool and principal BLASTn algorithm selected using the Cereal repeat sequences complete set database for comparisons. The DNA repeats in soybean sequence were then searched with default settings and results noted.

The second step involved, predictions of the eukaryotic gene features in the soybean *Rpp3* locus which included; conserved domains, start codon, splicing sites, exons, stop codon, and PolyA using two programs as follows;

i) Prediction of genes using the FGENESH program

The query soybean FASTA sequence with masked repeats from the censor tool was uploaded to FGENESH tool where gene prediction was performed. The search was first run against *Medicago truncatula* (model plant for legumes) and repeated with *Arabidopsis* (dicot plants) for comparisons, against two reference model organisms [23]. Data on the predicted genes and exons in the query soybean Chromosome 6 were collected.

ii) Prediction of genes using the DNA Subway program

The query Soybean sequence in FASTA format was uploaded to the DNA Subway web page and required fields entered appropriately stepwise in the pipeline that was ran as they became available. The DNA subway program was used for both gene prediction and detection of repetitive elements, unlike the other programs that only carried out one function. The Augustus, BLASTn and BLASTx algorithms incorporated in the DNA subway software were used to predict genes in a stepwise pipeline and final results viewed using a Java program.

The third step involved Annotation of genes in the *Rpp3* locus and functional protein predictions. The following eukaryotic gene features were identified in the *Rpp3* locus using the Augustus program; conserved domains, start codon, splicing sites, exons, stop codon and PolyA. For protein predictions, the messenger RNA (mRNA) transcript for each predicted gene was obtained by re-downloading only the exons sequences from the NCBI database and joined to form a transcript (this steps sliced out the introns from the genes). The transcript were separately uploaded to the NCBI database proteins searched using the BLASTx algorithms (searches protein databases using a translated nucleotide query) using the non-redundant protein sequences database (nr) with default settings in the NCBI. The roles and descriptions of proteins was searched from the related the genetic information of accession that gave the best hit.

The fourth step involved comparative genomics approach to establish shared synteny of the *Rpp3* locus. The soybean query chromosome 6 sequence downloaded was compared with three reference organisms' genomes namely: *Arabidopsis thaliana* (for dicots), *Medicago truncatula* (for legumes) and *Oryza sativa* (for monocots) in the NCBI database using BLASTn similarity search tool and later with the Fgenesh tool.

Finally, primer pairs were design for the *Rpp3* locus and associated QTLS as follows. The downloaded query soybean sequence that contained the Satt460 marker repeat (AT(8)TTATT(17)) and Augustus predicted genes from the respective strand (from base 43291218 – 43437779 in the genome) was used to design new primers using Primer 3 program [24]. The primers were designed following reviews of Kamel [25] to fulfil certain criteria such as primer length of typically 18-30 nucleotides, GC bases content, annealing and Melting Temperature (Tm) or Annealing temperature (Ta): for primers in the range 52-58°C which are unique for the target sequence for successful PCR [26].

Results

Repetitive sequences in Soybean *Rpp3* loci

Soybean chromosome 6 genomic DNA segment consisting of *Rpp3* gene is highly repeated with a few to hundreds of time with variable fragment sizes. Nine different types of repetitive sequences were found by the Censor tool (Table 3), predominated by transposable elements such as retrotransposons and DNA transposons. Other non transposable repeats and simple repeats e.g Integrated viruses and interspersed repeats had comparable frequencies. There were 61 simple repeats and only one SAT repeat with the shortest fragment of 68 bp length; this can be used to tag the different alleles at the *Rpp3* locus. However, the Trep tool found only transposable elements (DNA transposons and retrotransposons) as significant repetitive elements at the locus (Figure 1) with the latter repeat type predominant.

Genes predicted in the *Rpp3* locus

The *Rpp3* gene prediction using Augustus tool found 16 genes of variable sizes, exons on both DNA strands. More information on the genes is in Table 4, with their respective sequences positions in the NCBI database. The sizes of the genes (bp) ranged from hundred to tens of thousands. There was no direct relationship between size of the gene and number of exons per gene, for example the largest gene 6 (15009 bp) was comparable to a much smaller gene 12 (3756 bp) with 9 exons. The 16 genes were localised within a specific positions of the query sequence 146561bp (43291218-43437779 bp) close to Satt460 when compared to the whole 1.01×10^6 (bp) sequence downloaded (Figure 2) with over 90% of the query sequences with no genes predicted. The demonstration of the basic eukaryotic gene features annotations of sequences in Gene 1 is shown in Figure 3, with the position of marker Satt460 with an AT insertion in tight linkage, different from the *(AAT)8TAT(AAT)17* in literature review and soybean breeder's (http://soybeanbreederstoolbox.org/) databases.

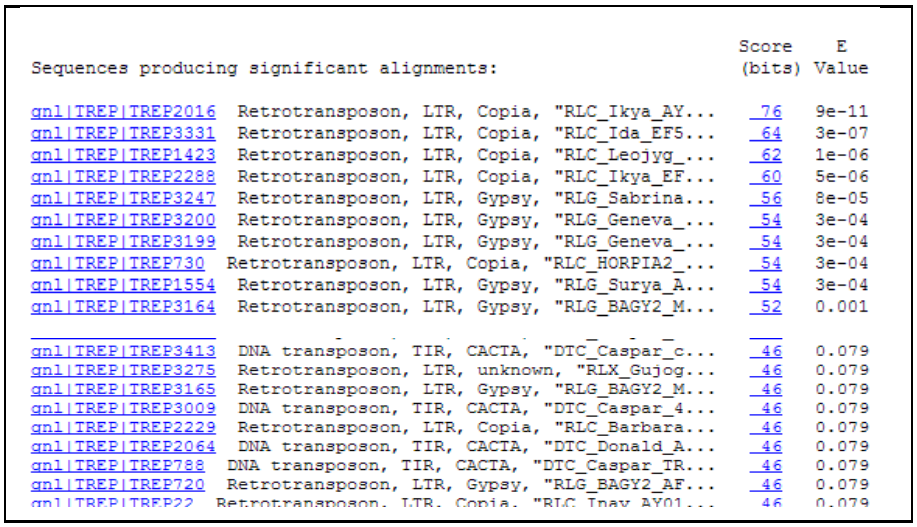
Functional annotations and protein prediction of the *Rpp3* locus

The query soybean transcript BLASTx results found involvement of the predicted protein coding genes in Soybean *Rpp3* locus in

	Repeat Class	No. Fragments	Length (bp)
1	Integrated Virus	4	271
2	Interspersed Repeat	13	2085
3	DNA transposon	598	129608
4	Endogenous Retrovirus	28	1986
5	LTR Retrotransposon	551	311707
6	Non-LTR Retrotransposon	150	21187
7	Repetitive Element	6	426
8	Simple Repeat	61	2174
9	SAT	1	68
	Total	**1412**	**469603**

Repeats classes in the *Rpp3* loci, DNA transposon and LTR retroposons are localised repeats also referred to as transposable elements, while Simple Repeat and SAT are short dispersed repeats in the soybean *Rpp3* genomic region. The other repeats are possibly recombinant DNA from the laboratory vector used in sequence

Table 3: Summary of class and names, number of fragments and size of repetitive DNA sequence in the presumed *Rpp3* gene in soybean chromosome 6 fragment generated in Censor tool.



Sequences producing significant alignments:		Score (bits)	E Value
gnl|TREP|TREP2016	Retrotransposon, LTR, Copia, "RLC_Ikya_AY...	76	9e-11
gnl|TREP|TREP3331	Retrotransposon, LTR, Copia, "RLC_Ida_EF5...	64	3e-07
gnl|TREP|TREP1423	Retrotransposon, LTR, Copia, "RLC_Leojyg_...	62	1e-06
gnl|TREP|TREP2288	Retrotransposon, LTR, Copia, "RLC_Ikya_EF...	60	5e-06
gnl|TREP|TREP3247	Retrotransposon, LTR, Gypsy, "RLG_Sabrina...	56	8e-05
gnl|TREP|TREP3200	Retrotransposon, LTR, Gypsy, "RLG_Geneva_...	54	3e-04
gnl|TREP|TREP3199	Retrotransposon, LTR, Gypsy, "RLG_Geneva_...	54	3e-04
gnl|TREP|TREP730	Retrotransposon, LTR, Copia, "RLC_HORPIA2_...	54	3e-04
gnl|TREP|TREP1554	Retrotransposon, LTR, Gypsy, "RLG_Surya_A...	54	3e-04
gnl|TREP|TREP3164	Retrotransposon, LTR, Gypsy, "RLG_BAGY2_M...	52	0.001
gnl|TREP|TREP3413	DNA transposon, TIR, CACTA, "DTC_Caspar_c...	46	0.079
gnl|TREP|TREP3275	Retrotransposon, LTR, unknown, "RLX_Gujog...	46	0.079
gnl|TREP|TREP3165	Retrotransposon, LTR, Gypsy, "RLG_BAGY2_M...	46	0.079
gnl|TREP|TREP3009	DNA transposon, TIR, CACTA, "DTC_Caspar_4...	46	0.079
gnl|TREP|TREP2229	Retrotransposon, LTR, Copia, "RLC_Barbara...	46	0.079
gnl|TREP|TREP2064	DNA transposon, TIR, CACTA, "DTC_Donald_A...	46	0.079
gnl|TREP|TREP788	DNA transposon, TIR, CACTA, "DTC_Caspar_TR...	46	0.079
gnl|TREP|TREP720	Retrotransposon, LTR, Gypsy, "RLG_BAGY2_AF...	46	0.079
gnl|TREP|TREP22	Retrotransposon, LTR, Copia, "RLC_Tnav_AY01...	46	0.079

Figure 1: Significant transposable elements in *Rpp3* locus revealed with the TREP tool including retrotransposons (LTR - Long terminal repeats), and DNA transposon repeats. LTRs are repetitive DNA sequences with hundreds to thousands of bases in retroviral DNA and in retrotransposons, flanking functional genes. They are used by viruses to insert their genetic sequences into the host genomes.

Gene	Size (bp)	Position of gene in NCBI		Exons per gene	DNA Strand
		Begins	Ends		
1	3975	43291218	43295193	10	+
2	2577	43310092	43312669	6	-
3	1872	43313547	43315419	2	-
4	7250	43321439	43328689	4	+
5	2896	43329323	43332219	6	-
6	15009	43338643	43353652	9	+
7	4559	43356003	43360562	5	+
8	11245	43364403	43375648	5	+
9	4044	43380645	43384689	11	-
10	1718	43388833	43390551	3	+
11	3918	43391783	43395701	5	+
12	3756	43401803	43405559	9	+
13	2623	43411236	43413859	2	-
14	8243	43417616	43425859	7	-
15	1015	43431743	43432758	2	+
16	321	43437458	43437779	1	-

Genes on positive (+) and negative (-) DNA strand are linked to marker Satt460 in cis and trans orientation respectively in the Soybean chromosome 6 genomic sequence. The 16 predicted genes size, corresponding downloadable Nucleotide positions in the NCBI database is shown in figure 2

Table 4: The genes predicted with size, positions from Augustus Algorithm prediction in the soybean query sequence.

many biological processes. The functional annotations were based on similarity sequence comparisons. Gene 1 and 12 had the same roles and both participate in many cellular and structural processes. Genes 1, 6, 7, 8, 10, 11, 12 and 14 have interesting biological implication, with Genes; 1 and 12 related in functions and both code for various biological and structural roles in Soybean. The role of gene 13 suggests storage proteins mobilization in seeds. The functional predictions for the genes were independent of the size of the transcript used. Some of transcripts produced significant hits but, had no functional proteins deduced (e.g genes 5 and 15) but had homology to *Phaseolus vulgaris*, which provides additional degree of confidence in synteny and functional gene discovery among related crops.

Comparative genomics and the *Rpp3* locus shared Synteny

The soy bean sequence Blast comparisons to *Arabidopsis thaliana* database found Chromosome 5 as best hit, with expected-value (E-value) of 3e-128 and 76% sequence identity while for *Oryza sativa,* the best hit was Chromosome 2, *Oryza sativa japonica*, with E-value of 2e-21 and 84% sequence identity. The Fgenesh tool predicted 248 genes (115 in positive (+) chain and 133 in negative (-) chain) with 796 exons (365 in +chain and 431 in –chain) using the *Medicago truncatula* as the reference organism, whereas *Arabidopsis* database revealed 166 genes (74 in +chain and 92 in –chain) with 731 exons (336 in +chain and 395 in –chain).

Primers designed for the *Rpp3* locus and QTLs linked to marker Satt460

The list of alternative primer pair sequences designed with potential to amplify the *Rpp3* locus and the Augustus program predicted genes that are linked to marker Satt460 are shown in Table 5. The primer

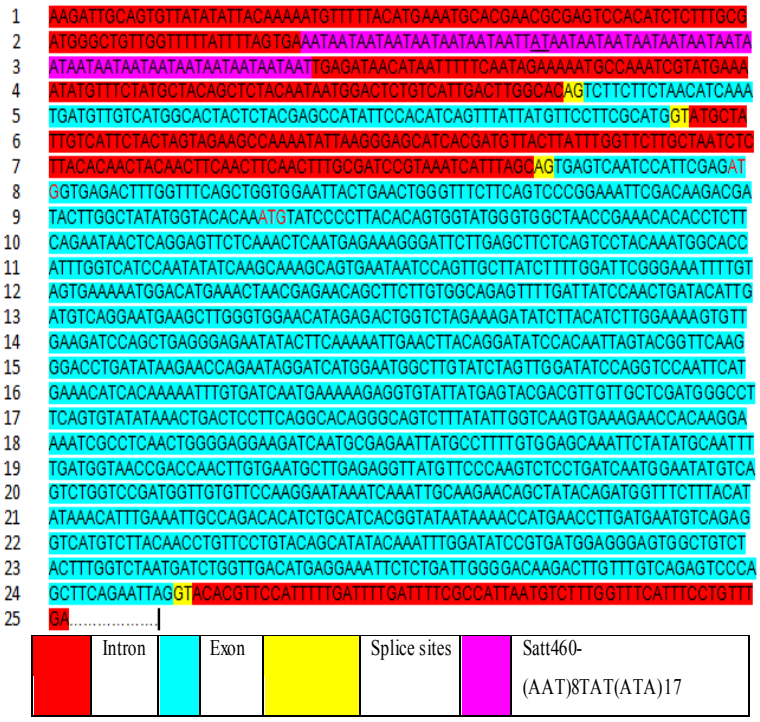


Figure 2: Manual annotation of gene 1 predicted with Augustus software and showing tight linkage to soybean rust disease marker Satt460 (SSR repeat) with an AT base insertion located within the intron (non coding region). Two exons (coding regions) in the predicted gene 1 is displayed with different sizes and interspersed with introns. The splice sites are not part of the coding region but show the probable regions of DNA slicing of mRNA post-translation modifications during protein synthesis.

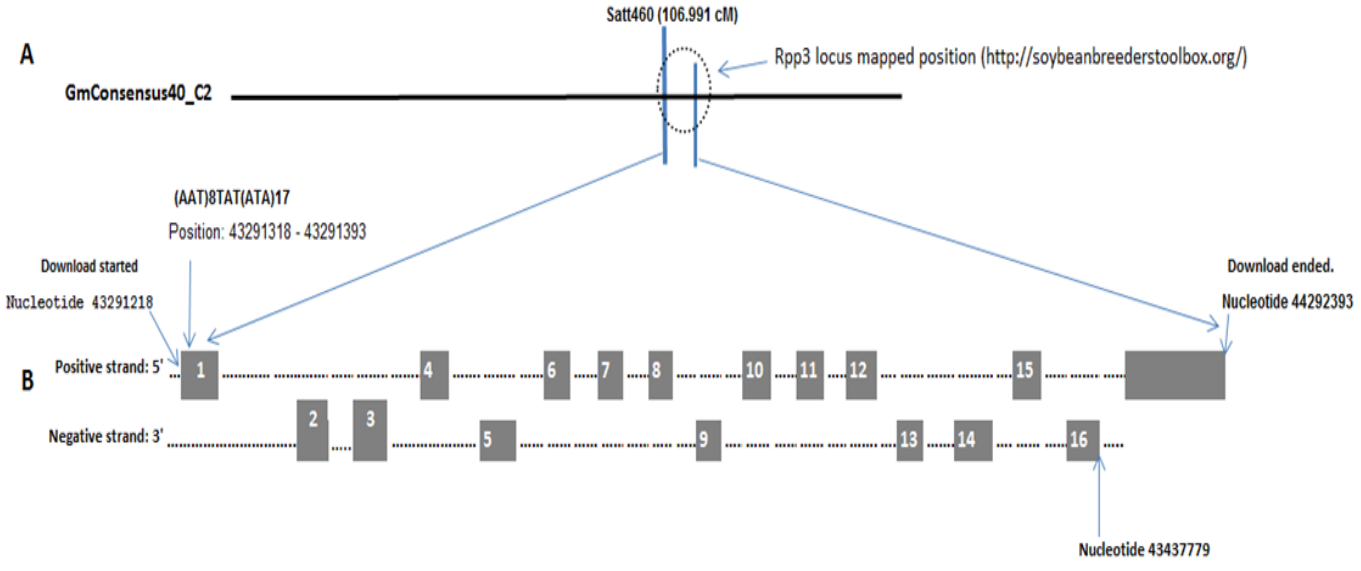


Figure 3: A. The GmConsensus40_C2 genetic Linkage map of Soybean with position of Satt460 and presumed *Rpp3* locus linked within a centimorgan (cM), this region was downloaded for subsequent analyses. **B.** Physical genomic sequence map for *Glycine max* chromosome 6 genomic scaffoldin the NCBI database downloaded as in materials and methods. Subsequent analyses with Augustus predicted 16 genes (small rectangles) which are tightly linked to locus Satt460 ((AAT)8TAT(ATA)17)) the marker for *Rpp3* locus soybean rust disease. The order and positions of predicted genes is shown for both strands and reconcile with table 4. The broken lines are intergenic nucleotides in the loci.

sequences designed are short with different nucleotide sizes, ranging from 18-24 bp. All the primers had the desired GC content range of 45-60 % except the forward primers for genes 7, 10 and 14. The expected product sizes range from 152 bp for gene 14 to 286 bp in gene 8, and can easily be separated by 1% routine agarose gel in electrophoresis. For Introns smaller than 230 bp, sequences from the flanking bases (exons) were included in the search to increase numbers of GC bases for successful operation of Primer 3 program. The 16th gene predicted

Gene	Intron size (bp)	Input sequence (bp)	Primer	Primer Sequence	TM (°C)	GC (%)	Product size (bp)
1	136	601	F	AGTCCACATCTCTTTGCG	54.6	50	220
			R	CCAAGTCAATGACAGAGTCC	55.5	50	
2	95	251	F	CACCAACGTAGGGCTACAC	56.6	58	297
			R	CTTTAAAGGTGAGGGCGT	55.4	50	
3	1261	350	F	CATCAGCATACTCTCCTTCC	54.8	50	282
			R	TAACCGTAAGGTGTCTCCAC	55.2	50	
4	4899	700	F	TGGACTAGCCAATGATGG	55.3	50	234
			R	AACACTAAAACCCCCTCG	55	50	
5	88	280	F	AGCCTCTGCTAGTGCCTCTG	60	60.0	132
			R	AGCCTGCAGAGCTAGGGTTT	60.5	55	
6	6122	950	F	GGGCTTGGCTTAAATCTTCC	60	50	240
			R	ATCAAATGGCTTCCATCTCG	60	45	
7	81	350	F	CACCCCAAAATAACCCAAAA	59.5	40	121
			R	TGCTAACGAAGAACGCTTGA	59.8	45	
8	74	420	F	TTCACAACCCTATTGTGGCA	60	45	286
			R	GTCCATGATTGCCTCAGCTT	60.2	50	
9	274	350	F	CTCCAATGGTCTTGGATCGT	60	50	191
			R	AAGGAAGTACTGAGTGCGGC	59.5	55	
10	483	350	F	TGTGTGCTTGTGTTTGTCTCTT	58.5	41	218
			R	CAAACACACTGAGCCACAGAC	59.4	52.4	
11	476	230	F	GGGCTGATCCAAGAGACAAA	60.2	50	200
			R	ATCGACATCCCTTCCAACAA	60.3	45	
12	73	230	F	TATGGCTACCCATCTGCCTC	60.1	55	189
			R	TCATTCCCACCATCTTGGAT	60.1	45	
13	1689	420	F	ACACAACAATTGGTCTGCCA	60	45	227
			R	GAGGAATCTTAGGGAACCGC	60	55	
14	252	280	F	TGAGCATACATATTCTGCAACAAA	59.7	33.3	152
			R	GGGTTAGTTGGACTTTGGGG	60.6	55	
15	82	280	F	GTACCAGGCAGAAGGGAACA	60.1	55.0	218
			R	AGTGTTTGGGAACCTGAACG	60	50	

Product size corresponds to the expected band size in a gel- electrophoresis assay, following successful amplifying of each predicted genes. Primer F is forward primer and R is reverse prime for the corresponding predicted gene in soybean. Primers highlighted with GC content less than 45% need to be extended as explained in the discussion section

Table 5: The predicted genes, with corresponding introns sizes, and designed primer pairs, with melting temperature (TM), GC bases content and product size genes that are linked to marker Satt460 and the putative *Rpp3* locus.

was intronless, hence had no primer was designed for it. The other QTLs linked to Satt460 (http://soybeanbreederstoolbox.org) are also targeted by the new primer sets designed.

Discussion

The objective of this study was to annotate the *Rpp3* locus to generate information for enhancing marker assisted selection (MAS) in soybean. Windsor and Mitchell [27] recommended, exploration of particular sets of genes and transcribed sequences soon after getting a draft or complete genomic sequence of an organism. The study shows that the *Rpp3* locus sequences is highly repeated with its underlying genes located downstream of marker Satt460 as illustrated below.

Repetitive sequences in soybean *Rpp3* gene locus

The nucleotide repeats around *Rpp3* gene are predominantly of dispersed repeat type (transposable element) and simple repeats. The findings agree to general concept by Nunberg et al. [28] that transposable DNA's comprise a significant proportion of the repetitive DNA found in eukaryotic genomes. They reported retroelements to vary in size from 11 to 14kb and duplicated a few hundred times in the soybean genome. In earlier related studies, Vodkin et al. [29] identified the first transposable element called Tgm in soybean. Lin et al. [30] explored the distribution of repetititve sequences in soybean genome and found them largely localised to the pericentromic region. The other repeats in the study i.e Integrated Virus, Endogenous Retrovirus are not segments of the genome but possibly recombinant DNA from the laboratory and comprises artificial DNA molecules such as cloning vectors. Some of the repeats unique to the Censor software, unlike the TREP software was due to very short sequences masked by the much longer significant transposons during the BLASTn search. Earlier DNA-DNA renaturation studies suggested that approximately 40-60% of the soybean genome sequence is repetitive [31,32], large and complex with high proportions of repetitive elements [33], with relatively few specific repetitive sequences [28] compared to findings in this study.

Molecular mechanisms like unequal crossing over, rolling circle amplification, replication slippage and mutations operating over a long time through selection are the probable sources of repeated DNA sequences in soybean. The crops genome was also reported to have exhibited two duplication events approximately 15 million year ago [34,35]. In future there might be need to relate the repetitive elements in present day soybean to its wild relatives to help broaden understanding of the crops genetic diversity and post domestication syndrome. In overall, flowering plants including soybean contain arrays of repetitive elements and genes assembled into different sets

of chromosomes [36]. Practically the DNA sequences polymorphism in crop genomes is the basis for use of repeats as molecular markers for taxonomic and phylogenetic studies [37]. In addition, mutations in transposable elements may create novel genes or loss of fitness in an organism Satyawada et al. [38] creating genetic variations. The marker Satt460 with sequence AT(8)TTATT(17) has an AT base insertion in its sequence and is located within the intron in close proximity to the predicted genes. The difference (AT base insertion) between the Satt460 in the soybean breeders database and previous studies (e.g Monteros et al. [12] and Hyten et al. [15]) to the one annotated in this study is possibly a mutation or sequencing error in this part of genome. Tandem repeats in the genome play significant structural and functional roles besides tagging genes of economic importance to molecular plant breeders [38].

Gene predictions and related quantitative trait loci (*QTLs*)

Gerstein et al. [39] defined a gene as a segment of genomic sequences encoding a coherent set of potentially overlapping functional products. In this study, sixteen genes were predicted around the putative *Rpp3* gene locus of soybean using *ab initio* gene predictions method (i.e. without making use of external evidence about the gene structure of the input sequence) using the AUGUSTUS tool. The results suggest that the *Rpp3* genes are non-overlapping, with variations in size and number of exons and not uniformly distributed across the genomic sequence analysed. In similar studies in rice, Nagaki et al. [40] found genes not uniformly distributed across the genomes regardless of their size. The paleoploid nature of soybean, with 2n = 40 [41], could be the source of the many genes identified in this study. And it is therefore expected that any given gene will be approximately four times in the genome [28]. The QTLs associated with rust resistance and tagged with maker Satt460 include; seed total oil content, seed weight, leaf shape, reaction to *Phytophthora megasperma f sp glycinea infection*, seed genistein content, pod number, main stem branching, (http://soybase.org/cmap/cgibin/cmap/feature?feature_acc=GmConsensus40_C2_Satt460).

Jain and Brar [42] reported occurrence of undesirable genomic portions (QTLs) in improved varieties through introgressive hybridization and attributed this to lack of knowledge on genes underlying the QTLs controlling the targeted agronomic traits. This study generally therefore attempted to demonstrate how to address this gap for plant breeders. The genes predicted on the positive sequence strand are in cis-orientation to the Satt460 marker and in coupling with the listed QTLs. This analogy will practically improve soybean MAS breeding strategies i.e amplified bands in gel electrophoresis could simultaneously tag both the underlying major genes and quantitative trait influenced but this needs verifications.

Comparative genomics and shared synteny of Soybean *Rpp3* genes

The comparison of soybean sequence to model crops (*Arabidopsis thaliana* and *Oryza sativa*) show sequence homology. The Cereal repeat repository in Trep database has conserved transposable repeats between cereals and soybean. This could point to highly conserved earlier shared sequence blocks in a common ancestor, although the crops diverged millions of years ago. Grant et al. [43] demonstrated synteny between *Arabidopsis* and soybean using sequences of mapped soybean RFLP probes and *Arabidopsis* genomic sequence. These findings were surprising, given the millions of years since the divergence of their lineages. In other studies, Yan et al. [44] found only three of 50 soybean contigs (6%) to possess microsynteny with *Arabidopsis*, whereas 54% showed microsynteny with *Medicago truncatula*. Shoemaker et al. [45] suggested cross-referencing soybean to model legumes to speed soybean genomics advances. The comparisons of a crop genome to model plants, helps to identifying conserved regions and provides a foundation for accelerating prediction novel genes in other plant taxa [42]. In addition Korf [46] recommends discovery of genes in related species by comparing to genomes to detect evolutionary pressures for conservation purposes. This is because the forces of natural selection cause genes and other functional elements to undergo mutation at a slower rate than the rest of the genome, since such mutations are more likely to negatively impact the organism than ones elsewhere.

Primer pairs designed

The forward and reverse primers presented in this study were not validated due to economic reasons, and hence the success for their amplifications and polymorphism would be an over speculation. The primers with GC base content less than 45% need to be extended. The GC % is an important characteristic of DNA and provides information about the strength of annealing thus should range between 45-60 percent in primers [47,48]. The absence of the original soybean sequence from which Satt460 was developed limited the discovery of other alternative more specific primers for screening soy bean rust disease. The very close proximity of targeted introns (markers) targeted by the primers and marker Satt460, of less than a centimorgan, on a positive side, gives promises of successful amplification of putative genes in soybean *Rpp3* locus in PCR assays. The designed primers therefore provide additional tools for screening soybean rust resistance and associated agronomic traits. They can also be extended to phylogenetic studies (i.e for germplasm characterisation) after validations since they are specifically designed to amplify predicted conserved-regions of the genome. In previous studies, Hyten et al. [15] designed 48 primers between nucleotides 1,077,201 and 1,977,200 in soybean scaffold 60 for amplifying the *Rpp3* locus sequences. They reported seven out of the 48 primers to have multiple amplicons in the soybean genome. This study used a segment of the same soybean genome to elucidate gene functions and related the genes with associated QTL to enhance MAS in the crop. The utility of these resources however depend on coordination of data assimilation into bioinformatics systems and training in the practical operation of those resources in crop improvement and other areas of science.

Functional annotation and protein predictions of the *Rpp3* gene sequences

The function of the protein coding genes in the *Rpp3* locus in Soybean was sought in this study. Gene 1 and 12 have the same biological roles, suggesting duplication of chromosome 6 in Soybean at these loci with no loss in function. The genes; 6, 7, 8 showed transcription activation, while gene 10 acts in reverse sense as a deactivator. This is expected since; the sequence analysed had predominantly DNA transposons (repetitive sequences) from our structural annotations. It was interesting to note that all the genes with functions derived are situated on the positive strand of chromosome 6 genomic sequences (Table 4). The other genes transcripts (genes 2,4,5,9 and 15) that translated into proteins with significant hits but no putative roles discovered are located on the negative strand (5' to 3' strand) due to the concept of the dogma of mRNA translation starting from upstream and proceeding downstream during protein synthesis. This implies that, in such a scenario, genes of this kind possibly needed

to first be reverse transcribed prior functional annotation, and the resultant complimentary DNA (cDNA) used in future studies. It is possible that the function of these genes has not yet been determined, despite being deposited in the NCBI or they simply don't code for any functional protein. The genes expressing immune responses (i.e 1, 12 and 14) and seed protein (gene 13) may be of interest to plant breeders, developing cultivars with disease resistance and nutritional levels in the grains. Soybean is a subject of ongoing functional genomics projects as introduced. This study thus, accurately reflects the current contribution of genomics to the understanding function of proteins on the genome scale. Through such functional annotation, discovery of novel genes of Soybean has been demonstrated in this study. The selection of these genes can then proceed efficiently via MAS for incorporation into new cultivars. Lack of previous functional annotation studies conducted in in Soybean *Rpp*3 locus did not allow for comparisons of our findings. Nevertheless determining if a sequence is functional is different from determining the function of the gene or its product [46]. The latter demands *in vivo* experimentation through gene knockout and other assays not addressed in this study. Bioinformatics has however made it possible to predict the function of genes based on sequence information alone and there is room for improvement in future studies.

Conclusions and Recommendations

The predominant repeats in the *Rpp3* gene are DNA transposons that are localised in the DNA fragment studied. DNA transposons are repetitive sequences in the genome, which can be amplified using short primer sequences designed as demonstrated by the study. These could be more applicable to diversity studies as a polymorphic parameter among soybean populations in time and space. Sixteen non-overlapping genes predicted are linked to marker Satt460 for rust resistance in soybean. 15 primers are designed for the predicted genes but need to be validated for use in Marker assisted breeding. Lectin protein was predicted which confirms disease resistance role of the *Rpp3* gene. The results show that comparative analysis of closely related species can be valuable in understanding a genome. We strongly recommend similar studies on the other five genes conferring resistance to Soybean rust disease for effective gene pyramiding to develop varieties with more durable resistance.

Acknowledgement

This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Dennis Okii is grateful for support by the Kirkhouse Trust foundation (UK) for the training in Applied Bioinformatics at the University of California, Davis, U.S.A. Special thanks go to Dr. Jill Wegrzyn, Mr. Josh Hegarty, Dr. James Kami and Profs: David Neale and Paul Gepts who taught the concepts used in this study.

References

1. Edwards D, Henry RJ, Edwards KJ (2012) Preface: advances in DNA sequencing accelerating plant biotechnology. Plant Biotechnol J 10: 621-622.
2. Edwards D, Batley J (2010) Plant genome sequencing: applications for crop improvement. Plant Biotechnol J 8: 2-9.
3. Berkman PJ, Skarshewski A, Lorenc MT, Lai K, Duran C, et al. (2011) Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. Plant Biotechnol J 9: 768-775.
4. Imelfort M, Duran C, Batley J, Edwards D (2009) Discovering genetic polymorphisms in next-generation sequencing data. Plant Biotechnol J 7: 312-317.
5. Allen AM, Barker GL, Berry ST, Coghill JA, Gwilliam R, et al. (2011) Transcript-specific, singlenucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). Plant Biotechnol J 9: 1086-1099.
6. Hiremath PJ, Farmer A, Cannon SB, Woodward J, Kudapa H, 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 Biotechnol J 9: 922-931.
7. Malory S, Shapter FM, Elphinstone MS, Chivers IH, Henry RJ (2011) Characterizing homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes. Plant Biotechnol J 9: 1131-1140.
8. Nock CJ, Waters DL, Edwards MA, Bowen SG, Rice N, et al. (2011) Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnol J 9: 328-333.
9. Marek LF, Mudge J, Darnielle L, Grant D, Hanson N, et al. (2001) Soybean genomic survey: BAC-end sequences near RFLP and SSR markers. Genome 44: 572-581.
10. Yang XB, Tschanz AT, Dowler WM, Wang TC (1991) Development of yield loss models in relation to reduction of components of soybean infected with *Phakopsora pachyrhizi*. Phytopathology 81: 1420-1426.
11. Sinclair JB, Hartman GL (1995) Proceedings of the Soybean Rust Workshop, Urbana, IL. 9–11 Aug. Natl. Soybean Res. Lab. Publ. 1, Urbana-Champaign, IL.
12. Monteros MJ, Missaoui AM, Phillips DV, Walker DR, Boerma HR (2007) Mapping and confirmation of the Hyuuga red-brown lesion resistance gene for Asian soybean rust. Crop Science 47: 829-834.
13. Miles MR, Morel W, Ray JD, Smith JR, Frederick RD, et al. (2008) Adult plant evaluation of soybean accessions for resistance to *Phakopsora pachyrhizi* in the field and green house in Paraguay. Plant Disease 92: 96-105.
14. Maphosa M, Talwana H, Tukamuhabwa P (2012) Enhancing soybean rust resistance through Rpp2, Rpp3 and Rpp4 pair wise gene pyramiding. African Journal of Agricultural Research 7: 4271-4277.
15. Hyten DL, Smith JR, Frederick RD, Tucker ML, Song Q, et al. (2009) Bulked segregant analysis using the GoldenGate assay to locate the Rpp3 locus that confers resistance to soybean rust in soybean. Crop Science 49: 265- 271.
16. Ray JD, Smith JR, Morel W, Bogado N, Walker DR (2011) Genetic Resistance to Soybean Rust in PI567099A is at or Near Rpp3 locus. Journal of Crop Improvement 25: 219-231.
17. Hyten DL, Hartman GL, Nelson RL, Frederick RD, Concibido VC, et al. (2007) Map location of the Rpp1 locus that confers resistance to soybean rust in soybean. Crop Science 47: 837-838.
18. Silva DC, Yamanaka N, Brogin RL, Arias CA, Nepomuceno AL, et al. (2008) Molecular mapping of two loci that confer resistance to Asian rust in soybean. Theor Appl Genet 117: 57-63.
19. Yamanaka N, Silva DCG, Passianotto ALL, Nogueira LM, Polizel AM, et al. (2008) Identification of DNA markers and characterisation of the genes for resistance against Asian soybean rust. In: Kudo H, Suenaga K, Soares RM, Toledo A (Eds.), Facing the challenge of soybean rust in South America. Japan International Research Center for Agricultural Sciences, Japan.
20. Garcia A, Calvo ES, de Souza Kiihl RA, Harada A, Hiromoto DM, et al. (2008) Molecular mapping of soybean rust (*Phakopsora pachyrhizi*) resistance genes: discovery of a novel locus and alleles. Theor Appl Genet 117: 545-553.
21. Kendrick MD, Harris DK, Ha BK, Hyten DL, Cregan PB, et al. (2011) Identification of a second Asian soybean rust resistance gene in Hyuuga soybean. Phytopathology 101: 535-543.
22. Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 20: 119-121.
23. Vassilev D, Leunissen J, Atanassov A, Nenov A, Dimov G (2005) Applications of bioinformatics in plant breeding. Biotechnology & Biotechnological Equipment Special Issue 19: 139-152.
24. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.
25. Kamel AA (2003) Bioinformatic tools and guideline for PCR primer design. African Journal of Biotechnology 2: 91-95.
26. Wu DY, Ugozzoli L, Pal BK, Qian J, Wallace RB (1991) The effect of temperature and oligonucleotide primer length on the specificity and efficiency of amplification by the polymerase chain reaction. DNA Cell Biol 10: 233-238.

27. Windsor AJ, Mitchell-Olds T (2006) Comparative genomics as a tool for gene discovery. Curr Opin Biotechnol 17: 161-167.

28. Nunberg A, Bedell JA, Budiman MA, CitekA RW, Clifton SW, et al. (2006) Survey sequencing of soybean elucidates the genome structure, composition and identifies novel repeats. Functional Plant Biology 33: 765-773.

29. Vodkin LO, Khanna A, Clough S, Shealy R, Philip R, et al. (2002) Structural and functional genomics projects in soybean. Plant Molecular Biology Reporter Supplement 18: S1.

30. Lin J, Jacobus BH, SanMiguel P, Walling JG, Yuan Y, et al. (2005) Pericentromeric regions of soybean (Glycine max L. Merr.) chromosomes consist of retroelements and tandemly repeated DNA and are structurally related and evolutionarily labile. Genetics 170: 1221-1230.

31. Goldberg RB (1978) DNA sequence organization in the soybean plant. Biochem Genet 16: 45-68.

32. Gurley WB, Hepburn AG, Key JL (1979) Sequence organization of the soybean genome. Biochim Biophys Acta 561: 167-183.

33. Stacey G, Vodkin L, Parrott WA, Shoemaker RC (2004) National Science Foundation-sponsored workshop report. Draft plan for soybean genomics. Plant Physiol 135: 59-70.

34. Schlueter JA, Dixon P, Granger C, Grant D, Clark L, et al. (2004) Mining EST databases to resolve evolutionary events in major crop species. Genome 47: 868-876.

35. Shoemaker RC, Schlueter J, Doyle JJ (2006) Paleopolyploidy and gene duplication in soybean and other legumes. Curr Opin Plant Biol 9: 104-109.

36. Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot 95: 127-132.

37. Smith DB, Flavell RB (1974) The relatedness and evolution of repeated nucleotide sequences in the genomes of some Gramineae species. Biochem Genet 12: 243-256.

38. Satyawada RR, Seema T, Deepika E, Keisham M, Marlykynti H (2010) DNA repetitive sequences-types, distribution and function: A review. Journal of Cell and Molecular Biology 7: 1-11.

39. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, et al. (2007) What is a gene, post-ENCODE? History and updated definition. Genome Res 17: 669-681.

40. Nagaki K, Cheng Z, Ouyang S, Talbert PB, Kim M, et al. (2004) Sequencing of a rice centromere uncovers active genes. Nat Genet 36: 138-145.

41. Goldblatt P (1981) (Eds) Cytology and phylogeny of Leguminosae. Advances in legume systematic Part 2. Royal Botanic Gardens.

42. Jain SM, Brar DS (2010) (Eds) Molecular Techniques in Crop Improvement. Springer Science and Business Media.

43. Grant D, Cregan P, Shoemaker RC (2000) Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. Proc Natl Acad Sci U S A 97: 4168-4173.

44. Yan HH, Mudge J, Kim DJ, Larsen D, Shoemaker RC, et al. (2003) Estimates of conserved microsynteny among the genomes of Glycine max, *Medicago truncatula* and *Arabidopsis thaliana*. Theor Appl Genet 106: 1256-1265.

45. Shoemaker RC (2003) The Status of Soybean Genomics and Its Role in the Development of Soybean Biotechnologies. AgBioForum 6: 4-7.

46. Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5: 59.

47. Dieffenbach CW, Lowe TMJ, Dveksler GS (1995) General Concepts for PCR Primer Design. In: Dieffenbach CW, Dveksler GS (Eds) PCR Primer, A Laboratory Manual. Cold Spring Harbor Laboratory Press, New York.

48. Stanke M, Steinkamp R, Waack S, Morgenstern B (2004) AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 32: W309-312.