Genetics and Molecular Biology, 35, 1 (suppl), 203-211 (2012)
Copyright © 2012, Sociedade Brasileira de Genética. Printed in Brazil
www.sbg.org.br

Research Article

A web-based bioinformatics interface applied to the GENOSOJA Project: Databases and pipelines

Leandro Costa do Nascimento[1], Gustavo Gilson Lacerda Costa[1], Eliseu Binneck[2],
Gonçalo Amarante Guimarães Pereira[1] and Marcelo Falsarella Carazzolle[1,3]

*[1]Laboratório de Genômica e Expressão, Departamento de Genética, Evolução e Bioagentes,
Instituto de Biologia, Universidade Estadual de Campinas, Campinas, SP, Brazil.*
[2]Empresa Brasileira de Pesquisa Agropecuária, Londrina, PR, Brazil.
*[3]Centro Nacional de Processamento de Alto Desempenho em São Paulo,
Universidade Estadual de Campinas, Campinas, SP, Brazil.*

Abstract

The Genosoja consortium is an initiative to integrate different omics research approaches carried out in Brazil. Basically, the aim of the project is to improve the plant by identifying genes involved in responses against stresses that affect domestic production, like drought stress and Asian Rust fungal disease. To do so, the project generated several types of sequence data using different methodologies, most of them sequenced by next generation sequencers. The initial stage of the project is highly dependent on bioinformatics analysis, providing suitable tools and integrated databases. In this work, we describe the main features of the Genosoja web database, including the pipelines to analyze some kinds of data (ESTs, SuperSAGE, microRNAs, subtractive cDNA libraries), as well as web interfaces to access information about soybean gene annotation and expression.

Key words: bioinformatics, database, gene expression, soybean, Genosoja.

Introduction

Soybean is a legume of great economic importance in the international market, with a world production of almost 260 million tons for the 2009/2010 harvest. Brazil appears as the world's second largest producer, with a share of about 25%, and the crop is responsible for 10% of the country's total exports.

In recent years, the Brazilian soybean crop has been constantly threatened by climatic constraints (especially long periods of drought) and some attacks by pathogens - such as Asian Rust disease (Yorinori *et al.*, 2005), resulting in millions in losses for producers. Such constraints increase the importance of the breeding programs, established to discover new techniques for planting and prevention, increase production and lower costs. High-throughput sequencing technologies (like 454 pyrosequencing, Illumina/Solexa and ABI/SOLiD) represent significant advantages in these areas by producing millions of reads that can be used to measure levels of gene expression, allowing the identification of new genes or novel splice variants. On the other hand, it is necessary to intensify efforts by bioinformatics groups to develop new pipelines and data integration.

To this end, the Brazilian Soybean Genome Consortium (Genosoja) was established in 2007 with the goal of integrating several institutions currently working with soybean genomics in Brazil. The project promotes the search for solutions regarding possible treats, and to improve the soybean production process, emphasizing stresses that affect domestic production, like the occurrence of droughts, and pathogen attacks such as Asian Rust disease.

Among the main objectives of the Genosoja consortium is the creation of a relational database, integrating the results achieved by different methodologies and research groups working in the project. Despite the existence of other integrated soybean databases, such as SoyXpress (Cheng and Stromvik, 2008) and the "Soybean full-length cDNA database" (Umezava *et al.*, 2008), as well as free pipelines available for data integration, such as Distributed Annotation System (DAS) (Dowell *et al.*, 2001; Jenkinson *et al.*, 2008) none of them integrates data from multiple experiments or provides transcriptome data from high-

Send correspondence to Gonçalo Amarante Guimarães Pereira. Laboratório de Genômica e Expressão, Departamento de Genética, Evolução e Bioagentes, Instituto de Biologia, Universidade Estadual de Campinas, Cidade Universitária Zeferino Vaz, 13083-970 Campinas, SP, Brazil. E-mail: goncalo@unicamp.br.

throughput sequencing technologies like the database described in this work.

In light of this context, we created a soybean database connecting public soybean data (like ESTs and genomic sequences) and project data (like SuperSAGE tags, microRNAs and subtractive cDNA libraries). This database offers search tools for users, including keyword searches, statistical comparisons, automatic annotation (using some protein databases such as NR, Uniref, KEGG and Pfam), gene ontology classification and gene expression profiles under several conditions. Moreover, searches by sequence homology are possible using the local BLAST. All data are stored in a Fedora Linux machine, running the MySQL database server. The web interfaces (http://www.lge.ibi.unicamp.br/soybean) are based on a combination of CGI scripts using Perl language (including BioPerl module) and the Apache Web Server. As soon as the private data are published, the database will be freely available.

Methods, Results and Discussion

Public soybean data

In order to construct the Genosoja database we first collected all soybean data available at public biology databases. The genome of the cultivar Williams 82 and their predicted genes (66,153 sequences) were downloaded from the Phytozome (Schmutz *et al.*, 2010). One full-length cDNA library from the Japanese cultivar Nourin2 was downloaded from the "Soybean full-length cDNA database". From NCBI (National Center for Biotechnology Information) we obtained 1,276,813 EST sequences (sequenced by SANGER and pyrosequencing technologies) and their equivalent GenBank files. All sequences were renamed in accordance to libraries, tissues and cultivars. This information was extracted from the GenBank files using homemade PERL scripts (Supplementary Material Figure S1). The bdtrimmer software (Baudet and Dias, 2005) was used to exclude ribosomal, vector, low quality and short (less than 100 bp) sequences. The EST assembly process was divided into two steps: (1) the ESTs were mapped into the soybean genome using the BLASTn algorithm (Altschul *et al.*, 1997) (e-value cutoff of $1e^{-10}$) and (2) all reads that aligned in same region of the reference were assembled together using the CAP3 program (Huang and Madan, 1999). The final result consists of 60,747 unigenes (30,809 *contigs* and 29,938 *singlets*). The effort to obtain the unigenes from assembled ESTs was important to increase the databases with information on untranslated regions (UTR), alternative splicing variants and gene expression profiling.

The Autofact program (Koski *et al.*, 2005) was used to perform an automatic annotation of the predicted genes and the assembled unigenes. The main contribution of Autofact is the capacity to resume the annotation based on sequence similarity searches in several databases. For this, we used the BLASTx procedure (e-value cutoff of $1e^{-5}$) to align the genes against certain protein databases, including: non-redundant (NR) database of *NCBI*, swissprot - databases containing only manually curated proteins (Suzek *et al.*, 2007), uniref90 and uniref100 - databases containing clustered sets of proteins from UniProt, Pfam - a database of protein families (Bateman *et al.*, 2002) and KEGG - a database of metabolic pathways (Kanehisa and Goto, 2000). The Autofact pipeline assigned function to 85% of the protein dataset. Figure 1 shows the complete pipeline of the public soybean data analysis.

Using the description of the origin of the ESTs (tissues and conditions), normalization procedures and statistical data analysis (Audic and Claverie, 1997), it was possible to infer differential gene expression among the assembled unigenes. This approach, called Electronic Northern, allows the users to compare gene expression profiles between two or more libraries and the results are available through a web interface (Figure 2).

Finally, the users can perform keyword and BLAST searches directly from the EST reads using the Gene Projects software (Carazzolle *et al.*, 2007). This software also allows the user to perform assembly and annotation in these reads in an effort to improve unigene assembly. After generating a login/password it is possible to work on specific projects which users can develop and organize thematically by adding sequences to the assembly. After the assembly it is possible to view and to edit the results, improving the quality of the contigs.

Solexa SuperSAGE data

The Genosoja project generated three libraries using SuperSAGE methodology and these were sequenced by Illumina/Solexa technology. One library was constructed exploring gene expression in plants (Brazilian cultivar PI561356 - resistant) infected by the fungus *Phakopsora pachyrhizi* (Asian Rust disease) and two samples of plants (Brazilians cultivars: BR 16 - susceptible and Embrapa 48 - resistant) submitted to drought stress - for descriptions see Soares-Cavalcanti *et al.* (2012, this issue) and Wanderley *et al.* (2012, this issue). In total, the SuperSAGE approach generated 4,373,053 tags with 26 bp each.

Initially the tags of each sample were grouped in unique sequences. The unique sequences that presented low read counts (read count < 2) were discarded from the list. The Audic-Claverie statistic (Audic and Claverie, 1997) with a 95% confidence level (cutoff of 0.05) was used to identify tags as *up-regulated* (more expressed in the

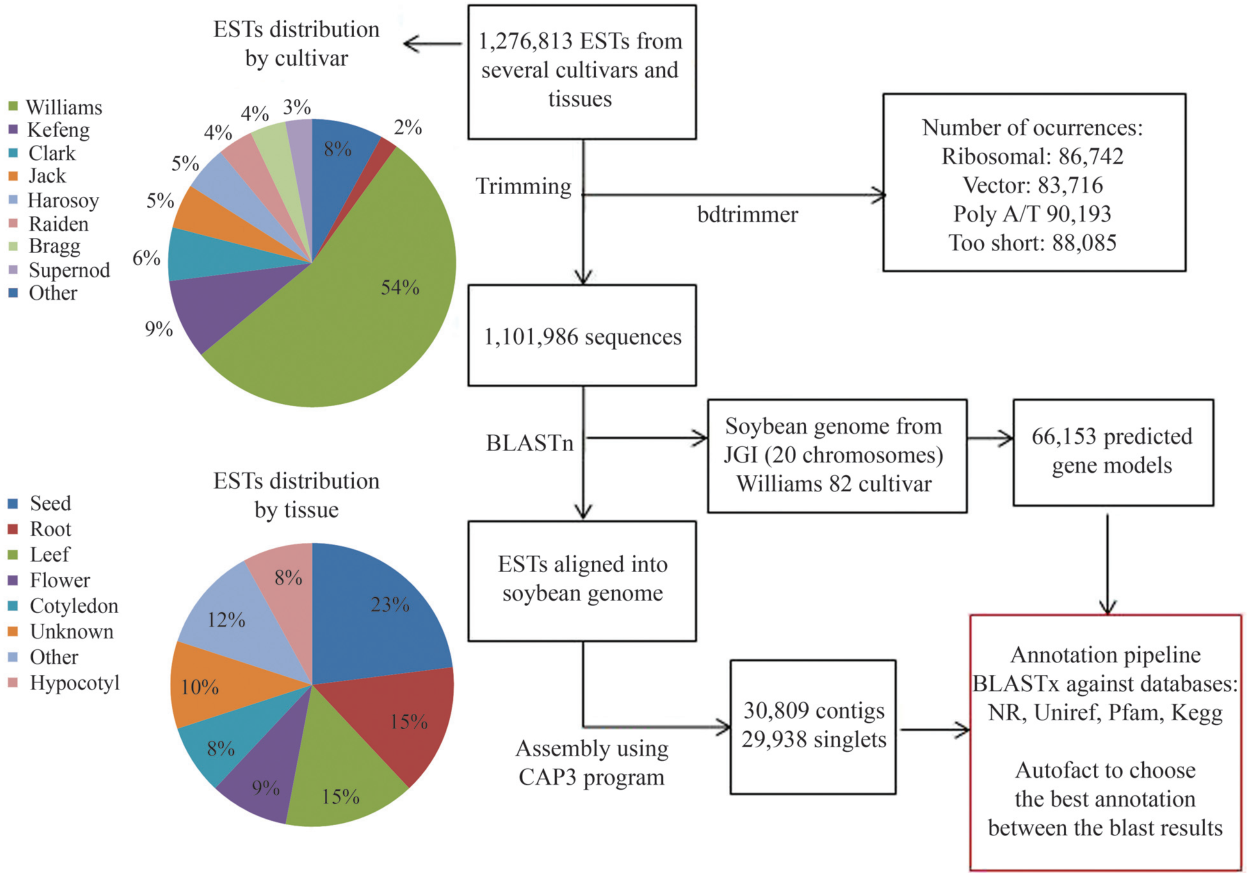
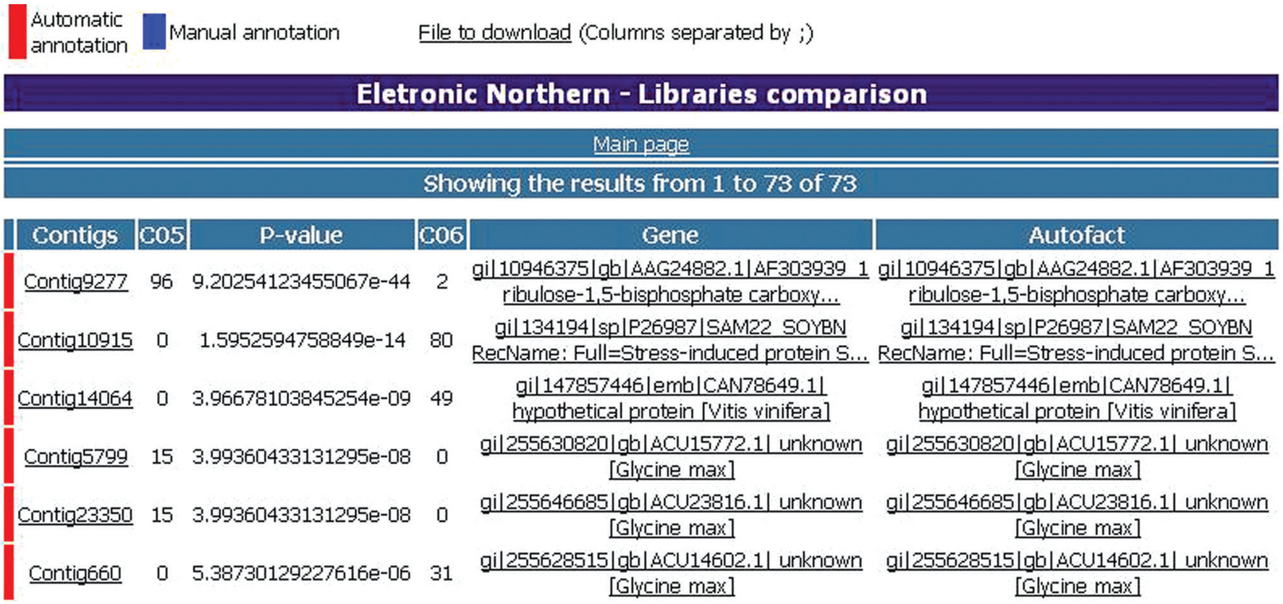

ESTs distribution by cultivar
Williams
Kefeng
Clark
Jack
Harosoy
Raiden
Bragg
Supernod
Other
8%
2%
54%
9%
6%
5%
5%
4%
4%
3%
1,276,813 ESTs from several cultivars and tissues
Trimming
bdtrimmer
Number of ocurrences:
Ribosomal: 86,742
Vector: 83,716
Poly A/T 90,193
Too short: 88,085
1,101,986 sequences
BLASTn
Soybean genome from JGI (20 chromosomes) Williams 82 cultivar
66,153 predicted gene models
ESTs distribution by tissue
Seed
Root
Leef
Flower
Cotyledon
Unknown
Other
Hypocotyl
23%
15%
15%
9%
8%
10%
12%
8%
ESTs aligned into soybean genome
Assembly using CAP3 program
30,809 contigs
29,938 singlets
Annotation pipeline
BLASTx against databases:
NR, Uniref, Pfam, Kegg
Autofact to choose the best annotation between the blast results

Figure 1 - Complete pipeline of the public soybean data analysis. We found many occurrences of vector and poly A/T sequences in the *NCBI* ESTs. After trimming, a reference assembly was performed using 1,101,986 sequences. Moreover, the predicted Williams 82 genes (66,153 sequences) and the assembled unigenes (60,747 sequences) were automatically annotated using the AutoFACT pipeline based on certain BLASTx results against several protein databases.


🟥 Automatic annotation 🟦 Manual annotation File to download (Columns separated by ;)

Eletronic Northern - Libraries comparison

Main page

Showing the results from 1 to 73 of 73

Contigs	C05	P-value	C06	Gene	Autofact
Contig9277	96	9.20254123455067e-44	2	gi|10946375|gb|AAG24882.1|AF303939_1 ribulose-1,5-bisphosphate carboxy...	gi|10946375|gb|AAG24882.1|AF303939_1 ribulose-1,5-bisphosphate carboxy...
Contig10915	0	1.59525947588849e-14	80	gi|134194|sp|P26987|SAM22_SOYBN RecName: Full=Stress-induced protein S...	gi|134194|sp|P26987|SAM22_SOYBN RecName: Full=Stress-induced protein S...
Contig14064	0	3.96678103845254e-09	49	gi|147857446|emb|CAN78649.1| hypothetical protein [Vitis vinifera]	gi|147857446|emb|CAN78649.1| hypothetical protein [Vitis vinifera]
Contig5799	15	3.99360433131295e-08	0	gi|255630820|gb|ACU15772.1| unknown [Glycine max]	gi|255630820|gb|ACU15772.1| unknown [Glycine max]
Contig23350	15	3.99360433131295e-08	0	gi|255646685|gb|ACU23816.1| unknown [Glycine max]	gi|255646685|gb|ACU23816.1| unknown [Glycine max]
Contig660	0	5.38730129227616e-06	31	gi|255628515|gb|ACU14602.1| unknown [Glycine max]	gi|255628515|gb|ACU14602.1| unknown [Glycine max]

Figure 2 - Electronic northern interface. With this tool it is possible to infer gene expression using an assembly of ESTs. A statistical test (p-value) is performed in real time when comparing two libraries. The description of the libraries and tissue ESTs were obtained from a GenBank sequences file using a specifically made PERL script (Supplementary Material Figure S1). Furthermore, a file with the results shown in the interface is available for download.

treated library) or *down-regulated* (more expressed in the control library).

In order to connect the unique tag with a gene sequence, the SOAP2 aligner program (Li *et al.*, 2009) was used to align the unique tags with three databases (shown previously): (*i*) assembled unigenes (60,747 sequences), (*ii*) predicted genes (66,153 sequences) and *(iii*) the soybean genome. The program has been configured to allow for up to 2 mismatches in the alignments (SNPs can generate mismatches in the alignment, especially in this case because the assembled unigenes are generated by ESTs from different cultivars) and return all optimal alignments. In cases where more than one optimal alignment exists we decided to use the results from assembled unigenes, because they contain the UTR regions (a large part of the SuperSAGE tags are in the 3' UTR region), followed by the alignment with the predicted genes and, in the last case, the genome alignment was considered. Figure 3 shows the pipeline used in the SuperSAGE analysis.

For the sample submitted under drought stress (BR 16 and Embrapa 48 cultivars) we mapped 84.3% of the unique tags (Table 1), whereby the remaining 15.7% could represent new soybean genes specific to Brazilian cultivars. Similar results were obtained from a sample infected by Asian Rust, which yielded a total of 21,338 unique tags (20.42%) (Table 1) that did not align with any soybean database. In this case the unique tags from fungi genes may have contributed to increase this value. We tried to map

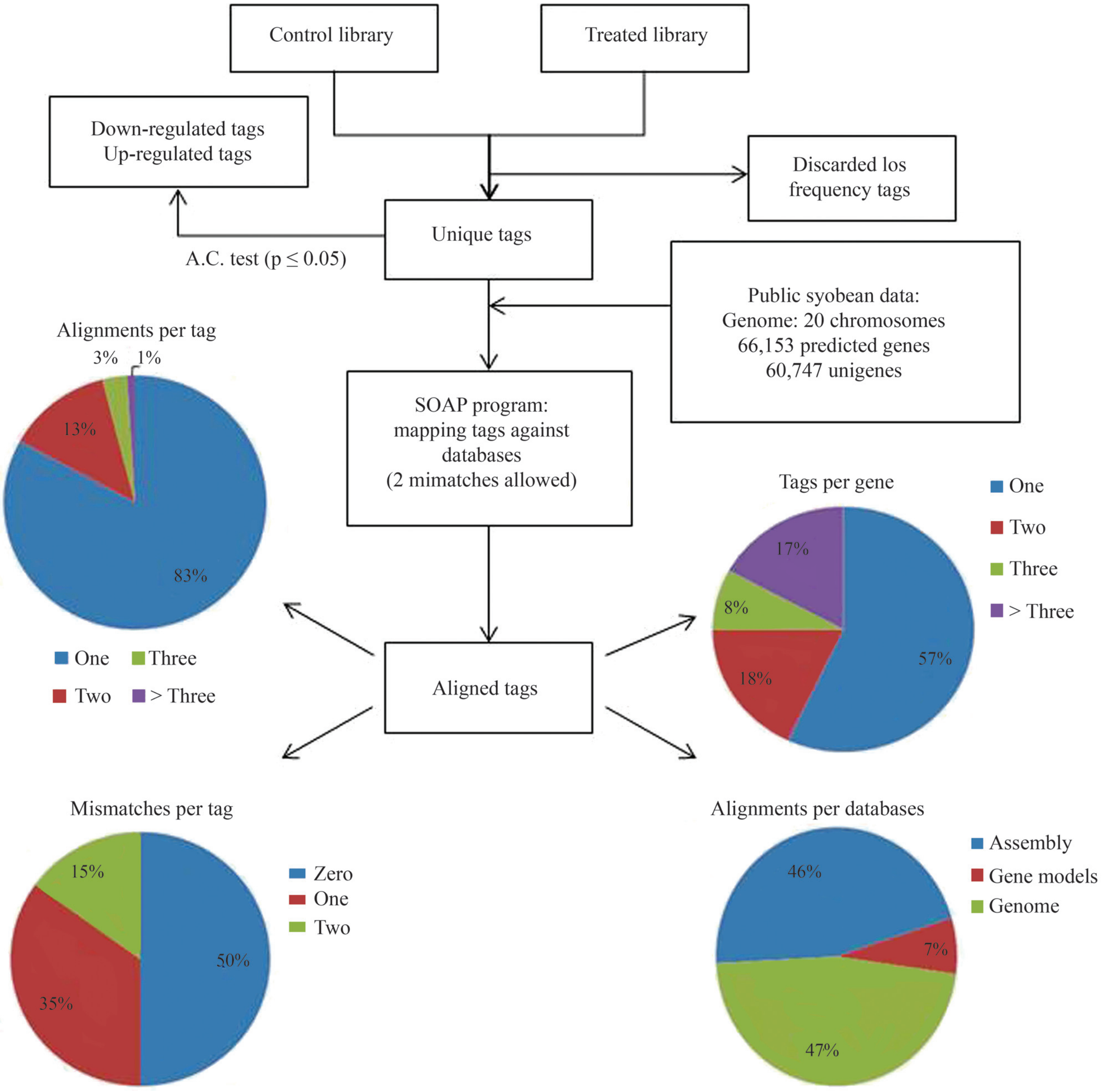
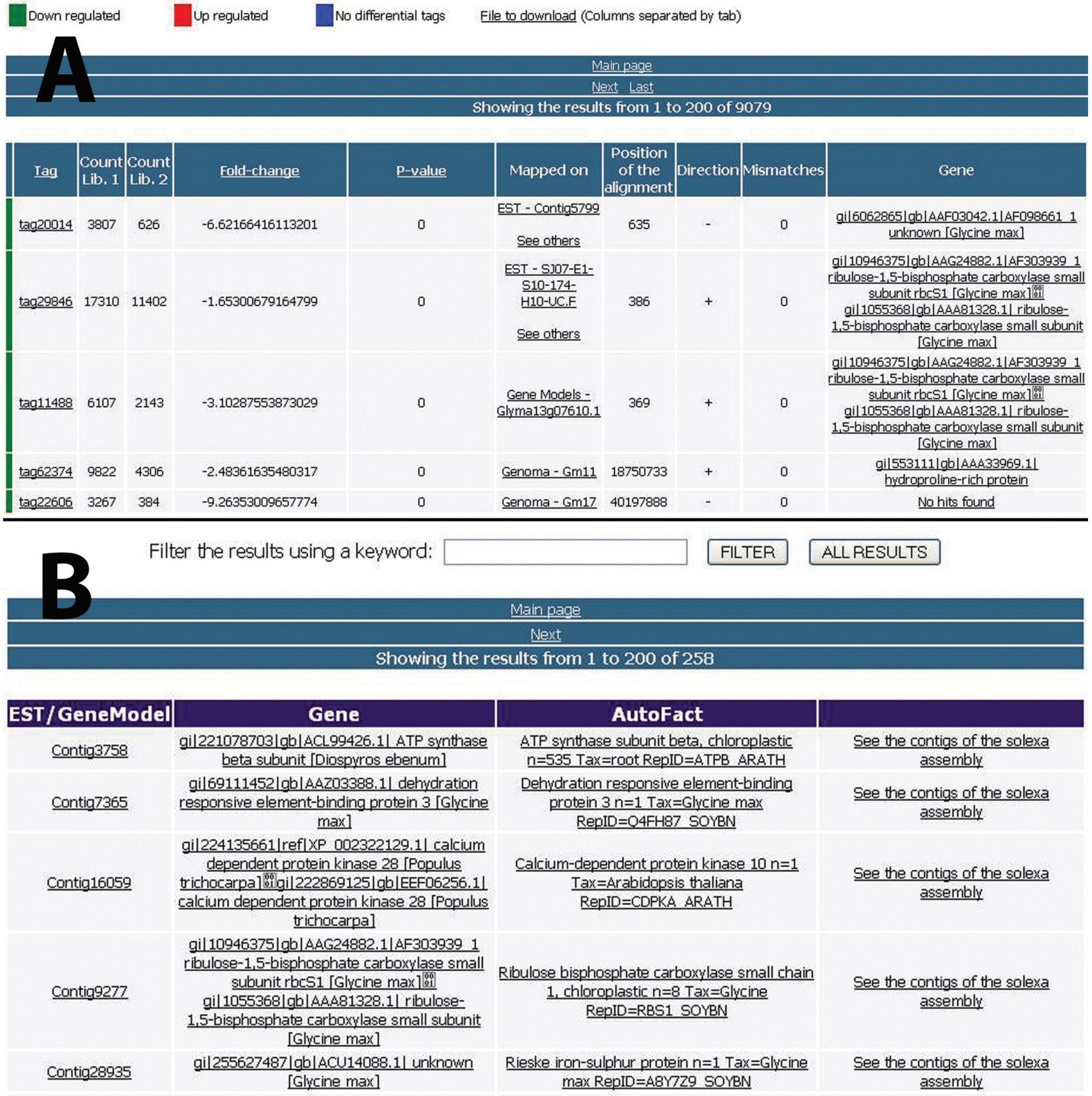


Figure 3 - Complete pipeline of the SuperSAGE data analysis. All tags (control and treated libraries) of each sample were grouped in unique tags. The unique tags were aligned against the public soybean databases using the SOAP2 aligner. Most of the tags have one alignment (83%) and up to one mismatch (85%). A summary of the total, unique, mapped and differential tags is shown in Table 1.

Table 1 - Summary of Solexa SuperSAGE data deposited in the Genosoja databank.

	Total tags (control)	Total tags (treated)	Unique tags	Mapped tags	Differential tags
Asian Rust	813,205	885,439	104,725	83,337 (79.58%)	15,761 (15.05%)
Drought BR 16	1,092,374	509,465	89,205	75,233 (84.34%)	14,450 (16.20%)
Drought Embrapa 48	653,352	419,218	74,833	63,083 (84.30%)	25,364 (33.89%)



🟩 Down regulated 🟥 Up regulated 🟦 No differential tags File to download (Columns separated by tab)

Showing the results from 1 to 200 of 9079

Tag	Count Lib. 1	Count Lib. 2	Fold-change	P-value	Mapped on	Position of the alignment	Direction	Mismatches	Gene
tag20014	3807	626	-6.62166416113201	0	EST - Contig5799 See others	635	-	0	gi|6062865|gb|AAF03042.1|AF098661_1 unknown [Glycine max]
tag29846	17310	11402	-1.65300679164799	0	EST - SJ07-E1-S10-174-H10-UC.F See others	386	+	0	gi|10946375|gb|AAG24882.1|AF303939_1 ribulose-1,5-bisphosphate carboxylase small subunit rbcS1 [Glycine max] gi|1055368|gb|AAA81328.1| ribulose-1,5-bisphosphate carboxylase small subunit [Glycine max]
tag11488	6107	2143	-3.10287553873029	0	Gene Models - Glyma13g07610.1	369	+	0	gi|10946375|gb|AAG24882.1|AF303939_1 ribulose-1,5-bisphosphate carboxylase small subunit rbcS1 [Glycine max] gi|1055368|gb|AAA81328.1| ribulose-1,5-bisphosphate carboxylase small subunit [Glycine max]
tag62374	9822	4306	-2.48361635480317	0	Genoma - Gm11	18750733	+	0	gi|553111|gb|AAA33969.1| hydroproline-rich protein
tag22606	3267	384	-9.26353009657774	0	Genoma - Gm17	40197888	-	0	No hits found

Filter the results using a keyword:

Showing the results from 1 to 200 of 258

EST/GeneModel	Gene	AutoFact	
Contig3758	gi|221078703|gb|ACL99426.1| ATP synthase beta subunit [Diospyros ebenum]	ATP synthase subunit beta, chloroplastic n=535 Tax=root RepID=ATPB_ARATH	See the contigs of the solexa assembly
Contig7365	gi|69111452|gb|AAZ03388.1| dehydration responsive element-binding protein 3 [Glycine max]	Dehydration responsive element-binding protein 3 n=1 Tax=Glycine max RepID=Q4FH87_SOYBN	See the contigs of the solexa assembly
Contig16059	gi|224135661|ref|XP_002322129.1| calcium dependent protein kinase 28 [Populus trichocarpa] gi|222869125|gb|EEF06256.1| calcium dependent protein kinase 28 [Populus trichocarpa]	Calcium-dependent protein kinase 10 n=1 Tax=Arabidopsis thaliana RepID=CDPKA_ARATH	See the contigs of the solexa assembly
Contig9277	gi|10946375|gb|AAG24882.1|AF303939_1 ribulose-1,5-bisphosphate carboxylase small subunit rbcS1 [Glycine max] gi|1055368|gb|AAA81328.1| ribulose-1,5-bisphosphate carboxylase small subunit [Glycine max]	Ribulose bisphosphate carboxylase small chain 1, chloroplastic n=8 Tax=Glycine RepID=RBS1_SOYBN	See the contigs of the solexa assembly
Contig28935	gi|255627487|gb|ACU14088.1| unknown [Glycine max]	Rieske iron-sulphur protein n=1 Tax=Glycine max RepID=A8Y7Z9_SOYBN	See the contigs of the solexa assembly

Figure 4 - Web interfaces. (A) Results for the SuperSAGE analysis. For each unique tag are available: tag count in control and treated libraries (columns 3 and 4), fold-change (column 5), p-value (column 6), the correspondent gene (column 7), alignment information (columns 8, 9 and 10) and gene annotation (column 11). (B) Results for one subtractive library of the project, showing all genes found in the library and their respective annotation. The interface allows the user to search using a keyword in the annotation results.

these tags against *Phakopsora pachyrhizi* databases, but we did not find any fungus genes, probably due to the limited amount of fungus data available in the literature. The genome of this fungus, for example, has an estimated size of 500 Mb, but there is only 50 Mb available at *NCBI*.

We constructed a web interface for SuperSAGE analysis (Figure 4A). This interface shows, for each tag, the count number in both libraries (control and treated), the correspondent gene and its annotation (NR and Autofact result), as well as the position and the number of mismatches in the alignment. The user can filter the results using a keyword or gene name.

Solexa cDNA subtractive libraries data

Twenty-two cDNA subtractive libraries from different cultivars were sequenced in the Genosoja context, using many treatments with different time courses (Table 2) (Rodrigues *et al.*, 2012, this issue). The reads were generated by Illumina/Solexa technology with read lengths of 45 or 76 bp, depending on the library.

In order to identify the genes in these libraries, the reads were mapped into soybean genes. First, we aligned the sequences against the unigenes using the SOAP2 aligner configured to allow up to two mismatches, discarding fragments with “Ns”, and returning all optimal alignments. The sequences that did not align with unigenes were aligned against the predicted genes with the same parameters. A web interface (Figure 4B) provides users with all genes identified in each library and enables searches by gene name and keywords (in annotation results).

Solexa microRNA data

The Genosoja project generated eight small RNA libraries from soybean - four of the plants with Asian Rust disease (Brazilians cultivar PI561356 - resistant and Embrapa 48 - susceptible) and four under drought stress (Brazilians cultivars BR 16 - susceptible and Embrapa 48 - resistant) (Molina *et al.*, 2012, this issue). These libraries were sequenced using Illumina/Solexa technology and for each library the reads size range from 19 to 24 bp (Table 3).

Initially, the reads were grouped into unique sequences and read frequencies computed. The unique sequences that presented low read counts (read count = 2) were discarded from the list, as they were possibly caused by sequencing errors. In order to perform differential expression analysis between libraries, both a normalization

Table 2 - Summary of Solexa cDNA data from subtractive libraries deposited in the Genosoja databank.

	Genotype	Time course	Read length	Reads	Aligned reads (%)	Genes
Asian Rust	PI1356 - resistant	12, 24 and 48 h	76 bp	5,185,015	82.65	3,103
Asian Rust	PI1356 - resistant	72 and 96 h	76 bp	5,000,616	81.43	1,303
Asian Rust	PI1356 - resistant	192 h	76 bp	4,700,869	71.32	1,318
Asian Rust	PI230970 - resistant	1 and 6 h	76 bp	4,679,963	79.87	948
Asian Rust	PI230970 - resistant	12 and 24 h	76 bp	4,878,530	79.44	950
Asian Rust	PI230970 - resistant	48 and 72 h	76 bp	4,335,862	78.87	3,309
Virus	CD206 - resistant	5 and 13 days	76 bp	5,963,145	31.67	1,855
Virus	BRSGO - susceptible	6 and 13 days	76 bp	5,345,985	81.42	1,541
Nitrogen*	MG/BR 46	-	76 bp	4,621,072	75.11	6,815
Nitrogen*	MG/BR 46	-	76 bp	5,343,969	77.02	18,921
Drought - leaf	BR 16 - sensitive	25-50 min	45 bp	1,854,641	81.13	1,560
Drought - leaf	BR 16 - sensitive	75-100 min	45 bp	519,031	80.09	2,009
Drought - leaf	BR 16 - sensitive	125-150 min	45 bp	2,035,320	81.01	3,124
Drought - root	BR 16 - sensitive	25-50 min	45 bp	2,486,569	65.71	258
Drought - root	BR 16 - sensitive	75-100 min	45 bp	2,458,847	76.83	600
Drought - root	BR 16 - sensitive	125-150 min	45 bp	2,428,923	74.57	657
Drought - leaf	Embrapa 48 - tolerant	25-50 min	76 bp	5,144,645	79.66	10,495
Drought - leaf	Embrapa 48 - tolerant	75-100 min	76 bp	5,644,473	81,57	17,810
Drought - leaf	Embrapa 48 - tolerant	125-150 min	76 bp	5,359,395	80.53	8,970
Drought - root	Embrapa 48 - tolerant	25-50 min	76 bp	3,095,694	82.34	3,187
Drought - root	Embrapa 48 - tolerant	75-100 min	76 bp	5,731,156	74.72	17,218
Drought - root	Embrapa 48 - tolerant	125-150 min	76 bp	5,545,375	78.63	17,520

* Inoculated with *B. japonicum.*

Table 3 - Data from Solexa MicroRNA libraries deposited in the Genosoja databank.

			Sequence sizes					
			19 bp	20 bp	21 bp	22 bp	23 bp	24 bp
Drought stress	Resistant	Control	327,448	271,772	531,595	357,980	203,722	208,377
		Treated	71,011	72,628	154,808	87,326	77,045	177,626
	Susceptible	Control	89,040	91,816	215,419	128,524	142,446	200,087
		Treated	266,165	220,714	353,003	244,641	138,051	250,213
Asian Rust	Resistant	Control	91,908	205,404	1,177,303	394,378	175,063	285,064
		Treated	100,045	155,849	779,788	426,383	187,926	859,624
	Susceptible	Control	115,824	236,750	921,964	340,129	167,306	540,949
		Treated	123,423	190,799	962,676	363,983	86,230	176,753

and statistical significance analysis were applied using DEGseq software (Wang *et al.*, 2009) considering a confidence level of 95% (cutoff of 0.05). Table 4 presents the number of unique and differential sequences in each library. For the statistical significance analysis, the treated over control libraries were considered.

To identify microRNAs from the small RNAs dataset it is necessary to identify the pre-microRNA by alignment of small RNAs (unique sequences) into the soybean genome assembly, followed by secondary structure identification. This alignment was performed using SOAP2 configured to allow for exact alignments only. The upstream and downstream genomic sequences of the read alignment position, 300 bp each in size, were extracted from the genome using homemade PERL scripts (Supplementary Material Figure S2). These genomic regions were aligned against the reverse complement of its respective tag (rc-tag) using the Smith-Waterman (Smith and Waterman, 1981) algorithm with two gaps and four mismatches allowed. The resulting sequences were considered pre-microRNA candidates, and the secondary structure was manually curated, resulting in 256 microRNAs (Figure 5) (Kulcheski *et al.*, 2011).

Finally, the microRNA target prediction was performed using the Smith-Waterman algorithm (3 mismatches allowed) to align the 256 microRNAs against the assembled unigenes (shown previously). We considered only alignments in the 5'-3' direction obtained by comparison of the unigenes with the NR database using BLASTx. This methodology was able to identify targets for 169 microRNAs, most of which (39%) presented one or two targets (Figure 5).

Conclusions

In this work we presented all the bioinformatics analysis and pipelines used in the Genosoja project. The web-based interface constructed and described herein represents an important tool to help in the discovery of genes and new drugs that will enable increased soybean productivity. This system's use of common references (genome, assembled unigenes and predicted genes) facilitates the incorporation of new data from other sequencing methodologies or experimental conditions. Moreover, the bioinformatic pipeline discussed herein can also be applied to any genomic project, regardless of the organism.

Table 4 - Unique and differential sequences in microRNA libraries generated by the Genosoja consortium.

			Sequence sizes					
			19 bp	20 bp	21 bp	22 bp	23 bp	24 bp
Drought stress	Resistant	Unique	725	665	448	522	448	231
		Differential	79.30%	77.30%	78.10%	76.60%	76.80%	89.20%
	Susceptible	Unique	719	652	448	516	442	170
		Differential	75.50%	75.60%	88.00%	81.80%	82.10%	95.30%
Asian Rust	Resistant	Unique	588	524	427	456	386	208
		Differential	54.30%	60.90%	72.10%	55.50%	46.40%	57.70%
	Susceptible	Unique	590	537	435	461	373	220
		Differential	63.40%	57.90%	76.30%	69.00%	67.80%	73.20%

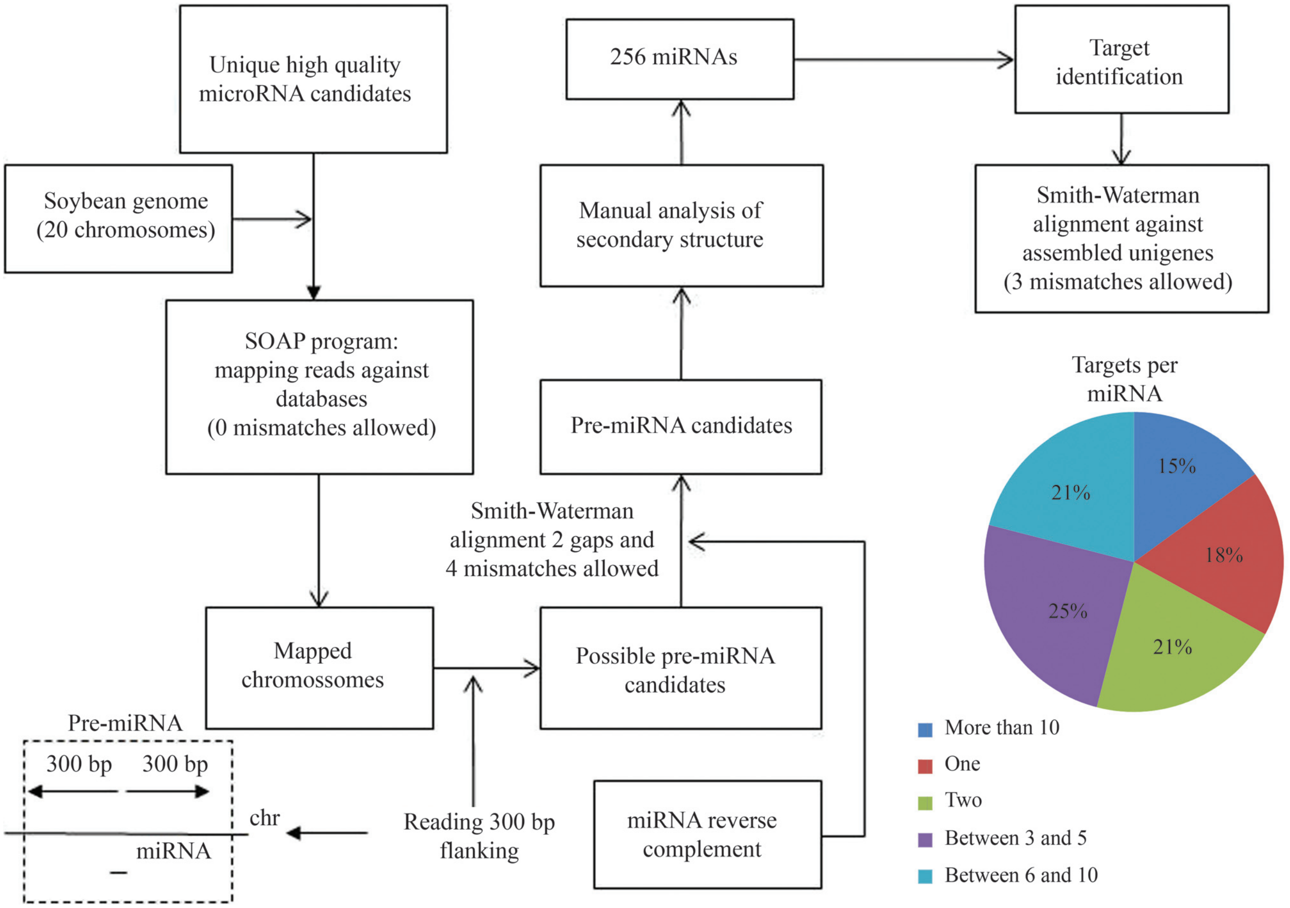

Unique high quality microRNA candidates
Soybean genome (20 chromosomes)
SOAP program: mapping reads against databases (0 mismatches allowed)
Mapped chromossomes
Pre-miRNA
300 bp
300 bp
chr
miRNA
Reading 300 bp flanking
Possible pre-miRNA candidates
miRNA reverse complement
Smith-Waterman alignment 2 gaps and 4 mismatches allowed
Pre-miRNA candidates
Manual analysis of secondary structure
256 miRNAs
Target identification
Smith-Waterman alignment against assembled unigenes (3 mismatches allowed)
Targets per miRNA
15%
18%
21%
25%
21%
🟦 More than 10
🟥 One
🟩 Two
🟪 Between 3 and 5
🟦 Between 6 and 10

Figure 5 - Pipeline used for microRNA prediction. The microRNA candidates (unique tags, Table 3) were mapped against the soybean genome with the SOAP2 aligner (no mismatches allowed). The genomic region (300 bp upstream and downstream of the alignment position) of the SOAP alignment was mapped with the reverse complement of the original microRNA using the Smith-Waterman algorithm. After manual analysis of the secondary structure, 256 microRNAs were identified. Finally, the Smith-Waterman algorithm (3 mismatches allowed) was used to identify targets in the assembled unigenes (it was possible to identify targets for 169 microRNAs).

Acknowledgments

The authors would like to acknowledge all research by the Brazilian Soybean Genome Consortium (Genosoja) involved in the generation of data used to construct the database presented in this paper. Moreover, we thank CNPQ (Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil) for financial support to this work.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Audic S and Claverie JM (1997) The significance of digital gene expression profiles. Genome Res 7:986-995.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Griffiths-Jones S, Howe KL, Marshall M and Sonnhammer ELL (2002) The Pfam protein families database. Nucleic Acids Res30:276-280.

Baudet C and Dias Z (2005) New EST trimming strategy in: Brazilian Symposium on Bioinformatics, 2005. Lect Notes Bioinf 3594:206-209.

Carazzolle MF, Formighieri EF, Digiampietri LA, Araujo MRR, Costa GLL and Pereira GAG (2007) Gene projects: A genome web tool for ongoing mining and annotation applied to CitEST. Genet Mol Biol 30(suppl):1030-1036.

Cheng KCK and Stromvik MV (2008) SoyXpress: A database for exploring the soybean transcriptome. BMC Genomics 9:e368.

Dowell RD, Jorkest RM, Day A, Eddy SR and Stein L (2001) The distributed annotation system. BMC Bioinformatics 2:e7.

Huang X and Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9:868-877.

Jenkinson AM, Albrecht M, Birney E, Blankenburg H, Down T, Finn RD, Hermjakib H, Hubbard TJP, Jimenez RC, Jones P, *et al.* (2008) Integrating biological data - The Distributed Annotation System. BMC Bioinformatics 9(Suppl 8):S3.

Kanehisa M and Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res28:27-30.

Koski LB, Gray LW, Lang BF and Burger G (2005) AutoFACT: An automatic functional annotation and classification tool. BMC Bioinformatics 6:e151.

Kulcheski FR, Oliveira LFV, Molina LG, Almerao MP, Rodrigues FA, Marcolino J, Barbosa JF, Stolf-Moreira R, Nepomuceno AL, Marcelino-Guimaraes FC, *et al.* (2011) Identification of novel soybean microRNAs involved in abiotic and biotic stress. BMC Genomics 12:e307.

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

Molina L, Cordenonsi G, Loss G, Oliveira LFV, Carvalho K, Kulcheski F and Margis R (2012) Metatranscriptomic analysis of small RNAs present in soybean deep sequencing libraries. Genet Mol Biol 35(suppl 1):292-303.

Rodrigues FA, Marcolino J, Carvalho JFC, Nascimento LC, Neumaier N, Farias JRB, Carazzolle MF, Marcelino FC and Nepomuceno AL (2012) Using subtractive libraries to prospect differentially expressed genes in soybean plants submitted to water deficit. Genet Mol Biol 35(suppl 1):304-314.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, *et al.* (2010) Genome sequence of the paleopolyploid soybean. Nature 463:178-183.

Smith TF and Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195-197.

Soares-Cavalcanti NM, Belarmino LC, Kido EA, Pandolfi V, Marcelino-Guimaraes FC, Rodrigues F, Pereira GAG and Benko-Iseppon AM (2012) Overall picture of expressed Heat Shock Factors in *Glycine max*, *Lotus japonicus* and *Medicago truncatula.* Genet Mol Biol 35(suppl 1): 247-259.

Suzek BE, Huang H, McGarvey P, Mazumber R and Wu CH (2007). Uniref: Comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23:1282-1288.

Umezawa T, Sakurai T, Totoki Y, Toyoda A, Seki M, Ishiwata A, Akiyama K, Kurotani A, Yoshida T, Mochida K, *et al.* (2008) Sequencing and analysis of approximately 40,000 soybean cDNA clones from a full-length-enriched cDNA library. DNA Res 15:333-346.

Wang L, Feng Z, Wang X, Wang X and Zhang X (2009) DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26:136-138.

Wanderley-Nogueira AC, Belarmino LC, Soares-Cavalcanti NM, Bezerra-Neto JP, Kido EA, Pandolfi V, Abdelnoor RV, Binneck E, Carazzole MF and Benko-Iseppon AM (2012) An overall evaluation of the Resistance (*R*) and Pathogenesis Related (*PR*) superfamilies in soybean, as compared with *Medicago* and *Arabidopsis*. Genet Mol Biol 35(suppl 1):260-271.

Yorinori JT, Paiva WM, Frederick RD, Costamilan LM and Bertagnolli PF (2005) Epidemics of soybean rust (Phakopsora pachyrhizi) in Brazil and Paraguay from 2001 to 2003. Plant Disease 89:675-677.

Internet Resources

NCBI, http://www.ncbi.nlm.nih.gov/ (October 10, 2011).

Phytozome, http://www.phytozome.net/soybean.php (October 10, 2011).

Soybean full-length cDNA database, http://rsoy.psc.riken.jp/ (October 10, 2011).

SOAP2 aligner, http://soap.genomics.org.cn/ (October 10, 2011).

Genosoja database, http://www.lge.ibi.unicamp.br/soybean/ (October 10, 2011).

SoyXpress database, http://soyxpress.agrenv.mcgill.ca (October 10, 2011).

Supplementary Material

The following online material is available for this article:

Figure S1 - Perl script to extract information about sequences from GenBank files.

Figure S2 - Perl script to extract the upstream and downstream genomic sequences of the read alignment position.

This material is available as part of the online article from http://www.scielo.br/gmb.

```
#! /usr/bin/perl -w

use Bio::SeqIO;

################################################################################
#####################
# Additional_file_1.pl
# author: Leandro Costa do Nascimento
# E-mails: leandro@lge.ibi.unicamp.br or l.costa.nascimento@gmail.com

# Article: A web-based bioinformatics interface applied to Genosoja Project: 
databases and pipelines
# Nascimento et al., 2011
# Bioinformatics - Genomics and Expression Laboratory (LGE) 
http://www.lge.ibi.unicamp.br
# GENOSOJA database: http://www.lge.ibi.unicamp.br/soja

# Usage: perl Additional_file_1.pl <folder_with_the_files> <gbk_file> 
<output_fasta>
# Bugs: Probably many! =D
################################################################################
#####################

### Global variables - Don't edit
#########################################################################
my $accession = "";
my $cultivar = "";
my $tissue = "";
my $cult = 0;
my $tis = 0;
my $conf = 0;
my $confirma = 0;
my $sequencia = "";
my $cria = 0;

my $dir = "";
my $gbk = "";
my $new = "";
my %cont;
################################################################################
###########################

### Parameters section
################################################################################
####
sub show_parameters{
   print "Usage: perl Additional_file_1.pl <folder_with_the_files> <gbk_file> 
<output_fasta>\n\n";
   exit(0);
}

($dir, $gbk, $new) = @ARGV;

if(!(defined($dir))){
   show_parameters();
}

if(!(defined($gbk))){
   show_parameters();
}

if(!(defined($new))){
   show_parameters();
}
################################################################################
###########################

open SEQ, "<$dir/$gbk";
```

```
while(<SEQ>){
   chomp;
   my $linha = $_;

   if(/ACCESSION\s+(.*)/){
      $accession = $1;
      $cultivar = "";
      $tissue = "";
      $sequencia = "";
      $cria = 0;
      $confirma = 0;
      $cult = 0;
      $tis = 0;
   }

   if(/^\s+\/cultivar\=\"([^\"]+)\"/){
      $cultivar = $1;
      $cultivar =~ s/\r//g;
      $cultivar =~ s/\n//g;

      if($cultivar ne ""){
         $cult = 1;
      }
   }

   if(/^\s+\/tissue\_type\=\"([^\"]+)\"/){
      $tissue = $1;
      $tissue =~ s/\r//g;
      $tissue =~ s/\n//g;

      if($tissue ne ""){
         $tis = 1;
      }
   }

   if(/^\/\//){
      $conf = 0;
      $cria = 1;
   }

   if($conf == 1){
      $linha =~ s/\s//g;
      $linha =~ s/\d//g;
      $linha =~ s/\n//g;
      $linha =~ s/\r//g;
      $sequencia = $sequencia . $linha;
   }

   if(/^ORIGIN\s+$/){
      $conf = 1;
   }

   if(($cria == 1) && ($cult == 1) && ($tis == 1)){
      open NEW, ">>$new";
         print NEW ">$accession cultivar:$cultivar tissue:$tissue\n";
         print NEW "$sequencia\n";
      close NEW;

      $accession = "";
      $cultivar = "";
      $tissue = "";
      $sequencia = "";
      $cria = 0;
      $confirma = 0;
      $cult = 0;
      $tis = 0;
   }
```

```
    }
close SEQ;
```

```
#! /usr/bin/perl -w
use Bio::SeqIO;

################################################################################
#####################
# Additional_file_2.pl
# author: Leandro Costa do Nascimento
# E-mails: leandro@lge.ibi.unicamp.br or l.costa.nascimento@gmail.com

# Article: A web-based bioinformatics interface applied to Genosoja Project:
databases and pipelines
# Nascimento et al., 2011
# Bioinformatics - Genomics and Expression Laboratory (LGE)
http://www.lge.ibi.unicamp.br
# GENOSOJA database: http://www.lge.ibi.unicamp.br/soja

# Usage:  perl Additional_file_2.pl <tags_file> <number_bases> <fasta_genome>
<soap_command>
# Bugs: Probably many! =D
################################################################################
#####################

### Global variables - Don't edit
#########################################################################
my $genome_file = "";
my $tags_file = "";
my $number_bases = 0;
my $soap_align_command = "";

my %alinhados = ();
my %ok = ();
my %cromossomo_13 = ();
my %alinhamentos_cromossomo_13 = ();
################################################################################
###########################

### Parameters section
################################################################################
####
sub show_parameters{
   print "Usage:  perl Additional_file_2.pl <tags_file> <number_bases>
<fasta_genome> <soap_command>\n\n";
   print "tags_file: file with the possible microRNAs in fastq format\n";
   print "number_bases: the script will get bases before and after the aligment
according to this parameter\n";
   exit(0);
}

($tags_file, $number_bases, $genome_file, $soap_align_command) = @ARGV;

if(!(defined($tags_file))){
   show_parameters();
}

if(!(defined($number_bases))){
   show_parameters();
}

if(!(defined($genome_file))){
   show_parameters();
}

if(!(defined($soap_align_command))){
   show_parameters();
}
################################################################################
###########################
```

```
### Edit this variables - if you want
#####################################################################
my $soap_file = "$tags_file\_X_genome.soap";
my $new = "$tags_file\_X_genome.fasta";
################################################################################
###########################

### Soap section
################################################################################
##########
print "Running the soap software to align the reads with the reference\n";
system("$soap_align_command -a $tags_file -D $genome_file.index -o $soap_file -r
2 -v 0");
################################################################################
###########################

### Searching for alignments
##############################################################################
print "Searching for alignments in the soap output file\n";
open FILE, "<$soap_file";
   while(<FILE>){
      chomp;
      my @linha = split(/\t/, $_);

      my $tag = $linha[0];
      my $sinal = $linha[6];
      my $referencia = $linha[7];
      my $position = $linha[8];

      if($referencia eq "Gm13"){
          if(defined($cromossomo_13{$tag})){
             $cromossomo_13{$tag}++;
          }
          else{
             $cromossomo_13{$tag} = 1;
          }

          if($cromossomo_13{$tag} >= 3){
             $alinhamentos_cromossomo_13{$tag} = "";
             next;
          }
          else{
             my $temp = $cromossomo_13{$tag};

             if(defined($ok{$tag})){
                $temp += $ok{$tag};
             }

             if(defined($alinhamentos_cromossomo_13{$tag})){
                $alinhamentos_cromossomo_13{$tag} .=
";$tag\_$temp,$position,$sinal";
             }
             else{
                $alinhamentos_cromossomo_13{$tag} =
"$tag\_$temp,$position,$sinal";
             }
          }
      }
      else{
         if(defined($ok{$tag})){
            $ok{$tag}++;
         }
         else{
            $ok{$tag} = 1;
         }

         if(defined($alinhados{$referencia})){
            $alinhados{$referencia} .= ";$tag\_$ok{$tag},$position,$sinal";
         }
         else{
            $alinhados{$referencia} = "$tag\_$ok{$tag},$position,$sinal";
         }
      }
   }
close FILE;

foreach(keys %alinhamentos_cromossomo_13){

   if($alinhamentos_cromossomo_13{$_} ne ""){
      if(defined($alinhados{Gm13})){
         $alinhados{Gm13} .= ";$alinhamentos_cromossomo_13{$_}";
      }
      else{
         $alinhados{Gm13} = "$alinhamentos_cromossomo_13{$_}";
      }
   }
}
################################################################################
###########################

### Getting the final sequences
###########################################################################
print "Getting the final sequences\n";
my $inseq = Bio::SeqIO-> new(-file   => "<$genome_file", -format => "fasta" );
while (my $seq = $inseq->next_seq){
   my $agora = $seq->display_id;

   if(defined($alinhados{$agora})){
      my @split = split(/\;/, $alinhados{$agora});

      # running in the separation of the tags by ";"
      foreach(@split){
         # running in the separation of the tag name and position by ","
         my @array = split(/\,/, $_);

         my $inicio = $array[1] - $number_bases;
         my $fim = $array[1] + $number_bases;
         my $tamanho = $seq->length;

         if($inicio < 1){
            $inicio = 1;
         }

         if($fim > $tamanho){
            $fim = $tamanho;
         }

         my $new_seq = $seq->subseq($inicio, $fim);
         open NEW, ">>$new";
            # tag reference:initial position in the reference.end position in
the reference alignment direction
            print NEW ">$array[0] $agora:$array[1] $array[2]\n";
            print NEW "$new_seq\n";
         close NEW;
      }
   }
}
################################################################################
############################
```