Sunflower Functional Genome Database, a curated unigene database to support functional diversity studies in sunflower


Background

SFGD is a secondary database integrating EST information from *Helianthus annuus*. Curated sequences were assembled into 41,013 sequences. Unigene annotation was inferred by sequence alignments, function description were integrated with the Gene Ontologies, and the Enzyme Commission identifiers for mapping onto metabolic pathways. *In silico* probe selection to custom design an oligonucleotide microarray and for the discovery of gene-rich SNPs were implemented as tools for functional diversity studies, including abiotic and biotic gene expression profiling and association studies in sunflower.

Methods

ESTs sequences were curated using EMBOSS suit [1], assembly was done using CAP3 program [2]. Unigene orientation was inferred using BLASTX against the RefSeq protein database [3]. Functional annotation was done using Blast2GO [4], Interpro [5] and KEGG [6]. For the custom Gene Expression design, eArray application was used [7]. Design parameter: Probe Length: 60bp, Probe per Target: 1. Masking of vector sequences, Probe Orientation: sense. Design Options: Best Probe Methodology with a 3´-end Bias, microarray format 4x44K. Eighty genes previously characterized were used as a probe group. SNP detection was performed running a custom pipeline with the CAP3 output [2] file as input. This pipeline finds positions covered by at least 2 distinct alleles in EST alignments with a minimum of 5 sequences, and a minimum of 50bp on either side of the SNP. The pipeline formats the SNP substitution and flanking region according to the requirements of the SequenceList file for Illumina genotyping technology. A relational MySQL database [8] is being designed in order to index, manage and store the functional genome data for sunflower.

Results

Sequences were assembled with CAP3 [2] into 28,089 singletons and 12,924 contigs. About 22,000 unigenes were annotated; a preliminary scan of the GO terms [9] and KEGG annotations [6] showed that the main biochemical pathways are represented. The microarray design comprises a total number of 45,220 features, with 1,417 Agilent controls, 80 probes 10 times replicated and 43,003 non-control features. Regarding polymorphism detection, 1283 SNPs in 695 contigs that fulfill GoldenGate Illumina requirements were found.

Future work

Information on the ESTs libraries, sequences, unigenes, functional annotations, microarrays probes and SNP will be accessible via a user-friendly web interface. Microarrays synthesis and Illumina Golden Gate assay are under way to be tested and validated as tools for transcriptomic and association analysis assessing biotic and abiotic stress responses over different *H. annuus* accessions.
Acknowledgements

This research was supported by CONICET-PIP 5788, ANPCyT/FONCYT, INTA-PPR AEBIO 241001 y 245001, INTA-PE AEBIO 245711 y 245732, ANPCyT-PAE 2007 PME-00024 y PID-00073.

References

6. **KEGG: The Kyoto Encyclopedia of Genes and Genomes.**
7. eArray (https://earray.chem.agilent.com/earray/)
8. MySQL (http://www.mysql.com/)