Plant Co-expression Annotation Tool: a tool to identify targets for proof of concept in Genetically Modified crop breeding pipelines

Marcos José Andrade Viana, Adhemar Zerlotini Neto, Mauricio de Alvarenga Mudadu

UFMG/EMBRAPA

Abstract

The development of Genetically Modified crops (GM) includes the discovery of candidate genes through bioinformatics analysis using genome data, biological pathways, gene expression, and others. Proteins of unknown function (PUFs) are interesting targets for proof of concept in GM crops breeding pipelines. Many PUFs are species specific and may participate in important biological pathways for organism survival. One way of inferring the function of PUFs (e.g.: relating them to factors of interest, like abiotic stresses), is through orthology and gene expression correlations by using co-expression networks. The purpose of this work is to characterize PUFs to be used in GM crops pipelines using orthology, co-expression networks and other tools. To date, we have downloaded, analyzed, and processed genomic data of 53 organisms from Phytozome, as well as the genome of the resurgent Boea Hygrometrica plant from NCBI, totaling 1, 862, 010 genes, 2, 332, 974 mRNA and 2, 332, 974 proteins. Diamond blast, against the NCBI's nr database, and InterproScan were used to discover 72, 266 PUFs for all organisms. PUFs were found by selecting proteins that have no matches within both diamond and interproscan searches. To construct the co-expression networks, RNA-seq data related to abiotic stress, like heat, drought, dehydration and osmotic stress were downloaded from the GEO / NCBI database: 16 samples from Glycine max, 14 from Zea mays 33 from Arabidopsis thaliana and 28 from Oryza sativa. This data was used to construct co-expression networks and clusters of transcripts with correlated expression using the LSTrAP software. Orthology was used to annotate the PUFs. In this regard, we have constructed 164, 267 orthologous groups using OrthoMCL software with the proteins from the 53 genomes. All the data obtained were stored in a database provided by the machado software (https://github.com/lmb-embrapa/machado). A web interface named "Plant Co-expression Annotation Tool" is under development to provide queries to mine PUFs from all 53 plant species. The tool provides analysis such as comparative functional annotation searches, expression values, biological pathways and ontologies. A search example was performed using the Oropetium thomaeum genome, a plant that is resistant to ressecation, and we found 10 PUFs correlated with abiotic stresses through orthology and co-expression networks. In summary, we believe our tool can be valuable for finding interesting targets to be used as proof of concept in GM crop breeding pipelines.

Funding: EMBRAPA/UFMG