

BIOINFORMATICS

Investigating Arabidopsis genes of unknown function by experimental and bioinformatic approaches

C.D. Town, B.A. Underwood, J.C. Redman, W.A. Moskal, Jr., H.C. Wu, W. Wang, E.L. Monaghan, H. Quan, J. Zhang, Y. Xiao

J. Craig Venter Institute, 9704 Medical Center Drive, Rockville MD 20850 USA

cdtown@jcv.org

keywords: RACE, novel genes, quantitative real time PCR (qPCR), reporter constructs, co-expression analysis

For several years, we have been investigating genes for which there was little or no experimental evidence supporting either their structure or function. At the conclusion of the whole genome annotation at TIGR, there were several thousand genes annotated as “hypothetical”, meaning that their annotation relied solely on *ab initio* predictions by gene-finding algorithms. We used 5’ and 3’ RACE to generate cDNA evidence in support of the structure and expression of ~1,500 of these genes and several hundred novel genes predicted by TwinScan and/or Eu’Gene resulting in the confirmation of some gene structures, the modification of others and the addition of new genes in subsequent versions of the annotation. The study also produced extensive evidence of alternate splicing. Of the 26,751 protein coding genes in TAIR6 annotation, over 8,000 have both molecular function and biological process annotated as “unknown”. Co-expression analysis provides a powerful tool to infer the function of unknown genes with those of known function from, although the method has not yet been applied to large-scale annotation in Arabidopsis. Publicly available datasets from Affymetrix ATH1 expression array combined with massively parallel signature sequencing have provided statistically significant expression values over diverse tissues for just over 22,000 distinct protein-coding genes. Because many of the genes of unknown function are expressed at levels that are too low to be effectively profiled by hybridization-based methods, the goal of our current NSF 2010 project is to generate expression profiles by quantitative real time PCR for 4,000+ Arabidopsis genes for which expression data are unavailable. To date, we have performed quantitative real time PCR on ~1,400 genes that either lack reliable expression data from or are not represented on the ATH1 array using cDNAs from leaf, root and T87 cell culture and seedlings treated with IAA, SA and salt. Over 90% of the genes were expressed in one of our current cDNA populations and ~ 30% of them showed differential expressions in at least 2 out of 6 conditions. Quantitating expression of genes expressed at low levels and in only some tissues has revealed an interesting pitfall of qPCR. Primer pairs that faithfully report transcript levels in some tissues have been found to mis-prime in RNA samples from which the transcript is absent. We have developed quality control protocols to discriminate between true and false signals with good accuracy. As more data accumulate, we will begin co-expression analysis with the public ATH1 data and assign these low-expressing genes of unknown function to known pathways and processes. In addition to the qPCR, we have developed a high throughput pipeline to generate promoter-reporter constructs for 1,000 low-expressing genes of unknown function and transform Arabidopsis plants. So far, promoters from 397 genes have been cloned into a GFP reporter construct, 288 have been transferred into Agrobacterium and 224 have been transformed into Arabidopsis. In many of these lines, GFP expression is localized to small regions of tissues and cell types. Both the qPCR and reporter construct data can be found at <http://www.tigr.org/tdb/e2k1/ath1/qpcr/index.shtml>.