

Whole-genome transcriptional analysis of Arabidopsis using massively-parallel signature sequencing (MPSS)

Blake Meyers

Delaware Biotechnology Institute, University of Delaware, 15 Innovation Way, Newark, DE 19711, USA.

We have generated a collection of 36,991,173 17-base sequence tags or “signatures” representing transcripts from the model plant Arabidopsis. These data were derived by massively parallel signature sequencing (MPSS) from 14 libraries and comprised 268,132 distinct sequences. For each library, comparable data were obtained with 20-base signatures. We developed a method for handling these data and for comparing these signatures to the annotated Arabidopsis genome. As part of this procedure, 858,019 potential or “genomic” signatures were extracted from the Arabidopsis genome and classified based on the position and orientation of the signatures relative to annotated genes. A comparison of genomic and expressed signatures matched 67,724 signatures predicted to be derived from distinct transcripts and expressed at significant levels. Expressed signatures were assigned to 19,088 of 29,084 annotated genes. A comparison of the representation of four-base words in the genomic and expression signatures demonstrated that ~7.7% of genomic signatures were under-represented in the expression data. These signatures contained one of 20 four-base words in either MPSS sequencing frame that did not sequence well. More than 89% of the sum of the expressed signature abundances matched the Arabidopsis genome, and many of the unmatched signatures found in high abundances were predicted to match to previously uncharacterized transcripts. We have developed a publicly available database and interface with which to view the MPSS transcriptional data and the genomic locations for these signatures (<http://mpss.udel.edu/at>). These data are the first large-scale quantitative expression data for plants in the public domain.