P602

Whole Genome Tiling Arrays: a New Tool for Genome Interpretation

Jun Lim

Department of Molecular Biotechnology, Konkuk University 143-701

Functional genomic analysis of any organism with a complete genome sequence requires accurate gene structure information and a complete gene inventory. Using the reference plant, *Arabidopsis thaliana*, a dual experimental strategy was employed for experimental verification of its genome annotation and construction of its ORFeome. The genome sequence of *Arabidopsis thaliana* serves as a reference for plants. The initial identification of transcriptional units in the *Arabidopsis* genome sequence was carried out largely by ab initio gene predictions, sequence homology, sequence motif analysis, and other non-experimental methods. This led to an estimate of 25,500 protein-coding genes. While gene prediction software has steadily improved, the ability of these programs to precisely determine gene structures in sequenced genomes remains unsatisfactory. A recent attempt to verify experimentally the accuracy of the *Arabidopsis* genome annotation with conventional molecular approaches proved quite inefficient in identifying full-length open-reading frame (ORFs). Global experimental approaches are expected to greatly improve genome annotation. A genome sequence that is empirically annotated can provide the foundation for the determination of its ORFeome, the complete set of ORF clones for all protein coding genes. Access to a "gold standard" cDNA/ORF clone collection, representing the entire *Arabidopsis* proteome, is urgently needed as a common resource for research. Here we report the experimental definition of the transcriptional units for all Arabidopsis genes by full-length cDNA discovery and by hybridization of RNA populations to whole genome arrays (WGAs) Dramatic improvements in genome annotation have been achieved for several organisms by means of sequences of fl-cDNAs. For Arabidopsis, three collections of fl-cDNAs have been used for this purpose. Of the 26,828 predicted genes in the *Arabidopsis* genome, 25,540 were annotated as protein-coding, with the rest being annotated as pseudo and partial genes. *Full-length (fl) cDNAs have been* sequenced and this information was used to correct the initial computational gene predictions in the genome sequence. However, even after extensive cDNA library development and ESTs sequencing, fl-cDNAs have been discovered for approximately two-thirds of the ~25,500 predicted *Arabidopsis* genes. Therefore, a second strategy was used to verify the accuracy of computational genome annotation. We next designed a set of 12 oligonucleotide arrays representing ~94% of the *Arabidopsis* genome sequence (110 Mb). Each array contains ~ 834,000 25-mer oligos. Four RNA populations were hybridized to these arrays, and a transcription map for the entire *Arabidopsis* genome was determined. In addition, we were also able to detect a large number (~54%) of transcriptionally active sites corresponding to the chromosomal locations where the newly discovered genes were found. Transcriptional activity was also detected in 2,000 intergenic regions (23%) that were thought to be devoid of transcriptional activity. The results show that the transcriptional activity across the chromosomes using four different mRNAs is quantitatively distinct. Using RNA populations from various tissues and whole genome high-density oligonuclotide tiling arrays, the transcriptional activity of individual chromosomes was examined. Chromosome and strand specific transcription was detected in various *Arabidopsis* tissue. This unbiased approach allows accurate determination of gene structures and identification of large number of novel transcription units. The dual strategy allowed the construction of approximately 30% of the *Arabidopsis* ORFeome. The data from genome tiling arrays have the potential to serve as a guide for the isolation of fl-cDNAs for each of the "untouched" Arabidopsis genes, and to allow the construction of the remaining plant ORFeome. More interestingly, approximately 7,600 annotated genes (~30% of all annotated genes) were identified that showed significant antisense RNA expression suggesting that double-strand RNA formation may be a general phenomenon in plant cells. We found complementary patterns of tissue-specific expression of sense and antisense RNAs for many genes, indicating a possible biological role for these transcripts. In our database search and comparison to the tiling chip results, we found that approximately 480 candidates of natural antisense transcripts (NATs) in *Arabidopsis*. Of approximately 480 NATs, 141 appeared to be conserved in rice, which suggest the conservation of their biological functions. Studies for expression and function of 141 NATs are in progress. In particular, T-DNA insertional mutations in putative promoter regions for NATs are under extensive investigation to elucidate the molecular mechanisms of gene expression regulated by NATs.