Techniques & Applications

# An integrated *Arabidopsis* annotation database for Affymetrix Genechip® data analysis, and tools for regulatory motif searches

Majid Ghassemian, David Waner, Jason Tchieu, Michael Gribskov and Julian I. Schroeder

**Genome-scale sequencing projects have provided the essential information required for the construction of entire genome chips or microarrays for RNA expression studies. The *Arabidopsis* and rice genomes have been sequenced and whole-genome oligonucleotide arrays are being manufactured. These should soon become available to researchers. Expression studies using genomic-scale expression arrays are providing us with a vast quantity of information at a rapid pace. The rate-limiting step in this type of experiments is not the data generation step but rather the data analysis component of experiments. We report improvements that should facilitate the analysis of Affymetrix Genechip® expression data.**

The recent availability of *Arabidopsis* oligonucleotide arrays from Affymetrix has provided many plant laboratories with a wealth of expression data that can be used to address important biological questions. One of the interesting types of analysis is the identification of novel *cis*-elements that regulate the expression of genes in response to various experimental treatments. By identifying subsets of genes that have a common expression profile, it should be possible to identify conserved motifs in the promoter or other regions. When we attempted this type of analysis for Genechip® experiments performed in our laboratory, we found that there were several practical obstacles to be overcome.

## Obstacles

First, the *Arabidopsis* annotation file provided by Affymetrix only contains the GenBank identifiers, and it can be time consuming to search for each gene individually in GenBank and to manually extract the promoter sequence from the bacterial artificial clone (BAC) sequence entry. This task is easier with the Munich Information Center for Protein Sequences (MIPS) *Arabidopsis thaliana* Database (MATDB) (http://mips.gsf.de/proj/thal/db/index.html) than with other databases because the MATDB entries include 500 bp of sequence upstream and downstream of the predicted coding sequence. However, the MATDB cannot be searched using GenBank identifiers so it would be useful to have a list of the standard *Arabidopsis* Genome Initiative (AGI) identifiers for each probe on the chip.

The second problem was that because the genome databases are being constantly improved, we found that the original Affymetrix annotation table was outdated. Although it is possible to look up the current annotation in online databases for a small set of interesting genes, it would be better to have an updated table of annotations.

## Solutions

To address these problems, and to streamline the process of analyzing Genechip® data, we are making the following resources available to the research community:

- A table of AGI identifiers corresponding to each Affymetrix Probe set. These are now the standard identifiers for *Arabidopsis* genes and proteins.
- A table of gene, protein and promoter sequences for each probe on the chip (except those that could not be unambiguously identified).
- Web-based software to generate a list of promoter sequences in FASTA format given a set of Affymetrix probes or AGI gene identifiers.
- *Arabidopsis* support files for Genespring™, enabling plant researchers to display expression data on chromosome maps, which can aid in expression analysis, in searches for DNA motifs, and in map-based identification of potential genes in disruption mutants.

To determine the AGI protein identifier corresponding to each probe, Perl scripts were written (D. Waner and J. Tchieu, unpublished) to extract protein identifiers from the annotation file and to find the corresponding protein entries in the GenBank protein database. The MATDB database was searched to find matches for the protein sequences from GenBank to find the corresponding AGI identifiers for most of the probes. In cases where a perfect protein sequence match was not found, corresponding genes were found by BLAST (basic local alignment search tool) search, followed by manual checking of the alignments. Cases where there were multiple high-scoring matches in the MIPS database were resolved by manually comparing the MIPS gene annotations with the annotations in the original Affymetrix probe file. As an additional quality control, we BLASTed each Affymetrix probe sequence against the corresponding nucleotide entry in our database. Entries with discrepancies that could not be resolved were removed from our tables. Note that all our tables ultimately derive from the annotations, not from the probe sequences themselves, therefore the accuracy of our information is dependent on the correct identification of the target genes by their assigned annotations. In spite of the above precautions, errors are likely to exist because of possible errors in the design of the Genechip® and existing errors in database annotations.

We have prepared spreadsheet files giving the AGI identifiers; predicted protein sequence, nucleotide sequence and 500 bp of upstream genomic sequence for each of the Affymetrix probe sets (except those for which sequences could not be found). These files are available for downloading at the following web site (www-biology.ucsd.edu/labs/schroeder/genechip.html).

In addition, to facilitate regulatory motif searches, we have created a web-based tool for obtaining promoter sequences for a given list of Affymetrix probe IDs or AGI identifiers. By uploading a text file containing a list of Affymetrix probe identifiers, and specifying a desired sequence length, the server will return a list of promoter sequences in FASTA format. The page also includes an option

 1360-1385/01/$ – see front matter  PII: S1360-1385(01)02092-1

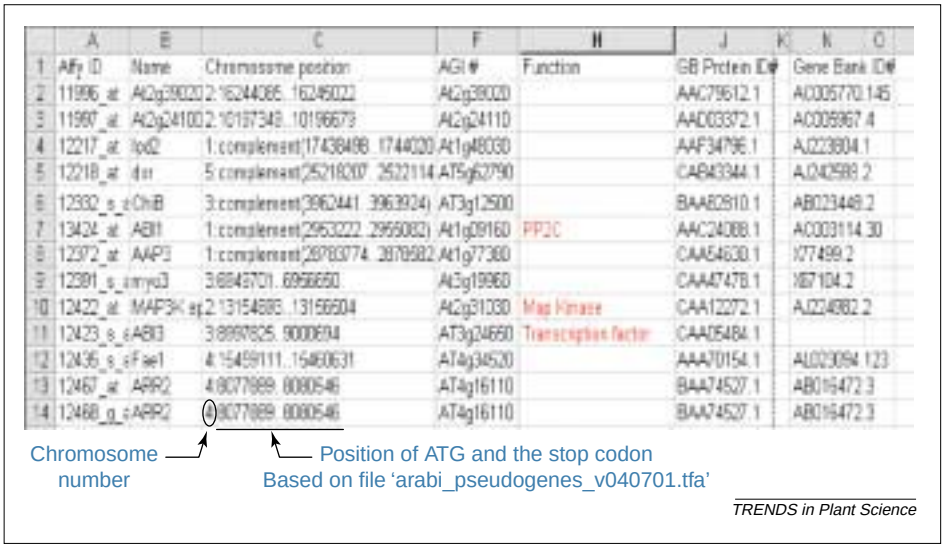


Fig. 1. Contents of the Genespring™ index file (displayed on a Microsoft™ Excel spreadsheet). Column A corresponds to the Affymetrix probe identifiers. Column B represents common gene names. Column C identifies the position of each open-reading-frame in reference to the *Arabidopsis* chromosomes. Column F is the AGI or MIPS identifier for corresponding probes. Column J shows the GenBank protein identifier. Column N shows the GenBank DNA identifier.

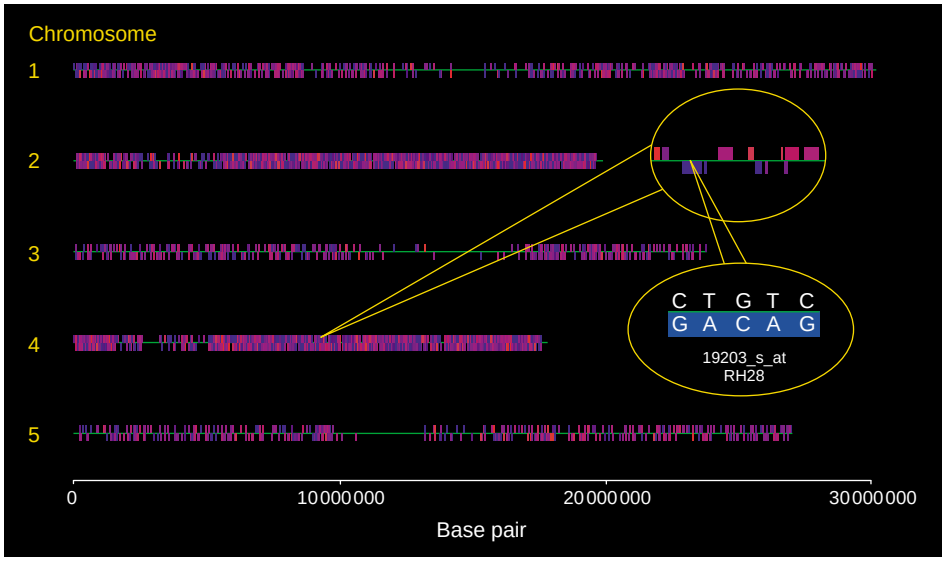


Fig. 2. Physical position of Affymetrix *Arabidopsis* Genechip® probes at single base pair resolution. When the coding region is on the forward strand the gene marker is displayed on top of the line, and when the coding region is on the reverse strand the gene marker is placed below the line. The color and the intensity of the color are representative of expression levels of a specific gene with reference to a control condition. Up-regulation is indicated in red and down regulation of expression is indicated in blue.

to automatically send the sequences to the MEME/MAST server at the San Diego Supercomputer Center. MEME and MAST are motif discovery and search tools that help in the identification of common motifs in sets of unaligned sequences[1,2]. These tools are described at the MEME website (http://meme.sdsc.edu/).

For users of the Genespring™ analysis software, we have created a 'Genome File' for *Arabidopsis*. This consists of a file containing the assembled genomic sequence for each of the five chromosomes, and an index file giving the position and orientation of each gene (Fig. 1, column C) based on the MIPS *Arabidopsis* chromosome sequence file of 'arabi_pseudogenes_v040701.tfa'. This enables the user to display data from Genechip® expression profiles on chromosome maps, with color-coding indicating the expression levels derived from experiments (Fig. 2). In addition, built-in tools can be used for analysis of regulatory motifs. The Genome File could be useful for applications such as the identification of candidate gene deletions after initial mapping of mutations to chromosomal regions, by searching for genes in the mapped region that are expressed in the wild type and absent in the mutant lines. These features were previously unavailable for *Arabidopsis* Genechip® users. Genespring™ can also link these sequences to external web sites for motif analysis. The Genespring™ files are available for download at the Schroeder Laboratory web site (www-biology.ucsd.edu/labs/schroeder/genechip.html). It should be noted that because the DNA sequence to the *Arabidopsis* chromosomes are constantly updated, the position of genes on chromosomes might vary based on the reference sequence file used.

Additional information (as it becomes available) can be added by users of the index file to enhance its effectiveness in data mining. Presently, we are trying to integrate cluster trees based on protein sequence homologies for all *Arabidopsis* proteins into this list. This tree assignment function for probes should provide clues to biological function of unknown genes based on close homologies to known ones.

**Acknowledgements**

M.Gh. and D.W. contributed equally to this work. This work was supported by an NSF genome grant (DBI-077379) to J.I.S. and M.Gr. and NIEHS (1P42ES10337), NIH (RO1 GM60396-01) and TMRI Syngenta-UC Biostar grants to J.I.S., and NSERC PDF to M.Gh. Protein identifiers from the annotation file kindly provided by Syngenta Corporation, www.tmri.org/gene_exp_web. M.Gr. is a member of the San Diego Supercomputer Center University of California and J.T. is affiliated with San Diego Supercomputer Center University and the Division of Biology, University of California, San Diego.

**References:**

1 Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28–36, AAAI Press, Menlo Park, CA, USA

2 Bailey, T.L. and Gribskov, M. (1998) Combining evidence using *p*-values: application to sequence homology searches. *Bioinformatics* 14, 48–54

**Majid Ghassemian***
**David Waner**
**Jason Tchieu**
**Michael Gribskov**
**Julian I. Schroeder**
Division of Biology, Cell and Developmental Biology Section and Center for Molecular Genetics, University of California, San Diego, La Jolla, CA 92093-0116, USA.
*e-mail: majidg@biomail.ucsd.edu