Supplemental File METHODS

MICROARRAY EXPRESSION PROFILE ANALYSIS OF THE LOW OXYGEN RESPONSE IN *ARABIDOPSIS* ROOT CULTURES

Erik Jan Klok, Iain W. Wilson, Dale Wilson, Scott C. Chapman, Rob M. Ewing, Shauna C. Somerville, W. James Peacock, Rudy Dolferus, Elizabeth S. Dennis

PATTERN ANALYSIS OF DATA

In plant breeding research, pattern analysis commonly refers to the complementary use of clustering and principal component analysis (PCA; Basford and Cooper, 1998). There is a clear mathematical association between the Euclidean distance measures determined in clustering and the correlation matrix utilised in PCA (Cooper and DeLacy, 1994), which can be utilised in interpreting the results.

Processing via the tRMA software (accessible at www.pi.csiro.au/gena/trma) resulted in a matrix of genes (rows) by treatments (columns), where the gene expression ratios had been corrected for spatial variation and standardised across treatments. This data matrix was processed by both cluster and principal component analysis using R scripts that utilised publicly available statistical libraries (Chapman et al., 2001). While clustering was used to group genes, we present here only the clustering of treatments.

A dissimilarity matrix was computed from the data matrix, using squared Euclidean distance as the dissimilarity measure. Hierarchical agglomerative clustering was applied to the matrix using minimisation of the within group sums of squares (Ward, 1963) as the fusion criterion. In clustering genes, the number of clusters to be inspected was derived from examination of the sums of squares explained as the number of cluster increased. To cluster treatments as in the dendrogram in the paper, the dissimilarity matrix was transposed prior to repeating the clustering. The data matrix was also subjected to principal component analysis with the results presented in a biplot (Gabriel, 1971).

Bi-plots incorporate the gene effects (scores) as points, and the treatments (loadings) as vectors (Gabriel, 1971; Chapman et al., 2001). Vectors that are close together are highly correlated, in terms of the gene effects observed for each treatment, while vectors that are orthogonal are poorly correlated. Points that are near the origin of the bi-plot are either close to the 'average' expression level for all treatments or are poorly explained by the PCA. Points (genes) that are close to the head of a vector have high positive expression values in that treatment, while genes that are on the opposite side of the origin, relative to the head of the vector, have negative expression values for that treatment. The relative expression level of any combination of gene and treatment can be determined by a perpendicular projection of a point onto a vector.

MOTIF SEARCHES

The 5' gene regions of the clustered genes were retrieved by performing BLASTN queries of the respective cDNA clones against the complete *Arabidopsis* genome sequence. The sequences of the 5' regions (up to 2000 bp) were used to obtain shared motifs by finding common short sequences (6-8 bp) that are over-represented in the 5' regions within a gene cluster, compared to all genes outside the cluster.

In more detail, clones on the microarray were mapped to their cognate genomic open reading frames using the respective ESTs as BLASTN queries against the complete *Arabidopsis* genome sequence. Summary information of these BLASTN queries is available as a flat file from the TAIR ftp site (http://www.*Arabidopsis*.org). Essentially, we took ESTs corresponding to the clones indicated and for each one used BLASTN against the genome sequence. Those ESTs matching an annotated ORF were 'assigned' to that ORF.

Two different sets of upstream sequences corresponding to putative promoter regions were constructed as follows. Genome annotation data was retrieved from the TIGR ftp site (ftp://www.tigr.org/pub/data/a_thaliana/). For the fixed-length 500 bp sequence set, an upstream 500 bp sequence was extracted for each annotated 'transcription unit' in the genome (from -500 to -1 relative to the translation start codon). For the 2,000 bp set of sequences, an upstream sequence was extracted with a minimum length of 50 bp and a maximum length of 2,000 bp, also directly upstream from the translation start codon. The set of 500 bp fragments can be queried via the web site of the *Arabidopsis* Functional Genomics Consortium (AFGC; http://afgc.stanford.edu/afgc_html/site2.htm).

The set of 500 bp sequences was created as follows. The sequences were 'quality-controlled' in order to determine their exact location (i.e., directly upstream from the ATG). First, fragments for which the annotated start coordinate did not correspond to 'ATG' or for which the annotated stop codon did not correspond to a stop codon were rejected. Second, BLASTX (DNA translated in 3 frames vs. protein) querying the set of 500 bp fragments against all predicted *Arabidopsis* proteins was used to exclude those fragments matching an upstream ORF (presumably cases for which the upstream intergenic sequence for a given ORF is less than 500 bp long). Finally, the set of 500 bp sequences was BLASTX queried against non-*Arabidopsis* protein sequences (NCBI) to exclude sequences corresponding to as yet unannotated ORFs in the *Arabidopsis* genome. The 500 bp sequence set consists of 21,038 upstream fragments and the 2,000 bp set of 25,459 fragments.

REFERENCES

Basford, K.E. and Cooper, M. (1998). Genotype x environment interactions and some considerations of their implications for wheat breeding in Australia. Aust. J. Agric. Res. **49**, 153-174.

Chapman, S.C., Schenk, P., Kazan, K., Manners, J.M. (2001). Using biplots to interpret gene expression patterns in plants. Bioinformatics (in press).

Cooper, M., and DeLacy, I. H. (1994). Relationships among analytical methods used to study genotypic variation andgenotype-by-environment interaction in plant breeding multi-environment experiments. Theor. Appl. Genet. **88**, 561-572.

Gabriel, K.R. (1971) The biplot graphic display of matrices with application to principal component analysis. Biometrika **58**, 453-467.

Ward, J.H. (1963) Hierarchical grouping to optimize an objective function. J. Amer. Stat. Assoc. **58**, 236-244.