Local Coexpression Domains of Two to Four Genes in the Genome of Arabidopsis[1][w]

Xin-Ying Ren, Mark W.E.J. Fiers, Willem J. Stiekema, and Jan-Peter Nap*

Applied Bioinformatics, Plant Research International, NL–6700 AA Wageningen, The Netherlands (X.-Y.R., M.W.E.J.F., J.-P.N.); Laboratory of Molecular Biology, Plant Sciences Group, Wageningen University and Research Centre, NL–6703 HA Wageningen, The Netherlands (X.-Y.R., W.J.S.); and Centre for BioSystems Genomics, NL–6708 PB, Wageningen, The Netherlands (W.J.S., J.-P.N.)

Expression of genes in eukaryotic genomes is known to cluster, but cluster size is generally loosely defined and highly variable. We have here taken a very strict definition of cluster as sets of physically adjacent genes that are highly coexpressed and form so-called local coexpression domains. The Arabidopsis (*Arabidopsis thaliana*) genome was analyzed for the presence of such local coexpression domains to elucidate its functional characteristics. We used expression data sets that cover different experimental conditions, organs, tissues, and cells from the Massively Parallel Signature Sequencing repository and microarray data (Affymetrix) from a detailed root analysis. With these expression data, we identified 689 and 1,481 local coexpression domains, respectively, consisting of two to four genes with a pairwise Pearson's correlation coefficient larger than 0.7. This number is approximately 1- to 5-fold higher than the numbers expected by chance. A small (5%–10%) yet significant fraction of genes in the Arabidopsis genome is therefore organized into local coexpression domains. These local coexpression domains were distributed over the genome. Genes in such local domains were for the major part not categorized in the same functional category (GOslim). Neither tandemly duplicated genes nor shared promoter sequence nor gene distance explained the occurrence of coexpression of genes in such chromosomal domains. This indicates that other parameters in genes or gene positions are important to establish coexpression in local domains of Arabidopsis chromosomes.

The combination of DNA sequence and expression data has revealed the existence of chromosomal domains of similarly expressed genes in several genomes, such as in yeast (*Saccharomyces cerevisiae*; Cohen et al., 2000), fly (*Drosophila melanogaster*; Spellman and Rubin, 2002), worm (*Caenohabditis elegans*; Roy et al., 2002; Lercher et al., 2003), human (*Homo sapiens*; Caron et al., 2001; Lercher et al., 2002; Versteeg et al., 2003), and more recently in the genome of the plant Arabidopsis (*Arabidopsis thaliana*; Birnbaum et al., 2003; Williams and Bowles, 2004). These analyses have focused on coexpression (Cohen et al., 2000; Spellman and Rubin, 2002; Lercher et al., 2003; Williams and Bowles, 2004), high expression (Caron et al., 2001), and so-called localized expression domains (Birnbaum et al., 2003), defined as spatial and/or temporal chromosomal domains of coordinated induction and repression in gene expression. Chromosomal domains (or clusters or regions) of similarly expressed (or coexpressed or coregulated or correlated) genes have been identified using sliding windows of either a given sequence length (number of nucleotides; Lercher et al., 2002) or of a given number of genes (Spellman and Rubin, 2002; Lercher et al., 2003; Williams and Bowles, 2004). Major experimental differences exist in the size of the window used for analysis and therefore the fraction of the genome evaluated as chromosomal domain. To determine the similarity between different expression profiles, the Pearson's correlation coefficient (R) was used and the average of all pairwise Rs over the expression values across experiments or tissues was evaluated over all windows and chromosomes (Cohen et al., 2000; Spellman and Rubin, 2002; Lercher et al., 2003).

In such genome-wide analyses, four different types of gene organization may account for high coexpression without giving evidence for the presence of chromosomal domains. These four types are: (1) overlapping genes (Cohen et al., 2000), (2) tandemly duplicated genes, (3) homologous genes (Spellman and Rubin, 2002; Lercher et al., 2003), or (4) genes in the same operon (Roy et al., 2002; Lercher et al., 2003). Generally, these four gene configurations have been analyzed separately for their contribution to coexpression. The remaining genes, if coexpressed, might be an indication of the existence of chromosomal domains. Housekeeping genes (Lercher et al., 2002; Roy et al., 2002; Lercher et al., 2003), genes with similar functions in different biological processes (Cohen et al., 2000; Spellman and Rubin, 2002), genes involved in the same metabolic pathway (Birnbaum et al., 2003), or

[1] This work was supported by the Dutch Organization for Scientific Research (in the framework of the project Wageningen Phytoinformatics: The Added Value from Plants) and by the Centre of BioSystems Genomics, which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research.

* Corresponding author; e-mail janpeter.nap@wur.nl; fax 31–317–418094.

[w] The online version of this article contains Web-only data.

Article, publication date, and citation information can be found at www.plantphysiol.org/cgi/doi/10.1104/pp.104.055673.

genes involved in the same biological process (Williams and Bowles, 2004) have all been identified in these chromosomal domains. Therefore, there does not appear to be a clear functional classification of genes present in such chromosomal domains.

The molecular mechanisms responsible for coordinated expression of neighboring genes are not well understood (Hurst et al., 2004). Coexpressed adjacent genes in yeast could not be explained solely by upstream activating sequences and are not due to divergently transcribed promoter regions, although the extent of physical vicinity seems to be important (Cohen et al., 2000). In worm, coexpression of genes could not be attributed to unrecognized operons or read-through transcription (Roy et al., 2002). Neither gene duplication nor common functionality was identified as the main cause for coexpression of neighboring genes in the Arabidopsis genome (Williams and Bowles, 2004). It is generally assumed that the coordinated expression of genes in chromosomal domains represents gene regulation at the level of specialized chromatin and chromosome structure. In Arabidopsis, limited chromosomal clustering of coregulated genes associated genome organization with gene regulation (Birnbaum et al., 2003). Analyses of the phenomenon in transgenic plants also indicated the importance of chromosomal context for proper gene expression (Mlynarova et al., 1994, 1995).

We here present the identification and analysis of local coexpression domains in the Arabidopsis genome. Local coexpression domains are here defined as chromosomal regions where physically adjacent genes have high correlated expression across all experiments. This definition focuses on the behavior of neighboring genes. Using the Munich Information Center for Protein Sequences (MIPS) Arabidopsis genome annotation (Schoof et al., 2002) and two types of whole-genome expression data, Massively Parallel Signature Sequencing (MPSS; Meyers et al., 2004) and an Affymetrix microarray (MA; Birnbaum et al., 2003), we have analyzed the coexpression of neighboring genes to identify local coexpression domains. Our results contrast with the genome-wide identification of more global coexpression domains, consisting of clusters up to 20 genes with a median cluster size of 100 kb. In such domains, coexpression was defined as a significant deviation from the averaged *R* (Williams and Bowles, 2004). This difference underlines the importance of distinguishing the size dimension of the chromosomal domains considered.

RESULTS

Chromosomal Coexpression Maps Reveal Local Coexpression Domains

The combination of the MIPS annotation of the Arabidopsis genome with the available MPSS and MA data resulted in a collection of 16,144 gene pairs with MPSS expression values and 18,443 pairs with MA expression values that could be analyzed. A more detailed description of the data sets generated is given in Table I and in "Materials and Methods." For visualization purposes, the overall whole-genome coexpression data were plotted in chromosomal coexpression maps as introduced by Cohen et al. (2000) for each chromosome of Arabidopsis. Figure 1 shows the chromosomal coexpression map of an area of 80 genes on chromosome 1 for which both MPSS (Fig. 1A) and MA (Fig. 1B) data were available. Genes with positively correlated expression are indicated in green. Genes that have correlated expression and are physically close together form green regions along the diagonal of the chromosomal coexpression map. Examples of such regions are indicated with a blue box (Fig. 1, A and B). Comparison of the same genomic regions in the MPSS (Fig. 1A) and MA (Fig. 1B) data show that regions can have different coexpression patterns in different data sets. Subsets of neighboring genes having high coexpression in the MPSS data set (Fig. 1A, blue box) showed low coexpression in the MA data set (Fig. 1B, yellow box), while subsets of

Table I. *Description of expression data used for whole-genome analysis*

	MPSS	MA
Genes with Expression		
Total	20,041	21,940
Overlapping	39	34
Without expressed neighbor(s)	851	651
Represented in pairs	19,151	21,255
Adjacent Pairs		
Total	16,144	18,443
Tandemly duplicated pairs	1,928 (11.9%)[a]	2,278 (12.4%)[a]
Coexpressed	905 (5.6%)[b]	1,800 (9.8%)[b]
Total excluding tandemly duplicated	14,216	16,165
Coexpressed excluding tandemly duplicated	689 (4.8%)[c]	1,481 (9.2%)[c]
Coexpressed Adjacent Pairs		
Total	905	1,800
Tandemly duplicated pairs	216 (23.9%)[d]	319 (17.7%)[d]
Tandemly Duplicated Pairs		
Total	1,928	2,278
Coexpressed	216 (11.2%)[e]	319 (14.0%)[e]

[a]The percentage of tandem duplicated pairs relative to the total number of adjacent pairs. [b]The percentage of coexpressed adjacent pairs relative to the total number of adjacent pairs. [c]The percentage of coexpressed adjacent pairs excluding tandemly duplicated relative to the total number of adjacent pairs excluding tandemly duplicated pairs. [d]The percentage of coexpressed tandemly duplicated pairs relative to the total number of coexpressed adjacent pairs. [e]The percentage of coexpressed tandemly duplicated pairs relative to the total number of tandem duplicated pairs.

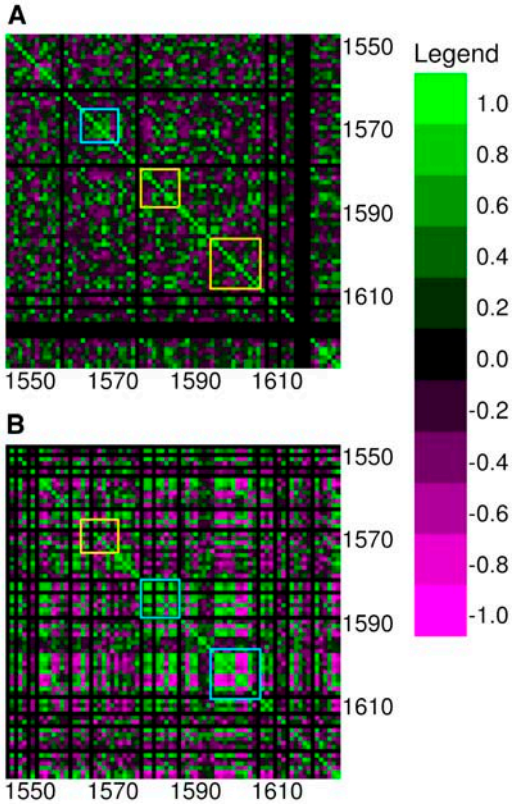

Figure 1. Chromosomal coexpression map of the Arabidopsis genome. The expression of each gene is correlated with all other genes on the same chromosome using a color coded representation of *R*. Green is positive correlation ($R > 0$), magenta is anticorrelation ($R < 0$), and black shows no correlation ($R = 0$), no expression, or missing data. A, Coexpression map of a small part of chromosome 1 using MPSS expression data, showing the 80 genes from At1g16240 (rank ID 1550) to At1g17090 (rank ID 1630). B, Coexpression map of the same 80 genes on chromosome 1 using MA expression data. The blue boxes in A and B indicate regions of blocks of coexpressed adjacent genes. The yellow boxes in A and B indicate the equivalent regions in the other data set.

neighboring genes in MA having high coexpression (Fig. 1B, blue boxes) have low coexpression in the MPSS data set (Fig. 1A, yellow boxes). Such differences in coexpression are likely to reflect the biological differences between the MPSS and MA data sets, although it cannot be fully excluded that technical differences between whole-genome expression profiling with MPSS and Affymetrix MA have also contributed in part to the differences observed. The MPSS data cover plant tissues and organs, while the MA data refer to defined root cells. The averaged expression over the biological material sampled in a data set may influence coexpression patterns of neighboring genes.

The number of local coexpression domains in the Arabidopsis genome and the number of genes involved were calculated. Two genes were considered to be adjacent, so present in a local coexpression domain, if their rank IDs (see "Materials and Methods") were consecutive with a difference of one and their pairwise *R* exceeded 0.7. Notably tandemly duplicated genes are known to influence coexpression statistics (Zhu, 2003; Hurst et al., 2004). Therefore, the occurrence of tandemly duplicated pairs was determined with pairwise protein BLAST using a cutoff of $E < 2 \times 10^{-1}$ (Lercher et al., 2002, 2003; Fukuoka et al., 2004; Williams and Bowles, 2004). This criterion has a false error rate of about 10% (Lercher et al., 2002; Williams and Bowles, 2004). In both the MPSS and MA expression data sets, only about 12% of all adjacent pairs (1,928 for MPSS and 2,278 for MA) are tandemly duplicated (Table I), of which only 11% to 14% are coexpressed (216 for MPSS and 319 for MA). This implies that in either expression data set only approximately 20% of the coexpressed pairs consist of tandemly duplicated genes. Only a minority of 11% to 14% of all tandemly duplicated gene pairs in the Arabidopsis genome are coexpressed (with $R > 0.7$), reflecting gene divergence after duplication (Williams and Bowles, 2004). As about 5% to 9% from all adjacent pairs excluding the tandemly duplicated pairs (689 for MPSS and 1,481 for MA) are coexpressed, a tandemly duplicated pair is about 2-fold (that is, 11%–14% relative to 5%–9%) more likely to be coexpressed than a nontandemly duplicated adjacent pair. Further analyses of the subpopulation of tandemly duplicated gene pairs do not indicate that a particular transcriptional orientation of the tandemly duplicated genes has a significantly higher inclination to be coexpressed (data not shown). In subsequent analyses, the subpopulations were evaluated with and without tandemly duplicated genes. The results are summarized in Table II. Depending on the expression data set considered, 5% to 9% of all nonduplicated gene pairs consist of coexpressed neighboring pairs (689 for MPSS and 1,481 for MA). These pairs tend to be spread throughout the genome (Fig. 2; MPSS data). The MA data set gave similar results (data not shown). Only 58 coexpressed pairs were common between the MPSS and MA sets out of 11,144 total common pairs (excluding tandemly duplicated pairs). These common coexpressed pairs are also widespread throughout the genome (Fig. 2).

In addition to the number of coexpressed gene pairs (duplets), the number of coexpressed triplets, quadruplets, and pentaplets in the Arabidopsis genome was determined (Table II; Fig. 2) using the strict criterion of highly correlated expression ($R > 0.7$) of all members in a multiplet. Triplet and quadruplet coexpression domains were considerably rarer (Table II), whereas coexpressed pentaplets did not occur in either the MPSS or the MA data set. To evaluate the significance of the observed numbers of the local coexpression domains in Arabidopsis, these numbers were compared with the numbers of pairs, triplets, and quadruplets obtained from randomized sets using the cumulative binomial distribution (Cohen et al., 2000). Such comparisons indicated that in all cases examined, local coexpression domains ranging from two to four genes occur in the Arabidopsis genome significantly more often than expected by chance alone (Table II).

Table II. *Number of local coexpression domains ranging two to four genes*

	Arabidopsis Genome		Random Genome (100×)	
	Total[a]	Coexpressed[b]	Average[c]	*P* Value[d]
Pairs				
MPSS+[e]	16,144	905 (5.60%)	676 ± 25	1.52×10^{-18}
MPSS−[f]	14,216	689 (4.85%)	588 ± 24	2.88×10^{-6}
MA+[g]	18,443	1,800 (9.76%)	1,352 ± 33	1.80×10^{-34}
MA−[h]	16,165	1,481 (9.16%)	1,211 ± 31	5.96×10^{-16}
Triplets				
MPSS+	13,142	52 (0.40%)	22.6 ± 4.7	7.95×10^{-8}
MPSS−	12,392	42 (0.34%)	19.6 ± 4.1	6.33×10^{-6}
MA+	15,634	113 (0.72%)	70.9 ± 8.0	9.70×10^{-7}
MA−	14,493	107 (0.74%)	70.9 ± 8.8	1.39×10^{-5}
Quadruplets				
MPSS+	10,718	5 (0.05%)	0.76 ± 0.89	9.88×10^{-4}
MPSS−	10,403	5 (0.05%)	0.72 ± 0.92	7.84×10^{-4}
MA+	13,282	8 (0.06%)	4.39 ± 2.38	5.81×10^{-2}[i]
MA−	12,866	7 (0.05%)	4.50 ± 1.85	8.24×10^{-2}[i]

[a]Total number of pairs, triplets, and quadruplets in each data set. [b]Coexpressed pairs, triplets, and quadruplets in each data set. Percentages in the brackets are coexpressed relative to the total. [c]Average and SD from 100 times randomizations. [d]*P* value according to the cumulative binomial distribution (Cohen et al., 2000) for obtaining such result by chance. $P < 0.01$ is considered significant. [e]MPSS data set including tandem duplicates. [f]MPSS data set excluding tandem duplicates. [g]MA data set including tandem duplicates. [h]MA data set excluding tandem duplicates. [i]Not significant.

Excluding tandem duplicates, coexpressed adjacent genes also occurred significantly more often than in random sets (Table II). Tandem duplicates are therefore not an important explanation for the occurrence of local coexpression domains in the Arabidopsis genome.

Local Coexpression Domains Are Not Solely Explained by Gene Orientation and/or Gene Distance

Apart from tandem duplications, gene orientation and gene distance could also explain the occurrence of local coexpression domains. If promoter sharing is an important mechanism for coexpression in the Arabidopsis genome, divergently transcribed gene pairs (←*gene A gene B*→) should be overrepresented in the subpopulation of coexpressed pairs, compared to coexpressed pairs that are tandemly (*gene A*→*gene B*→ or ←*gene A*←*gene B*, so two possibilities) or convergently (*gene A* →←*gene B*) transcribed. For all three orientation groups, the number of pairs and the number of coexpressed pairs were determined (Table III; Fig. 3, A and B). These results show that the Arabidopsis genome contains about twice as many tan-

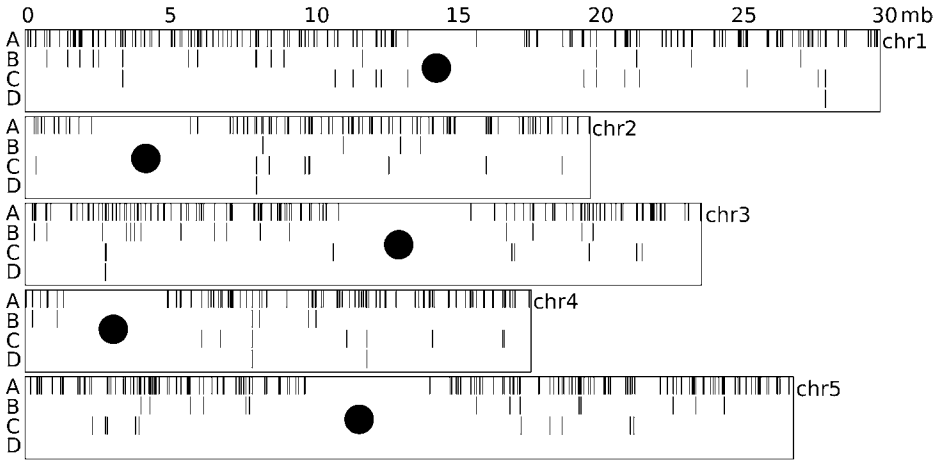

Figure 2. Distribution of local coexpression domains over the Arabidopsis chromosomes. Rectangles are schematic representations of chromosomes 1 to 5 from top to bottom, with black dots as centromeres. The numbers on the top show the scale in million bases along the chromosomes. Each gene is depicted as a black bar. Only the data sets excluding tandemly duplicated genes are shown. The letters on the left are: lane A, coexpressed pairs in the MPSS data set (689 pairs); lane B, common coexpressed pairs in both the MPSS and the MA data set (58 pairs); lane C, coexpressed triplets in the MPSS data set (42 triplets); and lane D, coexpressed quadruplets in the MPSS data set (5 quadruplets).

Table III. *Orientation of coexpressed gene pairs*

Orientation Groups[a]	Total[b]	Coexpressed[c]
MPSS		
tan − tandemly duplicated	6,979	322 (4.61%)
div − tandemly duplicated	3,541	191 (5.39%)
con − tandemly duplicated	3,696	176 (4.76%)
MA		
tan − tandemly duplicated	7,895	715 (9.06%)
div − tandemly duplicated	4,127	396 (9.60%)
con − tandemly duplicated	4,143	370 (8.93%)

[a]tan − tandemly duplicated, div − tandemly duplicated, and con − tandemly duplicated, respectively, are the subgroups of tandemly, divergently, and convergently transcribed pairs excluding tandem duplicates. [b]Total number of pairs in each direction group. [c]Number of coexpressed pairs in each direction group. Percentages in the brackets are number of coexpressed pairs relative to the total number of pairs. None of the proportions are significantly different from each other according to the z test for comparing population proportions.

demly transcribed pairs as divergently or convergently transcribed pairs. This is as expected, because the tandem orientation has two possibilities. For each orientation group, the fraction of coexpressed pairs relative to the total number of pairs in that group is plotted in Figure 3C. Expressed as a fraction relative to the total number of pairs in each orientation group, coexpressed divergently transcribed gene pairs occur in the same frequency as tandemly and convergently transcribed gene pairs (Fig. 3C). There are no significant differences in the proportions of coexpressed pairs between tandem and divergent, tandem and convergent, or divergent and convergent pairs (Table III). These results demonstrate that divergently transcribed gene pairs are not overrepresented in the subgroup of coexpressed gene pairs. Shared promoter sequences are therefore not a major explanatory variable for high coexpression between adjacent genes.

The closer two genes are, the higher the likelihood may be that their promoters influence each other, irrespective of gene orientation. Gene distance is here defined as the distance in nucleotides from the 5′ start ATG of one gene to the 5′ start ATG of the next gene. Thus defined, gene distance covers the distance between one coding sequence and promoter region for tandemly transcribed genes. For the other two gene orientations, this definition of gene distance results in the length of either the promoter sequence (in the case of divergently transcribed genes) or the inclusion of two coding regions (in the case of convergently transcribed genes). As a consequence, gene distance will favor divergently transcribed genes in the shorter distances and tandemly and convergently transcribed pairs in the larger distances. The subsequent distance analysis also distinguishes between gene orientations. All adjacent gene pairs (excluding the tandemly duplicated gene pairs) were sorted by gene distance and divided into consecutive bins of 1,000 pairs. This way, any influence of unequal numbers of pairs in distance bins was prevented. For each 1,000-pair bin, the number of tandemly, divergently, and convergently transcribed adjacent pairs was counted and plotted against the average gene distance (Fig. 4A). The average distance was calculated by averaging the gene distance of all pairs in each 1,000-pair bin. In the same way, the number of coexpressed adjacent pairs in each orientation group was counted and plotted (Fig. 4B). In both cases, it can be concluded that in the shorter gene distance classes divergently transcribed pairs occur more often than tandemly and convergently transcribed pairs, whereas in the larger gene distance classes divergently transcribed genes occur less often. Interestingly, coexpressed adjacent genes in any orientation could occur even at a gene distance as large as 12 kb. To be able to compare the relative occurrence of the orientation groups among all the 1,000-pair bins, the fraction of coexpressed pairs was plotted for each orientation group (Fig. 4C). The fractions of coexpressed pairs stay similar among three orientation groups and also stay similar over large gene distance range. Basically identical results were obtained for the MA data set (Fig. 4, D–F). Similar results were obtained using distance bins of 1-kb intervals or intergenic distances (data not shown). These results show that over a large gene distance range, the relative fraction of coexpressed pairs does not depend on gene distance, irrespective of gene orientation. Therefore, also gene distance and/or gene orientation are not important explanations for the occurrence of local coexpression domains in the Arabidopsis genome.

Genes in Local Coexpression Domains Scatter over Functional Categories

Having estimated the number of local coexpression domains in the Arabidopsis genome, the nature of the genes involved in such chromosomal domains was analyzed. The Arabidopsis Information Resource (TAIR)'s Gene Ontology (GO) using the high-level ontology terms known as GOslims developed for plants (Berardini et al., 2004) were used to characterize the genes in local coexpression domains. Genes in coexpressed triplets and quadruplets were not examined separately and pairs consisting of tandemly duplicated genes were not included in this analysis. Using the plant GOslim terms, the genes in coexpressed pairs were classified into the divisions for molecular function (15 categories), biological process (15 categories), and cellular components (16 categories). A pair was classified into a category if both members fell into the same category; otherwise, the pair was classified as ''not falling into the same category.'' Pairs of which one or both genes could not be classified were not included in the analysis. About 90% of all pairs (out of 14,216 for MPSS and 16,165 for MA; Table II) or coexpressed pairs (out of 689 for MPSS and 1,481 for MA; Table II) could be assigned to at least one GOslim category. Classification using the MIPS Functional Catalogue (Wu et al.,

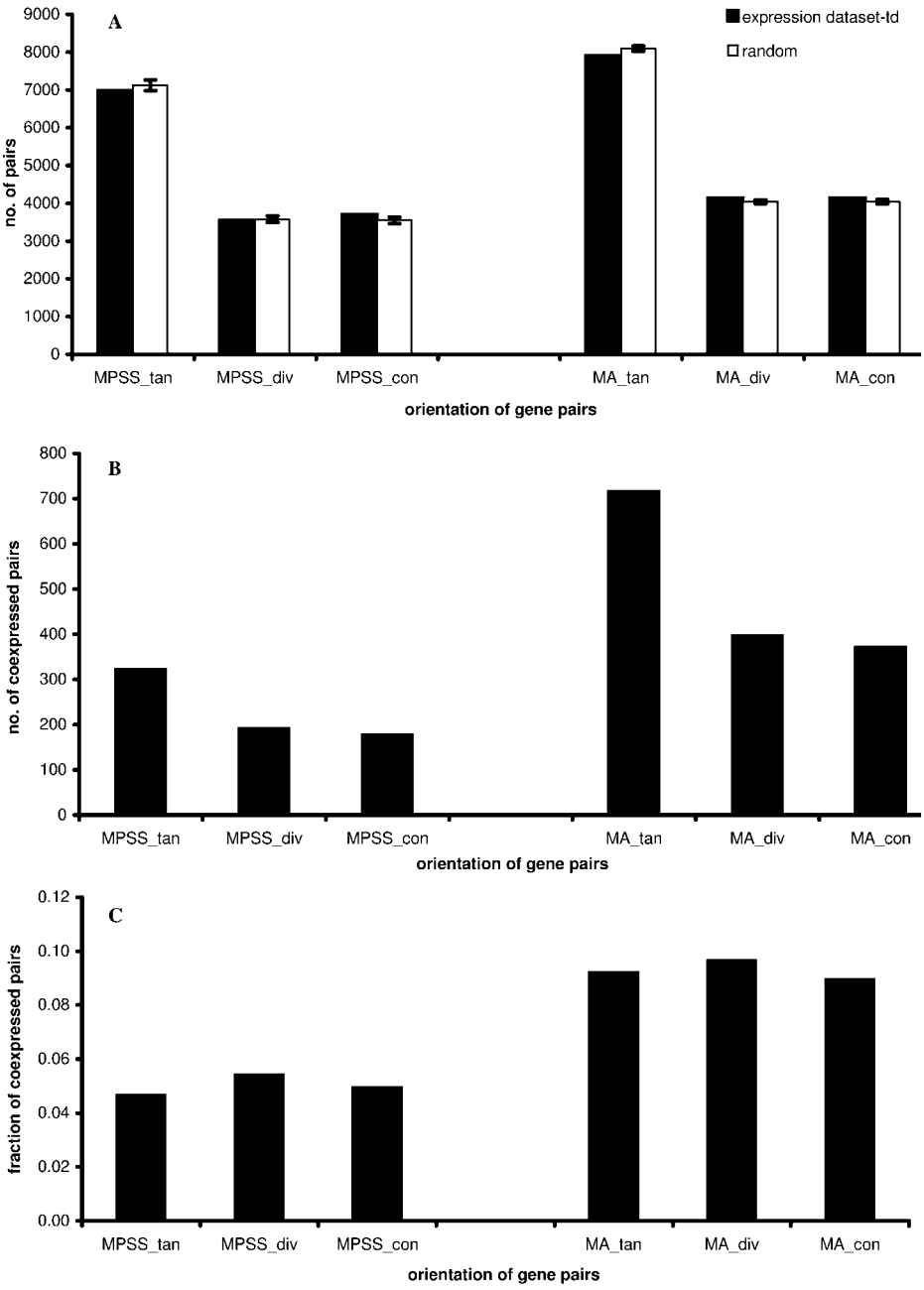

Figure 3. Orientation of genes in coexpressed pairs does not solely explain the occurrence of coexpression. The orientation groups based on the relative direction of transcription within a gene pair are tandem (tan), divergent (div), and convergent (con). Black bars are Arabidopsis expression data, white bars represent the averaged result from 100 randomizations. The *x* axis gives the expression data set used, either MPSS or MA, without tandemly duplicated genes. A, The number of pairs in each orientation group; B, the number of coexpressed pairs; C, the fractions of coexpressed pairs in each orientation group (given in B) relative to the total number of pairs in that corresponding orientation group (given in A). When corrected for the higher occurrence of tandemly oriented gene pairs, due to two possible orientations, none of the orientation groups is overrepresented in coexpressed pairs.

2002) covered much less (about only 30%) of the genes in pairs (data not shown). In each GOslim division, there are GOslim terms for ''unknown'' and ''other'' (Berardini et al., 2004). These should be considered less informative for the classification of pairs of genes. Therefore, we have distinguished a subclass of genes falling into the well-defined categories, excluding all categories with unknown and other. The results are summarized in Table IV. Considering the GOslim division for molecular function (GO_func), about 22% of the coexpressed pairs consist of genes that fall in the same functional category for both the MPSS and MA data sets (Table IV). For biological process (GO_proc), this is about 43% and for cellular component (GO_comp) this is 29%. When limited to the genes in categories that have no indication of other or unknown, about 6% to 7% of the pairs have genes that classify in the same category. Compared to the distribution of the genes of all pairs, the percentages of pairs in the same functional category do not differ significantly (at $P < 0.01$). Therefore, coexpressed pairs do not tend to fall more in the same GOslim category than other gene pairs (Table IV). Compared to what is expected on the basis of randomized genomes, the percentage of coexpressed genes falling in the same GOslim is not

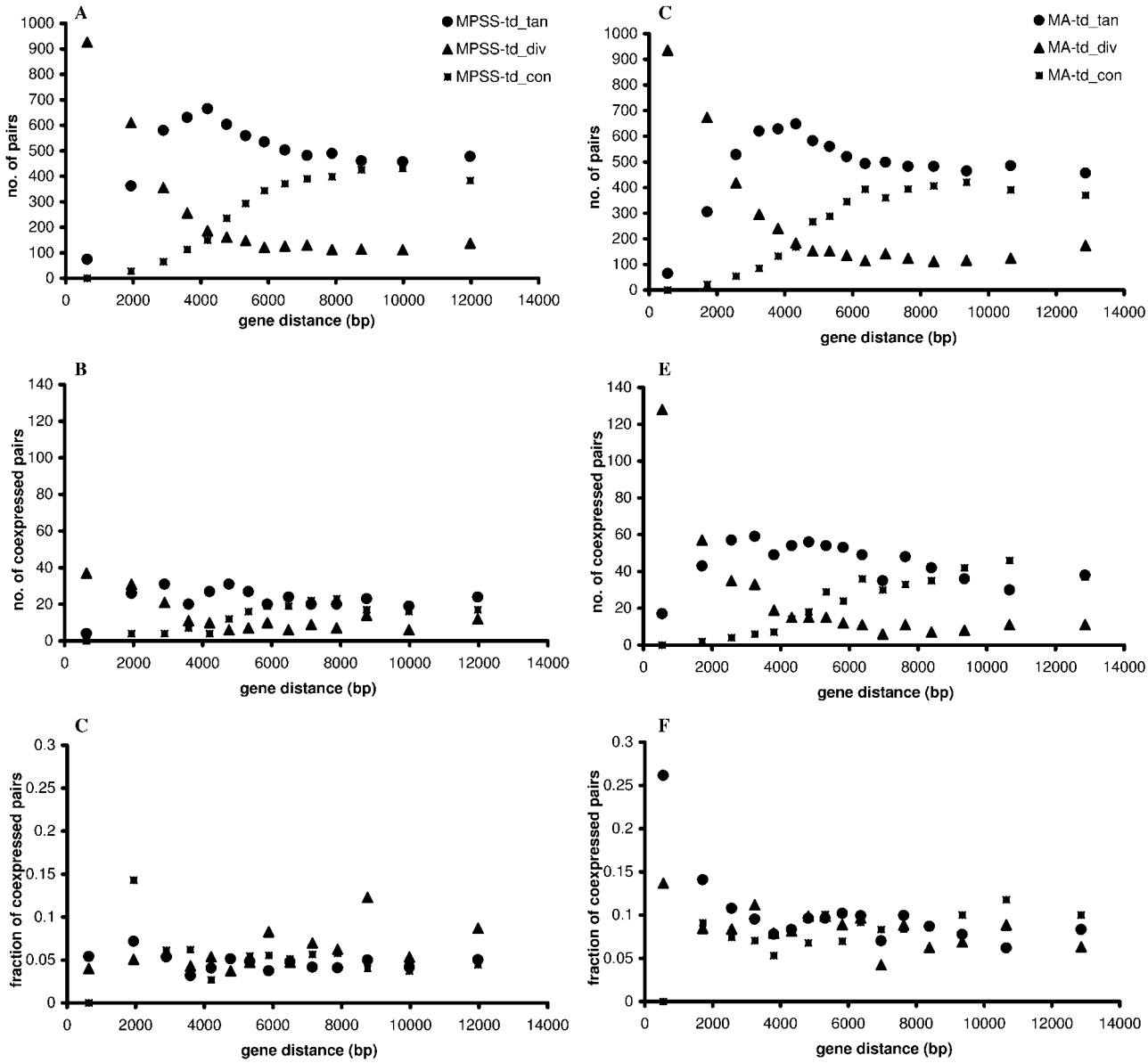

Figure 4. Gene distance of genes in coexpressed pairs does not solely explain the occurrence of coexpression. Gene distance, defined as start-to-start distance of adjacent gene pairs, is averaged for each 1,000-pair bin and plotted as function of gene orientation, subdivided into tandem pairs (tan; circles), divergent pairs (div; triangles), and convergent pairs (con; squares) for the MPSS data set (A–C) and the MA data set (D and E). A and D, Number of pairs; B and E, number of coexpressed pairs; C and F, the fraction of coexpressed pairs relative to the total number of pairs.

different from what is found in random genomes, with the notable exception of the percentage of genes that fall in the same category of well-defined biological processes. In both the MPSS and the MA data, about three times (6%–7% versus 2% expected) more coexpressed pairs occur in this category than expected on the basis of a random distribution. Within the category of well-defined biological processes, the category ''protein metabolism'' is overrepresented in both data sets: 43% (18 from 43) for MPSS and 61% (48 from 79) for MA of the pairs fall in this particular GOslim category (Table IV).

DISCUSSION

For different organisms, it has been demonstrated that appreciable numbers of genes in a genome occur in clusters characterized by correlated expression. Averaging coexpression over size-based or gene number-based windows showed that about 20% of the Drosophila genome resides in coexpression clusters (Spellman and Rubin, 2002). Within the Arabidopsis genome, such window-based coexpression clusters may consist of up to 20 genes (Birnbaum et al., 2003; Zhu, 2003; Williams and Bowles, 2004), while some

Table IV. *Distribution of gene pairs over GOslim categories*

	Genome All[a]	Coexpressed[b]	*P* Value[c]	Random Coexpressed[d]	*P* Value[e]
MPSS					
GO_func					
Covered[f]	12,920	624		537	
SameCat[g]	2,662 (20.6%)	136 (21.8%)	0.48	114 (21.2%)	0.80
SameKnCat[h]	1,041 (8.06%)	43 (6.89%)	0.26	43 (7.95%)	0.49
GO_proc					
Covered	12,927	623		537	
SameCat	5,366 (41.5%)	268 (43.0%)	0.46	225 (41.9%)	0.71
SameKnCat	867 (6.71%)	43 (6.90%)	0.86	11 (2.05%)	<0.0001[i]
GO_comp					
Covered	13,043	628		539	
SameCat	3,314 (25.4%)	181 (28.8%)	0.07	138 (25.6%)	0.22
SameKnCat	789 (6.05%)	42 (6.69%)	0.41	29 (5.38%)	0.35
MA					
GO_func					
Covered	14,804	1304		1,117	
SameCat	3,234 (21.8%)	286 (21.9%)	0.93	245 (21.9%)	1.0
SameKnCat	1,147 (7.75%)	99 (7.59%)	0.83	99 (8.86%)	0.26
GO_proc					
Covered	14,770	1,316		1,115	
SameCat	6,132 (41.5%)	552 (42.0%)	0.73	470 (42.2%)	0.92
SameKnCat	931 (6.30%)	79 (6.00%)	0.66	24 (2.15%)	<0.0001[i]
GO_comp					
Covered	14,756	1,317		1,115	
SameCat	3,806 (25.8%)	375 (28.5%)	0.04	289 (25.9%)	0.15
SameKnCat	811 (5.50%)	97 (7.37%)	0.012	77 (6.91%)	0.66

[a]Number of neighboring pairs included in the analysis. [b]Number of coexpressed pairs included in the analysis. [c]*P* value, the probability under the null hypothesis that the two population proportions are the same, derived from the standard normal tables of the z statistic for the difference of the population proportion between coexpressed pairs and all the pairs. [d]Number of coexpressed pairs in random sets. [e]*P* value, the probability under the null hypothesis that the two population proportions are the same, derived from the standard normal tables of the z statistic for the difference of the population proportion between coexpressed pairs of the Arabidopsis genome and coexpressed pairs in randomized sets. [f]Number of pairs of which both members are falling in a GOslim category. [g]Number of pairs of which both members fall into the same GOslim category. Percentage is the number of pairs relative to the total number of pairs covered. [h]Number of pairs of which both members fall into the same "known" GOslim category (excluding the categories with the indications "unknown" and "other"). Percentage is the number of pairs relative to the number of pairs covered. [i]Significant (two-tailed; $P < 0.01$).

evidence from quantitative trait loci studies suggested that clusters may be much larger (Khavkin and Coe, 1997). These data support the notion of higher-level genome organization that may range over distances up to several megabases (Hurst et al., 2004). Yet the concept of large coexpression clusters in such studies is based on a loose definition of the term cluster or chromosomal domain and associated terms such as neighboring. The process of summing and averaging may obscure local effects and underrate the presence and/or role of individual genes with different expression levels or expression patterns in large clusters. Therefore, it is perhaps not surprising that coexpression clusters are often associated with the activity of housekeeping (Lercher et al., 2002, 2003; Roy et al., 2002) or highly expressed (Caron et al., 2001; Versteeg et al., 2003) genes. Previous experience with transgene expression data indicated that the particular position of a single gene in a genome affects the expression of that gene. Depending on chromosomal context, two physically neighboring transgenes could be made to show correlated expression (Mlynarova et al., 2002). Therefore, we have here taken a very rigorous approach to the concept of cluster and analyzed the coexpression characteristics of genes that are physically adjacent in the genome according to genome annotation data.

Whole-genome chromosomal coexpression maps indicate the existence of numerous cases of local coexpression (Fig. 1), as was also shown in yeast (Cohen et al., 2000). Combining expression data and genome annotation, we identified 16,144 adjacent pairs of genes with sufficient expression data in the MPSS data set. The arbitrary criterion taken for inclusion of a gene in the analysis was detectable expression in at least one of the data libraries available. Although some genes may then have expression only in one library, around 80% of the genes have expres-

sion data in at least three different libraries and this is likely to yield reliable results. A major issue in such coexpression analyses is the occurrence of tandemly duplicated genes (Zhu, 2003; Hurst et al., 2004). Tandemly duplicated genes could be considered a trivial case of coexpression. All analyses, except when indicated, were performed with and without tandemly duplicated genes. From the 16,144 pairs, 12% were identified as tandemly duplicated genes. Only 11% of these tandemly duplicated gene pairs identified showed coexpression, which is 24% of all pairs with coexpression (Table I). The MA data set corroborates the MPSS findings: only 14% of the tandemly duplicated pairs showed coexpression. This suggests that, in contrast to inferences made for other genomes (Lercher et al., 2003), tandemly duplicated genes in the Arabidopsis genome are not a major cause of correlated expression of adjacent genes. Also the particular orientation of the tandemly duplicated genes, either tandemly, divergently, or convergently transcribed, was found to have no significantly higher inclination to be coexpressed, in contrast to the conclusions of the analyses of Williams and Bowles (2004). As the MA and MPSS data sets used in this study are biologically very different, their agreement with respect to the relative unimportance of tandemly duplicated genes in our analysis suggests that the data sets used in the respective analyses need to be considered. Careful future comparisons of data sets, gene coverage, and analytical methods used will have to reveal the cause of such differences.

From all nontandemly duplicated pairs in the MPSS data set, 4.9% shows coexpression. They are distributed over the whole genome (Fig. 2). Although this is a low percentage, randomization assays indicate that the number is significantly higher than to be expected by chance alone (Table II). There is a small yet significant fraction of the Arabidopsis genome that shows correlated expression between neighboring genes. Enlarging such local clusters by looking for series of consecutive genes that are correlated in all pairwise combinations reveals that there are few areas in the Arabidopsis genome that consist of more than two (up to four) genes (Table II; Fig. 2) with highly correlated expression. The size of these local coexpression domains is in agreement with local cluster sizes observed in yeast (Cohen et al., 2000) and worm (Roy et al., 2002). The MA data set, despite its technologically different approach for obtaining expression data and its biologically different experimental background, also showed local coexpression domains ranging from two to four genes distributed over the genome.

Over the whole genome, the two expression data sets show areas that have different coexpression patterns (Fig. 1) and in total only 58 coexpressed pairs were shared between both data sets (Fig. 2). These differences in coexpression and low number of shared pairs are likely to reflect the biological differences between the data sets. The MA data are well defined root cells and tissues, while the MPSS data concern more broad tissues and organs. Such biological differences will influence correlations in gene activity. Any expression data set will present a fixed average of expression over the sampled cells, tissues, organs, and experiments that should be taken into account when comparing such data sets.

To understand the possible causes for coexpression, the role of shared promoters and/or short gene distances was analyzed. The population of divergently transcribed genes does not contain a higher proportion of coexpressed genes compared to tandemly or convergently transcribed genes (Fig. 3; Table III). Promoter sharing is therefore not a likely explanation for the presence of local coexpression domains in the Arabidopsis genome, unlike the situation in the yeast genome (Cohen et al., 2000). Also, gene distance does not offer an important explanation for the occurrence of local coexpression domains. When corrected for gene orientation, the fraction of coexpressed genes does not depend on either gene orientation or gene distance (Fig. 4, C and F). Short gene distances ($<$1 kb) do not favor local coexpression and longer distances (up to 10 kb) need not necessarily be barriers to local coexpression. In this analysis, gene distance is defined as the distance from the 5′ start ATG of one gene to the 5′ start ATG of the next gene and includes the coding region of a gene (for tandemly transcribed gene pairs) or of both genes (for convergently transcribed gene pairs). Similar results were obtained when the intergenic distance, defined as the distance between the stop codon of one gene and the start codon of the next gene, was taken for analysis (data not shown). Therefore, the precise definition of gene distance in the analyses as presented does not affect the conclusions. The role between gene distance and correlation of expression has given different results in different studies. Some indicate that correlation declines with increasing distance (Cohen et al., 2000; Williams and Bowles, 2004), while others are less explicit and emphasize the role of relative genome compactness (Fukuoka et al., 2004). Analyses of the MA data with the TIGR5 annotation of the Arabidopsis genome had no significant effect on trends and conclusions (data not shown).

Previous studies suggested that clustering of functionally related genes may occur in all metazoans (including yeast, fly, worm, and human; Cohen et al., 2000; Lercher et al., 2002, 2003; Spellman and Rubin, 2002). A recent study (Williams and Bowles, 2004) demonstrated a significant enrichment for coexpressed genes in the same metabolic pathway, although this appeared not to be an explanation for the neighboring coexpression. In this study, a loose definition of neighboring was used, defining two genes as neighboring when they were within 10 genes of each other (Williams and Bowles, 2004). In worm, clusters of similarly expressed genes cover similar biological functions (Roy et al., 2002). In human, coexpression analysis over the whole genome was

shown to correlate with functional relatedness (Lee et al., 2004). In the expression data sets here analyzed with a gene ontology developed for plants (GOslim; Berardini et al., 2004), there is, however, no evidence that coexpressed genes in pairs are enriched in the same functional category compared to all genes in pairs (Table IV). This is also the case when compared to the percentages of coexpressed genes in random sets (Table IV). When the GOslim categories without "unknown" or "other" are used, only in the GOslim division covering "biological process" coexpressed gene pairs are about 3 times more frequently present than expected to occur by chance, notably in the GOslim biological process category "protein metabolism." In the other GOslim divisions, no such trend is present; coexpressed gene pairs are as frequently present as all gene pairs. (Table IV).

In our coexpression analyses of expression data, different libraries from either MPSS or MA data were combined irrespective of the biological characteristics of the material assayed. Therefore, the analyses have revealed the gene pairs that show stringent coexpression under a range of different (biological) conditions, cells, and/or tissue types. Combining more and different data sets, such as the data in various Arabidopsis expression repositories now available at TAIR (Rhee et al., 2003), National Center for Biotechnology Information (NCBI)'s Gene Expression Omnibus (Edgar et al., 2002), Genevestigator (Zimmermann et al., 2004), Stanford Microarray Database (Gollub et al., 2003), or the Arabidopsis Tissue-Specific Expression Database (Obayashi et al., 2004), will help to analyze the expression of genes over various conditions and cell types. Yet averaging more and different expression data sets would continue to favor the identification of gene pairs expressed under all conditions in as many cell and tissue types as available in expression repositories. Although this would reveal the expression potential of gene pairs in a genome, it would be much less informative for elucidating the whole-genome dynamics of coexpression. Local coexpression domains may be dynamic during growth and development of plants. In future analyses, it may therefore be worthwhile to analyze prechosen subsets of libraries and compare the local coexpression dynamics of different organs, tissues, or cells to identify time- or tissue-specific local coexpression domains.

True neighboring pairs can form local coexpression domains of two to four genes irrespective of gene orientation or gene distance. Having eliminated such configuration factors, a role of either the gene sequence itself or the DNA sequences surrounding these genes is suggested. In the transgenic situation, it was shown before that the expression of two unrelated genes became correlated when their surrounding DNA was supplied with a chromatin boundary element (Mlynarova et al., 2002). A next step of genome analysis will therefore be the detailed analysis of the sequences next to local coexpression domains. These may consist of boundary elements such as matrix-associated regions (Boulikas, 1995; Bell et al., 2001) and help to further define the (sequence) characteristics of such elements.

CONCLUSION

Defining local coexpression domains as genome areas with physically neighboring genes showing tight coexpression, we have here shown that the Arabidopsis genome contains a small yet significant number of coexpression domains that range from two to four genes. Neither tandemly duplicated genes nor divergently transcribed promoter regions nor short gene distances explain such local coexpression of adjacent genes. Either gene sequence or the surrounding DNA sequences are of importance for the coexpression pattern of such neighboring genes. Our study and the further unraveling of the relationships between local and global coexpression domains in relationship to surrounding DNA, gene regulation, and chromosome structure will help to gain understanding of the molecular mechanisms that establish local chromosomal domains of genes with high coexpression characteristics.

MATERIALS AND METHODS

Data Retrieval and Processing

The Arabidopsis (*Arabidopsis thaliana*) genome annotation from the March 2003 version of MIPS (Schoof et al., 2002) has 26,439 annotated genes on five chromosomes. Mitochondria and chloroplast genes were not taken into account in this study. There are 6,813, 4,181, 5,363, 3,987, and 6,095 genes on chromosome 1 to 5, respectively. The genes along each chromosome were sorted based on ascending start coordinates and were numbered consecutively. This established a rank number (rank ID) that helped to eliminate any discontinuity in the Arabidopsis Genome Initiative (AGI) numbers of the annotated genes and allowed analyzing physically adjacent genes. These rank IDs of genes were used to compare two different whole-genome expression data sets, MPSS expression data and MA expression data. Data are summarized in Table I. The MPSS data was obtained from the Arabidopsis MPSS website (http://mpss.udel.edu/at/java.html; Meyers et al., 2004). The MPSS data set has 14 libraries covering 5 plant tissues: callus, inflorescence, leaf, root, and silique. All MPSS 17-bp signatures that had a normalized expression abundance of at least 1 transcript per million in at least one of the 14 libraries were retrieved manually. Genes without MPSS signature or with no expression value of at least 1 transcript per million in any of the 14 libraries were not taken into consideration. With Python scripts, the MPSS signatures were mapped onto the MIPS genome annotation, based on an exact match of 17 bp, and assigned the corresponding chromosomal position. Each signature that was assigned more than once was removed from the data set. Each MPSS mapped signature was assigned to a class based on the genomic location and MIPS annotation. Seven different classes were defined according to the criteria on the MPSS website (http://mpss.udel.edu/at/java.html): class 1 (inside an annotated gene/feature); class 2 (within 250 bp 3′ of the annotated gene/feature); class 3 (anti-sense to annotated gene/feature); class 4 (between gene/feature); class 5 (within intron, sense strand); class 6 (within intron, anti-sense strand); and class 7 (within 17 bp of an exon boundary; spliced). With the precedence ranking of classifications: $1 = 7 > 2 > 5 > 3 > 6 > 4$ for signatures belonging to more than one possible class, every signature was assigned to only one class. The normalized expression values of both class 1 and class 2 signatures in the same library were summed and used as the expression value of the corresponding gene. Genes with neither class 1 nor class 2 signatures were considered to be not expressed and were not taken into consideration. This way, we obtained 20,041 genes having MPSS expression values, referred to as the MPSS data set.

The MA expression data were obtained from the on-line supplementary material of a Science article (Birnbaum et al., 2003). The MA data set based on the ATH1 GeneChip (Affymetrix, Santa Clara, CA) has expression data only from Arabidopsis root tissue, encompassing 15 different zones of the root that correspond to different cell types and tissues at progressive developmental stages. Genes not on the array were not taken into consideration. Genes on the array of which the AGI numbers could not be mapped to the MIPS genome annotation were also discarded. After mapping these gene expression data by their unique AGI numbers to the MIPS annotation, we obtained 21,940 genes having MA expression values, referred to as the MA data set. Analyses of the MA data with the TIGR5 annotation of the Arabidopsis genome had no major effect on trends and conclusions (data not shown).

In case of physically overlapping genes in either data set, the smaller one of the overlapping genes was removed from the data set, by which both gene and rank ID orders were maintained. For this reason, 39 and 34 genes were removed from the MPSS and the MA data set, respectively. The resulting data sets used for analysis consisted of 20,002 genes with MPSS expression data and 21,906 genes with MA expression data. Two genes were considered to be adjacent when their rank IDs were consecutive with a difference of one and when the genome sequence had no long stretches of N's in-between. In six cases (three on Chr1 and three on Chr2), the genome sequence was interrupted by a stretch of 60 or 120 N's. These genes were included in the subsequent analyses. Adjacent genes were considered per chromosome. With these criteria, 16,144 adjacent gene pairs were identified in the MPSS data set. These pairs comprised 19,151 genes with expression values. A total of 851 (that is, the difference between 20,002 and 19,151) genes in the MPSS data set had no neighbors with expression data. In the MA data set, 18,443 adjacent gene pairs in the MA data set were identified, comprising 21,255 genes and 651 isolated genes without expressed neighboring genes (Table I).

Tandemly duplicated genes were identified by local pairwise protein BLAST (BLASTP 2.2.6 [April 9, 2003]; Altschul et al., 1997) on all gene pairs in both data sets. A gene pair was considered to be a tandemly duplicated pair if BLASTP yielded $E < 2 \times 10^{-1}$ (Lercher et al., 2002, 2003; Fukuoka et al., 2004; Williams and Bowles, 2004). This criterion, developed on the basis of duplicated human (*Homo sapiens*) genes, removes about 90% of the related genes from a population and has a false positive rate of about 10% (Lercher et al., 2002; Williams and Bowles, 2004). This way, 1,928 and 2,278 adjacent pairs were identified in the MPSS and the MA data set, respectively. Most analyses were done for data sets including tandemly duplicated or excluding tandemly duplicated. Such exclusion implied that the tandemly duplicated pair was not included in the coexpression analysis, but the expression of each member of a tandemly duplicated gene pair was analyzed relative to its other neighbor.

Identification of Local Coexpression Domains

R was calculated between all adjacent pairs (duplets) of genes using the expression data from all libraries available in each data set. If *R* was higher than 0.7, the gene pair concerned was considered to be coexpressed. The value of $R > 0.7$ is generally considered a rule-of-thumb threshold (for example, see http://bbc.botany.utoronto.ca/affydb/BAR_instructions) and is used in various analyses (Cohen et al., 2000; Lee et al., 2004). When calculating the *R* values from a whole-genome all-against-all comparison (used to establish Fig. 1.) and plotting these as a histogram, the top 5% in this distribution may be used to derive a threshold for determining coexpression, analogous to the 5% upper tail in a normal distribution. For the MPSS data, the upper 5% cutoff is 0.65 and for the MA data 0.72. For convenience and comparability, the approximate average of $R > 0.7$ was chosen for analysis. With a lower threshold value for *R*, such as for example $R > 0.5$ (Blanc and Wolfe, 2004), the absolute numbers of the various categories of genes go up, but the relative results do not change dramatically from what is presented (data not shown).

The number of coexpressed adjacent pairs was counted. To evaluate the statistical significance of these numbers, they were compared with the number of coexpressed pairs from 100 randomizations of the population of expressed genes using the cumulative binomial distribution (Cohen et al., 2000). Preliminary analyses indicated that more than 100 randomizations did not result in significant changes in the numbers obtained (data not shown). In each round of randomization, nonadjacent pairs of genes were randomly selected with replacement from the list of expressed genes that have expressed neighbors until the same total number of pairs was obtained. For example, the MPSS data set has 16,144 gene pairs that are neighboring genes with expression values. One round of randomization on the MPSS data set consisted of 16,144 times of randomly picking two genes with replacement out of the list of genes represented in the 16,144 gene pairs, calculating *R* for each pair, and counting the number of pairs having $R > 0.7$. Similarly, coexpressed adjacent triplets, quadruplet, and pentaplets were identified as series of genes with consecutive IDs in which all possible [that is, (n!/(n – 2)!)*2; Smith, 2002] pairwise *R* were above the cutoff of 0.7. The significance of results was evaluated with randomizations equivalent to the procedure used in case of duplets.

The Role of Gene Direction and Gene Distance in Local Coexpression Domains

Adjacent gene pairs were separated into tandemly, divergently, and convergently transcribed pairs according to their relative direction of transcription. The number of coexpressed pairs in each orientation group was expressed as percentage relative to the total number of adjacent pairs in that group. Random pairs were made by randomly picking two nonadjacent genes from the list of expressed genes represented in pairs, analyzed for their orientation, and compared with the real genome using a variant of the two-sample *t* test for proportions for determining the significance of a difference between two population proportions (Ott and Longnecker, 2001). The test statistic is based on the z statistic from the normal distribution and is given by $(p1 - p2)/\sqrt{(p1*(1 - p1)/n1 + p2*(1 - p2)/n2)}$, with p1 and p2 the two sample proportions, n1 and n2 the two sample sizes, under the condition that n1*p1, n1*(1 − p1), n2*p2 and n2*(1 − p2) are all larger than 5. When $|z| > 2.575$, the two sample proportions are considered to be significantly different at the 1% level ($P < 0.01$). The z value is converted to a *P* value using standard normal tables.

For gene distance, the length in nucleotides from the 5′ start of one gene to the 5′ start of the next gene was used. The data sets excluding the tandemly duplicated gene pairs were analyzed. This way, there were 14,216 pairs in the MPSS-tandemly duplicated data set and 16,165 pairs in MA-tandemly duplicated data set. For each data set, gene pairs were sorted based on gene distance and bins of 1,000 pairs were taken and analyzed, excluding the last bin with less than 1,000 pairs. Per 1,000-pair bin, gene distance was calculated as the average over all 1,000 pairs. In total, 14 bins of 1,000 pairs for the MPSS-tandemly duplicated data set and 16 bins of 1,000 pairs in MA data set were analyzed. Within each 1,000-pair bin, the numbers of tandem, divergent, and convergent pairs were determined, as well as the numbers of coexpressed pairs within each orientation group. To be able to compare bins, the fraction of coexpressed pairs relative to the total number of pairs in each orientation group in each bin was calculated.

Functional Categorization of Genes Represented in Local Coexpression Domains

TAIR's GOslim, the GO developed for plants (Berardini et al., 2004), was used to classify the genes present in local coexpression domains. The categories for molecular function (15 GOslim categories), biological process (15 GOslim categories), and cellular component (16 GOslim categories) were taken in consideration. With Python scripts, the number of pairs of which both members could be classified in GOslim was determined from the total number of coexpressed pairs. From this, the number of pairs of which both members fall in the same GOslim category was determined, also with the help of Python scripts. The GOslim categories include several "unknown" and "other." These were considered to give less information about functional categorization and were set apart from the genes falling into a well-defined category. The percentages obtained were compared with random sets using the z test for the significance of difference between two proportions (Ott and Longnecker, 2001) as outlined above.

Upon request, all novel materials described in this publication will be made available in a timely manner for noncommercial research purposes, subject to the requisite permission from any third-party owners of all or parts of the material. Obtaining any permissions will be the responsibility of the requestor.

ACKNOWLEDGMENTS

We thank Nayelli Marsch-Martinez and Harrie Verhoeven (Bioscience, Plant Research International, Wageningen, The Netherlands), Blake Meyers

(University of Delaware), Jack Leunissen (Bioinformatics, Wageningen University, The Netherlands), and Huanming Yang (Beijing Genomics Institute, Beijing) as well as Roeland van Ham, Paulien Adamse, Oscar Vorst, and other members of the Applied Bioinformatics cluster of Plant Research International for helpful comments, suggestions, data, and discussions.

Received October 26, 2004; revised February 27, 2005; accepted February 28, 2005; published May 27, 2005.

LITERATURE CITED

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389–3402

Bell AC, West AG, Felsenfeld G (2001) Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. Science **291:** 447–450

Berardini TZ, Mundodi S, Reiser L, Huala E, Garcia-Hernandez M, Zhang PF, Mueller LA, Yoon J, Doyle A, Lander G, et al (2004) Functional annotation of the Arabidopsis genome using controlled vocabularies. Plant Physiol **135:** 745–755

Birnbaum K, Shasha DE, Wang JY, Jung JW, Lambert GM, Galbraith DW, Benfey PN (2003) A gene expression map of the Arabidopsis root. Science **302:** 1956–1960

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell **16:** 1679–1691

Boulikas T (1995) Chromatin domains and prediction of MAR sequences. Int Rev Cytol **162A:** 279–388

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, et al (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science **291:** 1289–1292

Cohen BA, Mitra RD, Hughes JD, Church GM (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet **26:** 183–186

Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res **30:** 207–210

Fukuoka Y, Inaoka H, Kohane IS (2004) Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. BMC Genomics **5:** 4

Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, Hebert JM, Hernandez-Boussard T, Jin H, Kaloper M, Matese JC, et al (2003) The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res **31:** 94–96

Hurst LD, Pal C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet **5:** 299–310

Khavkin E, Coe E (1997) Mapped genomic locations for developmental functions and QTLs reflect concerted groups in maize (*Zea mays* L.). Theor Appl Genet **95:** 343–352

Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Co-expression analysis of human genes across many micro array data sets. Genome Res **14:** 1085–1094

Lercher MJ, Blumenthal T, Hurst LD (2003) Co-expression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. Genome Res **13:** 238–243

Lercher MJ, Urrutia AO, Hurst LD, Cohen BA, Mitra RD, Hughes JD, Church GM (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet **31:** 180–183

Meyers BC, Lee DK, Vu TH, Tej SS, Edberg SB, Matvienko M, Tindell LD (2004) Arabidopsis MPSS. An online resource for quantitative expression analysis. Plant Physiol **135:** 801–813

Mlynarova L, Jansen RC, Conner AJ, Stiekema WJ, Nap JP (1995) The MAR-mediated reduction in position effect can be uncoupled from copy number-dependent expression in transgenic plants. Plant Cell **7:** 599–609

Mlynarova L, Loonen A, Heldens J, Jansen RC, Keizer P, Stiekema WJ, Nap JP (1994) Reduced position effect in mature transgenic plants conferred by the chicken lysozyme matrix-associated region. Plant Cell **6:** 417–426

Mlynarova L, Loonen A, Mietkiewska E, Jansen RC, Nap JP (2002) Assembly of two transgenes in an artificial chromatin domain gives highly coordinated expression in tobacco. Genetics **160:** 727–740

Obayashi T, Okegawa T, Sasaki-Sekimoto Y, Shimada H, Masuda T, Asamizu E, Nakamura Y, Shibata D, Tabata S, Takamiya K, et al (2004) Distinctive features of plant organs characterized by global analysis of gene expression in Arabidopsis. DNA Res **11:** 11–25

Ott RL, Longnecker M (2001) An Introduction to Statistical Methods and Data Analysis, Chapter 10, Categorical Data, Ed 5. Duxbury, Pacific Grove, CA, pp 482–485

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, et al (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res **31:** 224–228

Roy PJ, Stuart JM, Lund J, Kim SK (2002) Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. Nature **418:** 975–979

Schoof H, Zaccaria P, Gundlach H, Lemcke K, Rudd S, Kolesov G, Arnold R, Mewes HW, Mayer KF (2002) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. Nucleic Acids Res **30:** 91–93

Smith LI (2002) A Tutorial on Principal Components Analysis. http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf (August 27, 2004)

Spellman PT, Rubin GM (2002) Evidence for large domains of similarly expressed genes in the *Drosophila* genome. J Biol **1:** 5

Versteeg R, van Schaik BD, van Batenburg MF, Roos M, Monajemi R, Caron H, Bussemaker HJ, van Kampen AH (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res **13:** 1998–2004

Williams EJ, Bowles DJ (2004) Co-expression of neighboring genes in the genome of *Arabidopsis thaliana*. Genome Res **14:** 1060–1067

Wu CH, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu ZZ, Ledley RS, Lewis KC, Mewes HW, Orcutt BC, et al (2002) The Protein Information Resource: an integrated public resource of functional annotation of proteins. Nucleic Acids Res **30:** 35–37

Zhu T (2003) Global analysis of gene expression using GeneChip microarrays. Curr Opin Plant Biol **6:** 418–425

Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol **136:** 2621–2632