

# Gene Family Analysis of the Arabidopsis Pollen Transcriptome Reveals Biological Implications for Cell Growth, Division Control, and

## AQ1] Gene Expression Regulation<sup>1[w]</sup>

Cristina Pina<sup>2</sup>, Francisco Pinto, José A. Feijó\*, and Jörg D. Becker

Centro de Biologia do Desenvolvimento, Instituto Gulbenkian de Ciência, PT-2780-156 Oeiras, Portugal (C.P., J.A.F., J.D.B.); Biomathematics Group, Instituto Química e Biológica-Universidade Nova de Lisboa, PT-2780-156 Oeiras, Portugal (F.P.); and Faculdade de Ciências, Departamento de Biologia Vegetal,

AQ2] Universidade de Lisboa, Campo Grande, PT-1749-016 Lisbon, Portugal (J.A.F.)

AQ5] Upon germination, pollen forms a tube that elongates dramatically through female tissues to reach and fertilize ovules. While essential for the life cycle of higher plants, the genetic basis underlying most of the process is not well understood. We previously used a combination of flow cytometry sorting of viable hydrated pollen grains and GeneChip array analysis of one-third of the *Arabidopsis* (*Arabidopsis thaliana*) genome to define a first overview of the pollen transcriptome. We now extend that study to approximately 80% of the genome of *Arabidopsis* by using Affymetrix *Arabidopsis* ATH1 arrays and perform comparative analysis of gene family and gene ontology representation in the transcriptome of pollen and vegetative tissues. Pollen grains have a smaller and overall unique transcriptome (6,587 genes expressed) with greater proportions of selectively expressed (11%) and enriched (26%) genes than any vegetative tissue. Relative gene ontology category representations in pollen and vegetative tissues reveal a functional skew of the pollen transcriptome toward signaling, vesicle transport, and the cytoskeleton, suggestive of a commitment to germination and tube growth. Cell cycle analysis reveals an accumulation of G2/M-associated factors that may play a role in the first mitotic division of the zygote. Despite the relative underrepresentation of transcription-associated transcripts, nonclassical MADS box genes emerge as a class with putative unique roles in pollen. The singularity of gene expression control in mature pollen grains is further highlighted by the apparent absence of small RNA pathway components.

Pollen is the male partner in plant reproduction. Besides this fundamental biological function, pollen tubes have served as a useful model for the study of cell growth and morphogenesis (Feijó et al., 2001, 2004). These approaches focus on the role of the vegetative cell of pollen because of its extreme elongation without cell division. Early studies on the physiology and biochemistry of pollen (Mascarenhas, 1975) have been supported by small-scale genetic studies during the last two decades, revealing a limited number of genes that are thought or known to be involved in pollen development, pollen tube growth, and functions during interactions of pollen with

female tissues (Scott et al., 1991; McCormick, 1993, 2004; Twell, 2002). Based on the assumption that mature pollen grains already contain most of the transcripts needed for germination and tube growth (Mascarenhas, 1989; Honys and Twell, 2004), three recent studies have employed large-scale analysis of the *Arabidopsis* (*Arabidopsis thaliana*) pollen transcriptome to increase significantly our knowledge about the genetic basis of pollen germination and tube growth. Using Affymetrix 8K *Arabidopsis* GeneChips (representing about 8,000 genes), we compared the transcriptome of highly purified, cell-sorted pollen grains with those of four vegetative tissues (Becker et al., 2003), while Honys and Twell (2003) used the same oligonucleotide array for a comparison of the transcriptome of nonsorted pollen grains with those of four sporophytic developmental stages. In another approach, serial analysis of gene expression was used to profile the transcriptome of pollen under normal and chilling conditions (Lee and Lee, 2003). Our study revealed the expression of 1,584 genes in pollen grains, 10% of which were selectively expressed. Pollen expressed 3-fold fewer genes than any of the vegetative tissues profiled (Becker et al., 2003). All three studies came to the conclusion that the reduced transcriptome of pollen showed a high proportion of enriched or selectively expressed genes

<sup>1</sup> This work was supported by the Fundação para a Ciência e a Tecnologia (project no. POCTI/BCI/41725/2001 and fellowships SFRH/BPD/3619/2000 and SFRH/BD/6488/2001 to J.D.B. and F.P., respectively). C.P. is a student of the Gulbenkian PhD Program in Biomedicine (Portugal).

AQ3] <sup>2</sup> Present address: Medical Research Council Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK.

\* Corresponding author; e-mail jfeijo@fc.ul.pt; fax 351-21-4407970.

AQ4] <sup>[w]</sup> 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.104.057935](http://www.plantphysiol.org/cgi/doi/10.1104/pp.104.057935).

with proposed functions in signal transduction (Feijó et al., 2004), cell wall biosynthesis, and cytoskeletal dynamics (for review, see da Costa-Nunes and Grossniklaus, 2004).

The Affymetrix Arabidopsis 8K GeneChip was, however, short in many aspects, namely, coverage for the analysis of whole gene families and pathways to allow a conclusive functional analysis of the complement of pollen transcripts. Information on the whole transcriptome level is crucial to developing testable hypotheses for future research on pollen tube growth and morphogenesis. Development of the Affymetrix Arabidopsis ATH1 genome arrays covering more than 80% of the Arabidopsis genome triggered 2 approaches; while Honys and Twell (2004) studied the transcriptional changes during microgametogenesis in Arabidopsis Landsberg *erecta* (*Ler*), we focus our efforts on a comparative analysis of highly purified pollen grains of Arabidopsis Columbia (*Col-0*) and characterized the complement of mRNA with respect to functions of mature pollen grains and pollen tubes. In both studies, Affymetrix Arabidopsis ATH1 genome arrays covering more than 80% of the Arabidopsis genome were used. The comparison of the transcriptomes of cell-sorted hydrated pollen grains with those of flowers, leaves, seedlings, and siliques revealed the expression of a comparatively reduced set of genes with increased proportions of enriched and selectively expressed transcripts. The graphic snail view representation and principal component analysis of the studied transcriptomes showed a global distinctiveness of the pollen expression profile and considerable similarities among the remaining tissues. The uniqueness of the pollen transcriptome was dissected by analyzing the frequencies of representation of gene ontology (GO) categories in the different tissues, which revealed that gene expression in the male gametophyte is skewed toward signaling, vesicle trafficking, cell wall metabolism, transport across membranes, and the cytoskeleton and, therefore, committed to germination and tube growth. We used gene lists extracted from The Arabidopsis Information Resource (TAIR) and recent publications to analyze the representation of a number of functional classes in the various Arabidopsis tissues. The extensive representation of each gene family on the ATH1 array allows formulation of hypotheses, here illustrated by the analysis of nonclassical MADS box genes, small RNA pathways, and cell cycle components.

## RESULTS AND DISCUSSION

### Pollen Grains Accumulate a Restricted and Unique Set of Genes

We used Arabidopsis GeneChip ATH1 oligonucleotide arrays, representing 22,750 annotated genes of the Arabidopsis genome, to establish the gene expression profile of fluorescence-activated cell sorted, hydrated

pollen grains (Becker et al., 2003) in comparison to vegetative tissues (all samples with biological duplicates). In a recent study, the ATH1 GeneChip has been shown to have a sensitivity of approximately one transcript per cell and a high technical reproducibility and concordance with reverse transcription-PCR results (Redman et al., 2004). Microarray suite (MAS) 5.0 software was used to obtain present and absent calls, and all subsequent analyses were performed with DNA-Chip Analyzer 1.3. The results were highly reproducible, with calculated correlation coefficients for the various replicates consistently above 0.99. Expression data for all genes represented on the array and all tissues tested are available in Supplemental Table I.

The different vegetative tissues presented close mean percentages of present calls: flowers, 68%; leaves, 62%; seedlings, 68%; and siliques, 69%. Pollen grains, on the other hand, presented a much lower mean level of present calls: 29%, corresponding to 6,587 genes. This may reflect to a certain extent the fact that vegetative tissues comprise a variety of differentiated cell types, while the pollen transcripts originate from single cells, or simply may reflect the specialization of pollen transcripts in keeping cell viability and allowing future germination, pollen tube growth, and fertilization of the female gamete. The notion that single cells express a restricted repertoire of transcripts is supported by a recent study analyzing different single cell types on 8K Arabidopsis genome arrays, in which 16% and 18% of the genes were expressed in guard cells and mesophyll cells, respectively (Leonhardt et al., 2004).

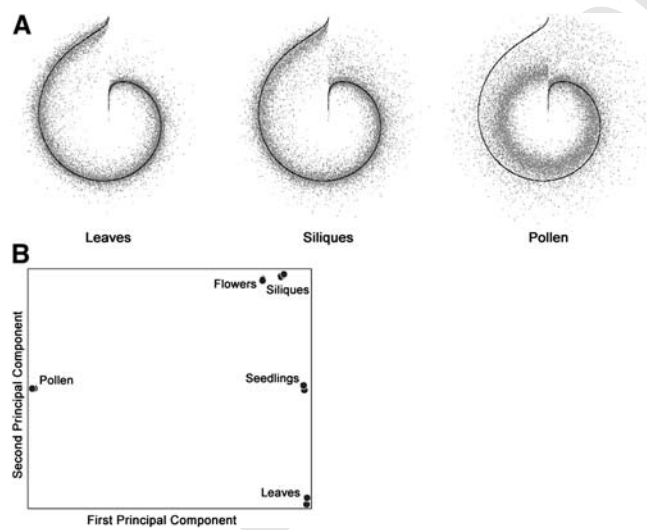
Honys and Twell (2004) concluded that 7,235 genes are expressed in mature pollen grains of Arabidopsis *Ler*. In addition to ecotype-specific differences in expression (Schmid et al., 2003), the discrepancy between our results and theirs can be explained to a large extent by the fact that Honys and Twell used the old MAS 4.0 detection algorithm to determine present and absent calls for their GeneChip raw datasets, but the empirical MAS 4.0 detection algorithm in comparison to the statistical MAS 5.0 detection algorithm yields more false-positive calls (Liu et al., 2002). Further possible sources for discrepancies in relation to our study are the omission of cell sorting and the cold storage of pollen before RNA extraction by Honys and Twell (2004), since autolysis observed in mature Arabidopsis pollen (Yamamoto et al., 2003) might ultimately lead to nonviable pollen with ongoing RNA degradation, and cold storage can provoke the degradation of specific RNAs in pollen (Wang et al., 2004).

Exclusion of genes with absent calls in all 10 arrays and with inconsistent expression levels within the replicates resulted in a list of 18,321 genes for further analysis. A snail view representation (Becker et al., 2003) of these genes underlines the striking differences of expression in the pollen transcriptome when compared to the transcriptome of a vegetative tissue (Fig. 1A). The uniqueness of the pollen transcriptome in

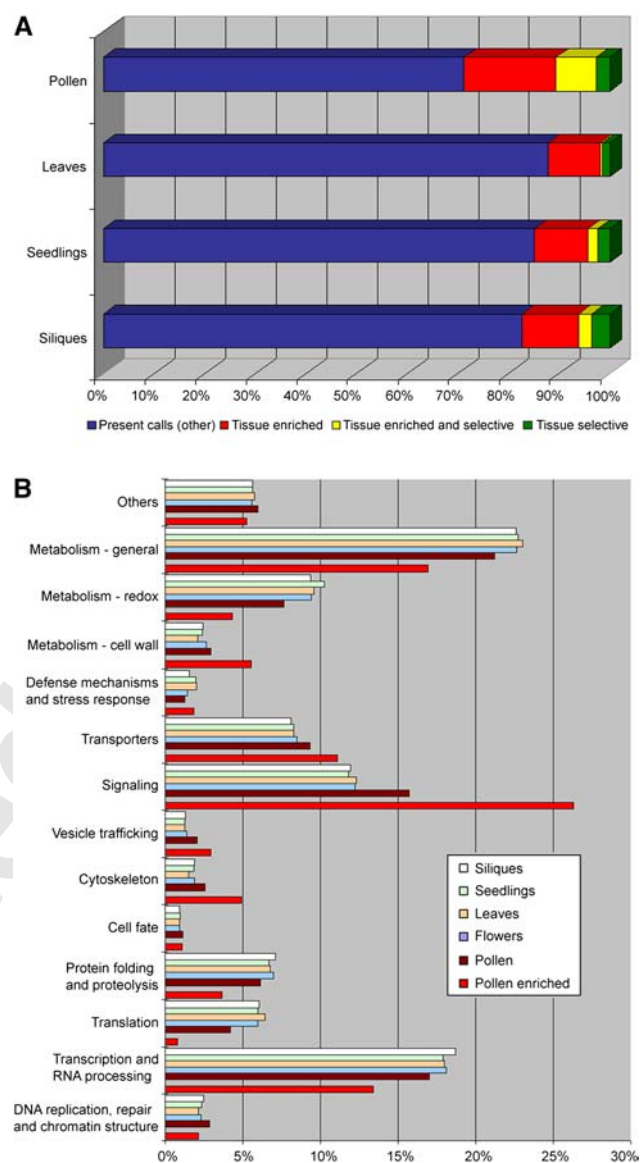
comparison to the vegetative ones is further underscored by a principal component analysis with 2,000 of the 3,734 genes called present in all arrays (Fig. 1B).

The 18,321 filtered genes were initially analyzed in terms of presence, enrichment, and selectivity of expression in the individual tissues. Tissue-selective genes are genes called present in only one of the tissues considered for this study. Comparisons of pollen grains to other *Arabidopsis* tissues excluded the flower samples because these contain pollen. The results of the comparisons of pollen with the vegetative tissues are given in Supplemental Table I. A false discovery rate of 0% (150 permutations) in all comparisons of the 2 pollen arrays to a set of 2 arrays of a vegetative tissue type confirms the robustness of the comparison criteria used.

The vast majority of the 6,587 genes expressed in pollen are also expressed in vegetative tissues; only 11% of pollen-expressed genes are pollen selective (739 genes). This proportion is nonetheless considerably larger than the proportion of selectively expressed genes in vegetative tissues, 2%, 4%, and 6%, respectively, in leaves, seedlings, and siliques (Fig. 2A). A comparison of our list of 739 pollen selectively expressed genes with sporophytic ATH1 datasets of more than 100 different developmental stages, tissues, and stress conditions in *Arabidopsis* ecotype Col-0 (AtGenExpress, excluding datasets of samples that



**Figure 1.** Snail view representation and principal component analysis of tissue-dependent gene expression patterns. A, 18,321 genes used for a snail view representation were ranked counterclockwise from top according to decreasing mean expression in seedlings. For each tissue, the mean pattern in seedlings (black line) is coplotted with the mean pattern of the specific tissue (gray dots). The radius encodes the logarithm of gene expression (values <1 were set to 1 for better visualization). B, Two thousand of the 3,734 genes expressed in the vegetative tissues and in pollen were analyzed by projecting the 5 tissue types (replicate datasets are shown) on the first 2 principal components. The first principal component separates pollen from the vegetative tissues, while the second principal component shows a further separation of the vegetative tissues within themselves.



**Figure 2.** Proportions of enriched and selectively expressed genes and classification of biological activities in *Arabidopsis* tissues. A, Genes called present in each of the *Arabidopsis* tissues were analyzed in terms of enriched and/or selective expression. Tissue-enriched genes have expression levels in any particular tissue that are at least 1.2-fold higher than in all the remaining tissues; selectively expressed genes are those called present only in the particular tissue analyzed. B, 8,463 genes represented on the ATH1 GeneChip that were classified into at least one GO category (as of September 2003) were regrouped into 14 biological activity classes. The transcription and RNA-processing classification includes transcription factors, basic transcriptional machinery, and RNA-processing factors; the cell fate category covers cell cycle, cell differentiation, and apoptosis-related molecules; transporters refers to membrane channels. The proportion of the genes expressed in each tissue assigned to the different activity classes is represented. The separated bars show this classification for the pollen-enriched genes.

contain pollen) reduces the list by more than one-half (data not shown). This finding supports the notion that genes we have classified here as pollen selectively expressed might also be expressed in tissues or under conditions we have not tested. Twenty-six percent (1,717 genes) of the pollen-expressed genes consist of pollen-enriched transcripts (Fig. 2A). About one-third of the pollen-enriched transcripts are also pollen selective and more than three-fourths of the selectively expressed genes in pollen are enriched transcripts as well. Those remaining consist of genes with comparable low levels of expression that are only considered to be present in the male gametophyte. Pollen has been described to continuously accumulate mRNA molecules from the stage of microspore mitosis all through the process of pollen grain maturation, and at least some of these transcripts should encode proteins involved in pollen germination and tube growth (Mascarenhas, 1990), which may provide an explanation for the high proportion of pollen-enriched transcripts.

#### The Pollen Transcriptome Is Functionally Skewed Toward Germination and Tube Growth

To determine whether the accumulated transcripts were functionally skewed toward the processes of pollen tube germination and growth, we classified genes expressed in the various *Arabidopsis* tissues and pollen-enriched genes into broad biological activities. We considered the 8,463 genes represented on the ATH1 GeneChip that were classified into one or more of GO categories (as of September 2003 annotation) and grouped them into 14 biological activity classes (Fig. 2B). Transcriptomes of vegetative tissues and flowers are covered in the different categories in a broadly identical way. Pollen-expressed genes, however, behave differently. Translation-related transcripts and, to a lesser extent, transcription-related ones, are proportionately underrepresented in pollen, while signaling, vesicle trafficking, the cytoskeleton, and membrane transport are proportionately overrepresented. The latter categories are clearly involved in the regulation of polarized tip growth of pollen tubes upon germination and along the female reproductive tract (Hepler et al., 2001). Cell wall metabolism is represented in an at least comparable level to vegetative tissues, while general and oxidative metabolism is underrepresented in the pollen expression profile. The apparently discordant behavior among categories of metabolic activity may be explained by the fact that cell wall remodeling is a particular feature of pollen tube growth. Furthermore, cell wall metabolism, the cytoskeleton, and signaling classes are overrepresented in pollen-enriched transcripts, as are transporter molecules and vesicle traffic elements. Enzymes involved in remodeling of the cell wall are among the genes with the highest expression values in pollen (Supplemental Table I). A similar expression pattern was previously described (Becker et al., 2003; Honys and Twell, 2003; Lee and Lee, 2003). Lee and Lee (2003) also analyzed

pollen-transcribed genes and their relative abundance in terms of functional classification in comparison to leaf. Data from the more comparable categories, namely, cellular biogenesis and signal transduction, revealed an overrepresentation in pollen-expressed genes. The reverse pattern was observed for classes like energy and protein synthesis, which is in agreement with our data. The transcription class in pollen, however, behaves similarly to leaf, both in terms of proportion of genes and of proportion of transcripts.

The dramatic increase in GO annotations of *Arabidopsis* genes (as of December 2003, covering 16,024 genes) prompted us to compare the frequency of individual GO terms in the 5 *Arabidopsis* tissues analyzed on a more global scale and add an analysis of statistical significance. The three categories of GO terms, function, process, and component, were considered separately and each level of the GO graph discriminated (see "Materials and Methods"). We performed tissue-to-tissue comparisons and calculated odds ratios as a measurement of the enrichment ( $>1$ ) or impoverishment ( $<1$ ) of term frequencies in any tissue 1 relative to any tissue 2. Statistical significance was determined using a two-tailed test for proportion difference between the two tissues (Supplemental Table II).

Table I shows statistically significant differences in frequencies of GO terms between pollen and the three vegetative tissues: leaves, seedlings, and siliques. The terms overrepresented in pollen can be assigned to the previously discussed biological activity classes of signaling, vesicle trafficking, and membrane transport. Terms impoverished in pollen grains relate to transcription, protein synthesis, and turnover and metabolism, again in agreement with the previous classification. Translation initiation factors are significantly enriched in the male gametophyte, while protein biosynthesis, ribonucleoprotein, and ribosome terms are impoverished; all these terms contribute to the biological activity class of translation and should not therefore contradict the detected underrepresentation of this class in pollen. The specific increased frequency of translation initiation factors when compared to the vegetative tissues may even reinforce the mechanistic notion that pollen stores functionally specialized transcripts, which then can be translated into proteins upon contact with the female tissues.

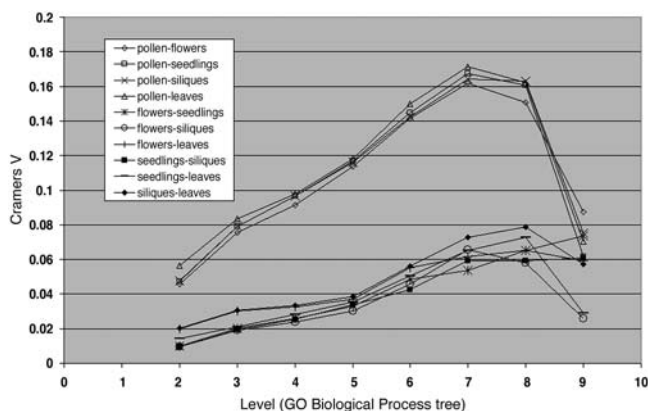
We calculated overall dissimilarities between any two *Arabidopsis* tissues discriminated by level and GO category using Cramer's V measure of association strength. This varies between 0 and 1, the former corresponding to equal relative frequency of terms in both tissues analyzed. Figure 3 shows the intertissue distances for every level of the GO process terms. Pollen clearly differs more from any of the vegetative tissues (including flowers) than these differ from one another, a finding reproduced in intertissue distances for function and component GO terms (data not shown). Overall, data on classification of biological activities in *Arabidopsis* pollen and vegetative tissues clearly support the idea of specialization of pollen

**Table 1.** GO terms differentially represented in pollen grains as compared to vegetative tissues

Relative frequencies of function, process, and component terms at the different levels of the GO trees were compared pairwise in pollen grains, leaves, seedlings, and siliques. For each comparison, genes were classified as either over- or underrepresented in one particular tissue; statistically significant differences between any two tissues were determined by using a two-tailed test for proportion differences ( $P$ -value  $< 0.05$ ). The terms for which statistical significance was found in all three pollen-to-vegetative tissue comparisons are shown.

[AQ13]

Function Term		
Level	Overrepresented in Pollen ( $P < 0.05^*$ )	Underrepresented in Pollen ( $P < 0.05^*$ )
2		Transcription regulator
3	Nucleotide binding	Structural protein
4	Intracellular transporter	Transcription factor
4	Purine nucleotide binding	Peptidase
	Transferase, transferring phosphorus-containing groups; EC 2.7.-.-	
	Protein kinase	
	Hydrolase, acting on acid anhydrides	
	Calcium binding	
	Translation initiation factor	
	Small GTPase regulatory/interacting protein	
5	Adenyl nucleotide binding	Endopeptidase
	Hydrolase, acting on acid anhydrides, in phosphorus-containing anhydrides	Aspartic-type endopeptidase
	Guanyl nucleotide binding	
	Phosphoric ester hydrolase	
	P-type ATPase	
6	ATP binding	Pepsin A
	ATPase	
	GTP binding	
	Phosphoric monoester hydrolase	
	GTPase	
7	Small monomeric GTPase	
	Protein phosphatase	
8	Protein Ser/Thr phosphatase	
Process Term		
Level	Overrepresented in Pollen ( $P < 0.05^*$ )	Underrepresented in Pollen ( $P < 0.05^*$ )
2	Cellular process	
3	Cell growth and maintenance	
	Cell communication	
4	Transport	Biosynthesis
	Signal transduction	Photosynthesis, light reactions
	Cell organization and biogenesis	
5	Protein transport	Macromolecule biosynthesis
	Intracellular transport	Protein biosynthesis
	Intracellular signaling cascade	
6	Intracellular protein traffic	
	Dephosphorylation	
	Protein amino acid dephosphorylation	
	Small GTPase-mediated signal transduction	
	Ribonucleotide biosynthesis	
	Nucleoside triphosphate metabolism	
	Nucleoside triphosphate biosynthesis	
	ER-to-Golgi transport	
Component Term		
Level	Overrepresented in Pollen ( $P < 0.05^*$ )	Underrepresented in Pollen ( $P < 0.05^*$ )
1		Cellular component
2	Unlocalized	Cell
3	Protein Ser/Thr phosphatase complex	Intracellular
4		Cytoplasm
		Ribonucleoprotein
5	Cytoskeleton	Ribosome
6		Chloroplast



**[AQ12] Figure 3.** Plot of intertissue distances for the different levels of the GO tree of biological process terms. Overall differences in term frequencies between any two *Arabidopsis* tissues at each level of the GO biological process tree were calculated using Cramer's V and interpreted as a measure of the distance or dissimilarity between the two tissues. Results vary between 0 and 1, lower values reflecting more similar relative frequencies of terms. A total of 16,024 genes represented on the ATH1 GeneChip and classified into GO categories (as of December 2003) were included in the analysis. The plot is representative of results obtained for biological function and component GO trees.

transcripts and show a disproportionate accumulation of mRNA molecules coding for proteins known to be important for pollen tube growth.

#### The Extensive Representation of Gene Families on the ATH1 Array Permits Detailed Functional Analyses

Apart from a global characterization, we were interested in a more detailed analysis of gene families and specific pathways with respect to their representation in the pollen transcriptome. Gene lists extracted from TAIR and recent publications were used to analyze gene families in the functional classes of signaling, transcription, transporter/channel, cell wall, cytoskeleton, vesicle trafficking, cell cycle, and small RNA pathways (Supplemental Table III). Coverage of the genes belonging to individual gene families and pathways on the ATH1 array frequently reached 80% and more, thus allowing a meaningful study of gene families and pathways in *Arabidopsis* pollen. In a complementary approach, the MAPMAN tool (Thimm et al., 2004) was used to map changes in expression levels in pollen compared to the vegetative tissues on metabolic pathways and other processes. The respective MAPMAN data file is available (see Supplemental Table IV). This detailed analysis presents a basis for valuable insights into biological processes of pollen, highlighted here by our results on nonclassical MADS box transcription factors, small RNA pathways, and the cell cycle in pollen.

#### Nonclassical MADS Box Genes Are Preferentially Expressed in the Male Gametophyte

Our GO analysis had revealed that transcription factors and RNA-processing machinery are underrep-

resented in pollen-expressed genes, a tendency becoming even more apparent when pollen-enriched transcripts are considered. This pattern is maintained across individual families of transcription factors (Supplemental Table III, transcription) with the noteworthy exception of MADS box genes, a family of transcription factors conserved in plants, fungi, and animals (Schwarz-Sommer et al., 1990).

**[AQ6]**

Of the 110 MADS box genes known (Alvarez-Buylla et al., 2000a; De Bodt et al., 2003a, 2003b; Kofuji et al., 2003; Martinez-Castilla and Alvarez-Buylla, 2003; Parenicova et al., 2003), 79 are represented in the Affymetrix ATH1 GeneChip, and 49 of these were included in our filtered gene list (Supplemental Table III, transcription). AGL6 and AGL7/AP1 are called present in flowers but absent in pollen grains, leaves, seedlings, and siliques, and were therefore excluded from further analysis. Seventeen (36%) out of the 47 MADS box genes are pollen expressed and 9 (53%) of these are pollen enriched, which stands for a clear overrepresentation of this family in the transcriptome of the *Arabidopsis* male gametophyte. Such overrepresentation of pollen-expressed and -enriched genes is restricted to the subgroup of genes that do not contain a classic MIKC gene structure: type I and MIKC\* genes (Table II; De Bodt et al., 2003a, 2003b; Kofuji et al., 2003; Parenicova et al., 2003). Four of the five MIKC\* genes included in our gene list are called present in pollen; all of these are pollen enriched, and three of the four genes are selectively expressed in pollen. Type I (non-MIKC\*) genes are also overrepresented in pollen grains; 6 of the 13 genes analyzed are expressed in pollen, 4 of these being both pollen enriched and undetectable in vegetative tissues. The function of both groups of nonclassical MADS box genes is unknown. Expression of the homeotic-like classic type II genes, on the contrary, is not overall increased in the *Arabidopsis* male gametophyte. These genes are known to be involved in the regulation of meristem and floral organ identity and cell type specification in floral organs; a distinct set of type II proteins are integrators of different pathways in the specification of flowering timing (Jack, 2001).

The preferential expression of the MIKC\* genes AGL104, AGL66, AGL30, and AGL65 in pollen grains is highly suggestive of a common or concurrent function in the male gametophyte and one might even speculate about a more generalized role in gametophytic tissues. Parenicova et al. (2003) detected AGL104 expression in septa and developing ovules as well as in developing anthers by *in situ* hybridization. Putative male gametophytic roles may also be inferred for type I AGL29, AGL84, At4g14530, and AGL49 genes, which we demonstrated to be preferentially expressed in pollen grains. Kofuji et al. (2003) detected similar expression patterns for the 4 MIKC\* genes and for AGL29 with macroarrays and northern blotting. Our analysis also identified seven classic type II genes as being expressed in pollen (Table II). Only one member of this group, AGL18, is pollen enriched.

**Table II.** Differential expression of type I MADS box transcription factors and cell cycle genes in pollen in comparison to vegetative tissues

The first column shows whether a gene is enriched in pollen when compared to the vegetative tissues, leaves, seedlings, and siliques, and the second gives the lower confidence bound of the fold change (average of the comparisons of pollen to the vegetative tissues). The third column depicts whether a gene is selectively expressed in pollen (present detection call only in pollen), followed by the Affymetrix probe set and the TAIR locus (AGI ID) assigned to this probe set in the fourth and fifth columns. The sixth column depicts the gene annotation. In the following columns, the expression value of the gene (weighted average of duplicates) and its detection call (P or A) are given for the respective cell type or tissue. For more details, see Supplemental Table III (transcription and cell cycle).

Transcription MADS Box Transcription Factors (Type I)													
Enriched	FC	Selective	Probe Set	AGI ID	Annotation	Pollen	Call	Leaf	Call	Seedling	Call	Silique	Call
X	19.8	X	266994_at	At2g34440	AGL29 (M/Mα)	915	P	33	A	30	A	32	A
X	4.5	X	248602_s_at	At5g49420	AGL84 (M/Mα)	49	P	6	A	5	A	15	A
X	2.6	X	245591_at	At4g14530	Type I MADS box (M)	62	P	19	A	17	A	21	A
X	2.0	X	263735_s_at	At1g60040	AGL49 (MADS-like/Mβ)	226	P	93	A	73	A	92	A
			247767_at	At5g58890	AGL82 (MADS-like/Mβ)	72	P	68	P	66	P	57	P
			247631_at	At5g60440	AGL62 (M/Mα)	41	P	27	P	21	P	30	P
Absent call in pollen: AGL23 (At1g65360), AGL28 (At1g01530), AGL33 (At2g26320), AGL37 (At1g65330), AGL40 (At4g36590), AGL87 (At1g22590), AGL96 (At5g06500)													
MADS Box Transcription Factors (MIKC*)													
Enriched	FC	Selective	Probe Set	AGI ID	Annotation	Pollen	Call	Leaf	Call	Seedling	Call	Silique	Call
X	49.5	X	255952_at	At1g22130	AGL104 (O/Mδ/MIKC*)	1569	P	24	A	23	A	21	A
X	13.7	X	262179_at	At1g77980	AGL66 (O/Mδ/MIKC*)	1331	P	75	A	62	A	54	A
X	4.5	X	266793_at	At2g03060	AGL30 (O/Mδ/MIKC*)	495	P	106	A	90	A	89	A
X	4.1		261423_at	At1g18750	AGL65 (O/Mδ/MIKC*)	1015	P	255	P	214	P	148	P
Absent call in pollen: AGL94 (At1g69540)													
MADS Box Transcription Factors (Type II)													
Enriched	FC	Selective	Probe Set	AGI ID	Annotation	Pollen	Call	Leaf	Call	Seedling	Call	Silique	Call
X	12.1		251623_at	At3g57390	AGL18 (MIKC)	4993	P	246	P	318	P	796	P
			264872_at	At1g24260	AGL9 (MIKC)	168	P	49	A	90	P	1318	P
			251312_at	At3g61120	AGL13 (MIKC)	100	P	71	A	62	P	50	P
		X	245819_at	At1g26310	AGL10 (MIKC)	99	P	105	A	93	A	70	A
			251672_at	At3g57230	AGL16 (MIKC)	80	P	365	P	231	P	179	P
			255014_at	At4g09960	AGL11 (MIKC)	50	P	203	A	166	A	2681	P
			247224_at	At5g65080	AGL68 (MIKC)	39	P	76	P	52	A	47	P
Absent call in pollen: AGL15 (At5g13790), ABS/TT16 (At5g23260), AGL31 (At5g65050), AGL5 (At2g42830), AG (At4g18960), AGL8 (At5g60910), AGL24 (At4g24540), AGL2 (At5g15800), AGL17 (At2g22630), FLC (At5g10140), FCL2 (At5g65070), AGL12 (At1g71692), AGL42 (At5g62165), AGL1 (At3g58780), AP3 (At3g54340), PI (At5g20240), AGL20 (At2g45660), AGL4 (At3g02310), AGL22 (At2g22540), AGL19 (At4g22950), AGL27 (At1g77080)													
Cell Cycle wCDK (Cyclin-Dependent Kinases)													
Enriched	FC	Selective	Probe Set	AGI ID	Annotation	Pollen	Call	Leaf	Call	Seedling	Call	Silique	Call
			262802_at	At1g20930	CDKB2;2	500	P	340	A	418	P	443	P
			256372_at	At1g66750	CDKD;2	481	P	402	P	400	P	399	P
			253756_at	At4g28980	CDKF;1	377	P	570	P	599	P	532	P
			247334_at	At5g63610	CDKE;1	372	P	461	P	402	P	606	P
			266401_s_at	At2g38620	CDKB1;2	260	P	158	A	385	P	553	P
			255909_at	At1g18040	CDKD;3	246	P	116	P	183	P	219	P
			252337_at	At3g48750	CDKA;1	111	P	672	P	647	P	553	P
Absent call in pollen: CDKB2;1 (At1g76540), CDKC;1 (At5g10270), CDKC;2 (At5g64960)													
Cyclin													
Enriched	FC	Selective	Probe Set	AGI ID	Annotation	Pollen	Call	Leaf	Call	Seedling	Call	Silique	Call
X	5.4		246385_at	At1g77390	CYCA1;2	599	P	67	A	97	P	85	A
X	1.9		262752_at	At1g16330	CYCB3;1	337	P	123	A	160	P	147	P
X	1.8		246938_at	At5g25380	CYCA2;1	87	P	30	A	44	P	52	P
			249129_at	At5g43080	CYCA3;1	332	P	212	P	287	P	248	P
			261780_at	At1g76310	CYCB2;4	190	P	147	A	291	P	238	P
			259290_at	At3g11520	CYCB1;3	146	P	99	A	323	P	669	P
			264043_at	At2g22490	CYCD2;1	142	P	251	P	236	P	158	P

(Table continues on following page.)

**Table II.** (Continued from previous page.)

Cyclin													
Enriched	FC	Selective	Probe Set	AGI ID	Annotation	Pollen	Call	Leaf	Call	Seedling	Call	Silique	Call
			245739_at	At1g44110	CYCA1;1	134	P	114	P	578	P	489	P
			264697_at	At1g70210	CYCD1;1	88	P	149	P	139	P	111	P
			246762_at	At5g27620	CYCH;1	66	P	159	P	225	P	195	P
			247190_at	At5g65420	CYCD4;1	51	P	30	A	63	P	57	P
Absent call in pollen: CYCA2;2 (At5g11300), CYCA2;3 (At1g15570), CYCA2;4 (At1g80370), CYCA3;2 (At1g47210), CYCA3;4 (At1g47230), CYCB1;1 (At4g37490), CYCB1;4 (At2g26760), CYCB2;1 (At2g17620), CYCB2;2 (At4g35620), CYCD3;1 (At4g34160), CYCD3;2 (At5g67260), CYCD3;3 (At3g50070), CYCD5;1 (At4g37630), CYCD6;1 (At4g03270)													
CKS (CDK/Cyclin Interactors and Regulatory Proteins)													
Enriched	FC	Selective	Probe Set	AGI ID	Annotation	Pollen	Call	Leaf	Call	Seedling	Call	Silique	Call
X	2.7		264061_at	At2g27970	CKS2	1196	P	203	P	804	P	845	P
			264070_at	At2g27960	CKS1	156	P	295	P	335	P	344	P
Absent call in pollen: KRP1 (At2g23430), KRP2 (At3g50630), KRP3 (At5g48820), KRP5 (At3g24810)													
RBR1, WEE1, and E2F-DP													
Enriched	FC	Selective	Probe Set	AGI ID	Annotation	Pollen	Call	Leaf	Call	Seedling	Call	Silique	Call
X	2.0		257524_at	At3g01330	DEL3	320	P	124	A	152	P	143	P
			256268_at	At3g12280	RBR1	255	P	255	P	344	P	320	P
			262106_at	At1g02970	WEE1	199	P	142	P	170	P	172	P
Absent call in pollen: E2Fa (At2g36010), E2Fb (At5g22220), E2Fc (At1g47870), Dpa (At5g02470), DPb (At5g03410), DEL2 (At5g14960)													

FC, Fold change; AGI, Arabidopsis Genome Initiative; P, Present; A, absent.

This gene was previously shown to have a gametophytic pattern of expression (Alvarez-Buylla et al., 2000b) and is phylogenetically related to the embryonically expressed AGL15, both clustering independently from the homeotic-like type II genes in several phylogenetic analyses (Alvarez-Buylla et al., 2000b; Kofuji et al., 2003; Martinez-Castilla and Alvarez-Buylla, 2003; Parenicova et al., 2003). This not only reinforces the notion of the distinctness of MADS box family genes expressed in pollen, but also raises the question of whether de novo transcription during germination is actually higher than indicated by early inhibitor studies (Mascarenhas, 1975). Supporting this notion, our data indicate that, for each transcription factor family, at least one member shows enriched expression in pollen.

#### Are Small RNA Pathways Inactivated in Tricellular and Mature Pollen?

The down-regulation of genes or even absent calls for transcripts in the pollen transcriptome can contain as much biological information as the up-regulation of transcriptional activity. Small RNA pathways in plants have been the focus of intensive studies during the last years and their importance has been shown for processes as diverse as antiviral defense, genome rearrangement, formation of heterochromatin, and developmental timing and patterning (Carrington and Ambros, 2003; Finnegan and Matzke, 2003; Lai, 2003). Recent studies have elucidated the function of specific genes in the formation and activity of micro-RNA (miRNA) and short-interfering RNA (siRNA; Xie et al., 2004), and in the control of de novo DNA

methylation (Chan et al., 2004). RNA interference (RNAi) has been successfully used for many studies of developmental pathways. However, while RNAi driven by the pollen-specific promoter LAT52 was shown to silence target genes when stably transformed into tobacco (*Nicotiana tabacum*; A. Cheung, personal communication), there are no references in the literature that the same can be achieved by transient transformation of mature pollen grains. This prompted us to analyze the expression of genes involved in small RNA pathways in Arabidopsis pollen. Surprisingly, all of the 15 genes analyzed, including Argonaute 1, 2, 4, and 7, Dicer-like 1-3 and RNA-dependent RNA polymerase 1, 2, and 6, were called absent in pollen, while the majority of them are expressed in the vegetative tissues tested (Table III; Supplemental Table III, small RNA pathways). This indicates that small RNA pathways might be inactivated in mature pollen grains.

Gupta et al. (2002) have reported the successful use of RNAi constructs to suppress the expression of AtPTEN1 in Arabidopsis during the late stage of pollen development. There may be consequences of the RNAi pathway being functional at least during some stages of pollen development, although no experimental data are available on the durability of these proteins. We reanalyzed data obtained by Honys and Twell (2004) regarding microgametogenesis from unicellular microspores to the tricellular pollen state. Most of the transcripts were called present in unicellular microspores and in bicellular pollen (Table III), similar to the vegetative tissues. In tricellular pollen, though, all of the 15 genes analyzed with the exception of Argonaute 1 were called absent. It is likely that the Argonaute 1 transcripts detected in tricellular pollen



**Table III.** Expression pattern of genes involved in small RNA pathways in vegetative tissues, mature pollen, and microgametogenesis

The first column indicates the annotation. The second and the third columns depict the Affymetrix probe set and the TAIR locus (AGI ID) assigned to this probe set. In the following columns, the call in the particular tissues and phases of microgametogenesis are symbolized by “+” (present call) and “–” (absent call). UNM, Uninuclear microspore; BCP, bicellular pollen; TCP, tricellular pollen; MPG, mature pollen grain (this study); \*, dataset of Honys and Twell (NASC arrays database), reanalyzed using MAS 5.0 and DNA-chip analyzer 1.3.

Annotation	Probe Set	AGI ID	Seedling	Leaf	Silique	Flower	UNM*	BCP*	TCP*	MPG
AGO1	262246_at	At1g48410	+	+	+	+	+	+	+	–
AGO2	262548_at	At1g31280	+	+	+	+	–	+	–	–
AGO4	266314_at	At2g27040	+	+	+	+	+	+	–	–
AGO7	256293_at	At1g69440	+	+	+	+	–	–	–	–
Putative AGO1 protein	264066_at	At2g27880	–	–	+	+	+	+	–	–
AGO1-related protein	267641_at	At2g32940	+	–	+	+	+	–	–	–
AGO1-similar protein	246025_at	At5g21150	–	–	+	+	+	+	–	–
DCL1	261584_at	At1g01040	+	+	+	+	+	+	–	–
DCL2	258863_at	At3g03300	+	+	+	+	+	+	–	–
DCL3	252716_at	At3g43920	+	+	+	+	+	+	–	–
HEN1	254449_at	At4g20910	+	+	+	+	+	+	–	–
RDR1	262888_at	At1g14790	+	+	+	+	+	+	–	–
RDR2	254933_at	At4g11130	+	+	+	+	+	+	–	–
RDR6	252261_at	At3g49500	+	+	+	+	+	+	–	–
SDE3	261381_at	At1g05460	+	+	+	+	+	+	–	–

samples are derived from impurities (12% of bicellular pollen; Honys and Twell, 2004) and not from tricellular pollen. Taken together, the results indicate that specific down-regulation of transcripts encoding proteins of small RNA pathways occurs during microgametogenesis in Arabidopsis. Furthermore, no other Arabidopsis Col-0 dataset of the AtGenExpress database shows the complete absence of the 15 transcripts analyzed here (data not shown). In support of our hypothesis of the inactivation of small RNA pathways in mature pollen, findings in other studies could be explained by this inactivation. First, early studies on RNA synthesis during pollen germination and tube growth of *Tradescantia paludosa* and lily have indicated that small RNAs about the same size as tRNA accumulate in pollen tubes, although no tRNA seems to be synthesized (Mascarenhas, 1975). The absence of RNA silencing in mature and germinating pollen would allow the accumulation of siRNA and/or miRNA precursors in pollen tubes, which would fall into this size range. Second, endogenous double-stranded RNAs in wild rice are maintained at a constant low concentration in every tissue of the plants, but their copy number increases by more than 10-fold in pollen grains (Moriyama et al., 1999), indicating a possible lack of copy number control through RNA silencing in pollen. The apparent absence of small RNA pathways in Arabidopsis mature pollen would affect cytoplasmic siRNA silencing (defense against viruses), the silencing of endogenous mRNAs by miRNAs, and heterochromatin silencing (Baulcombe, 2004), with possible broad implications for genetic and epigenetic processes in male gametophytes and fertilization.

At this point, we cannot exclude that the sperm cells still contain transcripts of the small RNA pathways that might not be detected in our study due to dilution

effects. In line with this notion, sequencing of almost 5,000 expressed sequence tags from a maize sperm cDNA library (Engel et al., 2003) has revealed expressed sequence tags with homologies to Arabidopsis Argonaute 1, 2, and 4 and Dicer-like 3 transcripts.

Our observations for RNA silencing in pollen and the lack of reports on successful silencing by ballistic bombardment of mature pollen so far suggest that RNAi-based studies in pollen grains should be initiated in early-stage pollen conditions. This could be achieved by stable transformation with an inducible RNAi construct under an early pollen-specific promoter (e.g. LAT52) or by ballistic bombardment and induction of an inducible construct at an early pollen stage. Transcripts resulting from de novo transcription after pollen germination might not be possible to silence with any of these strategies. Alternatively, antisense constructs might be successful as shown recently (Guyon et al., 2004).

#### Control of Cell Cycle Progression: Pollen Seems To Be Arrested in G1, While Expressing Transcripts for G2/M Transition

Sperm cells in Arabidopsis are in the S-phase at anthesis and continue through the cell cycle during pollen tube growth to reach G2 just prior to fertilization (Friedman, 1999). The vegetative nucleus is thought to be arrested in G1. The vegetative tissue profiles in this study result from several different cell types in different phases of the cell cycle. Especially for the seedling and silique samples, which show higher growth rates than the fully developed rosette leaves, the expression data for cell cycle genes should show average values of all phases of the cell cycle and thus represent a good baseline for the analysis of cell cycle

genes in pollen. The ATH1 GeneChip contains probe sets for more than 90% (55) of the core cell cycle genes predicted in a recent genome-wide analysis in Arabidopsis (Vandepoele et al., 2002). Twenty-three of the 55 genes analyzed are expressed in pollen grains (Table II). Combining these data with a recent schematic overview of the mechanistic regulation of the G1/S transition in plants (De Veylder et al., 2003), a hypothesis emerges as to how pollen grains are kept arrested in G1 (Fig. 4A): The block could be achieved by a combination of absence of essential components needed for the G1/S transition and an up-regulation of potential repressors and of potential factors of increased cell cycle duration.

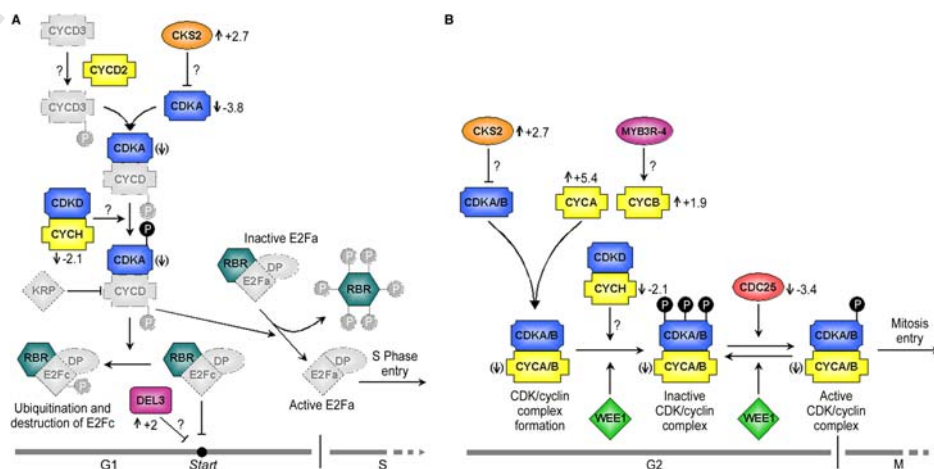
While our analysis is affected by the notion that the presence or absence of a given transcript does not necessarily have mechanistic implications for the encoded protein, we do, however, find a coherent pattern of a putative regulatory scheme expressed in our analysis of the pollen transcriptome.

Most striking is the complete absence of transcripts encoding for the subunits of the heterodimeric adenovirus E2 promoter-binding protein-dimerization partner (E2F-DP) transcription factor required for the activation of cell cycle and DNA synthesis genes (Fig. 4A). In addition, no D3-type cyclin is expressed in pollen, including CYCD3;1, which has been characterized as promoting S-phase entry (Schnittger et al., 2002; Dewitte et al., 2003). D-type cyclins can form complexes with A-type cyclin-dependent kinases (CDK). CDKA 1 is called present in pollen, but with a much lower expression level than in the vegetative tissues. The retinoblastoma (RB) tumor suppressor-related protein RBR1 is expressed in pollen, but may be kept in its nonphosphorylated state due to the absence of the CDKA/CYCD complex. In their nonphosphorylated form, RB proteins can repress the promoter activity of E2F-DP target genes (Weinberg, 1995). The expression of E2F-responsive genes in pollen seems to be repressed not only by the simple lack of E2F-DP transcripts and the nonphosphorylated

form of RBR1, but also by the up-regulation of expression of the gene encoding the DP-E2F-like protein DEL3 (Fig. 4A). DEL proteins lack a transcriptional activation domain and can bind to E2F-DP binding sites as monomers, antagonizing transcriptional activation by E2F-DP proteins (Kosugi and Ohashi, 2002; Mariconti et al., 2002). The expression of DEL1 and DEL3 before the transcription of S-phase genes (Mariconti et al., 2002) gives them a potential role in restraining cell proliferation by repressing transcription in G1. The CDK interacting proteins CKS1 and CKS2 are expressed in pollen, with CKS2 being enriched. Overexpression of CKS1 leads to growth inhibition in roots through increased cell cycle duration, associated with an equal extension of both the G1 and G2 phases (De Veylder et al., 2001). If we assume a similar function for CKS2, its overexpression in pollen might be an additional factor inhibiting cell cycle progression.

The detailed analysis of the complement of mRNAs in pollen encoding CDKs and cyclins reveals a rather unexpected picture. Following a schematic overview of the mechanistic regulation of the G2/M transition (De Veylder et al., 2003), pollen seems to feature most of the transcripts needed for the G2/M transition, although they seem to be kept in their inactive state (Fig. 4B). The transcripts include CDKA/B, CYCA, CYCB, CDKD, and CYCH (Table II). CYCA1;2, CYCA2;1, and CYCB3;1 are enriched in pollen. The latter two transcripts have been characterized recently as showing a peak in transcription in the M-phase of the cell cycle (Menges et al., 2003). In addition, the c-myc-like transcription factor AtMYB3R-4 (At5g11510; Supplemental Table III) is expressed in pollen. AtMYB3R-4 is the closest homolog of the tobacco NtmybA1 and NtmybA2 genes. NtmybA1 and NtmybA2 are specifically expressed at the G2/M transition and act as activators on M-specific activator elements in the promoters of plant B-type cyclins (Ito et al., 2001). The CDKA/B-CYCA/B complex is supposedly kept in its inactive form through phosphorylation by WEE1

**Figure 4.** Overview of potential cell cycle regulation in pollen. Pollen gene expression data for core genes of the cell cycle (see also Table II) are overlaid on a schematic overview of the G1/S (A) and G2/M (B) transitions. For proteins colored light gray, no transcripts were detected in pollen (absent call). Transcripts expressed in pollen (present call) are either expressed at a comparable level to the vegetative tissues or enriched (arrow up; fold change value) or depleted (arrow down; fold change value). Question marks indicate pathways that are still not fully demonstrated experimentally. (Modified from De Veylder et al., 2003).



kinase, expressed in pollen at levels equal to the levels in the vegetative tissues. A potentially antagonistic dual-specificity CDC25 phosphatase (At5g03455) has recently been described in *Arabidopsis* (Landrieu et al., 2004), but its expression is down-regulated in pollen, which would favor the inactive form of the complex. In addition, the overexpression of CKS2 might inhibit G2/M cell cycle progression in a similar way as assumed for the G1/S transition (Fig. 4B). In light of G2 karyogamy in *Arabidopsis*, our data might indicate a key role for pollen-derived transcripts or proteins during the first mitosis after fertilization. They could complement

G2/M factors of the egg cell and thus initiate the cell cycle in G2/M in the zygote. Whether the pollen transcripts or proteins would get into the egg cell via the released pollen cytoplasm or via the sperm cells remains to be clarified, e.g. by transcriptional profiling of isolated *Arabidopsis* sperm cells.

In support of our results on cell cycle transcripts in pollen, Hennig et al. (2004) have analyzed expression levels of the same set of core cell cycle genes during three stages of flower and fruit development in *Arabidopsis* and have come to the conclusion that S-phase genes were underrepresented and G2/M-phase genes enriched in a set of genes they identified as specifically expressed during reproduction. Since more than 80% of the genes in this floral-specific set were expressed in anthers and pollen (Hennig et al., 2004), it is likely that their result is largely based on the transcription levels of cell cycle genes in pollen as we describe them here.

## CONCLUSIONS

Our comparative study has increased the number of genes identified recently as expressed in *Arabidopsis* pollen (Becker et al., 2003) by more than 4-fold, with 26% of these 6,587 genes being enriched and 11% being selectively expressed. Pollen expresses a comparatively reduced set of genes; such a set is, however, tightly linked to its main biological roles of germination and tube growth and establishes pollen as a preferred model for the study of cell growth and morphogenesis (Feijó et al., 2001, 2004). The high genome coverage of the ATH1 GeneChip allowed the analysis of gene families and pathways. We have used this added value to set forth hypotheses regarding MADS box transcription factors, small RNA pathways, and cell cycle regulation in pollen. But the gene lists presented here may serve as starting points for the analysis of other aspects of pollen germination and pollen tube growth, including signal transduction, cell wall biosynthesis, and vesicle trafficking. In combination with studies on pollen physiology and biochemistry, the genetic basis unveiled in this study should bring us a step closer to a system's view of pollen germination and tube growth, in particular, and apical cell growth, in general.

## MATERIALS AND METHODS

### Plant Material and Growth Conditions

*Arabidopsis* (*Arabidopsis thaliana*) ecotype Col-0 was used in this study. To minimize interplant variability, tissues from a minimum of 12 plants were pooled for each RNA extraction. For the seedling samples, seeds were surface sterilized and then spread on petri dishes containing B-5 medium (Duchefa, Haarlem, The Netherlands) solidified with 0.8% phytagar (Duchefa). The seeds were cold treated for 3 d at 4°C to ensure uniform germination. The plates were transferred to short-day conditions (8 h of light at 21°C–23°C) and grown in a horizontal position for 7 d. Per seedling sample, more than 25 seedlings from 5 petri dishes were collected after 7 d of growth.

Plants for the leaf, flower, silique, and pollen samples were grown on soil for 12 weeks in short-day conditions (8 h of light at 21°C–23°C) and then changed to long-day conditions (16 h of light) to induce flowering. After bolting, more than 20 rosette leaves from different plants per leaf sample were collected. Flowers and siliques were harvested 2 weeks later. Twenty siliques and flowers, respectively, from different plants were pooled per sample.

### Isolation and Fluorescence-Activated Cell Sorting of Pollen Grains

Isolation and purification of hydrated pollen grains by flow cytometric cell sorting were carried out as described in Becker et al. (2003).

### RNA Isolation, Target Synthesis, and Hybridization to Affymetrix GeneChips

Total RNA was extracted using the RNeasy mini kit (Qiagen, Hilden, Germany). Concentration and purity were determined by spectrophotometry and integrity was confirmed using an Agilent 2100 bioanalyzer with a RNA 6000 nano assay (Agilent Technologies, Palo Alto, CA).

RNA was processed for use on Affymetrix (Santa Clara, CA) *Arabidopsis* ATH1 genome arrays, according to the manufacturer's Small Sample Labeling Protocol version II described in the technical note "GeneChip Eukaryotic Small Sample Target Labeling Assay Version II." Briefly, 100 ng of total RNA were used in a reverse transcription reaction (SuperScript II; Invitrogen, Paisley, UK) to generate first-strand cDNA. After second-strand synthesis, double-stranded cDNA was used in an in vitro transcription reaction to generate cRNA (MEGAscript T7 kit; Ambion, Austin, TX); 400 ng of this cRNA were used for a second cDNA synthesis, followed by a second in vitro transcription reaction to generate biotinylated cRNA (ENZO BioArray High-Yield RNA Transcript Labeling kit; ENZO Diagnostics, Farmingdale, NY). Size distribution of the cRNA and fragmented cRNA, respectively, was assessed using an Agilent 2100 bioanalyzer with a RNA 6000 nano assay.

Fifteen micrograms of cRNA were used in a 300-μL hybridization containing added hybridization controls; 200 μL of mixture were hybridized on arrays for 16 h at 45°C. Standard posthybridization wash and double-stain protocols (EukGE-WS2v4) were used on an Affymetrix GeneChip Fluidics Station 400. Arrays were scanned on an Affymetrix GeneChip scanner 2500.

### GeneChip Data Analysis

Scanned arrays were analyzed first with Affymetrix MAS 5.0 software to obtain absent/present calls and for subsequent analysis with dChip 1.3 (<http://www.dchip.org>, Wong Lab, Harvard). The following conditions were applied to ensure reliability of the analyses: First, each GeneChip experiment was performed with biological replicates. Second, we used a samplewise normalization to the median median probe cell intensity (CEL) of all arrays: For each sample, the median CEL intensity of one of the replicates was scaled to the median median CEL intensity of all arrays (in this study, 10 arrays with a median median CEL intensity of 147). Then the remaining replicates were normalized to this array (baseline) applying an Invariant Set Normalization Method (Li and Wong, 2001). In this way, a robust normalization could be achieved with calculated correlation coefficients for the various replicates consistently above 0.99. Third, normalized CEL intensities of the 10 arrays were used to obtain model-based gene expression indices based on a plasma membrane-only model (Li and Hung Wong, 2001). Potential array outliers that were detected in all replicates of one tissue/cell type were not called array outliers because we assumed that these genes might be expressed in a

[AQ7]

[AQ8]

tissue-specific manner and therefore are not true array outliers. Replicate data for the same sample were weighted genewise by using inverse-squared SE as weights. Only genes called present in at least 1 of the 10 arrays and within replicate arrays called present within a variation of  $0 < \text{median}(\text{SD}/\text{mean}) < 0.5$  were kept for downstream analysis (18,321 genes). Thus, genes called absent in all arrays and genes with highly inconsistent expression levels within replicate arrays were excluded. Finally, all genes compared were considered to be differentially expressed if they were called present in at least one of the arrays and if the 90% lower confidence bound of the fold change between experiment and baseline was above 1.2. The lower confidence bound criterion means that we can be 90% confident that the fold change is a value between the lower confidence bound and a variable upper confidence bound. Li and Hung Wong (2001) have shown that the lower confidence bound is a conservative estimate of the fold change and therefore more reliable as a ranking statistic for changes in gene expression. This criterion has been used in other gene expression studies (Ramalho-Santos et al., 2002; Becker et al., 2003).

Annotations for the approximately 22,750 genes represented on the Arabidopsis ATH1 genome array were obtained from the NetAffx database (www.affymetrix.com) as of June 2003.

## GO Analysis

The vocabulary used in the annotation of Arabidopsis gene properties is in agreement with the GO consortium. This consortium defines the list of terms that may be used and organizes them in a graph, where each term is a node and relations between terms are represented by edges linking the corresponding nodes. Edges are directed; that is, they depart from general terms to more specific ones. The GO graph is divided into 3 main trees: (1) molecular function terms (describing the function of the gene product at a biochemical level); (2) biological process terms (identifying the pathway or cellular task to which the gene product contributes); and (3) cellular component terms (describing the compartment or localization where the gene product is found in the cell). The specificity of a given term can be measured by its level: the minimum number of edges connecting it to the more general term in the tree, that is, the root node (e.g. the molecular function tree root node is "molecular function").

For any term in the GO directed acyclic graph, the number of associated genes in the tissue-expressed sets was determined. A gene was considered to be annotated with a given term if it was associated with that exact term or with any of its offshoots. Only GO terms occurring in at least one of the tissues were subsequently analyzed. For every term in each of the possible pairwise tissue comparisons, a normal proportions difference test was performed (Daniel, 1999). According to the null hypothesis, the frequency of occurrence of the GO term in both tissue gene lists was the same. Small test *P*-values are indicative of divergence in the distribution of genes associated with the respective GO terms in the two tissues compared. As the main goal of this procedure is to identify which annotations are specifically related to pollen, only GO terms with *P*-values lower than 0.05 in all 3 pollen-to-vegetative tissue comparisons were considered significantly over- or underrepresented. This stringent selection can be interpreted as a biological correction for the multiple testing error rate. Classic corrections, like Bonferroni procedures, are known to be too conservative in gene annotation analysis (Shah and Fedoroff, 2004). In addition to the significance testing for each individual GO term, overall annotation frequency dissimilarities between tissue gene lists were calculated. For each GO directed acyclic graph level and for each of the three GO categories (molecular function, biological process, and cellular component) a Cramer's *V* value was determined (Daniel, 1999). This measure of association strength varies between 0 and 1. In this case, a strong association means very different annotation frequencies in both tissues and a Cramer's *V* close to 1. If the 2 tissues have present gene lists similarly annotated, the Cramer's *V* should be close to 0.

Sequence data from this article have been deposited with the EMBL/GenBank data libraries under accession numbers ■■■.

## ACKNOWLEDGMENTS

We are grateful to Jorge Carneiro (Instituto Gulbenkian de Ciência) for his help with the snail views.

Received December 8, 2004; returned for revision February 2, 2005; accepted February 3, 2005.

## LITERATURE CITED

- Alvarez-Buylla ER, Liljegren SJ, Pelaz S, Gold SE, Burgeff C, Ditta GS, Vergara-Silva E, Yanofsky MF (2000a) MADS-box gene evolution beyond flowers: expression in pollen, endosperm, guard cells, roots and trichomes. *Plant J* 24: 457–466
- Alvarez-Buylla ER, Pelaz S, Liljegren SJ, Gold SE, Burgeff C, Ditta GS, Ribas de Pouplana L, Martinez-Castilla L, Yanofsky MF (2000b) An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. *Proc Natl Acad Sci USA* 97: 5328–5333
- Baulcombe D (2004) RNA silencing in plants. *Nature* 431: 356–363
- Becker JD, Boavida LC, Carneiro J, Haury M, Feijo JA (2003) Transcriptional profiling of Arabidopsis tissues reveals the unique characteristics of the pollen transcriptome. *Plant Physiol* 133: 713–725
- Carrington JC, Ambros V (2003) Role of microRNAs in plant and animal development. *Science* 301: 336–338
- Chan SW, Zilberman D, Xie Z, Johansen LK, Carrington JC, Jacobsen SE (2004) RNA silencing genes control de novo DNA methylation. *Science* 303: 1336
- da Costa-Nunes JA, Grossniklaus U (2004) Unveiling the gene-expression profile of pollen. *Genome Biol* 5: 205
- Daniel WW (1999) *Biostatistics: A Foundation for Analysis in the Health Sciences*. Wiley, New York
- De Bodt S, Raes J, Florquin K, Rombauts S, Rouze P, Theissen G, Van de Peer Y (2003a) Genomewide structural annotation and evolutionary analysis of the type I MADS-box genes in plants. *J Mol Evol* 56: 573–586
- De Bodt S, Raes J, Van de Peer Y, Theissen G (2003b) And then there were many: MADS goes genomic. *Trends Plant Sci* 8: 475–483
- De Veylder L, Beemster GT, Beeckman T, Inze D (2001) CKS1At overexpression in *Arabidopsis thaliana* inhibits growth by reducing meristem size and inhibiting cell-cycle progression. *Plant J* 25: 617–626
- De Veylder L, Joubes J, Inze D (2003) Plant cell cycle transitions. *Curr Opin Plant Biol* 6: 536–543
- Dewitte W, Riou-Khamlichi C, Scofield S, Healy JMS, Jacqumard A, Kilby NJ, Murray JAH (2003) Altered cell cycle distribution, hyperplasia, and inhibited differentiation in Arabidopsis caused by the D-type cyclin CYCD3. *Plant Cell* 15: 79–92
- Engel ML, Chaboud A, Dumas C, McCormick S (2003) Sperm cells of *Zea mays* have a complex complement of mRNAs. *Plant J* 34: 697–707
- Feijó JA, Costa SS, Prado AM, Becker JD, Certal AC (2004) Signalling by tips. *Curr Opin Plant Biol* 7: 589–598
- Feijó JA, Sainhas J, Holdaway-Clarke T, Cordeiro MS, Kunkel JG, Hepler PK (2001) Cellular oscillations and the regulation of growth: the pollen tube paradigm. *Bioessays* 23: 86–94
- Finnegan EJ, Matzke MA (2003) The small RNA world. *J Cell Sci* 116: 4689–4693
- Friedman WE (1999) Expression of the cell cycle in sperm of Arabidopsis: implications for understanding patterns of gametogenesis and fertilization in plants and other eukaryotes. *Development* 126: 1065–1075
- Gupta R, Ting JT, Sokolov LN, Johnson SA, Luan S (2002) A tumor suppressor homolog, AtPTEN1, is essential for pollen development in Arabidopsis. *Plant Cell* 14: 2495–2507
- Guyon V, Tang W-h, Monti MM, Raiola A, Lorenzo GD, McCormick S, Taylor LP (2004) Antisense phenotypes reveal a role for SHY, a pollen-specific leucine-rich repeat protein, in pollen tube growth. *Plant J* 39: 643–654
- Hennig L, Gruissem W, Grossniklaus U, Kohler C (2004) Transcriptional programs of early reproductive stages in Arabidopsis. *Plant Physiol* 135: 1765–1775
- Hepler PK, Vidali L, Cheung AY (2001) Polarized cell growth in higher plants. *Annu Rev Cell Dev Biol* 17: 159–187
- Hony D, Twell D (2003) Comparative analysis of the Arabidopsis pollen transcriptome. *Plant Physiol* 132: 640–652
- Hony D, Twell D (2004) Transcriptome analysis of haploid male gametophyte development in Arabidopsis. *Genome Biol* 5: R85
- Ito M, Araki S, Matsunaga S, Itoh T, Nishihama R, Machida Y, Doonan JH, Watanabe A (2001) G2/M-phase-specific transcription during the plant cell cycle is mediated by c-Myb-like transcription factors. *Plant Cell* 13: 1891–1905
- Jack T (2001) Plant development going MADS. *Plant Mol Biol* 46: 515–520
- Kofuji R, Sumikawa N, Yamasaki M, Kondo K, Ueda K, Ito M, Hasebe M (2003) Evolution and divergence of the MADS-box gene family

[AQ11]

[AQ9]

[AQ10]

- based on genome-wide expression analyses. *Mol Biol Evol* **20**: 1963–1977
- Kosugi S, Ohashi Y** (2002) E2Fs, E2F-like repressors of *Arabidopsis* that bind to E2F sites in a monomeric form. *J Biol Chem* **277**: 16553–16558
- Lai EC** (2003) microRNAs: runts of the genome assert themselves. *Curr Biol* **13**: R925–R936
- Landrieu I, da Costa M, De Veylder L, Dewitte F, Vandepoele K, Hassan S, Wieruszkeski JM, Faure JD, Van Montagu M, Inze D, et al** (2004) A small CDC25 dual-specificity tyrosine-phosphatase isoform in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* **101**: 13380–13385
- Lee JY, Lee DH** (2003) Use of serial analysis of gene expression technology to reveal changes in gene expression in *Arabidopsis* pollen undergoing cold stress. *Plant Physiol* **132**: 517–529
- Leonhardt N, Kwak JM, Robert N, Waner D, Leonhardt G, Schroeder JI** (2004) Microarray expression analyses of *Arabidopsis* guard cells and isolation of a recessive abscisic acid hypersensitive protein phosphatase 2C mutant. *Plant Cell* **16**: 596–615
- Li C, Hung Wong W** (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* **2**: R32
- Li C, Wong WH** (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA* **98**: 31–36
- Liu WM, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho MH, Baid J, et al** (2002) Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* **18**: 1593–1599
- Mariconti L, Pellegrini B, Cantoni R, Stevens R, Bergounioux C, Cella R, Albani D** (2002) The E2F family of transcription factors from *Arabidopsis thaliana*. Novel and conserved components of the retinoblastoma/E2F pathway in plants. *J Biol Chem* **277**: 9911–9919
- Martinez-Castilla LP, Alvarez-Buylla ER** (2003) Adaptive evolution in the *Arabidopsis* MADS-box gene family inferred from its complete resolved phylogeny. *Proc Natl Acad Sci USA* **100**: 13407–13412
- Mascarenhas JP** (1975) The biochemistry of angiosperm pollen development. *Bot Rev* **41**: 259–314
- Mascarenhas JP** (1989) The male gametophyte of flowering plants. *Plant Cell* **1**: 657–664
- Mascarenhas JP** (1990) Gene activity during pollen development. *Annu Rev Plant Physiol Plant Mol Biol* **41**: 317–338
- McCormick S** (1993) Male gametophyte development. *Plant Cell* **5**: 1265–1275
- McCormick S** (2004) Control of male gametophyte development. *Plant Cell* **16**: S142–S153
- Menges M, Hennig L, Gruissem W, Murray JAH** (2003) Genome-wide gene expression in an *Arabidopsis* cell suspension. *Plant Mol Biol* **53**: 423–442
- Moriyama H, Horiuchi H, Koga R, Fukuhara T** (1999) Molecular characterization of two endogenous double-stranded RNAs in rice and their inheritance by interspecific hybrids. *J Biol Chem* **274**: 6882–6888
- Parenicova L, de Folter S, Kieffer M, Horner DS, Favalli C, Busscher J, Cook HE, Ingram RM, Kater MM, Davies B, et al** (2003) Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in *Arabidopsis*: new openings to the MADS world. *Plant Cell* **15**: 1538–1551
- Ramalho-Santos M, Yoon S, Matsuzaki Y, Mulligan RC, Melton DA** (2002) “Stemness”: transcriptional profiling of embryonic and adult stem cells. *Science* **298**: 597–600
- Redman JC, Haas BJ, Tanimoto G, Town CD** (2004) Development and evaluation of an *Arabidopsis* whole genome Affymetrix probe array. *Plant J* **38**: 545–561
- Schmid M, Uhlenhaut NH, Godard F, Demar M, Bressan R, Weigel D, Lohmann JU** (2003) Dissection of floral induction pathways using global expression analysis. *Development* **130**: 6001–6012
- Schnittger A, Schobinger U, Bouyer D, Weigl C, Stierhof YD, Hulskamp M** (2002) Ectopic D-type cyclin expression induces not only DNA replication but also cell division in *Arabidopsis* trichomes. *Proc Natl Acad Sci USA* **99**: 6410–6415
- Schwarz-Sommer Z, Huijser P, Nacken W, Saedler H, Sommer H** (1990) Genetic control of flower development by homeotic genes in *Antirrhinum majus*. *Science* **250**: 931–936
- Scott R, Hodge R, Paul W, Draper J** (1991) The molecular biology of anther differentiation. *Plant Sci* **80**: 167–191
- Shah NH, Fedoroff NV** (2004) CLENCH: a program for calculating Cluster ENrichment using the gene ontology. *Bioinformatics* **20**: 1196–1197
- Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY, Stitt M** (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J* **37**: 914–939
- Twiss D** (2002) The developmental biology of pollen. In D O'Neill, JA Roberts, eds, *Plant Reproduction*, Vol 6. Sheffield Academic Press Ltd, Sheffield, UK, pp 86–153
- Vandepoele K, Raes J, De Veylder L, Rouze P, Rombauts S, Inze D** (2002) Genome-wide analysis of core cell cycle genes in *Arabidopsis*. *Plant Cell* **14**: 903–916
- Wang M-L, Hsu C-M, Chang L-C, Wang C-S, Su T-H, Huang Y-JJ, Jiang L, Jauh G-Y** (2004) Gene expression profiles of cold-stored and fresh pollen to investigate pollen germination and growth. *Plant Cell Physiol* **45**: 1519–1528
- Weinberg RA** (1995) The retinoblastoma protein and cell cycle control. *Cell* **81**: 323–330
- Xie Z, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, Zilberman D, Jacobsen SE, Carrington JC** (2004) Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol* **2**: E104
- Yamamoto Y, Nishimura M, Hara-Nishimura I, Noguchi T** (2003) Behavior of vacuoles during microspore and pollen development in *Arabidopsis thaliana*. *Plant Cell Physiol* **44**: 1192–1201