

**UCLA**

**UCLA Electronic Theses and Dissertations**

**Title**

Using Genome-Wide Approaches to Dissect Seed Development

**Permalink**

<https://escholarship.org/uc/item/8fn6j7zv>

**Author**

Le, Brandon

**Publication Date**

2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Using Genome-Wide Approaches To Dissect Seed Development

A dissertation submitted in partial satisfaction of the  
requirements for the degree Doctor of Philosophy  
in Molecular, Cell, and Developmental Biology

by

Brandon Hieu Le

2013



© Copyright by  
Brandon Hieu Le  
2013

## ABSTRACT OF THE DISSERTATION

Using Genome-Wide Approaches to Dissect Seed Development

by

Brandon Hieu Le

Doctor of Philosophy in Molecular, Cell, and Developmental Biology

University of California, Los Angeles, 2013

Professor Robert B. Goldberg, Chair

Seeds are a major source of food for human and animal consumption and can account for a large majority of the caloric intake by human. A detailed understanding of seed development should provide strategies for improving seed yield or enhance seed nutrients. Seed development consists of three major structure -- the embryo, endosperm, and seed coat -- all developing in parallel. Each of these structures undergo major developmental and physiological changes during seed development; however, it is not known how the different developmental programs in these structure are coordinated to form a seed.

How are genes regulated during seed development? What are the transcription factors playing a major role in controlling seed development? Are there genome-wide methylation changes during seed development, and if so, how do these changes affect seed activity? These were the driving questions that formed the basis of my Ph.D. research. To begin to address these questions, I used a variety of genomics tools [e.g., Sanger sequencing, next generation sequencing (NGS), GeneChip microarrays, laser capture microdissection (LCM)], with several plant species [e.g., *Arabidopsis*, soybean, scarlet runner bean (SRB)], to characterize seed transcriptomes and methylomes on a genome-wide basis. In addition, I used

reverse genetics to knockout seed-specific transcription factors and determine what role they may play in seed development. Collectively, data generated from my Ph.D. research can contribute to understanding gene activity during seed development.

The dissertation of Brandon Hieu Le is approved.

Ann M. Hirsch

Matteo Pellegrini

John J. Harada

Robert B. Goldberg, Committee Chair

University of California, Los Angeles

2013

*In loving memory of my dad, David  
To my mom, Nancy and my wife, Bekah*

*I dedicate the following Vietnamese proverb to my parents*

*Cong Cha nhu nui Thai Son  
Nghia Me nhu nuoc trong nguon chay ra  
Mot long tho Me kinh Cha  
Cho tron chu hieu moi la dao con*

## TABLE OF CONTENTS

List of Figures .....	viii
List of Tables .....	xv
Supplementary Files .....	xvi
Acknowledgements .....	xvii
Vita .....	xxi
Publications and Presentations .....	xxi
<b>Introduction</b>	
Introduction to the Dissertation .....	1
<b>Chapter One</b>	
Using Genomics to Dissect Legume Seed Development .....	10
References .....	21
<b>Chapter Two</b>	
Global Analysis of Gene Activity in <i>Arabidopsis</i> Seed Development and Identifying Seed-Specific Transcription Factors .....	24
References .....	32
<b>Chapter Three</b>	
Global Analysis of <i>lec1</i> Mutant During <i>Arabidopsis</i> Seed Development .....	41
References .....	53
<b>Chapter Four</b>	
Dynamic Changes in DNA Methylation During Soybean Seed Development ...	62
References .....	83
<b>Appendices</b>	
<b>A</b> - Comprehensive Developmental Profiles of Gene Activity in Regions and Subregions of the <i>Arabidopsis</i> Seed .....	119
References .....	129
<b>B</b> - Identification of Quantitative Trait Loci Controlling Gene Expression	

During the Innate Immunity Response of Soybean .....	138
References .....	149
<b>C</b> - Identification of Putative <i>Arabidopsis</i> DEMETER Target Genes By GeneChip Analysis .....	151
References .....	156
<b>D</b> - Genes Directly Regulated By LEAFY COTYLEDON2 Provide Insight Into the Control of Embryo Maturation and Somatic Embryogenesis .....	157
References .....	163
<b>E</b> - Differentiation and Degeneration of Cells That Play a Major Role in Tobacco Anther Dehiscence .....	164
References .....	186
<b>F</b> - An Atlas of Gene Activity in Soybean Seed Regions, Compartments, and Tissues During Development .....	188
<b>G</b> - Gene Activity in Different Regions of a Globular Stage SRB Embryo By EST Sequencing .....	197
<b>H</b> - ESTDB - A Web-Based Relational Database For Analyzing and Storing DNA Sequences .....	205
<b>I</b> - Functional Annotation of Genes Represented on the <i>Arabidopsis</i> and Soybean GeneChip Arrays .....	214
<b>J</b> - Design and Characterization of the Affymetrix Soybean Whole Genome Transcript Array .....	222
<b>K</b> - Webbook - A Web-Based Lab Notebook As An Undergraduate Teaching Tool .....	226
<b>L</b> - Design and Implementation of Web-Based Relational Databases .....	232

## LIST OF FIGURES

### Chapter One

Figure 1	Soybean seed development .....	12
Figure 2	Diversity of legume seed size and embryo morphology .....	13
Figure 3	Genomic strategy for identifying legume seed gene regulatory networks .....	15
Figure 4	Using SRB as a genomics engine to uncover genes active early in embryogenesis .....	16
Figure 5	Using LCM and transcriptional profiling to identify genes required to make a soybean seed .....	17
Figure 6	Identifying DNA sequences important for suspensor transcription.....	18

### Chapter Two

Figure 1	Schematic representation of <i>Arabidopsis</i> seed development and stages of the life cycle used for GeneChip analysis .....	26
Figure 2	Genes active before, during, and after <i>Arabidopsis</i> seed development .....	26
Figure 3	Quantitative regulation of mRNAs shared by all stages of seed development .....	28
Figure 4	Identification of <i>Arabidopsis</i> seed-specific mRNAs .....	28
Figure 5	Seed-specific TF gene activity in different <i>Arabidopsis</i> seed compartments, regions, and tissues .....	30
Figure S1	Correlation of GeneChip data between biological replicates and polysomal and total RNA populations .....	36
Figure S2	Functional classification of probe sets on the Affymetrix <i>Arabidopsis</i> ATH1 22K GeneChip .....	37
Figure S3	GeneChip detection limit and correlation between GeneChip data and qRT-PCR .....	37



Figure S4	Genes active in <i>Arabidopsis</i> floral and vegetative organs .....	38
Figure S5	Functional distribution and stage specificity of the seed-specific mRNA population .....	38
Figure S6	Network of chalazal endosperm transcription factor genes .....	39
Figure S7	Localization of seed-specific transcription factor mRNAs within specific seed compartments .....	40

### Chapter Three

Figure 3-1	Genes active in <i>lec1</i> seeds from fertilization to maturation and post-germination .....	56
Figure 3-2	Pairwise comparison of WT and <i>lec1</i> seed stages and seedling .....	57
Figure 3-3	LEC1 represses seedling genes during seed maturation .....	58
Figure 3-4	LEC1 regulates many seed-specific TFs during seed development .....	60

### Chapter Four

Figure 4-1	Soybean seed development .....	89
Figure 4-2	Genome-wide methylation changes during maturation, dormancy, and post-germination .....	91
Figure 4-3	Seed-specific genes are not affected by the DNA methylation changes .....	93
Figure 4-4	Local DNA methylation changes during seed development .....	95
Figure 4-5	DNA methylation is maintained in endoreduplicating cells .....	96
Figure 4-6	Genome-wide methylation changes among seed parts .....	97
Figure 4-S1	Proportions of cytosines and methylcytosines detected .....	99
Figure 4-S2	Average methylation distribution in genes and TEs .....	101
Figure 4-S3	Transcript abundance of methyltransferases and seed genes .....	102

### Appendix A

Figure 1	Gene activity in <i>Arabidopsis</i> seed regions and subregions throughout development .....	121
----------	--	-----

Figure 2	PCA of seed subregion mRNA populations .....	123
Figure 3	Dominant patterns of gene expression during seed development .....	124
Figure 4	Functions of subregion-specific genes .....	124
Figure 5	CZE is a unique seed subregion developmentally .....	125
Figure 6	Functions of CZE-delayed coexpressed gene sets .....	126
Figure 7	Predicted transcriptional modules regulating maturation in seeds .....	127
Figure S1	Microdissection of seed subregions during Arabidopsis seed development .....	133
Figure S2	Comparison of relative RNA levels and promoter activities of endosperm-expressed genes .....	134
Figure S3	Hierarchichal clustering of seed subregion mRNA populations .....	135
Figure S4	DPs of gene expression during seed development .....	136
Figure S5	Predicted transcriptional modules of coexpressed gene sets .....	137

## Appendix B

Figure 1	Oxidative burst triggered by 1 $\mu$ M flg22 (A and D), 50 $\mu$ g/mL chitin (B and E), or a mixture of 1 $\mu$ M flg22 + 50 $\mu$ g/mL chitin (C and F) in LD00- 2817P (LD), LDX01-1-65 (LX), Ripley (Rip), and EF59 (EF) leaf discs measured in relative luminescence units (RLUs) .....	141
Figure 2	Flower diagram showing numbers of overlapping and non- overlapping MAMP-responsive genes among LD00-2817P (LD), LDX01-1-65 (LX), Ripley, and EF59 .....	142
Figure 3	qRT-PCR validation of the MAMP-responsive genes in four contrasting soybean genotypes treated over 30 min with MAMPs (flg22 + chitin) ...	142
Figure 4	MapMan overview of regulation showing the regulation (blackarrows), protein modification (greenarrows), regulation of transcription (red arrows), hormones (green dotted arrows), enzyme families (black dotted arrows), and transport pathway (red dotted arrows) genes that	

	are preferentially expressed in MAMP-treated leaves from LD00-2817P (A), LDX01-1-65 (B), Ripley (C), and EF59 (D) genotypes .....	143
Figure 5	Bacterial population (A and C) and <i>S. sclerotiorum</i> actin mRNA transcripts levels (B and D) in LD00-2817P (A and B) and LDX01-1-65 (C and D) leaf tissues recovered 0 and 3, or 2 d after inoculation (DPI) with <i>Psg</i> (5.3 × 10 <sup>5</sup> CFU/mL, 10 μM MgCl <sub>2</sub> ) or <i>S. sclerotiorum</i> , respectively .....	144
Figure 6	Localization of QTLs providing variation of MTI responses .....	145
Figure 7	Variation of the pathogen resistance levels in MAMP-treated contrasting F3 lines, defined based on higher (A and B) or lower (C and D) expression of MAMP-responsive genes .....	145
<b>Appendix C</b>		
Figure 1	Scatter plots of wild type and CaMV:DME pollen and stamen .....	154
Figure 2	Analysis of RNA profiles .....	154
Figure 3	Functional categorization of wild type Arabidopsis whole genome (A) and genes induced by DME (B) .....	155
Figure 4	Semi-quantitative RT-PCR validation of putative DME target genes ....	155
<b>Appendix D</b>		
Figure 1	Profiles of RNAs induced by LEC2 that are prevalent during the maturation phase .....	159
Figure 2	qPCR data validate DNA microarray results and reveal an early role for several LEC2-induced RNAs .....	159
Figure 3	LEC2 target genes contain upstream RY motifs .....	160
Figure 4	LEC2 binds specifically to RY motifs upstream of LEC2 target genes ..	160
<b>Appendix E</b>		
Figure 1	Schematic representation of tobacco anther development based upon histological studies at the light microscope level (Satina and Blakeslee	

	1941; Joshi et al. 1967; Koltunow et al. 1990) .....	166
Figure 2	Establishment of anther shape from the stamen primordia .....	170
Figure 3	Development of the anther notch at stages -5 to -1 .....	171
Figure 4	Cells of the circular cell cluster during tobacco anther development ....	172
Figure 5	Number of cells in the circular cell cluster and stomium during anther development .....	173
Figure 6	Circular cell cluster degeneration and stomium differentiation .....	174
Figure 7	Stomium cells during tobacco anther development .....	176
Figure 8	Plasmodesmata in the anther notch region during anther development .....	177
Figure 9	Late events in stomium development .....	178
Figure 10	Stomium cell death and anther dehiscence .....	180
Figure 11	Using laser capture microdissection (LCM) to isolate stomium cells and detect specific mRNAs .....	181
Figure 12	Summary of events that occur during tobacco anther notch-region development .....	182
Figure 13	Cell lineages participating in tobacco anther dehiscence .....	184
Figure 14	Models for the specification and differentiation of the circular cell cluster and stomium .....	185
<b>Appendix F</b>		
Figure F-1	Genes active in compartments, regions, and tissues throughout soybean seed development .....	190
Figure F-2	Seed Gene Network website ( <a href="http://seedgenenetwork.net/">http://seedgenenetwork.net/</a> ) .....	192
Figure F-3	Seed Gene Network website -- Search Form .....	193
Figure F-4	Seed Gene Network website -- Search Results output .....	194
Figure F-5	Seed Gene Network website -- Heat map visualization of the GeneChip datasets .....	195

Figure F-6	Seed Gene Network website -- Bar plot visualization of the GeneChip datasets .....	196
------------	--	-----

## Appendix G

Figure G-1	Key questions in early embryo development .....	201
Figure G-2	Scarlet runner bean as a model system to study early embryo development .....	201
Figure G-3	Strategy for discovering genes active during early plant embryo development .....	202
Figure G-4	Phaseolus coccineus embryo EST website ( <a href="http://estdb.biology.ucla.edu/PcEST">http://estdb.biology.ucla.edu/PcEST</a> ) .....	203
Figure G-5	mRNA localization of selected molecular markers by in-situ hybridization analysis .....	204

## Appendix H

Figure H-1	ESTDB Sequence Analysis website ( <a href="http://estdb.biology.ucla.edu/~goldberg">http://estdb.biology.ucla.edu/~goldberg</a> ) .....	207
Figure H-2	ESTDB Sequence Analysis website - Sequence Processing .....	208
Figure H-3	ESTDB Sequence Analysis website - Summary and Search Form .....	209
Figure H-4	ESTDB Sequence Analysis website -- Project Summary .....	210
Figure H-5	ESTDB Sequence Analysis website -- Sequence Record Page .....	211
Figure H-6	ESTDB Sequence Analysis website -- Contig Assembly Summary .....	212
Figure H-7	ESTDB Sequence Analysis website -- Multiple Sequence Alignment View .....	213

## Appendix I

Figure I-1	Distribution of 8,247 features on the Arabidopsis AtGenome1 (8K) GeneChip array into major functional categories .....	216
Figure I-2	Distribution of transcription factor families on the Arabidopsis AtGenome1 (8K) GeneChip array .....	217

Figure I-3	Distribution of 22,747 features on the <i>Arabidopsis</i> ATH1-121501 (22K) GeneChip array into major functional categories .....	218
Figure I-4	Distribution of transcription factor families represented on the <i>Arabidopsis</i> ATH1-121501 (22K) GeneChip array .....	219
Figure I-5	Distribution of 37,593 features on the Soybean Genome GeneChip array into functional categories .....	220
Figure I-6	Assignment of 2,832 features on the Soybean Genome GeneChip array into transcription factor families .....	221
<b>Appendix J</b>		
Figure J-1	Illustration of the association between probe id and the predicted gene model .....	225
<b>Appendix K</b>		
Figure K-1	Webbook website ( <a href="http://estdb.biology.ucla.edu/webbook">http://estdb.biology.ucla.edu/webbook</a> ) .....	227
Figure K-2	Webbook website -- Project Page .....	228
Figure K-3	Webbook website -- Experiment Page .....	229
Figure K-4	Webbook website -- Protocols Page .....	230
Figure K-5	Webbook website -- Lab Stocks .....	231
<b>Appendix L</b>		
Figure L-1	The Goldberg Lab website ( <a href="http://www.mcdb.ucla.edu/Research/Goldberg">http://www.mcdb.ucla.edu/Research/Goldberg</a> ) .....	234
Figure L-2	Seed Gene Network website ( <a href="http://seedgenenetwork.net">http://seedgenenetwork.net</a> ) .....	235
Figure L-3	Webbook website ( <a href="http://estdb.biology.ucla.edu/webbook">http://estdb.biology.ucla.edu/webbook</a> ) .....	236
Figure L-4	ESTDB Sequence Analysis website ( <a href="http://estdb.biology.ucla.edu/~goldberg">http://estdb.biology.ucla.edu/~goldberg</a> ) .....	237
Figure L-5	<i>Arabidopsis thaliana</i> GeneChip Project website ( <a href="http://estdb.biology.ucla.edu/genechip/">http://estdb.biology.ucla.edu/genechip/</a> ) .....	239
Figure L-6	<i>Arabidopsis thaliana</i> GeneChip Project website -- Search Form .....	240

Figure L-7	<i>Phaseolus coccineus</i> ESTs (PcEST) website ( <a href="http://estdb.biology.ucla.edu/PcEST/">http://estdb.biology.ucla.edu/PcEST/</a> ) .....	242
Figure L-8	<i>Phaseolus coccineus</i> ESTs (PcEST) website -- Search Form .....	243

## LIST OF TABLES

### Chapter Four

Table 4-1. Abbreviation of seed stages, organs, and tissues in this study .....	98
Table 4-S1. Sequencing summary -- related to Figure 4-1 .....	103
Table 4-S2. Bulk methylation statistics for the nuclear and chloroplast genome -- related to Figures 2, 5, 6 .....	104
Table 4-S3. Pairwise comparison summary statistics -- related to Figures 4-2, 4-5, and 4-6 .....	105
Table 4-S4. List of differentially methylated regions (DMRs) -- related to Figure 4-4 ....	113

### Appendix A

Table 1. Abbreviation for developmental stages and seed subregions .....	122
--	-----

### Appendix D

Table 1. LEC2 activates a gene with an upstream RY motif in planta .....	161
--	-----

### Appendix E

Table 1. Major events within the notch region during tobacco anther development .....	167
Table 2. Comparison of anther notch region development and the dehiscence program in tobacco and <i>Arabidopsis thaliana</i> .....	183



## SUPPLEMENTARY MATERIALS

### **Chapter Two (Datasets published in Le et al. 2010 PNAS)**

Dataset S1 - Dataset S1 contains Tables S1-S10.

Dataset S2 - Dataset S2 contains Tables S11-S20.

### **Chapter Three (Unpublished datasets)**

Dataset 3-1 - Raw data and correlation analysis of the *lec1* seed development GeneChip data.

Dataset 3-2 - This dataset contains Tables showing comparisons between wild type and *lec1* seed stages.

## ACKNOWLEDGEMENTS

I would like to thank my mentor, Professor Bob Goldberg, for teaching me how to think critically about science, and giving me numerous opportunities to explore and drive my own curiosity. Not only have you taught me how to do science, both at the bench and the computer, but you have instilled in me a passion for teaching that I never knew I had before. I will miss our 2 AM discussions about experiments and data.

I would like to thank members of the Seed Institute, both past and present, for providing me intellectual stimulation year after year at Lake Arrowhead and UC Davis. I would especially like to thank John Harada and Bob Fischer for the constant motivation, intellectual drive and support during my graduate years.

I would like to thank past and present Goldberg lab members for the close collaborations and learning experiences. I want to thank Anhthu Bui, with whom I interacted on a daily basis over the majority of my Ph.D. studies. You have taught me many experimental techniques and we have tackled many new and exciting technologies together. Thank you for being a mentor at the bench and a friend. Thank you to Tomokazu and Mari Kawashima for being my long time roommates and friends, and for teaching me about the Japanese culture. I am indebted to Min Chen for her continual support these past few years, for helping with multiple computational tasks, and for introducing me to programming. I want to thank Kelli Henry, Jungim Hur, Jer-Young Lin, Ann Amores, Eden Maloney, Elaine Chiu and Jennifer Kwan for the many memories in and outside the lab. I also want to thank Joseph Means for his daily hard work maintaining the lab and for giving me a break from lab work with discussions about sports and other daily life activities.

I would like to acknowledge financial support from the National Institutes of Health Training Grant in Genomic Analysis and Interpretation T32HG002536 and grants awarded to Professor Bob Goldberg from Ceres, Inc., the Department of Energy, and the National Science Foundation Plant Genome Program.

**Chapter One:** I am grateful to my co-authors (Javier Wagmaister and Tomokazu Kawashima) for their intellectual contribution to this review. I want to thank Koen Weterings, Nestor Apuya, and Anhthu Bui for starting the SRB system that became the basis of all the genomics work in our lab. I am grateful to John Harada and Bob Goldberg for critical comments and scientific contributions.

**Chapter Two:** This project was truly collaborative and I am grateful to members of the Harada lab (Linda Kwong, Julie Pelletier) for their relentless and diligent efforts to harvest the numerous seed samples required for this project. I am especially thankful to Steve Horvath and Zeke Fang for showing us how to analyze microarray data. Javier Wagmaister and Kelli Henry contributed seed-specific-TF-promoter::GFP plants. Mark Belmonte, Ryan Kirkbride, and Julie Pelletier contributed LCM data from Arabidopsis seed compartments. I would like to thank my co-authors, Chen Cheng and Anhthu Bui, for helping with the GeneChip work and data analysis. Lastly I would like to thank Gary Drews, Bob Fischer, Jack Okamuro, John Harada, and Bob Goldberg for critical comments and suggestions on the project.

**Chapter Three:** This is another large collaborative project involving our lab and John Harada's lab. I would like to thank Julie Pelletier and other members of the Harada lab for characterizing and harvesting precious *lec1* seed materials. I am grateful to John Harada for letting me take the lead on the analysis of these GeneChip data. I also want to thank John Harada and Bob Goldberg for critical comments and guidance with this project.

**Chapter Four:** Min Chen and Jer-Young Lin, Anhthu Bui, and Tzung-Fu Hsieh made significant contributions to this project. Anhthu Bui and Tzung-Fu Hsieh initiated the project and provided the first BS-Seq libraries. Jer-Young Lin and Min Chen generated all the BS-DNA-Seq libraries included in the study. In addition, Min Chen contributed her computational expertise, wrote scripts and set up the analysis pipeline. Pao-Yang Chen provided the BS-Seeker program for aligning BS-DNA-Seq reads and provided suggestions and comments on data analysis. Julie Pelletier and Ryan Kirkbride generated small RNA libraries and contributed to the analysis of

small RNA data. Jungim Hur assisted with the capture of ABPY and ADPY tissues in the endoreduplication story (Figure 4-5). Kelli Henry provided RNA-Seq datasets used in the study. Matteo Pellegrini, Bob Fischer, John Harada, and Bob Goldberg contributed critical comments, suggestions, and insights to the project.

Finally, I would like to thank my brothers -- Steven, Anthony, and John -- for their continual support. I am extremely thankful to my parents -- David and Nancy Le -- for giving me the opportunity to reach my true potential when we emigrated from Vietnam. Thank you David and Shawna Charney for your constant encouragement and support. Thank you Oscar and Emmy for providing me daily amusement and joy (and for refraining from accidentally deleting my thesis when sitting on the computer). And lastly, a special thank you to my wife, Bekah, for her love and support on this long but wonderful journey.

## VITA

### EDUCATION

**2000**            **University of California, Los Angeles**  
B.S., Mathematics & Applied Sciences

### AWARDS

**2008 - 2010**        NIH Genome Analysis Fellowship

**2004**            UCLA Staff Award for Outstanding Service

### RESEARCH EXPERIENCE

**2000 - 2005**        **Staff Research Associate I to III, University of California, Los Angeles**  
Dr. Bob Goldberg

### TEACHING EXPERIENCE

**2005 - 2011**        **Bioinformatics Instructor, University of California, Los Angeles**  
*Honors Collegium 70AL: "Gene Discovery Laboratory"*

**2008**            **Teaching Assistant, University of California, Los Angeles**  
*Honors Collegium 70A: "Genetic Engineering in Medicine, Agriculture, and Law"*

**2007**            **Teaching Assistant, University of California, Los Angeles**  
*MCDB 150: "Plant Chemical and Molecular Communication"*

**2003 - 2004**        **Teaching Assistant, University of California, Los Angeles**  
*Honors Collegium 70A: "Genetic Engineering in Medicine, Agriculture, and Law"*

**2003 - 2004**        **Teaching Assistant, University of California, Los Angeles**  
*Honors Collegium 70AL: "Gene Discovery Laboratory"*

### PUBLICATIONS

1. Belmonte, M.F.<sup>#</sup>, Kirkbride, R.C. <sup>#</sup>, Stone, S.L., Pelletier, J.M., Bui, A.Q., Yeung, E.C., Hashimoto, M., Fei, J., Harada, C.M., Munoz, M.D., **Le, B.H.**, Drews, G.N., Brady, S.M., Goldberg, R.B., Harada, J.J., Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed, *PNAS* **110**, E435–44 (2013). <sup>#</sup> Authors contributed equally to this work.

2. Valdés-López, O., Thibivilliers, S., Qiu, J., Xu, W., Nguyen, T., Libault, M., **Le, B.H.**, Goldberg, R.B., Hill, C., Hartman, G., Diers, B., Stacey, G., Identification of quantitative trait loci controlling gene expression during the innate immunity response of soybean, *Plant Physiol* **157**, 1975–1986 (2011).

3. **Le, B.H.**<sup>#</sup>, Cheng, C.<sup>#</sup>, Bui, A.Q.<sup>#</sup>, Wagmaister, J.A., Henry, K.F., Pelletier, J., Kwong, L., Belmont, M., Kirkbride, R., Horvath, S., Drews, G.N., Fischer, R.L., Okamuro, J.K., Harada, J.J., and Goldberg, R.B., Global analysis of gene activity during Arabidopsis seed development and

identification of seed-specific transcription factors, *PNAS* **107**, 8063–8070 (2010). # Authors contributed equally to this work.

4. Ohr, H., Bui, A.Q., **Le, B.H.**, Fischer, R.L., Choi, Y., Identification of putative Arabidopsis DEMETER target genes by GeneChip analysis, *Biochem Biophys Res Commun* **364**, 856–860 (2007).

5. **Le, B.H.**#, Wagmaister, J.A.#, Kawashima, T.#, Bui, A.Q., Harada, J.J., and Goldberg, R.B., Using genomics to study legume seed development, *Plant Physiol* **144**, 562–574 (2007).

6. Braybrook, S.A., Stone, S.L., Park, S., Bui, A.Q., **Le, B.H.**, Fischer, R.L., Goldberg, R.B. and Harada, J.J., Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis, *PNAS* **103**, 3468–3473 (2006).

7. Sanders, P.M., Bui, A.Q., **Le, B.H.**, Goldberg R.B., Differentiation and degeneration of cells that play a major role in tobacco anther dehiscence, *Sex Plant Reprod* **17**, 219–241 (2005).

## INTRODUCTION TO THE DISSERTATION

**How Are Genes Regulated During Seed Development?**

**What Transcription Factors Are Playing a Major Role in Controlling Seed Development?**

**Are There Genome-Wide Methylation Changes During Seed Development,  
and if so, How Do These Changes Affect Gene Activity?**

These were the driving questions that formed the basis of my Ph.D. research and graduate career. To begin to address these questions, I used a variety of genomics tools [e.g., Sanger sequencing, next generation sequencing (NGS), GeneChip microarrays, laser capture microdissection (LCM)], with several plant species [e.g., *Arabidopsis*, soybean, scarlet runner bean (SRB)], to characterize seed transcriptomes and methylomes on a genome-wide basis. In addition, I used reverse genetics to knockout seed-specific transcription factors and determine what role they may play in seed development.

### ***Thesis Chapters and Appendices***

My Ph.D. thesis is divided into four chapters and several appendices. The chapters describe work that was the major focus of my Ph.D. research. By contrast, the appendices contain summaries of other published and unpublished research projects that I have been involved with in a collaborative basis during my graduate career. Below is a brief summary of the content of each chapter and appendix.

### **Chapter One – Using Genomics to Dissect Legume Seed Development**

This chapter contains a review published in *Plant Physiology*. This review highlights the genomics tools available to study seed development in legumes, and provides a summary of the transcriptome analysis that I carried out on the embryo proper and suspensor of early stage Scarlet Runner Bean embryos.

**Chapter Two – Global Analysis of Gene Activity in Arabidopsis Seed Development and Identifying Seed-Specific Transcription Factors**

This chapter, published in the *Proceedings of the National Academy of Sciences (PNAS)*, presents the results of a collaborative NSF-funded project between our lab at UCLA and John Harada's lab at UC Davis to identify all of the genes required to "make a seed." This work documents the global changes in gene activity that occur throughout Arabidopsis seed development as well as the entire plant life cycle, ultimately highlighting 48 seed-specific transcription factors that may play critical roles in regulating seed development. My role in this project was to use computational approaches to analyze and describe all of the transcriptome datasets, as well as identify seed-specific transcription factors. Furthermore, in conjunction with Anhthu Bui, I annotated the *Arabidopsis* AtGenome1 (8K) and the ATH1 (22K) Affymetrix GeneChip arrays, sorting each probe set into functional categories. I also used reverse genetics to knockout seed-specific TFs in order to study their role during seed development.

**Chapter Three – Global Analysis of *lec1* Mutant During Arabidopsis Seed Development**

This chapter summarizes the unpublished results of a project between our lab and John Harada's lab at UC Davis, that focuses on how a mutation in LEC1, a global regulator of seed development, affects gene activity during *Arabidopsis* seed development and life cycle. For this project, I used computational approaches to analyze and describe all of the *lec1* transcriptome datasets, and described how the *lec1* mutation affects global gene activity and the activity of seed-specific transcription factors genes.

**Chapter Four – Dynamic Changes in DNA Methylation During Soybean Seed Development**

This chapter represents the major focus of my Ph.D. research efforts during the past three years. It was carried out in collaboration with Min Chen and Jer-Young Lin of our lab, and funded by an NSF Plant Genome Project to Bob Goldberg, John Harada, and Matteo Pellegrini to identify the regulatory processes controlling seed development. A draft manuscript is presented that contains the results of whole-genome methylome analyses during seed



development – from fertilization through dormancy – and the post-germination phase of soybean. In addition, it contains data that describe the methylomes of different seed tissues before and after endoreduplication, as well as major seed parts such as seed coat, cotyledons, and axis. My role was to provide intellectual leadership, designed experiments and used computational approaches to analyze and describe all methylome datasets.

***Appendix A through E (Published co-authored work from collaborative projects in which I played a role)***

**Appendix A** – Belmonte, M.F., Kirkbride, R.C., Stone, S.L., Pelletier, J.M., Bui, A.Q., Yeung, E.C., Hashimoto, M., Fei, J., Harada, C.M., Munoz, M.D., **Le, B.H.**, Drews, G.N., Brady, S.M., Goldberg, R.B., Harada, J.J. (2013). **Comprehensive developmental profiles of gene activity in regions and subregions of the *Arabidopsis* seed.** PNAS 110, E435–E444

This paper characterizes gene activity in every *Arabidopsis* seed compartment, tissue, and region throughout all of seed development using laser capture microdissection and GeneChip microarrays. This project is part of the NSF Plant Genome Program grant awarded to Bob Goldberg, John Harada, and Steve Horvath to identify all the genes active in every tissue, region, and compartment throughout seed development. I provided characterization of the GeneChip used for the study and consulted on the data analysis approaches.

**Appendix B** – Valdés-López, O., Thibivilliers, S., Qiu, J., Xu, W.W., Nguyen, T.H.N., Libault, M., **Le, B.H.**, Goldberg, R.B., Hill, C.B., Hartman, G.L., et al. (2011). **Identification of quantitative trait loci controlling gene expression during the innate immunity response of soybean.** Plant Physiol 157, 1975–1986.

This project used a Soybean Affymetrix Whole Genome Transcript array to identify genes regulated during the plant innate immunity response. I collaborated with Affymetrix to design the whole genome soybean GeneChip array for use by the entire soybean community.

**Appendix C** –\_Ohr, H., Bui, A.Q., **Le, B.H.**, Fischer, R.L., and Choi, Y. (2007). **Identification of putative Arabidopsis DEMETER target genes by GeneChip analysis.** Biochem Biophys Res Commun 364, 856–860.

This project identified *DEMETER* (*DME*) target genes by analyzing the results of *Arabidopsis* GeneChip experiments. *DME* is a DNA glycosylase that is primarily expressed in the central cell and is required for removing methylated cytosines from the maternal allele, establishing imprinted genes in the endosperm following double fertilization. Putative targets were identified by ectopically expressing *DME* in pollen and stamen, where *DME* is normally not active, and searching for up-regulated genes. I helped carry out the GeneChip analysis.

**Appendix D** – Braybrook, S.A., Stone, S.L., Park, S., Bui, A.Q., **Le, B.H.**, Fischer, R.L., Goldberg, R.B. and Harada, J.J. (2006). **Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis.** PNAS 103, 3468-3473.

This project identified *LEC2* target genes by expressing an inducible form of *LEC2* in seedlings. Transcripts present after *LEC2* induction were profiled using *Arabidopsis* GeneChip. I generated and analyzed the wild type GeneChip developmental data series used in the paper.

**Appendix E** – Sanders, P.M., Bui, A.Q., **Le, B.H.**, Goldberg, R.B. (2005). **Differentiation and degeneration of cells that play a major role in tobacco anther dehiscence.** Sex Plant Reprod, 17, 219-241

This published work characterized the differentiation and degeneration of stomium cells during anther dehiscence using light and transmission electron microscopy. I carried out LCM of stomium regions from tobacco anthers and performed quantitative PCR.

***Appendix F through K (Unpublished Research Projects and Web Databases)***

**Appendix F – An Atlas of Gene Activity in Soybean Seed Regions, Compartments, and Tissues During Development**

This project is a collaborative effort between our lab and John Harada's at UC Davis as part of the NSF Plant Genome Program grant to identify all the genes required to "make a soybean seed." This project involved profiling gene activity in every tissue, region, and compartment of a soybean seed from fertilization to maturation using LCM coupled with soybean GeneChip arrays. For this project, with the help of Anhthu Bui and Javier Wagmaister, I annotated > 37,000 features on the soybean GeneChip array, classifying each probe set feature into functional categories and established computational approaches to analyze these transcriptome datasets. I also generated RNA-Seq data from whole seeds and embryo regions to compare against the GeneChip datasets. The results from this project are currently being prepared for a manuscript where I will be a co-author with the senior author on the project, Jungim Hur.

#### **Appendix G – Gene Activity In Different Regions of a Globular Stage SRB Embryo By EST Sequencing**

This project was part of a DOE grant awarded to Bob Goldberg to study gene activity in the embryo proper and suspensor regions of a globular stage embryo. Embryo proper and suspensor regions were hand dissected from globular stage SRB embryos. Sanger EST sequencing and 454 NGS was used to profile gene activity in these embryonic regions. My role for this project involved isolating DNA from plasmid clones, analyzing EST sequence data using BLAST, establishing a web-based relational database for the processing and storage of sequence data, and carrying out *in-situ* hybridization to study mRNA localization patterns within the embryonic regions. Part of this work is published in the *Plant Physiology* review (Chapter one). Additionally, I was also involved in sequencing and annotating the SRB genome using 454 and Illumina NGS technologies.

#### **Appendix H – ESTDB - A Web-Based Relational Database For Analyzing And Storing DNA Sequences**

This project involved the development of a web-based relational database for the high-throughput analysis of Sanger sequences from the SRB EST sequencing project (Appendix G) with Anhthu Bui, Harry Hahn, and Bob Goldberg. Several functions of this web-based sequence

analysis toolkit includes batch download of sequence data directly from the sequencing facility, batch BLAST of EST sequences against publicly available databases (e.g. NCBI non-redundant database), storing BLAST results and sequence annotations, and assembly of ESTs into clusters using the CAP3 contig assembly program. I worked closely with Harry Hahn, our system administrator, to develop and add the necessary tools needed for sequence analysis.

### **Appendix I – Functional Annotation of Genes Represented on the Arabidopsis and Soybean GeneChip Arrays**

This project involved the annotation of every gene feature on the *Arabidopsis* and Soybean GeneChip arrays and the categorization of these features into functional categories including the identification and classification of TF genes based on TF gene families. These annotations and functional categories guided the interpretation and analysis of the GeneChip datasets. The *Arabidopsis* and soybean GeneChip arrays were annotated with help from Anhthu Bui and Javier Wagmaister. My role included carrying out BLAST analysis of each array sequence against the public sequence databases, assigning annotations and functional categories to each unique sequence feature, and categorizing TFs into TF families. These annotations are available through our lab web site for *Arabidopsis* (<http://seedgenenetwork.net/presentation#arabidopsis>) and soybean (<http://seedgenenetwork.net/presentation#soybean/VT>).

### **Appendix J – Design And Characterization of the Affymetrix Soybean Whole Genome Transcript Array**

This project was a collaboration between our lab and Affymetrix with consultation from members of the soybean community, both academic and commercial, and the Joint Genome Institute (JGI) to design a next generation whole genome transcript (WT) array for soybean. The first generation soybean Genome array was designed using available full-length cDNAs and ESTs with 37,000 probe sets interrogating ~25,000 distinct genes/transcripts. Using the published whole soybean genome sequence, we designed the soybean WT array to interrogate ~ 66,000 genes including high and low confidence gene models. I provided computational analysis and

characterization of the probe sequences on the new array. Files containing the characterization of the WT array and the association of probe sets with predicted gene models are available on our lab SeedGeneNetwork website (<http://seedgenenetwork.net/presentation#soybeanWT>).

### **Appendix K – Webbook - A Web-Based Lab Notebook As An Undergraduate Teaching Tool**

This project involved the development of a web-based database that serves as an electronic lab notebook for teaching undergraduates how to do research. The tool was developed as part of Bob Goldberg's undergraduate lab course, Honors Collegium 70AL - Gene Discovery Laboratory. This web database stores students data entry and houses protocols for carrying out research at the bench. Uploaded data (e.g., gel images) are stored for each experiment and are easily accessible. The webbook was programmed by Harry Hahn with original input design provided by myself, Anhthu Bui, and Bob Goldberg. I provided leadership in the design, maintenance, and implementation of the site.

### **Appendix L – Design and Implementation of Web-Based Relational Databases**

As part of an NSF-funded project, to make data (e.g. transcriptomes, methylomes) available to the general research community, I worked closely with Harry Hahn, Weihong Yan, Min Chen, Anhthu Bui, and Bob Goldberg to create, design, and implement the following web-based databases in addition to our lab website.

*Lab* - <http://www.mcdb.ucla.edu/Research/Goldberg>

I designed and created the lab web site to highlight the research being carried out in the lab as well as Bob Goldberg's unique teaching techniques and courses.

*Seed Gene Network* - <http://seedgenenetwork.net>

I designed and assisted with the creation of this website as a portal for the dissemination of data generated through the NSF-funded project to profile gene activity in every compartment, tissue, and cell types across different stages of soybean seed development. This website is constantly updated to include integration of next-generation sequencing data (mRNA-Seq, smRNA-Seq, BS-DNA-Seq) to examine gene regulation during seed development from a systems biology

view. The website was created by Harry Hahn and is maintained by Weihong Yan, Min Chen, and myself. This website hosts data generated from projects in Chapter Four and appendices A, F, I, and J.

*Lab Webbook* - <http://estdb.biology.ucla.edu/webbook>

I designed and assisted with the creation of this website to serve as a central repository of lab protocols, lab stock information (e.g. DNA, RNA, seeds) as well as lab experimental notes and data. This website was also used to teach undergraduate students how to do science as part of Bob Goldberg's HC70AL - Gene Discovery Laboratory course (see Appendix K).

*ESTDB* - <http://estdb.biology.ucla.edu/~goldberg>

I designed and implemented this web-based DNA sequence analysis tool for high-throughput semi-automated annotation and analysis of EST sequences incorporating tools for DNA sequence analysis (BLAST - sequence alignment, CAP3, contig assembly program, PFAM - protein family database, PHRED - base calling program for original sequence trace files). The website was programmed by Harry Hahn, with input from myself, Anhthu Bui, and Bob Goldberg. This website was developed as part of the SRB EST sequencing project (Appendix G).

*Arabidopsis Seed Development GeneChip* - <http://estdb.biology.ucla.edu/genechip>

This website host all GeneChip data generated from profiling *Arabidopsis* seed development from before fertilization through maturation and other times in the plant life cycle (see Chapter two). The website has several features including the ability to browse the datasets and to carry out comparative analysis between different datasets. The website was programmed by Harry Hahn with design and directions from myself, Anhthu Bui, and Bob Goldberg.

*Phaseolus coccineus ESTs (PcESTs)* - <http://estdb.biology.ucla.edu/PcEST>

This website was developed to provide to the research community a collection of ESTs identified from different regions (embryo proper and suspensor) of a SRB globular stage embryo (see Appendix G). All the information on this website was summarized and annotated by myself and Anhthu Bui. On this website, the user can browse the datasets or use BLAST to identify ESTs

with homology to a queried sequence of interest. The website was programmed by Harry Hahn with design, data, and summaries provided by myself, Anhthu Bui, and Bob Goldberg.

## CHAPTER ONE

### USING GENOMICS TO STUDY LEGUME SEED DEVELOPMENT



# Using Genomics to Study Legume Seed Development<sup>1</sup>

Brandon H. Le<sup>2</sup>, Javier A. Wagmaister<sup>2</sup>, Tomokazu Kawashima<sup>2</sup>, Anhthu Q. Bui, John J. Harada, and Robert B. Goldberg\*

Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, California 90095 (B.H.L., J.A.W., T.K., A.Q.B., R.B.G.); and Section of Plant Biology, Division of Biological Sciences, University of California, Davis, California 95616 (J.J.H.)

Seeds are essential for flowering plant reproduction because they protect, nourish, and contain the developing embryo that represents the next sporophytic generation. In addition, seeds contain energy resources that sustain the young sporophyte during germination before photosynthesis begins. In legumes, food reserves stored in embryonic cotyledons make seeds important as a food source for both human and animal consumption. For example, soybean (*Glycine max*) is now one of the most important seed crops in the world (Wilcox, 2004). Research on legume seed development has led to direct applications, such as seeds with more nutrients (Kinney, 1998; Wang et al., 2003; Krishnan, 2005), reduced allergens (Herman et al., 2003), and novel constituents, such as edible vaccines (Moravec et al., 2007). In the current genomic era, it is now possible to begin to understand what genes are required to make a legume seed and how regulatory networks are interconnected in legume genomes to program seed formation. In the future, this information should permit novel approaches to breed and engineer legume seeds with new agronomic traits and, most importantly, help provide a sustainable food supply for a growing human population. This *Update* outlines how our laboratories have been using legumes and functional genomics to identify genes that program legume seed development.

Seed development is triggered by a novel double-fertilization process that leads to the differentiation of the embryo, endosperm, and seed coat, which are the major compartments of the seed (Fig. 1, A–C; Goldberg et al., 1994; Miller et al., 1999; Gehring et al., 2004; Laux et al., 2004; Moise et al., 2005). These compartments have different origins and play distinct roles in seed formation. The maternally derived seed coat differentiates from the ovule integuments that surround the

embryo sac and plays a major role in protecting the embryo and transferring nutrients from the maternal plant to the developing embryo (Fig. 1, A and C; Murray, 1987; Borisjuk et al., 2004; Moise et al., 2005). By contrast, the embryo and endosperm are direct descendents of the fertilized egg and central cell, respectively. The endosperm proliferates to occupy most of the postfertilization embryo sac and nourishes the embryo early in development (Gehring et al., 2004). In many flowering plants, such as legumes, the endosperm is absorbed by the embryo during development and is not present in the mature seed (Fig. 1, A–C; Goldberg et al., 1994). After fertilization, the zygote divides asymmetrically, giving rise to a small apical cell that develops into the embryo proper and a large basal cell that forms the suspensor. The suspensor is a terminally differentiated structure that supports and nourishes the embryo proper and degenerates later in development (Yeung and Meinke, 1993). The embryo proper, on the other hand, represents the new sporophytic generation and contains the shoot and root meristems that are responsible for generating organ systems of the mature plant after seed germination (Fig. 1C; Goldberg et al., 1994; Laux et al., 2004).

### MAJOR QUESTIONS REMAIN UNANSWERED IN SEED DEVELOPMENT

Many developmental and physiological events occur within each seed compartment during development (Fig. 1B) and are programmed, in part, by the activity of different genes (Goldberg et al., 1989, 1994; Stangeland et al., 2003; Gehring et al., 2004; Haughn and Chaudhury, 2005). Seed development, therefore, is the result of a mosaic of distinct gene expression programs occurring in parallel in different seed compartments (e.g. embryo, endosperm, seed coat) as well as within specific regions and tissues (e.g. embryo proper, suspensor, epidermis). What these programs are and how they are integrated into unique regulatory networks within the plant genome remain major unanswered questions (Fig. 1D). Specifically, it is not yet known what genes in different seed compartments play important roles in cell fate specification, differentiation, and morphogenesis during early seed and embryo development. Molecular identification and

---

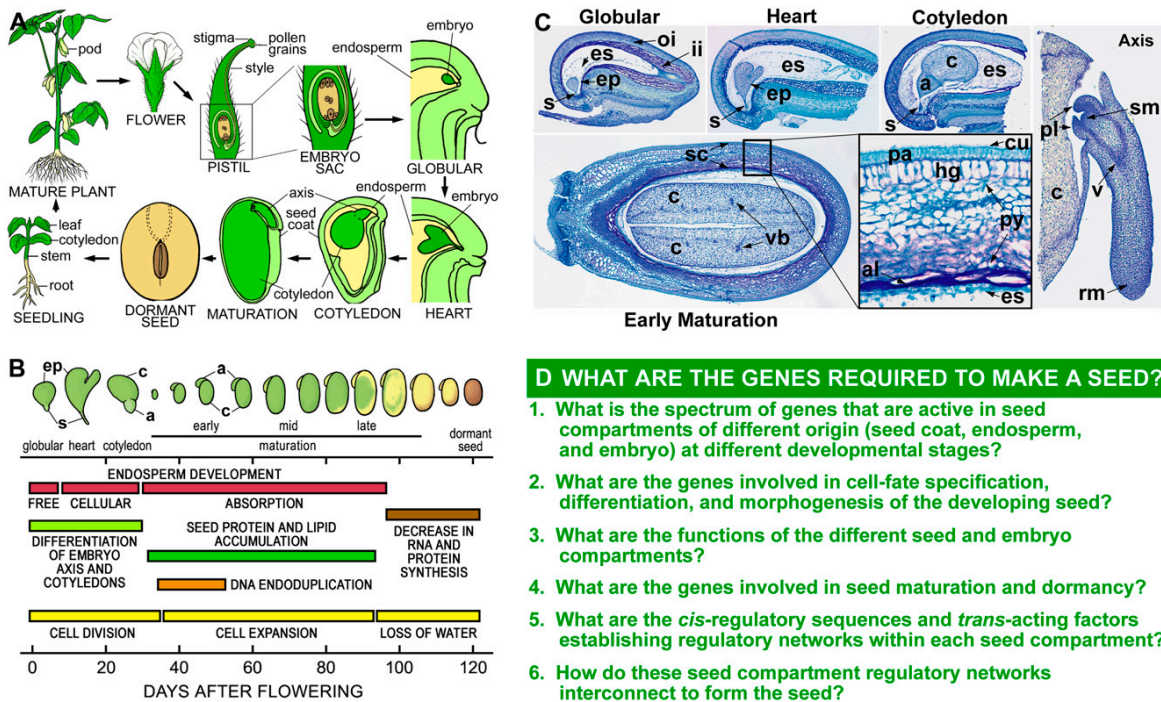
<sup>1</sup> This work was supported by the National Science Foundation Plant Genome Program (grant no. DBI-0501720), the Department of Energy (grant no. DE-FG03-97ER20263), and Ceres Inc. T.K. is a recipient of a Nakajima Foundation predoctoral fellowship.

<sup>2</sup> These authors contributed equally to the article.

\*Corresponding author; e-mail bobg@ucla.edu; fax 310-825-8201.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors ([www.plantphysiol.org](http://www.plantphysiol.org)) is: Robert B. Goldberg (bobg@ucla.edu).

[www.plantphysiol.org/cgi/doi/10.1104/pp.107.100362](http://www.plantphysiol.org/cgi/doi/10.1104/pp.107.100362)



**Figure 1.** Soybean seed development. A, Cartoon depicting soybean life cycle. B, Schematic representation of soybean seed development. Embryo morphologies and developmental events were adapted and modified from Goldberg et al. (1989). C, Paraffin transverse 10- $\mu$ m sections of soybean globular, heart, cotyledon, and early maturation seeds. Inset contains a magnified view (40 $\times$ ) of the seed coat. Axis longitudinal section was obtained from an early maturation seed. D, Major unanswered questions in seed development. a, Axis; al, aleurone; c, cotyledon; cu, cuticle; ep, embryo proper; es, endosperm; hg, hourglass cells; ii, inner integument; oi, outer integument; pa, palisade layer; pl, plumule; py, parenchyma; rm, root meristem; s, suspensor; sc, seed coat; sm, shoot meristem; v, vascular tissues; vb, vascular bundle.

characterization of these genes will help identify regulatory networks that program and coordinate the development of each seed compartment. In addition, it is not known what the functions are of many genes that are expressed in different seed compartments. Identifying the function of compartment-specific genes should provide new insight into their roles in seed development. At present, new genomic resources allow seed biologists to use global gene expression profiling and comparative genomics to answer many questions that only a short time ago seemed out of reach. These questions, and others (Fig. 1D), are challenging the field of seed biology, and their answers should provide new insights into the process of seed development and lead to improved seeds for human and animal consumption.

**LEGUMES ARE AN EXCELLENT MODEL SYSTEM TO STUDY SEED DEVELOPMENT**

Legumes represent one of the largest and most diverse families of flowering plants, with approximately

20,000 species classified (Doyle and Luckow, 2003). There are three subfamilies in legumes and the largest, Papilionoideae, contains most of the model species in which different aspects of plant biology have been studied. The most common legume models are peanut (*Arachis hypogaea*), Lotus (*Lotus japonicus*), Medicago (*Medicago truncatula*), soybean (*Glycine max*), scarlet runner bean (SRB; *Phaseolus coccineus*), common bean (*Phaseolus vulgaris*), pea (*Pisum sativum*), and broad bean (*Vicia faba*). The latter five species have been used historically to study seed and embryo development (Goldberg et al., 1989; Johnson et al., 1994; Coste et al., 2001; Weterings et al., 2001; Weber et al., 2005).

Several features make legumes an excellent model system to study seed and embryo development. For example, many legumes, such as soybean and peanut, are food crops of major economic importance. The mature seeds of these legumes are rich in proteins, carbohydrates, and oils, and accumulate to high nutritional value. These stored seed food reserves make legumes, such as soybean, the second most important crop for human nutrition and animal feed (Rubel et al., 1972; Duranti and Gius, 1997; Graham and Vance,

2003). One advantage of using crop models to study seed biology is to be able to modify traits of agronomic importance, such as improved seed nutritional composition, reduced allergen levels, or increased seed number and size (Kinney, 1998; Herman et al., 2003; Wang et al., 2003; Gupta et al., 2006). In addition, legume seed biology has been studied for more than 150 years using descriptive, physiological, biochemical, molecular, and genetic approaches (see below). These studies have provided a solid intellectual framework for using legume models to study and dissect seed development in our current genomic era. The recent development of genomic tools, such as genome sequences, ESTs, oligonucleotide and cDNA microarrays, and comprehensive databases, such as the Legume Information System (<http://www.comparative-legumes.org>), make legumes an excellent model to study seed development at a global scale (VandenBosch and Stacey, 2003; Gepts et al., 2005; Gonzales et al., 2005). These genomic tools allow comparative genomic analyses in closely related species (Zhu et al., 2005) and should facilitate the identification and investigation of genes important for seed development.

One of the most fascinating characteristics of legumes is that collectively they produce a large range of seed sizes (Fig. 2A). For example, some legume seeds are giants and are excellent models for developmental studies, particularly during early stages of seed development. The large size of SRB globular-stage seed and embryo allows manipulation and isolation of embryonic regions, such as the embryo proper and suspensor, using hand-dissection techniques (Walbot et al.,

1972; Sussex et al., 1973; Weterings et al., 2001). Due to their size, large quantities of cells and tissues from these SRB embryonic regions can be obtained, facilitating molecular and biochemical studies. Manipulation of seeds and embryos at early stages of development is difficult in other model plant species with smaller seeds, such as *Arabidopsis* (*Arabidopsis thaliana*), making many legumes particularly useful to study early developmental seed biology.

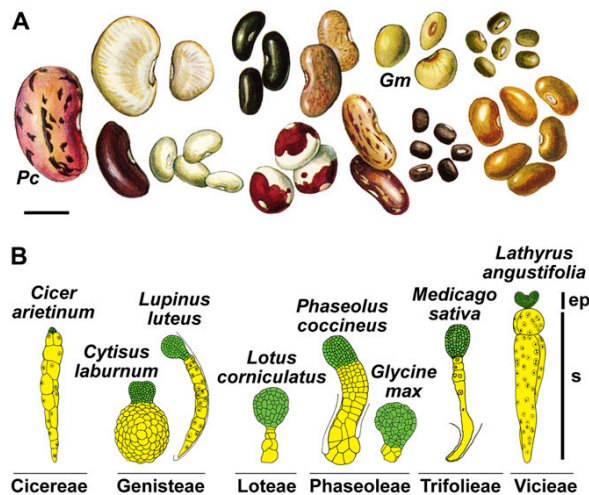
A second novel feature of legumes is that their embryos show a wide range of morphological forms (Fig. 2B). For example, two closely related species, soybean and SRB, have morphologically distinct suspensors. The soybean suspensor is small, consisting of a few cells, whereas the SRB suspensor is much larger and contains several hundred cells (Fig. 2B). The variety in size and shape of legume seeds and embryos makes them excellent models for comparative morphological studies using a functional genomics approach. This strategy can lead to a better understanding of the function, evolution, and diversity of legume seeds and their corresponding compartments.

#### LEGUMES HAVE BEEN USED TO STUDY SEED DEVELOPMENT FOR MORE THAN 150 YEARS

Historically, legumes have been used to address important questions of seed and embryo development. In fact, early work with legumes contributed to the development of major ideas in biology. For example, during the early 1800s, Matthias Schleiden used several legumes, including *Medicago* and *Vicia*, to investigate the endosperm and describe the process of seed development (Schleiden and Vogel, 1838, 1842). These studies contributed to his role in establishing the cell theory. In the mid-1800s, Gregor Mendel used peas to study the inheritance of phenotypic variation, including seed color and shape, leading to his Laws of Inheritance and the establishment of modern-day genetics (Mendel, 1865).

From the late 1800s to the middle of the 1900s, legumes were used to describe the processes of seed and embryo development, including the cellular events that occur before and after fertilization, early embryo cell cleavages, and endosperm differentiation. For example, Guignard's compendium of more than 40 legume species described the rich diversity of legume embryo and suspensor morphologies (Fig. 2B; Guignard, 1882). These studies, and others, contributed to our overall understanding of seed and embryo development at the descriptive level (Martin, 1914; Brown, 1917; Cooper, 1938).

Studies on legume seed formation transitioned from descriptive anatomy to experiments at the molecular, biochemical, and physiological levels during the 1970s (Dure, 1975), although ultrastructural and histochemical studies of legume seed development continued (Johansson and Walles, 1993; Nishizawa et al., 1994; Duval et al., 1995). Work in many legumes provided



**Figure 2.** Diversity of legume seed size and embryo morphology. A, Legume seed shape and size variation. Scale bar = 1 cm. Adapted and modified from Vaughan et al. (1997). B, Variation in embryo-proper (ep) and suspensor (s) morphologies within and between legume tribes. Embryos are not drawn to scale. Images were adapted and modified from Martin (1914), Lersten (1983), and Chamberlin et al. (1994). *Pc*, *P. coccineus*; *Gm*, *G. max*.

some of the earliest measurements of RNA, DNA, carbohydrates, lipids, and protein levels in seeds (Rubel et al., 1972; Clutter et al., 1974; Hill and Breidenbach, 1974; Davies, 1976; Pattee et al., 1981; Singh et al., 1981; Adams et al., 1982; Dhillon and Miksche, 1983). These studies provided new insights into the processes by which food reserves accumulate and are stored in seeds, as well as demonstrating that genome endoreplication processes occur in specific seed compartments (e.g. cotyledon, suspensor). In addition, legumes such as SRB were used to dissect and manipulate the embryo proper and suspensor experimentally to study the role of growth hormones (e.g. GA) in early embryo development (Sussex et al., 1973; Cionini et al., 1976; Alpi et al., 1979).

During this same period, our laboratory used RNA-excess DNA-RNA hybridization experiments to show that approximately 14,000 to 18,000 diverse mRNAs are present in soybean embryos at different developmental stages (Goldberg et al., 1981b, 1989). We also demonstrated that most diverse mRNA species are present throughout seed development, but that small numbers of mRNAs, including those encoding storage proteins, are regulated quantitatively at specific developmental stages (Goldberg et al., 1981a). Seed protein genes were chosen as models to investigate gene regulation during legume seed development because they encode superprevalent mRNAs that could be easily identified and isolated and because of their importance as a food source for human and animal consumption.

Research on genes active in legume seed development exploded during the late 1970s and 1980s when it became possible to clone and study individual mRNAs and genes and reinsert them into plants using newly developed transformation techniques (Bevan et al., 1983; Estrella-Herrera et al., 1983; Fraley et al., 1983). In addition, Murai et al. (1983) demonstrated that the common bean *phaseolin* seed storage protein gene could be transferred to sunflower (*Helianthus annuus*) cells and expressed. This sunflower experiment showed that gene cloning and *Agrobacterium tumefaciens* transformation techniques could be combined to transfer foreign genes into plant cells and study their function. Research in several laboratories with legume seed protein genes, such as  $\beta$ -conglycinin, glycinin, Kunitz trypsin inhibitor, and lectin, showed that their mRNA accumulation patterns are regulated temporally and spatially (Goldberg et al., 1981a, 1983; Meinke et al., 1981; Rerie et al., 1992) and controlled by both transcriptional and posttranscriptional processes (Evans et al., 1984; Beach et al., 1985; Chappell and Chrispeels, 1986; Walling et al., 1986). Subsequent work determined that cis-regulatory sequences flanking legume seed protein genes could confer embryo-specific expression patterns in heterologous plants, such as tobacco (*Nicotiana tabacum*) and petunia (*Petunia hybrida*; Chen et al., 1986; Okamura et al., 1986; Jofuku et al., 1987; Higgins et al., 1988; Naito et al., 1988; Baumlein et al., 1992; Wohlfarth et al., 1998; Chandrasekharan et al., 2003).

This work provided insights into the mechanisms controlling gene activity during seed development and showed that the cis-regulatory sequences and transcription factors controlling legume seed protein gene expression are highly conserved in other plant species.

At present, the remarkable development of new genomic resources makes it possible to study legume gene expression during seed and embryo development at a global level. Currently, *Medicago*, *Lotus*, and soybean cDNA and oligonucleotide microarrays are available (Endo et al., 2002; Vodkin et al., 2004; Firnhaber et al., 2005). An increasing collection of legume seed transcriptome and proteomic data (Thibaud-Nissen et al., 2003; Firnhaber et al., 2005; Hajdich et al., 2005; Buitink et al., 2006; Dhaubhadel et al., 2007), cDNA library collections, and public EST databases (Journet et al., 2002; Shoemaker et al., 2002; Asamizu et al., 2004; Firnhaber et al., 2005; Ramirez et al., 2005) have helped to provide a global view of gene activity at specific seed developmental stages. Sequences of the *Medicago*, *Lotus*, soybean, common bean, and peanut genomes (Broughton et al., 2003; Gepts et al., 2005; Young et al., 2005; Jackson et al., 2006) should provide an invaluable resource for identifying and characterizing genes that play critical roles during seed and embryo development in the near future.

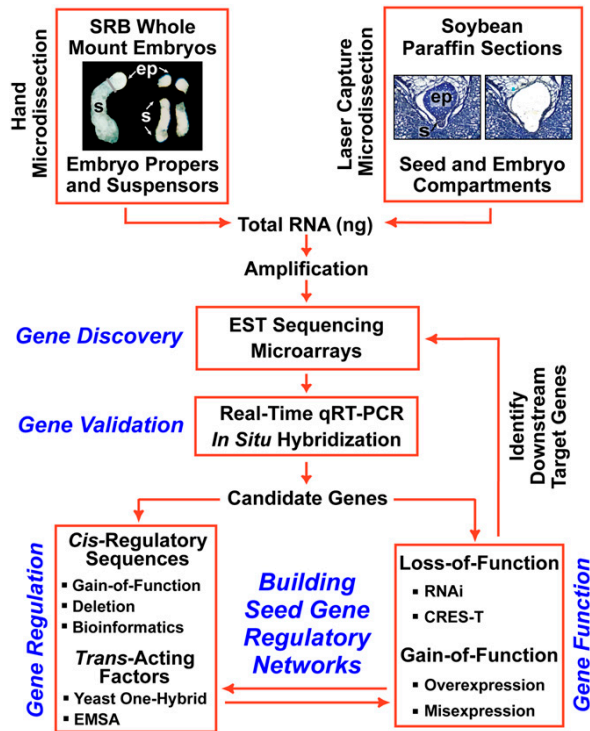
#### AN NOVEL STRATEGY TO DISSECT REGULATORY NETWORKS PROGRAMMING EARLY SEED AND EMBRYO DEVELOPMENT

Our laboratory has developed a genomics strategy to begin to identify the regulatory networks that program legume seed and embryo development (Fig. 3). One aspect of this strategy is to use the giant SRB embryo, pioneered by Sussex and colleagues (Walbot et al., 1972; Sussex et al., 1973), as an entry point to dissect the molecular events that occur during early embryogenesis (Fig. 4). More recently, we incorporated soybean into our strategy (Fig. 5) because it contrasts with SRB in terms of early embryo morphology (Fig. 2B) and because our laboratory has worked on soybean embryo development for more than 25 years (Goldberg et al., 1989). Recent development of soybean and SRB genomic resources allows both of these legumes to be used as excellent systems to identify genes that play important roles in the differentiation of unique seed and embryo compartments and to build compartment-specific regulatory networks that are crucial for seed formation (Figs. 4–6).

#### USING SRB AS A GENOMICS ENGINE TO DISSECT EARLY EMBRYOGENESIS

Our laboratory has utilized SRB as a model system (Weterings et al., 2001) to identify genes and regulatory networks programming early embryo developmental events using a genomics approach (Figs. 3 and 4). We have focused on the question of how the





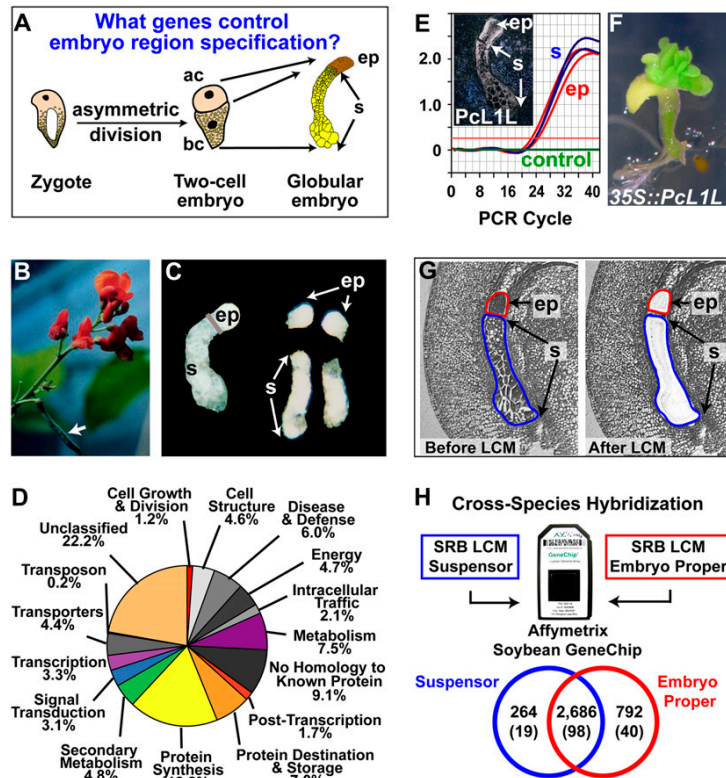
**Figure 3.** Genomics strategy for identifying legume seed gene regulatory networks.

embryo-proper and suspensor regions are specified from the apical and basal cells of a two-cell embryo, respectively (Fig. 4A; Weterings et al., 2001). SRB is unique in this regard because of its giant seed and large embryo (Fig. 2A), permitting hand dissection of embryo-proper and suspensor regions at early developmental stages (Fig. 4, B and C; Walbot et al., 1972; Sussex et al., 1973; Weterings et al., 2001). We prepared cDNA libraries from hand-dissected embryo-proper and suspensor regions of SRB globular-stage embryos (Fig. 3; Weterings et al., 2001) and ESTs were sequenced to determine what genes are active in the embryo proper and suspensor. Our SRB early embryo EST database is available at both <http://estdb.biology.ucla.edu/PcEST> and GenBank (accession nos. CA896559–CA916678). Figure 4D shows the distribution of >16,000 SRB suspensor ESTs grouped by functional categories, illustrating the large diversity of genes that are active in a terminally differentiated suspensor. Surprisingly, more than 300 suspensor transcription factor ESTs were identified that are distributed into a variety of transcription factor gene families (A.Q. Bui, B.H. Le, and R.B. Goldberg, unpublished data), providing a glimpse into the spectrum of regulatory genes active in one region of a legume embryo shortly after fertilization. What roles these transcription factors play in suspensor differentiation and function remain to be determined.

As part of our strategy, we carried out real-time quantitative reverse transcription (qRT)-PCR and in situ hybridization experiments (Fig. 3) to quantify and localize the accumulation of SRB mRNAs during early embryo development. One advantage of using giant SRB embryos is that qRT-PCR can be used to quantify mRNA levels in different regions of a single embryo. For example, PcL1L mRNA uncovered in our EST studies is localized throughout the globular-stage SRB embryo (Fig. 4E, inset; Kwong et al., 2003). *PcL1L* is a relative of the Arabidopsis *LEAFY COTYLEDON1 (LEC1)* CAAT-box-binding transcription factor gene that is a critical regulator of embryo development (Lotan et al., 1998). Real-time qRT-PCR studies on regions dissected from a single SRB embryo at the globular-stage show similar PcL1L mRNA prevalences in the embryo-proper and suspensor regions (Fig. 4E). Remarkably, Arabidopsis plants transformed with a full-length PcL1L cDNA under the control of the cauliflower mosaic virus 35S gene promoter (R.W. Kwong, J. Pelletier, and J.J. Harada, unpublished data) formed ectopic embryo-like structures on seedlings (Fig. 4F), a phenotype similar to that obtained with a 35S::*AtLEC1* chimeric gene (Lotan et al., 1998). These results suggest that *PcL1L* is an important regulator of SRB embryo development, illustrating the power of the SRB system as a gene discovery engine to uncover regulatory genes that play essential roles in seed development.

#### USING SOYBEAN TO IDENTIFY GENES REQUIRED TO MAKE A SEED

Even though the giant SRB embryo is a novel system for studying the specification of embryo regions early in development, this legume does have some limitations for global studies of seed formation. For example, although we have produced an EST dataset for SRB globular-stage embryo regions (<http://estdb.biology.ucla.edu/PcEST>), few other genomic resources are available. In addition, transformation procedures have not yet been developed for SRB. Finally, because SRB is not a major food crop, it is unlikely that a genome project will be carried out to sequence the SRB genome. To complement our use of SRB as a gene discovery engine and overcome these deficiencies, we decided to go back to the future (Goldberg et al., 1989) and use soybean to dissect genes important for making a seed. At the present time, a large number of genomic resources have been developed for soybean, including microarrays, EST databases, and genome sequences (Shoemaker et al., 2002; Vodkin et al., 2004; Jackson et al., 2006). In addition, transformation procedures are well established for soybean, making it feasible to address questions of gene function (Ko et al., 2006; Olhoft et al., 2006). Finally, the unique morphological differences between soybean and SRB embryos (e.g. suspensor size and shape) should permit structure-function questions of legume embryo diversity to be studied (Fig. 2B).



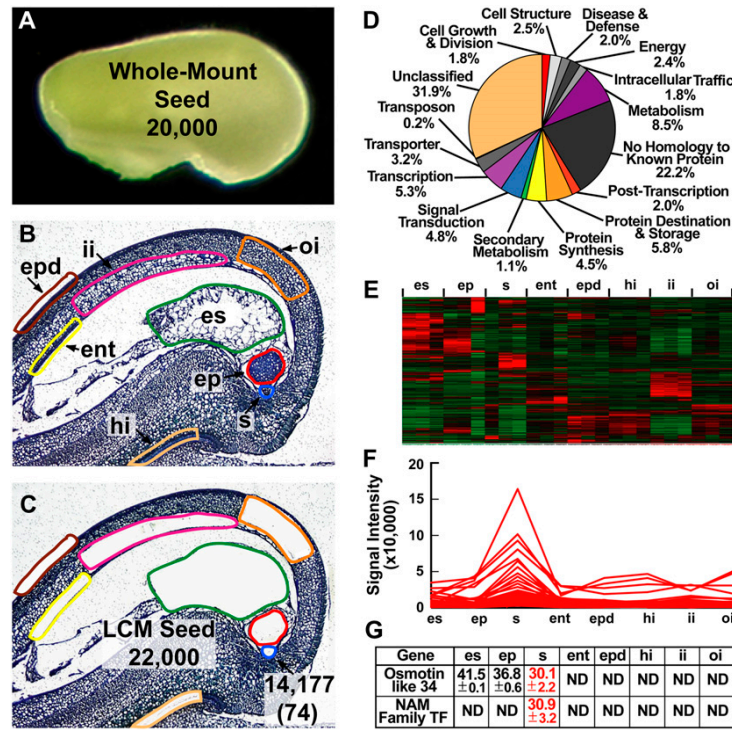
**Figure 4.** Using SRB as a genomics engine to uncover genes active early in embryogenesis. A, Model for the specification of the embryo proper and suspensor, adapted and modified from Weterings et al. (2001). ac, Apical cell; bc, basal cell; ep, embryo proper; s, suspensor. B, SRB plant with a pod indicated by the arrow. C, Hand-dissected SRB globular-stage embryos before and after separating the embryo proper and suspensor. D, Functional category distribution of SRB suspensor ESTs. E, Real-time qRT-PCR validation of Pcl1L mRNA accumulation pattern. Inset, In situ hybridization of SRB globular-stage embryo mRNA using a Pcl1L antisense probe. In situ data were taken from Kwong et al. (2003). F, Pcl1L overexpression in transgenic Arabidopsis seedlings. G, SRB globular-stage seed paraffin sections before and after capturing the embryo proper and suspensor by LCM. H, Cross-species hybridization of SRB embryo-proper and suspensor mRNA captured in G with an Affymetrix Soybean GeneChip containing 37,593 soybean probe sets. Only SRB sequences with a high similarity to soybean EST sequences on the GeneChip will hybridize. Therefore, the proportion of heterologous SRB suspensor and embryo-proper transcripts detected on the GeneChip is lower than that detected for homologous soybean RNAs (Fig. 5C). Venn diagram of transcripts detected in the suspensor and embryo proper is shown. Numbers in parentheses refer to the number of transcription factor gene transcripts detected. Data are available at <http://estdb.biology.ucla.edu/seed>.

Although we have been able to take advantage of the giant SRB embryo to dissect by hand embryo-proper and suspensor regions (Figs. 3 and 4C), this approach is time consuming and not practical with legumes that have smaller embryos, such as soybean (Fig. 2B). One way to overcome the limitations of hand dissection and to be able to isolate different regions from any legume seed and embryo regardless of size (Fig. 2) is to make use of spectacular progress in laser capture microdissection (LCM) technology (Day et al., 2005; Nelson et al., 2006). LCM technology makes it possible to study gene activity in the entire seed because any seed compartment, region, or tissue can be isolated easily throughout development (Fig. 5). LCM technology has been used successfully in rice (*Oryza sativa*), maize (*Zea mays*), tobacco, soybean, and Arabidopsis in which a variety of plant tissues and cell types have been isolated and studied, including those in an early Arabidopsis embryo (Asano et al., 2002; Kerk et al., 2003; Nakazono et al., 2003; Casson et al., 2005; Klink et al., 2005; Sanders et al., 2005; Spencer et al., 2007).

We have been using LCM with soybean seeds to identify all the genes required to make a seed (Figs. 3 and 5). In combination with GeneChip technology, we can investigate the global gene activity profiles in different compartments of the entire seed. For example, we used LCM to isolate the endosperm, suspensor, embryo proper, endothelium, inner integument, outer

integument, epidermis, and hilum from a globular-stage soybean seed (Fig. 5, B and C). We hybridized RNAs isolated from each of these seed regions, as well as from intact globular-stage seeds (Fig. 5, A–C), with soybean Affymetrix GeneChips (J.A. Wagmaister, X. Wang, A.Q. Bui, B.H. Le, and R.B. Goldberg, unpublished data). We then compared the spectrum of diverse transcripts present in the entire globular-stage soybean seed (Fig. 5A) to those obtained with the eight laser-captured seed regions (Fig. 5, B and C). These data are available at <http://estdb.biology.ucla.edu/seed> as part of our National Science Foundation (NSF) Plant Genome Research Project.

Approximately 20,000 diverse transcripts were found to be present in the whole-mount globular-stage soybean seed (Fig. 5A), a value close to that which we obtained more than a quarter of a century ago using Rot curve hybridization technology (Goldberg et al., 1981b, 1989). A smaller number of diverse mRNAs were found to be present in individual soybean globular-stage seed regions. For example, approximately 14,000 diverse transcripts were detected in the suspensor, including those that encode about 700 transcription factors (Fig. 5C). These values are comparable to those obtained with each of the other globular-stage seed regions (i.e. 14,000–17,000 diverse transcripts and 600–800 transcription factor mRNAs; J.A. Wagmaister, X. Wang, A.Q. Bui, and R.B. Goldberg, unpublished data). Soybean suspensor transcripts are distributed



**Figure 5.** Using LCM and transcriptional profiling to identify genes required to make a soybean seed. A, Globular-stage soybean seed showing the approximate number of transcripts detected in the entire seed using the Affymetrix Soybean GeneChip (Whole-Mount Seed). B and C, Globular-stage soybean seed paraffin sections before (B) and after (C) capturing the highlighted seed compartments by LCM. The approximate total number of diverse transcripts detected collectively from LCM seed compartments is shown (LCM Seed) in addition to the total number of transcripts detected in the suspensor (arrow). The number in parentheses refers to suspensor transcripts not detected in other seed compartments at the level of the GeneChip (i.e. suspensor-specific transcripts). Raw data were deposited in the Gene Expression Omnibus (GEO) as data series GSE6414 (<http://www.ncbi.nlm.nih.gov/geo>) and can also be accessed at <http://estdb.biology.ucla.edu/seed>. D, Functional category distribution of soybean suspensor transcripts detected by GeneChip analysis. E, Unsupervised hierarchical clustering of the top 2,000 most varying transcripts detected in all globular-stage seed compartments using dChip version 1.3 (Li and Wong, 2001). F, Supervised cluster analysis of suspensor developmentally regulated transcripts. Values represent the mean and SD of the threshold cycle (Ct) for two biological replicates with two technical replicates each. Ct values were adjusted to an 18S rRNA internal control. One Ct cycle represents a 2-fold difference in RNA prevalence. Lower Ct values indicate higher RNA levels. ent, Endothelium; ep, embryo proper; epd, epidermis; es, endosperm; hi, hilum; ii, inner integument; oi, outer integument; s, suspensor.

into functional categories (Fig. 5D) similar to what we observed from the analysis of giant SRB suspensor ESTs (Fig. 4D). These findings indicate that there is a large diversity of biological functions in the small soybean suspensors (Fig. 5D) and that there does not appear to be any apparent differences in the functional groupings of soybean and SRB suspensor mRNAs, despite large differences in size and morphology (Fig. 2B).

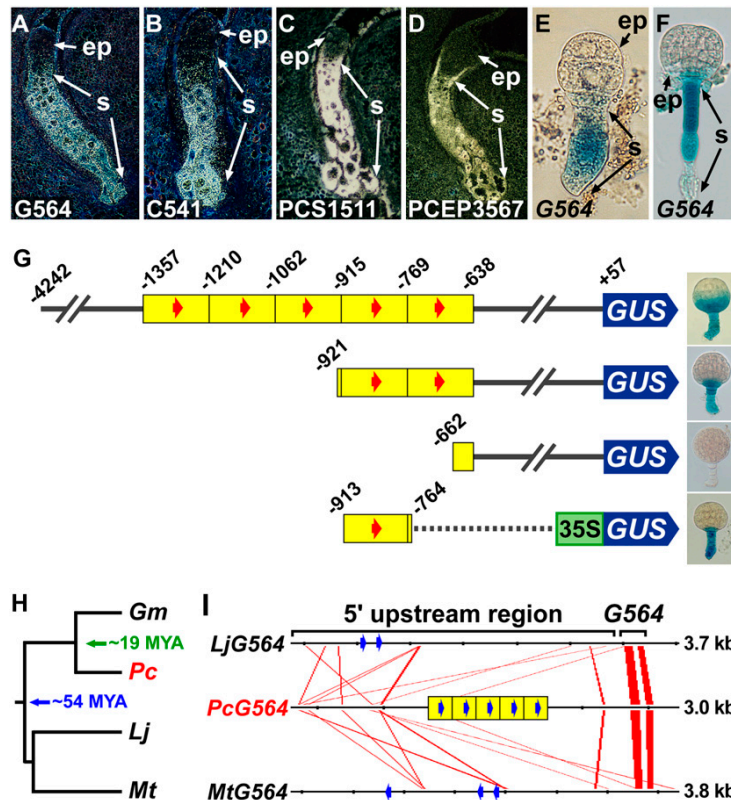
We estimated independently that there are approximately 22,000 diverse transcripts present in a soybean globular-stage seed (Fig. 5, B and C) by taking the union of each individual seed mRNA set captured by LCM (Fig. 5C) in close agreement with the result obtained with intact globular-stage seeds (Fig. 5A). These data indicate that (1) at least 20,000 to 22,000 diverse mRNAs are required to make a globular-stage soybean

seed; (2) the majority of the diverse mRNAs present in each globular-stage seed region are shared with other regions; and (3) there are small sets of seed region-specific mRNAs.

We used hierarchical clustering to determine whether sets of mRNAs that are shared between different soybean globular-stage seed regions are coregulated at a quantitative level (Fig. 5, E and F). Our analyses identified groups of shared coregulated mRNAs that accumulate at a higher level in a particular seed region (Fig. 5E). For example, approximately 65 suspensor transcripts accumulate at a 2-fold or higher level in the suspensor compared with other seed regions (Fig. 5F).

Finally, comparison of diverse mRNAs present in each region of a soybean globular-stage seed identified small sets of region-specific transcripts. For example,





**Figure 6.** Identifying DNA sequences important for suspensor transcription. A to D, In situ hybridization of SRB globular-stage embryos using antisense probes from *G564* (A), *C541* (B), *PCS1511* (C), and *PCEP3567* (D) cDNAs. *G564* and *C541* in situ hybridization data were taken from Weterings et al. (2001). E and F, GUS enzymatic activity in transgenic tobacco (E) and Arabidopsis (F) embryos carrying a *G564::GUS* chimeric gene. G, *G564* 5'-deletion and gain-of-function analyses in transgenic tobacco plants. The number indicates position relative to the transcription start site (+1). Yellow blocks indicate approximately 150-bp tandem duplications. Red arrows indicate a 10-bp motif (GAAAAG<sup>C</sup>/GAA) identified to be conserved in the upstream sequences of *G564* and *C541* (Weterings et al., 2001). Deletion data were taken from Weterings et al. (2001). The -913 to -764 gain-of-function construct was made by fusing this *G564* upstream region with a cauliflower mosaic virus 35S::GUS chimeric gene and transforming tobacco plants according to Koltunow et al. (1990). H, Legume phylogenetic tree, including SRB, soybean, *Lotus*, and *Medicago*. Soybean and *Lotus/Medicago* diverged from SRB approximately 19 and 54 million years ago, respectively (Lavin et al., 2005). I, Conserved regions among legume *G564* genes. Red lines indicate closely related sequences displayed by FamilyRelationsII (20-bp window size, 75% similarity or greater; Brown et al., 2005). Blue arrows indicate motifs identified by MEME (20-bp window size; Bailey and Elkan, 1994) to be significantly enriched in the *G564* upstream regions. ep, Embryo proper; *Gm*, *G. max*; *Lj*, *L. japonicus*; *Mt*, *M. truncatula*; MYA, million years ago; *Pc*, *P. coccineus*; s, suspensor.

74 mRNAs were detected in the suspensor that are undetectable in other seed regions at the level of the GeneChip (Fig. 5C; J.A. Wagmaister, X. Wang, A.Q. Bui, B.H. Le, and R.B. Goldberg, unpublished data). Real-time qRT-PCR experiments validated the GeneChip specificity of two of these suspensor transcripts—an osmotin-like 34 mRNA and a mRNA encoding a member of the NAM transcription factor family (Fig. 5G; C. Cheng, J.A. Wagmaister, A.Q. Bui, and R.B. Goldberg, unpublished data). Taken together, this example illustrates that it is possible to combine LCM and GeneChip technologies to profile the spectrum of mRNAs that are present in any legume seed compartment and region throughout development (Fig. 3). The challenge will be to identify which mRNAs play a critical role in the differentiation of each seed region and how their corresponding genes are organized into regulatory networks in the soybean genome (Fig. 3).

#### USING COMPARATIVE GENOMICS TO IDENTIFY SEED mRNAs IN DIVERSE LEGUME SPECIES

Legumes exhibit a wide range of diversity in seed size and embryo morphology (Fig. 2), providing an outstanding opportunity to use LCM and functional genomics to compare the mRNA sets present in the

different legume seeds. For example, it should be possible to isolate and compare the RNA sets present in legume suspensors that vary greatly in size and form (e.g. SRB, *Medicago*, *Lotus*, and soybean; Fig. 2B) and address the question of what role, if any, variation in suspensor size plays in legume embryo development. Although it is unlikely that GeneChips will be constructed containing diverse mRNAs for each legume species, it is possible to take advantage of the close relationship between legumes at the DNA and RNA levels to use cross-species hybridization approaches to compare gene activity within the seed regions of any legume. Cross-species hybridization approaches have been applied successfully in plants, animals, and fungi where species have diverged from a common ancestor more than 75 million years ago, such as pig and human (Chismar et al., 2002; Moody et al., 2002; Adjaye et al., 2004; Becher et al., 2004; Ji et al., 2004; Wang et al., 2004; Chalmers et al., 2005; Nowrousian et al., 2005). This evolutionary distance is greater than that separating SRB, *Medicago*, *Lotus*, and soybean, which have been shown to diverge from a common ancestor approximately 54 million years ago (Fig. 6; Lavin et al., 2005).

To test this approach, we carried out cross-species hybridization using laser-captured SRB suspensors and embryo-proper RNAs hybridized with soybean GeneChips (Fig. 4, G and H). Our results indicated that



most diverse SRB embryo mRNAs that are conserved enough to be detected by the soybean GeneChip are shared by SRB embryo-proper and suspensor regions (Fig. 4H). By contrast, small sets of mRNAs, including those encoding transcription factors, are specific to each region of the globular-stage SRB embryo at the level of the GeneChip (Fig. 4H). The results from these LCM cross-species hybridization studies complement those generated by sequencing hand-dissected SRB embryo-proper and suspensor ESTs (Figs. 3 and 4D). In addition, they identified 1,000 new embryo proper- and suspensor-specific mRNAs, including those encoding approximately 60 transcription factors that might play key roles in embryo region specification during early embryogenesis (Fig. 4A). Our data indicate that cross-species hybridization using mRNAs from diverse legumes can be successful in identifying genes active during seed development at a global level. Coupling LCM with cross-species hybridization using available legume GeneChips and microarrays (e.g. soybean) should provide an entry point for identifying genes that play important roles in the development of model legume seeds, such as soybean, *Medicago*, and *Lotus*, as well as in those of nonmodel legumes where few genomic resources are available (Fig. 2B).

#### IDENTIFYING REGULATORY NETWORKS REQUIRED TO PROGRAM A LEGUME SEED

Using the genomics strategy that we developed to study the early stages of legume seed and embryo development (Fig. 3), we identified genes, including those encoding transcription factors, that are active specifically in the embryo proper and suspensor of SRB and soybean globular-stage embryos (Figs. 4, D and H, and 5, D and G). We also identified genes that are active specifically in other compartments of the seed (e.g. endosperm, integuments, hilum; Fig. 5C). What DNA sequences and transcription factors regulate compartment-specific genes within a seed and how compartment-specific genes are organized into regulatory networks within a plant genome are important questions of seed biology (Fig. 1D).

As a first step to uncover regulatory networks that operate within legume seeds, we used *in situ* hybridization to identify mRNAs from our EST database that accumulate specifically in the SRB suspensor (Figs. 3 and 4). For example, *G564*, *C541*, *PCS1511*, and *PCEP3567* mRNAs accumulate at a high level in the suspensor of SRB globular-stage embryos (Fig. 6, A–D; Weterings et al., 2001; A.Q. Bui, Y. Bi, and R.B. Goldberg, unpublished data). *G564* and *C541* mRNAs encode proteins with unknown functions (Weterings et al., 2001). By contrast, *PCS1511* and *PCEP3567* mRNAs encode GA 3 $\beta$ -hydroxylase and a homeodomain transcription factor related to *WOX9* (Haecker et al., 2004), respectively. Other mRNAs, such as those encoding additional enzymes in the GA biosynthetic pathway (e.g. *ent*-kaurene synthase, *ent*-kaurene oxidase, GA 20-oxidase), show a similar accumulation pattern in

SRB globular-stage embryos (A.Q. Bui, Y. Bi, and R.B. Goldberg, unpublished data), suggesting that their corresponding genes might be organized into a suspensor regulatory network. To begin dissecting suspensor regulatory networks, our laboratory analyzed in detail the *G564* gene 5'-upstream region (Fig. 3; Weterings et al., 2001).

We showed that approximately 4.2 kb of the *G564* upstream region activates transcription in the suspensor of transgenic tobacco globular-stage embryos, demonstrating that suspensor-specific expression is controlled primarily at the transcriptional level (Fig. 6E; Weterings et al., 2001). In addition, the *G564* upstream region also activates suspensor transcription in transgenic Arabidopsis embryos (Fig. 6F; X. Wang, T. Kawashima, and R.B. Goldberg, unpublished data), suggesting that the machinery regulating suspensor-specific transcription is conserved in flowering plants. *G564* 5'-deletion and gain-of-function analyses identified regions important for suspensor transcription (Fig. 6G; Weterings et al., 2001). The *G564* upstream region possesses five approximately 150-bp tandem duplications (Fig. 6G). Each duplication is capable of activating suspensor transcription, indicating that cis-regulatory sequences within each duplicated fragment are sufficient to direct transcription in the suspensor (Weterings et al., 2001; T. Kawashima, Y. Bi, and R.B. Goldberg, unpublished data). Computational analysis uncovered a conserved 10-bp sequence (GAAAAG<sup>C</sup>/T<sup>T</sup>GAA) in the upstream regions of the SRB *G564* and *C541* genes (Fig. 6G, red arrows; Weterings et al., 2001), suggesting that this motif might play an important role in regulating suspensor-specific transcription during early embryogenesis. If so, the 10-bp motif might be conserved in the upstream regions of other SRB suspensor-specific genes and their orthologs in closely related legumes.

The spectacular increase in legume genome sequences enables comparative approaches to be used to identify conserved cis-regulatory sequences among related legume species. For example, we uncovered *G564* orthologs in soybean, *Lotus*, and *Medicago* (Fig. 6, H and I; T. Kawashima and R.B. Goldberg, unpublished data). Soybean separated from SRB approximately 19 million years ago and from *Lotus* and *Medicago* approximately 54 million years ago (Fig. 6H; Lavin et al., 2005). Results obtained from two different computational analyses uncovered short conserved regions between the 5'-upstream DNA sequences of *G564* genes in SRB, *Lotus*, and *Medicago* (Fig. 6I). The first approach used FamilyRelationsII to identify blocks of similar DNA sequences shared between *G564* upstream regions (Fig. 6I, red lines; Brown et al., 2005). In addition, this program showed that *G564* structure is conserved (two exons and one intron) in these three legumes (Fig. 6I), suggesting strongly that the *G564* genes are orthologous. Blocks of similar DNA sequences in the upstream regions of orthologous genes have been shown to contain cis-regulatory sequences (Yuh et al., 2002). The closely related sequence blocks found in the legume *G564* upstream regions might also contain cis-regulatory

sequences important for suspensor-specific transcription. We also used Multiple Em for Motif Elicitation (MEME; Bailey and Elkan, 1994) to identify sequences significantly enriched in the *G564* upstream regions (Fig. 6I, blue arrows; T. Kawashima and R.B. Goldberg, unpublished data). Significantly, regions identified by MEME in SRB include the conserved 10-bp motif sequence (Weterings et al., 2001). Whether the 10-bp motif is an important suspensor cis-regulatory sequence and what trans-acting factors regulate transcription in the suspensor remain to be determined.

One or more of the suspensor-specific transcription factors that we identified using EST and LCM-GeneChip analyses (Figs. 4 and 5) might interact with the conserved 10-bp motif and other cis-regulatory sequences to control transcription in the suspensor of SRB and other legumes. Similarly, transcription factors specific to other seed compartments (Fig. 5) might play an important role in controlling transcription in different parts of the seed. To date, the molecular mechanisms by which region-specific transcription factors are interconnected to form seed regulatory networks remain unknown. Studying the function of region-specific transcription factors is essential for understanding the importance of these proteins in seed development and for uncovering downstream target genes to construct seed gene regulatory networks (Fig. 3).

Advances in soybean transformation procedures (Ko et al., 2006; Ölhoft et al., 2006) have made it possible to use loss-of-function and gain-of-function strategies to study gene function directly in soybean (Fig. 3). Although T-DNA is used commonly to generate loss-of-function alleles in *Arabidopsis* (Alonso et al., 2003), this technique might not be appropriate in soybean for several reasons. The soybean genome is 8 times larger than that of *Arabidopsis* (Arumuganathan and Earle, 1991) and contains a majority of repetitive sequences (Goldberg, 1978), requiring a large effort to generate T-DNA insertions at a saturation level. Because soybean transformation procedures are not as efficient as the seed transformation technique used in *Arabidopsis* (Clough and Bent, 1998), it would be challenging to produce large numbers of independent transgenic lines. In addition, soybean is a polyploid (Shoemaker et al., 1996; Hymowitz, 2004) and the presence of homeologous genes may complicate the interpretation of knockout results due to gene redundancy. A more productive approach is to utilize RNA interference (RNAi) knockdown strategies to study gene functions that have proven useful in a variety of eukaryotes, including soybean (Fig. 3; Subramanian et al., 2005; Amore and Davidson, 2006; Nunes et al., 2006). For example, Herman et al. (2003) used RNAi in soybean to eliminate allergenic proteins from soybean seeds. The advantage of RNAi is that it can be used to target specific genes and has the potential to knock down sets of closely related genes (Miki et al., 2005; Kaur et al., 2006). This approach, and an analogous one using chimeric repressors to knock down related genes (Fig. 3, CRES-T; Hiratsu et al., 2003), should be feasible

for studying the functions of transcription factors identified by LCM and GeneChip analysis in different compartments of a soybean seed, including those that are redundant in the soybean genome (Fig. 3). The sequence of the soybean genome (Jackson et al., 2006) combined with RNAi studies should make it possible to identify downstream target genes that are regulated by region-specific transcription factors at the global level, facilitating identification of seed and embryo regulatory networks (Fig. 3).

## FUTURE PERSPECTIVES

The study of legume seed development has become exciting due to the availability of new genomic resources and sophisticated techniques, such as LCM and RNA profiling using GeneChip arrays. We have identified genes that are unique to a particular seed compartment and that are coregulated within the context of the soybean globular-stage seed (Fig. 5, C, E, and F). The completion of the soybean genome sequence (Jackson et al., 2006) should allow us to identify conserved motifs among the upstream regions of these unique and coregulated genes, thus facilitating the identification of compartment-specific cis-regulatory sequences that connect seed genes into regulatory networks. In addition to the soybean genome sequence, other legume genome sequences will be available soon (Broughton et al., 2003; Gepts et al., 2005; Young et al., 2005). Genome sequences from diverse legume species will provide an invaluable resource for comparative analysis to identify conserved cis-regulatory sequences that connect seed genes into regulatory networks, an approach that has been successful in other eukaryotes, such as the sea urchin (Yuh et al., 2002; Bolouri and Davidson, 2003). Comparative analysis between legume genome sequences, combined with the GeneChip and EST data obtained from seed and embryo compartments of SRB and soybean (Figs. 4 and 5) and other plants (e.g. *Arabidopsis*; Casson et al., 2005), should facilitate the discovery of genes essential for seed and embryo development, including those important for specific legume traits, such as seed size and embryo morphology (Fig. 2). In addition, comparative analysis of legume genomes with other nonlegume genomes, such as *Arabidopsis*, rice, and poplar (*Populus* spp.), will advance the discovery of genes important for seed development in flowering plants (Graham et al., 2004; Zhu et al., 2005). Finally, once a soybean whole-genome GeneChip becomes available, the entire mRNA profiles of all seed and embryo compartments can be determined, completing the identification of seed- and embryo-specific genes. Taken together, remarkable advances in genomic resources will allow us to answer questions regarding seed and embryo development (Fig. 1D) that were not possible only a few years ago. It is now becoming realistic in this genomic era to understand what genes and regulatory networks are required to make a legume seed.

Sequence data from this article can be found in the GenBank/EMBL data libraries under accession numbers CA896559 to CA916678 (ESTs) and AF325187 (G564).

## ACKNOWLEDGMENTS

We are grateful to all the members of our laboratory, past and present, who have helped to establish soybean and SRB as powerful systems to investigate seed development. We particularly acknowledge Dr. Koen Weterings, Dr. Yuping Bi, and Dr. Xingjun Wang for contributing to many of the experiments summarized in this *Update*. In addition, we thank Ms. Chen Cheng for carrying out the real-time qRT-PCR experiment presented in Figure 5G. We would also like to acknowledge Ian Sussex, Roger Beachy, Tim Hall, Maarten Chrispeels, Niels Nielsen, Lila Vodkin, Don Boulter, T.J. Higgins, Klaus Muntz, and Uli Wobus whose laboratories helped to provide a foundation for understanding gene activity during legume seed development.

Received March 28, 2007; accepted April 18, 2007; published June 6, 2007.

## LITERATURE CITED

- Adams CA, Norby SW, Rinne RW (1982) Protein modification and utilization of starch in soybean (*Glycine max* (L.) Merr.) seed saturation. *J Exp Bot* 33: 279–287
- Adjaye J, Herwig R, Herrmann D, Wruck W, Benkahl A, Brink TC, Nowak M, Carnwath JW, Hultschig C, Niemann H, et al (2004) Cross-species hybridisation of human and bovine orthologous genes on high density cDNA microarrays. *BMC Genomics* 5: 83
- Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen HM, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, et al (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* 301: 653–657
- Alpi A, Lorenzi R, Cionini PG, Bennici A, Damato F (1979) Identification of gibberellin A<sub>1</sub> in the embryo suspensor of *Phaseolus coccineus*. *Planta* 147: 225–228
- Amore G, Davidson EH (2006) cis-Regulatory control of cyclophilin, a member of the ETS-DRI skeletogenic gene battery in the sea urchin embryo. *Dev Biol* 293: 555–564
- Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. *Plant Mol Biol Rep* 9: 208–218
- Asamizu E, Nakamura Y, Sato S, Tabata S (2004) Characteristics of the *Lotus japonicus* gene repertoire deduced from large-scale expressed sequence tag (EST) analysis. *Plant Mol Biol* 54: 405–414
- Asano T, Masumura T, Kusano H, Kikuchi S, Kurita A, Shimada H, Kadowaki K (2002) Construction of a specialized cDNA library from plant cells isolated by laser capture microdissection: toward comprehensive analysis of the genes expressed in the rice phloem. *Plant J* 32: 401–408
- Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2: 28–36
- Baumlein H, Nagy I, Villarroel R, Inze D, Wobus U (1992) Cis-analysis of a seed protein gene promoter—the conservative RY repeat CATGCATG within the legumin box is essential for tissue-specific expression of a legumin gene. *Plant J* 2: 233–239
- Beach LR, Spencer D, Randall PJ, Higgins TJV (1985) Transcriptional and post-transcriptional regulation of storage protein gene-expression in sulfur-deficient pea-seeds. *Nucleic Acids Res* 13: 999–1013
- Becher M, Talke IN, Krall L, Kramer U (2004) Cross-species microarray transcript profiling reveals high constitutive expression of metal homeostasis genes in shoots of the zinc hyperaccumulator *Arabidopsis halleri*. *Plant J* 37: 251–268
- Bevan MW, Flavell RB, Chilton M (1983) A chimeric antibiotic resistance gene as a selectable marker for plant cell transformation. *Nature* 304: 184–187
- Bolouri H, Davidson EH (2003) Transcriptional regulatory cascades in development: initial rates, not steady state, determine network kinetics. *Proc Natl Acad Sci USA* 100: 9371–9376
- Borisjuk L, Rolletschek H, Radchuk R, Weschke W, Wobus U, Weber H (2004) Seed development and differentiation: a role for metabolic regulation. *Plant Biol* 6: 375–386
- Broughton WJ, Hernandez G, Blair M, Beebe S, Gepts P, Vanderleyden J (2003) Beans (*Phaseolus* spp.)—model food legumes. *Plant Soil* 252: 55–128
- Brown CT, Xie Y, Davidson EH, Cameron RA (2005) Paircomp, Family-RelationsII and Cartwheel: tools for interspecific sequence comparison. *BMC Bioinformatics* 6: 70
- Brown MM (1917) The development of the embryo-sac and of the embryo in *Phaseolus vulgaris*. *Bull Torrey Bot Club* 44: 535–544
- Buitink J, Leger JJ, Guisle I, Vu BL, Wullemme S, Lamirault G, Le Bars A, Le Meur N, Becker A, Kuester H, et al (2006) Transcriptome profiling uncovers metabolic and regulatory processes occurring during the transition from desiccation-sensitive to desiccation-tolerant stages in *Medicago truncatula* seeds. *Plant J* 47: 735–750
- Casson S, Spencer M, Walker K, Lindsey K (2005) Laser capture microdissection for the analysis of gene expression during embryogenesis of *Arabidopsis*. *Plant J* 42: 111–123
- Chalmers AD, Goldstone K, Smith JC, Gilchrist M, Amaya E, Papalopolu N (2005) A *Xenopus tropicalis* oligonucleotide microarray works across species using RNA from *Xenopus laevis*. *Mech Dev* 122: 355–363
- Chamberlin MA, Horner HT, Palmer RG (1994) Early endosperm, embryo, and ovule development in *Glycine max* (L.) Merr. *Int J Plant Sci* 155: 421–436
- Chandrasekharan MB, Bishop KJ, Hall TC (2003) Module-specific regulation of the beta-phaseolin promoter during embryogenesis. *Plant J* 33: 853–866
- Chappell J, Chrispeels MJ (1986) Transcriptional and posttranscriptional control of phaseolin and phytohemagglutinin gene expression in developing cotyledons of *Phaseolus vulgaris*. *Plant Physiol* 81: 50–54
- Chen ZL, Schuler MA, Beachy RN (1986) Functional analysis of regulatory elements in a plant embryo-specific gene. *Proc Natl Acad Sci USA* 83: 8560–8564
- Chismar JD, Mondala T, Fox HS, Roberts E, Langford D, Masliah E, Salomon DR, Head SR (2002) Analysis of result variability from high-density oligonucleotide arrays comparing same-species and cross-species hybridizations. *Biotechniques* 33: 516–518, 520, 522, 524
- Cionini PG, Bennici A, Alpi A, Damato F (1976) Suspensor, gibberellin and in vitro development of *Phaseolus coccineus* embryos. *Planta* 131: 115–117
- Clough SJ, Bent AF (1998) Floral dip: a simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. *Plant J* 16: 735–743
- Clutter M, Brady T, Walbot V, Sussex I (1974) Macromolecular synthesis during plant embryogenesis—cellular rates of RNA synthesis in diploid and polytene cells in bean embryos. *J Cell Biol* 63: 1097–1102
- Cooper DC (1938) Embryology of *Pisum sativum*. *Bot Gaz* 100: 123–132
- Coste E, Ney B, Crozat Y (2001) Seed development and seed physiological quality of field grown beans (*Phaseolus vulgaris* L.). *Seed Sci Technol* 29: 121–136
- Davies DR (1976) DNA and RNA contents in relation to cell and seed weight in *Pisum sativum*. *Plant Sci Lett* 7: 17–25
- Day RC, Grossniklaus U, Macknight RC (2005) Be more specific! Laser-assisted microdissection of plant cells. *Trends Plant Sci* 10: 397–406
- Dhaubhadel S, Gijzen M, Moy P, Farhangkhoe M (2007) Transcriptome analysis reveals a critical role of CHS7 and CHS8 genes for isoflavonoid synthesis in soybean seeds. *Plant Physiol* 143: 326–338
- Dhillon SS, Miksche JP (1983) DNA, RNA, protein and heterochromatin changes during embryo development and germination of soybean (*Glycine max* L.). *Histochem J* 15: 21–37
- Doyle JJ, Luckow MA (2003) The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. *Plant Physiol* 131: 900–910
- Duranti M, Gius C (1997) Legume seeds: protein content and nutritional value. *Field Crops Res* 53: 31–45
- Dure LS (1975) Seed formation. *Annu Rev Plant Physiol Plant Mol Biol* 26: 259–278
- Duval M, Pepin R, Job C, Derpierre C, Douce R, Job D (1995) Ultrastructural localization of the major biotinylated protein from *Pisum sativum* seeds. *J Exp Bot* 46: 1783–1786
- Endo M, Matsubara H, Kokubun T, Masuko H, Takahata Y, Tsuchiya T, Fukuda H, Demura T, Watanabe M (2002) The advantages of cDNA microarray as an effective tool for identification of reproductive organ-specific genes in a model legume, *Lotus japonicus*. *FEBS Lett* 514: 229–237
- Estrella-Herrera L, Depicker A, Van Montagu M, Schell J (1983) Expression of chimeric genes transferred into plant cells using a Ti-plasmid-derived vector. *Nature* 303: 209–213

- Evans LS, Dimitriadis L, Hinkley DA (1984) Seed protein quantities of field-grown soybeans exposed to simulated acidic rain. *New Phytol* **97**: 71–76
- Firnhaber C, Puhler A, Kuster H (2005) EST sequencing and time course microarray hybridizations identify more than 700 *Medicago truncatula* genes with developmental expression regulation in flowers and pods. *Planta* **222**: 269–283
- Fraleigh RT, Rogers SG, Horsch RB, Sanders PR, Flick JS, Adams SP, Bittner ML, Brand LA, Fink CL, Fry JS, et al (1983) Expression of bacterial genes in plant cells. *Proc Natl Acad Sci USA* **80**: 4803–4807
- Gehring M, Choi Y, Fischer RL (2004) Imprinting and seed development. *Plant Cell* **16**: S203–S213
- Gepts P, Beavis WD, Brummer EC, Shoemaker RC, Stalker HT, Weeden NF, Young ND (2005) Legumes as a model plant family. Genomics for food and feed report of the Cross-Legume Advances Through Genomics Conference. *Plant Physiol* **137**: 1228–1235
- Goldberg RB (1978) DNA sequence organization in the soybean plant. *Biochem Genet* **16**: 45–68
- Goldberg RB, Barker SJ, Perez-Grau L (1989) Regulation of gene expression during plant embryogenesis. *Cell* **56**: 149–160
- Goldberg RB, Depaiva G, Yadegari R (1994) Plant embryogenesis—zygote to seed. *Science* **266**: 605–614
- Goldberg RB, Hoschek G, Ditta GS, Breidenbach RW (1981a) Developmental regulation of cloned superabundant embryo mRNAs in soybean. *Dev Biol* **83**: 218–231
- Goldberg RB, Hoschek G, Tam SH, Ditta GS, Breidenbach RW (1981b) Abundance, diversity, and regulation of mRNA sequence sets in soybean embryogenesis. *Dev Biol* **83**: 201–217
- Goldberg RB, Hoschek G, Vodkin LO (1983) An insertion sequence blocks the expression of a soybean lectin gene. *Cell* **33**: 465–475
- Gonzales MD, Archuleta E, Farmer A, Gajendran K, Grant D, Shoemaker R, Beavis WD, Waugh ME (2005) The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic Acids Res* **33**: D660–665
- Graham MA, Silverstein KAT, Cannon SB, VandenBosch KA (2004) Computational identification and characterization of novel genes from legumes. *Plant Physiol* **135**: 1179–1197
- Graham PH, Vance CP (2003) Legumes: importance and constraints to greater use. *Plant Physiol* **131**: 872–877
- Guignard L (1882) Recherches anatomiques et physiologiques sur des légumineuses. PhD thesis. University of Paris, Paris
- Gupta PK, Rustgi S, Kumar N (2006) Genetic and molecular basis of grain size and grain number and its relevance to grain productivity in higher plants. *Genome* **49**: 565–571
- Haecker A, Gross-Hardt R, Geiges B, Sarkar A, Breuninger H, Herrmann M, Laux T (2004) Expression dynamics of WOX genes mark cell fate decisions during early embryonic patterning in *Arabidopsis thaliana*. *Development* **131**: 657–668
- Hajdúch M, Ganapathy A, Stein JW, Thelen JJ (2005) A systematic proteomic study of seed filling in soybean. Establishment of high-resolution two-dimensional reference maps, expression profiles, and an interactive proteome database. *Plant Physiol* **137**: 1397–1419
- Haughn G, Chaudhury A (2005) Genetic analysis of seed coat development in *Arabidopsis*. *Trends Plant Sci* **10**: 472–477
- Herman EM, Helm RM, Jung R, Kinney AJ (2003) Genetic modification removes an immunodominant allergen from soybean. *Plant Physiol* **132**: 36–43
- Higgins TJV, Newbigin EJ, Spencer D, Llewellyn DJ, Craig S (1988) The sequence of a pea vicilin gene and its expression in transgenic tobacco plants. *Plant Mol Biol* **11**: 683–695
- Hill JE, Breidenbach RW (1974) Proteins of soybean seeds II. Accumulation of major protein components during seed development and maturation. *Plant Physiol* **53**: 747–751
- Hiratsuka K, Matsui K, Koyama T, Ohme-Takagi M (2003) Dominant repression of target genes by chimeric repressors that include the EAR motif, a repression domain, in *Arabidopsis*. *Plant J* **34**: 733–739
- Hymowitz T (2004) Speciation and cytogenetics. In HR Boerma, J Specht, eds, *Soybeans: Improvement, Production, and Uses*, Ed 3. American Society of Agronomy, Madison, WI, pp 97–129
- Jackson SA, Rokhsar D, Stacey G, Shoemaker RC, Schmutz J, Grimwood J (2006) Toward a reference sequencing of the soybean genome: a multiagency effort. *Crop Sci* **46**: S55–S61
- Ji W, Zhou W, Gregg K, Yu N, Davis S, Davis S (2004) A method for cross-species gene expression analysis with high-density oligonucleotide arrays. *Nucleic Acids Res* **32**: e93
- Jofuku KD, Okamoto JK, Goldberg RB (1987) Interaction of an embryo DNA binding protein with a soybean lectin gene upstream region. *Nature* **328**: 734–737
- Johansson M, Walles B (1993) Functional-anatomy of the ovule in broad bean, *Vicia faba* L. 2. Ultrastructural development up to early embryogenesis. *Int J Plant Sci* **154**: 535–549
- Johnson S, Liu CM, Hedley CL, Wang TL (1994) An analysis of seed development in *Pisum sativum* XVIII. The isolation of mutants defective in embryo development. *J Exp Bot* **45**: 1503–1511
- Journet EP, van Tuinen D, Gouzy J, Crespeau H, Carreau V, Farmer MJ, Niebel A, Schiex T, Jaillon O, Chatagnier O, et al (2002) Exploring root symbiotic programs in the model legume *Medicago truncatula* using EST analysis. *Nucleic Acids Res* **30**: 5579–5592
- Kaur J, Sebastian J, Siddiqi I (2006) The *Arabidopsis*-mei2-like genes play a role in meiosis and vegetative growth in *Arabidopsis*. *Plant Cell* **18**: 545–559
- Kerk NM, Ceserani T, Tausta SL, Sussex IM, Nelson TM (2003) Laser capture microdissection of cells from plant tissues. *Plant Physiol* **132**: 27–35
- Kinney AJ (1998) Plants as industrial chemical factories—new oils from genetically engineered soybeans. *Fett Lipid* **100**: 173–176
- Klink VP, Alkharouf N, MacDonald M, Matthews B (2005) Laser capture microdissection (LCM) and expression analyses of *Glycine max* (soybean) syncytium containing root regions formed by the plant pathogen *Heterodera glycines* (soybean cyst nematode). *Plant Mol Biol* **59**: 965–979
- Ko TS, Korban SS, Somers DA (2006) Soybean (*Glycine max*) transformation using immature cotyledon explants. *Methods Mol Biol* **343**: 397–405
- Koltunow AM, Truettner J, Cox KH, Wallroth M, Goldberg RB (1990) Different temporal and spatial gene expression patterns occur during anther development. *Plant Cell* **2**: 1201–1224
- Krishnan HB (2005) Engineering soybean for enhanced sulfur amino acid content. *Crop Sci* **45**: 454–461
- Kwong RW, Bui AQ, Lee H, Kwong LW, Fischer RL, Goldberg RB, Harada JJ (2003) LEAFY COTYLEDON1-LIKE defines a class of regulators essential for embryo development. *Plant Cell* **15**: 5–18
- Laux T, Wurschum T, Breuninger H (2004) Genetic regulation of embryonic pattern formation. *Plant Cell (Suppl)* **16**: S190–S202
- Lavin M, Herendeen PS, Wojciechowski MF (2005) Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. *Syst Biol* **54**: 575–594
- Lersten NR (1983) Suspensors in Leguminosae. *Bot Rev* **49**: 233–257
- 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
- Lotan T, Ohto M, Yee KM, West MA, Lo R, Kwong RW, Yamagishi K, Fischer RL, Goldberg RB, Harada JJ (1998) *Arabidopsis* LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells. *Cell* **93**: 1195–1205
- Martin JN (1914) Comparative morphology of some Leguminosae contributions from the Hull Botanical Laboratory. *Bot Gaz* **58**: 154–167
- Meinke DW, Chen J, Beachy RN (1981) Expression of storage-protein genes during soybean seed development. *Planta* **153**: 130–139
- Mendel G (1865) Versuche über Pflanzenhybriden. In *Verhandlungen des naturforschenden Vereines*, Vol Bd. IV für das Jahr. Abhandlungen, Brünn, Czech Republic, pp 3–47
- Miki D, Itoh R, Shimamoto K (2005) RNA silencing of single and multiple members in a gene family of rice. *Plant Physiol* **138**: 1903–1913
- Miller SS, Bowman LAA, Gijzen M, Miki BLA (1999) Early development of the seed coat of soybean (*Glycine max*). *Ann Bot (Lond)* **84**: 297–304
- Moise JA, Han S, Gudynaite-Savitch L, Johnson DA, Miki BLA (2005) Seed coats: structure, development, composition, and biotechnology. In *Vitro Cell Dev Biol Plant* **41**: 620–644
- Moody DE, Zou Z, McIntyre L (2002) Cross-species hybridisation of pig RNA to human nylon microarrays. *BMC Genomics* **3**: 27
- Moravec T, Schmidt MA, Herman EM, Woodford-Thomas T (2007) Production of *Escherichia coli* heat labile toxin (LT) B subunit in soybean seed and analysis of its immunogenicity as an oral vaccine. *Vaccine* **25**: 1647–1657
- Murai N, Sutton DW, Murray MG, Slightom JL, Merlo DJ, Reichert NA, Senguptagopalan C, Stock CA, Barker RF, Kemp JD, et al (1983) Phaseolin gene from bean is expressed after transfer to sunflower via tumor-inducing plasmid vectors. *Science* **222**: 476–482

- Murray DR (1987) Nutritive role of seed coats in developing legume seeds. *Am J Bot* 74: 1122–1137
- Naito S, Dube PH, Beachy RN (1988) Differential expression of conglycinin alpha-subunit and beta-subunit genes in transgenic plants. *Plant Mol Biol* 11: 109–123
- Nakazono M, Qiu F, Borsuk LA, Schnable PS (2003) Laser-capture microdissection, a tool for the global analysis of gene expression in specific plant cell types: identification of genes expressed differentially in epidermal cells or vascular tissues of maize. *Plant Cell* 15: 583–596
- Nelson T, Tausta SL, Gandotra N, Liu T (2006) Laser microdissection of plant tissue: What you see is what you get. *Annu Rev Plant Biol* 57: 181–201
- Nishizawa NK, Mori S, Watanabe Y, Hirano H (1994) Ultrastructural-localization of the basic 7S globulin in soybean (*Glycine Max*) cotyledons. *Plant Cell Physiol* 35: 1079–1085
- Nowrousian M, Ringelberg C, Dunlap JC, Loros JJ, Kuck U (2005) Cross-species microarray hybridization to identify developmentally regulated genes in the filamentous fungus *Sordaria macrospora*. *Mol Genet Genomics* 273: 137–149
- Nunes ACS, Vianna GR, Cuneo F, Amaya-Farfan J, de Capdeville G, Rech EL, Aragao FJL (2006) RNAi-mediated silencing of the myo-inositol-1-phosphate synthase gene (GmMIP1) in transgenic soybean inhibited seed development and reduced phytate content. *Planta* 224: 125–132
- Okamoto JK, Jofuku KD, Goldberg RB (1986) Soybean seed lectin gene and flanking nonseed protein genes are developmentally regulated in transformed tobacco plants. *Proc Natl Acad Sci USA* 83: 8240–8244
- Olhoft PM, Donovan CM, Somers DA (2006) Soybean (*Glycine max*) transformation using mature cotyledonary node explants. *Methods Mol Biol* 343: 385–396
- Pattee HE, Young CT, Giesbrecht FG (1981) Seed size and storage effects on carbohydrates of peanuts. *J Agric Food Chem* 29: 800–802
- Ramirez M, Graham MA, Blanco-Lopez L, Silvente S, Medrano-Soto A, Blair MW, Hernandez G, Vance CP, Lara M (2005) Sequencing and analysis of common bean ESTs. Building a foundation for functional genomics. *Plant Physiol* 137: 1211–1227
- Rerie WG, Newbigin EJ, Higgins TJV (1992) Genes encoding seed globulins in legumes. In IM Morrison, ed, *Advances in Plant Cell Biochemistry and Biotechnology: A Research Annual*, Vol 1. Jai Press, Greenwich, CT, pp 53–104
- Rubel A, Rinne RW, Canvin DT (1972) Protein, oil, and fatty-acid in developing soybean seeds. *Crop Sci* 12: 739–741
- Sanders PM, Bui AQ, Le BH, Goldberg RB (2005) Differentiation and degeneration of cells that play a major role in tobacco anther dehiscence. *Sex Plant Reprod* 17: 219–241
- Schleiden MJ, Vogel T (1838) Über das Albumen, insbesondere der Leguminosen, Vol 19: Pt I. Verhandlungen der Kaiserlich Leopoldinisch-Carolinischen Deutschen Akademie der Naturforscher, Bonn
- Schleiden MJ, Vogel T (1842) Über das Albumen, insbesondere der Leguminosen, Vol 19: Pt II. Verhandlungen der Kaiserlich Leopoldinisch-Carolinischen Deutschen Akademie der Naturforscher, Breslau, Poland
- Shoemaker R, Keim P, Vodkin L, Retzel E, Clifton SW, Waterston R, Smoller D, Coryell V, Khanna A, Erpelding J, et al (2002) A compilation of soybean ESTs: generation and analysis. *Genome* 45: 329–338
- Shoemaker R, Polzin K, Labate J, Specht J, Brummer EC, Olson T, Young N, Concibido V, Wilcox J, Tamulonis JP, (1996) Genome duplication in soybean (*Glycine* subgenus *soja*). *Genetics* 144: 329–338
- Singh U, Jambunathan R, Saxena NP (1981) Changes in carbohydrates, amino-acids and proteins in developing seed of chickpea. *Phytochemistry* 20: 373–378
- Spencer MWB, Casson SA, Lindsey K (2007) Transcriptional profiling of the Arabidopsis embryo. *Plant Physiol* 143: 924–940
- Stangeland B, Salehian Z, Aalen R, Mandal A, Olsen OA (2003) Isolation of GUS marker lines for genes expressed in Arabidopsis endosperm, embryo and maternal tissues. *J Exp Bot* 54: 279–290
- Subramanian S, Graham MY, Yu O, Graham TL (2005) RNA interference of soybean isoflavone synthase genes leads to silencing in tissues distal to the transformation site and to enhanced susceptibility to *Phytophthora sojae*. *Plant Physiol* 137: 1345–1353
- Sussex I, Clutter M, Walbot V, Brady T (1973) Biosynthetic activity of suspensor of *Phaseolus coccineus*. *Caryologia* 25: 261–272
- Thibaud-Nissen F, Shealy RT, Khanna A, Vodkin LO (2003) Clustering of microarray data reveals transcript patterns associated with somatic embryogenesis in soybean. *Plant Physiol* 132: 118–136
- VandenBosch KA, Stacey G (2003) Summaries of legume genomics projects from around the globe. Community resources for crops and models. *Plant Physiol* 131: 840–865
- Vaughan JG, Geissler C, Nicholson B, Nicholson B (1997) *The New Oxford Book of Food Plants*. Oxford University Press, New York
- Vodkin LO, Khanna A, Shealy R, Clough SJ, Gonzalez DO, Philip R, Zabala G, Thibaud-Nissen F, Sidorous M, Stromvik MV, et al (2004) Microarrays for global expression constructed with a low redundancy set of 27,500 sequenced cDNAs representing an array of developmental stages and physiological conditions of the soybean plant. *BMC Genomics* 5: 73
- Walbot V, Clutter M, Sussex IM (1972) Reproductive development and embryogeny in *Phaseolus*. *Phytomorphology* 22: 59–68
- Walling L, Drews GN, Goldberg RB (1986) Transcriptional and post-transcriptional regulation of soybean seed protein mRNA levels. *Proc Natl Acad Sci USA* 83: 2123–2127
- Wang TL, Domoney C, Hedley CL, Casey R, Grusak MA (2003) Can we improve the nutritional quality of legume seeds? *Plant Physiol* 131: 886–891
- Wang Z, Dooley TP, Curto EV, Davis RL, VandeBerg JL (2004) Cross-species application of cDNA microarrays to profile gene expression using UV-induced melanoma in *Monodelphis domestica* as the model system. *Genomics* 83: 588–599
- Weber H, Borisjuk L, Wobus U (2005) Molecular physiology of legume seed development. *Annu Rev Plant Biol* 56: 253–279
- Weterings K, Apuya NR, Bi Y, Fischer RL, Harada JJ, Goldberg RB (2001) Regional localization of suspensor mRNAs during early embryo development. *Plant Cell* 13: 2409–2425
- Wilcox JR (2004) World distribution and trade of soybean. In RH Boerma, JE Specht, eds, *Soybeans: Improvement, Production, and Uses*, Ed 3. American Society of Agronomy, Madison, WI, pp 1–14
- Wohlfarth T, Braun H, Kirik V, Kollé K, Czihal A, Tewes A, Luerssen H, Misera S, Shutov A, Baumlein H (1998) Regulation and evolution of seed globulin genes. *J Plant Physiol* 152: 600–606
- Yeung EC, Meinke DW (1993) Embryogenesis in angiosperms—development of the suspensor. *Plant Cell* 5: 1371–1381
- Young ND, Cannon SB, Sato S, Kim D, Cook DR, Town CD, Roe BA, Tabata S (2005) Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. *Plant Physiol* 137: 1174–1181
- Yuh CH, Brown CT, Livi CB, Rowen L, Clarke PJ, Davidson EH (2002) Patchy interspecific sequence similarities efficiently identify positive cis-regulatory elements in the sea urchin. *Dev Biol* 246: 148–161
- Zhu H, Choi HK, Cook DR, Shoemaker RC (2005) Bridging model and crop legumes through comparative genomics. *Plant Physiol* 137: 1189–1196

## CHAPTER TWO

### GLOBAL ANALYSIS OF GENE ACTIVITY DURING ARABIDOPSIS SEED DEVELOPMENT AND IDENTIFICATION OF SEED-SPECIFIC TRANSCRIPTION FACTORS

# Global analysis of gene activity during *Arabidopsis* seed development and identification of seed-specific transcription factors

Brandon H. Le<sup>a,1</sup>, Chen Cheng<sup>a,1</sup>, Anthu Q. Bui<sup>a,1</sup>, Javier A. Wagmaister<sup>a,2</sup>, Kelli F. Henry<sup>a</sup>, Julie Pelletier<sup>b</sup>, Linda Kwong<sup>b</sup>, Mark Belmonte<sup>b</sup>, Ryan Kirkbride<sup>b</sup>, Steve Horvath<sup>c</sup>, Gary N. Drews<sup>d</sup>, Robert L. Fischer<sup>e</sup>, Jack K. Okamoto<sup>f</sup>, John J. Harada<sup>b</sup>, and Robert B. Goldberg<sup>a,3</sup>

<sup>a</sup>Department of Molecular, Cell, and Developmental Biology, and <sup>d</sup>Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095; <sup>b</sup>Section of Plant Biology, Division of Biological Sciences, University of California, Davis, CA 95616; <sup>c</sup>Department of Biology, University of Utah, Salt Lake City, UT 84112; <sup>e</sup>Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720; and <sup>f</sup>United States Department of Agriculture, Agricultural Research Service, Beltsville, MD 20705

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2001.

Contributed by Robert B. Goldberg, March 19, 2010 (sent for review February 19, 2010)

**Most of the transcription factors (TFs) responsible for controlling seed development are not yet known. To identify TF genes expressed at specific stages of seed development, including those unique to seeds, we used Affymetrix GeneChips to profile *Arabidopsis* genes active in seeds from fertilization through maturation and at other times of the plant life cycle. Seed gene sets were compared with those expressed in prefertilization ovules, germinating seedlings, and leaves, roots, stems, and floral buds of the mature plant. Most genes active in seeds are shared by all stages of seed development, although significant quantitative changes in gene activity occur. Each stage of seed development has a small gene set that is either specific at the level of the GeneChip or up-regulated with respect to genes active at other stages, including those that encode TFs. We identified 289 seed-specific genes, including 48 that encode TFs. Seven of the seed-specific TF genes are known regulators of seed development and include the *LEAFY COTYLEDON (LEC)* genes *LEC1*, *LEC1-LIKE*, *LEC2*, and *FUS3*. The rest represent different classes of TFs with unknown roles in seed development. Promoter- $\beta$ -glucuronidase (*GUS*) fusion experiments and seed mRNA localization GeneChip datasets showed that the seed-specific TF genes are active in different compartments and tissues of the seed at unique times of development. Collectively, these seed-specific TF genes should facilitate the identification of regulatory networks that are important for programming seed development.**

embryo | transcriptome | mRNA localization

Seed development in higher plants begins with a double fertilization process that occurs within the ovule and ends with a dormant seed primed to become the next plant generation (1, 2). The major events that occur during seed development are shown schematically in Fig. 1 and are described elsewhere in detail (1–8). Genetic studies with *Arabidopsis* have uncovered several genes that play major roles in seed development (9–11), including those that govern endosperm formation (12, 13), embryo differentiation (14, 15), and seed coat development (8). In addition, molecular studies with *Arabidopsis* and other plants have identified the *cis*-control regions of several genes active during seed development, particularly those encoding storage proteins, and the transcription factors (TFs) that play a role in their regulation (16–22). Nevertheless, the identities of most regulators of seed development and their direct targets are largely unknown.

To date, many studies have been carried out by using microarrays to identify genes that are expressed at different times of the plant life cycle (23–26). Only a few of these studies, however, have focused exclusively on seeds and/or embryos to identify important regulators and processes required for seed development (27–31). In this paper, we present results of Affymetrix GeneChip experiments that profile genes that are active before, during, and after *Arabidopsis* seed for-

mation. Our experiments identified 48 TF genes that are active exclusively, or at elevated levels, in seeds. These seed-specific TF genes encode several classes of TFs, are active at different developmental times, and may be important for controlling stage-specific biological events during seed formation. Chimeric promoter- $\beta$ -glucuronidase (*GUS*) transgene experiments and laser capture microdissection (LCM) (32) GeneChip datasets demonstrated that the seed-specific TF genes are active in specific seed compartments and tissues, suggesting that they may play an important role in the differentiation and/or function of unique seed parts. Our data represent an important step toward identifying gene regulatory networks (33) in the *Arabidopsis* genome that are responsible for programming seed development. What these seed-specific TF genes do and how they are integrated into regulatory networks remain to be determined.

## Results

**Overview and GeneChip Analysis.** We carried out GeneChip hybridization experiments (*Materials and Methods*) by using mRNAs isolated from *Arabidopsis* unfertilized ovules (OV); seeds containing (i) zygotes (24H), (ii) globular-stage embryos (GLOB), (iii) cotyledon-stage embryos (COT), (iv) mature green embryos (MG), and (v) postmature green embryos (PMG); and post-germination seedlings (SDLG) to identify genes that are active before, during, and after seed development (Figs. 1 and 24 and *SI Materials and Methods*). These mRNAs represent gene sets that are active during periods when major events occur within the seed—including embryo differentiation, endosperm formation, seed coat development, storage reserve accumulation, and maturation (Fig. 1). We applied a stringent protocol to analyze our GeneChip data and restricted our analysis to mRNAs for which the detection call by the Microarray Analysis Suite (MAS) 5.0 software was P (Present) in both biological replicates to reduce the inclusion of false positives (*Materials and Methods*). Only probe sets with consensus detection calls of PP were considered to represent

Author contributions: B.H.L., A.Q.B., G.N.D., R.L.F., J.K.O., J.J.H., and R.B.G. designed research; B.H.L., C.C., A.Q.B., J.A.W., K.F.H., J.P., L.K., M.B., and R.K. performed research; S.H. contributed new reagents/analytic tools; B.H.L., C.C., and R.B.G. analyzed data; and B.H.L. and R.B.G. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, [www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo) (accession no. GSE680).

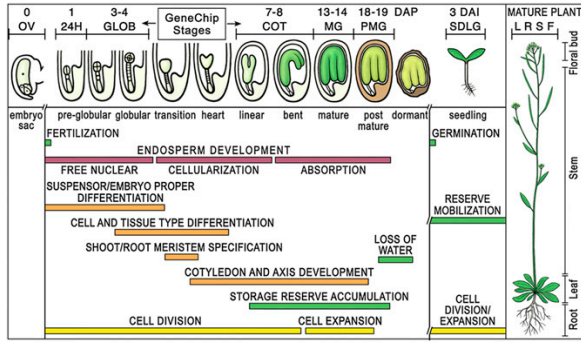
<sup>1</sup>B.H.L., C.C., and A.Q.B. contributed equally to this work.

<sup>2</sup>Present address: Bayer CropSciences, Buenos Aires, Argentina, 1605.

<sup>3</sup>To whom correspondence should be addressed. E-mail: bobg@ucla.edu.

This article contains supporting information online at [www.pnas.org/cgi/content/full/1003530107/DCSupplemental](http://www.pnas.org/cgi/content/full/1003530107/DCSupplemental).





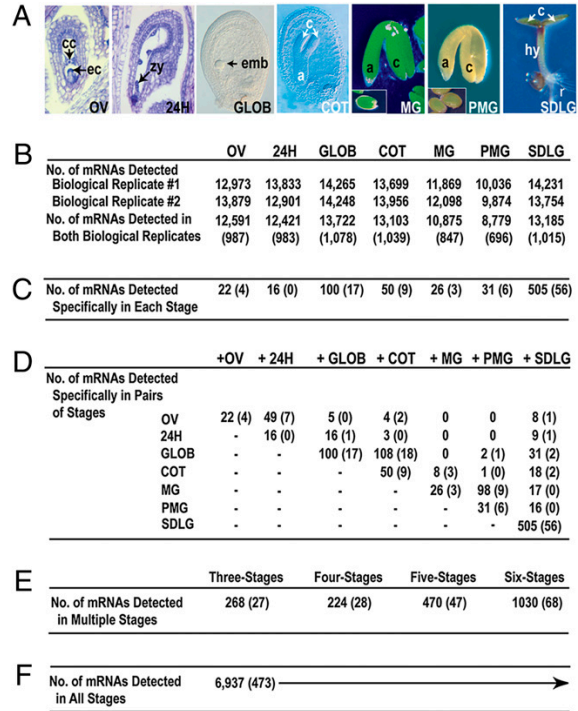
**Fig. 1.** Schematic representation of *Arabidopsis* seed development and stages of the life cycle used for GeneChip analysis. Seed cartoons were adapted from Bowman and Mansfield (57) and are not drawn to scale. Developmental events were modified from Goldberg et al. (1). Stages used for GeneChip analysis are described in *SI Materials and Methods*. Numbers correspond to days after pollination (DAP) or days after imbibition (DAI). Brackets mark the range of embryo stages included in each GeneChip seed sample. OV, unfertilized ovule; 24H, 24-h postpollination seed; GLOB, globular-stage seed; COT, cotyledon-stage seed; MG, mature-green-stage seed; PMG, postmature-green-stage seed; SDLG, seedling; L, leaf; R, root; S, stem; F, floral buds.

mRNAs present in any given developmental stage. Probe sets with discordant MAS 5.0 detection calls between biological replicates [e.g., P and A (Absent), or PA] were assigned a consensus detection call of Insufficient (INS) and removed from further analysis across all datasets used for comparative analysis (*Materials and Methods*). Thus, the seed stage-specific and seed-specific mRNA sets presented in this paper represent the minimum number of genes that are active at specific periods of seed development.

**mRNAs Present Before, During, and After Seed Development.** We detected  $\approx 9,000$ – $14,000$  unique mRNAs present at different developmental stages (Fig. 2B). Pearson correlation coefficients between biological replicates ranged from 0.96 to 0.99 (Fig. S1), indicating excellent concordance with each other. The number of mRNAs did not vary significantly in the periods before and after fertilization (OV and 24H), including the early stages of seed development (GLOB and COT) [ $P > 0.30$ , Analysis of Variance (ANOVA)]. By contrast, the number of mRNAs detected in MG and PMG seed stages decreased significantly ( $P < 0.001$ , ANOVA), and the values we obtained agreed with those in the late seed development AtGenExpress dataset (25). After germination, there was a significant increase in gene activity compared with seed MG and PMG stages ( $P < 0.001$ , ANOVA) (Fig. 2B). Collectively, we detected 15,563 diverse mRNAs throughout seed development (24H to PMG), and 16,701 mRNAs before, during, and after seed formation (OV to SDLG).

The average Pearson correlation coefficients between GeneChip experiments using mRNAs from different stages ranged from 0.97 for OV and 24H samples to 0.20 for PMG and SDLG samples. In general, Pearson correlation coefficients decreased as the seed stage pairs became more distant to each other developmentally. For example, the average correlation coefficients between GLOB and COT, GLOB and MG, and GLOB and PMG mRNAs were 0.87, 0.41, and 0.24, respectively. We grouped features on the *Arabidopsis* ATH1 GeneChip array into functional categories (34) (Fig. S24) and determined that specific functional groups were enriched or reduced in the mRNA populations at each developmental period (Table S1 in Dataset S1).

Taken together, these data show that the number of diverse seed mRNAs decreased significantly when the seed began preparing for dormancy, major changes in gene activity occurred



**Fig. 2.** Genes active before, during, and after *Arabidopsis* seed development. (A) Bright-field (OV, 24H), Nomarski (GLOB, COT), and whole-mount (MG, PMG, SDLG) photographs of prefertilization ovule, seed stages, and post-germination seedling used for GeneChip analysis, respectively. OV and 24H seed samples were visualized from 10  $\mu$ m stained paraffin sections (58). Insets show seeds used to dissect whole-mount MG and PMG embryos. Embryo in COT seed is at the linear cotyledon (LCOT) stage (Fig. 1). (B) Number of mRNAs detected at each stage of development. Numbers for biological replicates 1 and 2 indicate the number of probe sets with a MAS 5.0 detection call of P in each experiment (*Materials and Methods*). The number for both biological replicates indicates a consensus probe set detection call of PP and was used for subsequent analysis (*Materials and Methods*). Scatter plots and correlation coefficients comparing biological replicates are presented in Fig. S1. (C–F) Minimum number of specific and shared mRNAs at each developmental stage. The stringent filtering process used for these analyses is outlined in *Materials and Methods*. A total of 8,510 probe sets with INS (e.g., PA) or marginal (MM) consensus calls between biological replicates listed in B were removed across all developmental stages (*SI Materials and Methods*). The remaining probe sets were used to determine the number of stage-specific mRNAs (C) and mRNAs shared by two stages (D), three to six stages (E), or all stages (F). mRNAs (6,178 of the 6,937) shared by all stages (F) varied quantitatively across development ( $P < 0.05$ , ANOVA). Number in parentheses indicates TF mRNAs. The identities of mRNAs in each category (e.g., seed-stage-specific) are listed in Tables S3, S4, and S7 of Dataset S1. a, axis; c, cotyledon; cc, central cell; ec, egg cell; emb, embryo; hy, hypocotyl; r, roots; zy, zygote.

across seed development, and at least 16,000 genes are active throughout seed development.

**TF mRNAs Present Before, During, and After Seed Development.** We detected  $\approx 700$ – $1,000$  diverse TF mRNAs in each developmental stage, representing 36–55% of TF mRNAs represented on the *Arabidopsis* ATH1 GeneChip (Fig. 2B). Fewer TF mRNAs were detected in the MG and PMG stages compared with early seed stages (24H to COT) and postgermination SDLG, reflecting the decrease in mRNAs as a whole late in seed development (Fig. 2B). The proportion of TF transcripts relative to total mRNAs within a population, however, was the same for all stages (i.e.,  $\approx 8\%$ ). Collectively, we detected 1,327 diverse TF mRNAs throughout seed development



(24H to PMG) and 1,455 TF mRNAs before, during, and after seed formation (OV to SDLG).

We annotated features on the *Arabidopsis* ATH1 GeneChip array corresponding to major TF families to determine the spectrum of TF mRNAs present before, during, and after seed development (Fig. S2B). All major TF families were represented in the mRNA population at each developmental stage. However, significant differences were observed in the representation of specific TF families (Table S2 in Dataset S1). Taken together, these data suggest that at least 1,300 diverse TF mRNAs are required to program all of seed development, the number of TF mRNAs decreases before dormancy, and that the representation of specific TF mRNA families differs at specific developmental periods.

#### Each Seed Developmental Stage Has a Small Set of Specific mRNAs.

We identified a small number of mRNAs specific to each stage at the level of the GeneChip, including those encoding TFs from a variety of different families (Fig. 2C and Table S3 in Dataset S1). The stage-specific mRNAs included a range of functional categories, although the majority encoded proteins that were either unclassified or had no known function (Table S3 in Dataset S1). Approximately half of the seed-stage-specific mRNAs were also found to be seed-specific (see Fig. 4 below), suggesting that they play important roles in seed development. The largest numbers of seed-stage-specific mRNAs were observed in the GLOB and COT stages when major differentiation and morphological events occur during seed development (Fig. 1). For example, 100 GLOB-specific and 50 COT-specific mRNAs were identified, including 17 and 9 stage-specific TF mRNAs, respectively (Fig. 2C and Table S3 in Dataset S1). The GLOB-specific TF mRNAs included those encoding AUXIN RESPONSE FACTOR21 (ARF21, AT1G34410), LATERAL ORGAN BOUNDARIES35 (LBD35, AT5G35900), LBD15 (AT2G40470), and MINISEED3 (MINI3, AT1G55600)—the latter playing a major role in seed size (35). By contrast, we identified <75 mRNAs that were specific for the 24H, MG, and PMG stages, including 9 TF mRNAs (Fig. 2C). After germination, >500 SDLG-specific mRNAs were observed, including 56 SDLG-specific TF mRNAs (Fig. 2C and Table S3 in Dataset S1).

We compared the mRNA sets represented in pairs of developmental stages to determine whether there were seed-period-specific mRNAs in addition to those unique to individual stages (Fig. 2D). We observed that pairs of seed stages that were close to each other developmentally (e.g., OV-24H, GLOB-COT) had small sets of mRNAs that were not detected at the level of the GeneChip at other stages of development (Fig. 2D and Table S4 in Dataset S1). For example, OV and 24H seeds had 49 specific mRNAs that were absent in other stages as well as in post-germination SDLG. Similarly, GLOB and COT seeds had 108 mRNAs, including 18 encoding TFs, that were not detected at any other stage investigated, whereas MG and PMG seeds had 98 mRNAs, including 9 TFs, that were absent at other developmental periods. Neither 24H and MG nor 24H and PMG seeds, by contrast, had any detectable pair-specific mRNAs (Fig. 2D). Analysis of Gene Ontology (GO) terms enriched in both the seed-stage-specific and seed-period-specific mRNA sets (SI Materials and Methods) indicated that the GLOB and COT mRNA sets were enriched in sequences encoding gene regulatory functions (i.e., TFs) ( $P < 0.01$ ), whereas the MG and PMG mRNA sets were enriched in sequences encoding seed dormancy and embryo developmental functions [e.g., late embryo abundant (lea) proteins] ( $P < 0.01$ ) (Table S5 in Dataset S1).

We used the Harada-Goldberg LCM GeneChip dataset (Materials and Methods) to determine where the seed-stage-specific and seed-period-specific mRNAs were localized within the seed (Table S6 in Dataset S1). Most of the GLOB- and GLOB-COT-specific mRNAs were localized either within the endosperm or the seed coat. Very few were present exclusively within the embryo, although many of the endosperm mRNAs were also colocalized within the suspensor. Most

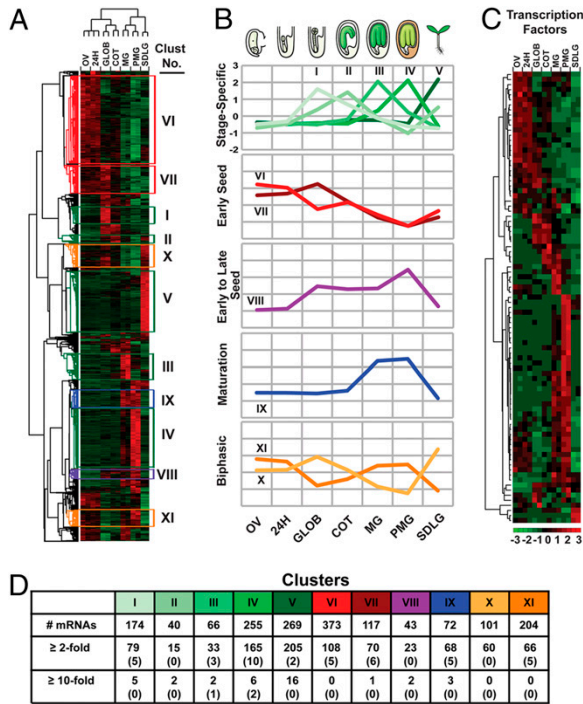
embryo mRNAs were probably below the detection limit of our GeneChip experiments (see below) because of their small contribution to the entire seed mRNA population at these developmental stages (Figs. 1 and 2A). By contrast, the MG- and MG-PMG-specific mRNAs were localized primarily within the embryo or seed coat at these stages of seed development when the embryo is fully differentiated and occupies the majority of space within the seed (Figs. 1 and 2A).

We carried out quantitative reverse transcription PCR (qRT-PCR) experiments with 19 seed-stage-specific mRNAs (Fig. 2C) and seed-period-specific mRNAs (Fig. 2D), including 13 TF mRNAs, to determine whether they were present at other stages, but below the detection limit of the GeneChip (Table S10 in Dataset S1). In our experiments, the detection level was  $\approx 2 \times 10^{-5}$ , or one transcript per 200,000 (Fig. S3A), which is similar to that determined by others using Affymetrix GeneChips (36, 37). All tested mRNAs were validated by qRT-PCR in their target stages at levels similar to those observed with the GeneChip (Table S10 in Dataset S1 and Fig. S3B; Pearson correlation coefficient = 0.76). Most mRNAs tested ( $\approx 70\%$ ), however, were also detected in one or more other stages at greatly reduced levels ( $\approx 10$ – $10,000$  fold) (Table S10 in Dataset S1). The small number of stage-specific mRNAs determined by qRT-PCR to be present in other stages at levels close to that of the target stage, or with slightly reduced levels (2- to 3-fold reduction), had GeneChip signal intensities bordering on the limits of detection (Table S10 in Dataset S1 and Fig. S3C). Taken together, these results indicate that each stage and period of seed development has a small set of mRNAs, including those encoding TFs, that is either absent from other stages or present at highly reduced levels, and that many of these mRNAs are localized in specific parts of the seed.

#### Most Diverse Seed mRNAs Are Present Before, During, and After Seed Development.

By contrast with the small number of seed-stage- and seed-period-specific mRNAs (Fig. 2C and D), most seed mRNAs ( $\approx 7,000$ ), including those encoding TFs ( $\approx 500$ ), were detected before (OV), during (24H to PMG), and after seed formation (SDLG) (Fig. 2F and Table S7 in Dataset S1), indicating that most diverse seed mRNA sequences are present from fertilization through dormancy. A large number of mRNAs were also observed in mosaic combinations of three to six stages (e.g., OV, 24H, and PMG; COT, PMG, and SDLG), although the total number ( $\approx 2,000$ ) was significantly less than those mRNAs shared across development (Fig. 2E and Table S4 in Dataset S1). Most of the shared mRNAs (4,237 of 6,937 or  $\approx 61\%$ ), including those encoding TFs (301 of 473 or  $\approx 64\%$ ), changed significantly in prevalence by at least 2-fold during at least one period of development (e.g., OV-24H) ( $P < 0.01$ ,  $t$  test).

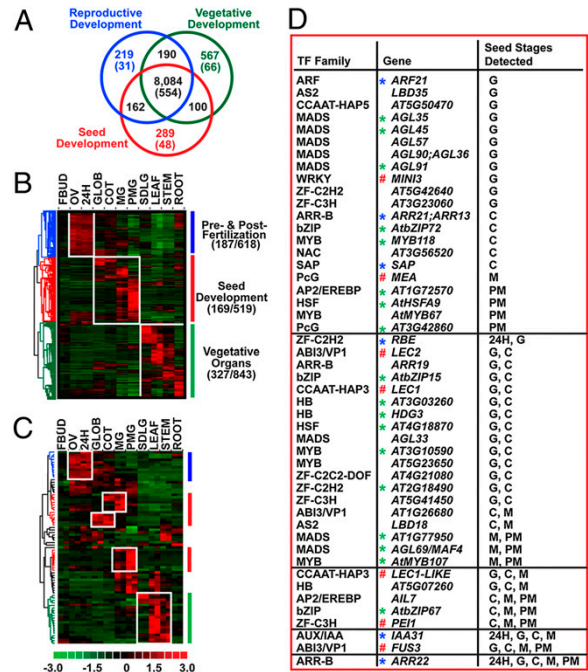
We carried out unsupervised hierarchical clustering analysis on a subset of 2,000 shared mRNAs that had the largest standard deviation in signal levels across all developmental stages (i.e., varied the most quantitatively) to identify up-regulated sets of mRNAs (Fig. 3A and Table S7 in Dataset S1). The different mRNA samples (e.g., OV to SDLG) clustered according to their temporal relationships during seed formation, reflecting the correlation coefficients between developmentally similar and dissimilar mRNA populations (Fig. 3A, top brackets). For example, the OV and 24H shared mRNAs were more similar in their quantitative profiles than the OV and SDLG shared mRNA sets. We identified 11 prominent mRNA clusters in the shared mRNA population (clusters I to XI), including those encoding TFs, demonstrating that complex patterns of quantitative mRNA changes occurred during seed development coinciding with unique developmental events (Fig. 3A–C and Table S7 in Dataset S1). Shared mRNAs grouped into seed-stage-up-regulated clusters (GLOB, COT, MG, and PMG) (I–IV) and seed-period up-regulated clusters for (i) early seed development (OV to GLOB) (VI and VII), (ii) early to late seed development (GLOB to PMG) (VIII), (iii) late seed development (MG and PMG) (IX), and (iv) postgermination SDLG (V) (Fig. 3A and B). In addition, two biphasic mRNA clusters (X and XI) were identified that were up-regulated in temporally



**Fig. 3.** Quantitative regulation of mRNAs shared by all stages of seed development. (A) Unsupervised hierarchical clustering of GeneChip samples (Fig. 1) and probe sets with a consensus detection call of PP in all stages of development (Fig. 2F) was carried out by using dChip 1.3 (56) as described in *Materials and Methods*. Only the top 2,000 probe sets with the most varying signals across all stages were included in the clustering analysis (*SI Materials and Methods*). Numbered boxes highlight individual clusters of coregulated mRNAs. The identities of the top 2,000 mRNAs used for this clustering analysis and the mRNAs in each cluster are listed in [Table S7 of Dataset S1](#). GO terms that are enriched significantly in each cluster ( $P < 0.01$ ) are presented in [Tables S8 and S9 of Dataset S1](#). (B) Graphical representation of cluster mRNA accumulation patterns. Lines represent the average mRNA accumulation pattern for all mRNAs in each cluster. (C) Unsupervised hierarchical clustering of 89 TF mRNAs included in the top 2,000 most varying probe sets shared by all stages of seed development (Fig. 2F) presented in A. The identities of TF mRNAs in each cluster are listed in [Tables S7 and S9 of Dataset S1](#). (D) The number of mRNAs in each cluster shown in B and the number of mRNAs per cluster that increased significantly in prevalence  $\geq 2$ -fold and  $\geq 10$ -fold relative to the mean signal intensity of each cluster ( $P < 0.05$ ). Scale from  $-3$  (green) to  $+3$  (red) represents the relative number of standard deviations from the mean signal intensity for each probe set across all developmental stages.

distinct developmental periods. Cluster X mRNAs were up-regulated in GLOB seeds and postgermination SDLG, whereas cluster XI contained mRNAs that were highly prevalent in OV-24H and MG-PMG seeds (Fig. 3A and B). Approximately 30% (cluster XI) to 94% (cluster IX) of the mRNAs in each cluster increased in prevalence by at least 2-fold, with a smaller proportion of mRNAs (1–5%) in most clusters increasing  $>10$ -fold ( $P < 0.05$ ,  $t$  test) (Fig. 3D). The highest mRNA abundance change for the shared mRNAs was  $\approx 35$ -fold and occurred during maturation (MG and PMG, cluster IX) and after seed germination (SDLG, cluster V) ([Table S9 in Dataset S1](#)).

GO term analysis of cluster mRNAs showed enrichment for sequences programming distinct processes at specific stages of seed development ([Tables S8 and S9 in Dataset S1](#)). For example, the mRNAs present in cluster I (GLOB), cluster II (COT), and cluster III (MG) were enriched for sequences involved in carbohydrate metabolic processes ( $P < 7.7 \times 10^{-6}$ ), rhamnose biosynthesis ( $P <$



**Fig. 4.** Identification of *Arabidopsis* seed-specific mRNAs. (A) GeneChip data obtained for all stages of the life cycle (Fig. 2 and Fig. S4) were partitioned into three groups: (i) reproductive development [OV and floral buds (FBUD) (blue circle)], (ii) seed development [(24H, GLOB, COT, MG, and PMG) (red circle)], and (iii) vegetative development [SDLG, leaf, stem, and root] (green circle). Processing and filtering of the data are outlined in *SI Materials and Methods*. Number in parentheses indicates number of TF mRNAs. The identities of seed-specific, seed-specific TF, reproductive-organ-specific, and vegetative-organ-specific mRNAs are listed in [Table S13 of Dataset S2](#). GO terms that are enriched significantly in the seed-, reproductive-, and vegetative-specific mRNA sets ( $P < 0.01$ ) are presented in [Table S15 of Dataset S2](#). Seed development accumulation patterns and representation of functional groups for the 289 seed-specific mRNAs are shown in Fig. S5. (B and C) Unsupervised hierarchical clustering of mRNAs (B) and TF mRNAs (C) shared by all periods of the life cycle [i.e., intersection of mRNA sets in A] was carried out by using dChip 1.3 (56) as described in *SI Materials and Methods* and Fig. 3A legend. Only the top 2,000 probe sets with the most varying signals across all periods of the life cycle were included in the clustering analysis shown in B (*SI Materials and Methods*). All 77 TF mRNAs included in the top 2,000 most varying probe sets shared by all life cycle periods (B) were used for the clustering analysis shown in C. Blue, red, and green bars highlight mRNA clusters that are up-regulated in (i) OV and 24H seeds, (ii) GLOB, COT, MG, and PMG seeds, and (iii) SDLG and vegetative organs, respectively. The number of mRNAs in each cluster that increased significantly in prevalence  $\geq 2$ -fold relative to the mean signal intensity of each cluster ( $P < 0.05$ ) is listed next to the bars in B. The identities of mRNAs shared throughout the life cycle (A) and those that are present in up-regulated clusters (B and C) are listed in [Table S17 of Dataset S2](#). GO terms that are enriched significantly in each cluster ( $P < 0.01$ ) are presented in [Table S18 of Dataset S2](#). (D) TF families and stage specificity of seed-specific TF mRNAs identified in A. Seed-specific TF mRNAs were classified into families as shown in Fig. S5. Homozygous T-DNA insertion lines for TF genes marked with a \* did not produce a detectable seed phenotype (green \* by us and blue \* by others) ([Table S20 in Dataset S2](#)). Mutations in seed-specific TF genes marked with a # were shown previously by us (e.g., *lec 1*, *lect1-like*, *lec2*, *mea*) and by others (e.g., *pei1*, *fus3*), to produce a seed-defective phenotype ([Table S20 in Dataset S2](#)). Scale from  $-3$  (green) to  $+3$  (red) represents the relative number of standard deviations from the mean signal intensity for each probe set across all developmental stages.

$9.6 \times 10^{-4}$ ), and fatty acid metabolism ( $P < 5 \times 10^{-4}$ ), respectively. These GO terms most likely reflect key biological events that occur in seeds during these time periods; for example, starch accumulation

in the outer integument seed coat layer (GLOB), mucilage formation in the seed coat (COT), and fatty acid synthesis in the embryos during maturation (MG) (5, 38). Each cluster contained TF mRNAs (Fig. 3 C and D) that were up-regulated >2-fold in prevalence and may be important for regulating the GO-term biological processes that occur during the corresponding developmental period (Fig. 3C and Tables S8 and S9 in Dataset S1). For example, AtMyb61 mRNA (AT1G09540), ATAF1 mRNA (AT1G01720), ATbZIP53 mRNA (AT3G62420), and SPLAYED (SYD) mRNA (AT2G28290) were prevalent in cluster I (GLOB), cluster IV (PMG), cluster IX (MG and PMG), and cluster XI (OV-24H and MG-PMG), respectively. These TF mRNAs have been shown to regulate seed coat mucilage extrusion (AtMyb61) (39), ABA response (ATAF1) (40), maturation gene expression (ATbZIP53) (41), and cotyledon boundary-shoot meristem formation (SYD) (42), respectively (Table S9 in Dataset S1).

Taken together, these data show that (i) most diverse seed mRNAs, including TF mRNAs, are present before, during, and after seed formation, (ii) shared seed mRNAs undergo significant prevalence changes and are grouped into stage- and period-specific clusters of coup-regulated mRNA sets, and (iii) coup-regulated mRNAs within each cluster encode proteins involved in important seed processes.

**Most Seed mRNA Sequences Are Present Throughout the Plant Life Cycle.** We carried out GeneChip hybridization experiments by using mRNAs isolated from floral buds (FBUD), leaves (L), stems (S), and roots (R) (Fig. S4) and compared the combined reproductive (FBUD, OV) and vegetative (L, R, S, SDLG) mRNA populations with those present throughout seed development (24H, GLOB, COT, MG, and PMG) to determine the representation of seed mRNAs at other times of the life cycle (Fig. 4A and Table S13 in Dataset S2). Most diverse seed mRNA sequences were represented in the reproductive and vegetative organs of the mature plant (Fig. 4A), in addition to being present before (OV) and after seed germination (SDLG) (Fig. 2). We identified a minimum of 8,084 diverse mRNAs, including 562 TF mRNAs, that were shared by the seed, floral, and vegetative mRNA populations (Fig. 4A) by using our stringent filtering criterion (*Materials and Methods*). A smaller number of life cycle mosaic mRNAs (100–200) were shared by different combinations of two of the three developmental periods (e.g., seed and vegetative development; Fig. 4A and Table S16 in Dataset S2). Unsupervised hierarchical clustering analysis of the top 2,000 most varying life cycle-shared mRNAs (*Materials and Methods*) indicated there were clusters of up-regulated mRNA sets, including those encoding TFs, specific for each period of the life cycle, including seed development (Fig. 4 B and C and Table S17 in Dataset S2). Clusters of up-regulated GLOB, COT, MG, and PMG seed mRNAs were identified in the life cycle-shared mRNA population, including those encoding TFs (Fig. 4 B and C), that contained sequences also present in up-regulated shared seed mRNA clusters at the same developmental stages (Fig. 3). For example, ~80% of the mRNAs that were present within the PMG cluster of the shared seed mRNA population (cluster IV, Fig. 3A and B) were also found within the PMG seed cluster of the life cycle shared mRNA population (Fig. 4B). GO term analysis of the up-regulated seed mRNAs (Fig. 4B) showed enrichment for sequences involved in processes that occur during seed development and maturation (e.g., water deprivation, fatty acid biosynthesis; Table S18 in Dataset S2). Thus, many stage- and period-specific seed mRNAs were up-regulated in the context of both seed development and the entire plant life cycle.

Collectively, we detected a total of 18,504 diverse mRNAs, including 1,675 TF mRNAs, during seed, reproductive, and vegetative periods of the life cycle at the level of the GeneChip, a value only 20% higher than the 15,563 diverse mRNAs present during seed development. Taken together, these data indicate that (i) there is a large overlap in the mRNA populations of developing seeds, from fertilization through dormancy, with those present in floral and veg-

etative organs of the mature plant, (ii) there is a small set of life cycle-shared mRNAs that is up-regulated during specific periods of seed development, and (iii) at least 18,504 diverse mRNAs are required to program the sporophytic phase of the plant life cycle, a value consistent with the 20,000 diverse mRNAs present in the AtGenExpress GeneChip database (25).

**Seeds Contain a Small Set of Specific mRNAs Enriched for Sequences Encoding Seed-Specific TFs.** By contrast with the large overlap between the mRNA populations present throughout the life cycle, we identified a small set of mRNAs that was specific at the level of the GeneChip for each developmental period, including seeds from fertilization through dormancy (Fig. 4A and Table S13 in Dataset S2). For example, we identified 289 mRNAs that were seed-specific and not detected at other times of development, including 48 seed-specific TF mRNAs (Fig. 4A and D and Table S13 in Dataset S2). qRT-PCR experiments with 36 of the seed-specific TF mRNAs indicated that they were either absent from mature plant organs (L, R, S, FBUD) or represented in one or more organs at highly reduced levels. In general, the seed-specific TF mRNAs were reduced in prevalence by 100- to 60,000-fold when present in floral and vegetative mRNA populations (Table S11 in Dataset S2). For example, five seed-specific TF mRNAs [AtbZIP72 (AT5G07160), heat shock protein AT-HSFA9 (AT5G54070), AGL33 (AT2G26320), myb-related protein (AT5G23650), and Homeobox (HB) protein (AT5G07260)] were undetectable at both the GeneChip and qRT-PCR levels in mature plant organs. AT-HSFA9 mRNA had been shown previously to be seed-specific (43). By contrast, LEC1 and LEC2 mRNAs were absent from L, R, and S at the qRT-PCR level but present at a 600-fold (LEC2 mRNA) to 3,000-fold (LEC1 mRNA) reduced levels in FBUD (Table S11 in Dataset S2). Comparison of our seed-specific TF mRNA set against relevant *Arabidopsis* gene expression datasets (e.g., AtGenExpress) validated that the seed-specific TF mRNAs uncovered here were either not detectable at other periods of the life cycle or present in one or more mature plant organ systems at reduced levels.

The seed-specific mRNAs were distributed into all major functional groups, although 25% encoded proteins that were either unclassified or had no known function (Fig. S5A and Table S14 in Dataset S2). Remarkably, the largest known functional group represented in the seed-specific mRNA population was transcription (18%), reflecting a significant enrichment (48/289) in TF mRNAs ( $P < 1 \times 10^{-4}$ , Fisher Exact Test) (Table S14 in Dataset S2), including 17 of the 23 major TF families represented on the GeneChip (compare Figs. S2B and S5B). Three TF families, ARR-B, MADS, and CCAAT, were overrepresented significantly in the seed-specific TF mRNA population ( $P < 0.02$ , Fisher Exact test) (Table S14 in Dataset S2), whereas others (e.g., Trihelix, GRAS, bHLH) had no representatives (Fig. S5B). Analysis of GO term enrichment categories encoded by the seed-specific mRNA set also showed an enrichment for sequences involved in transcriptional regulation ( $P < 0.01$ ), seed development ( $P < 0.001$ ), somatic embryogenesis ( $P < 0.01$ ), and oil storage ( $P < 0.001$ ) (Table S15 in Dataset S2). As expected, mRNAs encoding known seed-specific protein markers, such as storage proteins, lea proteins, and oleosin, were also enriched in the seed-specific mRNA population (Table S13 in Dataset S2), reflecting the overrepresentation of sequences in the protein destination and storage category ( $P < 4 \times 10^{-5}$ , Fisher Exact Test) (Fig. S5A and Table S14 in Dataset S2).

Approximately 80% of the seed-specific mRNAs, including those encoding TFs, were either stage specific (e.g., 24H, GLOB, COT, MG, or PMG) or period specific for two contiguous stages (e.g., 24H-GLOB, GLOB-COT, MG-PMG) during seed development (Fig. 4D and Fig. S5C). Most were either GLOB- or COT-stage-specific, or were present specifically during the GLOB-COT period of development (Fig. 4D and Fig. S5C), mirroring what was observed with the seed-stage-specific mRNA set that included half of the seed-specific mRNAs (Fig. 2 B and C). Together, these data



indicate that there is a small set of seed-specific mRNAs, enriched for sequences encoding TFs, that is either absent from mature plant organs or present at highly reduced levels, and that most of these mRNAs accumulate at specific times of seed development when major events required for seed formation occur (Fig. 1).

**Seed-Specific TF mRNAs Are Localized in Different Seed Compartments.**

We used chimeric seed-specific promoter-*GUS* transgenes to localize seed-specific TF gene transcriptional activity within seed compartments (i.e., embryo, endosperm, seed coat) from fertilization through dormancy (Fig. 5A and Fig. S64). Several unique transcriptional patterns were observed, including transcription in the (i) entire embryo, (ii) embryonic organs including cotyledons and axis, (iii) endosperm, and (iv) chalazal endosperm (Fig. 5A). The latter transcriptional pattern was observed for nine different seed-specific promoter-*GUS* reporter genes (Fig. S64). In general, the embryo transcriptional patterns correlated with the accumulation of corresponding seed-specific TF mRNAs late in development (COT-PMG), whereas the endosperm/seed coat transcriptional patterns correlated with seed-specific TF mRNAs that accumulated early in development (GLOB-COT) (Fig. 5A and Fig. S7).

We used the Harada-Goldberg LCM GeneChip dataset to localize the seed-specific TF mRNAs within all major seed compartments and tissues during development (*Materials and Methods*), including the

(i) embryo (embryo proper and suspensor), (ii) endosperm (peripheral, micropylar, and chalazal), and (iii) seed coat (general and chalazal) (Fig. 5B and C, Figs. S6B and S7, and Table S19 in Dataset S2). Each seed-specific TF mRNA had a unique temporal- and compartment-specific accumulation pattern (Fig. 5B and C, and Fig. S7). For example, *AtbZIP67* mRNA (AT3G44460) accumulated in the (i) embryo proper, (ii) peripheral, micropylar, and chalazal endosperm regions, and (iii) general and chalazal seed coat during the heart (HRT) to MG stages (Fig. 5B and C). By contrast, *myb* TF mRNA AT3G10590 was detected primarily in the chalazal endosperm and seed coat regions during the preglobular (PG) to MG period of development (Fig. 5B and C). In general, both the whole seed mRNA accumulation patterns and the *GUS* transgene localization profiles were congruent with both the temporal and spatial TF mRNA accumulation patterns identified by using the Harada-Goldberg LCM GeneChip dataset (Fig. 5A and C and Fig. S6). Together, these data indicate that the seed-specific TF mRNAs are localized within unique seed compartments at precise times during seed development and that transcriptional processes are primarily responsible for generating the seed-specific TF mRNA accumulation patterns.

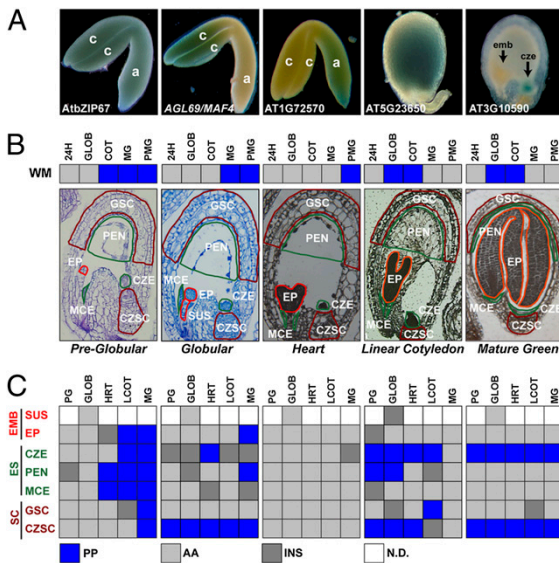
**Mutations in Genes Encoding Most Seed-Specific TFs Do Not Result in a Detectable Phenotype.**

Seven of the 48 TF mRNAs uncovered in our seed-specific TF mRNA set encoded important regulatory proteins that resulted in seed-lethal phenotypes when their corresponding genes were mutated, including *LEC1*, *LEC2*, *FUS3*, *PEI1*, *MINISEED3*, and *MEDEA* (Fig. 4D and Table S20 in Dataset S2) (35, 44–46). Mutations (47) in an additional 24 seed-specific TF genes did not result in a seed-lethal phenotype or detectably alter seed development (Fig. 4D and Table S20 in Dataset S2). Taken together, these data indicate that the seed-specific TF gene set is enriched for important known regulators of seed development, but that mutant alleles of most seed-specific TF genes do not yield a seed phenotype and, as such, the functions of most seed-specific TF genes investigated here are not yet known.

**Discussion**

We profiled *Arabidopsis* mRNA sets before, during, and after seed formation, and compared these mRNA sets to those from mature plant organ systems to uncover key transcriptional regulators of seed development. Our experiments showed that at least 16,000 mRNAs are required to program *Arabidopsis* seed development—from fertilization through dormancy. This is undoubtedly a lower limit due to (i) our use of whole seed mRNAs, (ii) the inability of the GeneChip to detect rare sequences in complex mRNA populations such as whole seeds, and (iii) the incompleteness of the Affymetrix ATH1 GeneChip, which is missing ~20% of known *Arabidopsis* genes (36). If we assume that sequences not included on the GeneChip represent a random collection of genes, then at least 19,000 diverse mRNAs are required to program seed development and make an *Arabidopsis* seed.

Most diverse seed mRNA sequences are present before fertilization, persist from zygote formation through dormancy, and are represented after seed formation in germinating seedlings and mature plant organ systems. Thus, most seed mRNAs are used in different developmental contexts throughout the plant life cycle, although significant quantitative changes occur in individual mRNA prevalences that correspond with specific seed developmental stages and/or periods of the life cycle. The reduction in the number of seed mRNAs detected during late development (MG-PMG) is probably due to mRNA turnover resulting from the general shutdown in transcriptional processes as the seed enters dormancy (48). More than a generation ago, we (49) and others (50), used RNA/cDNA hybridization experiments to investigate mRNA populations during seed development. Our conclusions from that era, using primitive technology that provided the foundation for questions being addressed currently with sophisticated genomics approaches, are in remarkable agreement with those reported here. That is, most diverse seed mRNA sequences



**Fig. 5.** Seed-specific TF gene activity in different *Arabidopsis* seed compartments, regions, and tissues. (A) Localization of GUS enzyme activity in seeds and embryos of transgenic lines carrying different seed-specific TF gene upstream regions (Fig. 4D) fused with the *GUS* reporter gene (*SI Materials and Methods*). Squares in the horizontal bars below each GUS-stained embryo or seed show the GeneChip MAS 5.0 consensus call for the seed-specific TF gene in whole-mount seeds at different developmental stages (Fig. 2). Blue and gray squares represent consensus detection calls of PP and AA, respectively (see *Materials and Methods*). (B) Bright-field photographs of *Arabidopsis* 5–7 μm paraffin seed sections at different developmental stages. Highlighted areas represent compartments, regions, and tissues captured by LCM (32). (C) Seed-specific TF mRNA localization within seeds at different stages of development. Blue, light gray, and dark gray squares indicate GeneChip MAS 5.0 consensus detection calls of PP, AA, and INS, respectively (see *Materials and Methods*). White squares indicate not determined (N.D.). a, axis; c, cotyledon; cze, chalazal endosperm; czsc, chalazal seed coat; emb, embryo; ep, embryo proper; es, endosperm; gsc, general seed coat; mce, micropylar endosperm; pen, peripheral endosperm; sc, seed coat; sus, suspensor; PG, preglobular stage seed; HRT, heart-stage seed; LCO, linear cotyledon-stage seed.

persist throughout development, are represented in mature plant organs, and stage-specific quantitative changes occur in specific seed mRNA sets (49, 50).

We identified a small set of mRNAs, significantly enriched for sequences encoding TFs, which is either specific for seeds at the qRT-PCR level or present in one or more mature plant organs at levels significantly below those of seed mRNAs that are shared with other periods of the life cycle. These seed-specific mRNAs represent <2% of the total mRNAs present during seed development (289/15,500). At least half of the seed-specific mRNAs, including the seed-specific TF mRNAs, accumulate at specific stages of seed development when major events required for seed formation occur (e.g., GLOB, COT, MG). The remaining seed-specific mRNAs accumulate within temporally contiguous periods that correspond with key seed developmental events as well (e.g., GLOB-COT). Most of the seed-stage- and seed-specific mRNAs identified here accumulate within the GLOB-COT period of seed development, correlating with the period when the majority of *Arabidopsis* embryo-defective mutants arrest in seed development (51)—a time when critical morphogenetic and biochemical events occur that are required for embryo and seed formation (2) (Fig. 1). Other mRNAs are specific for the MG-PMG period of seed development when maturation occurs. Almost all of the GLOB-COT-specific mRNAs are localized within the endosperm, whereas the MG-PMG mRNAs are represented within the embryo, although these mRNAs can also be present in other regions of the seed. The temporal and spatial mRNA accumulation patterns for the seed-specific mRNAs correlate with biological processes that occur uniquely within seeds during the plant life cycle; that is, the formation of a triploid endosperm and a maturation period when seeds accumulate high levels of food reserves, prepare for dehydration, and enter dormancy (2, 4).

We identified 48 seed-specific TF mRNAs that most likely play important roles in regulating seed development. This is a lower limit of the number of seed-specific regulators because of the stringent filtering process we used in comparing GeneChip datasets generated in this study (*Materials and Methods*). If we lower our stringency to include discordant MAS 5.0 consensus calls containing one P and consider TF genes not present on the GeneChip, then the number of seed-specific TF mRNAs could approach 100 or more, but is still a small proportion of the 1,400 diverse TF mRNAs that we detected throughout seed development. Seed TF mRNAs that are shared with other periods of the life cycle clearly play important roles in seed development; however, the seed-specific TF mRNAs uncovered here probably guide processes unique to seeds.

The functions of most seed-specific TF mRNAs uncovered here are not known; however, the seed-specific TF mRNA set is enriched for known regulators of seed development (e.g., LEC1, LEC2, L1L, FUS3, MEDEA) that were uncovered in genetic screens for embryo defective mutants, strongly suggesting that the other seed-specific TF mRNAs will play critical regulatory roles as well. These regulators have been shown to be critical for controlling events unique to seeds; that is endosperm formation and maturation (15, 52). The critical question is, of course, what role do the remaining seed-specific TF mRNAs in our dataset play in seed development? The localization patterns of most of these seed-specific TF mRNAs also suggest that they are involved in regulating either the differentiation and/or function of the endosperm early in seed development or events required for maturation in either specific embryo regions and/or the seed coat late in development. One clue as to the function of several seed-specific TF mRNAs is the observation that nine are localized in the chalazal endosperm layer during the GLOB-COT phase and correlate with the transcription of their corresponding genes (Fig. S6). This seed layer plays an important role in embryo development transferring critical nutrients from maternal to embryonic tissues (2), and it is possible that the nine chalazal-specific TF genes form a regulatory

network required for the differentiation and/or function of this seed region. Analysis of the chalazal endosperm-specific TF gene promoters shows an enrichment for a CARGCW8AT motif ( $P < 1 \times 10^{-4}$ ), which is an AGL15 (AT5G13790) TF binding site (Fig. S6.4) (53). AGL15 mRNA is present in the Harada-Goldberg LCM chalazal endosperm dataset, suggesting that AGL15 might act upstream of the chalazal endosperm-specific TF genes and play a role in activating at least one chalazal endosperm gene regulatory network.

The roles of most regulatory genes in controlling seed development and how seed gene sets are organized into regulatory networks are not well understood. The seed-specific TF genes uncovered in our study should provide an important starting point for understanding how gene activity is coordinated during seed development to make a seed. Clearly, how specific compartments of the seed are differentiated and the roles that seed-specific TF genes play in this process remain to be determined.

## Materials and Methods

**Plant Material.** Detailed information on (i) growth of *Arabidopsis* plants, (ii) stages of seed development, and (iii) characteristics of plant material are presented in *SI Materials and Methods*.

**RNA Isolation and Affymetrix GeneChip Hybridization.** Details of RNA isolation, biotinylated cRNA synthesis, and hybridization with Affymetrix *Arabidopsis* ATH1 22K GeneChips (36, 54) are presented in *SI Materials and Methods*. Two biological replicates were analyzed for each sample and processed at the same time to reduce variability. Signal intensities and detection calls [(P), (A), or (M)] were determined by using Affymetrix MAS 5.0 software default parameters (36, 54). All of the ATH1 22K GeneChip data were deposited in the Gene Expression Omnibus (GEO) as Series GSE680. Experiments were also carried out by using the first generation Affymetrix *Arabidopsis* AtGenome1 8K GeneChip (54, 55) with OV, 24H, COT, and MG seed RNAs. These data are not discussed here in detail but are also deposited in GEO as part of Series GSE680.

**Analysis of GeneChip Hybridization Data.** The consensus call for each probe set was assigned as PP, AA, or MM by combining the detection calls of both biological replicates (36, 54). Probe sets with different detection calls between biological replicates (e.g., P in replicate 1 and A in replicate 2) were assigned a consensus detection call of INS, or insufficient. We applied a stringent filter to our data by using only probe sets with a consensus call of PP (i.e., P in both biological replicates), and removing probe sets that had consensus calls of either INS or MM from all datasets used for comparative analysis. A detailed description of the process we used to analyze, filter, and compare the results of our GeneChip hybridization experiments, including hierarchical clustering using dChip 1.3 software (56) and GO term enrichment analysis, is presented in *SI Materials and Methods*.

**Real-Time Quantitative RT-PCR Validation of GeneChip Data.** qRT-PCR reactions were carried out by using the procedures described in *SI Materials and Methods*. Primer sequences used for the qRT-PCR reactions are listed in *Table S12 of Dataset S2*. qRT-PCR data were evaluated using three criteria: (i) observed vs. expected qRT-PCR product T<sub>m</sub>, (ii) technical replicate Ct value reproducibility, and (iii) observed vs. expected qRT-PCR product size. Only qRT-PCR results that satisfied all three criteria were considered reliable and used in this paper.

**Localization of TF Gene Activity in Specific Seed Compartments.** TF promoter-*GUS* transgene experiments were carried out according to procedures outlined in *SI Materials and Methods*. Localization of mRNAs to specific seed compartments were determined by using the Harada-Goldberg LCM GeneChip datasets that are deposited in GEO as Series GSE12402, GSE11262, GSE15160, GSE12403, and GSE15165 for PG, GLOB, HRT, LCOT, and MG stages of seed development, respectively.

**ACKNOWLEDGMENTS.** We thank Weimin Deng for maintaining our *Arabidopsis* plants, Zixing Fang for assistance with the dChip analysis, and Zugen Chen for outstanding advice and help with the GeneChip hybridizations. This work was supported by grants from the National Science Foundation Plant Genome Program and Ceres (to R.B.G. and J.J.H.), US Public Health Service National Research Service Award GM07104 (K.F.H.), and National Institutes of Health Training Grant in Genomic Analysis and Interpretation T32HG002536 (B.H.L.).

1. Goldberg RB, de Paiva G, Yadegari R (1994) Plant embryogenesis: Zygote to seed. *Science* 266:605–614.
2. Raghavan V (2006) *Double Fertilization: Embryo and Endosperm Development in Flowering Plants* (Springer, Berlin).
3. Goldberg RB, Barker SJ, Perez-Grau L (1989) Regulation of gene expression during plant embryogenesis. *Cell* 56:149–160.
4. Harada JJ (1997) *Seed maturation and the control of germination. Cellular and Molecular Biology of Seed Development, Advances in Cellular and Molecular Biology of Plants*, eds Larkins BA, Vasil IK (Kluwer Academic Publishers, Dordrecht, the Netherlands), Vol 4, pp 545–592.
5. Baud S, Boutin J-P, Miquel M, Lepiniec L, Rochat C (2002) An integrated overview of seed development in *Arabidopsis thaliana* ecotype WS. *Plant Physiol Biochem* 40: 151–160.
6. Olsen OA (2004) Nuclear endosperm development in cereals and *Arabidopsis thaliana*. *Plant Cell* 16(Suppl):S214–S227.
7. Laux T, Würschum T, Breuninger H (2004) Genetic regulation of embryonic pattern formation. *Plant Cell* 16(Suppl):S190–S202.
8. Haughn G, Chaudhury A (2005) Genetic analysis of seed coat development in *Arabidopsis*. *Trends Plant Sci* 10:472–477.
9. Jenik PD, Gillmor CS, Lukowitz W (2007) Embryonic patterning in *Arabidopsis thaliana*. *Annu Rev Cell Dev Biol* 23:207–236.
10. Devic M (2008) The importance of being essential: EMBRYO-DEFECTIVE genes in *Arabidopsis*. *C R Biol* 331:726–736.
11. Meinke D, Muralla R, Sweeney C, Dickerman A (2008) Identifying essential genes in *Arabidopsis thaliana*. *Trends Plant Sci* 13:483–491.
12. Gehring M, Choi Y, Fischer RL (2004) Imprinting and seed development. *Plant Cell* 16 (Suppl):S203–S213.
13. Huh JH, Bauer MJ, Hsieh TF, Fischer RL (2008) Cellular programming of plant gene imprinting. *Cell* 132:735–744.
14. Breuninger H, Rikirsch E, Hermann M, Ueda M, Laux T (2008) Differential expression of WOX genes mediates apical-basal axis formation in the *Arabidopsis* embryo. *Dev Cell* 14:867–876.
15. Braybrook SA, Harada JJ (2008) LECs go crazy in embryo development. *Trends Plant Sci* 13:624–630.
16. Braybrook SA, et al. (2006) Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis. *Proc Natl Acad Sci USA* 103:3468–3473.
17. Chandrasekharan MB, Bishop KJ, Hall TC (2003) Module-specific regulation of the beta-phaseolin promoter during embryogenesis. *Plant J* 33:853–866.
18. Kroj T, Savino G, Valon C, Giraudat J, Parcy F (2003) Regulation of storage protein gene expression in *Arabidopsis*. *Development* 130:6065–6073.
19. Mönke G, et al. (2004) Seed-specific transcription factors ABI3 and FUS3: Molecular interaction with DNA. *Planta* 219:158–166.
20. Yamamoto A, et al. (2009) *Arabidopsis* NF-YB subunits LEC1 and LEC1-LIKE activate transcription by interacting with seed-specific ABRE-binding factors. *Plant J* 58: 843–856.
21. Kawashima T, et al. (2009) Identification of cis-regulatory sequences that activate transcription in the suspensor of plant embryos. *Proc Natl Acad Sci USA* 106:3627–3632.
22. Vasil V, et al. (1995) Overlap of Viviparous1 (VP1) and abscisic acid response elements in the Em promoter: G-box elements are sufficient but not necessary for VP1 transactivation. *Plant Cell* 7:1511–1518.
23. Benedito VA, et al. (2008) A gene expression atlas of the model legume *Medicago truncatula*. *Plant J* 55:504–513.
24. Ma L, et al. (2005) Organ-specific expression of *Arabidopsis* genome during development. *Plant Physiol* 138:80–91.
25. Schmid M, et al. (2005) A gene expression map of *Arabidopsis thaliana* development. *Nat Genet* 37:501–506.
26. Jiao Y, et al. (2009) A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. *Nat Genet* 41:258–263.
27. Girke T, et al. (2000) Microarray analysis of developing *Arabidopsis* seeds. *Plant Physiol* 124:1570–1581.
28. Ruuska SA, Girke T, Benning C, Ohlrogge JB (2002) Contrapuntal networks of gene expression during *Arabidopsis* seed filling. *Plant Cell* 14:1191–1206.
29. Spencer MW, Casson SA, Lindsey K (2007) Transcriptional profiling of the *Arabidopsis* embryo. *Plant Physiol* 143:924–940.
30. Verdier J, et al. (2008) Gene expression profiling of *M. truncatula* transcription factors identifies putative regulators of grain legume seed filling. *Plant Mol Biol* 67:567–580.
31. Day RC, Herridge RP, Ambrose BA, Macknight RC (2008) Transcriptome analysis of proliferating *Arabidopsis* endosperm reveals biological implications for the control of syncytial division, cytokinin signaling, and gene expression regulation. *Plant Physiol* 148:1964–1984.
32. Kerk NM, Ceserani T, Tausta SL, Sussex IM, Nelson TM (2003) Laser capture microdissection of cells from plant tissues. *Plant Physiol* 132:27–35.
33. Davidson EH, Levine MS (2008) Properties of developmental gene regulatory networks. *Proc Natl Acad Sci USA* 105:20063–20066.
34. Mayer K, et al. (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* 402:769–777.
35. Luo M, Dennis ES, Berger F, Peacock WJ, Chaudhury A (2005) MINISEED3 (MINI3), a WRKY family gene, and HAIKU2 (IKU2), a leucine-rich repeat (LRR) KINASE gene, are regulators of seed size in *Arabidopsis*. *Proc Natl Acad Sci USA* 102:17531–17536.
36. 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.
37. Lockhart DJ, et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 14:1675–1680.
38. Windsor JB, Symonds VV, Mendenhall J, Lloyd AM (2000) *Arabidopsis* seed coat development: morphological differentiation of the outer integument. *Plant J* 22: 483–493.
39. Penfield S, Meissner RC, Shoue DA, Carpita NC, Bevan MW (2001) MYB61 is required for mucilage deposition and extrusion in the *Arabidopsis* seed coat. *Plant Cell* 13: 2777–2791.
40. Jensen MK, et al. (2008) Transcriptional regulation by an NAC (NAM-ATAF1,2-CUC2) transcription factor attenuates ABA signalling for efficient basal defence towards *Blumeria graminis* f. sp. *hordei* in *Arabidopsis*. *Plant J* 56:867–880.
41. Alonso R, et al. (2009) A pivotal role of the basic leucine zipper transcription factor bZIP53 in the regulation of *Arabidopsis* seed maturation gene expression based on heterodimerization and protein complex formation. *Plant Cell* 21:1747–1761.
42. Kwon CS, Chen C, Wagner D (2005) WUSCHEL is a primary target for transcriptional regulation by SPLAYED in dynamic control of stem cell fate in *Arabidopsis*. *Genes Dev* 19:992–1003.
43. Kotak S, Vierling E, Bäumlein H, von Koskull-Döring P (2007) A novel transcriptional cascade regulating expression of heat stress proteins during seed development of *Arabidopsis*. *Plant Cell* 19:182–195.
44. Harada JJ (2001) Role of *Arabidopsis* LEAFY COTYLEDON genes in seed development. *J Plant Physiol* 158:405–409.
45. Li Z, Thomas TL (1998) PEI1, an embryo-specific zinc finger protein gene required for heart-stage embryo formation in *Arabidopsis*. *Plant Cell* 10:383–398.
46. Sørensen MB, Chaudhury AM, Robert H, Bancharrel E, Berger F (2001) Polycomb group genes control pattern formation in plant seed. *Curr Biol* 11:277–281.
47. Alonso JM, et al. (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* 301:653–657.
48. Walling L, Drews GN, Goldberg RB (1986) Transcriptional and post-transcriptional regulation of soybean seed protein mRNA levels. *Proc Natl Acad Sci USA* 83:2123–2127.
49. Goldberg RB, Hoschek G, Tam SH, Ditta GS, Breidenbach RW (1981) Abundance, diversity, and regulation of mRNA sequence sets in soybean embryogenesis. *Dev Biol* 83:201–217.
50. Galau GA, Dure L, 3rd (1981) Developmental biochemistry of cottonseed embryogenesis and germination: Changing messenger ribonucleic acid populations as shown by reciprocal heterologous complementary deoxyribonucleic acid–messenger ribonucleic acid hybridization. *Biochemistry* 20:4169–4178.
51. McElver J, et al. (2001) Insertional mutagenesis of genes required for seed development in *Arabidopsis thaliana*. *Genetics* 159:1751–1763.
52. Kiyosue T, et al. (1999) Control of fertilization-independent endosperm development by the MEDEA polycomb gene in *Arabidopsis*. *Proc Natl Acad Sci USA* 96:4186–4191.
53. Tang W, Perry SE (2003) Binding site selection for the plant MADS domain protein AGL15: an in vitro and in vivo study. *J Biol Chem* 278:28154–28159.
54. Hennig L, Menges M, Murray JA, Grissem W (2003) *Arabidopsis* transcript profiling on Affymetrix GeneChip arrays. *Plant Mol Biol* 53:457–465.
55. Zhu T, Wang X (2000) Large-scale profiling of the *Arabidopsis* transcriptome. *Plant Physiol* 124:1472–1476.
56. Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biol*, 2: research0032.0031–0032.0011.
57. Bowman JL, Mansfield SG (1993) Embryogenesis: Introduction. *Arabidopsis: An Atlas of Morphology and Development*, ed Bowman JL (Springer, New York), pp 351–361.
58. Lotan T, et al. (1998) *Arabidopsis* LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells. *Cell* 93:1195–1205.

# Supporting Information

Le et al. 10.1073/pnas.1003530107

## SI Materials and Methods

**Growth of Plants.** *Arabidopsis thaliana* (L.) Heyn ecotype Wassilewskija (Ws-0) plants were grown in 4 × 4-inch pots at 20 °C with 50–70% relative humidity under continuous fluorescent light in a Conviron chamber at University of California, Davis. With the exception of roots and seedlings (see below), plant material was harvested from 39- to 45-day-old plants.

**Plant Material Used For GeneChip Experiments.** The seed stages used for GeneChip analysis are shown schematically in Fig. 1. Each developmental stage was represented by two biological replicates that were harvested independently. Unfertilized ovules (OV) contained mature embryo sacs that were synchronized developmentally by emasculating of stage 13 flowers (1) 24 h before collection and then hand dissected from pistils. Twenty-four hour post-fertilization seeds (24H) were collected from siliques 24 h after hand pollination of emasculated flowers and contained zygotes. Several unpollinated and pollinated pistils were left on plants to determine the extent of cross pollination and the efficiency of pollination. Approximately 1.7% cross pollination occurred for the ovule samples, whereas ≈0.9% cross pollination occurred for the 24H fertilized seeds. The efficiency of pollination was ≈85% for the 24H seed samples. Three to four days after pollination (DAP) seeds (GLOB), seven to eight DAP seeds (COT), 13–14 DAP seeds (MG), and 18–19 DAP seeds (PMG) contained primarily (i) pre-globular to globular stage white embryos, (ii) linear cotyledon (LCOT) to bent cotyledon stage green embryos, (iii) mature-green embryos that filled the seed, and (iv) postmature green embryos that were beginning to yellow, respectively (Figs. 1 and 2A). To avoid overlap between MG and PMG stages, three to five siliques separated siliques from which the MG and PMG seeds were harvested. To collect seedlings (SDLG), seeds were surface-sterilized, placed on plates (250 seeds per 100-mm plate) containing GM media (1× Murashige and Skoog (MS) salts, 1% sucrose, 1× B5 vitamins, 0.8% BactoAgar), stratified at 4 °C for 3 days, and allowed to germinate in a Percival growth chamber at 22 °C by using a 16-h day, 8-h night cycle. Seedlings were harvested 3 days after germination.

Wild-type and *lec1-1* leaves, stems, and floral buds were collected from plants grown similarly as those used for the ovule and seed collections. *lec1-1* homozygous plants were obtained by rescuing and germinating immature *lec1-1* homozygous seeds in culture (2). Harvested leaves contained young and old basal rosette leaves that were intact and green. No cauline, damaged, or discolored leaves were harvested. Stems were devoid of cauline leaves, floral buds, and developing siliques. Stems were cut into pieces by using a razor blade and frozen immediately in liquid nitrogen. Floral buds included the entire floral meristem containing stage 1–12 closed buds from primary and secondary inflorescences (1). To collect roots, wild-type seeds were surface-sterilized and stratified at 4 °C for 6 days. Approximately 200 seeds were germinated in liquid culture containing Gamborg's B-5 medium and grown with constant shaking (100 rpm) in a Percival growth chamber at 22 °C, using a 16-h day, 8-h night cycle. For *lec1-1* root collections, by contrast, *lec1-1* immature seeds were harvested from siliques (2) and sowed directly in liquid media with no stratification. Both wild-type and *lec1-1* roots were collected after 3 weeks in liquid culture.

We used wild-type and *lec1-1* leaves, stems, roots, and floral buds as biological replicates because (i) mutations in the *LEC1* gene (e.g., *lec1-1*) only affect seed development (2), (ii) *LEC1* mRNA is either undetectable using quantitative real-time reverse transcription PCR (qRT-PCR) (leaves, roots, and stems; Table S11 in Dataset S2) or is many thousand-fold less prevalent compared with

seeds (floral buds; Table S11 in Dataset S2), and (iii) comparisons of GeneChip data between wild-type and *lec1-1* leaf, root, stem, and floral bud biological replicates had correlation coefficients of 0.93–0.99 (Fig. S1).

**RNA Isolation and Affymetrix GeneChip Hybridization.** Total RNA was isolated from seeds and plant organs by using a hot borate procedure (3). Biotinylated cRNAs were synthesized from poly(A) RNAs according to the Affymetrix technical manual with modifications. Briefly, mRNA was isolated from total RNA [8 μg (ovules) or 15 μg (seeds, seedlings, organ systems)] by one round of poly(A) RNA selection by using oligo(dT) Dynabeads according to the manufacturer's protocol (DynaL Biotech). Poly(A) RNA was converted to double-stranded cDNA by using an Affymetrix T7-oligo (dT)<sub>24</sub> primer with the SuperScript Choice System according to the manufacturer's instructions (Invitrogen). Biotinylated cRNAs were synthesized by in vitro transcription by using T7 RNA polymerase (Enzo Life Sciences), purified, and then fragmented chemically according to the Affymetrix technical manual. Biotinylated cRNAs (15 μg) were hybridized with Affymetrix *Arabidopsis* ATH1 22K GeneChips (4) in the Affymetrix GeneChip Hybridization Oven at 45 °C for 16 h according to the standard array protocol in the Affymetrix technical manual. The EukGe-WS2v4 protocol was used for washing and staining arrays by using an Affymetrix Fluidics Station. Hybridized GeneChips were scanned by using a Hewlett-Packard GeneArray Scanner, and the probe set images were converted to signal intensities and calls [present (P), absent (A), or marginal (M)] by using Affymetrix Microarray Suite 5.0 (MAS 5.0) default parameters. The MAS 5.0 data were further analyzed by using dChip 1.3 (5), Microsoft Excel, and Microsoft Access software. GeneChips representing biological replicates were hybridized, washed, and scanned at the same time at the UCLA Microarray Core Facility (<http://microarray.genetics.ucla.edu>) to minimize variability. All original image files (.cel) and tab-delimited text files (.txt) containing MAS 5.0-analyzed data were deposited in the Gene Expression Omnibus (GEO) database ([www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo)) as series GSE680. We also created an interactive web site (<http://estdb.biology.ucla.edu/genechip>) that allows the expression data for any *Arabidopsis* gene investigated in our GeneChip experiments to be downloaded and plotted with respect to the developmental stages studied here.

An earlier set of experiments was carried out by using the first-generation Affymetrix *Arabidopsis* AtGenome1 8k GeneChip (6) for OV, 24H, COT, and MG RNAs (Fig. 1). These data were analyzed by using Affymetrix MAS 4.0 software and were deposited into the GEO database as part of series GSE680 (.cel and .txt files).

**Analysis of GeneChip Hybridization Data.** Signal intensities (relative mRNA prevalences) and signal detection calls (P, A, or M) were generated by using MAS 5.0 and imported into Microsoft Excel and Microsoft Access for further analysis. For comparative purposes, GeneChip data were scaled globally to a target intensity of 500 for all probe sets on the chip by using MAS 5.0 default parameters. Each probe set was manually assigned a consensus detection call in Microsoft Excel based on the MAS 5.0 detection calls of both biological replicates of an RNA sample. Probe sets with signal detection calls of P, A, or M in both biological replicates were assigned consensus detection calls of PP, AA, or MM, respectively. In general, the detection calls for biological replicates agreed with one another ≈91% of the time, on average, with a range of 88–93%. By contrast, probe sets with different, or discordant, detection calls for the two biological replicates (e.g., P and A; P and M) were assigned a consensus detection call of Insufficient (INS). On average, ≈9% of the probe



sets were assigned a consensus detection call of INS for a given pair of biological replicates, with a range of 7.4–11.7%. These percentages are in agreement with the discordance values (i.e., INS) reported by others using the Affymetrix *Arabidopsis* ATH1 22K GeneChip (7), and by Affymetrix in their tests of GeneChip reproducibility and detection call discordance ([www.affymetrix.com/support/technical/technotes/manufacturing\\_quality\\_technote.pdf](http://www.affymetrix.com/support/technical/technotes/manufacturing_quality_technote.pdf)). Our analysis of probe sets with INS detection calls showed that the majority of these probe sets had (i) signal intensities close to, or below, the GeneChip detection limit (Fig. S3), (ii) were distributed randomly across the GeneChip, and (iii) were generated independently of the hybridized RNA sample (i.e., hybridization of the same RNA sample with multiple GeneChips generated distinct probe sets with INS calls).

**Filtering of GeneChip Data For PP Consensus Calls.** Because of the uncertainty in INS calls, only probe sets with detection calls of PP (i.e., P in both biological replicates) were considered to represent a mRNA present in any given developmental stage. Microsoft Excel was used to remove probe sets with calls of INS or MM to compare gene activity between different developmental stages.

**Filtering of seed development datasets.** For the analysis of gene activity before, during, and after seed development (Fig. 2 and Dataset S1), we removed 8,510 probe sets with consensus detection calls of either INS or MM in at least one developmental stage from all sample datasets. An additional 4,164 probe sets with consensus detection calls of AA across all developmental stages were also removed, leaving 10,062 probe sets passing all filters (44% of the ATH1 GeneChip probe sets). These probe sets were used for the analysis of results presented in Figs. 2 and 3.

**Filtering of organ system datasets.** For the comparative analysis of vegetative organ and floral bud GeneChip data (Fig. S4), we removed 5,140 probe sets with detection calls of MM or INS from all sample datasets and an additional 3,831 probe sets with consensus detection calls of AA, leaving 13,775 probe sets (61% of the ATH1 GeneChip probe sets). These probe sets were used for the analysis of data presented in Fig. S4.

**Filtering of seed, reproductive, and vegetative development datasets.** For our analysis of gene activity throughout the *Arabidopsis* life cycle (Fig. 4 and Dataset S2), we removed 10,359 probe sets with a consensus detection call of MM or INS in at least one developmental stage, and an additional 2,748 probe sets with a consensus detection call of AA across all developmental stages. The remaining 9,611 probe sets (42% of the ATH1 GeneChip probe sets) with a detection call of PP in at least one developmental stage were used for the data presented in Fig. 4 and Dataset S2.

**Identifying developmental stage-specific genes from filtered datasets.** To identify life cycle-stage-specific genes (e.g., seed-specific transcription factor genes), we grouped biological samples into three groups: (i) reproductive development—consisting of floral buds and OV, (ii) seed development—consisting of 24H, GLOB, COT, MG, and PMG seed stages, and (iii) vegetative development—consisting of SDLG, leaves, roots, and stems (Fig. 4 and Dataset S2). For each group, we designated a gene as specific if the probe set had (i) a consensus call of PP in at least one developmental stage within the group and (ii) a call of AA across all samples in the other two groups. For example, seed-specific genes were defined as having consensus calls of PP in at least one seed stage (24H, GLOB, COT, MG, or PMG), and had consensus calls of AA in all other developmental stages (i.e., OV, floral bud, leaf, root, and stem).

**Identifying seed- and organ-specific genes from filtered datasets.** To identify seed-stage or organ-specific gene sets (e.g., COT-specific, seedling-specific) (Figs. 2 and 4), we used Microsoft Excel to filter probe sets with a detection call of PP in one sample and AA across all other samples. Likewise, to identify multiple-stage-specific gene sets (e.g., COT plus PMG specific), we filtered for probe sets with a detection call of PP in two or more stages and AA in the remaining stages (Fig. 2 and Dataset S1).

**Identifying shared genes from filtered datasets.** To identify genes shared by developmental stages being analyzed (e.g., seed development, all stages of the life cycle) (Figs. 2 and 4 and Datasets S1 and S2), we filtered for probe sets with a detection call of PP in all biological samples being compared (e.g., seed development: OV, 24H, GLOB, COT, MG, PMG, SDLG). Shared gene sets were then analyzed further by using hierarchical clustering analysis (see below).

**Hierarchical Clustering.** Unsupervised hierarchical clustering analysis was performed by using dChip 1.3 software (5) on probe sets that had consensus detection calls of PP across all developmental stages being compared (i.e., shared gene sets). Hierarchical clustering was carried out with the 6,937 probe sets that had PP consensus detection calls before, during, and after seed development (Fig. 2) and the 8,084 probe sets that had PP consensus calls at all stages of the life cycle (Fig. 4). In some analyses both the GeneChip samples and probe sets (mRNAs) were clustered independently (e.g., Fig. 3A). In other analyses (e.g., Fig. 4C), only probe set data were clustered and samples orders were fixed manually. Shared probe sets were rank ordered by their signal intensity standard deviation values across all developmental stages being compared. Only the top 2,000 most-varying probe sets were selected for hierarchical clustering (Table S7 in Dataset S1 and Table S17 in Dataset S2). Before clustering, the signal intensities of a probe set across all samples were standardized in dChip 1.3 to have a mean of zero and a standard deviation of one. Hierarchical distance measurements between genes were defined as 1-correlation coefficient of standardized intensity values (5). To identify genes that were up-regulated at least 2-fold within each cluster (Table S7 in Dataset S1 and Table S17 in Dataset S2), we used the Compare sample function in dChip 1.3 with the following parameters:  $E/B > 2$  or  $E/B < 2$ ,  $|E-B| > 100$ ; Student's *t* test  $P < 0.05$ ; and 500 permutations to obtain a false discovery rate (FDR), where B and E represent baseline and experimental groups, respectively (5). The FDRs for most clusters shown in Figs. 3 and 4 were  $< 5\%$ , with the exception of clusters III and IV in Fig. 3 that had FDRs  $< 10\%$ .

**Gene Ontology (GO) Term Enrichment Analysis.** Gene lists were analyzed for GO term enrichment by using the Biomap tool available at the VirtualPlant web site (<http://virtualplant.bio.nyu.edu/cgi-bin/vpweb/virtualplant.cgi>) and the hypergeometric distribution function to calculate the term enrichment *P* value. Only GO terms with a  $P < 0.01$  are listed in Tables S6 and S8 of Dataset S1 and S14 and S18 of Dataset S2, respectively.

**GeneChip Sensitivity Experiments.** cDNA plasmids (1  $\mu$ g) representing the *LEC1* gene (AT1G21970; ref. 2) and the *RAP2.1* gene (AT1G46768; ref. 8) were linearized, and biotinylated cRNAs were synthesized by using T7 RNA polymerase according to the manufacturer's instructions (Enzo Life Sciences). *LEC1* and *RAP2.1* biotinylated cRNAs were treated with RNase-free DNase I (Promega) to remove DNA, and the cRNAs were purified by using Qiagen RNeasy minicolumns. Different amounts of biotinylated *LEC1* and *RAP2.1* cRNAs were mixed with 15  $\mu$ g of biotinylated *lec1-1* seed cRNA that was synthesized from a pool of 24H, COT, MG, and PMG *lec1-1* poly(A) seed mRNAs in ratios ranging from 1:100 to 1:1,000,000 of the total biotinylated *lec1-1* cRNA mass. We used *LEC1* and *RAP2.1* cRNAs for our GeneChip sensitivity experiments because (i) *LEC1* is deleted in *lec1-1* plants (2) and (ii) *RAP2.1* is not active in *lec1-1* seeds at the level of the GeneChip (GEO Series GSE 1051). Using this strategy, we could simulate the exact conditions under which all of our GeneChip hybridization experiments were carried out. Each spiked biotinylated cRNA population (e.g., 1:100, 1:1,000, etc.) was hybridized with replicate *Arabidopsis* ATH1 22K GeneChips as outlined above in the section on *Affymetrix*



**GeneChip Hybridization.** The intensity values of LEC1 and RAP2.1 probe sets in each spiked biotinylated cRNA population were generated by using MAS 5.0 software and then graphed relative to their fractions of the *lec1-1* RNA mass (Fig. S3).

**Real-Time Quantitative RT-PCR Validation of GeneChip Data.** Total RNA (5 µg) was treated with RNase-free DNase I according to manufacturer's protocol (Ambion), extracted with a 1:1 mixture of phenol and chloroform (24:1 chloroform-isoamyl alcohol; ref. 9), and then precipitated by using 2 µL of Pellet Paint (Novagen) as carrier. RNA concentrations were determined by using a Nanodrop Spectrophotometer (NanoDrop Technologies). First-strand cDNA was synthesized from the DNase-treated RNA (1 µg) by using Invitrogen SuperScript II reverse transcriptase (RT) following manufacturer's instructions (+RT reaction). A parallel control reaction was carried out without RT (-RT reaction) to monitor for genomic DNA contamination in the RNA sample. +RT and -RT reaction solutions were diluted 1:10 with TE buffer [10 mM Tris, 1 mM EDTA (pH 8.0)], and 1 µL of each reaction solution was added to 12.5 µL of 2× SYBR green supermix (Bio-Rad) containing the relevant primer pairs (Table S12 in Dataset S2) according to the manufacturer's instructions. A Bio-Rad iCycler was used to monitor the real-time quantitative reverse transcription PCR (qRT-PCR) for 50 cycles.

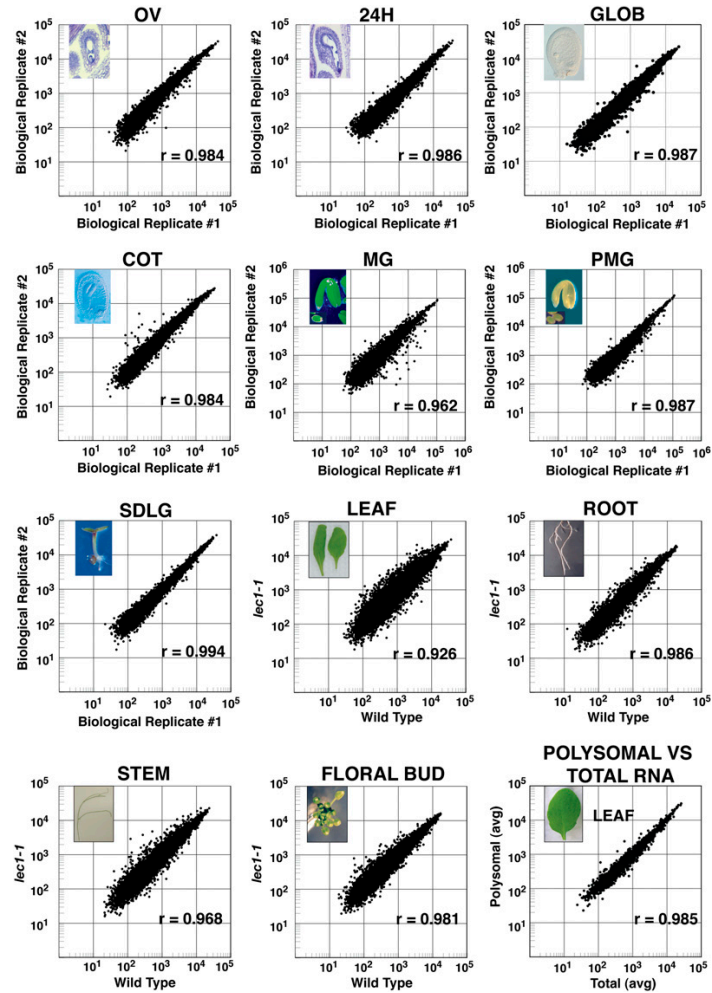
Primer pairs for qRT-PCR validation experiments (Table S12 in Dataset S2) were selected from the 11 sense and anti-sense 25-mer Affymetrix *Arabidopsis* ATH1 22K GeneChip probe set sequences that were the target of each mRNA being validated. That is, the same target sequences on the GeneChip that hybridized with the mRNA were used to generate primers for the qRT-PCRs. Specific primer pairs were selected by using Beacon Designer 2.07 software (Premier Biosoft International) and were synthesized by Invitrogen. A standard calibration curve ( $10^{-7}$  to 1 ng) was constructed by using 10-fold serial dilutions of *Phaseolus coccineus* LEC1-like plasmid DNA (10) for each set of qRT-PCRs to correlate threshold values (Ct) with mass amounts of target cDNA sequence. Under our qRT-PCR conditions 0.1 fg of target DNA sequence had a qRT-PCR Ct of 34. Duplicate technical reactions were carried out for each biological replicate, making a total of four qRT-PCR Ct values per RNA sample (i.e., two technical replicates per each biological replicate).

**Seed-Specific Transcription Factor Promoter-GUS Reporter Gene Localization Experiments.** Seed-specific transcription factor (TF) gene upstream sequences (Fig. 5 and Fig. S6) were amplified by using iProof DNA polymerase, following the manufacturer's protocol (Bio-Rad). Up to 4 kb of DNA sequence was amplified depending on the distance between the target TF gene translation start site and the 3' end of the next upstream gene. Amplified DNA fragments were cloned directionally into the Gateway pENTR/D-TOPO vector (Invitrogen) and then recombined into a destination vector that contained the *Escherichia coli* β-glucuronidase (*GUS*) gene (pBGWFS7; ref. 11).

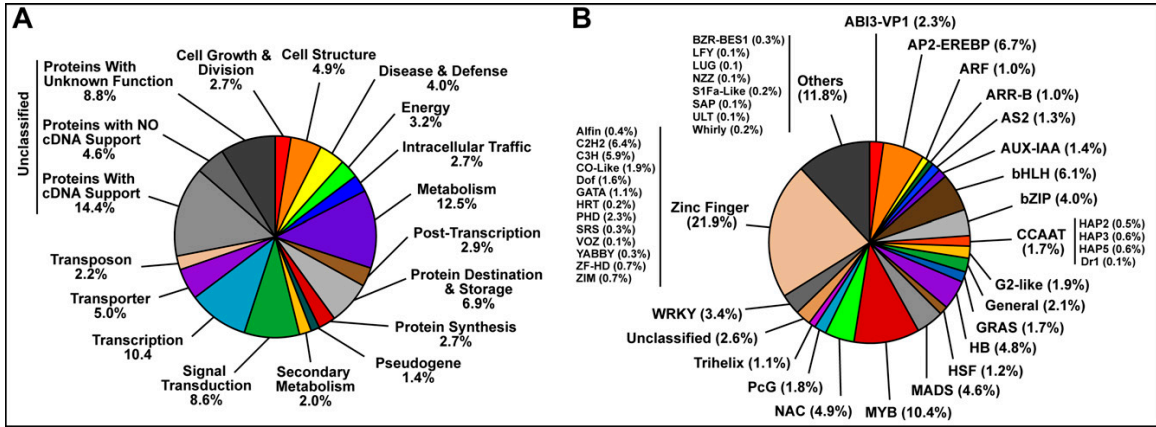
**Transformation into Arabidopsis plants.** Expression vectors containing the chimeric TF promoter-GUS genes were transferred into *Agrobacterium tumefaciens* strain LBA4404 by electroporation. *Arabidopsis* plants (Columbia-0) were grown at a density of six to eight plants per pot in the UCLA Plant Growth Center by using a 16-h day, 8-h night cycle at 22 °C for 3–4 weeks, and were transformed with *Agrobacterium* using the floral dip method (12). In brief, the main inflorescence of each plant was removed when the primary stem was ≈10 cm long. Plants were dipped twice (2 and 7 days after inflorescence removal) in a solution of 10% sucrose, 0.5× MS salts, and 0.05% silwet L-77 solution containing LBA4404 *Agrobacterium* cells that were grown for 48 h. Transformed T1 plants were selected for glufosinate resistance (i.e., Basta) by germinating seeds in soil that was moistened with water containing 1 part per 3,500 of glufosinate. Seeds from several independently transformed T1 lines were harvested, and at least four independent T2 lines per construct were used for GUS localization analysis (Fig. 5).

**GUS localization in developing seeds.** Seeds and embryos were dissected from siliques at different times of development (Fig. 5 and Fig. S6) and incubated in a solution containing GUS substrate [50 mM phosphate buffer (pH 7), 0.1% Triton X-100, 10 mM EDTA, 0.5 mM potassium ferricyanide, 0.5 mM potassium ferrocyanide, 1 mM 5-bromo-4-chloro-3-indolylglucuronide (X-glc) (13)] for 24 h at 37 °C. GUS localization patterns were observed by using either a dissecting microscope (Olympus SZH) or a compound microscope (Leica DM 5000B). Pictures of seed and embryo GUS-localization patterns were taken with a Leica DC500 digital camera and generated by using Leica Firecam 1.2 software.

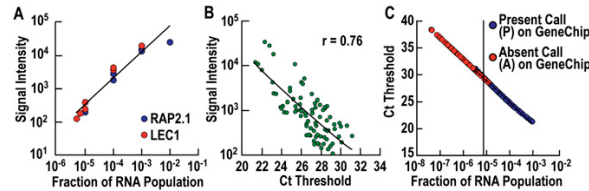
1. Smyth DR, Bowman JL, Meyerowitz EM (1990) Early flower development in *Arabidopsis*. *Plant Cell* 2:755–767.
2. Lotan T, et al. (1998) *Arabidopsis* LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells. *Cell* 93:1195–1205.
3. Wan CY, Wilkins TA (1994) A modified hot borate method significantly enhances the yield of high-quality RNA from cotton (*Gossypium hirsutum* L.). *Anal Biochem* 223:7–12.
4. 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.
5. Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* 2:research0032.0031–0032-0011.
6. Zhu T, Wang X (2000) Large-scale profiling of the *Arabidopsis* transcriptome. *Plant Physiol* 124:1472–1476.
7. Hennig L, Menges M, Murray JA, Grissem W (2003) *Arabidopsis* transcript profiling on Affymetrix GeneChip arrays. *Plant Mol Biol* 53:457–465.
8. Okamoto JK, Caster B, Villarreal R, Van Montagu M, Jofuku KD (1997) The AP2 domain of APETALA2 defines a large new family of DNA binding proteins in *Arabidopsis*. *Proc Natl Acad Sci USA* 94:7076–7081.
9. Sevag MG, Lackman DB, Smolens J (1938) The isolation of the components of Streptococcal nucleoproteins in serologically active form. *J Biol Chem* 124:425–436.
10. Kwong RW, et al. (2003) LEAFY COTYLEDON1-LIKE defines a class of regulators essential for embryo development. *Plant Cell* 15:5–18.
11. Karimi M, Inzé D, Depicker A (2002) GATEWAY vectors for *Agrobacterium*-mediated plant transformation. *Trends Plant Sci* 7:193–195.
12. Clough SJ, Bent AF (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J* 16:735–743.
13. Jefferson RA, Kavanagh TA, Bevan MW (1987) GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO J* 6:3901–3907.
14. Mayer K, et al. (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* 402:769–777.
15. Guo A, et al. (2005) DATF: A database of *Arabidopsis* transcription factors. *Bioinformatics* 21:2568–2569.
16. Davuluri RV, et al. (2003) AGRIS: *Arabidopsis* gene regulatory information server, an information resource of *Arabidopsis* cis-regulatory elements and transcription factors. *BMC Bioinformatics* 4:25.
17. Riaño-Pachón DM, Ruzicic S, Dreyer I, Mueller-Roeber B (2007) PlnTFDB: An integrative plant transcription factor database. *BMC Bioinformatics* 8:42.
18. Iida K, et al. (2005) RARTF: Database and tools for complete sets of *Arabidopsis* transcription factors. *DNA Res* 12:247–256.
19. O'Connor TR, Dyreson C, Wyrick JJ (2005) Athena: A resource for rapid visualization and systematic analysis of *Arabidopsis* promoter sequences. *Bioinformatics* 21:4411–4413.



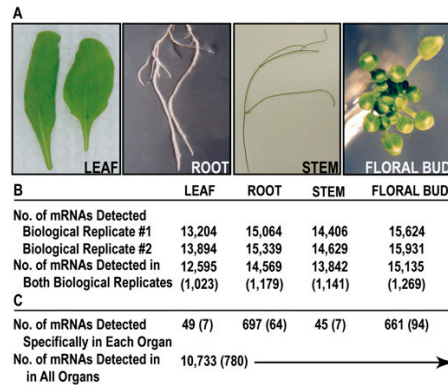
**Fig. S1.** Correlation of GeneChip data between biological replicates and polysomal and total RNA populations. The MAS 5.0 GeneChip signal intensities ([SI Materials and Methods](#)) from either biological replicates or polysomal and total RNAs were plotted on X-Y scatter plots. Only probe sets with consensus detection calls of PP were plotted ([SI Materials and Methods](#)). The Pearson correlation coefficient ( $r$ ) for each pair of RNA samples was calculated using Microsoft Excel. Wild type and *lec1-1* samples were used to obtain scatter plots and Pearson correlation coefficients for the leaf, root, stem, and floral bud samples ([SI Materials and Methods](#)).



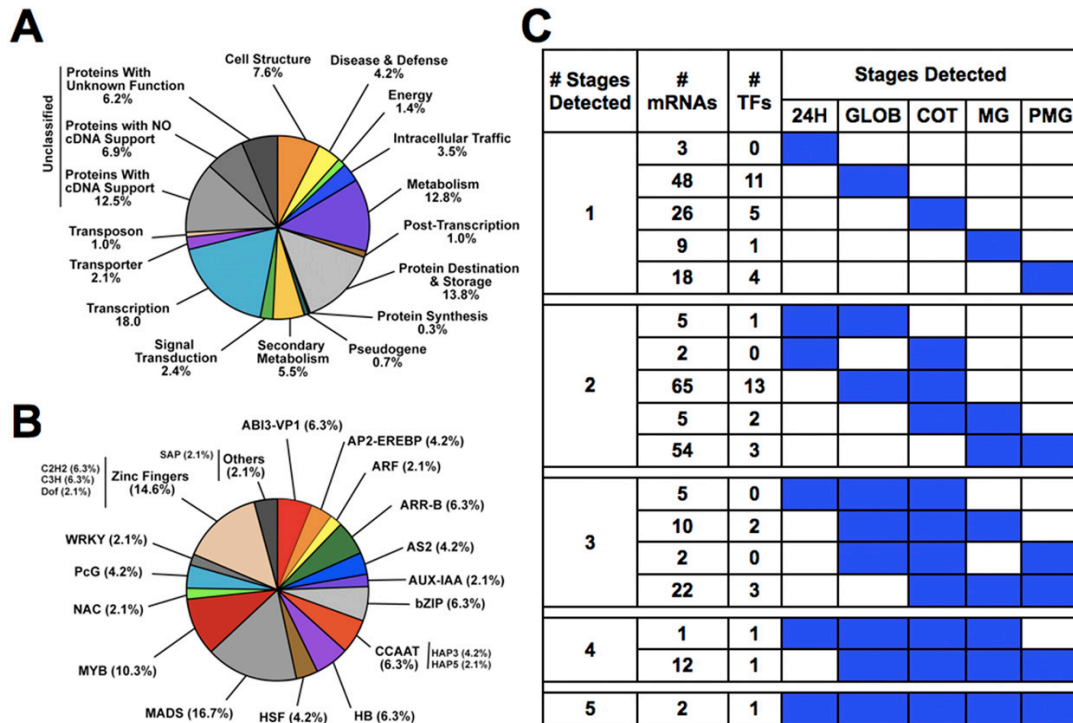
**Fig. S2.** Functional classification of probe sets on the Affymetrix *Arabidopsis* ATH1 22K GeneChip. (A) Probe sets on the array were classified manually into functional groups by using annotations from TIGR [The Institute for Genomic Research] (<http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=arab>), GO [Gene Ontology Consortium], ([www.geneontology.org](http://www.geneontology.org)), and TAIR [The *Arabidopsis* Information Resource], (<http://arabidopsis.org>) following the MIPS [Munich Information Center for Protein Sequences], (<http://mips.helmholtz-muenchen.de/plant/athal/>) classification scheme with modifications (14). Number represents percent of probe sets in each category of the 22,746 probe sets on the Affymetrix ATH1 22K GeneChip (4). (B) Probe sets corresponding to TF genes were classified into TF families by using publicly available *Arabidopsis* TF databases [Database of *Arabidopsis* Transcription Factors (DATF) (15); *Arabidopsis* Gene Regulatory Information Server (AGRIS) (16); Plant Transcription Factor Database (PlnTFDB) (17); and RIKEN *Arabidopsis* Transcription Factor database (RARTF) (18)]. The 1,954 probe sets representing TF genes were classified into more than 69 TF families. Not all TF families are shown in the figure. Numbers in parentheses represent percent of TF family probe sets out of the 1,954 probe sets representing TF genes. A Microsoft Excel file containing the classification of all genes on the Affymetrix *Arabidopsis* ATH1 22K GeneChip with respect to functional categories and transcription factor families can be downloaded directly from our *Arabidopsis* GeneChip web site at <http://estdb.biology.ucla.edu/genechip/>.



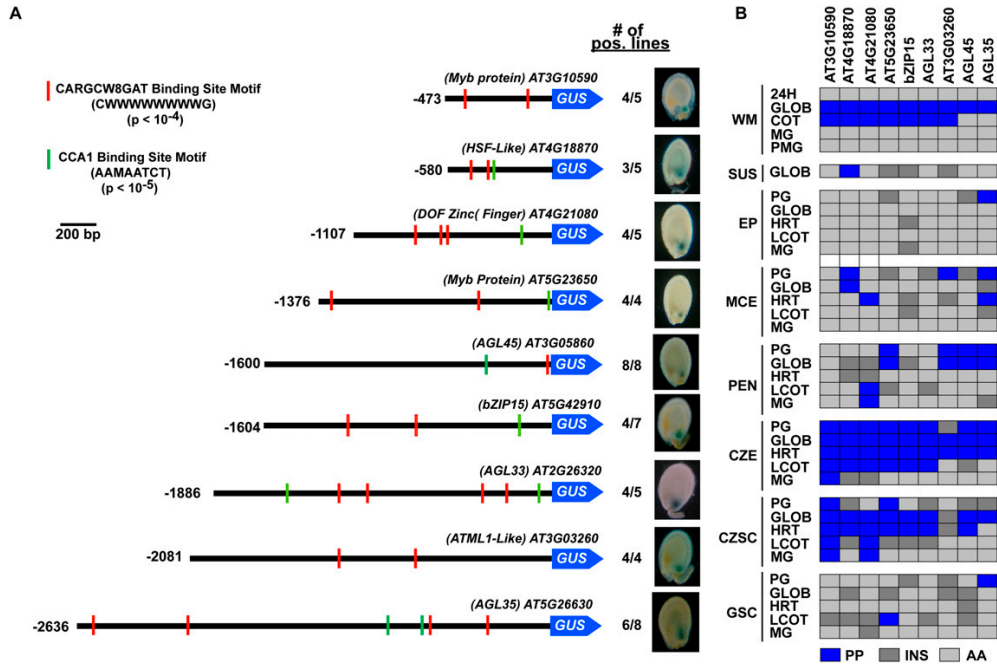
**Fig. S3.** GeneChip detection limit and correlation between GeneChip data and qRT-PCR. (A) Transcript detection limit using the *Arabidopsis* ATH1 22K GeneChip. LEC1 and RAP2.1 biotinylated cRNAs were mixed with 15  $\mu$ g of biotinylated *lec1-1* seed cRNAs at varying dilution levels (1:100–1:1,000,000) and hybridized with replicate Affymetrix ATH1 22K GeneChips as described in *SI Materials and Methods*. GeneChip MAS 5.0 signal intensities of the LEC1 and RAP2.1 cRNAs were plotted relative to their spiked-in prevalences in the *lec1-1* RNA population (i.e., 1:100, 1:1,000, etc). Only GeneChip data points with consensus detection calls of PP for each replicate dilution were plotted (*SI Materials and Methods*). Consensus detection calls of AA were obtained with the 1:500,000 and 1:1,000,000 dilutions and were not plotted. The Pearson correlation coefficient ( $r$ ) of the spiked-in LEC1 and RAP2.1 cRNA dilution series was 0.96. Blue and red circles represent RAP2.1 and LEC1 spiked-in cRNA data points, respectively. (B) Correlation between GeneChip signal intensities and real-time qRT-PCR Ct values. Average Ct values from qRT-PCR experiments were plotted against the mean GeneChip signal intensities of several mRNAs, including seed-stage-specific TFs (Dataset S1). Only GeneChip data with consensus detection calls of PP were plotted (*SI Materials and Methods*). The Pearson correlation coefficient ( $r$ ) calculated from the regression line was 0.76. (C) Correlation between Ct value and mRNA prevalence. The relationship between Ct value and fraction of RNA population for a specific mRNA was extrapolated from the relationships between (i) GeneChip signal intensity, (ii) Ct value, and (iii) fraction of mRNA population shown in A and B. We determined the predicted fraction of RNA population ( $z$ ) for a given Ct value ( $x$ ) by using the regression line equations obtained from A and B. Ct values ( $x$ ) obtained from qRT-PCR analysis were plotted against the fraction of RNA population ( $z$ ) for mRNAs with a GeneChip MAS 5.0 consensus detection calls of either present (blue circles) or absent (red circles). The vertical line represents the GeneChip detection limit ( $2 \times 10^{-5}$ ) taken from A.



**Fig. S4.** Genes active in *Arabidopsis* floral and vegetative organs. (A) Bright-field photographs of vegetative and floral organs used for GeneChip analysis (*SI Materials and Methods*). (B) Number of mRNAs detected by the GeneChip in different organ systems. Numbers for biological replicates 1 and 2 indicate the number of probe sets with detection calls of P by the MAS 5.0 software in each experiment. The number for both biological replicates indicates probe sets with consensus detection calls of PP (*SI Materials and Methods*). Biological replicates 1 and 2 were obtained from the organ systems of wild-type and *lec1-1* plants, respectively (*SI Materials and Methods*). Scatter plots and correlation coefficients comparing GeneChip data for the biological replicates are shown in Fig. S1. (C) Number of organ-specific and shared mRNAs in different vegetative and floral organs. Data analysis was carried out as described in Fig. 2 legend and *SI Materials and Methods*. 6,336 of the 10,733 shared mRNAs varied quantitatively across all organ systems (ANOVA,  $P < 0.05$ ). Numbers in parentheses indicate TF mRNAs.

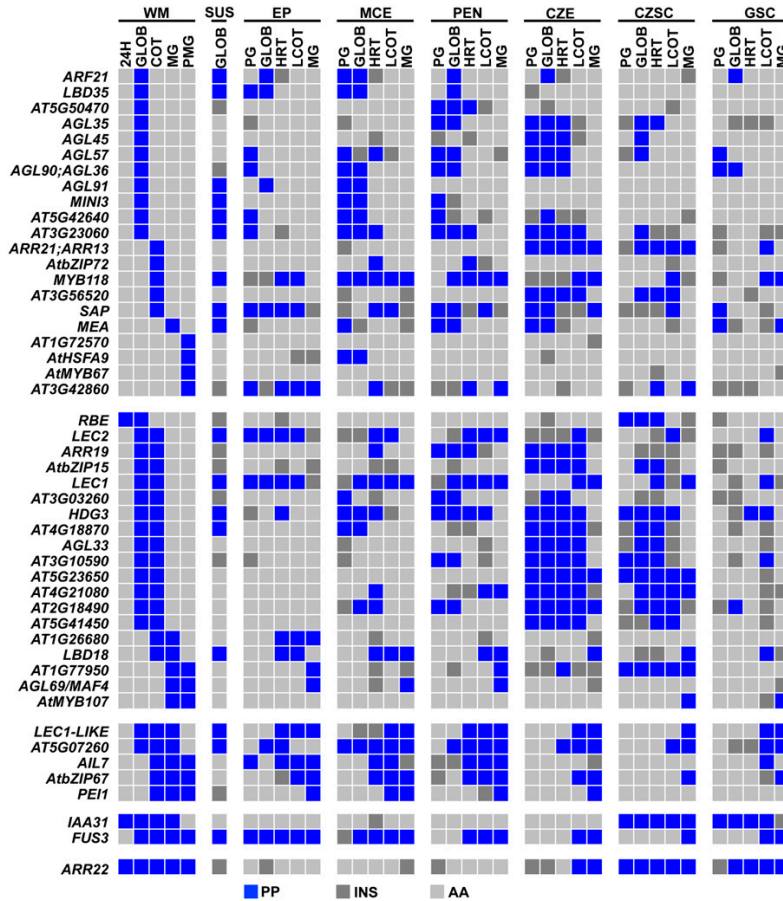


**Fig. S5.** Functional distribution and stage specificity of the seed-specific mRNA population. (A) The 289 seed-specific mRNAs (Fig. 4) were divided into functional categories as outlined in Fig. S2 legend. (B) TF families represented in the 48 seed-specific TF mRNAs (Fig. 4) were categorized according to Fig. S2 legend. (C) Representation of the 289 seed-specific mRNAs at different stages of seed development. Blue and white rectangles indicate the presence (PP) or absence (AA) of a given mRNA at the detection level of the GeneChip (Fig. S2) (*SI Materials and Methods*).



**Fig. 56.** Network of chalazal endosperm transcription factor genes. (A) The 5' upstream region of seed-specific TF genes was fused to the *GUS* reporter gene, and *GUS* enzyme activity was localized within whole mount *Arabidopsis* seeds as described in *SI Materials and Methods*. Upstream regions are not drawn to scale. Thick black bars represent seed-specific TF gene upstream regions. The number to the left indicates the number of nucleotides 5' to the gene translation start site, and the relative size of the upstream region used for analysis. Seeds and embryos from at least four T2 transgenic lines for each seed-specific TF gene upstream region were analyzed. Fractions indicate number of lines with *GUS*-positive results in the chalazal endosperm of whole mount seeds over the total number of lines observed. Enriched TF binding sites were identified by using the Athena database and data-mining tools that search for overrepresented TF binding sites in a set of *Arabidopsis* upstream regions (e.g., chalazal endosperm-specific genes) (<http://www.bioinformatics2.wsu.edu/Athena>) (19). (B) Chalazal endosperm TF mRNA localization within seeds at different stages of development using the Harada-Goldberg Lab GeneChip datasets (*Materials and Methods*). Bright-field photographs of *Arabidopsis* 5–7  $\mu\text{m}$  paraffin seed sections at different developmental stages highlighting areas captured by LCM are shown in Fig. 5B. Blue, light gray, and dark gray squares indicate GeneChip MAS 5.0 consensus detection calls of PP, AA, and INS, respectively (*SI Materials and Methods*). These data are a subset of those shown in Fig. 57. CZE, chalazal endosperm; CZSC, chalazal seed coat; EP, embryo proper; GSC, general seed coat; MCE, micropylar endosperm; PEN, peripheral endosperm; SUS, suspensor; WM, whole mount; PG, preglobular-stage seed; HRT, heart-stage seed; LCOT, linear cotyledon-stage seed.





**Fig. S7.** Localization of seed-specific transcription factor mRNAs within specific seed compartments. Seed-specific TFs were localized to individual seed compartments by using the Harada-Goldberg Lab LCM GeneChip datasets (*Materials and Methods*). Bright-field photographs of *Arabidopsis* 5–7  $\mu\text{m}$  paraffin seed sections at different developmental stages highlighting areas captured by LCM are shown in Fig. 5B. Blue, light gray, and dark gray squares represent GeneChip MAS 5.0 consensus detection calls of PP, AA, and INS, respectively (*SI Materials and Methods*). A small subset of these data are also presented in Fig. 5C [(i.e., AT1G72570, AT3G10590, AT5G23650, AGL69 (*MAF4*), and *AtbZIP67*)]. CZE, chalazal endosperm; CZSC, chalazal seed coat; EP, embryo proper; GSC, general seed coat; MCE, micropylar endosperm; PEN, peripheral endosperm; SUS, suspensor; WM, whole mount; PG, preglobular-stage seed; HRT, heart-stage seed; LCOT, linear cotyledon-stage seed.

## Other Supporting Information Files

[Dataset 1 \(XLS\)](#)

[Dataset 2 \(XLS\)](#)

## CHAPTER THREE

### GLOBAL ANALYSIS OF *LEC1* MUTANT GENE ACTIVITY DURING ARABIDOPSIS SEED DEVELOPMENT

## Abstract

*LEAFY COTYLEDON1 (LEC1)* is a master regulator of embryogenesis and seed maturation in *Arabidopsis*. *lec1* mutants have defects in suspensor development and seed maturation, are desiccation intolerant, and germinate precociously. To further dissect the role of LEC1 during seed development on a genome-wide level, we profiled the *Arabidopsis lec1* seed transcriptome from fertilization through maturation using GeneChip microarrays. By comparing *lec1* seed transcriptomes with those of wild type seeds, we show that LEC1 plays an essential role during early embryogenesis and seed maturation, that the *lec1* mutation causes significant changes in seed gene activity, including seed-specific transcription factors, and that LEC1 acts to repress genes that will be activated after dormancy ends during seed germination.

## Introduction

*LEAFY COTYLEDON1 (LEC1)* is a member of an *Arabidopsis* transcription factor gene class that plays a major role in regulating seed development. These genes include *ABSCISIC ACID INSENSITIVE3 (ABI3)*, *FUSCA3 (FUS3)*, and *LEC2*, and are important for regulating both early seed development and the maturation program that is responsible for the synthesis of storage proteins and oils that will be utilized by the germinating seedling (West et al., 1994). Mutations in the *LEC1* gene cause profound pleiotropic effects on embryo development both early and late in seed development (Lotan et al., 1998). During early embryogenesis, the *lec1* mutation results in improper suspensor cell division, and in combination with other *lec* mutants (e.g. *fus3*, *abi3*), can cause secondary embryos to form from an irregular suspensor (Lotan et al., 1998).



These data suggest that LEC1 might be important in suppressing the embryogenic potential of the suspensor during early embryogenesis (Lotan et al., 1998). During late embryogenesis, the *lec1* mutation affects the ability of the embryo to tolerate desiccation, and post-germination programs that are normally repressed during seed development are initiated prematurely (Lotan et al., 1998; West et al., 1994). Thus, the *LEC1* gene performs major regulatory roles during both early and late embryogenesis.

GeneChip microarrays have been used as a tool for studying gene activity on a genome-wide level during *Arabidopsis* seed development (Belmonte et al., 2013; Le et al., 2010). Previously, we used the *Arabidopsis* ATH1 GeneChip microarray to profile gene activity before, during, and after *Arabidopsis* seed formation, and identified a large number of seed-specific transcription factor genes that may be essential for regulating seed development (Le et al., 2010). To explore the repertoire of genes that might be regulated by *LEC1*, we used the *Arabidopsis* ATH1 GeneChip to profile gene activity in *lec1* seeds from fertilization through maturation. We identified more than 12,000 mRNAs present at any given *lec1* seed stage, and identified mRNA sets that are regulated by LEC1 either directly or indirectly. We also show that a large number of seedling mRNAs are present in *lec1* seeds during maturation, suggesting that LEC1 is important for repressing post-germination programs during seed maturation. Lastly, we show that many seed-specific TFs are affected in the *lec1* mutant background highlighting LEC1 as a master regulator of seed development in *Arabidopsis*.

## **Results**

### **Overview of *lec1* Seed Collection**

Previously we profiled gene activity in *Arabidopsis* wild-type seeds before, during, and after seed development using the *Arabidopsis* ATH1 GeneChip assays (Le et al., 2010). To study the role of LEC1 during and after seed development, we collected *lec1* mutant seeds from seed stages equivalent to those studied during wild-type development (**Figure 3-1A**). Mature *lec1* embryos are altered morphologically and, therefore, proper staging was required in order for *lec1* seeds to be at stages equivalent to those in wild type (see **Materials and Methods**). Several seeds within a silique were dissected to examine embryo development before harvesting seeds within the silique.

We used *Arabidopsis* ATH1 GeneChips to profile genome-wide gene activity in *lec1* seeds throughout development from fertilization through maturation. Because *lec1* mutant seeds show embryo defects during both early and late seed development (Lotan et al., 1998; Meinke et al., 1994; West et al., 1994), we isolated mRNAs from *lec1* seeds containing (i) zygotes 24hr post-pollination (24H), (ii) globular-stage embryos (GLOB) (iii) cotyledon-stage embryos (COT), (iv) mature green embryos (MG), (v) post-mature green embryos (PMG), and three days after imbibition (DAI) seedlings (SDLG) (**Figure 3-1A**). For comparative purposes, we used the GeneChip data analysis methods established in Le et al. (Le et al., 2010).

### **mRNAs Detected During *lec1* Seed Development**

We detected approximately 12,000-14,000 mRNAs at any given stage during *lec1* seed and seedling development (**Figure 3-1B**). Pearson correlation coefficients for the biological replicates ranged from 0.96 to 0.99 indicating that they were in strong agreement with each other (**Dataset 3-1**). The number of diverse mRNAs in *lec1* seeds

remained constant throughout seed development [ $P > 0.91$ , Analysis of Variance (ANOVA)]. Each *lec1* seed stage had a small number of stage-specific mRNAs including transcription factor mRNAs (**Figure 3-1C**). Most mRNAs were present in either multiple seed stages or throughout *lec1* seed development (**Figure 3-1C to 3-1E**), similar to findings during wild type seed development (Le et al., 2010). There was a small overlap between WT and *lec1* stage-specific mRNAs suggesting that LEC1 might play a role in regulating stage-specific mRNAs in wild type seeds.

### ***LEC1 is Important For Regulating Seed Maturation in Arabidopsis***

To further dissect the role of *LEC1* during seed development, we compared the developmental profiles between WT and *lec1* seeds. Whereas the number of diverse mRNAs in *lec1* seed stages remained constant throughout seed development (**Figure 3-1B**), the number of diverse mRNAs decreased significantly during seed maturation in wild type seeds (Le et al., 2010). Pearson correlation coefficients between WT and *lec1* replicates ranged from 0.91 to 0.98 for 24H, GLOB, COT, and SDLG and 0.78 for MG and PMG stages, respectively, suggesting very little differences in gene activity between WT and *lec1* seeds early in development, but large differences during seed maturation (**Dataset 3-1**). In agreement with these results, the number of diverse mRNAs detected were similar between WT and *lec1* 24H, GLOB, and COT seed stages [e.g. 12,421 (WT 24H) vs 12,440 (*lec1* 24H)] but were distinct for MG and PMG seed stages [e.g. 8,779 (WT PMG) vs 12,006 (*lec1* PMG)] (**Figure 3-1B**) (Le et al., 2010). Collectively, these results suggest that *LEC1* performs dual roles during seed maturation and early seed formation.

### **LEC1 Is Essential For Repressing Seedling Genes During Seed Maturation**

We carried out pairwise comparison between WT and *lec1* seeds at each developmental stage to identify mRNA sets that are detected specifically in WT or *lec1* seeds (**Figure 3-2 and Dataset 3-2**). We identified a small number of WT- and *lec1*-specific mRNAs for 24H and COT seed stages, and a much larger number for MG and PMG seed stages, at the level of the GeneChip. For example, we identified 119 mRNAs detected in WT but not *lec1* 24H seeds, and 89 mRNAs detected in *lec1* but not WT 24H seeds, respectively (**Figure 3-2**). By contrast, we identified 104 and 2,142 WT- and *lec1*-specific mRNAs at the PMG stage (**Figure 3-2**), respectively, and this increase in *lec1*-specific mRNAs was coupled with precocious germination of *lec1* seeds, as evident by radical protrusion from the seed (**Figure 3-1A**, inset in PMG). We compared the list of *lec1*-specific mRNAs identified at the 24H, COT, MG, and PMG seed stages to mRNAs detected in wild type seedlings to determine whether the large number of *lec1*-specific mRNAs might be related to post-germination seedling mRNAs (**Figure 3-2**). We detected 49% (44/89), 27% (11/41), and 43% (43/99) WT SDLG-specific mRNAs represented in the *lec1*-specific mRNA sets at the 24H, GLOB, and COT seed stages, respectively. Significantly however, 70% (512/726), and 83% (1,780/2,142) of the *lec1*-specific mRNAs at the MG and PMG seed stages, respectively, were detected in the WT SDLG mRNA population. These results suggest that most *lec1*-specific mRNAs at the maturation stages are seedling mRNAs and that *LEC1* is essential for repressing seedling genes transcription during seed development.

### **Several Seed-Specific TF Genes Are Regulated by *LEC1* Directly or Indirectly**

Previously, we identified 289 seed-specific mRNAs, including 48 TF mRNAs. Many of these seed-specific TF mRNAs, including *LEC1* mRNA, have been shown to play an essential role during seed development (Le et al., 2010). To assess the effects of the *lec1* mutation on seed-specific TF mRNAs, we compared their temporal accumulation patterns and quantitative levels during WT and *lec1* seed development (**Figure 3-4**). 24 TFs have inconsistent detection calls (see Materials and Methods), and, therefore, whether the *lec1* mutation affects these TF mRNAs could not be established (**Figure 3-4**). Five TFs showed no changes in mRNA levels in both WT and *lec1* seeds suggesting that these TFs are not regulated by *LEC1*. By contrast, 19 TFs showed either quantitative changes, temporal changes, or qualitative changes in the *lec1* seed (**Figure 3-4**). Six TFs showed quantitative changes in mRNA accumulation in the *lec1* mutant, including PEI1 and bZIP67. Both PEI1 and bZIP67 transcripts accumulate in wild type MG stage seeds but their mRNA levels were reduced in the *lec1* MG seeds. bZIP67 has been shown to interact directly with *LEC1* and *LEC1-LIKE* to regulate the *sucrose synthase 2 (SUS2)* gene (Yamamoto et al., 2009). Six TFs showed changes in the temporal accumulation of mRNAs in the *lec1* mutant, including *LEC1-LIKE* mRNA. *LEC1-LIKE* mRNA accumulates as early as the GLOB stage seed and peaks at MG stage (**Figure 3-4**) in WT seeds but could only be detected, at the level of the GeneChip, in the *lec1* COT stage seed. Seven TF mRNAs were not detected in the *lec1* seeds throughout development including *LEC1* and *AGL45* mRNAs. The absence of detectable *LEC1* transcripts in the *lec1* seeds confirmed that the mutant is a null and does not transcribe the *LEC1* gene. Approximately 80% (19/24) of the seed-specific TF

mRNAs were affected by the *lec1* mutation indicating that *LEC1* acts as a master regulator seed development.

## **Discussion**

Our genome-wide profile of *lec1* mutant seeds across development provides an in-depth look into *Arabidopsis* seed development in the absence of a major regulator of embryogenesis. Previous characterizations of *lec1* mutant seeds suggest that LEC1 has essential regulatory roles during early and late embryogenesis. In early embryogenesis, it was suggested that LEC1 might serve to repress the embryogenic potential of the suspensor (Lotan et al., 1998). In our comparison of early seed stages (24H, GLOB), we found small number of mRNAs detectable in WT but not *lec1* seeds indicating that these mRNAs might be important for understanding the role of LEC1 during early embryogenesis. Interestingly, YUCCA3 and PIN4, two genes involved in auxin biosynthesis and auxin polar transport, respectively, were not detectably expressed in the *lec1* mutant seeds (Friml et al., 2002; Zhao et al., 2001). Using the *Arabidopsis* LCM GeneChip dataset (Belmonte et al., 2013), we find that YUCCA3 and PIN4 mRNAs are detected in the suspensor of WT globular stage embryo. Mutation in auxin signaling and transport genes result in defective embryo formation, indicating that auxin plays a major role in cell specification during early embryogenesis. Furthermore, it has been proposed that polar auxin transport is important for controlling suspensor fate (Larsson et al., 2008). Therefore, LEC1 might repress the embryogenic potential of the suspensor by ensuring proper auxin biosynthesis and transport during early embryogenesis.

Most of the major effects of the *lec1* mutation are observed during seed maturation, when the embryo is accumulating seed storage protein reserves and preparing for desiccation and dormancy. *lec1* embryos are desiccation intolerant and also germinate precociously, taking on seedling identity during seed development. We found that a large number of seedling mRNAs accumulate prematurely in MG and PMG stage *lec1* seeds indicating that LEC1 normally represses the post-germinative program during seed maturation. Since LEC1 is a transcriptional regulator and a large number of seedling mRNAs are present in the *lec1* PMG seeds, this suggests that LEC1 indirectly suppress the post-germinative program through the activation of other downstream genes. What downstream genes are regulated directly by LEC1 remains to be determined.

Previously, we identified a set of seed-specific transcription factors, which includes LEC1 (Le et al., 2010). LEC1 mRNAs are detected from pre-globular stage to maturation, indicating that LEC1 may have broad regulatory roles during seed development. Several seed-specific TFs were undetected in the *lec1* seeds throughout development, suggesting that LEC1 may play a role in directly or indirectly activating these genes during seed development. Some of these seed-specific TFs were detected early in development (AGL45, AGL57) and overlap with the LEC1 mRNA accumulation pattern. Interestingly, AGL45 mRNAs accumulate primarily in the chalazal endosperm (CZE) from preglobular to heart stage seeds but LEC1 mRNA does not accumulate in the chalazal endosperm until linear cotyledon and mature green stage (Belmonte et al., 2013; Le et al., 2010). The CZE possessed the largest number of seed-specific mRNAs early in seed development (Belmonte et al., 2013). Taken together, these results

suggest that (i) LEC1 likely regulates AGL45 indirectly, (ii) LEC1 could be important for the indirect activation of other seed-specific genes in the CZE early in seed development and (iii) LEC1 likely has an important role in the chalazal endosperm late in seed development. Another seed-specific TF, MYB67, is active late in PMG stage seeds, and does not overlap with LEC1 mRNA accumulation pattern, suggesting that MYB67, too, might be regulated by LEC1 indirectly. The 48 seed-specific TFs uncovered by our microarray study represents a minimal set of seed-specific TFs based on the sensitivity of the GeneChip technology, the stringency of our analysis, and the representation of only 80% of the predicted *Arabidopsis* genes on the array (Le et al., 2010). Perhaps with more sensitive transcriptome profiling technology, like next generation sequencing (RNA-Seq), we can identify additional seed-specific TFs, some of which might be regulated directly by LEC1 in a regulatory network controlling *Arabidopsis* embryogenesis and seed development.

## **Materials and Methods**

### **Plant Material and Growth**

*Arabidopsis lec1-1* homozygous plants were obtained by rescuing and germinating immature *lec1-1* homozygous seeds in culture (Lotan et al., 1998). The *lec1-1* mutant allele is a null allele resulting from a complete deletion of the LEC1 gene (Lotan et al., 1998) and is in the ecotype Wassilewskija (Ws-0). All plant materials were grown in a Conviron chamber and harvested at the University of California, Davis as previously described (Le et al., 2010).



## Characterization of WT and *lec1* Seed Collection

The seed stages harvested in this study are indicated in **Figure 3-1**. For each seed developmental stage, we included two biological replicates that were harvested independently. Twenty-four hour post fertilization *lec1* seeds (24H) were collected from siliques 24 hr after hand pollination of emasculated flowers, and contained zygotes as previously described for WT 24H seeds (Le et al., 2010). Several unpollinated and pollinated pistils were left on plants to measure the extent of cross-pollination and pollination success. Approximately 2.3% cross pollination occurred for the 24H fertilized seed with a pollination efficiency of ~ 70%. Three to four days after pollination (DAP) seeds (GLOB) and seven to eight DAP (COT) seeds contained globular stage embryo and torpedo to linear cotyledon embryos, respectively. We found a clear boundary in *lec1* at eight DAP, when the embryo is just starting to bend. Therefore, *lec1* COT seeds were targeted for seven to eight DAP. Since the *lec1* embryo at the MG stage was morphologically and phenotypically different from WT MG embryo (which is bright green and completely fills the seed coat), we first needed to characterized silique, seed, and embryo phenotypes between 13 and 18 DAP for WT and *lec1* plants. In WT, the MG stage is reached at 13DAP with no noticeable change until 16DAP when the embryo color begins to fade. Therefore, we targeted *lec1* MG seeds for 13-14DAP. In *lec1* mutant seeds, the changes were more subtle, but based on silique and embryo morphology, we could target the equivalent embryo stage by starting from older to younger siliques. We find the first green silique without a yellow tip and examine the seeds inside for transparent seed coat. We selectively move up to younger siliques until we find embryos that were not cramped in the seed coat. Seeds from the next

three younger siliques were then harvested. To ensure there is no overlap between MG and PMG stage seeds, PMG seeds were harvested at the same time and generally three to five siliques separates the two stages. PMG seeds were harvested from 18-19DAP seeds containing very mature but not fully dried seed. Both WT and *lec1* silique morphology and yellowing of the seed coat was very similar and the same criteria was used for both genotypes. The first yellow silique was opened and seeds were observed for a yellow seed coat. If seeds containing yellow seed coats were found, seeds from the next two to three older siliques were harvested. Seeds from shattered siliques were excluded. *lec1* seedlings were harvested from rescued *lec1* immature seeds grown on GM media (1X MS salts, 1% sucrose, 1X B5 vitamins, 0.8% BactoAgar) and cultured in a Percival cabinet at 22°C, 16hrs day:8hrs night. *lec1* seedlings were harvested four full days after the immature seed rescue.

### **RNA Isolation and Affymetrix GeneChip Hybridization**

Total RNA isolation, cRNA synthesis, and GeneChip hybridization were carried out as described previously (Le et al., 2010).

### **GeneChip Data Analysis**

GeneChip CEL files were processed and downstream data analysis including the generation of signal intensities and detection calls and specific analysis were carried out as previously described (Le et al., 2010).

### **Data Submission**

All *lec1* seed development data have been deposited into the Gene Expression Omnibus (GEO) as series GSE 1051.

## References

Belmonte, M.F., Kirkbride, R.C., Stone, S.L., Pelletier, J.M., Bui, A.Q., Yeung, E.C., Hashimoto, M., Fei, J., Harada, C.M., Munoz, M.D., et al. (2013). Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed. *Proceedings of the National Academy of Sciences* 110, E435–E444.

Friml, J., Benková, E., Blilou, I., Wisniewska, J., Hamann, T., Ljung, K., Woody, S., Sandberg, G., Scheres, B., and Jürgens, G. (2002). AtPIN4 Mediates Sink-Driven Auxin Gradients and Root Patterning in Arabidopsis. *Cell* 108, 661–673.

Larsson, E., Sitbon, F., and Arnold, von, S. (2008). Polar auxin transport controls suspensor fate. *Plant Signal Behav* 3, 469–470.

Le, B.H., Cheng, C., Bui, A.Q., Wagmaister, J.A., Henry, K.F., Pelletier, J., Kwong, L.W., Belmonte, M.F., Kirkbride, R., Horvath, S., et al. (2010). Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. *Proceedings of the National Academy of Sciences* 107, 8063–8070.

Lotan, T., Ohto, M., Yee, K., West, M., Lo, R., Kwong, R., Yamagishi, K., Fischer, R.L., Goldberg, R.B., and Harada, J.J. (1998). Arabidopsis LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells. *Cell* 93, 1195–1205.

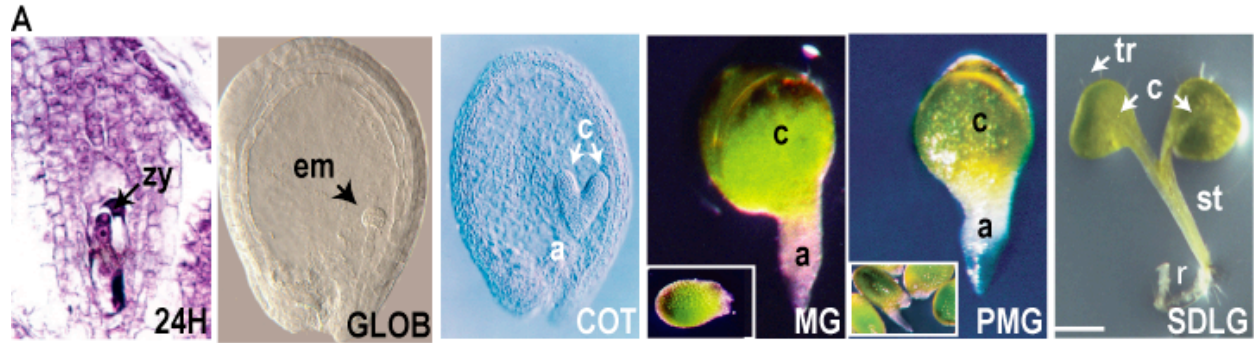
Meinke, D.W., Franzmann, L., Nickle, T., and Yeung, E. (1994). Leafy Cotyledon Mutants of Arabidopsis. *The Plant Cell* 6, 1049–1064.

West, M., Yee, K., Danao, J., Zimmerman, J., Fischer, R.L., Goldberg, R.B., and Harada, J.J. (1994). LEAFY COTYLEDON1 Is an Essential Regulator of Late Embryogenesis and Cotyledon Identity in Arabidopsis. *The Plant Cell* 1731–1745.

Yamamoto, A., Kagaya, Y., Toyoshima, R., Kagaya, M., Takeda, S., and Hattori, T. (2009). Arabidopsis NF-YB subunits LEC1 and LEC1-LIKE activate transcription by interacting with seed-specific ABRE-binding factors. *The Plant Journal* 58, 843–856.

Zhao, Y., Christensen, S.K., Fankhauser, C., Cashman, J.R., Cohen, J.D., Weigel, D., and Chory, J. (2001). A role for flavin monooxygenase-like enzymes in auxin biosynthesis. *Science* 291, 306–309.

**Figure 3-1. Genes active in *lec1* seeds from fertilization to maturation and post-germination.** (A) Bright-field (24H), Nomarski (GLOB, COT), and whole-mount (MG, PMG, SDLG) photographs of seed stages and post-germination seedling used for GeneChip analysis, respectively. 24H seed samples were visualized from 10  $\mu$ m stained paraffin sections (58). Insets show seeds used to dissect whole-mount MG and PMG *lec1* embryos. (B) Number of mRNAs detected at each stage of development. Numbers for biological replicates 1 and 2 indicate the number of probe sets with a MAS 5.0 detection call of P in each experiment (Materials and Methods). The number for both biological replicates indicates a consensus probe set detection call of PP and was used for subsequent analysis (Materials and Methods). (C–E) Minimum number of specific and shared mRNAs at each developmental stage. (C) mRNAs detected in a specific stage or in all stages. mRNAs shared by two stages (D) and three to five stages (E). Number in parentheses indicates TF mRNAs. a, axis; c, cotyledon; em, embryo; r, roots; st, stem; tr, trichome; zy, zygote.



**B**

	24H	GLOB	COT	MG	PMG	SDLG
No. of mRNAs Detected						
Biological Replica #1	13,557	13,154	13,492	12,552	12,434	14,383
Biological Replica #2	12,959	14,056	13,123	13,429	13,955	14,707
No. of mRNAs Detected in Both Biological Replicas	12,440	12,796	12,454	11,837	12,006	13,802

**C**

	24H	GLOB	COT	MG	PMG	SDLG
No. of mRNAs Detected Specifically in Each Stage	86 (5)	54 (2)	45(4)	9 (0)	20 (4)	315 (26)
No. of mRNAs Detected in All Stages	9,073 (422) →					

**D**

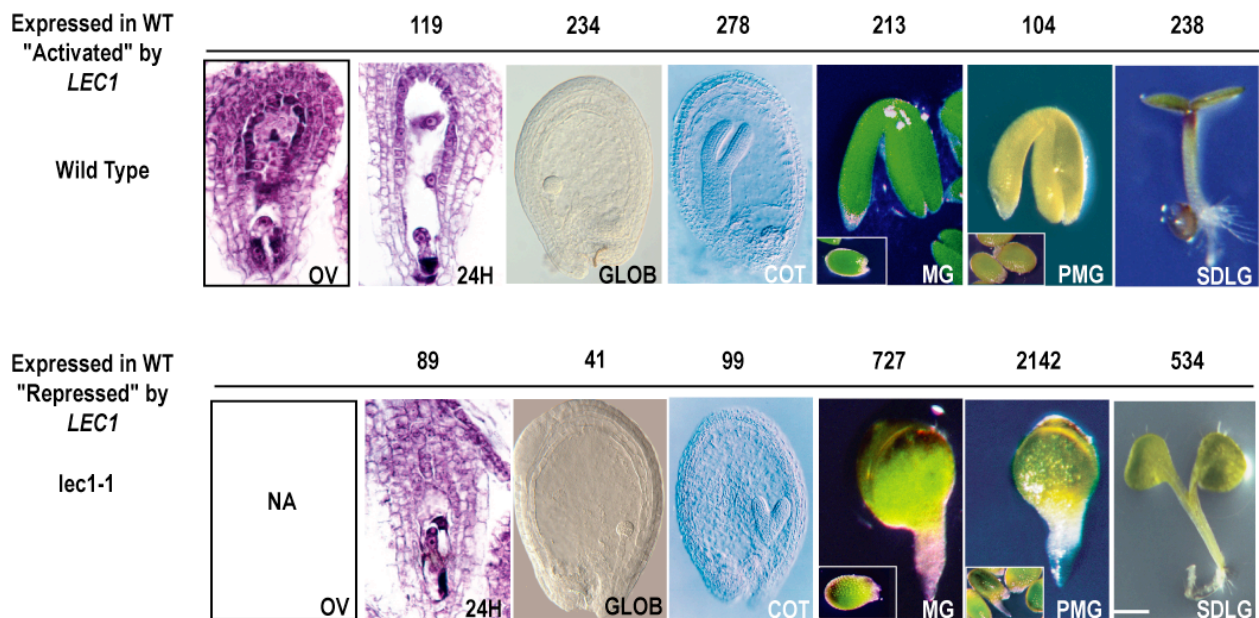
		+ 24H	+ GLOB	+ COT	+ MG	+ PMG	+ SDLG
No. of mRNAs Detected Specifically in Pairs of Stages	24H	86 (5)	32(3)	5(1)	1(0)	1 (0)	27 (2)
	GLOB	-	54 (2)	103(14)	1(0)	1 (0)	14 (1)
	COT	-	-	45 (4)	1(0)	1(0)	8 (0)
	MG	-	-	-	9 (0)	62 (6)	3 (0)
	PMG	-	-	-	-	20 (4)	28 (2)
	SDLG	-	-	-	-	-	315(26)

**E**


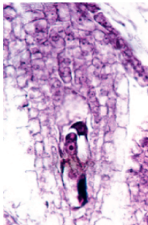


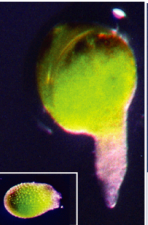
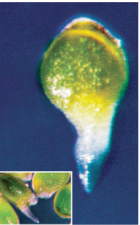
	Three-Stages	Four-Stages	Five-Stages
No. of mRNAs Detected in at Least	367 (32)	190 (22)	261 (19)

**Figure 3-2. Pairwise comparison of WT and *lec1* seed stages and seedling.**

Pairwise comparison of equivalent seed stages between WT and *lec1* seeds and seedlings was carried out to identify mRNAs detected in one genotype and not the other at the specific equivalent seed stage. Wild type seed developmental data for the comparison was obtained from Le et al. (Le et al., 2010). Wild type seed and seedling images were taken from Le et al. (Le et al., 2010). *lec1* seed and seedling images are the same as in Figure 3-1.



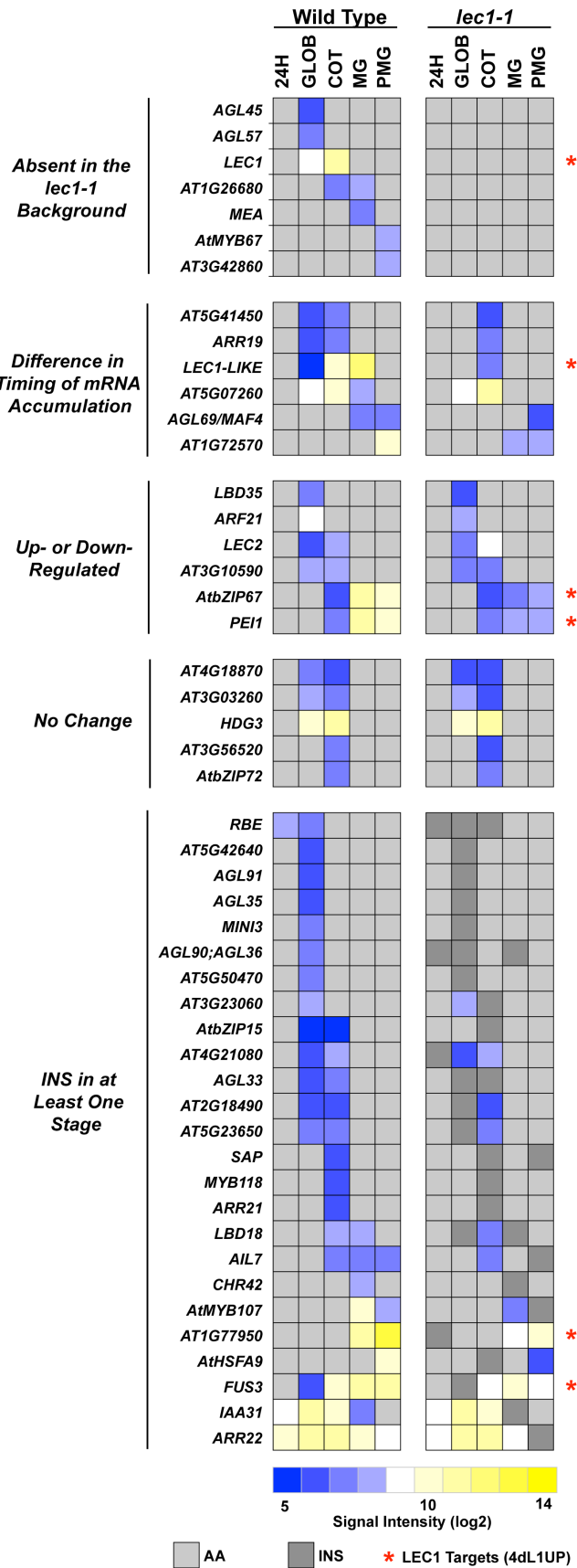
**Figure 3-3. LEC1 represses seedling genes during seed maturation.** mRNAs detected in *lec1* seeds but are not detected in WT seeds (see Figure 3-2) were compared against the mRNA sets present in WT seedling. WT seedling image and data was obtained from Le et al. (Le et al., 2010). Percentage calculation in the third row is defined as the number in row two divided by the number in row one.

	Wild Type	lec1-1				
	SDLG	24H	GLOB	COT	MG	PMG
						
# of mRNAs Detected in <i>lec1</i> Seed Stage NOT Detected in Wild Type Seed Stage		89	41	99	727	2142
# of mRNA Detected in Wild Type Seedling NOT Detected in Wild Type Seed		44	11	43	512	1780
Percent of WT Seedling mRNAs Detected in <i>lec1</i> Seed Stage		49.4%	26.8%	43.4%	70.4%	83.1%



**Figure 3-4. LEC1 regulates many seed-specific TFs during seed development.**

The list of 48 seed-specific TFs and the corresponding detection and transcript levels were obtained from Le et al. (Le et al., 2010). Heatmap representation of seed-specific TFs activity in WT and *lec1* seeds. AA, probe set with detection call of A in both biological replicates; INS, probe set with inconsistent detection call between biological replicates. \* 4dL1UP (Four day induction, LEC1 up-regulated) corresponds to genes that are up-regulated from *LEC1* induction studies and are putative direct LEC1 targets.



## Tables and Datasets

Dataset 3-1 - Raw GeneChip data from *lec1* seed and correlation coefficients of biological replicates

Dataset 3-2 - List of genes obtained from pairwise comparison of WT and *lec1* seed developmental stages

## CHAPTER FOUR

### DYNAMIC CHANGES IN DNA METHYLATION DURING SOYBEAN SEED DEVELOPMENT

## Introduction

Seeds are derived from a unique double fertilization event: the two sperm cells fuse with the egg cell and central cell to form the zygote and endosperm, respectively (Goldberg et al., 1989). The embryo initially undergoes a morphogenetic program with rapid cell division, establishing the two major embryonic organs: an axis -- containing a shoot and root meristem -- that will give rise to the mature plant after seed germination, and the cotyledon, a terminally differentiated organ with specialized functions for storage reserve production that senesces following germination and seedling development (**Figure 4-1**). A major shift occurs from a morphogenetic program to a maturation program that includes cessation of cell division, activation of cell expansion processes to enlarge the seed, accumulation of storage reserves, and preparation for desiccation (**Figure 4-1**). During maturation, cotyledon cells enlarge and undergo a unique endoreduplication process that facilitates the production of highly prevalent seed storage protein RNAs that are utilized during germination and seedling growth. Finally, during dormancy, metabolic and developmental processes are suspended, RNA and protein synthesis terminates, and the quiescent seed awaits the optimal environment for germination and seedling growth (Goldberg et al., 1989), a novel seed property that allows seed survival even after 2,000 years (Sallon et al., 2008). Genetic and genomic studies of seed have uncovered regulators of seed development (e.g. LEAFY COTYLEDON1, FUSCA3), identified temporally-regulated seed-specific transcriptional processes, determined that both maternal and paternal genome contributes to early seed development (Harada and Pelletier, 2012), and that imprinting plays a crucial role in regulating seed size and early development. However, very little is known about the crucial epigenetic regulators and processes controlling seed development, and how genetic and epigenetic mechanisms work together to regulate seed formation.

Recently, an explosion of whole genome methylome studies have shown the essential roles that DNA methylation plays in eukaryotic differentiation and development. For example, in animals, DNA methylation has been shown to be important in imprinting, X-chromosome inactivation, and silencing of tumor-suppressor genes (Bird, 2002). Recent genome-wide

methylome studies report distinct methylation patterns between different cell types (e.g. undifferentiated vs differentiated cells) (Laurent et al., 2010; Lister et al., 2009), and organs (e.g. honeybee queens vs workers brains) (Foret et al., 2012; Lyko et al., 2010), and have identified new imprinted loci (Xie et al., 2012). In plants, methylation has been shown to be important for development because DNA methylation mutants have pleiotropic effects (Zhang and Jacobsen, 2006). In addition, methylation play an important role in imprinting, which is critical for early seed development (Köhler et al., 2012), and recent methylome studies have identified novel imprinted loci on a genome-wide basis (Gehring et al., 2009; Hsieh et al., 2009; Hsieh et al., 2011). However, the full extent that methylation plays during seed development (e.g. different developmental phases, tissues) in addition to regulating imprinting is largely unknown.

We profiled the DNA methylation landscape during soybean seed development at single-base resolution to investigate the extent to which DNA methylation plays a role throughout all of seed development -- from fertilization through dormancy and post-germination (**Figure 4-1**). Soybean seeds were investigated because they are (i) an excellent system for dissecting basic processes controlling seed development (Goldberg et al., 1989; Le et al., 2007) and (ii) a major source of food and fuel worldwide, making up ~70% of world protein meal consumption (<http://soystats.com/2012>). We profiled soybean methylomes from (i) globular stage through maturation, dormancy and post-germination, (ii) among different seed parts at distinct developmental stages, and (iii) between different regions within cotyledons undergoing differential endoreduplication (**Figure 4-1**). We addressed the following questions: (i) are there DNA methylation changes during seed development, (ii) is there differential methylation in different seed parts, and (iii) is there a link between DNA methylation and endoreduplication?

We found significant global CHH methylation increase in whole seeds from maturation to dormancy and decrease in post-germination seedling, which is coupled with global seed coat CHH hypomethylation relative to the rest of the seed. In addition, DNA methylation patterns are maintained during endoreduplication. These results suggest a role for DNA methylation in regulating seed dormancy, a role for the seed coat in regulating embryo and seed development,

and a unique correlation between endoreduplication and DNA methylation. Collectively, these data provide a comprehensive spatial and temporal profile of the DNA methylation landscape during soybean seed development from fertilization to dormancy and uncovers new roles for DNA methylation during seed development.

## **Results**

### **Single-Base Resolution of Soybean Seed Methylomes From Fertilization Through Dormancy and Post-Germination**

We profiled the soybean seed DNA methylation landscape at a single-base resolution from six stages of development using bisulfite sequencing (BS-Seq) to obtain a comprehensive methylation profile from fertilization through dormancy and post-germination (**Figure 4-1; Table 4-1**). We studied the methylomes of whole seeds and seedlings representing four major developmental phases: (1) differentiation and morphogenesis [globular stage (GLOB) seeds], (2) maturation [early- (EM), mid- (MM), and late-maturation stage (LM) seeds], (3) dormancy [dry (DRY) seeds], and (4) post-germination [whole seedlings (SDLG) and seedling cotyledons (SDLG-COT) six days after imbibition] (**Figure 4-1; Table 4-1**). We also dissected embryonic axis (AXIS), embryonic cotyledons (COT), and seed coat (SC) from EM and MM seeds to compare the methylomes between different seed organs and regions at distinct developmental stages. Lastly, we used laser capture microdissection (LCM) of EM stage seeds to dissect different seed coat layers in order to compare the methylomes of different seed coat tissues as well as to obtain abaxial (ABPY) and adaxial (ADPY) cotyledon parenchyma tissues – two tissues showing differential endoreduplication within the same embryonic organ. Collectively, our BS-Seq datasets provide a comprehensive spatial and temporal profile of the DNA methylation landscape across the entire soybean genome throughout all of seed development.

### **Deep Sequence Coverage of the Soybean Methylome**

We generated 144 to 557 million Illumina reads for each seed stage, region, organ, and tissue, representing seven to 35 times coverage of the ~1 Gb soybean genome (**Table 4-S1**). We detected 259 to 287 million cytosines representing 89% to 98% of all detectable genomic cytosines with at least one read (**Table 4-S1 and Supplementary Information**), and obtained an average sequence depth of 4X to 19X per cytosine. We checked the conversion efficiency of the bisulfite treatment by examining the conversion of C to T in the unmethylated chloroplast and spiked-in lambda genomes. We observed on average BS conversion efficiency of unmethylated C to T greater than 99.4% for both the chloroplast and lambda genomes indicating a high conversion efficiency for our bisulfite treatment. Additionally, we observed that 9%, 12%, and 79% of the seed methylome is present in CG, CHG, and CHH contexts (where H = A, C, T), which is similar to the proportion of CG, CHG, and CHH sites in the soybean genome (**Figure 4-S1A**). These results indicate that our datasets represent an unbiased, deep representation of the soybean seed methylome.

### **A Small Fraction of the Soybean Seed Genome is Methylated**

To determine the genome-wide soybean seed methylation level, we calculated the fractional methylation levels (i.e.  $C/(C+T)$ ; see Materials and Methods) and observed an average genome-wide methylation level of 11.4% for all detected cytosines, with 54.7%, 34.3%, and 2.3% of cytosines being methylated, on average, in CG, CHG, and CHH contexts, respectively (**Figure 4-S2A; Table 4-S2**). To determine the distribution of these methylated cytosines along the soybean seed genome, we calculated fractional methylation levels in 500 kb windows spanning all 20 chromosomes for each DNA context separately. We observed high methylation levels primarily along the TE-dense pericentromeric region including the centromere and low methylation levels primarily at the gene-rich chromosomal arms across all 20 chromosomes (**Figure 4-2A**). In agreement with these observations, average methylation along TEs were relatively high compared to genes in all DNA context (**Figure 4-2C and 4-2D**). Taken together,



these data suggests that a small fraction of the soybean seed genome is methylated, primarily within TEs.

## **Global CHH Methylation Changes During and After Seed Development**

### ***CHH methylation increases during seed development and decreases after germination***

To determine whether global DNA methylation changes occur during soybean seed development, we profiled whole seed methylomes at the GLOB, EM, MM, LM, and DRY stages, representing the differentiation, maturation, and dormancy phases (**Figure 4-1**). We did not observe any significant global DNA methylation changes in any DNA context during the transition from morphogenesis to maturation (i.e. GLOB to EM), (**Figure 4-2A and 4-2B**). Interestingly, during maturation, CHH methylation levels increased significantly from the EM to MM stages and peaked at the DRY stage, when the embryo becomes dormant. For example, CHH methylation levels increased ~3-fold during maturation from the EM to MM stages ( $p < 0.001$ , t-test) (**Figure 4-2A and 4-2B; Tables 4-S2 and 4-S3**). The increase in CHH methylation was associated with an overall increase in the proportion of mCHH versus mCG or mCHG sites detected during maturation and dormancy even though the absolute number of mCG and mCHG remains constant (**Figure 4-S1B**). By contrast, the average CG and CHG methylation levels remained relatively unchanged from fertilization through dormancy ( $p > 0.05$ , t-test) (**Figure 4-2B; Table 4-S3**).

We asked whether the CHH hypermethylation during seed maturation and dormancy was maintained following germination. We profiled SDLG and SDLG-COT methylomes, representing the transitional state from a dormant embryo to a highly active seedling. Because the dormant seed is comprised of mostly cotyledon tissue, the SDLG-COT provides a direct access to methylation changes in the embryo post-germination. Interestingly, CHH methylation levels were not maintained but decreased following germination. Specifically, CHH methylation levels decreased 2.4-fold and 1.7-fold in the SDLG and SDLG-COT, respectively ( $p < 0.001$ , t-test) (**Figure 4-2A and 4-2B; Table 4-S3**). By contrast, CG and CHG methylation levels

remained unchanged relative to other seed stages. Taken together, these results indicate that DNA methylation changed dramatically during seed development and CHH hypermethylation during seed development was reversed following seed germination.

### ***CHH methylation changes occur primarily within transposable elements***

We scanned the genome to determine what sequences were primarily affected by CHH hypermethylation during seed development. We examined the genome-wide methylation distribution during and after seed development across the 20 chromosomes and observed major differences in CHH methylation primarily in TE-dense regions including the pericentromere (**Figure 4-2B; Figure 4-S2**). Furthermore, we did not observe major differences in gene body methylation for all DNA context (**Figure 4-2C**) but observe significant changes in TE methylation in the CHH context (**Figure 4-2D**). We looked at different TE classes (e.g., DNA transposons, retrotransposons) and observed that the dramatic changes observed in CHH methylation occurred in all TEs regardless of type and sizes (**Figure 4-S2**). For example, CHH methylation levels increased in both DNA and retrotransposons, including Gypsy, Copia, and Mutator family – with no apparent differences observed between TE classes (**Figure 4-S2**). SDLG-COT showed overall hypomethylation in all DNA context relative to the DRY seed both within genes and TEs (**Figures 4-2C and 4-2D**) suggesting global hypomethylation of the soybean cotyledon genome post-germination. These results suggest that CHH hypermethylation during seed development and hypomethylation during post-germination mainly affects TE methylation.

### **Major regulators of seed development are unaffected by DNA methylation**

DNA methylation have been shown to be an important regulatory mechanism for gene expression during plant differentiation and development. To determine whether seed-specific transcription regulators including the LEAFY COTYLEDON genes (LEC1, LEC1-LIKE, LEC2, FUS3) and the ABA-INSENSITIVE 3 (ABI3) gene, as well as seed-storage protein genes (e.g. Glycinin, conglycinin) and seedling-specific genes are regulated at the DNA methylation level,

we examined the transcriptional and DNA methylation changes during and after seed development (**Figure 4-3**). RNA-Seq data indicated that most of the transcriptional regulator transcripts and storage protein reserves accumulate primarily during seed maturation and is not present post-germination, whereas ICL and CAB1 accumulated predominantly in the SDLG and SDLG-COT (**Figure 4-3A and 4-3C**). The transcript accumulation changes dynamically during seed maturation with most storage protein gene transcript accumulation peaking during mid-maturation. However, examination of the DNA methylation status within these genes and 5 kb upstream or downstream of each gene indicated that these seed- and seedling-specific genes lack DNA methylation within the gene body (**Figure 4-3B and 4-3D**). CG and CHG methylation around these genes remain constant during and after seed development while CHH methylation showed increased methylation for some genes from mid-maturation to dormancy as was observed in our global analysis (**Figure 4-2**). Although the accumulation of seed-specific regulators, storage protein genes, and seedling-specific genes transcripts changed dynamically during development, these changes are not associated with any major changes in the DNA methylation status within or near these genes suggesting that many of these genes are regulated at the transcriptional or post-transcriptional but not at the DNA methylation level in agreement with observations made two decades earlier (Walling et al., 1986).

### **Localized DNA Methylation changes in all sequence context during seed development**

Global methylation analysis indicates that major methylation differences occur in the CHH context during seed development. To determine if there were local DNA methylation changes in the soybean genome during seed development, we carried out pairwise comparisons of seed stage methylomes to identify differentially methylated regions (DMRs) along the genome (see **Supporting Materials and Methods**) in different DNA context. We identified 1,859, 1,767, and 2,142 DMRs in the CG, CHG, and CHH context, respectively, representing ~ two million bases or 0.2% of the entire genome (**Figure 4-4A and Table 4-S4**). The average and largest DMRs have lengths ranging from 220 to 260 bp and 500 to 1,100 bp, respectively, for all DNA context,

indicating there are no changes in DNA methylation over large segments of the genome during seed development. Interestingly, many CG-DMRs and CHG-DMRs were identified between the DRY seed and early seed stages as well as seedlings. For example, we identified 1,030 and 572 CG-DMRs between the GLOB-DRY and DRY-SDLG comparisons, respectively, suggesting major DNA changes from morphogenesis to dormancy and post-germination. These DMRs were classified into four categories representing major genomic features: genes, TEs, promoters, and intergenic regions. The majority of CG-DMRs were identified within or upstream of a gene (56.3%) while 39.8% and 3.9% were in intergenic regions and TEs, respectively. On the other hand, most CHH-DMRs were identified in intergenic regions (68.8%) and a smaller proportion in genes (6.6%) or TEs (12.4%). To further explore the methylation changes during seed development, we clustered the DMRs and observed several major patterns of local DNA methylation changes (**Figure 4-4B**). For example, in the CG-DMRs, there DNA methylation changes during seed development, primarily during MM, LM, and DRY seed stages and in the SDLG (Patterns 2 and 3), similar to the changes observed in CHH methylation on a global scale, indicating major changes occurring during seed desiccation, preparation for dormancy, and post-germination. Interestingly, for these CG-DMRs, we observed varying DNA methylation changes in the SDLG-COT, with DMRs showing similar patterns between the DRY seed and SDLG-COT (Patterns 2a and 3a) while Patterns 2b and 3b showed differences in methylation between the DRY seed and SDLG-COT (**Figure 4-4B**). Furthermore, we observed DMRs showing the presence (Pattern 1b) or absence (1a) of DNA methylation, respectively, in the seedling cotyledon relative to other seed stages (**Figure 4-4B and 4-4C**). Since there are no cell divisions occurring in the seedling cotyledons except for DNA synthesis, the presence or absence of DNA methylation marks post-germination suggests that active and/or passive DNA methylation might be involved post-germination in the seedling cotyledons. To determine if these DMRs affect changes in transcript accumulation during seed development, we examined DMRs within 1 kb upstream of a gene, under the hypothesis that DNA methylation changes in the promoter can have dramatic effects on transcription. We clustered the methylation changes

for the promoter CG-DMRs and the RNA-seq transcript levels for the corresponding downstream genes (**Figure 4-4D**) and did not observe major correlation between the DNA methylation changes and the corresponding transcript accumulation patterns, indicating that the roles of these DMRs during seed development and post-germination remains to be determined.

### **Endoreduplication and DNA methylation are coupled during seed development**

Since the increase in DNA methylation coincide with the onset of endopolyploidization in the cotyledons, we asked whether the DNA methylation landscape is maintained following endoreduplication. To address this question directly, we used LCM to capture ABPY (endoreduplicating) and ADPY (non-endoreduplicating) cotyledonary parenchyma tissues undergoing differential endoreduplication (**Figure 4-5A and 4-5B and Table 4-1**) at the EM stage (Li, 2004). We analyzed DNA sequence coverage along the entire genome for ABPY and ADPY tissues (**Figure 4-5C**), and did not observe any major differences in genome coverage indicating that there is uniform DNA replication along the genome in endoreduplicating cells. That is, all DNA sequences in the genome are replicated to the same extent. There were no significant differences in DNA methylation between the ABPY and ADPY (**Figure 4-5D**) tissues in all DNA context. We independently confirmed these results using whole cotyledons (endoreduplicating) and axis (non-endoreduplicating) from EM and MM stages (**Figure 4-6B**) (Dhillon and Miksche, 1983). For example, DNA methylation around seed storage protein genes are unchanged between endoreduplicating and non-endoreduplicating tissues (**Figure 4-5E**). These results indicate that DNA methylation is maintained in all DNA contexts during endoreduplication and is highly coordinated during DNA synthesis in the presence or absence of cell division.

### **Global hypomethylation at CHH sites in the seed coat**

#### ***The Seed Coat is Hypomethylated Relative to the Embryo***

To determine where in the seed CHH hypermethylation occurs, we profiled three major seed parts present during maturation: embryonic cotyledons, embryonic axis, and seed coat, from EM and MM stage seeds, representing the stages where the initial CHH hypermethylation during seed development appeared (**Figure 4-6A; Materials and Methods**). Similar to whole seed methylomes at EM and MM stages, we did not observe differences in CG or CHG methylation among seed parts from both seed stages (t-test  $p < 0.01$ , ratio  $\geq 1.5$ ) (**Figure 4-6B**), indicating that globally, CG and CHG methylation are affected equally in all seed compartments. By contrast, we observed significant changes in CHH methylation between seed parts within the same stage (e.g. AXIS versus SC) and also between EM and MM stages (**Figure 4-6B**). We observed CHH methylation increasing from EM to MM stage within the same seed parts; that is, CHH methylation increased by developmental stage, irrespective of seed tissue (**Figure 4-6B to 4-6D**). For example, there was a 1.6-fold, 2.0-fold, and 1.6-fold increase in CHH methylation from EM to MM stage for AXIS, COT, and SC, respectively (**Table 4-S3**). Surprisingly, we noticed that CHH methylation is hypomethylated in the SC relative to the embryo parts (AXIS and COT) (t-test  $p < 0.001$ , mean ratio  $\geq 1.5$ ). The SC is hypomethylated 1.8-fold relative to the AXIS (CHH context,  $p < 0.001$ , Student's t-test) at the EM stage and is 1.8-fold and 1.6-fold hypomethylated relative to both the AXIS and COT (t-test  $p < 0.001$ ), respectively, at the MM stage. There are very little changes in gene body methylation in all DNA context among the different seed parts (**Figure 4-6C**). However, significant differences in CHH methylation of TEs was observed between different seed parts (**Figure 4-6D**). For example, TE CHH methylation increased in MM stage AXIS and COT compared to the SC (**Figure 4-6D**), irrespective of TE families (**Figure 4-S2**). These results suggest that CHH hypermethylation during seed development is regulated temporally throughout the seed although differential seed part CHH methylation occurs.

***Seed Coat CHH Hypomethylation is Associated With Increase Small RNAs Abundance at TEs***

The reduction in DNA methylation is generally associated with the activation of TEs as shown in DNA methyltransferases mutant studies (Lippman et al., 2003; Lister et al., 2008). In Arabidopsis, TE transcripts accumulate abundantly in the seed coat of mature seeds compared to the embryo and this accumulation pattern coincide with a reduction in RdDM mRNAs (Belmonte et al., 2013). Interestingly, we found that RdDM mRNAs were also reduced in the SC compared to the AX of soybean EM stage seeds (**Figure 4-S3**) suggesting that the SC hypomethylation might be due to a reduction in RdDM mRNAs and more importantly, that this process is conserved between Arabidopsis and soybean. To determine whether the SC hypomethylation is correlated with siRNA abundance and indirectly TE activity, we profiled smRNA populations from EM-AX, EM-COT, and EM-SC. The seed smRNA populations consisted primarily of 21 and 24 nt smRNAs, with a slight reduction in 24nt smRNAs in the SC (**Figure 4-6E**). We aligned the smRNA sequences to TEs and found that 24nt siRNAs were abundant in TEs in the SC compared to the AX and that this abundance correlated with the decrease in CHH methylation (**Figure 4-6F**). Taken together, these results suggest that the SC hypomethylation might be associated with a reduction in RdDM mRNAs, and in turn, affect the methylation of TEs and its activation in the SC.

## **Discussion**

We presented genome-wide methylome of seeds from after fertilization to maturation and dormancy. These data represents the first DNA methylation study in plants across all of seed development. These developmental seed methylome data showed dynamic changes in DNA methylation during maturation and peaking during dormancy. Increase methylation during dormancy have been observed in many system and is not specific to seed. DNA hypermethylation was observed during bud setting (dormancy) in *Castanea sativa* (Santamaría et al., 2009), tuber meristem dormancy in potato (Law and Suttle, 2003), and induction of dormancy in strawberry (Zhang et al., 2011) suggesting that DNA hypermethylation during dormancy might be a highly conserved mechanism. One possible role for the observed

increase in DNA methylation may be in chromatin compaction and condensation during desiccation and dormancy. Seed maturation and dormancy is characterized by an increase in chromatin condensation (van Zanten et al., 2011), and reduced transcriptional activity (Le et al., 2010). Chodavarupu et al. shows that nucleosomes-bound DNA are highly methylated as compared to the flanking DNA (Chodavarupu et al., 2010) and single-molecule study indicates that DNA methylation increases nucleosome compaction and rigidity (Choy et al., 2010) further supporting the hypothesis that increase DNA methylation during seed maturation and dormancy is associated with chromatin compaction and condensation. Furthermore, the increase CHH methylation could also be important for silencing TEs to maintain genome integrity during desiccation and dormancy as plant TEs are preferentially methylated through the RNA-dependent DNA methylation (RdDM) pathway (Law and Jacobsen, 2010). These processes are not necessarily mutually exclusive. The exact role for increase DNA methylation during soybean seed development remains to be determined.

Following seed germination, there is a large decrease in DNA methylation. In addition to soybean (this study), decrease in DNA methylation after dormancy was observed in other plant species including wheat (Meng et al., 2011), pepper (Portis et al., 2004), rapeseed (Lu et al., 2006), and *Silene latifolia* (Zluvova et al., 2001) suggesting that this process is highly conserved between monocots and dicots. In *Arabidopsis*, chromatin condensation is released within 24 hours after germination suggesting that decrease in DNA methylation could be necessary for chromatin de-condensation and transcriptional activation in the germinating seedling (van Zanten et al., 2011). How is DNA demethylation occurring following germination? One possibility is that active DNA demethylation is involved since decrease in DNA methylation was observed in the seedling cotyledons prior to cell division (Zluvova et al., 2001). In plants, active DNA methylation is carried out by the DNA glycosylase family of base-excision repair enzymes including DEMETER (DME), REPRESSOR OF SILENCING1 (ROS1), DEMETER-LIKE2 (DML2), and DML3. In soybean, because of ancient polyploidization events, there are two putative orthologs for each DME and ROS1 gene (Zemach et al., 2010). mRNA-Seq datasets



indicate that these putative DME and ROS1 orthologs mRNAs are detected in the seedling cotyledon (data not shown) suggesting that active DNA demethylation might occur in the seedling cotyledon. However, we cannot rule out the possibility that passive DNA demethylation might also occur in the germinated cotyledons since there are evidence that although soybean seedling cotyledon cells do not divide, there is an increase in DNA content and endopolyploid nuclei within five to 10 days after germination (Dhillon and Miksche, 1983), indicating DNA replication do occur. It is possible that both active and passive demethylation processes are occurring in the seedling cotyledons following germination.

In our study, global methylation in the CHH but not CHG or CG context decreased in the seedling and seedling cotyledons. Why are only CHH methylation but not CG methylation affected? There is a general notion that CG methylation are stably maintained whereas non-CG methylation including CHH methylation are subjected to developmental regulation (Zhang and Jacobsen, 2006). CHH methylation is mediated through the RdDM pathway, are triggered by siRNAs (Law and Jacobsen, 2010), and both the establishment and maintenance of CHH methylation require constant small RNA targeting. The fact that we observed CHH methylation increasing during dormancy and reversibly decreasing following germination further support this notion. Because dormancy is a temporary transitional period between the mature embryo and post-germinative development, CHH methylation allows for transient developmental regulation as compared to more stable or permanent developmental programs.

The reduction of CHH methylation in the soybean seed coat relative to the embryo resembles other systems including the Arabidopsis sperm cell and vegetative cell in mature pollen (Calarco et al., 2012; Ibarra et al., 2012; Slotkin et al., 2009), and the Arabidopsis and rice endosperm and embryo in developing seeds (Hsieh et al., 2009; Zemach et al., 2010). Both the vegetative cell and endosperm serve as companion cells that protects the integrity of the gamete and zygotic genome, respectively, early in development. Likewise, the seed coat could serve as companion cells that protects the integrity of the embryo genome later during development (e.g. maturation and desiccation), since hypomethylation of the seed coat was

more drastic during mid-maturation compared to early-maturation stage. Similar to the companion cells, vegetative cell and endosperm, the seed coat does not contribute genetic material to the zygotic genome (embryo) that will develop into the next generation plant and therefore could tolerate the activation of TEs that could disrupt the integrity of these genomes. Although amiRNAs movement between the vegetative cell and sperm cell (Slotkin et al., 2009) and the central cell and egg cell (Ibarra et al., 2012) have been demonstrated experimentally, it remains unknown how siRNAs move between cells or systemically (Melnyk et al., 2011). In the other two system, the companion cells are connected symplastically to the male and female gametes, respectively, providing a direct connection for cell-to-cell movement via the plasmodesmata. However, the maternal seed coat tissue is not connected to the embryo symplastically and all transfer of substrates between the seed coat and embryo travel apoplastically. Nevertheless, The parallel characteristic of the maternal seed coat with the male and female gamete companion cells, vegetative cell and endosperm, respectively, raises an intriguing possibility for the epigenetic regulation of the embryo by the maternal tissue during seed development.

## **Experimental Procedures**

### **Plant Growth and Tissue Collection**

Soybean plants (*Glycine max* (L.) cv. Williams 82) were grown at 22°C with a 16-hour light to 8-hour dark cycle in the UCLA Plant Growth Center. Seeds were staged based on seed length, weight, and embryonic characteristics. GLOB, EM, MM, and LM seeds have seed length of 1.0-1.5 mm, 6.0-7.0mm, 110-120mm, and 120-140mm, respectively. In addition, MM and LM seeds weighed between 150-250mg and 230-350mg, respectively. Furthermore, MM embryo have green cotyledons and a yellow axis tip while LM embryo are completely yellow. DRY seeds have an average dry weight of 155mg. SDLG were collected six days after imbibition, having an average hypocotyl length of 8 cm. Cotyledons (COT) were dissected from SDLG six days after imbibition and weighed between 250-400mg. Embryonic axis (AXIS), cotyledons (COT), and seed coat (SC) were manually separated from EM and MM seeds. AXIS and COT were harvested without the plumule. SC were separated from the embryo but might contain remnants of the aleurone and crushed endosperm. Whole seeds and tissues were harvested and quickly frozen in liquid nitrogen. Frozen tissues were ground to a fine powder using a mortar and pestle and stored at -80°C.

### **Laser Capture Microdissection**

EM-stage seeds were harvested, cut in half transversely, fixed in ethanol:acetic acid, dehydrated, infiltrated and embedded in paraffin solution containing Paraplast-X-Tra tissue embedding medium (Fisher Scientific) according to the methods of Kerk et al. (Kerk et al., 2003). Ten micron paraffin cross-sections were cut for each seed halves on a rotary Histocut 820 microtome (Reicher-Jung) and floated in DEPC-treated water to stretch ribbons containing seed serial sections. Seed sections were placed on PEN-foil slides (Leica Microsystems) and de-paraffinized with two two-minutes xylene treatments before LCM. Tissue sections were captured using a Leica LMD6000 microdissection scope (Leica Microsystems) into a PCR tube cap containing DNA isolation solution included in the FFPE DNA isolation kit (Qiagen, Valencia,

CA). Based on endoreduplication studies by Li et al. (Li and Nielsen, 2004), ~350 micron from the cotyledons ends are excluded within each cross-section. Four cell layers excluding the epidermis and three cell layers including the epidermis were captured for ABPY and ADPY, respectively. The epidermis layer of the cotyledons show no endoreduplication and was excluded from the ABPY collection (Li and Nielsen, 2004).

### **BS-DNA-Seq Library**

Genomic DNA was isolated from hand dissected or laser captured seed tissue using the DNEASY Plant Mini kit (Qiagen, Valencia, CA) and FFPE DNA isolation kit (Qiagen, Valencia, CA), respectively, according to the manufacturer's instructions. Approximately 100 nanogram to one microgram of genomic DNA was subjected to library preparation following the methods of Hsieh et al. with modifications (Hsieh et al., 2009). Specifically, three nanogram of unmethylated cl857 Sam7 Lambda DNA (GenBank Accession NC\_001416; Promega, Madison, WI) was spiked-in with the genomic DNA prior to DNA sonication to serve as an internal control for estimating bisulfite conversion efficiency. In one sample, EM-SC, we spiked-in methylated cl857 Sam7 Lambda DNA to estimate the extent of bisulfite over-conversion (i.e. methylcytosines converted to uracil). To generate methylated Lambda DNA, unmethylated Lambda DNA was treated with CpG methyltransferase (NEB, Ipswich, MA), purified following the protocols of Marmur et al. (Marmur, 1961), and validated with methylation-sensitive BstUI enzyme (NEB, Ipswich, MA) according to manufacturer's instructions. Adapter-ligated genomic DNA was subjected to two rounds of bisulfite (BS) treatment using the EpiTect Bisulfite Conversion kit (Qiagen, Valencia, CA). BS-treated DNA was purified and amplified for 10 cycles using ExTaq (Takara, Shiga, Japan) DNA polymerase. PCR-amplified DNA fragments were size selected using the AMPure XP beads (Beckman).

### **mRNA-Seq Library**

Total RNA was isolated from whole seeds and seed parts using the Concert Plant RNA Reagent (Invitrogen, Carlsbad, CA) according to manufacturer's instructions. Total RNA was treated with RNase-free DNaseI (Ambion, Austin, TX) and subjected to two rounds of poly-A+ RNA selection using oligo-dT25 magnetic beads (Dynabeads; Invitrogen, Carlsbad, CA). Approximately 100 nanogram of poly-A+ RNA was subjected to mRNA-Seq library preparation according to the Illumina mRNA-Seq protocol (Part # 1004898 Rev. D). Adapter-ligated cDNAs were size selected on an agarose gel and purified cDNAs were amplified by PCR for 15 cycles. Following cDNA amplification, the DNA was subjected to end repair, A' tail addition, and adapter ligation according to the Illumina mRNA-Seq protocol. For EM seed parts (i.e. axis, cotyledons, seed coat), 20ng of total RNA was amplified with Nugen Ovation Pico WTA system v.1 (Nugen, San Carlos, CA). Double-stranded cDNA was generated using WT-Ovation Exon Module (Nugen, San Carlos, CA) and quantified with Picogreen on the Nanodrop ND-3300 (Nanodrop Technologies). One microgram of double-stranded cDNA was fragmented using NEB Fragmentase for 15 minutes at 37oC for the standard protocol (NEB, Ipswich, MA). The Illumina TruSeq DNA Sample prep kit was used to prepare the Illumina library with modifications (Illumina, San Diego, CA). Specifically, the Covaris shearing step was omitted, the final PCR enrichment step was performed using Agilent Pfu Turbo Cx DNA polymerase (Agilent, Santa Clara, CA) instead of the TruSeq PCR mix, and 12 cycles of PCR were performed.

### **Small RNA-Seq Library**

Total RNA was isolated from EM AXIS, COT, and SC using the Concert Plant RNA Reagent (Invitrogen, Carlsbad, CA) according to manufacturer's instructions. The TruSeq Small RNA Sample Preparation kit was used to generate libraries with modifications (Illumina, San Diego, CA). Specifically, 250 ng total RNA was used as input and 15 PCR cycles was used for the final PCR enrichment step.

### **Sequencing**

Single-end 50-bp or 100-bp reads were generated for each library by the UCLA Genome Sequencing Center (<http://gsc.ucla.edu/>) or the Broad Stem Cell Research Center High Throughput Sequencing Core using the Illumina Genome Analyzer Iix or HiSeq2000 sequencing machines.

### **BS-DNA-Seq Sequence Processing and Alignment**

Read sequences were aligned to a reference genome using the BS Seeker program (Chen et al., 2010) allowing for up to two mismatches. The reference genome consisted primarily of soybean scaffold sequences including the 20 chromosomes, mitochondrion sequences, and other unanchored sequences (Schmutz et al., 2010). Eighty-one scaffolds containing chloroplast sequences was replaced with the 152,218 bp fully sequenced soybean chloroplast genome sequence (DQ317523) (Saski et al., 2005). In addition, the 48,502 bp cl857 Sam7 Lambda genome (NC\_001416) (Sanger et al., 1982), and the 5,386 bp  $\phi$ -X174 (PhiX) genome (NC\_001422) (Sanger et al., 1977) were included as extra chromosomes so that reads from the spiked-in Lambda and PhiX control could align. Only reads that mapped uniquely to the reference genome were retained for further analysis. The BS-Seeker output was subjected to post-processing that consisted of two steps: (1) To reduce PCR amplification bias for each library, clonal reads (i.e. reads containing identical 5' mapped position and exact nucleotide sequence) were collapsed and all but one read was retained (Lister et al., 2008); (2) Reads containing three or more consecutive cytosines in the CHH context, that are likely not bisulfite converted (Cokus et al., 2008), were removed.

### **Methylation Call**

For each cytosine in the reference genome, the read depth is defined as the total number of reads covering the genomic cytosine with a C (representing a methylated cytosine) or a T (representing an unmethylated cytosine). Next, we calculated the methylation level for each cytosine position as  $C / (C + T)$ . We further restricted our downstream analysis to only cytosines

that have at least two reads with methylated cytosines. For 500 kb windows, we calculated the methylation level for the window as the sum of  $C/(C+T)$  for all cytosines from both strands within the window.

### **Differentially methylated regions (DMRs)**

DMRs were defined similar to the approach taken by Laurent et al. (Laurent et al., 2010) with modifications. To reduce variation that might occur due to uneven sequencing depth between tissue samples, we restricted our analysis to only cytosines that were detected in both samples being compared and have read depth  $\geq 5$ . For each common cytosine, we calculated the methylation difference between two BS-DNA-Seq data sets (eg. GLOB - EM), and generated 200bp sliding windows with a 100 step size along the chromosomes for each DNA context. Windows containing no methylated cytosines were removed from further analysis. The average delta methylation for each window was calculated and follows a normally distributed population. The mean and standard deviation of the methylation differences for all windows for each DNA context were calculated and standardized using the following formula where  $n$  is the number of cytosines within each window. A DMR was defined by three parameters: number of cytosines in the window, the mean delta methylation difference, and the z-score. For CG and CHG context, windows containing at least eight cytosines, have mean delta methylation  $\geq 0.50$ , and  $z \geq 10$  were defined as DMR. For the CHH context, the window must have at least eight cytosines, mean delta methylation  $\geq 0.15$ , and  $z \geq 16$  (see **Supporting Information**). DMRs that overlapped by 1bp were merged. DMRs were associated to specific gene or TE if the DMR regions overlap with a gene or TE by one bp or is within 1 kb upstream of the gene or TE and are listed in **Table 4-S4**.

### **mRNA-Seq Sequence Processing and Analysis**

Sequences was subjected to quality filtering. Sequences was first filtered based on the Illumina purity filter and trimmed at the 5' and 3' based on positions with error-rate  $> 0.1\%$ . Second,

rRNA reads were identified by mapping trimmed reads against a rRNA database using Bowtie (version 0.12.7) and removed from further analysis (Langmead et al., 2009). The remaining high-quality reads were mapped to the soybean genome (Glyma1.0) using Bowtie allowing for two mismatches. Only uniquely mapped reads (i.e. reads that map to one unique genomic locus) were used for subsequent analysis. Expression values for each gene within a dataset was normalized as Reads per kilobase per million (RPKM) value according to Mortazavi et al. (Mortazavi et al., 2008).

### **Small RNA Sequence Processing And Analysis**

50-bp small RNA sequences were first selected for reads containing adapters. Reads were trimmed at the adapter site and reads between 18 – 24 nucleotides were kept for further analysis. Trimmed reads were first filtered by removing reads that mapped to rRNA and tRNA sequences. The remaining trimmed reads were aligned to the genome using Bowtie (version 0.12.7) allowing no mismatch in the alignment (-v 0) but allowing matches to multiple positions within the genome (-a). Distinct small RNA sequences were quantified and normalized using the methods of Lee et al. (Lee et al., 2012).

### **Data Availability**

All data have been submitted to the NCBI's Gene Expression Omnibus (GEO) database, <http://ncbi.nlm.nih.gov/geo/> (accession no. GSE34637, GSE37893, and GSE37895)



## References

- Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & Development* 16, 6-21.
- Calarco, J.P., Borges, F., Donoghue, M.T.A., Van Ex, F., Jullien, P.E., Lopes, T., Gardner, R., Berger, F., Feijo, J.A., Becker, J.D., et al. (2012). Reprogramming of DNA Methylation in Pollen Guides Epigenetic Inheritance via Small RNA. *Cell* 151, 194-205
- Chen, P.-Y., Cokus, S.J., and Pellegrini, M. (2010). BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics* 11, 203.
- Chodavarapu, R.K., Feng, S., Bernatavichute, Y.V., Chen, P.-Y., Stroud, H., Yu, Y., Hetzel, J.A., Kuo, F., Kim, J., Cokus, S.J., et al. (2010). Relationship between nucleosome positioning and DNA methylation. *Nature* 466, 388–392.
- Choy, J.S., Wei, S., Lee, J.Y., Tan, S., Chu, S., and Lee, T.-H. (2010). DNA Methylation Increases Nucleosome Compaction and Rigidity. *J. Am. Chem. Soc.* 132, 1782–1783.
- Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452, 215-219.
- Dhillon, S.S., and Miksche, J.P. (1983). DNA, RNA, protein and heterochromatin changes during embryo development and germination of soybean (*Glycine max L.*). *Histochem. J.* 15, 21–37.
- Foret, S., Kucharski, R., Pellegrini, M., Feng, S., Jacobsen, S.E., Robinson, G.E., and Maleszka, R. (2012). DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. *Proc. Natl. Acad. Sci. USA* 109, 4968-4973.
- Gehring, M., Bubb, K.L., and Henikoff, S. (2009). Extensive demethylation of repetitive elements during seed development underlies gene imprinting. *Science* 324, 1447-1451.
- Goldberg, R.B., Barker, S.J., and Perez-Grau, L. (1989). Regulation of gene expression during plant embryogenesis. *Cell* 56, 149-160.
- Harada, J.J., and Pelletier, J. (2012). Genome-wide analyses of gene activity during seed development. *Seed Sci. Res.* 22, S15-S22.

Hsieh, T.-F., Ibarra, C.A., Silva, P., Zemach, A., Eshed-Williams, L., Fischer, R.L., and Zilberman, D. (2009). Genome-wide demethylation of Arabidopsis endosperm. *Science* 324, 1451-1454.

Hsieh, T.-F., Shin, J., Uzawa, R., Silva, P., Cohen, S., Bauer, M.J., Hashimoto, M., Kirkbride, R.C., Harada, J.J., Zilberman, D., et al. (2011). Regulation of imprinted gene expression in Arabidopsis endosperm. *Proc. Natl. Acad. Sci. USA* 108, 1755-1762.

Ibarra, C.A., Feng, X., Schoft, V.K., Hsieh, T.-F., Uzawa, R., Rodrigues, J.A., Zemach, A., Chumak, N., Machlicova, A., Nishimura, T., et al. (2012). Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science* 337, 1360–1364.

Kerk, N., Ceserani, T., Tausta, S., Sussex, I., and Nelson, T. (2003). Laser capture microdissection of cells from plant tissues. *Plant Physiol.* 132, 27-35.

Köhler, C., Wolff, P., and Spillane, C. (2012). Epigenetic mechanisms underlying genomic imprinting in plants. *Annual Review of Plant Biology* 63, 331-352.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Res* 19, 1639–1645.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Kin Sung, K.W., Rigoutsos, I., Loring, J., et al. (2010). Dynamic changes in the human methylome during differentiation. *Genome Res.* 20, 320-331.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* 11, 204–220.

Law, R.D., and Suttle, J.C. (2003). Transient decreases in methylation at 5“-cCGG-3” sequences in potato (*Solanum tuberosum* L.) meristem DNA during progression of tubers through dormancy precede the resumption of sprout growth. *Plant Mol Biol* 51, 437–447.

Le, B.H., Wagmaister, J.A., Kawashima, T., Bui, A.Q., Harada, J.J., and Goldberg, R.B. (2007). Using genomics to study legume seed development. *Plant Physiol.* 144, 562-574.

Le, B.H., Cheng, C., Bui, A.Q., Wagmaister, J.A., Henry, K.F., Pelletier, J., Kwong, L.W., Belmonte, M.F., Kirkbride, R., Horvath, S., et al. (2010). Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. *Proceedings of the National Academy of Sciences* 107, 8063–8070.

Lee, T.-F., Gurazada, S.G.R., Zhai, J., Li, S., Simon, S.A., Matzke, M.A., Chen, X., and Meyers, B.C. (2012). RNA polymerase V-dependent small RNAs in Arabidopsis originate from small, intergenic loci including most SINE repeats. *Epigenetics : Official Journal of the DNA Methylation Society* 7, 781–795.

Li, S., and Nielsen, N.C. (2004). Endoreduplication During Soybean Seed Development (Purdue, Indiana: Purdue University), pp. 1-106.

Lippman, Z., May, B., Yordan, C., Singer, T., and Martienssen, R. (2003). Distinct mechanisms determine transposon inheritance and methylation via small interfering RNA and histone modification. *PLoS Biol* 1, E67.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 133, 523-536.

Lister, R., Pelizzola, M., Downen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences *Nature* 462, 315-322.

Lu, G., Wu, X., Chen, B., Gao, G., Xu, K., and Li, X. (2006). Detection of DNA methylation changes during seed germination in rapeseed (*Brassica napus*). *Chinese Sci Bull* 51, 182–190.

Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C., and Maleszka, R. (2010). The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biol.* 8, e1000506.

Marmur, J. (1961). A Procedure for the Isolation of Deoxyribonucleic Acid from Microorganisms. *J. Mol. Biol.* 3.

Melnyk, C.W., Molnar, A., and Baulcombe, D.C. (2011). Intercellular and systemic movement of RNA silencing signals. *Embo J* 30, 3553–3563.

Meng, F.R., Li, Y.C., Yin, J., Liu, H., Chen, X.J., Ni, Z.F., and Sun, Q.X. (2011). Analysis of DNA methylation during the germination of wheat seeds. *Biol Plant* 1–7.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* 5, 621-628.

Portis, E., Acquadro, A., Comino, C., and Lanteri, S. (2004). Analysis of DNA methylation during germination of pepper (*Capsicum annuum* L.) seeds using methylation-sensitive amplification polymorphism (MSAP). *Plant Sci* 166, 169–178.

Sallon, S., Solowey, E., Cohen, Y., Korchinsky, R., Egli, M., Woodhatch, I., Simchoni, O., and Kislev, M. (2008). Germination, genetics, and growth of an ancient date seed. *Science* 320, 1464.

Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., and Petersen, G.B. (1982). Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.* 162, 729-773.

Sanger, F.F., Air, G.M.G., Barrell, B.G.B., Brown, N.L.N., Coulson, A.R.A., Fiddes, C.A.C., Hutchison, C.A.C., Slocombe, P.M.P., and Smith, M.M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265, 687-695.

Santamaría, M.E., Hasbún, R., Valera, M.J., Meijón, M., Valledor, L., Rodríguez, J.L., Toorop, P.E., Cañal, M.J., and Rodríguez, R. (2009). Acetylated H4 histone and genomic DNA methylation patterns during bud set and bud burst in *Castanea sativa*. *J Plant Physiol* 166, 1360–1369.

Saski, C., Lee, S.-B., Daniell, H., Wood, T.C., Tomkins, J., Kim, H.-G., and Jansen, R.K. (2005). Complete Chloroplast Genome Sequence of *Glycine max* and Comparative Analyses with other Legume Genomes. *Plant molecular biology* 59, 309-322.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178-183.

Slotkin, R.K., Vaughn, M., Borges, F., Tanurdzić, M., Becker, J.D., Feijo, J.A., and Martienssen, R.A. (2009). Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* 136, 461–472.

Thorvaldsdóttir, H.H., Robinson, J.T.J., and Mesirov, J.P.J. (2012). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14, 178–192.

van Zanten, M., Koini, M.A., Geyer, R., Liu, Y., Brambilla, V., Bartels, D., Koornneef, M., Fransz, P., and Soppe, W.J.J. (2011). Seed maturation in *Arabidopsis thaliana* is characterized by nuclear size reduction and increased chromatin condensation. *Proceedings of the National Academy of Sciences* 108, 20219–20224.

Walling, L., Drews, G., and Goldberg, R.B. (1986). Transcriptional and post-transcriptional regulation of soybean seed protein mRNA levels. *Proceedings of the National Academy of Sciences* 83, 2123–2127.

Xie, W., Barr, C.L., Kim, A., Yue, F., Lee, A.Y., Eubanks, J., Dempster, E.L., and Ren, B. (2012). Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* 148, 816-831.

Zemach, A., Kim, M.Y., Silva, P., Rodrigues, J.A., Dotson, B., Brooks, M.D., and Zilberman, D. (2010). Local DNA hypomethylation activates genes in rice endosperm. *Proceedings of the National Academy of Sciences* 107, 18729–18734.

Zhang, L., Wang, Y., Zhang, X., Zhang, M., Han, D., Qiu, C., and Han, Z. (2011). Dynamics of phytohormone and DNA methylation patterns changes during dormancy induction in strawberry (*Fragaria × ananassa* Duch.). *Plant Cell Rep* 31, 155–165.

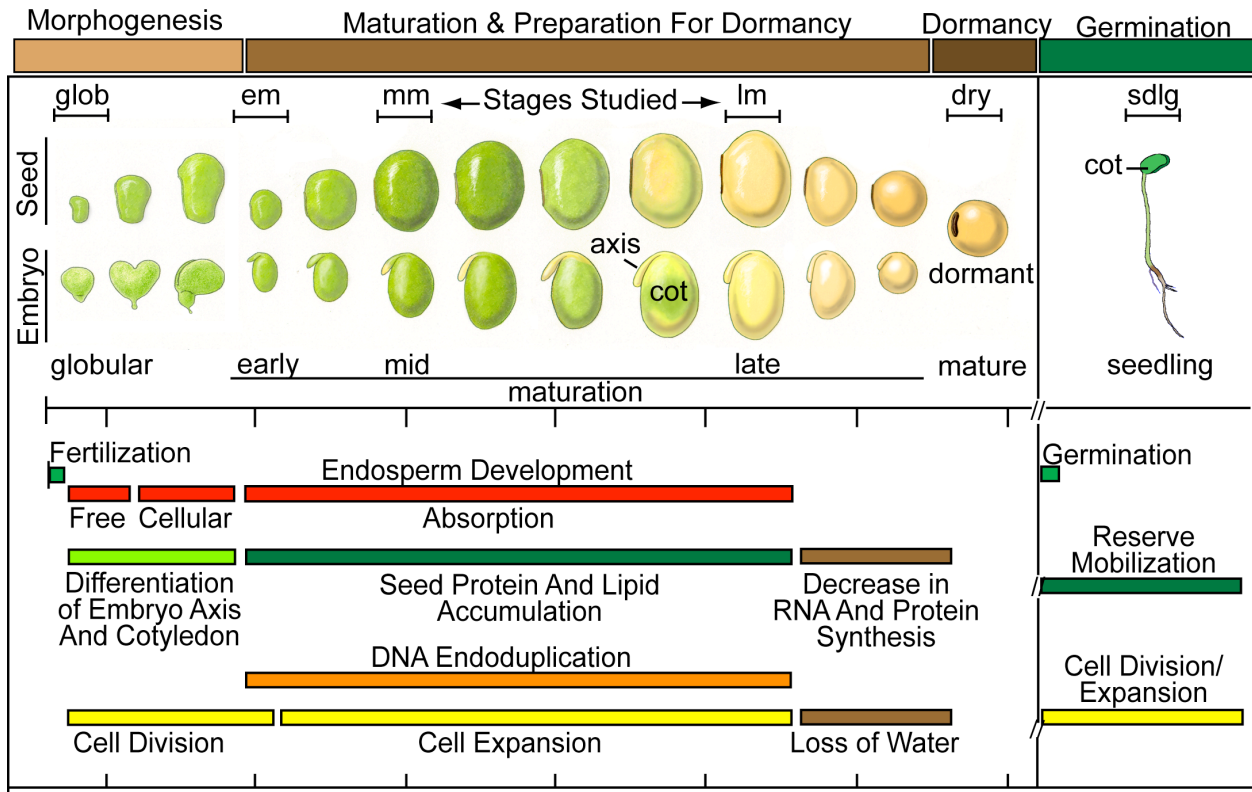
Zhang, X., and Jacobsen, S.E. (2006). Genetic analyses of DNA methyltransferases in *Arabidopsis thaliana*. *Cold Spring Harb Symp Quant Biol* 71, 439–447.

Zluvova, J., Janousek, B., and Vyskot, B. (2001). Immunohistochemical study of DNA methylation dynamics during plant development. *J Exp Bot* 52, 2265–2273.

### **Acknowledgement**

This work was supported by grants from the National Science Foundation Plant Genome Program (to R.B.G. and J.J.H.) and National Institutes of Health Training Grant in Genomic Analysis and Interpretation T32HG002536 (to B.H.L.)

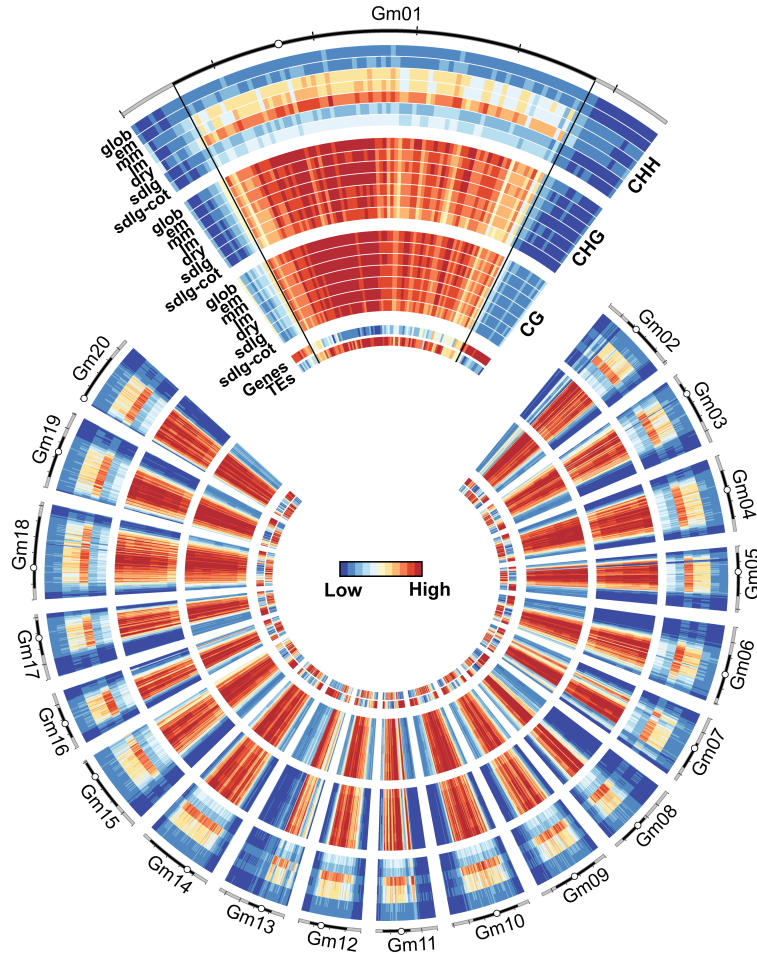
**Figure 4-1. Soybean seed development.** Cartoon of soybean seed and embryo development. Seed processes were adapted from (Goldberg et al., 1989). Seed and embryo images are not drawn to scale. axis, embryonic axis; cot, embryonic cotyledons; dry, dry seed; em, early-maturation stage seed; glob, globular stage seed; lm, late-maturation stage seed; mm, mid-maturation stage seed; sdlg, six days after imbibition seedling. See also **Table 4-1**.



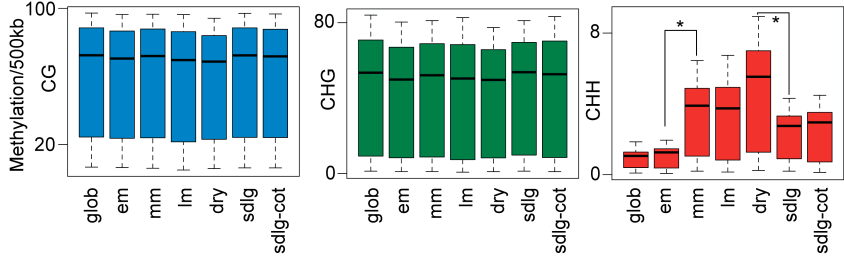
**Figure 4-2. Genome-wide methylation changes during maturation, dormancy, and post-germination.** (A) DNA methylation along the soybean genome across seed development is represented as a heatmap using Circos (Krzywinski et al., 2009). The heatmap represents the average DNA methylation level in each seed or seedlings (500 kb windows). For each DNA methylation context (CG, CHG, CHH), the highest intensities for all tracks were plotted based on the DRY seed (see Materials and Methods). For each chromosome, the black and gray bar represents the pericentromeric and euchromatic region, respectively, and the open circle indicates the relative position of the centromere. Chromosome, centromere, and pericentromere coordinates were obtained from the JGI Phytozome website (<http://phytozome.net>). Genes and TEs tracks represent densities of genes and TEs in 500kb windows along the soybean genome based on annotations from the JGI Phytozome website and the SoyTEdb database (Du et al., 2010), respectively. See **Table 4-1** for seed and seedling abbreviations. Gm, Glycine max (B) Boxplot of average genome-wide methylation level in 500kb windows across the 20 chromosomes for CG, CHG, and CHH. Each box represents the middle 50% of methylation levels. Black bar indicates the median methylation level and whiskers indicate 1.5 times the box length. \* indicates statistically significant comparisons ( $p < 0.001$ , t-test) and mean ratio  $\geq 1.5$ . For CG and CHG context, none of the stages have statistically different means and a mean ratio  $\geq 1.5$ . For the CHH context, two comparisons representing significant changes in DNA methylation are indicated. See **Table 4-S3** for all pairwise stage comparisons. Average methylation level (y-axis) are plotted 3kb upstream and 4kb downstream of the 5' and 3' ends of annotated genes (C) and TEs (D) indicated by the dashed lines. 19,170 high-confidence genes with annotated 5' and 3' UTRs (Schmutz et al., 2010) and 38,041 TEs identified in the SoyTEdb database (Du et al., 2010) were used for these plots. See also Related **Figures 4-S2 and 4-S3 and Tables 4-S2 and S3**.



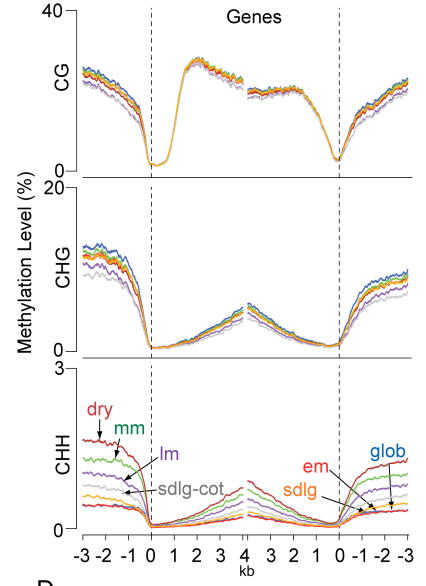
A



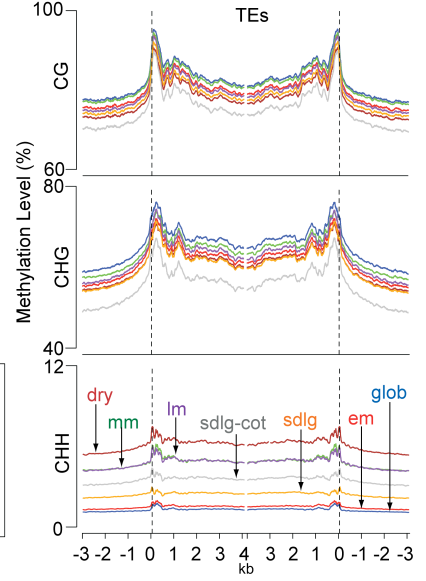
B



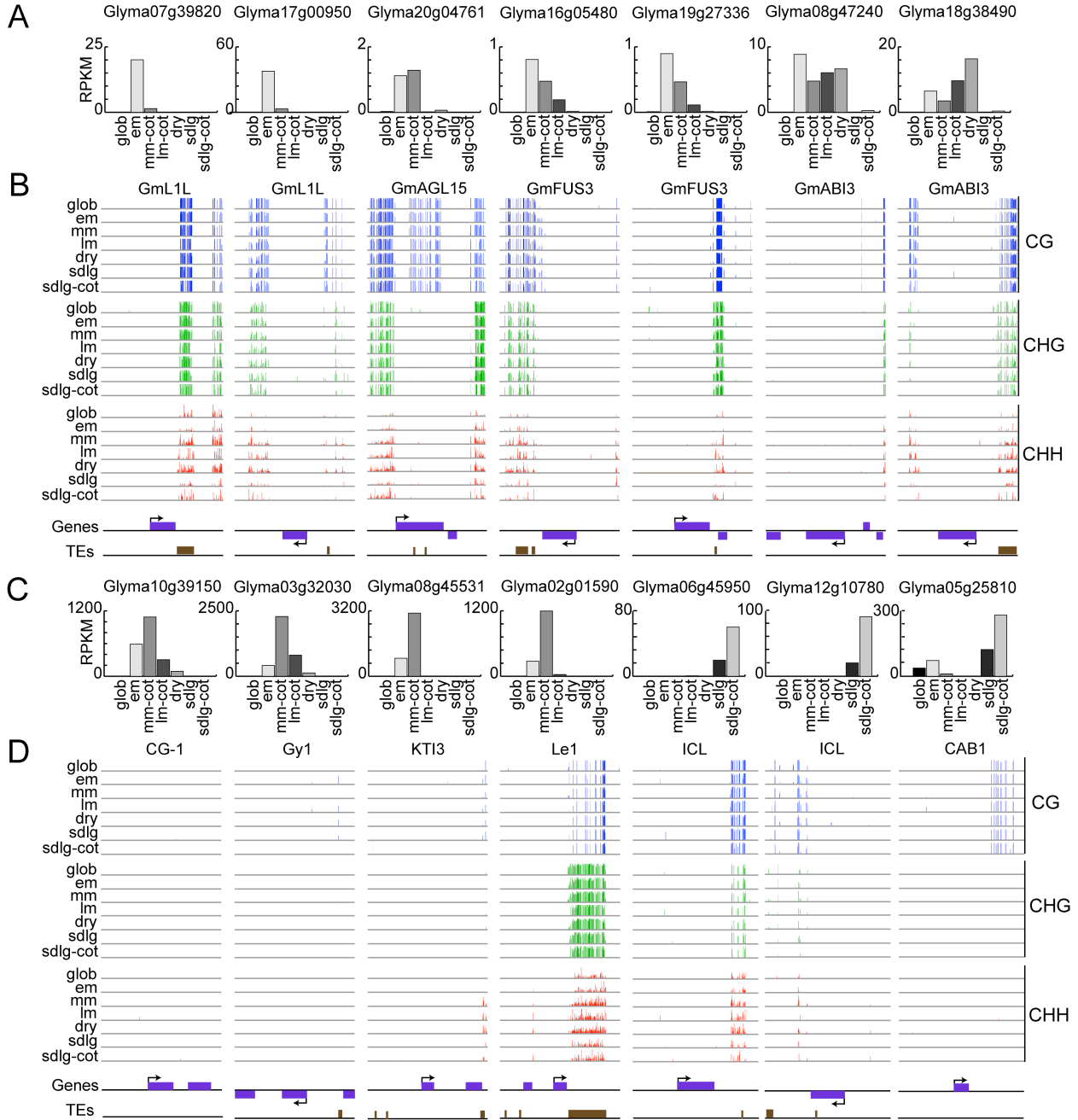
C



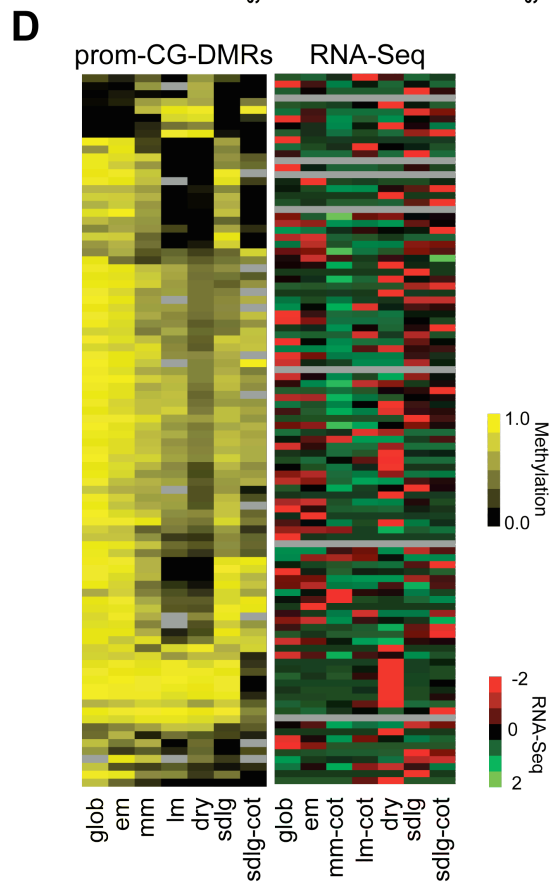
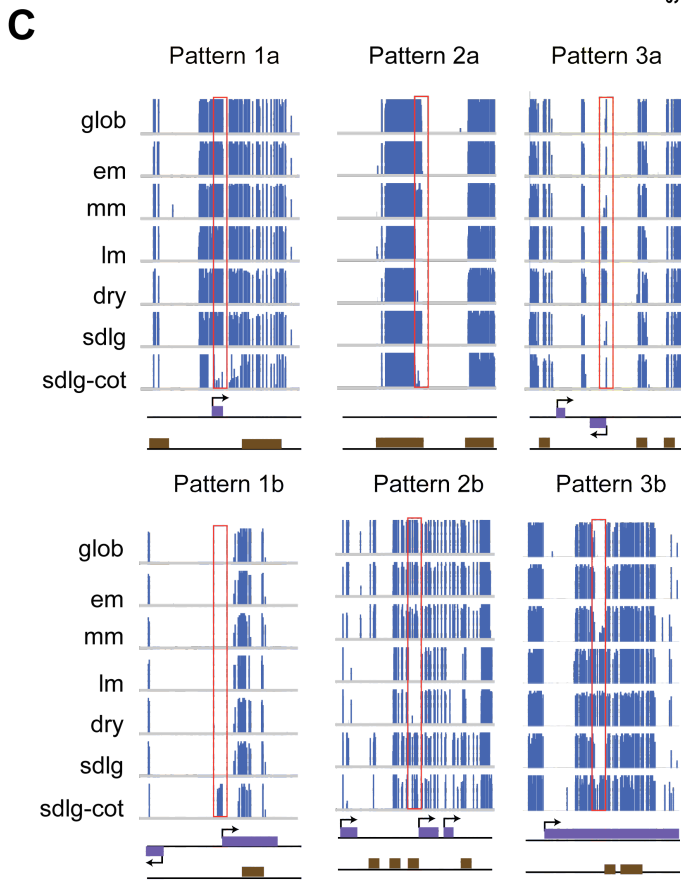
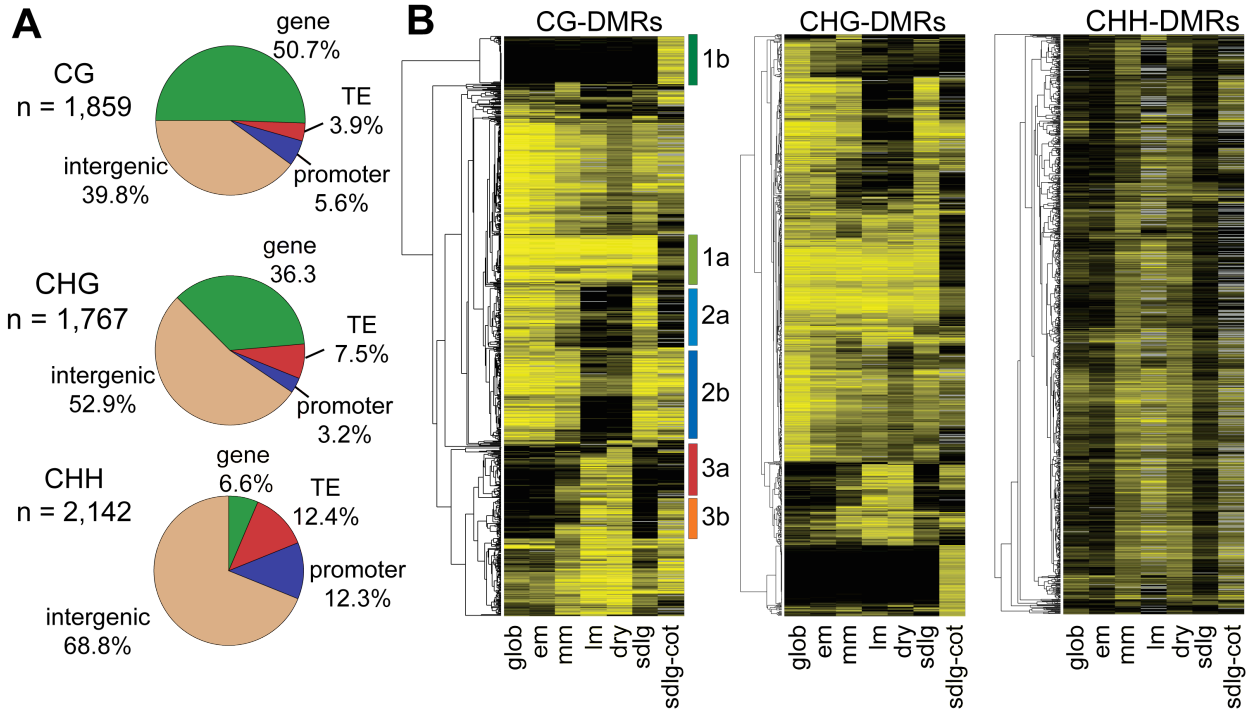
D



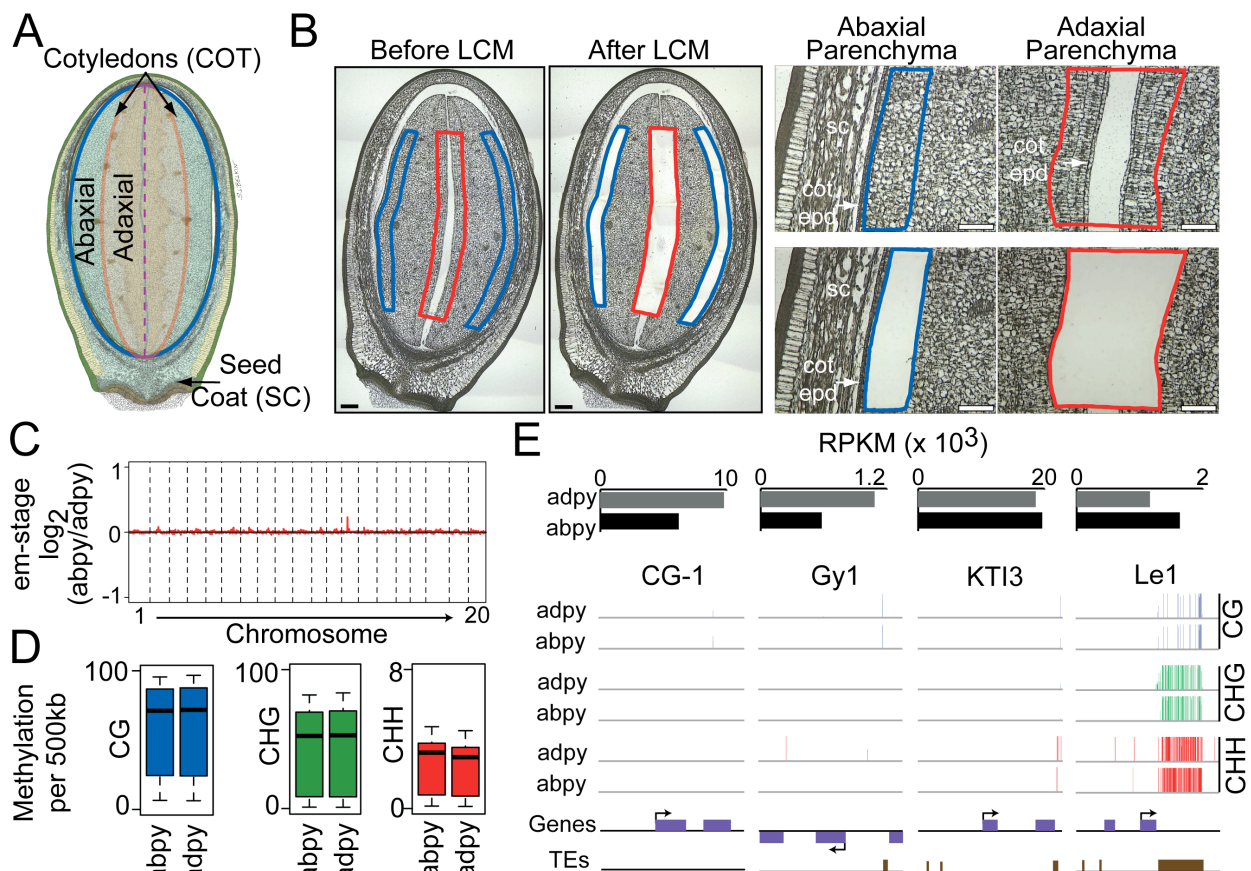
**Figure 4-3.** Seed-specific genes are not affected by the DNA methylation changes. RNA-Seq expression data (A and C) and DNA methylation data (B and D) for seed regulators (A and B) and seed storage protein genes (C and D). RNA-Seq data are represented as RPKM values. DNA methylation data are shown as barplots along the genome (B and D) where the height of the bar represents the methylation level ranging from 0 to 100. For the DNA methylation view, we obtained DNA methylation patterns from 5 kb upstream and downstream of each gene feature. Sample abbreviation is the same as in **Figure 4-2** and **Table 4-1**. ABI3, ABA-INSENSITIVE 3; AGL15, AGAMOUS-LIKE 15; CAB1, Chlorophyl A/B 1; CG-1, Conglycinin; FUS3, FUSCA3; Gy1, Glycinin 1; ICL, Isocitrate lyase; Le1, Lectin 1; L1L, LEC1-LIKE; KTI3, Kunitz Trypsin Inhibitor 3.



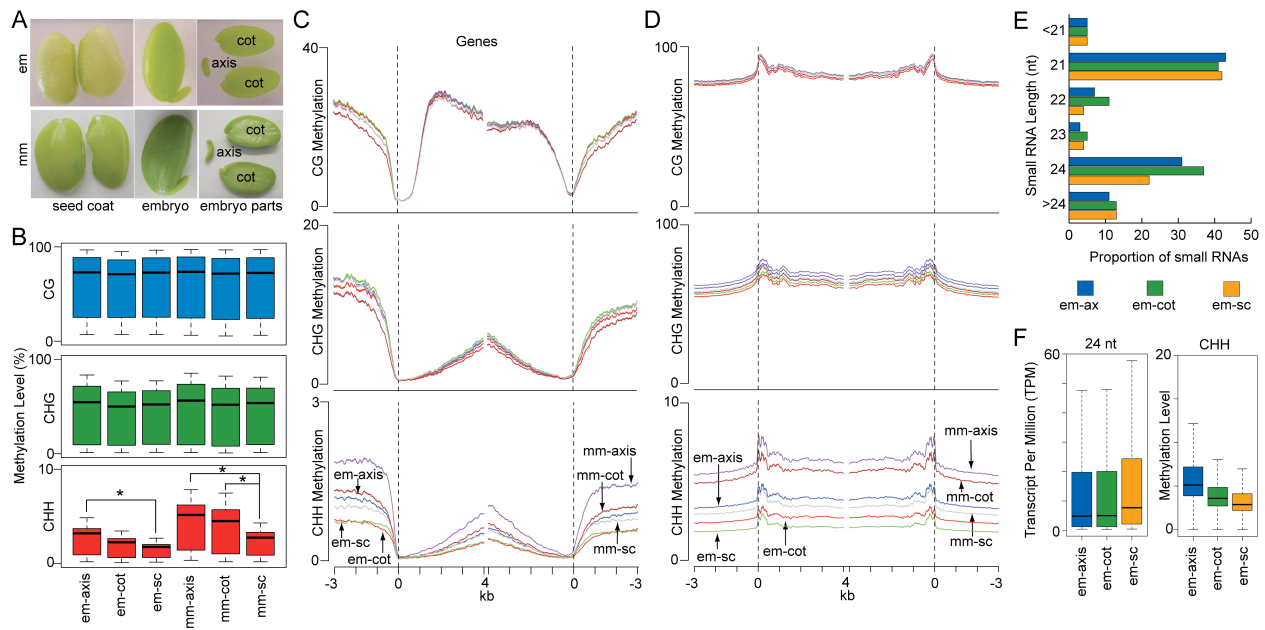
**Figure 4. Local DNA methylation changes during seed development.** (A) Proportion of differentially methylated regions (DMRs) in CG, CHG, CHH context during seed development. Gene and TE indicates DMRs overlapping with a gene or TE, respectively. Promoter is defined as 1kb upstream of a gene. Intergenic includes all DMRs that does not fit in the other categories. (B) Hierarchical clustering of average methylation level for CG-DMRs, CHG-DMRs, and CHH-DMRs across seed development. (C) Representative CG-DMR from each pattern observed in the CG-DMR clusters in (B). Purple and brown boxes represent genes and TEs, respectively. The red boxes highlight the CG-DMR region. (D) Hierarchical clustering of CG-DMRs located in the promoter (within 1kb upstream of a gene) and the standardized RNA-Seq RPKM expression of the associated downstream gene. See also Related **Table 4-S4**.



**Figure 4-5. DNA methylation is maintained in endoreduplicating cells.** LCM isolation of abaxial (ABPY) and adaxial (ADPY) parenchyma cells from embryonic cotyledons at EM stage. (A) Cross-section cartoon indicating the abaxial and adaxial regions of the embryonic cotyledons during early-maturation stage. (B) Paraffin cross section of whole abaxial (abpy) and adaxial (adpy) parenchyma regions from early-maturation stage embryonic cotyledons before and after LCM. Scale bars indicate 100 $\mu$ m. The cotyledon epidermis (cot-epd) was not captured from abpy but was captured from adpy (see Materials and Methods). (C) Uniform DNA replication along the genome in endoreduplicated cells. Log<sub>2</sub> ratio of mapped genomic DNA Illumina reads from abpy versus adpy in 500 kb windows along the chromosomes. (D) Boxplot of average genome-wide methylation level in 500kb windows in endoreduplicating (abpy) and non-endoreduplicating (adpy) cells. (E) RNA expression and DNA methylation of storage protein genes in embryonic cotyledon parenchyma tissues. See **Figure 4-2B** for plotting details, **Figure 4-3** for gene abbreviations, and also Related **Figure 4-S2** and **Table 4-S3**.



**Figure 4-6. Genome-wide methylation changes among seed parts.** (A) Seed coat (sc), embryonic axis (axis), and embryonic cotyledons (cot) isolated from early-maturation (em) and mid-maturation (mm) stage seeds. (B) Boxplot of average genome-wide methylation level in 500kb windows among different seed parts as previously described in Figure 2B. \* indicates statistically significant comparisons ( $p < 0.001$ , t-test) and mean ratio  $\geq 1.5$ . For CG and CHG context, none of the stages have statistically different means. (C-D) Methylation levels along genes (C) and TEs (D) are plotted as described previously in **Figure 4-2C and 4-2D**. (E) Distribution of small RNAs by length detected in the early maturation stage seed parts. (F) Distribution of 24 nt small RNAs mapped to TEs and TEs CHH methylation level in the seed parts. See also Related **Figure 4-S2 and Tables 4-S2 and 4-S3**.



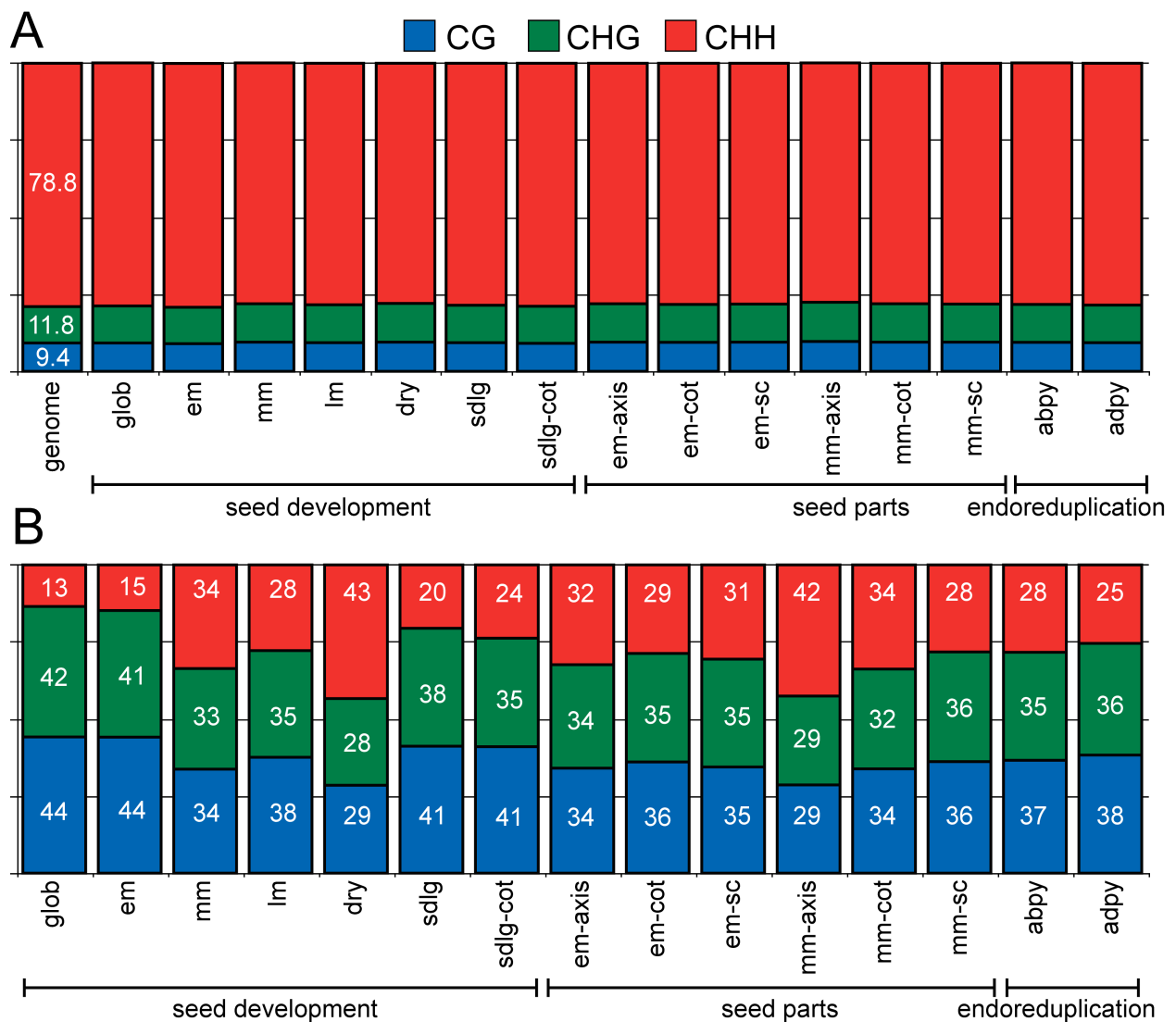
**Table 4-1. Abbreviation of Seed Stages, Organs, and Tissues in this Study.**

Description	Abbreviation
<i>Whole Seeds and Seedlings</i>	
Seeds containing globular embryo	GLOB
Seeds containing early-maturation embryo	EM
Seeds containing mid-maturation embryo	MM
Seeds containing late-maturation embryo	LM
Dry seeds	DRY
Seedlings six days after imbibition (6DAI)	SDLG
Cotyledons from 6DAI seedlings	SDLG-COT
<i>Seed Organs and Tissues</i>	
Embryonic cotyledons	COT
Embryonic axis	AX
Seed coat	SC
Seed coat parenchyma	SC-PY
Seed coat palisades	SC-PA
Seed coat hilum	SC-HI
Abaxial cotyledon parenchyma tissue	AB-COT-PY
Adaxial cotyledon parenchyma tissue	AD-COT-PY

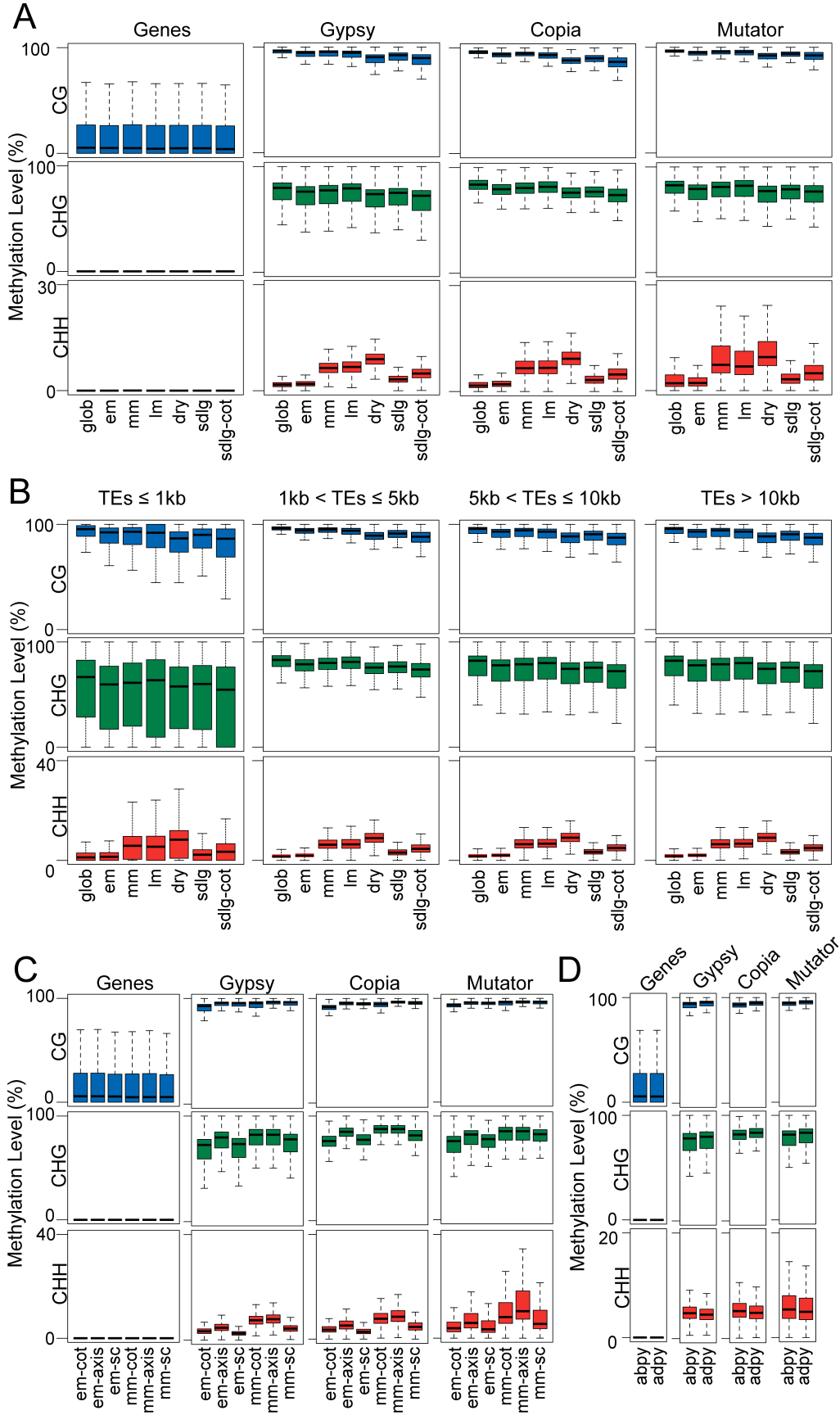


## Supplemental Figures

**Figure 4-S1. Proportion of cytosines and methylcytosines detected.** (A) Total proportions of cytosines in the CG, CHG, and CHH context detected in each seed stage, organ, or tissue. Proportion of cytosines in the CG, CHG, and CHH context for the genome was determined from the soybean genome sequence file (Schmutz et al., 2010). Proportions of cytosines for all seed samples was determined from all distinct cytosines with two or more uniquely mapped reads. (B) Proportions of methylcytosines in the different DNA context detected in each seed stage, organ, or tissue. Methylated cytosines were defined as described in the Materials and Methods.

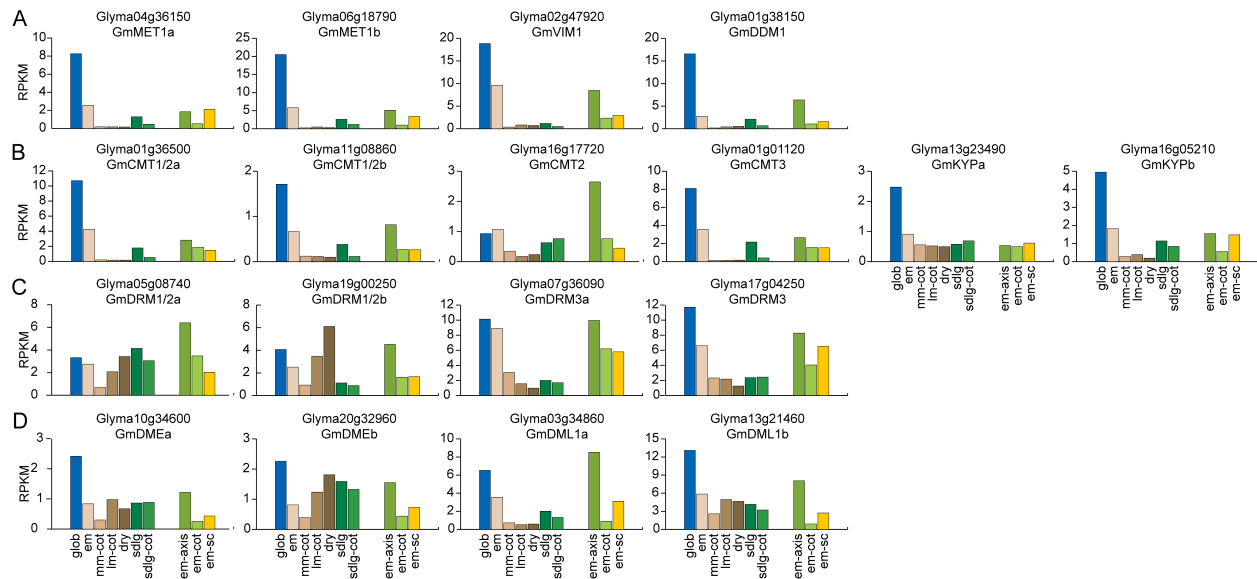


**Figure 4-S2. Average methylation distribution in genes and TEs.** Boxplots showing the average methylation level for all genes and TEs was carried out for seed developmental stages (A-B), seed region comparisons (C), and parenchyma tissues (D). For seed development (A-B), we examined average methylation based on TE family (A) and TE length (B). We calculated the bulk methylation level as  $C/(C+T)$  for all detected cytosines within each gene and TE feature (see Materials and Methods). Each box represents the middle 50% of methylation levels. Black bar indicates the median methylation level and whiskers indicate 1.5 times the box length.



### Figure 4-S3. Transcript Abundance of Methyltransferases and Seed Genes.

Transcripts for genes encoding proteins involved in DNA methylation or seed development were obtained from transcriptome (mRNA-Seq) data of whole seeds and seedlings. The bar plot represents standardized RPKM values. Transcripts from genes involved in CG (A), CHG (B), and CHH (C) methylation, and active DNA de-methylation (D).



**Table 4-S1. Sequencing summary – Related to Figure 4-1**

Sample <sup>1</sup>	Raw Reads	Mapped Reads <sup>2</sup>	Mapped Rate <sup>3</sup> (%)	Unique Reads <sup>4</sup>	Clonal Rate <sup>5</sup> (%)	Total Bases <sup>6</sup>	Total Coverage <sup>7</sup>	Cytosines Detected <sup>8</sup>	% Cytosines Detected <sup>9</sup>
<i>Seed Development</i>									
glob	194,931,579	132,930,048	68.2	111,107,273	5.4%	10,871,737,100	11.4	277,859,645	95.2%
em	191,051,920	140,126,157	73.3	108,622,354	9.3%	10,645,902,900	11.1	276,261,027	94.6%
mm	335,960,889	206,960,751	61.6	152,819,455	10.1%	14,927,414,000	15.6	280,947,779	96.2%
lm	143,738,842	99,778,977	69.4	83,160,747	8.6%	8,121,981,200	8.5	268,615,338	92.0%
dry	382,178,570	263,532,853	69.0	205,897,105	14.3%	20,142,040,900	21.1	282,871,386	96.9%
sdlg	207,092,516	133,593,946	64.5	118,094,233	9.9%	10,127,669,000	10.6	274,985,363	94.2%
sdlg-cot	150,748,539	95,096,721	63.1	81,911,885	11.4%	6,975,230,850	7.3	259,261,308	88.8%
<i>Seed Parts</i>									
em-axis	350,636,909	230,611,208	65.8	198,271,089	11.7%	19,403,572,000	20.3	283,801,796	97.2%
em-cot	393,928,206	262,074,005	66.5	226,330,829	11.1%	22,197,557,000	23.2	284,822,631	97.6%
em-sc	556,998,350	387,777,459	69.6	337,492,479	11.2%	33,048,826,500	34.6	286,851,358	98.3%
mm-axis	456,739,220	305,354,417	66.9	260,562,185	10.8%	25,463,521,200	26.7	283,948,065	97.3%
mm-cot	273,305,294	176,147,293	64.5	160,645,013	8.3%	15,683,746,500	16.4	276,519,472	94.7%
mm-sc	250,021,609	165,854,229	66.3	152,958,535	7.2%	14,921,840,200	15.6	281,293,260	96.3%
<i>Endoreduplication</i>									
abpy	196,409,743	142,061,308	72.3	132,072,468	15.1%	12,878,496,400	13.5	279,068,050	95.6%
adpy	245,286,721	176,244,437	71.9	108,813,545	10.5%	10,637,454,300	11.1	276,365,013	94.7%

1. Sample abbreviations are as described in the text.

2. Reads aligned uniquely to the reference genome using BS-Seeker (allowing for two mismatches). See Materials and Methods.

3. Percentage defined as number of mapped reads divided by total raw reads.

4. From the reads uniquely aligned to the reference genome, reads with the exact same sequence are considered clonals and all but one read is kept and is defined as a unique read.

5. Fraction of unique reads divided by mapped reads.

6. Total bases covered by all the unique reads.

7. Coverage is defined relative to the nuclear, chloroplast & mitochondrion genome estimated at 955Mb.

8. Total cytosines detected with at least one read.

9. % cytosines detected is defined relative to total theoretically detectable cytosines (291,956,640).

**Table 4-S2. Bulk Methylation Statistics For the Nuclear and Chloroplast Genome – Related to Figures 4-2 to 4-5**

Sample <sup>3</sup>	Nuclear Genome (Chr1-20) <sup>1</sup>							Chloroplast Genome <sup>2</sup>							
	Cytosines Filtered <sup>5</sup> (X Fold)	Cov <sup>6</sup> (X Fold)	mC Detected	mC (%)	C <sup>7</sup> (%)	CG <sup>7</sup> (%)	CHG <sup>7</sup> (%)	CHH <sup>7</sup> (%)	Cytosines Detected <sup>4</sup>	Cytosines Filtered <sup>5</sup> (X Fold)	Cov <sup>6</sup> (X Fold)	C <sup>7</sup> (%)	CG <sup>7</sup> (%)	CHG <sup>7</sup> (%)	CHH <sup>7</sup> (%)
<i>Seed Development</i>															
glob	246,456,146	5.7	32,809,985	13.3	10.7	57.1	36.8	0.9	28,365	20,194	90.0	0.3	0.7	0.4	0.3
em	242,283,618	5.6	31,348,541	12.9	9.6	53.2	32.5	1.0	28,389	20,406	188.4	0.6	0.9	0.6	0.6
mm	245,004,155	7.8	44,684,321	18.2	11.9	54.9	34.5	2.9	28,672	27,366	612.5	0.8	0.8	0.7	0.8
lm	218,693,101	4.3	33,412,760	15.3	11.7	53.8	33.7	3.0	28,534	22,839	247.5	0.7	0.9	0.7	0.7
dry	248,448,285	10.5	53,988,527	21.7	12.1	52.0	32.3	3.9	28,542	25,680	449.0	0.6	0.8	0.6	0.6
sdlg	235,558,095	5.3	33,565,983	14.2	10.7	53.6	32.9	1.6	28,626	27,207	507.5	1.1	1.1	1	1.1
sdlg-cot	202,827,760	3.7	26,547,035	13.1	10.0	49.4	28.4	2.2	28,669	27,126	603.5	0.8	0.8	0.8	0.8
<i>Seed Parts</i>															
em-axis	252,333,575	10.2	46,751,367	18.5	11.7	55.7	36.0	2.3	28,674	26,724	503.7	0.8	0.8	0.7	0.8
em-cot	256,167,519	11.6	44,589,632	17.4	10.6	54.5	33.1	1.7	28,665	27,010	588.5	0.7	0.8	0.7	0.7
em-sc	258,412,134	17.3	48,305,849	18.7	10.9	56.6	35.1	1.3	28,608	25,474	358.3	0.8	1	0.8	0.8
mm-axis	249,284,504	13.3	55,499,492	22.3	13.1	55.6	37.0	3.6	28,677	26,920	602.4	0.7	0.8	0.7	0.7
mm-cot	236,391,742	8.2	42,003,319	17.8	12.4	55.1	34.7	3.4	28,609	25,260	442.6	0.6	0.7	0.6	0.6
mm-sc	249,557,218	7.8	42,924,738	17.2	12.8	59.4	38.7	2.3	28,502	20,778	171.9	0.7	0.9	0.7	0.6
<i>Endoreduplication</i>															
abpy	243,798,570	6.7	40,190,338	16.5	11.5	54.6	34.7	2.4	28,515	25,327	313.3	0.8	0.9	0.8	0.8
adpy	237,678,336	5.6	36,667,728	15.4	11.2	54.7	34.7	2.2	28,485	23,497	219.3	0.9	1	0.9	0.9

1. For the nuclear genome, we only considered the 20 chromosome scaffolds assembled from the Soybean build version 1.0 (Schmutz et al., (2010) Nature). The nuclear genome is 950.1 Mb in size with 325,047,251 cytosines contained in 20 chromosomes.
2. The fully assembled chloroplast genome is 152 kb in size and contains 53,844 cytosines (GenBank Accession: DQ317523) (Saski et al., (2005) Plant Mol Biol).
3. Sample abbreviations are described in the text.
4. Total cytosines detected with at least one read.
5. Number of cytosines remaining after filtering (See Materials and Methods).
6. Coverage per DNA strand based on the filtered cytosines (See Table S1).
7. Bulk methylation for all cytosines (C), CG, CHG, and CHH, respectively, were determined as C/(C+T) where C and T represents total read with a cytosine (C) or a thymine (T). Bulk methylation calculations were based on the filtered cytosines.

**Table 4-S3. Pairwise comparison summary statistics – Related to Figures 3 to 5**

The genome was divided into 500 kb windows (n = 1,908). For each sample, we calculated the average methylation level of all cytosines within the window. Since the average methylation levels are normally distributed, we used the Student’s T-test to test the null hypothesis of no differences between the mean methylation level. \*\* denotes significant comparisons meeting the following criteria: p < 0.001 AND Ratio ≥ 1.5.

***Seed Development***

Context	Sample 1	Sample 2	Mean 1	Mean 2	Ratio	p-value	Significant?
CG	glob	glob	57.97	57.97	1.00	1	no
CG	glob	em	57.97	55.88	1.04	3.54E-02	no
CG	glob	mm	57.97	56.56	1.02	1.59E-01	no
CG	glob	lm	57.97	55.11	1.05	4.41E-03	no
CG	glob	dry	57.97	53.19	1.09	1.10E-06	no
CG	glob	sdlg	57.97	55.41	1.05	9.05E-03	no
CG	glob	sdlg-cot	57.97	52.15	1.11	2.19E-09	no
CG	em	glob	55.88	57.97	1.04	3.54E-02	no
CG	em	em	55.88	55.88	1.00	1	no
CG	em	mm	55.88	56.56	1.01	4.96E-01	no
CG	em	lm	55.88	55.11	1.01	4.39E-01	no
CG	em	dry	55.88	53.19	1.05	5.44E-03	no
CG	em	sdlg	55.88	55.41	1.01	6.28E-01	no
CG	em	sdlg-cot	55.88	52.15	1.07	1.02E-04	no
CG	mm	glob	56.56	57.97	1.02	1.59E-01	no
CG	mm	em	56.56	55.88	1.01	4.96E-01	no
CG	mm	mm	56.56	56.56	1.00	1	no
CG	mm	lm	56.56	55.11	1.03	1.50E-01	no
CG	mm	dry	56.56	53.19	1.06	5.91E-04	no
CG	mm	sdlg	56.56	55.41	1.02	2.43E-01	no
CG	mm	sdlg-cot	56.56	52.15	1.08	5.80E-06	no
CG	lm	glob	55.11	57.97	1.05	4.41E-03	no
CG	lm	em	55.11	55.88	1.01	4.39E-01	no
CG	lm	mm	55.11	56.56	1.03	1.50E-01	no
CG	lm	lm	55.11	55.11	1.00	1	no
CG	lm	dry	55.11	53.19	1.04	4.92E-02	no
CG	lm	sdlg	55.11	55.41	1.01	7.60E-01	no
CG	lm	sdlg-cot	55.11	52.15	1.06	2.28E-03	no
CG	dry	glob	53.19	57.97	1.09	1.10E-06	no

CG	dry	em	53.19	55.88	1.05	5.44E-03	no
CG	dry	mm	53.19	56.56	1.06	5.91E-04	no
CG	dry	lm	53.19	55.11	1.04	4.92E-02	no
CG	dry	dry	53.19	53.19	1.00	1	no
CG	dry	sdlg	53.19	55.41	1.04	1.99E-02	no
CG	dry	sdlg-cot	53.19	52.15	1.02	2.75E-01	no
CG	sdlg	glob	55.41	57.97	1.05	9.05E-03	no
CG	sdlg	em	55.41	55.88	1.01	6.28E-01	no
CG	sdlg	mm	55.41	56.56	1.02	2.43E-01	no
CG	sdlg	lm	55.41	55.11	1.01	7.60E-01	no
CG	sdlg	dry	55.41	53.19	1.04	1.99E-02	no
CG	sdlg	sdlg	55.41	55.41	1.00	1	no
CG	sdlg	sdlg-cot	55.41	52.15	1.06	5.72E-04	no
CG	sdlg-cot	glob	52.15	57.97	1.11	2.19E-09	no
CG	sdlg-cot	em	52.15	55.88	1.07	1.02E-04	no
CG	sdlg-cot	mm	52.15	56.56	1.08	5.80E-06	no
CG	sdlg-cot	lm	52.15	55.11	1.06	2.28E-03	no
CG	sdlg-cot	dry	52.15	53.19	1.02	2.75E-01	no
CG	sdlg-cot	sdlg	52.15	55.41	1.06	5.72E-04	no
CG	sdlg-cot	sdlg-cot	52.15	52.15	1.00	1	no
CHG	glob	glob	41.70	41.70	1.00	1	no
CHG	glob	em	41.70	38.69	1.08	8.50E-04	no
CHG	glob	mm	41.70	39.77	1.05	3.49E-02	no
CHG	glob	lm	41.70	39.23	1.06	7.28E-03	no
CHG	glob	dry	41.70	37.05	1.13	1.81E-07	no
CHG	glob	sdlg	41.70	38.19	1.09	8.28E-05	no
CHG	glob	sdlg-cot	41.70	34.71	1.20	1.42E-15	no
CHG	em	glob	38.69	41.70	1.08	8.50E-04	no
CHG	em	em	38.69	38.69	1.00	1	no
CHG	em	mm	38.69	39.77	1.03	2.23E-01	no
CHG	em	lm	38.69	39.23	1.01	5.45E-01	no
CHG	em	dry	38.69	37.05	1.04	5.65E-02	no
CHG	em	sdlg	38.69	38.19	1.01	5.57E-01	no
CHG	em	sdlg-cot	38.69	34.71	1.11	2.50E-06	no
CHG	mm	glob	39.77	41.70	1.05	3.49E-02	no
CHG	mm	em	39.77	38.69	1.03	2.23E-01	no
CHG	mm	mm	39.77	39.77	1.00	1	no
CHG	mm	lm	39.77	39.23	1.01	5.51E-01	no



CHG	mm	dry	39.77	37.05	1.07	1.87E-03	no
CHG	mm	sdlg	39.77	38.19	1.04	7.06E-02	no
CHG	mm	sdlg-cot	39.77	34.71	1.15	3.88E-09	no
CHG	lm	glob	39.23	41.70	1.06	7.28E-03	no
CHG	lm	em	39.23	38.69	1.01	5.45E-01	no
CHG	lm	mm	39.23	39.77	1.01	5.51E-01	no
CHG	lm	lm	39.23	39.23	1.00	1	no
CHG	lm	dry	39.23	37.05	1.06	1.32E-02	no
CHG	lm	sdlg	39.23	38.19	1.03	2.36E-01	no
CHG	lm	sdlg-cot	39.23	34.71	1.13	1.74E-07	no
CHG	dry	glob	37.05	41.70	1.13	1.81E-07	no
CHG	dry	em	37.05	38.69	1.04	5.65E-02	no
CHG	dry	mm	37.05	39.77	1.07	1.87E-03	no
CHG	dry	lm	37.05	39.23	1.06	1.32E-02	no
CHG	dry	dry	37.05	37.05	1.00	1	no
CHG	dry	sdlg	37.05	38.19	1.03	1.82E-01	no
CHG	dry	sdlg-cot	37.05	34.71	1.07	5.14E-03	no
CHG	sdlg	glob	38.19	41.70	1.09	8.28E-05	no
CHG	sdlg	em	38.19	38.69	1.01	5.57E-01	no
CHG	sdlg	mm	38.19	39.77	1.04	7.06E-02	no
CHG	sdlg	lm	38.19	39.23	1.03	2.36E-01	no
CHG	sdlg	dry	38.19	37.05	1.03	1.82E-01	no
CHG	sdlg	sdlg	38.19	38.19	1.00	1	no
CHG	sdlg	sdlg-cot	38.19	34.71	1.10	3.28E-05	no
CHG	sdlg-cot	glob	34.71	41.70	1.20	1.42E-15	no
CHG	sdlg-cot	em	34.71	38.69	1.11	2.50E-06	no
CHG	sdlg-cot	mm	34.71	39.77	1.15	3.88E-09	no
CHG	sdlg-cot	lm	34.71	39.23	1.13	1.74E-07	no
CHG	sdlg-cot	dry	34.71	37.05	1.07	5.14E-03	no
CHG	sdlg-cot	sdlg	34.71	38.19	1.10	3.28E-05	no
CHG	sdlg-cot	sdlg-cot	34.71	34.71	1.00	1	no
CHH	glob	glob	0.97	0.97	1.00	1	no
CHH	glob	em	0.97	1.08	1.11	8.15E-09	no
CHH	glob	mm	0.97	3.25	3.36	0	**
CHH	glob	lm	0.97	3.30	3.41	0	**
CHH	glob	dry	0.97	4.38	4.51	0	**
CHH	glob	sdlg	0.97	1.79	1.85	2.09E-170	**
CHH	glob	sdlg-cot	0.97	2.51	2.59	1.78E-279	**

CHH	em	glob	1.08	0.97	1.11	8.15E-09	no
CHH	em	em	1.08	1.08	1.00	1	no
CHH	em	mm	1.08	3.25	3.02	0	**
CHH	em	lm	1.08	3.30	3.07	0	**
CHH	em	dry	1.08	4.38	4.06	0	**
CHH	em	sdlg	1.08	1.79	1.66	4.97E-124	**
CHH	em	sdlg-cot	1.08	2.51	2.33	1.43E-239	**
CHH	mm	glob	3.25	0.97	3.36	0	**
CHH	mm	em	3.25	1.08	3.02	0	**
CHH	mm	mm	3.25	3.25	1.00	1	no
CHH	mm	lm	3.25	3.30	1.02	4.76E-01	no
CHH	mm	dry	3.25	4.38	1.34	1.25E-42	no
CHH	mm	sdlg	3.25	1.79	1.82	1.95E-153	**
CHH	mm	sdlg-cot	3.25	2.51	1.30	1.23E-34	no
CHH	lm	glob	3.30	0.97	3.41	0	**
CHH	lm	em	3.30	1.08	3.07	0	**
CHH	lm	mm	3.30	3.25	1.02	4.76E-01	no
CHH	lm	lm	3.30	3.30	1.00	1	no
CHH	lm	dry	3.30	4.38	1.32	5.07E-37	no
CHH	lm	sdlg	3.30	1.79	1.84	1.05E-143	**
CHH	lm	sdlg-cot	3.30	2.51	1.31	3.14E-35	no
CHH	dry	glob	4.38	0.97	4.51	0	**
CHH	dry	em	4.38	1.08	4.06	0	**
CHH	dry	mm	4.38	3.25	1.34	1.25E-42	no
CHH	dry	lm	4.38	3.30	1.32	5.07E-37	no
CHH	dry	dry	4.38	4.38	1.00	1	no
CHH	dry	sdlg	4.38	1.79	2.44	2.08E-249	**
CHH	dry	sdlg-cot	4.38	2.51	1.74	1.36E-122	**
CHH	sdlg	glob	1.79	0.97	1.85	2.09E-170	**
CHH	sdlg	em	1.79	1.08	1.66	4.97E-124	**
CHH	sdlg	mm	1.79	3.25	1.82	1.95E-153	**
CHH	sdlg	lm	1.79	3.30	1.84	1.05E-143	**
CHH	sdlg	dry	1.79	4.38	2.44	2.08E-249	**
CHH	sdlg	sdlg	1.79	1.79	1.00	1	no
CHH	sdlg	sdlg-cot	1.79	2.51	1.40	8.35E-55	no
CHH	sdlg-cot	glob	2.51	0.97	2.59	1.78E-279	**
CHH	sdlg-cot	em	2.51	1.08	2.33	1.43E-239	**
CHH	sdlg-cot	mm	2.51	3.25	1.30	1.23E-34	no

CHH	sdlg-cot	lm	2.51	3.30	1.31	3.14E-35	no
CHH	sdlg-cot	dry	2.51	4.38	1.74	1.36E-122	**
CHH	sdlg-cot	sdlg	2.51	1.79	1.40	8.35E-55	no
CHH	sdlg-cot	sdlg-cot	2.51	2.51	1.00	1	no

### Seed Parts

Context	Sample 1	Sample 2	Mean 1	Mean 2	Ratio	p-value	Significant
CG	em-axis	em-axis	57.71	57.71	1.00	1	no
CG	em-axis	em-cot	57.71	56.52	1.02	2.24E-01	no
CG	em-axis	em-sc	57.71	57.93	1.00	8.30E-01	no
CG	em-axis	mm-axis	57.71	56.94	1.01	4.43E-01	no
CG	em-axis	mm-cot	57.71	56.14	1.03	1.19E-01	no
CG	em-axis	mm-sc	57.71	58.91	1.02	2.33E-01	no
CG	em-cot	em-axis	56.52	57.71	1.02	2.24E-01	no
CG	em-cot	em-cot	56.52	56.52	1.00	1	no
CG	em-cot	em-sc	56.52	57.93	1.02	1.49E-01	no
CG	em-cot	mm-axis	56.52	56.94	1.01	6.79E-01	no
CG	em-cot	mm-cot	56.52	56.14	1.01	6.99E-01	no
CG	em-cot	mm-sc	56.52	58.91	1.04	1.51E-02	no
CG	em-sc	em-axis	57.93	57.71	1.00	8.30E-01	no
CG	em-sc	em-cot	57.93	56.52	1.02	1.49E-01	no
CG	em-sc	em-sc	57.93	57.93	1.00	1	no
CG	em-sc	mm-axis	57.93	56.94	1.02	3.25E-01	no
CG	em-sc	mm-cot	57.93	56.14	1.03	7.42E-02	no
CG	em-sc	mm-sc	57.93	58.91	1.02	3.23E-01	no
CG	mm-axis	em-axis	56.94	57.71	1.01	4.43E-01	no
CG	mm-axis	em-cot	56.94	56.52	1.01	6.79E-01	no
CG	mm-axis	em-sc	56.94	57.93	1.02	3.25E-01	no
CG	mm-axis	mm-axis	56.94	56.94	1.00	1	no
CG	mm-axis	mm-cot	56.94	56.14	1.01	4.38E-01	no
CG	mm-axis	mm-sc	56.94	58.91	1.03	5.22E-02	no
CG	mm-cot	em-axis	56.14	57.71	1.03	1.19E-01	no
CG	mm-cot	em-cot	56.14	56.52	1.01	6.99E-01	no
CG	mm-cot	em-sc	56.14	57.93	1.03	7.42E-02	no
CG	mm-cot	mm-axis	56.14	56.94	1.01	4.38E-01	no
CG	mm-cot	mm-cot	56.14	56.14	1.00	1	no
CG	mm-cot	mm-sc	56.14	58.91	1.05	6.10E-03	no
CG	mm-sc	em-axis	58.91	57.71	1.02	2.33E-01	no

CG	mm-sc	em-cot	58.91	56.52	1.04	1.51E-02	no
CG	mm-sc	em-sc	58.91	57.93	1.02	3.23E-01	no
CG	mm-sc	mm-axis	58.91	56.94	1.03	5.22E-02	no
CG	mm-sc	mm-cot	58.91	56.14	1.05	6.10E-03	no
CG	mm-sc	mm-sc	58.91	58.91	1.00	1	no
CHG	em-axis	em-axis	41.75	41.75	1.00	1	no
CHG	em-axis	em-cot	41.75	38.33	1.09	1.30E-04	no
CHG	em-axis	em-sc	41.75	39.61	1.05	1.70E-02	no
CHG	em-axis	mm-axis	41.75	42.28	1.01	5.81E-01	no
CHG	em-axis	mm-cot	41.75	39.69	1.05	2.56E-02	no
CHG	em-axis	mm-sc	41.75	42.26	1.01	5.82E-01	no
CHG	em-cot	em-axis	38.33	41.75	1.09	1.30E-04	no
CHG	em-cot	em-cot	38.33	38.33	1.00	1	no
CHG	em-cot	em-sc	38.33	39.61	1.03	1.31E-01	no
CHG	em-cot	mm-axis	38.33	42.28	1.10	1.48E-05	no
CHG	em-cot	mm-cot	38.33	39.69	1.04	1.22E-01	no
CHG	em-cot	mm-sc	38.33	42.26	1.10	7.71E-06	no
CHG	em-sc	em-axis	39.61	41.75	1.05	1.70E-02	no
CHG	em-sc	em-cot	39.61	38.33	1.03	1.31E-01	no
CHG	em-sc	em-sc	39.61	39.61	1.00	1	no
CHG	em-sc	mm-axis	39.61	42.28	1.07	3.51E-03	no
CHG	em-sc	mm-cot	39.61	39.69	1.00	9.31E-01	no
CHG	em-sc	mm-sc	39.61	42.26	1.07	2.64E-03	no
CHG	mm-axis	em-axis	42.28	41.75	1.01	5.81E-01	no
CHG	mm-axis	em-cot	42.28	38.33	1.10	1.48E-05	no
CHG	mm-axis	em-sc	42.28	39.61	1.07	3.51E-03	no
CHG	mm-axis	mm-axis	42.28	42.28	1.00	1	no
CHG	mm-axis	mm-cot	42.28	39.69	1.07	5.93E-03	no
CHG	mm-axis	mm-sc	42.28	42.26	1.00	9.84E-01	no
CHG	mm-cot	em-axis	39.69	41.75	1.05	2.56E-02	no
CHG	mm-cot	em-cot	39.69	38.33	1.04	1.22E-01	no
CHG	mm-cot	em-sc	39.69	39.61	1.00	9.31E-01	no
CHG	mm-cot	mm-axis	39.69	42.28	1.07	5.93E-03	no
CHG	mm-cot	mm-cot	39.69	39.69	1.00	1	no
CHG	mm-cot	mm-sc	39.69	42.26	1.06	4.69E-03	no
CHG	mm-sc	em-axis	42.26	41.75	1.01	5.82E-01	no
CHG	mm-sc	em-cot	42.26	38.33	1.10	7.71E-06	no
CHG	mm-sc	em-sc	42.26	39.61	1.07	2.64E-03	no

CHG	mm-sc	mm-axis	42.26	42.28	1.00	9.84E-01	no
CHG	mm-sc	mm-cot	42.26	39.69	1.06	4.69E-03	no
CHG	mm-sc	mm-sc	42.26	42.26	1.00	1	no
CHH	em-axis	em-axis	2.57	2.57	1.00	1	no
CHH	em-axis	em-cot	2.57	1.82	1.41	3.13E-68	no
CHH	em-axis	em-sc	2.57	1.44	1.79	9.34E-172	**
CHH	em-axis	mm-axis	2.57	4.03	1.57	1.01E-99	**
CHH	em-axis	mm-cot	2.57	3.76	1.46	4.15E-70	no
CHH	em-axis	mm-sc	2.57	2.43	1.06	2.53E-03	no
CHH	em-cot	em-axis	1.82	2.57	1.41	3.13E-68	no
CHH	em-cot	em-cot	1.82	1.82	1.00	1	no
CHH	em-cot	em-sc	1.82	1.44	1.27	1.04E-35	no
CHH	em-cot	mm-axis	1.82	4.03	2.22	1.43E-238	**
CHH	em-cot	mm-cot	1.82	3.76	2.07	3.47E-195	**
CHH	em-cot	mm-sc	1.82	2.43	1.34	2.30E-51	no
CHH	em-sc	em-axis	1.44	2.57	1.79	9.34E-172	**
CHH	em-sc	em-cot	1.44	1.82	1.27	1.04E-35	no
CHH	em-sc	em-sc	1.44	1.44	1.00	1	no
CHH	em-sc	mm-axis	1.44	4.03	2.80	0	**
CHH	em-sc	mm-cot	1.44	3.76	2.62	1.39E-285	**
CHH	em-sc	mm-sc	1.44	2.43	1.69	8.51E-153	**
CHH	mm-axis	em-axis	4.03	2.57	1.57	1.01E-99	**
CHH	mm-axis	em-cot	4.03	1.82	2.22	1.43E-238	**
CHH	mm-axis	em-sc	4.03	1.44	2.80	0	**
CHH	mm-axis	mm-axis	4.03	4.03	1.00	1	no
CHH	mm-axis	mm-cot	4.03	3.76	1.07	1.01E-03	no
CHH	mm-axis	mm-sc	4.03	2.43	1.66	5.79E-123	**
CHH	mm-cot	em-axis	3.76	2.57	1.46	4.15E-70	no
CHH	mm-cot	em-cot	3.76	1.82	2.07	3.47E-195	**
CHH	mm-cot	em-sc	3.76	1.44	2.62	1.39E-285	**
CHH	mm-cot	mm-axis	3.76	4.03	1.07	1.01E-03	no
CHH	mm-cot	mm-cot	3.76	3.76	1.00	1	no
CHH	mm-cot	mm-sc	3.76	2.43	1.55	6.22E-90	**
CHH	mm-sc	em-axis	2.43	2.57	1.06	2.53E-03	no
CHH	mm-sc	em-cot	2.43	1.82	1.34	2.30E-51	no
CHH	mm-sc	em-sc	2.43	1.44	1.69	8.51E-153	**
CHH	mm-sc	mm-axis	2.43	4.03	1.66	5.79E-123	**
CHH	mm-sc	mm-cot	2.43	3.76	1.55	6.22E-90	**

CHH	mm-sc	mm-sc	2.43	2.43	1.00	1	no
-----	-------	-------	------	------	------	---	----

***Endoreduplication***

Context	Sample 1	Sample 2	Mean 1	Mean 2	Ratio	p-value	Significant
CG	abpy	adpy	56.43	56.81	1.01	6.95E-01	no
CHG	abpy	adpy	40.27	40.68	1.01	6.55E-01	no
CHH	abpy	adpy	2.61	2.45	1.07	1.16E-03	no

**Table 4-S4. List of Differentially Methylated Regions (DMRs) – Related to Figure 4-4  
(Partial List)**

Chr	DMR Start	DMR End	DMR Size	Classification	Feature	Description
1	327501	327800	300	Intergenic	NA	
1	742201	742400	200	Intergenic	NA	
1	837901	838200	300	Intergenic	NA	
1	1050601	1050900	300	Gene	Glyma01g01440;	NA
1	1385601	1385800	200	Intergenic	NA	
1	1524901	1525100	200	Promoter	Glyma01g01960;	Syntaxin/t-SNARE_family_protein
1	1793801	1794000	200	Intergenic	NA	
1	2162301	2162500	200	Intergenic	NA	
1	2184301	2184600	300	Intergenic	NA	
1	2185801	2186000	200	Intergenic	NA	
1	2238101	2238400	300	Intergenic	NA	
1	2282001	2282200	200	Intergenic	NA	
1	2715601	2715800	200	Intergenic	NA	
1	2823101	2823400	300	Promoter	Glyma01g03310;	Peroxidase_superfamily_protein
1	3000301	3000500	200	Gene	Glyma01g03490;	NSP-interacting_kinase_3
1	3050301	3051000	700	Intergenic	NA	
1	3317901	3318200	300	Intergenic	NA	
1	3560801	3561100	300	Intergenic	NA	
1	3642701	3642900	200	Intergenic	NA	
1	3643501	3643800	300	Intergenic	NA	
1	3774001	3774200	200	Intergenic	NA	

1	3854001	3854200	200	Intergenic	NA	
1	3882001	3882300	300	Intergenic	NA	
1	4025001	4025200	200	Intergenic	NA	
1	4077201	4077400	200	Intergenic	NA	
1	4326001	4326400	400	Intergenic	NA	
1	4484101	4484300	200	Intergenic	NA	
1	4610401	4610700	300	Intergenic	NA	
1	4767201	4768000	800	Intergenic	NA	
1	5273201	5273400	200	Intergenic	NA	
1	5306401	5306700	300	Gene	Glyma01g05470;	Fatty_acid/ sphingolipid_desatur ase
1	5456401	5456600	200	Intergenic	NA	
1	5597501	5597700	200	TE	RLG_Gmr216_G m1-1;	
1	5721101	5721400	300	Intergenic	NA	
1	5816801	5817100	300	Intergenic	NA	
1	6301701	6301900	200	Intergenic	NA	
1	6547201	6547400	200	Intergenic	NA	
1	6746601	6746800	200	Intergenic	NA	
1	7013001	7013200	200	Intergenic	NA	
1	7090201	7090400	200	Intergenic	NA	
1	7666801	7667100	300	Intergenic	NA	
1	7726801	7727100	300	Intergenic	NA	
1	8019701	8019900	200	Intergenic	NA	
1	8256301	8256600	300	Intergenic	NA	
1	8575701	8576000	300	Gene	Glyma01g07770;	6- phosphogluconolact onase_1
1	8580501	8580800	300	Intergenic	NA	



1	8583301	8583600	300	Intergenic	NA	
1	9051701	9051900	200	Intergenic	NA	
1	9248501	9248800	300	Intergenic	NA	
1	9250901	9251200	300	Intergenic	NA	
1	9723301	9723500	200	Intergenic	NA	
1	9768601	9768800	200	Intergenic	NA	
1	11129801	11130100	300	Intergenic	NA	
1	11344201	11344400	200	Intergenic	NA	
1	11653201	11653400	200	Intergenic	NA	
1	11697001	11697200	200	Intergenic	NA	
1	11706001	11706500	500	Intergenic	NA	
1	11999201	11999400	200	Gene	Glyma01g09660;	Phosphoribulokinase / Uridine_kinase_fam ily
1	12091601	12091800	200	Intergenic	NA	
1	12420701	12420900	200	Intergenic	NA	
1	12530201	12530400	200	Gene	Glyma01g09900;	transferases;folic_acid_binding

## **Extended Analysis**

### **Simulation data for mappability**

Since the soybean genome is an ancient polyploid with > 60% of the genome represented by repetitive sequences (Schmutz et al., 2010), we asked what fraction of the nuclear genome sequences could we detect uniquely with 50 to 100 bp sequencing reads. Using a sliding window approach, we generated > 950 million 100-bp reads with a 99-bp overlap covering the entire genome. We next collapsed and removed identical redundant reads leaving 805 million (~85%) reads that are unique. The 805 million reads were aligned to the genome using BS-Seeker to mimic our data processing pipeline allowing for no mismatch. Approximately 782 million reads aligned to the genome uniquely covering 857 million bases or 90% of the nuclear genome including > 325 million cytosines or 90% of the genomic cytosines. These results suggest that although soybean is an ancient polyploid with a large repeat-rich genome, most of the sequences have diverged significantly and can be distinguished clearly from 100 bp reads indicating that whole genome BS sequencing can interrogate most, if not all, of the genome sequences. Since the unmethylated chloroplast genome is commonly used as a control for determining non-conversion rates from BS-treatment, we asked what fraction of the chloroplast genome can be detected uniquely from 100 bp reads. The soybean chloroplast genome is ~ 152 kb in size and consists of two inverted identical ~25kb repeat sequences and an 83kb and a 17kb single copy region. Using the same approach as the nuclear genome, we obtained > 99% of the 100bp reads are unique. However, when we aligned the unique reads to the chloroplast genome using BS-Seeker, ~58% of the reads aligned uniquely while 42% of the reads mapped to multiple locations in the chloroplast genome. Therefore, we are able to detect ~90kb covering 58% of the chloroplast genome and can detect ~14,000 cytosines or 53% of the total genomic cytosines.

### **Analysis of conversion and inappropriate conversion from the BS-DNA-Seq Libraries**

During bisulfite conversion experiments, there are two possible conversion errors that might occur: non-conversion (false positive) and inappropriate-conversion (false negative) errors. In non-conversion errors, unmethylated cytosines are not converted to thymine and would erroneously appear as a methylated cytosine in the final sequencing read. On the other hand, inappropriate conversion is derived from the conversion of methylated cytosines to thymine by the bisulfite treatment, usually due to long exposure to the bisulfite reagents. To assess the extent of non-conversion during bisulfite treatment, we spiked in unmethylated lambda DNA into each genomic DNA sample during the library preparation step (see **Materials and Methods**). We obtained at least 300X coverage of the lambda genome and detected up to 99.98% of the genomic cytosines, representing comprehensive coverage of the lambda genome. Overall, we obtained good conversion efficiency from 99.31% to 99.69%, similar to conversion rates from others using similar kits (Lister et al., 2009). For the assessment of inappropriate conversion, we spiked in methylated lambda DNA into one of the genomic DNA samples. The methylated lambda DNA was generated by treating unmethylated lambda DNA with CpG methyltransferase. We obtained ~150X coverage of the lambda genome, detecting 99.91% of all genomic lambda cytosines. We detected 92.6% of CG cytosines as methylated suggesting that 7.4% of the CG cytosines were inappropriately converted. However, we cannot rule out the possibility that our treatment of unmethylated lambda DNA with CpG methyltransferase was not 100 percent effective, indicating that the 7.4% might be a vast overestimate of inappropriate conversion. Genereux et al. observed 6.1 % inappropriate conversion under a 200 min BS treatment and under a high molarity and temperature condition (highMT) (Genereux et al., 2008). Taken together, these results suggest that our BS treatment reactions are sufficient at converting unmethylated cytosines to thymine while reducing inappropriate conversion of methylated cytosines to thymine.

### **Comparison of Dry Seed Against Dry Seed As a Control For DMR Detection**

To determine appropriate parameters for defining differentially methylated regions (DMRs), we processed four lanes of sequence data from DRY seeds, dividing the data into two sets, A and B, each containing two lanes of sequence data. The comparison between sets A and B will help determine the number of false positives and appropriate cut-offs to apply to distinct DNA samples. Under our current parameters: At least 8 informative cytosines, absolute average methylation difference  $\geq 15\%$ , and absolute z-score  $\geq 10$  (CG and CHG) or  $z \geq 16$  (CHH), we obtained 11, 4, and 3 DMRs in the CG, CHG, and CHH context, respectively, indicating a false positive rate of  $1.1 \times 10^{-4}\%$  (CG),  $4.2 \times 10^{-5}\%$  (CHG), and  $3.2 \times 10^{-5}\%$  (CHH). These results suggests that the DMRs obtained in our analysis between different tissues or seed stages will be significantly enriched for true positives.

### **Supplemental References**

Genereux, D.P., Johnson, W.C., Burden, A.F., Stöger, R., and Laird, C.D. (2008). Errors in the bisulfite conversion of DNA: modulating inappropriate- and failed-conversion frequencies. *Nucleic Acids Res* 36, e150.

Lister, R., Pelizzola, M., Downen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462, 315–322.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178-183.

## APPENDIX A

Belmonte, M.F., Kirkbride, R.C., Stone, S.L., Pelletier, J.M., Bui, A.Q., Yeung, E.C., Hashimoto, M., Fei, J., Harada, C.M., Munoz, M.D., **Le, B.H.**, Drews, G.N., Brady, S.M., Goldberg, R.B., Harada, J.J. (2013). **Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed.** PNAS 110, E435–E444

# Comprehensive developmental profiles of gene activity in regions and subregions of the *Arabidopsis* seed

Mark F. Belmonte<sup>a,1,2</sup>, Ryan C. Kirkbride<sup>a,1</sup>, Sandra L. Stone<sup>a,3</sup>, Julie M. Pelletier<sup>a</sup>, Anhthu Q. Bui<sup>b,4</sup>, Edward C. Yeung<sup>c</sup>, Meryl Hashimoto<sup>a</sup>, Jiong Fei<sup>a</sup>, Corey M. Harada<sup>a</sup>, Matthew D. Munoz<sup>a,5</sup>, Brandon H. Le<sup>p</sup>, Gary N. Drews<sup>d</sup>, Siobhan M. Brady<sup>a,e</sup>, Robert B. Goldberg<sup>b,6</sup>, and John J. Harada<sup>a,6</sup>

<sup>a</sup>Department of Plant Biology and <sup>e</sup>Genome Center, University of California, Davis, CA 95616; <sup>b</sup>Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, CA 90095; <sup>c</sup>Department of Biological Sciences, University of Calgary, Calgary, AB, Canada T2N 1N4; and <sup>d</sup>Department of Biology, University of Utah, Salt Lake City, UT 84112

Contributed by Robert B. Goldberg, December 20, 2012 (sent for review December 7, 2012)

**Seeds are complex structures that consist of the embryo, endosperm, and seed-coat regions that are of different ontogenetic origins, and each region can be further divided into morphologically distinct subregions. Despite the importance of seeds for food, fiber, and fuel globally, little is known of the cellular processes that characterize each subregion or how these processes are integrated to permit the coordinated development of the seed. We profiled gene activity genome-wide in every organ, tissue, and cell type of *Arabidopsis* seeds from fertilization through maturity. The resulting mRNA datasets offer the most comprehensive description of gene activity in seeds with high spatial and temporal resolution, providing unique insights into the function of understudied seed regions. Global comparisons of mRNA populations reveal unexpected overlaps in the functional identities of seed subregions. Analyses of coexpressed gene sets suggest that processes that regulate seed size and filling are coordinated across several subregions. Predictions of gene regulatory networks based on the association of transcription factors with enriched DNA sequence motifs upstream of coexpressed genes identify regulators of seed development. These studies emphasize the utility of these datasets as an essential resource for the study of seed biology.**

laser-capture microdissection | mRNA localization | transcriptome

The significance of seeds is reflected by their relevance to diverse biological areas. Evolutionarily, the ability of flowering plants to make seeds has conferred significant selective advantages, accounting, in part, for their dominance among the Plantae. The seed habit facilitates fertilization in nonaqueous environments, provides protection and nutrients for the developing embryo, and permits the embryo to remain quiescent until conditions are favorable for seedling development (1). Seeds are a key to global food security, because they account for the large majority of calories consumed by humans. An estimated 70–100% more food will need to be produced worldwide by 2050 without an appreciable increase in arable land and despite global climate change (2). A detailed understanding of seed development may enable the design of cogent strategies to enhance seed quality and yield.

The developmental significance of seeds is that they are complex yet elegant structures, consisting of embryo, endosperm, and seed-coat regions that are each divided into subregions (3). The complexity of the seed originates with its precursor, the ovule, which consists of the female gametophyte embedded within integument layers. Seed development is initiated with the fusion of the egg and central cells of the female gametophyte with two sperm cells from the pollen tube. This double fertilization, unique to flowering plants, produces the progenitors of the embryo and endosperm regions of the seed, respectively. Patterning and morphological differentiation occur in the embryo and endosperm regions early in seed development, during the morphogenesis phase. In many plants, including *Arabidopsis*, the embryo undergoes stereotypic cell-division patterns, differentiating into the embryo proper that becomes the body of the vegetative plant and the suspensor, an ephemeral structure that

serves as a conduit between the embryo proper and the seed coat (Figs. 1 *A–F*). The primary endosperm cell undergoes nuclear but not cell divisions, and nuclei migrate to form three subregions: micropylar, which is nearest the young embryo; peripheral, in the center of the endosperm region; and chalazal, at the pole opposite to the embryo. Cellularization of the endosperm proceeds in a wave-like manner from the micropylar to chalazal end (4). Ovule integument cells divide and differentiate into the distinct cell types of the seed coat that envelope the embryo and endosperm. Late in seed development during the maturation phase, the embryo accumulates storage macromolecules and becomes tolerant of desiccation. Although development of these subregions has been well-characterized morphologically, little is known of the cellular processes that occur in these subregions or how the development of the subregions is coordinated within the context of seed development.

A key to dissecting seed development is to obtain an integrated understanding of gene activity, and therefore the cellular processes that occur in seed regions throughout development.

## Significance

**Seeds are complex structures that are comprised of the embryo, endosperm, and seed coat. Despite their importance for food, fiber, and fuel, the cellular processes that characterize different regions of the seed are not known. We profiled gene activity genome-wide in every organ, tissue, and cell type of *Arabidopsis* seeds from fertilization through maturity. The resulting mRNA datasets provide unique insights into the cellular processes that occur in understudied seed regions, revealing unexpected overlaps in the functional identities of seed regions and enabling predictions of gene regulatory networks. This dataset is an essential resource for studies of seed biology.**

Author contributions: M.F.B., R.C.K., S.L.S., J.M.P., E.C.Y., R.B.G., and J.J.H. designed research; M.F.B., R.C.K., S.L.S., J.M.P., A.Q.B., E.C.Y., M.H., J.F., and M.D.M. performed research; C.M.H., B.H.L., G.N.D., and S.M.B. contributed new reagents/analytic tools; M.F.B., R.C.K., S.L.S., J.M.P., E.C.Y., S.M.B., and J.J.H. analyzed data; and M.F.B., R.C.K., and J.J.H. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, [www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo) (accession no. GSE12404).

<sup>1</sup>M.F.B. and R.C.K. contributed equally to the manuscript.

<sup>2</sup>Present address: Department of Biological Sciences, University of Manitoba, Winnipeg, MB, Canada R3T 2N2.

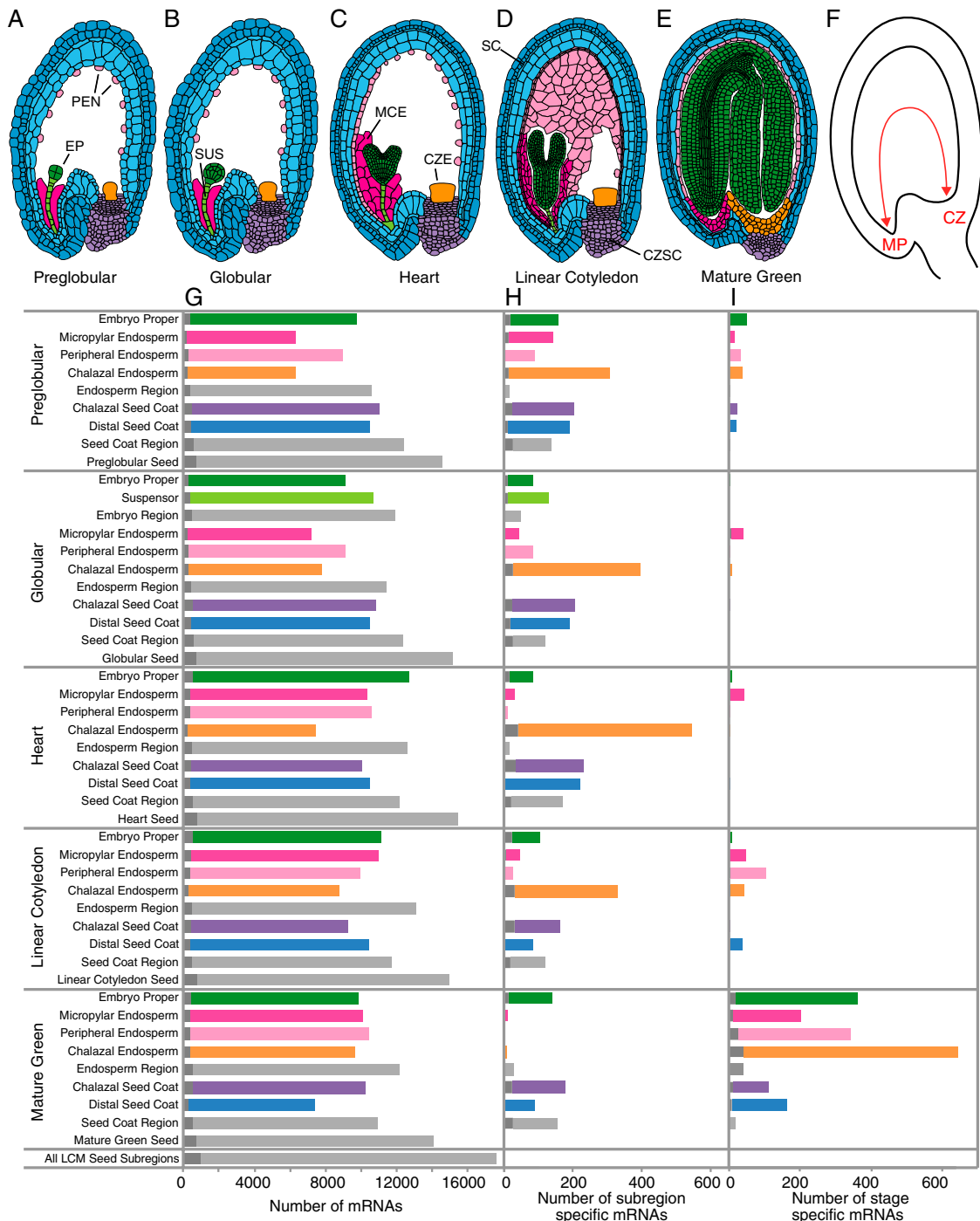
<sup>3</sup>Present address: Research Services, University of Saskatchewan, Saskatoon, SK, Canada S7N 4J8.

<sup>4</sup>Present address: BASF Plant Sciences, Research Triangle Park, NC 27709.

<sup>5</sup>Present address: Bioinformatics and Medical Informatics Graduate Program, San Diego State University, San Diego, CA 92182.

<sup>6</sup>To whom correspondence may be addressed. E-mail: [jjharada@ucdavis.edu](mailto:jjharada@ucdavis.edu) or [bobg@ucla.edu](mailto:bobg@ucla.edu).

This article contains supporting information online at [www.pnas.org/lookup/suppl/doi:10.1073/pnas.1222061110/-DCSupplemental](http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1222061110/-DCSupplemental).



**Fig. 1.** Gene activity in *Arabidopsis* seed regions and subregions throughout development. (A–F) Representation of *Arabidopsis* seeds at the (A) preglobular stage, (B) globular stage, (C) heart stage, (D) linear cotyledon stage, and (E) mature green stage. (F) Diagram of a seed showing micropylar (MP) and chalazal (CZ) poles. (G) Number of distinct mRNAs detected in seed subregions (colored bars), regions, and seeds (light gray bars) at different stages. Dark gray bars indicate the number of distinct transcription factor mRNAs. Lists of mRNAs and their levels are in [Dataset S2](#). (H) Number of distinct mRNAs that accumulate specifically in a subregion or region at a given stage. Subregion and region-specific mRNAs are listed in [Dataset S3](#). (I) Number of distinct mRNAs that accumulate at a specific stage in a subregion or region. Stage-specific mRNAs are listed in [Dataset S3](#). Abbreviations are given in Table 1.

We previously analyzed genes expressed in developing whole seeds at several developmental stages and identified seed-specific genes and transcription factors, and these data provided clues about temporally regulated processes that occur during seed development (5). Questions remain, however, about the processes that occur specifically in a subregion and the interactions among different regions. A number of recent studies have reported gene activity in specific seed regions at the whole-genome level, such as the embryo, endosperm, and seed coat (reviewed by ref. 6). However, these studies do not enable an integrated understanding of seed development, because most focused on a specific stage of seed development and few reported gene activity in more than one region. Here, we describe gene activity genome-wide in all subregions and regions of seeds of the model plant *Arabidopsis*, from fertilization through maturity. The temporal and spatial integration of cellular and physiological processes in multiple subregions and stages permits seminal insights into the developmental processes that characterize specific seed regions and the gene regulatory programs that underlie seed development. Use of uniform platforms of subregion isolation and RNA analyses permit direct comparisons of mRNA levels in different subregions and stages, enabling an integrated understanding of seed development.

## Results

**Spatial and Temporal Resolution of mRNA Profiles During Seed Development.** We profiled mRNA populations from six to seven seed subregions at five stages of development (Fig. 1 *A–E* and Fig. S1 *A–E*) to obtain the most comprehensive description of gene activity in seed development, representing 31 combinations of subregions and stages. Laser-capture microdissection (LCM) (*Materials and Methods*) was used to isolate the embryo proper (EP) and suspensor (SUS) of the embryo region, micropylar (MCE), peripheral (PEN), and chalazal (CZE) subregions of the endosperm region, and the chalazal (CZSC) and distal (SC) seed coat (Fig. S1 *F–Q* and Dataset S1, Table S1). The subregions were isolated in replicate at the preglobular, globular, and heart stages that collectively represent the morphogenesis phase (Fig. 1 *A–C*). Subregions isolated at the mature green stage correspond to the maturation phase (Fig. 1*E*), whereas the linear-cotyledon stage (Fig. 1*D*) is a transition between the two phases. All subregions and stages and their abbreviations are listed in Table 1. The ephemeral SUS was isolated only at the globular stage because an average of 1,700 captured subregions were needed for each biological replicate (Dataset S1, Table S2). mRNAs in each subregion were detected and quantified using stringent analyses of Affymetrix ATH1 GeneChip hybridization data (*Materials and Methods* and Dataset S2). These data are available at the Gene Expression Omnibus (GEO) database ([www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo)) as series GSE12404 and in an interactive form at <http://seedgenenetwork.net>. Exhaustive control experiments, including validation of relative mRNA levels using quantitative RT-PCR (qRT-PCR), comparisons of mRNA accumulation patterns with promoter activities, and confirmation that mRNA sequence amplification was unbiased, established that the LCM seed dataset is representative of subregion RNA populations, qualitatively and quantitatively (Fig. S2 and Dataset S1, Tables S3 and S4).

Fig. 1*G* summarizes gene activity in subregions, regions, and whole seeds at each developmental stage. We detected between ~6,000 and 13,000 distinct mRNAs in each subregion. The number of mRNAs detected in a region, calculated as the union of mRNAs present in each of its constituent subregions, was not appreciably higher than that of individual subregions. Similarly, the union of mRNAs present in embryo, endosperm, and seed-coat regions, representing whole-seed mRNA number, was not appreciably higher than that of a single region. These results indicate that there is substantial overlap in the genes expressed in regions and subregions of a seed. An average of ~14,800 distinct mRNAs was detected throughout the seed at each stage, and collectively a minimum of 17,594 distinct mRNAs were detected in at least one subregion and stage of seed develop-

**Table 1. Abbreviations for developmental stages and seed subregions**

Abbreviation	Region
<b>Stage identifiers</b>	
pg	Preglobular
g	Globular
h	Heart
lc	Linear cotyledon
mg	Mature green
<b>Subregion identifiers</b>	
EP	Embryo proper
SUS	Suspensor
MCE	Micropylar endosperm
PEN	Peripheral endosperm
CZE	Chalazal endosperm
CZSC	Chalazal seed coat
SC	Distal seed coat

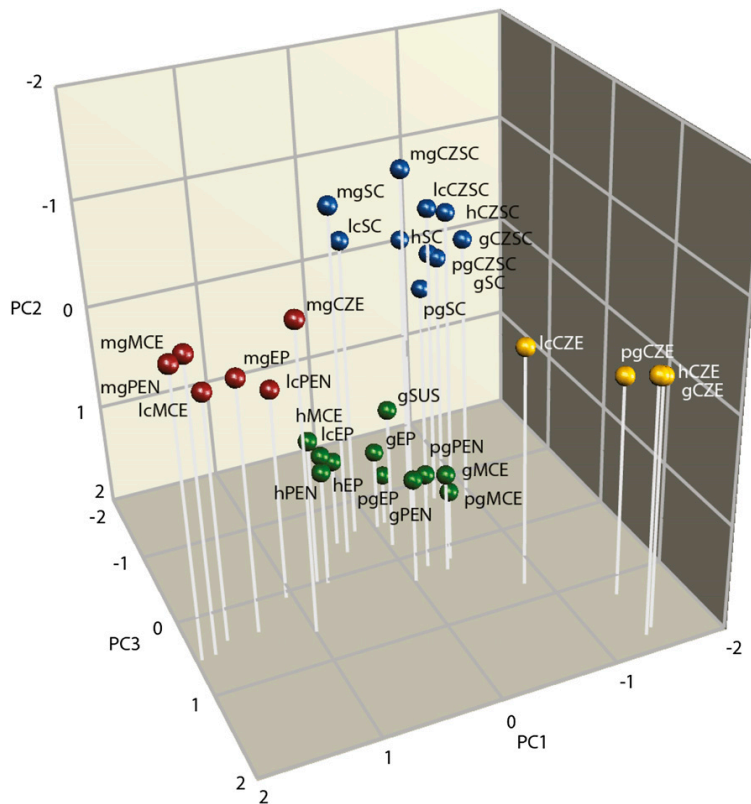
ment. These are minimum estimates given the stringency of the analyses, the absence of probes for ~17.5% of *Arabidopsis* genes on the ATH1 GeneChip, and the lack of consistent detection of mRNAs present at the lowest prevalence levels (fraction of mRNA < 10<sup>-5</sup>) in GeneChip experiments (5).

**Global Comparisons of mRNA Populations Reveal Functional Relationships Among Seed Subregions.** To provide a framework to understand how seed development is coordinated, we compared mRNA populations from all 31 combinations of subregions and stages globally using principal component analysis (PCA). The rationale was that subregions identified as being most closely associated in the analysis were likely to share the greatest similarity in overall gene expression and, therefore, cellular functions. Fig. 2 shows that four groups of subregions were identified by the analysis: (i) the EP, SUS, MCE, and PEN subregions at the preglobular to heart/linear cotyledon stage (green); (ii) CZE subregions at these early stages (yellow); (iii) all embryo and endosperm subregions late in development (red); and (iv) all CZSC and SC subregions at all stages (blue).

The finding that the MCE and PEN early in development shared greater similarity with both embryo subregions than with the CZE was surprising, because the embryo and endosperm arise from separate fertilization events. Late in development, all embryo and endosperm subregions, including the CZE, formed a group that was distinct from the same subregions early in development, suggesting that a major shift in gene expression occurs during the transition from early to late stages. The CZSC and SC at all stages grouped, suggesting that these subregions of the seed-coat region share greater similarity with each other than with embryo and endosperm subregions. Identical results were obtained using hierarchical clustering of mRNA populations from all subregions and stages (Fig. S3*A*), further supporting the biological significance of the groupings. The same relationships among subregions were obtained when mRNAs at each stage were clustered hierarchically (Fig. S3 *B–F*). Taken together, the results suggest that the maternally derived seed coat differs fundamentally from the embryo and endosperm that both arise from fertilization events and that embryo and endosperm subregions share a complex relationship that is dependent on spatial and temporal cues.

**Diverse Sets of Coexpressed Genes Underlie Seed Development. Subregion-specific gene sets.** We identified mRNAs that accumulate specifically in a subregion to begin to discover the gene-expression programs that underlie the complex relationship among subregions of the seed. We defined subregion-specific mRNAs as those that accumulated at a statistically significant ( $q < 0.001$ , mixed-model ANOVA), fivefold or higher level in one subregion relative to all others at a given stage. The fivefold cutoff value was based on the finding that fold-change values for





**Fig. 2.** PCA of seed subregion mRNA populations. PCA plot shows four distinct groups of subregion mRNA populations: subregions of the seed coat region at all stages (blue), EP, SUS, MCE, and PEN subregions at early stages (green), CZE subregions at early stages (yellow), and the EP and all endosperm subregions at the maturation phase (red). Principal components one through three collectively represent 55.6% of the variance in the dataset. Abbreviations are given in Table 1.

mRNAs significantly higher in one subregion versus all others ranged from 1.01 to 210, with a median of 4.7. Fig. 1H shows that between 0 (mature-green PEN) and 545 (heart CZE) subregion-specific mRNAs were identified (Dataset S3). Thus, few mRNAs accumulated specifically at the cell or tissue level relative to the total number of mRNAs in each subregion.

To determine how subregion-specific mRNAs changed over time, we clustered mRNAs from all subregions and stages to define 47 dominant expression patterns (DPs) (Fig. 3A and Fig. S4) (7) and assigned mRNAs to these patterns on the basis of correlation (Pearson's correlation > 0.8) (Dataset S4). Several of the coexpressed gene sets consisted of mRNAs that accumulated primarily in one subregion (Fig. 3A and Fig. S4) (DPs 15, 18, 20, 21, 24, 29, and 37), and an average of 70% of mRNAs in these gene sets were subregion-specific ( $\geq$ fivefold enrichment,  $q < 0.001$ ). The accumulation patterns show that some mRNAs accumulated predominantly in one subregion at several stages. In contrast, each subregion, with the exception of the mature-green PEN, contained between 6 and 135 mRNAs that accumulated subregion-specifically at only one stage. Thus, some genes were expressed subregion-specifically at a single stage, whereas others were expressed specifically over several stages. These temporal variations of subregion-specific expression patterns add complexity to the gene regulatory networks that operate during seed development.

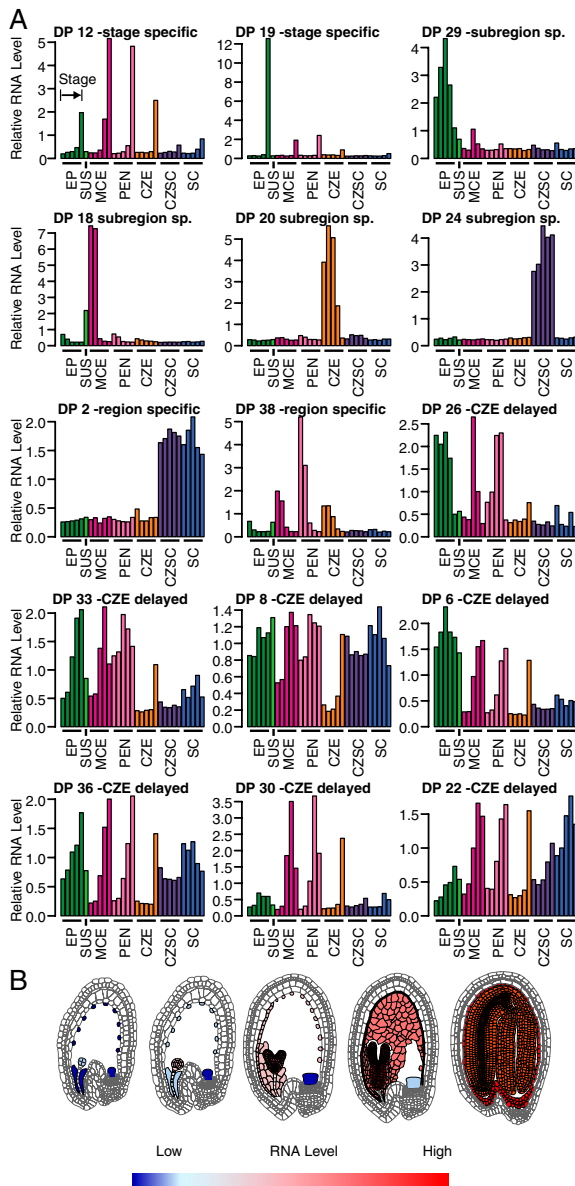
**Stage-specific gene sets.** Because global comparisons of mRNA populations suggested a concerted temporal shift in gene expression during development (Fig. 2), we identified mRNAs in each subregion that accumulated specifically at a single stage. We defined stage-specific mRNAs as those present at a statistically significant ( $q < 0.001$ , mixed-model ANOVA), fivefold or higher level at one stage relative to all others in a given subregion (Dataset S3). Fig. 1I shows that relatively few genes were expressed stage-spe-

cifically early during seed development. Rather, an average of 71% of the stage-specific mRNAs of each subregion accumulated at the mature-green stage. These results indicate that a major transition in gene activity occurs at the mature-green stage.

Many of the same mature-green stage-specific mRNAs accumulated in all embryo and endosperm subregions. Two coexpressed gene sets, DP 12 and DP 19 (Fig. 3A), consisted of mRNAs that accumulated primarily at the mature-green stage in all embryo and endosperm subregions and to a lesser extent in seed-coat subregions. An average of 58% and 66% of mRNAs in DP 12 and DP 19, respectively, were also designated as mature-green stage-specific mRNAs ( $\geq$ fivefold enrichment,  $q < 0.001$ ) in the EP, MCE, PEN, and CZE. By contrast, averages for the CZSC and SC were 16% and 14%, respectively. Taken together, these results suggest a common set of genes is coordinately up-regulated in embryo and endosperm subregions late in seed development.

**Roles of Subregion-Specific Genes in Seed Development.** We obtained insight into the cellular processes that characterize each subregion by identifying Gene Ontology (GO) terms and metabolic pathways that were significantly overrepresented ( $P < 0.001$ , hypergeometric distribution) (Dataset S3) among subregion-specific mRNAs. Fig. 4A lists GO terms and metabolic pathways for subregion-specific mRNAs that were overrepresented at two or more stages and/or for the DP gene set that exhibits subregion specificity. The analysis confirmed known functions for some subregions and provided a glimpse into the specific functions of other subregions whose roles in seed development were not known.

**Embryo proper.** EP-specific mRNAs were significantly enriched for GO terms known to be associated with patterning events that occur during embryo development, such as determination of bilateral symmetry and abaxial cell fate specification (8).



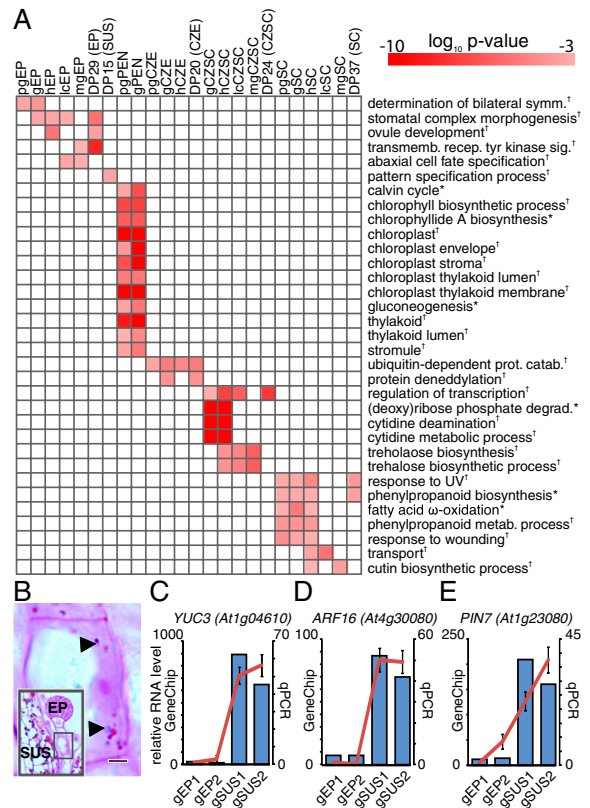
**Fig. 3.** Dominant patterns of gene expression during seed development. (A) Forty-seven DPs were identified using Fuzzy *K* means clustering of the 50% most variant mRNAs in all seed subregions and stages. Bar graphs depict median mRNA levels in each subregion (colored bars) at each stage (left to right, preglobular to mature-green stage). DPs representing the indicated stage-specific, subregion-specific, region-specific, and CZE-delayed coexpressed gene sets are shown. Remaining DPs are shown in Fig. S4, and mRNAs in all DPs are listed in Dataset S4. The average number of mRNAs in each DP gene set was 103. (B) Heat map of conceptualized CZE-delayed mRNA accumulation patterns in embryo and endosperm subregions. mRNA accumulation in seed coat subregions is not shown.

**Seed coat.** SC-specific mRNAs were overrepresented for processes associated with the synthesis of flavonoids that serve to provide protection for seeds against biotic and abiotic stresses (9). **Peripheral endosperm.** At the preglobular and globular stages, PEN-specific mRNAs were overrepresented for GO terms for

chloroplast compartments and metabolic pathways related to photosynthesis. These processes were not known to occur in the PEN, but their occurrence was validated functionally as discussed below.

**Chalazal seed coat.** mRNAs associated with trehalose and cytidine metabolism were overrepresented in the CZSC. Although trehalose plays an essential role in seed development (10), localization of key enzymes required for its synthesis to the CZSC was not known previously.

**Suspensor.** The SUS is an embryonic structure of 8–10 cells, and little is known about its cellular functions (11). Because SUS-specific mRNAs in DP 15 (Fig. S4) that were overrepresented for the GO term, pattern-specification process, encode efflux transporters for the hormone auxin, we analyzed a number of mRNAs involved in auxin signaling. Fig. 4 C–E shows that mRNAs encoding an auxin biosynthetic enzyme, YUC3, an auxin efflux carrier, PIN7, and a transcription factor responsive to auxin, ARF16, were more prevalent in the SUS than in the EP. These results are consistent with the SUS serving as a site of perception of the auxin gradient across the SUS and EP early in seed development that is essential for patterning of the embryo



**Fig. 4.** Functions of subregion-specific genes. (A) Heat map showing the *P* value significance of enrichment of GO terms (\*) or metabolic pathways (\*) for subregion-specific mRNAs (Dataset S3). The listed GO terms are for biological processes or cellular components and metabolic pathways that were overrepresented at two or more stages and/or for the DP gene set that exhibit subregion specificity. (B) Histochemical staining of starch granules (arrowheads) in the suspensor. (Scale bar, 3  $\mu$ m.) (Inset) The location of the enlarged area relative to the embryo proper and suspensor. (C–E) Relative mRNA levels determined in GeneChip experiments (bar plots) and qRT-PCR experiments (line plots) for the indicated genes involved in (C) auxin biosynthesis, (D) auxin response, and (E) polar auxin transport.

(12, 13). Additionally, they open the possibility that the SUS may serve as an auxin source for the gradient. We also discovered that mRNAs encoding all enzymes involved in starch biosynthesis were detected in the SUS (Dataset S1, Table S5) and demonstrated the presence of a functional pathway by showing that starch grains accumulate in SUS cells (Fig. 4B).

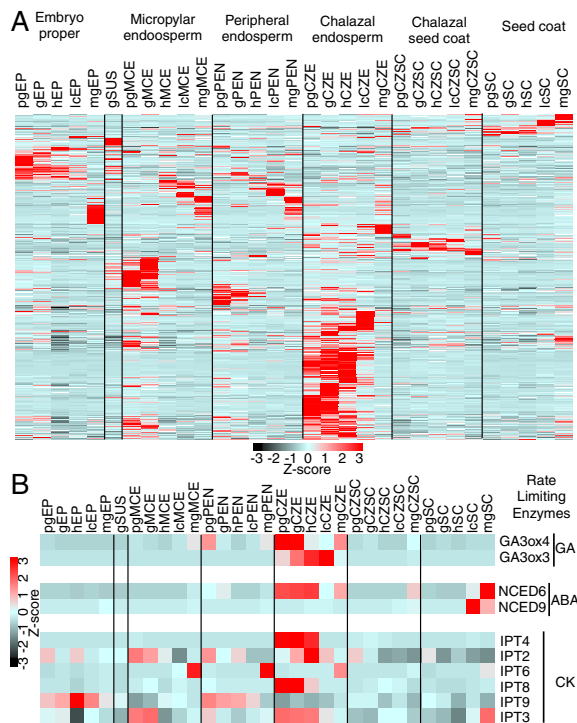
**Chalazal endosperm.** The CZE is a unique subregion developmentally. Early in seed development, the CZE possessed the largest number of subregion-specific mRNAs (Fig. 1H). Moreover, the CZE was highly enriched for mRNAs that were only detected during seed development. Seed-specific mRNAs were identified as those present in at least one subregion and stage of seed development but that were not detected in GeneChip experiments in seedlings, leaves, stems, or roots of vegetatively growing plants and flower buds or ovules of reproductively growing plants, as described previously (5). Of at least 1,316 seed-specific mRNAs (Dataset S2), the largest fraction accumulated predominantly in the CZE, as shown in Fig. 5A. Additionally, 244 of 788 CZE-specific mRNAs accumulated seed specifically. These results are consistent with our previous report that the promoters of several seed-specific transcription factor genes are active specifically in the CZE (5). Together, these gene expression patterns support the conclusion that the CZE differs fundamentally from other subregions early in seed development, suggesting that novel processes occur in the CZE.

Given the uniqueness of the CZE, we were interested to understand its role in seed development. CZE-specific mRNAs were overrepresented for the GO term ubiquitin-dependent protein catabolism (Fig. 4A), suggesting a potential regulatory role for the CZE. We also showed that rate-limiting enzymes for the biosynthesis of the hormones gibberellic acid, abscisic acid, and cytokinin, accumulated primarily, but not exclusively, in the CZE (Fig. 5B), consistent with other reports showing that hormone metabolism genes are expressed in the CZE (14–16). Because of the importance of these hormones for seed development, these results suggest that the CZE may serve as a communication hub that integrates developmental processes within the seed.

**Integration of Gene Activity and Cellular Function Across Subregions and Stages.** Gene sets temporally regulated in embryo and endosperm subregions. Although many gene sets expressed subregion- and stage-specifically were identified, we were interested to know the extent to which gene expression was coordinated across distinct subregions and stages during seed development. We identified gene sets that were coexpressed in several subregions and stages. The most prominent coexpression pattern, representing 11 DPs (Fig. 3A and Fig. S4) (DP 1, 6, 7, 8, 9, 14, 22, 26, 30, 33, and 36), involved mRNAs that accumulated in the EP and all endosperm subregions but whose accumulation in the CZE was delayed relative to the other subregions (Fig. 3B).

**Functions of CZE-delayed gene sets.** The expression patterns of CZE-delayed gene sets (Fig. 3B) suggest that specific cellular processes that occur in all embryo and endosperm subregions are delayed in the CZE. We identified significantly enriched ( $P < 0.001$ , hypergeometric distribution) GO terms and metabolic pathways for the CZE-delayed gene sets and showed that several gene sets were implicated to play important roles in seed development (Dataset S4). For example, the DP 26 gene set was overrepresented for GO terms related to cytokinesis and the phragmoplast, a cytoskeletal structure specific to dividing plant cells (Figs. 3A and 6E). Cytokinesis occurs in the embryo throughout the morphogenesis phase early in development. By contrast, the endosperm initially undergoes nuclear but not cell divisions, with subsequent cellularization and cell division occurring sequentially from the micropylar to the chalazal end (4). Thus, the DP 26 coexpression pattern is coincident with the patterns of cytokinesis during seed development.

Another CZE-delayed gene set, DP 33, was significantly enriched for GO terms and metabolic pathways related to photosynthesis and carbon metabolism, including chloroplast struc-

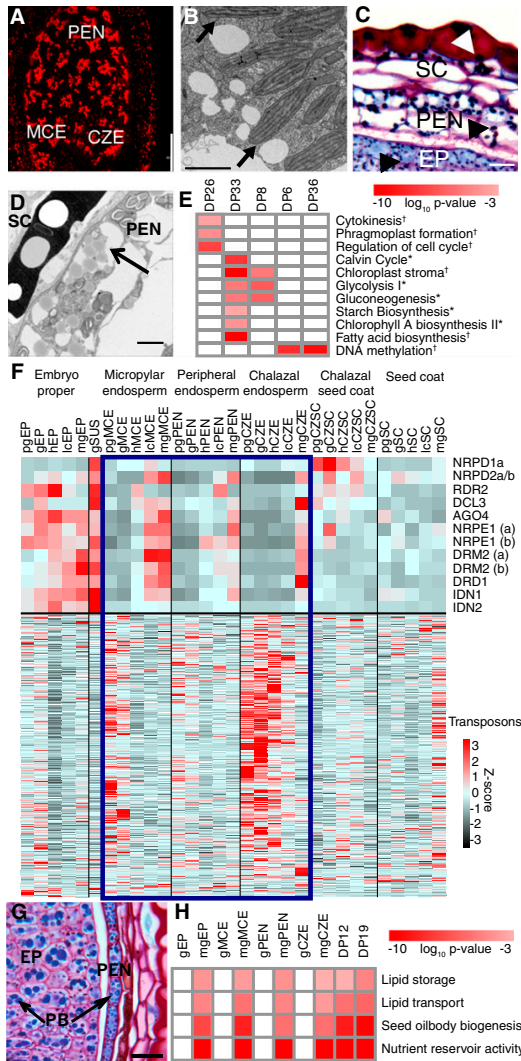


**Fig. 5.** CZE is a unique seed subregion developmentally. (A) Hierarchical clustering of seed-specific mRNAs. The largest number of seed-specific mRNAs accumulate primarily in the CZE. (B) Heat map depicting relative levels of mRNAs encoding rate-limiting enzymes for gibberellic acid (GA; GA3ox), abscisic acid (ABA; NCED), and cytokinin (CK; IPT) biosynthesis.

ture and function, glycolysis, gluconeogenesis, starch biosynthesis, and fatty acid biosynthesis (Figs. 3A and 6E, and Dataset S4). Similarly, DP 8 was associated with glycolysis and gluconeogenesis. These associations were surprising, because photosynthesis and carbon metabolism are known to occur in the embryo, but much less was known about these processes in the endosperm (17). In support of the prediction that starch and lipid biosynthesis occur in the endosperm, qRT-PCR experiments showed that mRNAs encoding all enzymes required for starch and fatty acid biosynthesis were detected in endosperm subregions (Dataset S1, Table S5). The enriched GO terms were predictive of cellular function, because differentiated chloroplasts, starch grains, and lipids were detected in endosperm cells (Fig. 6A–D). Thus, processes involved in photosynthesis and carbon metabolism that are known to characterize the embryo also occur in all three endosperm subregions, although these processes are delayed in the CZE.

Two other CZE-delayed gene sets, DP 6 and DP 36, were significantly enriched for the DNA methylation GO term (Figs. 3A and 6E, and Dataset S4). DNA methylation in plants is mediated primarily through three pathways involving METHYLTRANSFERASE1 (MET1), CHROMOMETHYLASE3 (CMT3), and the RNA-directed DNA methylation (RdDM) enzymes (18). Of the three, only mRNAs involved in RdDM exhibited a CZE-delayed accumulation pattern (Fig. 6F, Upper). DNA methylation functions primarily to silence transposon activity (19), and the endosperm is notable because transposons become hypomethylated in the central cell of the female gametophyte, the precursor of the endosperm (20). Consistent with these observations, transposon activity was high in all three endosperm subregions early in seed development but decreased late in seed development coincident with the increase





**Fig. 6.** Functions of CZE-delayed coexpressed gene sets. (A) Autofluorescent chloroplasts in the endosperm of a globular-stage seed. (B) Transmission electron microscopy of chloroplasts (arrows) in the PEN. (C) Histochemical staining of starch granules (arrowheads) in the SC, PEN, and EP. (D) Transmission electron micrograph of oil bodies (arrows) in cellularized PEN. (E) Heat map showing  $P$  value significance of enrichment of selected GO terms (\*) or metabolic pathways (\*) associated with the indicated CZE-delayed gene sets. (F) Heat maps showing mRNAs involved in the RNA-dependent DNA methylation pathway (Upper) and 1,155 probesets corresponding to transposons (Lower) (21). (G) Detection of protein bodies (PB) in the EP and PEN. (H) Heat map showing  $P$  values for GO-term enrichment of mature-green stage-specific mRNAs and the indicated DPs at the globular and mature-green stages. (Scale bars: 25  $\mu\text{m}$  in A, 0.5  $\mu\text{m}$  in B, 10  $\mu\text{m}$  in C, 3  $\mu\text{m}$  in D, and 10  $\mu\text{m}$  in G.)

in RdDM mRNA levels (Fig. 6F, Lower) (21). The anticorrelation between RdDM mRNAs and transposon activity opens the possibility that DNA methylation is required to silence transposons late in endosperm development.

**Reprogramming of Seed Development During the Maturation Phase. Maturation occurs in embryo and endosperm subregions.** The reprogramming of gene expression that occurs late in seed development

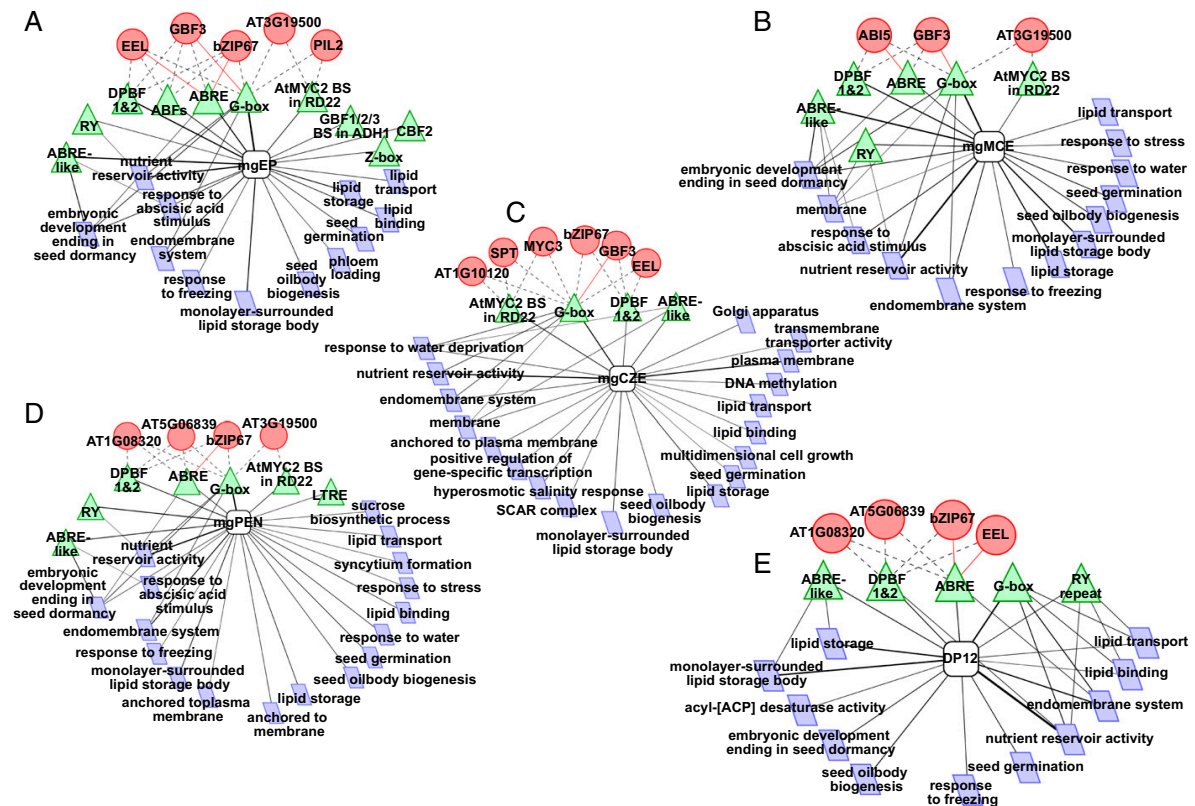
appears to be associated with the onset of the maturation phase. Several gene sets comprised of mRNAs that accumulated primarily at the mature-green stage, including mature-green stage-specific mRNAs in the EP, MCE, PEN, and CZE (Fig. 1I) and DP 12 and 19 (Fig. 3A), were all significantly enriched ( $P < 0.001$ , hypergeometric distribution) for GO terms characteristic of seed maturation, including nutrient reservoir activity, lipid storage, and seed oil body biogenesis, among others (Fig. 6H and Datasets S3 and S4).

The finding that the same sets of maturation-related mRNAs accumulated in the embryo and in all three endosperm subregions was unexpected. Although the embryo undergoes maturation, and lipids are known to accumulate in the endosperm (22), the extent to which the maturation program occurs in endosperm subregions was not known. We showed that cellular structures that accumulate in maturation-phase embryos, such as storage protein bodies (Fig. 6G) and oil bodies (Fig. 6D), were detected in endosperm subregions. These results provide compelling evidence that maturation processes associated with seed filling occur in the embryo and all three endosperm subregions and involve many of the same genes.

**Predicted regulatory circuitry controlling genes expressed during the maturation phase.** One key to improving seeds as food is to define the gene regulatory networks that control the accumulation and composition of storage products during the maturation phase. We developed a framework to predict transcriptional modules that link transcription factors with their potential coexpressed target genes (Materials and Methods). The strategy associates DNA sequence motifs that are significantly enriched in the upstream regions of coexpressed genes ( $P < 0.001$ , hypergeometric distribution) with coexpressed transcription factors known or predicted to bind the overrepresented motifs. We showed that several transcriptional modules were predicted for seed-coat region-specific genes linking enriched MYB, HD-ZIP IV, and AG DNA sequence motifs with transcription factors known to be involved in seed coat and ovule development, such as MYB5, GL2, SHP2, STK, SHP1, and SEP (Fig. S54 and Dataset S1, Table S6) (23, 24). Thus, the approach identified known developmental regulators.

Fig. 7E and Dataset S1, Table S6 show that the DP 12 gene set, consisting of mRNAs that accumulate predominantly in embryo and endosperm subregions during the maturation phase, defined a transcriptional module in which significantly enriched DNA motifs with a G-box core, including ABRE, ABRE-like, and DPBF1 and -2, were associated with bZIP transcription factors. Many of these overrepresented DNA sequence motifs are known to characterize the promoters of maturation expressed genes (25), and two of the associated transcription factors, EEL and bZIP67, play roles in regulating maturation genes (26, 27). Furthermore, transcriptional modules derived from coexpressed genes associated with a specific GO term were also identified. These submodules identify potential regulatory circuits that control processes associated with the GO term during the maturation phase. We also generated transcriptional modules for mature-green stage-specific genes expressed in the EP, MCE, PEN, and CZE and showed that there was substantial overlap in the enriched DNA motifs and associated transcription factors identified in each subregion (Fig. 7A–D). The results suggest that maturation processes are regulated similarly but not identically in the embryo and endosperm subregions.

Analyses of other coexpressed gene sets identified transcription factors known to play critical roles in seed development. For example, genes expressed primarily in the MCE [Fig. 3A (DP 18), Fig. S5D, and Dataset S1, Table S6] were enriched for the W-box DNA sequence motif that is predicted to associate with MINI-SEED3, a WRKY transcription factor that is expressed primarily in the MCE and is a regulator of seed size (28). Similarly, CCA1, a transcription factor involved in controlling seed dormancy as a central circadian clock regulator, was associated with the CCA1 binding-site motif that is enriched in the promoter of genes expressed early in endosperm development [Fig. 3A (DP 27), Fig. S5C, and Dataset S1, Table S6] (29). Thus, the transcriptional



**Fig. 7.** Predicted transcriptional modules regulating maturation in seeds. DNA motifs (green triangles) and GO terms (blue parallelogram) that are significantly overrepresented ( $P < 0.001$ , hypergeometric distribution) within the coexpressed gene set (open squiggle). Coexpressed transcription factors are represented as circles. Transcriptional modules were predicted for mature-green stage-specific genes in the (A) EP, (B) MCE, (C) CZE, and (D) PEN, and for (E) DP 12. All four mature green-stage subregions possess transcriptional modules in which bZIP transcription factors known to regulate maturation genes such as bZIP67 (AT3G44460), EEL (AT2G41070), or ABI5 (AT2G36270) are associated with overrepresented G box-like DNA motifs such as ABRE and DPBF1 and -2. Edges in red indicate known interactions between transcription factors and DNA motifs, whereas dashed lines represent predicted interactions. All enriched DNA motifs and GO terms are listed in [Dataset S1, Table S6](#).

modules identified key regulators of seed development, suggesting their utility as predictive tools to provide insight into gene regulatory networks controlling seed development.

### Discussion

We profiled RNA populations in every cell type, tissue, sub-region, and region of *Arabidopsis* seeds throughout development to obtain an integrated understanding of the processes that underlie seed development. A minimum of 17,594 distinct mRNAs were detected in at least one subregion and stage, indicating that at least 60% of the *Arabidopsis* genome is expressed during seed development. The use of LCM facilitated gene discovery. Compared with our previous analyses of mRNA populations in whole-mount seeds (5), the LCM profiling experiments detected more mRNAs throughout seed development (17,594 vs. 15,577), a higher average number of mRNAs present in seeds at each stage (14,800 vs. 11,780), and a higher number of seed-specific mRNAs (1,316 vs. 289). The dataset provides the most comprehensive description of gene activity during seed development.

**Coexpressed Gene Sets Inform the Cellular Processes that Underlie Seed Development.** The LCM profiling experiments describe global gene activity in seed subregions that were previously inaccessible to such analyses. Identification of both region-specific mRNAs and subregion-specific mRNAs (Fig. 1H) suggests that

subregions within the same region have both shared and distinct functions. For example, mRNAs that accumulate specifically in the seed-coat region are overrepresented for the GO terms flavonoid biosynthetic process and proanthocyanidin biosynthetic process ([Dataset S3](#)), suggesting that processes associated with responses to biotic and abiotic stresses occur in both the SC and CZSC (9). By contrast, the CZSC alone is overrepresented for subregion-specific mRNAs associated with trehalose and cytidine metabolism, but these mRNAs are not detected at substantial levels in the SC (Fig. 4A and [Dataset S3](#)). Thus, distinct gene sets are involved in controlling region-specific and subregion-specific functions.

Our data suggest that the functional differentiation of subregions within a region occur through at least two distinct sets of processes. First, genes expressed specifically within a subregion appear to play a significant role in specifying its function. GO terms and metabolic pathways enriched for mRNAs specifically expressed in the EP, SC, and PEN accurately predict functions that are known or that we have shown to occur in these subregions (Figs. 4 and 6). Many subregion-specific genes are active at the preglobular stage, suggesting that subregion identity is specified at the earliest stage of seed development (Figs. 1H and 3A, and [Fig. S4](#)). For example, consistent with our finding that many genes are expressed specifically in each endosperm sub-region at the preglobular stage, others have shown that the

MCE, PEN, and CZE can be differentiated morphologically at the 16-nuclei stage that corresponds to the zygote-stage of seed development (30). Second, subregion function is also influenced by temporal differences in the expression of gene sets. CZE-delayed gene sets consist of mRNAs that accumulate later in the CZE than in the other embryo and endosperm subregions (Fig. 3B). Delayed accumulation of these mRNAs accounts, in part, for the finding that the CZE differs from the EP, SUS, MCE, and PEN early in seed development (Fig. 2). Together, these results define gene sets with diverse coexpression patterns that contribute to the overall complexity of seed development.

**Gene Sets Associated with the Control of Seed Mass.** Seed mass is positively correlated with seedling survival and, therefore, is a determinant of plant fitness (31). The ability to modulate seed mass has important implications for altering crop yield. Several CZE-delayed gene sets are associated with processes that control seed mass (Figs. 3A and 6). For example, DP 26 is over-represented for mRNAs involved in cytokinesis. Because the timing of endosperm cellularization is correlated with seed mass (32, 33), with smaller seeds undergoing early cellularization, mechanisms that regulate this gene set are likely to be involved in controlling seed mass. Similarly, a second CZE-delayed gene set, DP 33, is enriched for mRNAs involved in photosynthesis and carbon metabolism, and photosynthetic activity in seeds is correlated with seed biomass (34).

Two other CZE-delayed gene sets, DP 6 and DP 36, which are associated with DNA methylation via the RdDM pathway, may also be related to the control of seed size (Figs. 3A and 6). A potential tie between the RdDM pathway and seed mass is that DNA methylation is implicated to control the expression of many imprinted genes in the endosperm, genes that are expressed specifically or preferentially from either maternal or paternal alleles (20, 35). Imprinted genes are predicted and, in one case, shown to be involved in controlling resource allocation to the embryo, a process that is critical in determining seed mass (36, 37). Moreover, imprinted genes are often flanked by transposons, and the methylation status of the transposable element is thought to determine the activity of many imprinted genes (38, 39). Consistent with the interpretation that transposons affect the activity of neighboring genes, NRPD1a, a component of the RdDM pathway, is required to silence genes encoding endosperm-specific transcription factors that are adjacent to transposons (40). Thus, accumulation of RdDM mRNAs late in endosperm development that appears to correlate with transposon silencing and, presumably, DNA methylation may also result in the silencing of imprinted genes (Fig. 6F). We estimate that 35 of 47 imprinted genes (39) are down-regulated in the endosperm coincident with the activation of RdDM genes. Because imprinted genes are thought to enable the endosperm to promote early embryo development, silencing of these genes late in seed development may be required to allow induction of the maturation phase in the endosperm.

**Coordinated Gene Expression in the Embryo and Endosperm and Its Relevance to the Origin of the Endosperm.** An overriding theme that emerges from this comprehensive developmental profile of mRNA populations is that there is extensive overlap in the gene-expression programs that characterize embryo and endosperm subregions. Although each subregion possesses mRNAs that accumulate specifically in that subregion (Fig. 1H), a global comparison of mRNA populations demonstrated unexpected similarities between embryo and endosperm subregions (Fig. 2). These similarities result, at least in part, from the large number of CZE-delayed genes that are coexpressed in EP, SUS, MCE, and PEN subregions early in seed development (Fig. 3 and Fig. S4). Thus, the same sets of genes that are associated with photosynthesis, carbon metabolism, cytokinesis, and DNA methylation are active in all embryo and endosperm subregions early in seed development, although their activity is delayed in the CZE. Consistent with the extensive overlap in embryo and endosperm gene activity, a large set of genes is up-regulated coordinately in

the embryo proper and all endosperm subregions during the transition from the morphogenesis to the maturation phase, and many of the same putative regulators operate in the two seed regions [Fig. 3 (DP 12 and 19) and Fig. 7]. These results suggest that there is substantial coordination of the biological processes that occur in embryo and endosperm regions.

Parallels in embryo and endosperm expression programs have implications for resolving longstanding questions about the evolution of seeds. The endosperm region is unique to angiosperms, and two major hypotheses have been advanced to explain its evolutionary origin (41, 42). One hypothesis is that the endosperm is a modified supernumerary embryo resulting from a second fertilization event that acquired embryo-nourishing functions. The second hypothesis proposes that the endosperm is homologous with the gymnosperm female gametophyte, the development of which is promoted by a second fertilization event. Morphological analysis of endosperm development in basal angiosperm taxa suggests that there are shared features of early embryo and endosperm development, including unequal division and polarization of the initial cell and differential development at the micropylar and chalazal poles (43). Our results demonstrating strong overlap in embryo and endosperm gene activity are consistent with an embryo-based evolutionary origin of the endosperm, although we cannot exclude the possibility of homology between the endosperm and the female gametophyte.

## Conclusions

The LCM seed dataset represents a robust resource to support studies of seed biology. We have demonstrated that the dataset can be used to identify sets of genes that are expressed in specific subregions and stages and other gene sets the expression patterns of which are integrated across multiple subregions and stages. Thus, these data define coexpressed gene sets with extremely high spatial and temporal resolution. Analysis of these coexpressed genes can accurately predict the biological function of seed subregions, providing fresh insights into the cellular processes that underlie seed development.

A key to understanding seed development is to define the regulatory circuitry that governs these diverse coexpressed gene sets. The dataset serves as a platform to identify the gene regulatory networks that operate during seed development, in part, by identifying the transcription factors that accumulate in spatially restricted locations within the seed at specific stages. We have shown that known regulators of seed development can be identified by the association of overrepresented DNA sequence motifs with coexpressed transcription factors.

The biological stories that we have presented demonstrate the utility of the dataset in uncovering new and significant information about the processes that underlie seed development. Although much remains to be learned about seed biology to obtain the basic information needed for the design of strategies to improve crops for agriculture and enhanced food security, we anticipate that this dataset will be a critical tool in enabling these discoveries.

## Materials and Methods

**Profiling Subregion mRNA Populations.** Siliques containing seeds of *Arabidopsis thaliana* (L.) Heynh, ecotype Wassilewskija (Ws-0) were staged according to criteria described previously (5) and detailed in [Dataset S1, Table S1](#). Details about seed collection, histological protocols, and microdissection using a Leica LMD6000 Laser Microdissection System (Leica Microsystems) are given in [SI Materials and Methods](#).

Total RNA was extracted, purified from captured microdissected subregions, analyzed, and amplified as described in [SI Materials and Methods](#). The number of captured subregions per biological replicate and total RNA yields are summarized in [Dataset S1, Table S2](#). Amplified cDNA was hybridized with the Affymetrix GeneChip ATH1 *Arabidopsis* Genome Array as previously described (5). The effects of amplification on relative RNA levels were determined using qRT-PCR experiments on cDNA amplified from 2 ng of total RNA and cDNA synthesized from 1 µg of total RNA (ThermoScript RT-PCR Systems). [Dataset S1, Table S4](#) shows that linear cDNA amplification did not appreciably alter the representation of mRNAs in the population.



**Data Analysis.** GeneChip hybridization data were analyzed and detection calls for mRNAs (present, absent, or marginal) were made as previously described (5). For quantitative comparisons of mRNA levels, signal intensities from all 75 GeneChip experiments were normalized using RMA (44). Correlation between RMA normalized biological replicates averaged 0.96 (Pearson's correlation, [Dataset S2](#)) and ranged between 0.93 and 0.98, which was higher than that obtained using other normalization methods. Relative RNA levels were validated with qRT-PCR experiments as previously described (45). DNA sequences and efficiencies of primer pairs used for qRT-PCR experiments and comparison of relative mRNA levels determined in GeneChip and qRT-PCR experiments are given in [Dataset S1, Table S3](#).

Mixed-model linear ANOVA, used to assess the significance of mRNA level comparisons in different samples, and identification of dominant expression patterns was done as previously described (7). Other analyses, including hierarchical clustering and bootstrapping analysis and PCA, are described in [SI Materials and Methods](#).

**Identification of Transcriptional Modules.** The software package, ChipEnrich (46), was modified to identify significantly enriched DNA sequence motifs upstream of coexpressed genes that were associated with transcription factors known or predicted to bind the motifs as described in [SI Materials and Methods](#). Files used to generate the transcriptional modules are in [Dataset S1, Table S6](#).

**Microscopy.** Procedures for light and fluorescence microscopy, transmission electron microscopy, and confocal laser scanning microscopy are described in [SI Materials and Methods](#).

**ACKNOWLEDGMENTS.** We thank Samantha Duong, Maichi Phan, Emilia Madejska, Xiaohua Lu, Alexander Olson, and Chen Cheng for technical assistance, and Bob Fischer for comments about the manuscript. This work was supported by grants from the National Science Foundation Plant Genome Program (to R.B.G. and J.J.H.) and a postdoctoral fellowship from the Natural Sciences and Engineering Research Council (to M.F.B.).

- Steeves TA (1983) The evolution and biological significance of seeds. *Can J Bot* 61(12): 3550–3560.
- Godfray HCl, et al. (2010) Food security: The challenge of feeding 9 billion people. *Science* 327(5967):812–818.
- Ohto M, Stone SL, Harada JJ (2007) Genetic control of seed development and seed mass. *Seed Development, Dormancy, and Germination*, eds Bradford KJ, Nonogaki H (Blackwell, Oxford), pp 1–24.
- Brown RC, Lemmon BE, Nguyen H, Olsen O-A (1999) Development of endosperm in *Arabidopsis thaliana*. *Sex Plant Reprod* 12(1):32–42.
- Le BH, et al. (2010) Global analysis of gene activity during *Arabidopsis* seed development and identification of seed-specific transcription factors. *Proc Natl Acad Sci USA* 107(18):8063–8070.
- Harada JJ, Pelletier J (2012) Genome-wide analyses of gene activity during seed development. *Seed Sci Res* 22(Suppl 51):S15–S22.
- Brady SM, et al. (2007) A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science* 318(5851):801–806.
- Capron A, Chatfield S, Provart N, Berleth T (2009) Embryogenesis: Pattern formation from a single cell. *Arabidopsis Book* 7:e0126.
- Pourcel L, Routaboul J-M, Cheynier V, Lepiniec L, Debeaujon I (2007) Flavonoid oxidation in plants: From biochemical properties to physiological functions. *Trends Plant Sci* 12(1):29–36.
- Schluepmann H, Paul M (2009) Trehalose metabolites in *Arabidopsis*-elusive, active and central. *Arabidopsis Book* 7:e0122.
- Kawashima T, Goldberg RB (2010) The suspensor: Not just suspending the embryo. *Trends Plant Sci* 15(1):23–30.
- Wang JW, et al. (2005) Control of root cap formation by MicroRNA-targeted auxin response factors in *Arabidopsis*. *Plant Cell* 17(8):2204–2216.
- Weijers D, et al. (2006) Auxin triggers transient local signaling for cell specification in *Arabidopsis* embryogenesis. *Dev Cell* 10(2):265–270.
- Hu JH, et al. (2008) Potential sites of bioactive gibberellin production during reproductive growth in *Arabidopsis*. *Plant Cell* 20(2):320–336.
- Lefebvre V, et al. (2006) Functional analysis of *Arabidopsis* NCED6 and NCED9 genes indicates that ABA synthesized in the endosperm is involved in the induction of seed dormancy. *Plant J* 45(3):309–319.
- Miyawaki K, Matsumoto-Kitano M, Kakimoto T (2004) Expression of cytokinin biosynthetic isopentenyltransferase genes in *Arabidopsis*: Tissue specificity and regulation by auxin, cytokinin, and nitrate. *Plant J* 37(1):128–138.
- Xiang DQ, et al. (2011) Genome-wide analysis reveals gene expression and metabolic network dynamics during embryo development in *Arabidopsis*. *Plant Physiol* 156(1): 346–356.
- Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 11(3):204–220.
- Lisch D (2009) Epigenetic regulation of transposable elements in plants. *Annu Rev Plant Biol* 60(1):43–66.
- Köhler C, Wolff P, Spillane C (2012) Epigenetic mechanisms underlying genomic imprinting in plants. *Annu Rev Plant Biol* 63(1):331–352.
- Slotkin RK, et al. (2009) Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* 136(3):461–472.
- Penfield S, et al. (2004) Reserve mobilization in the *Arabidopsis* endosperm fuels hypocotyl elongation in the dark, is independent of abscisic acid, and requires PHOSPHOENOLPYRUVATE CARBOXYKINASE1. *Plant Cell* 16(10):2705–2718.
- Colombo L, Battaglia R, Kater MM (2008) *Arabidopsis* ovule development and its evolutionary conservation. *Trends Plant Sci* 13(8):444–450.
- Li SF, et al. (2009) The *Arabidopsis* MYB5 transcription factor regulates mucilage synthesis, seed coat development, and trichome morphogenesis. *Plant Cell* 21(1): 72–89.
- Gutierrez L, Van Wuytswinkel O, Castelain M, Bellini C (2007) Combined networks regulating seed maturation. *Trends Plant Sci* 12(7):294–300.
- Bensmihen S, et al. (2002) The homologous ABIS and EEL transcription factors function antagonistically to fine-tune gene expression during late embryogenesis. *Plant Cell* 14(6):1391–1403.
- Yamamoto A, et al. (2009) *Arabidopsis* NF-YB subunits LEC1 and LEC1-LIKE activate transcription by interacting with seed-specific ABRE-binding factors. *Plant J* 58(5): 843–856.
- Luo M, Dennis ES, Berger F, Peacock WJ, Chaudhury A (2005) MINISEED3 (MINI3), a WRKY family gene, and HAIKU2 (IKU2), a leucine-rich repeat (LRR) KINASE gene, are regulators of seed size in *Arabidopsis*. *Proc Natl Acad Sci USA* 102(48): 17531–17536.
- Penfield S, Hall A (2009) A role for multiple circadian clock genes in the response to signals that break seed dormancy in *Arabidopsis*. *Plant Cell* 21(6):1722–1732.
- Brown RC, Lemmon BE, Nguyen H (2003) Events during the first four rounds of mitosis establish three developmental domains in the syncytial endosperm of *Arabidopsis thaliana*. *Protoplasma* 222(3–4):167–174.
- Westoby M, Jurado E, Leishman M (1992) Comparative evolutionary ecology of seed size. *Trends Ecol Evol* 7(11):368–372.
- Garcia D, et al. (2003) *Arabidopsis* haiku mutants reveal new controls of seed size by endosperm. *Plant Physiol* 131(4):1661–1670.
- Scott RJ, Spielman M, Bailey J, Dickinson HG (1998) Parent-of-origin effects on seed development in *Arabidopsis thaliana*. *Development* 125(17):3329–3341.
- Goffman FD, Alonso AP, Schwender J, Shachar-Hill Y, Ohlrogge JB (2005) Light enables a very high efficiency of carbon storage in developing embryos of rapeseed. *Plant Physiol* 138(4):2269–2279.
- Raissig MT, Baroux C, Grossniklaus U (2011) Regulation and flexibility of genomic imprinting during seed development. *Plant Cell* 23(1):16–26.
- Costa LM, et al. (2012) Maternal control of nutrient allocation in plant seeds by genomic imprinting. *Curr Biol* 22(2):160–165.
- Haig D, Westoby M (1991) Genomic imprinting in the endosperm: Its effect on seed development in crosses between species, and between different ploidies of the same species, and its implications for the evolution of apomixis. *Philos T R Soc B* 333(1266): 1–13.
- Gehring M, Bubbs KL, Henikoff S (2009) Extensive demethylation of repetitive elements during seed development underlies gene imprinting. *Science* 324(5933): 1447–1451.
- Hsieh TF, et al. (2009) Genome-wide demethylation of *Arabidopsis* endosperm. *Science* 324(5933):1451–1454.
- Lu J, Zhang C, Baulcombe DC, Chen ZJ (2012) Maternal siRNAs as regulators of parental genome imbalance and gene expression in endosperm of *Arabidopsis* seeds. *Proc Natl Acad Sci USA* 109(14):5529–5534.
- Baroux C, Spillane C, Grossniklaus U (2002) Evolutionary origins of the endosperm in flowering plants. *Genome Biol* 3(9):reviews1026.
- Friedman WE (2001) Developmental and evolutionary hypotheses for the origin of double fertilization and endosperm. *C R Acad Sci III* 324(6):559–567.
- Floyd Sandra K, Friedman William E (2000) Evolution of endosperm developmental patterns among basal flowering plants. *Int J Plant Sci* 161(S6):S57–S81.
- Irizarry RA, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4(2):249–264.
- Yamagishi K, et al. (2005) TANMEI/EMB2757 encodes a WD repeat protein required for embryo development in *Arabidopsis*. *Plant Physiol* 139(1):163–173.
- Orlando DA, Brady SM, Koch JD, Dinneny JR, Benfey PN (2009) Manipulating large-scale *Arabidopsis* microarray expression data: identifying dominant expression patterns and biological process enrichment. *Methods Mol Biol* 553:57–77.

# Supporting Information

Belmonte et al. 10.1073/pnas.1222061110

## SI Materials and Methods

**Plant Materials and Growth.** Wild-type *Arabidopsis* were grown in a peat-based medium (Sunshine Mix #1) at 22 °C in 50–70% relative humidity under constant light (100–150  $\mu\text{E}/\text{m}^2\cdot\text{s}^{-1}$ ). Plant material was harvested between 2:00 PM and 5:00 PM for consistency.

**Laser-Capture Microdissection.** Whole siliques or seeds dissected from siliques were collected and immediately fixed in 3:1 95% (vol/vol) ethanol:acetic acid at 4 °C under RNase-free conditions (1). Whole siliques were cut into ~3-mm segments before fixation to promote fixative penetration. The material was then vacuum-infiltrated for 30 min and fixed overnight at 4 °C. The plant material was rinsed three times with 70% (vol/vol) ethanol, dehydrated in a graded ethanol series (70%, 85%, 95%, 100%, 100% ethanol), and infiltrated with xylenes (1:3, 1:1, 3:1 xylenes: ethanol, followed by 100% xylenes twice). Samples were incubated with paraffin chips overnight at room temperature, at 42 °C for at least 1 h, and at 60 °C for 30 min. The xylenes-paraffin mixture was then replaced with 100% paraffin, and samples were infiltrated with changes in paraffin for 2–3 d at 60 °C.

Seeds or siliques were sectioned at either a 5- or 7- $\mu\text{m}$  thickness using a Leica RM2125RT rotary microtome (Leica Microsystems) and mounted on RNase-free polyethylene naphthalate (PEN)-membrane slides (Leica Microsystems). Use of 5- to 7- $\mu\text{m}$  sections minimized contamination from adjacent and underlying cell types during microdissection. Slides were dried at room temperature and deparaffinized twice in 100% xylenes for 1 min.

Each seed subregion was microdissected independently to minimize contamination from adjacent cell and tissue types (Fig. S1 F–Q). Subregions were captured through sequential serial sections to obtain an accurate representation of all mRNAs within a subregion. For example, serial sections encompassing the entire embryo proper of mature green-stage seeds were captured. The one exception is that the embryo proper (EP) and suspensor (SUS) of globular-stage seeds were microdissected from medial sections to avoid endosperm contamination. We also captured serial sections of entire whole seeds at each developmental stage. Two biological replicates were captured for each subregion at each developmental stage, with the exception of the globular-stage chalazal endosperm (CZE), and heart-stage CZE and chalazal seed coat (CZSC) for which three biological replicates were obtained. Each biological replicate consisted of captured subregions from at least 10 seeds. All subregions were captured within 1 mo of fixation to maximize RNA quality.

**Affymetrix GeneChip Hybridization Experiments.** Microdissected subregions were harvested directly into RNA extraction buffer and total RNA was extracted (RNAqueous-Micro; Ambion). Following treatment of the samples on the RNA purification column with RNase-free DNase (1:4 dilution of DNase I in RDD buffer; Qiagen), RNA levels were quantified (Quant-iT Ribogreen RNA Assay Kit; Invitrogen) using a ND-3330 Fluorespectrometer (Nano-Drop). The numbers of captured sections per bioreplicate ranged from 53 to 2,167, and total RNA yields were between 5 and 66 ng depending on the seed subregion, as detailed in Dataset S1, Table S2. Total RNA was analyzed by microcapillary electrophoresis (RNA 6000 Pico Chip, Agilent 2100 BioAnalyzer; Agilent Technologies), using whole-mount silique RNA as a control.

Two to 6 ng of total RNA was converted to cDNA using a linear amplification method (WT-Ovation Pico RNA Amplification

System; NuGEN Technologies), and the cDNA was fragmented and labeled (FL-Ovation cDNA Biotin Module V2; NuGEN Technologies). Five micrograms of amplified cDNA was hybridized with the Affymetrix GeneChip ATH1 *Arabidopsis* Genome Array as described by Le et al. (2).

To determine the number of distinct mRNAs in a seed subregion, we normalized GeneChip hybridization data and assigned present, absent, and marginal signal detection calls using MAS 5.0 software (Affymetrix) (3). mRNAs were designated as detected in a population if their signal detection calls were present in either both biological replicates or in at least two of three of the replicates. Signal detection calls and relative levels for mRNAs in all GeneChip experiments are given in Dataset S2.

**Global Comparisons of mRNA Populations.** Hierarchical clustering and bootstrapping analysis was conducted with RMA-normalized data for all GeneChip biological replicates using the pvclust package with default settings (<http://cran.r-project.org/web/packages/pvclust/pvclust.pdf>) (4). Principal component analysis was carried out on RMA-normalized, replicate averaged data using the prcomp function in R (5).

**Identification of Coexpressed Gene Sets. Subregion-specific and region-specific mRNAs.** A mRNA specific to a seed subregion was defined as one whose relative level is at least fivefold higher and significantly different ( $q < 0.001$ , mixed-model ANOVA) than those detected in all other subregions at a given developmental stage. Lists of subregion-specific mRNAs and their overrepresentation for Gene Ontology (GO) terms, DNA motifs, and metabolic pathways are given in Dataset S3.

Seed regions are defined as the embryo, consisting of the EP and SUS, the endosperm, consisting of the micropylar (MCE), peripheral (PEN), and CZE, and the seed coat, consisting of the SC and CZSC. Region-specific mRNAs are present at a fivefold or higher level in all subregions of a region and significantly different ( $q < 0.001$ , mixed-model ANOVA) than in all other seed subregions. Embryo region-specific mRNAs were only identified at the globular stage because the SUS subregion was not analyzed at other developmental stages. Lists of region-specific mRNAs and their enrichment for GO terms, DNA motifs, and metabolic pathways are given in Dataset S3.

**Stage-specific mRNAs.** A mRNA specific to a developmental stage is one whose level is fivefold higher and significantly different ( $q < 0.001$ , mixed-model ANOVA) at one developmental stage relative to all other stages in a given seed subregion. Stage-specific mRNAs and their enriched GO terms, DNA motifs, and metabolic pathways are listed in Dataset S3.

**Dominant expression pattern identification.** Dominant expression patterns (DPs) were identified essentially as previously described (6). RMA-normalized and averaged data from all seed subregions and developmental stages, but not from whole seeds, were filtered to remove probesets that did not exceed a minimum expression cutoff (RMA value of 15, >75% of RNAs designated present by MAS 5.0 analysis exceeded this cutoff value) in at least one seed subregion and developmental stage. To identify DPs, the sample variance of the remaining probesets were ranked, and the 50% most variant, corresponding to 8,047 RNAs, were retained for clustering analysis. Filtered data were clustered using the FKM implementation FANNY (<http://cran.r-project.org/web/packages/cluster/cluster.pdf>) (7) with a  $K$  of 50 and a probability for cluster membership ( $m$ -value = 0.44) that resulted in most probesets being assigned to a single cluster.



Attempts to use smaller  $K$  values eliminated clusters with significant patterns, whereas use of larger  $K$  values did not generate additional clusters with novel RNA accumulation patterns. The median RNA accumulation pattern for all mRNAs in a cluster was determined, and the 50 patterns were subjected to hierarchical clustering using  $1 - r$  ( $r$  = Pearson's correlation coefficient) as the distance metric. Members of distinct clusters that shared significant similarity (separated by a node with a tree height of less than 0.15 in the dendrogram) were combined, and a new median RNA accumulation pattern was determined and reanalyzed using hierarchical clustering. These procedures generated 47 different median RNA accumulation patterns. mRNAs were then reassigned to clusters based on correlation with the 47 median accumulation patterns. mRNAs whose levels were above the minimum expression cutoff, variance were among the top 75%, and accumulation pattern correlated strongly (Pearson's correlation  $> 0.8$ ) with a median RNA accumulation pattern were assigned to that cluster. Coexpressed gene sets corresponding to DPs contained an average of 104 mRNAs, ranging between 3 and 508. Lists of mRNAs and their enrichment for GO terms, DNA motifs, and metabolic pathways are given in [Dataset S4](#).

**Seed-Specific mRNAs.** Seed-specific mRNAs were those that are called as: (i) present in all or a majority of replicates for at least one seed subregion; and (ii) absent in all replicates of reproductive (ovules and floral buds) and vegetative organs (leaf, stem, root, and seedling) (2).

**Quantitative RT-PCR Experiments.** Results of GeneChip hybridization experiments were validated using quantitative RT-PCR (qRT-PCR) experiments. PCR amplification reactions and data analysis was done as described previously (8), except that 100 pg of amplified cDNA derived from microdissected seed subregions were used and data were normalized to *PP2AA3* levels, (At1g13320) (9). *PP2AA3* RNA levels were relatively constant in all seed subregions and developmental stages. Primer pairs for amplification of specific mRNAs were designed using Beacon Designer 3 (Premier Biosoft International). The DNA sequences and efficiencies of primer pairs used to validate GeneChip data and the corresponding relative mRNA levels are provided in [Dataset S1, Table S3](#).

**ChipEnrich.** We modified the ChipEnrich software program (10) to identify GO terms, metabolic pathways, transcription-factor families, and DNA sequence motifs that are overrepresented in coexpressed gene sets and to discover potential transcriptional modules. This Java program was developed originally to identify significantly enriched GO terms (2009 download) and transcription factor families from gene lists. Significance of enrichment is reported as  $P$  values calculated from the hypergeometric distribution (11) using the Apache Commons Math library (<http://jakarta.apache.org/commons/math>). The following functions were added to ChipEnrich which is available at <http://seedgenenetwork.net/presentation#software>.

**Metabolic pathway enrichment analysis.** Genes represented on the ATH1 GeneChip were annotated according to metabolic pathways described in the PATHWAYS database from AraCyc (2008 download). Enrichment was defined as the ratio of: (i) the number of AGI locus identifiers in the query list annotated as belonging to a pathway to (ii) the number of AGI locus identifiers associated with the pathway on the GeneChip compared with the ratio of (iii) the total number of AGI locus identifiers present in the query list to (iv) the total number of AGI locus identifiers present on the GeneChip.

**DNA motif enrichment analysis.** Gene sets were analyzed to identify enriched DNA sequence motifs known to interact with transcription factors (*Arabidopsis* Gene Regulation Information Server, [\[arabidopsis.med.ohio-state.edu/\]\(http://arabidopsis.med.ohio-state.edu/\), August 2009\) that are located in the region 1-kb upstream of the gene's transcription start site \(TAIR9, \[www.arabidopsis.org\]\(http://www.arabidopsis.org\)\) as described by others \(12, 13\). The background distribution was determined by identifying DNA motifs for all genes represented as singletons on the \*Arabidopsis\* ATH1 GeneChip \(see \[Dataset S1, Table S6\]\(#\) for a list of all DNA motifs used in this study\). Statistical enrichment \( \$P\$  value  \$< 0.001\$ \) was determined for each gene list using the hypergeometric distribution. Enriched DNA sequence motifs were also identified among genes overrepresented for a GO term within a gene list.](http://</a></p></div><div data-bbox=)

**Putative transcriptional modules.** To discover putative transcriptional modules, we associated significantly enriched DNA sequence motifs with transcription factors known or predicted to bind the motifs. We used known interactions between transcription factors and DNA motifs specified in AtcisDB (14) and defined by others in the literature and assumed that transcription factors of a particular family bind to the same DNA motif (6). Two variations of this approach were used. In the first approach, we associated DNA motifs significantly enriched within a coexpressed gene set with their cognate transcription factors that were included in the coexpressed gene set. In the second variation, we identified DNA motifs that were significantly enriched for genes corresponding to an overrepresented GO term and associated coexpressed transcription factors known or predicted to bind the enriched DNA motifs. Overrepresented GO terms, DNA motifs, and their associated transcription factors were compiled into two Cytoscape compatible files that were used as network and node attribute files, and the modules were visualized with Cytoscape. All files, including the network and node attribute files, used to generate the transcriptional modules, are found in [Dataset S1, Table S6](#).

Outputs are summarized in a text file, <significant.txt>, in which the gene set name is in the first column, enriched GO terms, DNA motifs, or transcription factor families are listed in the second column, and  $P$  values indicating the significance of enrichment are given in the third column. Each unique enriched category is set in a new row. If a DNA motif is significantly overrepresented within a gene list ( $P < 0.001$ ), it is also determined if the motif is enriched among genes significantly overrepresented for a GO term ( $P < 0.001$ ). In the <significant.txt> file, the overrepresented GO terms are listed in the first column, enriched DNA motifs are in the second column, and  $P$  values are in the third column. Transcription factors in the gene list (first column) that are predicted or known to bind with enriched DNA motifs (second column) are also listed in the <significant.txt> file. A separate node attribute file is also provided from ChipEnrich that describes whether a node (first column of <significant.txt> file) is a pattern, GO term, DNA motif, transcription factor family, or transcription factor.

The <significant.txt> file is designed to be used as the network file for the network graphing software, Cytoscape (version 2.6.3, [www.cytoscape.org](http://www.cytoscape.org)). The <node.txt> file is used as the attributes file (15).  $P$  values are also imported with the network file as edge attributes. For visualization purposes, a thicker line represents a lower  $P$  value, a dashed line represents a transcription factor with a predicted binding interaction, and a solid red edge is an experimentally determined transcription factor–DNA motif interaction. All files, including the network and attributes files, used to generate the transcriptional modules are found in [Dataset S1, Table S6](#).

**Analysis of GFP Activity.** Activities of selected promoters were evaluated using promoter-GFP chimeric genes generated previously (16). Transgenic seeds with GFP reporter genes were analyzed using a Zeiss Axioskop 2 plus compound microscope equipped with a FS10 FITC filter set (exciter PB 450–490; Carl Zeiss) (16). An exposure time of 500 ms was used for all images.

Seeds were analyzed at all stages of development, although images shown in Fig. S2 are primarily of globular-stage seeds. Seeds from at least four plants were examined for each promoter-GFP line.

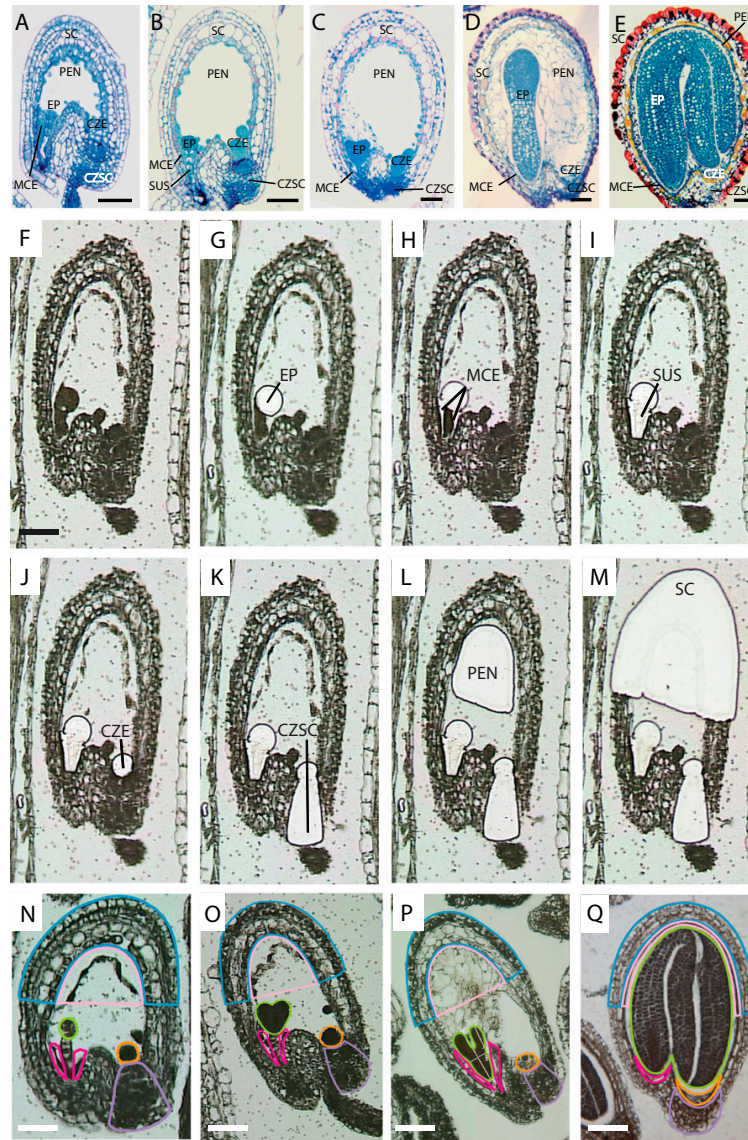
**Histology.** For light microscopy, samples were fixed in 2.5% (vol/vol) glutaraldehyde and 1.6% (wt/vol) paraformaldehyde buffered with 0.05 M phosphate buffer, pH 6.9, dehydrated with methyl cellosolve followed by two changes of absolute ethanol, and then infiltrated and embedded in Histo-resin (Leica) according to the methods of Yeung (17). Three micrometer-thick serial sections were stained with periodic acid-Schiff for total carbohydrates and counterstained with amido black 10B for protein or with Toluidine blue O for general observations.

Chloroplasts were localized by auto fluorescence using 633-nm excitation and a 650-nm filter using a Zeiss-700 confocal scanning laser microscope. Seeds from at least three different plants were used for observations. LSM Zen imaging software (Carl Zeiss) was used to construct 3D images from 40 optical slices in the z-dimension (z-stack). No other image enhancement was performed.

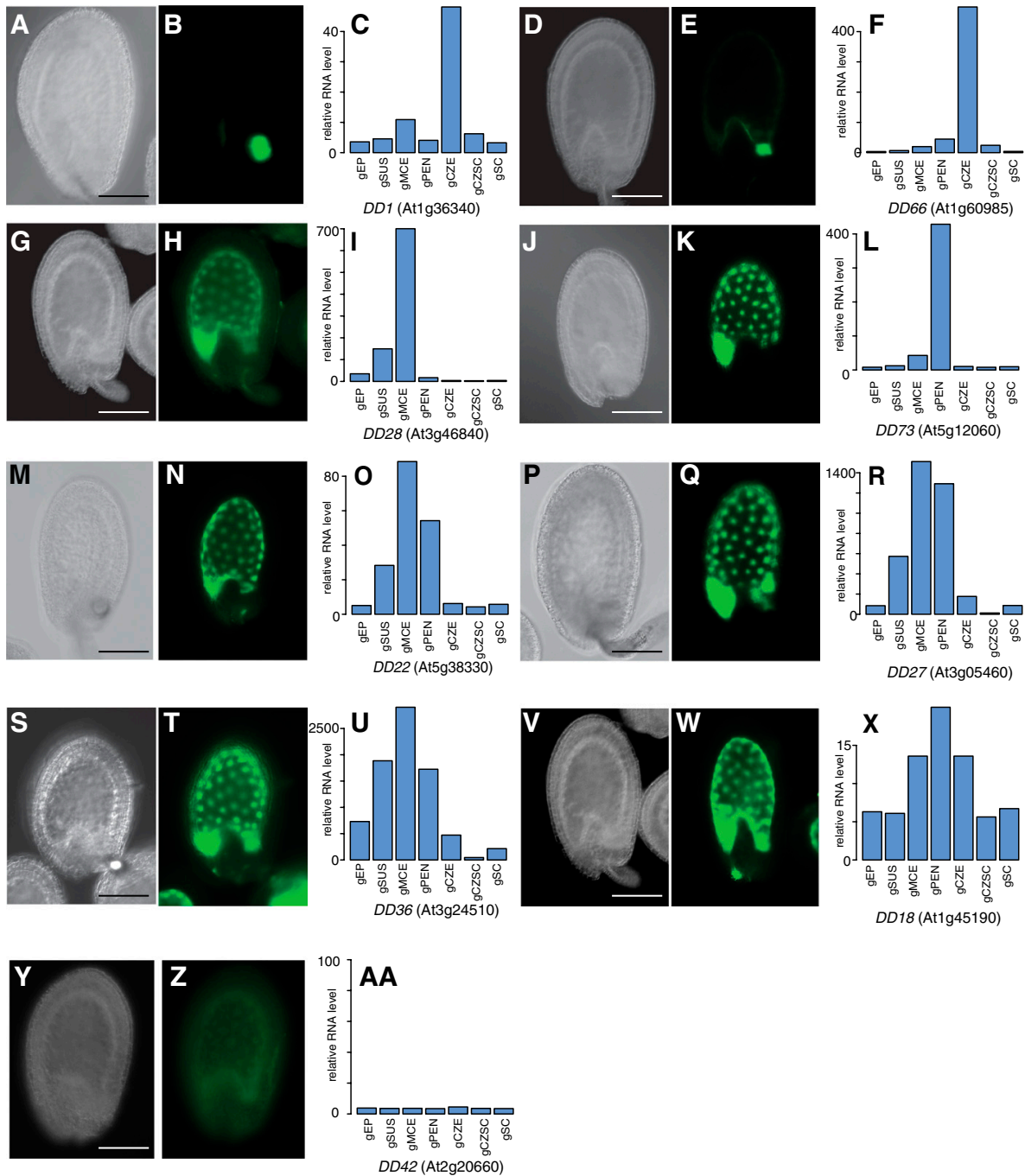
Oil bodies and chloroplast ultrastructure were visualized using transmission electron microscopy. Seeds were fixed in 2% (wt/vol) paraformaldehyde, 2.5% (vol/vol) glutaraldehyde and 0.1 M sodium phosphate buffer, pH 7.2 (Karnovsky's fixative), using a

PELCO Biowave microwave (Ted Pella) under vacuum (20 psi) as follows: 5 min at 000 W, 10 s at 200 W, 20 s at 155 W, 10 s at 250 W. Tissue was stored in Karnovsky's fixative until it was processed further. Tissue was rinsed with 0.05 M sodium phosphate buffer, pH 6.9, and postfixed with 1% (vol/vol) osmium tetroxide in 0.05 M sodium phosphate buffer, pH 6.9 for 2 h followed by microwave fixation under vacuum for 40 s at 250 W. The tissue was then incubated in 0.1% aqueous tannic acid for 30 min and rinsed and stained with 2% (wt/vol) aqueous uranyl acetate. Dehydration was accomplished by immersing the tissue for 20 min each in three changes of 95% (vol/vol) acetone followed by two changes of 100% acetone. The tissue was infiltrated in 3:1 and 2:1 acetone:Spurr's resin for at least 1 h each, 1:1 acetone:resin overnight, 1:2 acetone:resin for 24 h, and in two changes of pure resin for 24 h each. Samples were embedded in capsules and polymerized overnight at 70 °C. Thick sections were cut on a Leica Ultracut UCT Ultramicrotome (Leica Microsystem) at 400 nm and stained with Methylene blue and Azure B. Thin sections were cut using a Diatome Diamond Knife (Diatome) at 60–90 nm, transferred to copper grids and double-stained with 4% (wt/vol) alcoholic uranyl acetate and lead citrate. The sections were viewed on a Philips CM120 Biotwin Lens (FEI) and images were captured using a Gatan BioScan camera.

1. Kerk NM, Ceserani T, Tausta SL, Sussex IM, Nelson TM (2003) Laser capture microdissection of cells from plant tissues. *Plant Physiol* 132(1):27–35.
2. Le BH, et al. (2010) Global analysis of gene activity during *Arabidopsis* seed development and identification of seed-specific transcription factors. *Proc Natl Acad Sci USA* 107(18):8063–8070.
3. Hubbell E, Liu WM, Mei R (2002) Robust estimators for expression analysis. *Bioinformatics* 18(12):1585–1592.
4. Suzuki R, Shimodaira H (2006) Pvcust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22(12):1540–1542.
5. Team RDC (2012) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).
6. Brady SM, et al. (2007) A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science* 318(5851):801–806.
7. Kaufman L, Rousseeuw PJ (1990) *Finding Groups in Data: An Introduction to Cluster Analysis* (Wiley, New York).
8. Yamagishi K, et al. (2005) TANMEI/EMB2757 encodes a WD repeat protein required for embryo development in *Arabidopsis*. *Plant Physiol* 139(1):163–173.
9. Czechowski T, Stitt M, Altmann T, Udvardi MK, Scheible WR (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in *Arabidopsis*. *Plant Physiol* 139(1):5–17.
10. Orlando DA, Brady SM, Koch JD, Dinneny JR, Benfey PN (2009) Manipulating large-scale *Arabidopsis* microarray expression data: Identifying dominant expression patterns and biological process enrichment. *Methods Mol Biol* 553:57–77.
11. Gadbury GL, Garrett KA, Allison DB (2009) Challenges and approaches to statistical design and inference in high-dimensional investigations. *Methods Mol Biol* 553:181–206.
12. O'Connor TR, Dyreson C, Wyrick JJ (2005) Athena: A resource for rapid visualization and systematic analysis of *Arabidopsis* promoter sequences. *Bioinformatics* 21(24):4411–4413.
13. Vandepoele K, Quimbaya M, Casneuf T, De Veylder L, Van de Peer Y (2009) Unraveling transcriptional control in *Arabidopsis* using cis-regulatory elements and coexpression networks. *Plant Physiol* 150(2):535–546.
14. Davuluri RV, et al. (2003) AGRIS: *Arabidopsis* gene regulatory information server, an information resource of *Arabidopsis* cis-regulatory elements and transcription factors. *BMC Bioinformatics* 4:25.
15. Cline MS, et al. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2(10):2366–2382.
16. Steffen JG, Kang IH, Macfarlane J, Drews GN (2007) Identification of genes expressed in the *Arabidopsis* female gametophyte. *Plant J* 51(2):281–292.
17. Yeung EC (1999) The use of histology in the study of plant tissue culture systems: Some practical comments. *In Vitro Cell Dev Biol Plant* 35(2):137–143.

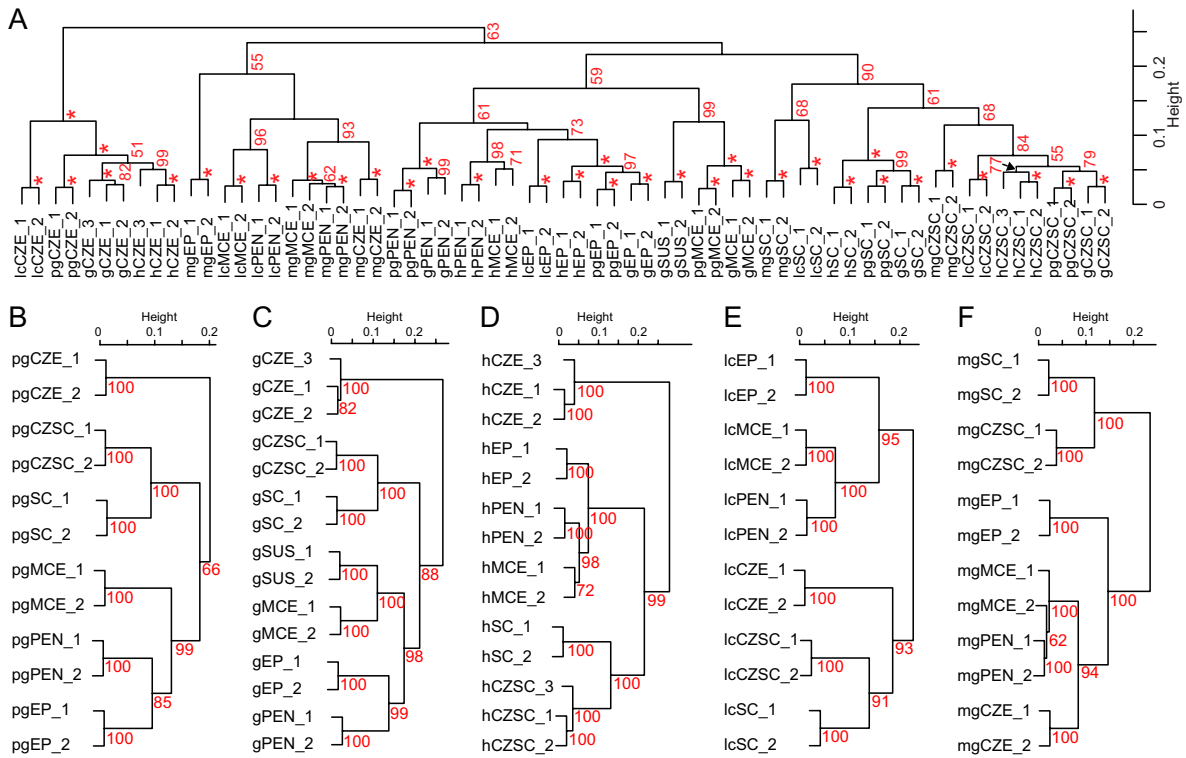


**Fig. S1.** Microdissection of seed subregions during *Arabidopsis* seed development. (A–E) Longitudinal sections of developing *Arabidopsis* seeds across five stages of development: (A) preglobular stage, (B) globular stage, (C) heart stage, (D) linear cotyledon stage, and (E) mature-green stage. (F–M) Order of subregion microdissection from a globular-stage seed. Medial longitudinal sections through the embryo proper and suspensor of a globular-stage embryo are shown. (N–Q) Outline of subregions captured at the (N) preglobular, (O) heart, (P) linear cotyledon, and (Q) mature-green stages of seed development. Subregions outlined are EP (green), MCE (dark pink), PEN (light pink), CZE (orange), CZSC (purple), and SC (blue). Abbreviations are given in Table 1. (Scale bars: A–E, 50  $\mu$ m; F–M, 45  $\mu$ m; N, 40  $\mu$ m; O, 60  $\mu$ m; P, 85  $\mu$ m; Q, 120  $\mu$ m.)

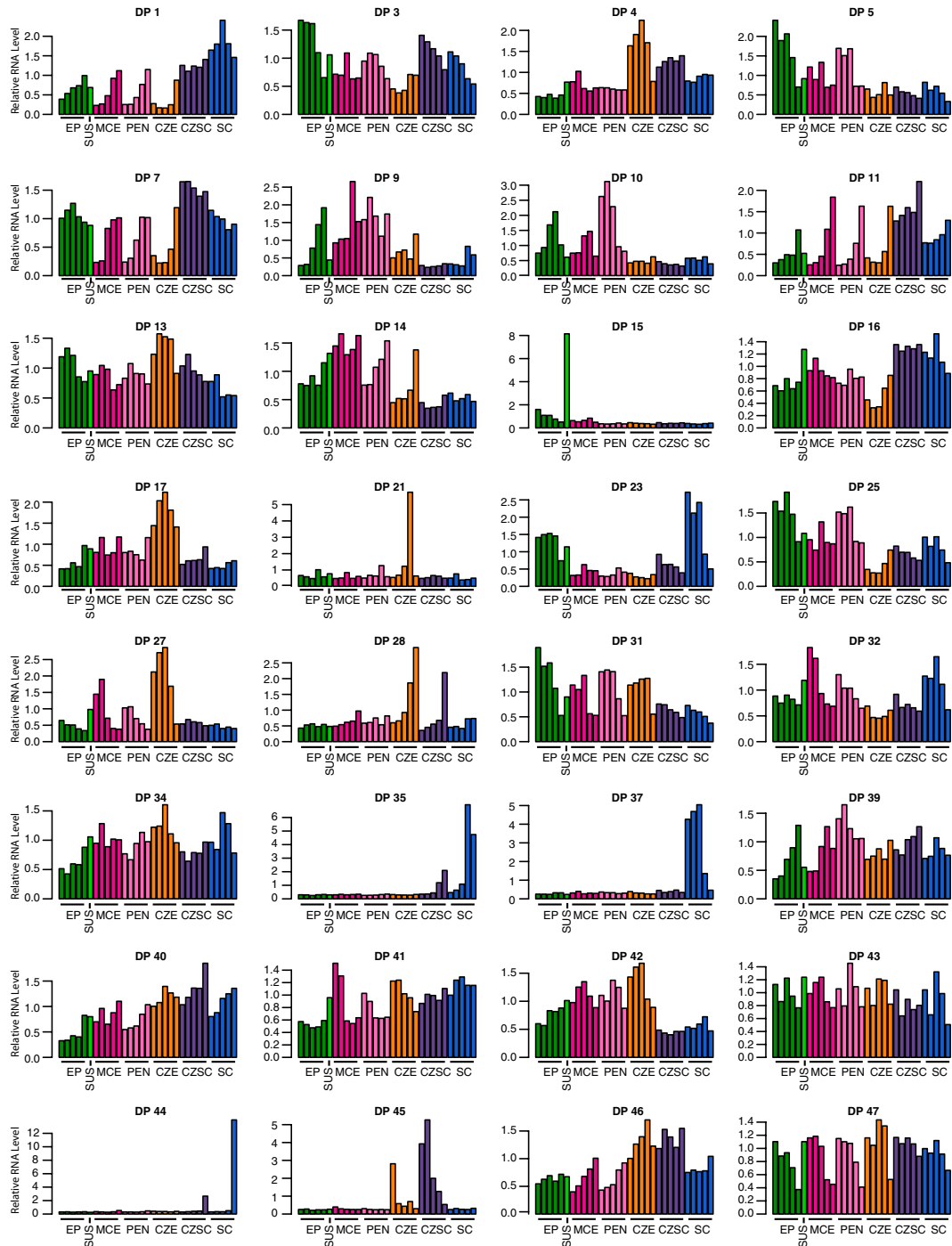


**Fig. S2.** Comparison of relative RNA levels and promoter activities of endosperm-expressed genes. (B, E, H, K, N, Q, T, W, Z) GFP activity of (A, D, G, J, M, P, S, V, Y) globular-stage seeds containing reporter genes fused with the indicated promoters. (C, F, I, L, O, R, U, X, AA) Relative RNA levels as determined by GeneChip hybridization experiments. (A–C) *DD1* (At1g36340) and (D–F) *DD66* (At1g60985) are active specifically in the CZE. (G–I) *DD28* (At3g46840) is active primarily in the MCE and SUS. (J–L) *DD73* (At5g12060), (M–O) *DD22* (At5g38330), and (P–Q) *DD27* (At3g05460) are active primarily in the MCE and PEN. (S–U) *DD36* (At3g24510) and (V–X) *DD18* (At1g45190) are active in all three endosperm subregions. (Y–AA) Promoter activity and mRNA accumulation for the negative control gene, *DD42* (At2g20660). Exposure times for fluorescence micrographs were 500 ms. (Scale bar, 100  $\mu$ m.)

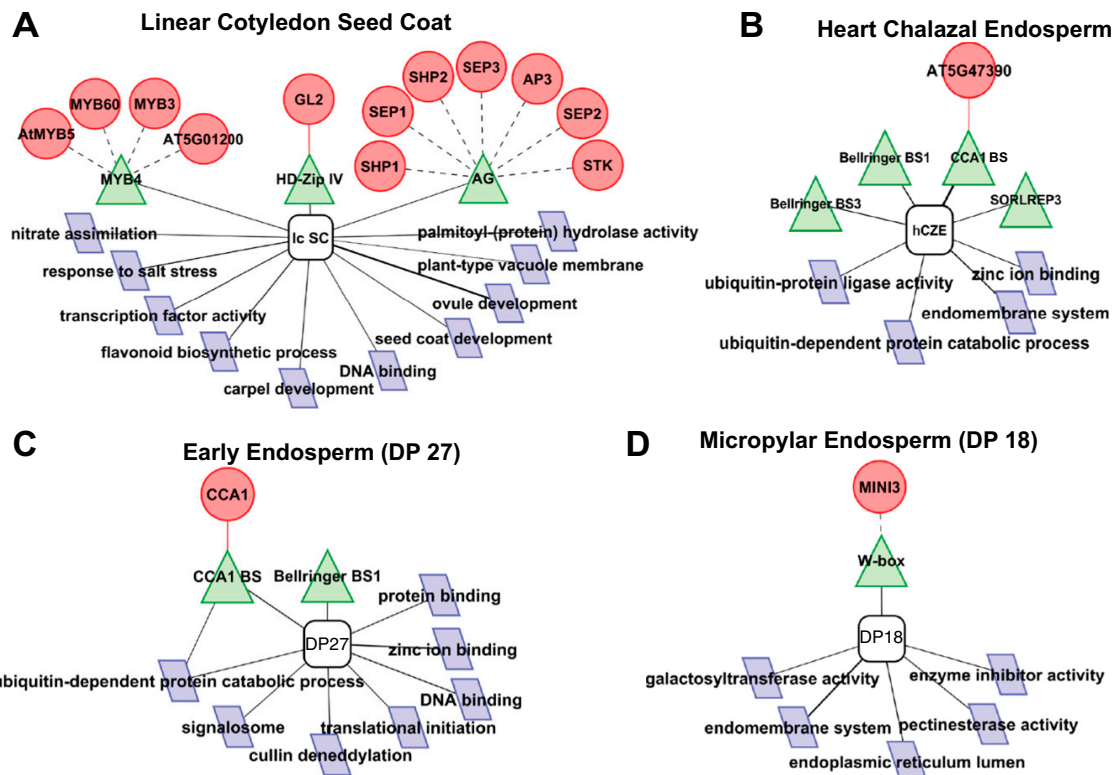




**Fig. S3.** Hierarchical clustering of seed subregion mRNA populations. (A) Correlation-based hierarchical clustering of all biological replicates of mRNAs populations in seed subregions. (B–F) Hierarchical clustering of mRNA populations at the (B) preglobular, (C) globular, (D) heart, (E) linear-cotyledon, and (F) mature-green stages. Bootstrap values are shown in red. Asterisks indicate a bootstrap value of 100. Numbers in the sample name indicate biological replicates.



**Fig. 54.** DPs of gene expression during seed development. DPs that are not shown in Fig. 3. These patterns were defined with Fuzzy *K* means clustering of the 50% most variant mRNAs in all combinations of seed subregions and stages.



**Fig. S5.** Predicted transcriptional modules of coexpressed gene sets. Squircles represent sets of coexpressed genes, parallelograms and triangles depict significantly enriched ( $P < 0.001$ , hypergeometric distribution) GO terms and DNA motifs, respectively, and circles correspond to coexpressed transcription factors predicted or known to interact with DNA motifs. Abbreviated GO terms are given in [Dataset S1, Table S6](#). (A) MYB, HD-ZIP, and MAD5 transcriptional modules based on mRNAs that accumulate specifically in the seed-coat region at the linear-cotyledon stage. (B) A CCA1-like transcription factor is associated with the CCA1 binding site among genes expressed specifically in the heart CZE. (C) Genes expressed in all endosperm subregions early in development (DP 27) are the basis for a transcriptional module with the CCA1 transcription factor associated with the CCA1 DNA motif. We previously reported overrepresentation of the CCA1 DNA motif for transcription factor mRNAs that accumulate specifically in the CZE (2). (D) A MINISEED1 module linking the transcription factor with the enriched W-box DNA motif in a MCE-specific gene set (DP 18).

## Other Supporting Information Files

[Dataset S1 \(XLS\)](#)  
[Dataset S2 \(XLSX\)](#)  
[Dataset S3 \(XLSX\)](#)  
[Dataset S4 \(XLSX\)](#)

## APPENDIX B

Valdés-López, O., Thibivilliers, S., Qiu, J., Xu, W.W., Nguyen, T.H.N., Libault, M., **Le, B.H.**, Goldberg, R.B., Hill, C.B., Hartman, G.L., et al. (2011). **Identification of quantitative trait loci controlling gene expression during the innate immunity response of soybean.** *Plant Physiol* 157, 1975–1986.



# Identification of Quantitative Trait Loci Controlling Gene Expression during the Innate Immunity Response of Soybean<sup>1[W][OA]</sup>

Oswaldo Valdés-López, Sandra Thibivilliers, Jing Qiu, Wayne Wenzhong Xu, Tran H.N. Nguyen, Marc Libault<sup>2</sup>, Brandon H. Le, Robert B. Goldberg, Curtis B. Hill, Glen L. Hartman, Brian Diers, and Gary Stacey\*

Department of Statistics (J.Q.) and Divisions of Biochemistry and Plant Sciences, National Center for Soybean Biotechnology, C.S. Bond Life Sciences Center (O.V.-L., S.T., T.H.N.N., M.L., G.S.), University of Missouri, Columbia, Missouri 65211; Minnesota Supercomputing Institute for Advanced Computational Research, University of Minnesota, Minneapolis, Minnesota 55455 (W.W.X.); Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, California 90095 (B.H.L., R.B.G.); and United States Department of Agriculture-Agricultural Research Service (G.L.H.) and Department of Crop Sciences (C.B.H., G.L.H., B.D.), University of Illinois, Urbana, Illinois 61801

Microbe-associated molecular pattern-triggered immunity (MTI) is an important component of the plant innate immunity response to invading pathogens. However, most of our knowledge of MTI comes from studies of model systems with relatively little work done with crop plants. In this work, we report on variation in both the microbe-associated molecular pattern-triggered oxidative burst and gene expression across four soybean (*Glycine max*) genotypes. Variation in MTI correlated with the level of pathogen resistance for each genotype. A quantitative trait locus analysis on these traits identified four loci that appeared to regulate gene expression during MTI in soybean. Likewise, we observed that both MTI variation and pathogen resistance were quantitatively inherited. The approach utilized in this study may have utility for identifying key resistance loci useful for developing improved soybean cultivars.

Plants, like animals, are constantly under threat from different pathogens. To counter this threat, plants have developed a sophisticated system to detect pathogens and trigger a strong defense response (Jones and Dangl, 2006; Boller and Felix, 2009; Segonzac and Zipfel, 2011). The plant immune system is generally considered to have two separate, but interacting components (Jones and Dangl, 2006; Zipfel and Robatzek, 2010). In one component, using systems analogous to those used in animals, the plant recognizes the invading pathogen through detection of conserved structural motifs, termed microbe- or pathogen-associated

molecular patterns (MAMPs or PAMPs, respectively). MAMPs are conserved molecules that may be essential for the pathogen's life cycle. Well-studied examples include lipopolysaccharides, chitin, flagellin (flg), and translation elongation factor Tu (EF-Tu; Boller and Felix, 2009; Segonzac and Zipfel, 2011; Thomma et al., 2011). MAMPs are recognized at the plant cell surface via pattern recognition receptors (PRRs). Once MAMPs are detected, a defense response, termed MAMP-triggered immunity (MTI), is mounted that leads to plant resistance (Zipfel, 2008; Katagiri and Tsuda, 2010). MTI generally produces broad resistance against a variety of nonadapted pathogens; however it is usually weak and can be overcome by the adapted pathogen. Perhaps due to this fact, plants have also developed a second means to recognize pathogens that results in stronger but more specific resistance. In this case, the plant recognizes specific effector proteins produced by the pathogen and often directly inserted into the plant's cytoplasm (Göhre and Robatzek, 2008; Katagiri and Tsuda, 2010). Recognition of the effector triggers this second line of plant defense against pathogens, hence the name effector-triggered immunity (ETI; Göhre and Robatzek, 2008; Segonzac and Zipfel, 2011). ETI is often associated with an accelerated and amplified MTI response and a hypersensitive cell death response at the infection site (Jones and Dangl, 2006). However, since effector proteins are species, race, pathotype, or strain specific and

<sup>1</sup> This work was supported by the U.S. Department of Energy, Office of Basic Energy Sciences (grant no. DE-FG02-08ER15309 to G.S.), the United Soybean Board (to G.S., B.D., and G.L.H.), the North Central Soybean Research Program (to G.S.), as well as a National Science Foundation Plant Genome Grant (to R.B.G.).

<sup>2</sup> Present address: Department of Botany and Microbiology, University of Oklahoma, Norman, OK 73019.

\* Corresponding author; e-mail stacey@missouri.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors ([www.plantphysiol.org](http://www.plantphysiol.org)) is: Gary Stacey (stacey@missouri.edu).

<sup>[W]</sup> The online version of this article contains Web-only data.

<sup>[OA]</sup> Open Access articles can be viewed online without a subscription.

[www.plantphysiol.org/cgi/doi/10.1104/pp.111.183327](http://www.plantphysiol.org/cgi/doi/10.1104/pp.111.183327)

often require a specific resistance protein for their recognition by the host, ETI is quite specific, often protecting plants from only specific genotypes of any given pathogen species (Segonzac and Zipfel, 2011).

MTI is characterized by a variety of plant responses, which include changes in ion flux across the plasma membrane, changes in cytoplasmic calcium levels, production of reactive oxygen species (ROS), callose deposition, modification of phytohormone concentrations, and induction or repression of different plant genes that condition antibiotic molecules, such as the phytoalexin glyceollin produced by soybean (*Glycine max*; Zipfel, 2009; Boudsocq et al., 2010; Lygin et al., 2010; Mersmann et al., 2010; Luna et al., 2011). These responses can be detected within minutes to hours after MAMP treatment, and the magnitude of the response may be plant-species dependent (Segonzac and Zipfel, 2011). The MTI response can be regulated at the transcriptional, posttranscriptional, and post-translational levels through chromatin modification, transcription factors, noncoding RNAs, and changes in protein glycosylation patterns (Alvarez et al., 2010; Zhang et al., 2011). We mention these examples to emphasize the complexity of the MTI response.

Several reports have demonstrated the ability of MTI to protect plants from a variety of bacterial and fungal pathogens (Gómez-Gómez et al., 1999; Segonzac and Zipfel, 2011). MTI is usually expressed as partial resistance and segregates as a quantitative trait that is inherited in an oligogenic or a polygenic manner with genetic effects being largely additive. In contrast, specific resistance (ETI) typically confers complete resistance and is usually pathotype specific and conditioned by single dominant resistance genes. However, ETI can be overcome by adaptation of the pathogen population through selection toward pathotypes that produce effectors not recognized by this system. In comparison to narrowly effective ETI-based resistance, broadly effective MTI-based resistance may be more durable with a broader spectrum of effectiveness against multiple pathogens or pathotypes of a pathogen. Indeed, in some host-parasite systems, such as *Sclerotinia* stem rot (*Sclerotinia sclerotiorum*) disease on soybean, partial resistance may be the only type of resistance available since no effective ETI system is known (Poland et al., 2008; Vuong et al., 2008). In a few cases, a direct link has been established between MTI and specific quantitative trait loci (QTLs) for partial resistance. For example, two major effect QTLs associated with resistance to *Pseudomonas syringae* pv *phaseolicola* were reported in *Arabidopsis thaliana*, one of which is flagellin-sensitive 2 (*FLS2*) that encodes a PRR receptor (Forsyth et al., 2010; Ahmad et al., 2011).

Soybean is a chief source of protein and oil for human and animal consumption, and is grown on about 6% of the world's arable land (Hartman et al., 2011). Soybean production in the United States increased from 75 million metric tons in 2000 to 91 million metric tons in 2010 with a concomitant increase in the value of the crop from \$13 to \$39 billion,

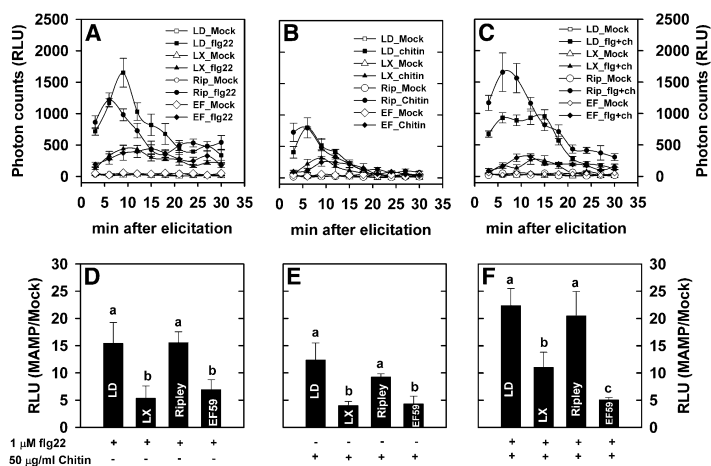
respectively (USDA, 2010). Data compiled for the 2010 growing season indicates that the total yield loss from a variety of diseases totaled approximately 7% of U.S. (<http://aes.missouri.edu/delta/research/soyloss.stm>) production, which equates to approximately \$5 billion.

The MTI response has been widely studied, mainly in the model plant *Arabidopsis* and a few crops, such as tomato (*Solanum lycopersicum*) and rice (*Oryza sativa*). Few studies have examined MTI in soybean. There is significant potential in harnessing soybean MTI to improve cultivars to withstand a variety of pathogens and, therefore, enhance soybean production. In this study, we evaluated the MTI responses in different soybean genotypes by challenging with two known MAMPs: (1) the conserved 22-amino acid peptide from bacterial flg22, representing bacterial pathogens, and (2) crab shell chitin, representing fungal pathogens that contain chitin in their cell walls. We found clear variation between soybean genotypes in their MAMP-triggered oxidative burst, as well as in the expression of MAMP-responsive genes. Based on this analysis and the availability of a well-characterized population of soybean lines, we chose two parental soybean genotypes, one with a strong MTI response (LD00-2817P; Diers et al., 2010) and one with a weak MTI response (LDX01-1-65; Diers et al., 2005). These two genotypes also show significant differences in pathogen susceptibility when infected after MAMP treatment. Analysis of the population of F3 lines from the cross between the two genotypes identified a single QTL associated with MAMP-triggered oxidative burst. Similarly, the analysis of gene expression across the F3 lines led to the identification of three QTLs associated with MAMP-triggered gene expression (i.e. expression QTLs [eQTLs]). These QTLs associated with MTI expression, as well as the general approach utilized in this study, may be useful in the genetic improvement of biotic stress resistance in soybean.

## RESULTS

### Variation in the Oxidative Burst across Soybean Genotypes

The oxidative burst triggered by applications of flg22, chitin, or a mixture of both MAMPs was measured in leaf discs from four different genotypes, originally chosen due to their use as parents of recombinant populations previously genotyped with molecular markers (Kim et al., 2011). This analysis showed significant differences in MAMP-triggered oxidative burst among the four genotypes (Fig. 1, A–C). The data also showed that the genotypes LD00-2817P and Ripley gave the strongest oxidative burst and the genotypes LDX01-1-65 and EF59 gave the weakest responses, which was also observed when the total ROS produced during the oxidative burst (hereafter called as total ROS production) was compared (Fig. 1, D–F). A pairwise



**Figure 1.** Oxidative burst triggered by 1  $\mu$ M flg22 (A and D), 50  $\mu$ g/ml chitin (B and E), or a mixture of 1  $\mu$ M flg22 + 50  $\mu$ g/ml chitin (C and F) in LD00-2817P (LD), LDX01-1-65 (LX), Ripley (Rip), and EF59 (EF) leaf discs measured in relative luminescence units (RLUs). A to C, Time kinetics of the MAMP-triggered oxidative burst. Each data point represents the average of five biological replicates with three technical replicates. Error bars represent  $\pm$  SE of the average. D to F, Total ROS produced during the oxidative burst over 30 min of MAMP treatment. Results are the average  $\pm$  SE of the ratio between MAMP/mock from five biological replicates with three technical replicates for each. Within each diagram, values sharing the common letters are not significantly different at  $P \leq 0.05$  by ANOVA.

comparison of the total ROS produced revealed that LDX01-1-65 treated with MAMP mixture (flg22 + chitin) produced significantly more ROS than EF59; no significant difference in the total ROS was observed between LD00-2817P and Ripley with this treatment, even though Ripley showed a stronger oxidative burst than LD00-2817P over the first 10 min of treatment (Fig. 1, C–F). Likewise, LD00-2817P and Ripley showed a similar oxidative burst and total ROS production when treated with either flg22 or chitin alone (Fig. 1, A and B). However, the oxidative burst and total ROS production triggered by chitin across genotypes, except for EF59, was significantly lower than that triggered by the mixture of MAMPs (Fig. 1, B and C). These results clearly indicate significant genotypic variation in the MTI response among soybean genotypes, which suggests that QTLs associated with this trait may be identified after further genetic analyses.

#### Development of an Affymetrix Soybean Whole-Genome Transcript Array to Evaluate Genome-Wide Gene Expression Changes during the MTI Response

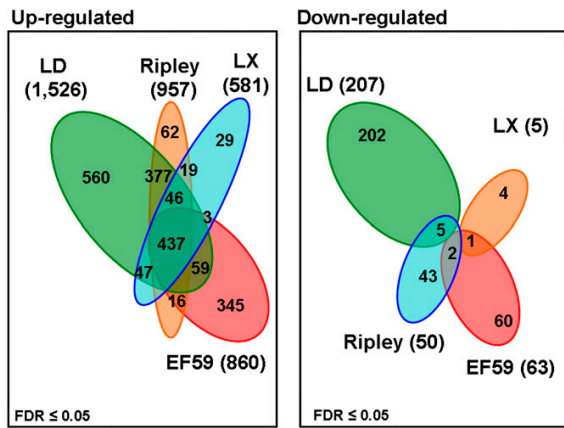
The first-generation Affymetrix Soybean Genome array was derived from a collection of over 350,000 ESTs and was predicted to allow transcriptional analysis of 35,611 distinct genes, or approximately 50% of the soybean genome ([http://media.affymetrix.com/support/technical/datasheets/soybean\\_datasheet.pdf](http://media.affymetrix.com/support/technical/datasheets/soybean_datasheet.pdf)). With publication of the full soybean genome sequence (Schmutz et al., 2010), it became apparent that the first-generation Affymetrix soybean array does not represent all genes in the soybean genome, and that unambiguous identification of duplicated homologous genes was not possible in every case using this array (Libault et al., 2010).

To overcome these issues and investigate the expression of all genes in the soybean genome, a custom Affymetrix soybean whole-genome transcript (sense

orientation) array (termed here Soybean WT array) was developed. The design of this array paid special attention to the ability to distinguish soybean paralogs such that the Soybean WT array allows quantification of the expression of every exon from all 66,000 predicted soybean transcripts (Schmutz et al., 2010). The specific design of the array also allows for the detection of (1) transcripts with undefined 3' ends, (2) non-polyadenylated mRNAs, (3) truncated transcripts, (4) alternative polyadenylation sites, and (5) alternative splicing events. A full description of the array design can be found in Supplemental Materials and Methods S1 and Supplemental Figure S5. To demonstrate the utility of this array, it was utilized to identify those soybean genes whose expression responded to MAMP treatment (see below and Supplemental Results S1).

#### Identification of MAMP-Responsive Soybean Genes

To identify MAMP-responsive genes the transcriptional profile of four soybean genotypes was analyzed after 30 min of MAMP treatment (this time point was chosen since it gave the strongest response to MAMP treatment in Arabidopsis; Navarro et al., 2004). We hybridized fragmented cDNA from MAMP- or mock-treated leaves to the Affymetrix Soybean WT array (24 GeneChip hybridization = four genotypes  $\times$  two treatments [MAMP or mock]  $\times$  three replicates). We used a linear mixed model (false discovery rate [FDR]  $< 0.05$ ) with an additional cutoff of a 2-fold ratio in pairwise comparisons (i.e. LD00-2817P<sub>Treatment</sub> versus LD00-2817P<sub>Control</sub>) to identify genes differentially expressed in each genotype. We observed considerable variation in the transcriptional profiles of the different genotypes (Fig. 2; Supplemental Tables S3–S5). On average, we detected 986 genes (both up-regulated and down-regulated) per genotype, with a minimum of 586 (LDX01-1-65) and a maximum of 1,733 (LD00-2817P; Fig. 2). Across all four genotypes, we detected a total of 4,249 differentially regulated genes. Only 437 common



**Figure 2.** Flower diagram showing numbers of overlapping and non-overlapping MAMP-responsive genes among LD00-2817P (LD), LDX01-1-65 (LX), Ripley, and EF59. Differentially expressed genes in each genotype were identified by linear mixed models at FDR < 0.05, with additional cutoff of 2-fold ratio in pairwise comparisons (MAMP-treated leaves versus mock-treated leaves). Over- and nonoverlapping genes were identified after a pairwise comparison between genotypes (i.e. LD versus LX).

genes were significantly, differentially regulated in all four genotypes, and 560, 29, 62, and 345 genes were specifically, differentially regulated in LD00-2817P, LDX01-1-65, Ripley, and EF59, respectively (Fig. 2; Supplemental Tables S6–S11). Likewise, a pairwise analysis revealed that LD00-2817P versus Ripley and LD00-2817P versus LDX01-1-65 showed 377 and 47 commonly up-regulated genes, respectively (Fig. 2; Supplemental Table S6). In contrast, only five and two common down-regulated genes were found among LD00-2817P versus Ripley and Ripley versus EF59 (Fig. 2).

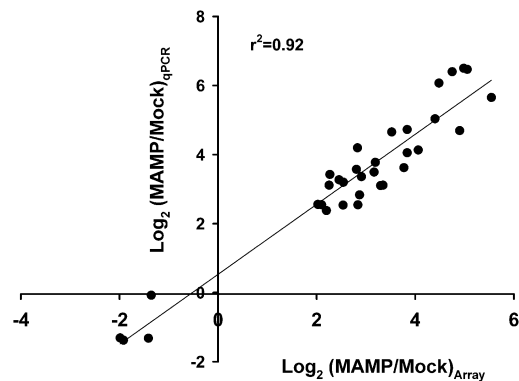
To confirm these DNA microarray results, the expressions of 40 randomly selected genes were analyzed via quantitative reverse transcription (qRT)-PCR (Fig. 3). The patterns of expression obtained using the Soybean WT GeneChip was confirmed for 35 genes (Fig. 3). This corresponds to about 90% validation of the Soybean WT GeneChip results by qRT-PCR.

Unique genes differentially expressed in each genotype were functionally classified using the MapMan gene functional classification system (Thimm et al., 2004; Usadel et al., 2009). Across all four genotypes, six functional categories were overrepresented: regulation, protein modification, regulation of transcription, hormones, enzyme families, and transport (Fig. 4). Likewise, most of the gene expression variation was detected in these functional categories (Fig. 4). Interestingly, we observed that most of the genes belonging to the regulation category encode different kinds of receptor-like kinases, including different Leu-rich repeat (LRR) and LysM receptors and calcium-dependent protein kinases (Supplemental Fig. S1). Likewise, the analysis revealed that LD00-2817P, Ripley, and EF59

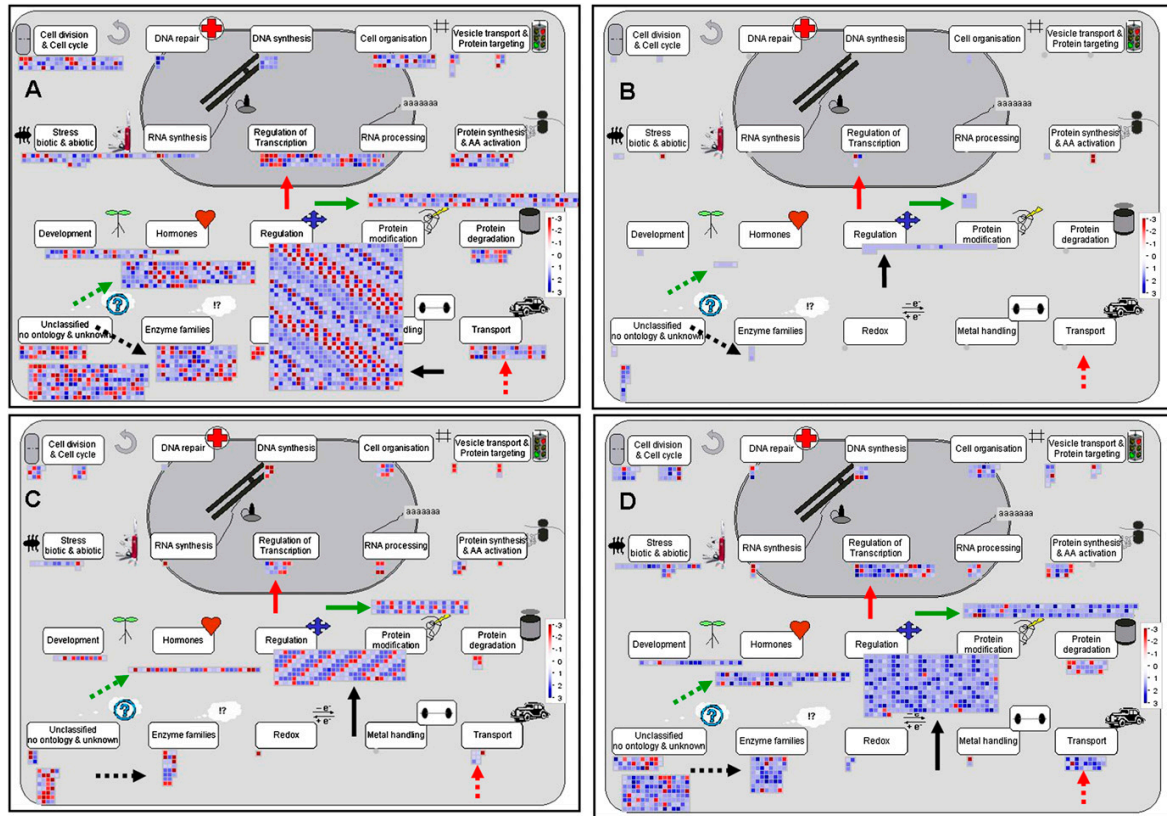
express several genes involved in auxin (indole-3-acetic acid), abscisic acid, brassinosteroid (BR), and ethylene synthesis, whereas LDX01-1-65 expresses genes involved only in indole-3-acetic acid and BR production (Fig. 4). Together, this transcriptional analysis is consistent with the differential response of these four soybean genotypes to MAMP treatment.

### MTI Can Induce Pathogen Resistance in Soybean

MAMP induction of gene expression may not always be translated into pathogen resistance. Therefore, it was important to show that the observed genotypic differences reflected measures of pathogen susceptibility. To address this issue, the mapping parents LD00-2817P and LDX01-1-65, crossed originally to develop a recombinant inbred population to map QTLs controlling soybean cyst nematode resistance (*Heterodera glycines* Ichinohe; Kim et al., 2011), which show a clear divergent MTI response (in this work), were selected for further analysis. Detached leaves from each of these two soybean-mapping parents were treated for 24 h with flg22, chitin, or a combination of both. After incubation, individual leaves were challenged separately with either *P. syringae* pv *glycinea* (*Psg*) or *S. sclerotiorum*. Three days after inoculation with *Psg*, no significant differences (FDR < 0.05) in colony-forming units (CFUs) were found between the two mock-treated parents, consistent with preliminary experiments in whole plants in the absence of any MAMP treatment (Fig. 5, A and C; Supplemental Fig. S2). However, significant differences (FDR < 0.05) were detected when mock and MAMP treatments were compared. For example, *Psg*



**Figure 3.** qRT-PCR validation of the MAMP-responsive genes in four contrasting soybean genotypes treated over 30 min with MAMPs (flg22 + chitin). A total of 40 randomly selected genes were used for qRT-PCR validation. Log<sub>2</sub> fold change values (MAMP/mock) from the qRT-PCR data were plotted against log<sub>2</sub> (MAMP/mock) hybridization intensity ratio values from the Soybean WT array (Array). *r*, Pearson correlation coefficient. Data are the average from three biological replicates with consistent results in each one. qPCR and Soybean WT values from the 40 selected genes are provided in the Supplemental Table S12.



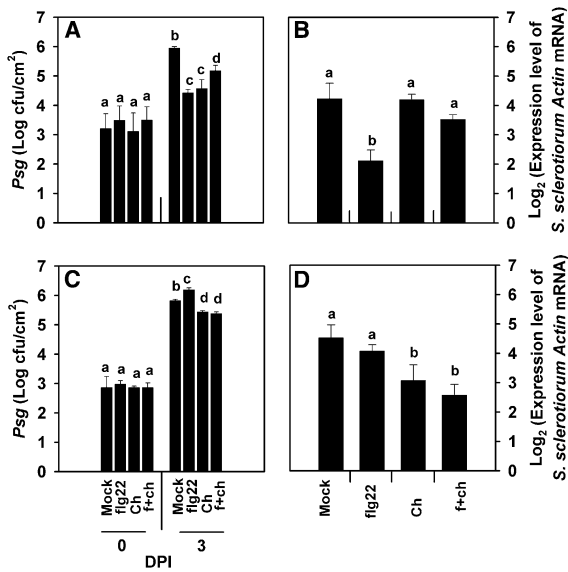
**Figure 4.** MapMan overview of regulation showing the regulation (black arrows), protein modification (green arrows), regulation of transcription (red arrows), hormones (green dotted arrows), enzyme families (black dotted arrows), and transport pathway (red dotted arrows) genes that are preferentially expressed in MAMP-treated leaves from LD00-2817P (A), LDX01-1-65 (B), Ripley (C), and EF59 (D) genotypes. Individual genes are represented by small squares. The  $\log_2$  (MAMP/mock) values for the differentially expressed genes (FDR < 0.05,  $\geq 2$ -fold difference) were false color coded by using a scale of  $-3$  to  $\pm 3$ . The intensity of blue and red colors indicates the degree of expression of the corresponding genes in MAMP-treated leaves. Color saturated at  $\pm 3$  (8-fold difference or higher).

log CFUs values were reduced 30% upon MAMP treatment of genotype LD00-2817P, while LDX01-1-65 plants showed a reduction of 7% (Fig. 5, A and C). Interestingly, we observed that flg22 treatment did not reduce colonization of *Psg* on LDX01-1-65 leaves (Fig. 5C). Somewhat similar results were found when plants were challenged with *S. sclerotiorum*. For example, flg22 pretreatment resulted in a significant decrease of *S. sclerotiorum* colonization in LD00-2817P, but not in LDX01-1-65 (Fig. 5, B and D). In contrast, chitin and the mixture of both MAMPs resulted in a decrease of *S. sclerotiorum* colonization in LDX01-1-65 but not LD00-2817P. Further statistical analysis on these data revealed that the difference between these two parents was statistically significant (FDR < 0.05), suggesting that interaction between genotype and treatment was relevant for soybean pathogen resistance triggered by MTI. Together, these results indicate that there was variable response by the two soybean genotypes to the

MAMP treatments, and in general, the parent showing the stronger MTI response (i.e. LD00-2817P) exhibited stronger *Psg* resistance after MAMP treatment.

#### Soybean MTI Segregates as a Multigenic, Quantitative Trait

To identify specific loci controlling the soybean MTI response, we analyzed a population of 97 F3 lines developed from a cross between LD00-2817P and LDX01-1-65, which gave divergent MTI response based on both ROS and gene expression analysis (see above). In the first experiment, we analyzed the individual MTI response across the 97 lines by quantifying the total ROS production. In a second experiment, we analyzed MAMP-triggered gene expression across the population for 25 selected genes that showed higher expression in MAMP-treated LD00-2817P when compared to MAMP-treated LDX01-1-65 (see “Materials



**Figure 5.** Bacterial population (A and C) and *S. sclerotiorum* actin mRNA transcripts levels (B and D) in LD00-2817P (A and B) and LDX01-1-65 (C and D) leaf tissues recovered 0 and 3, or 2 d after inoculation (DPI) with *Psg* ( $5 \times 10^5$  CFU/mL, 10 mM MgCl<sub>2</sub>) or *S. sclerotiorum*, respectively. Leaves were treated for 24 h with 1  $\mu$ M flg22, 50  $\mu$ g/mL chitin (Ch), or mixture of both flg22 and chitin (f + ch), or mock treated. Means  $\pm$  SE are from four biological replicates. Within each diagram, values sharing the common letters are not significantly different at FDR < 0.05 by ANOVA.

and Methods”). Levels of total ROS production among F3 lines were normally distributed ( $P < 0.05$ ); whereas, the expression of only some of the MAMP-responsive genes showed a normal distribution.

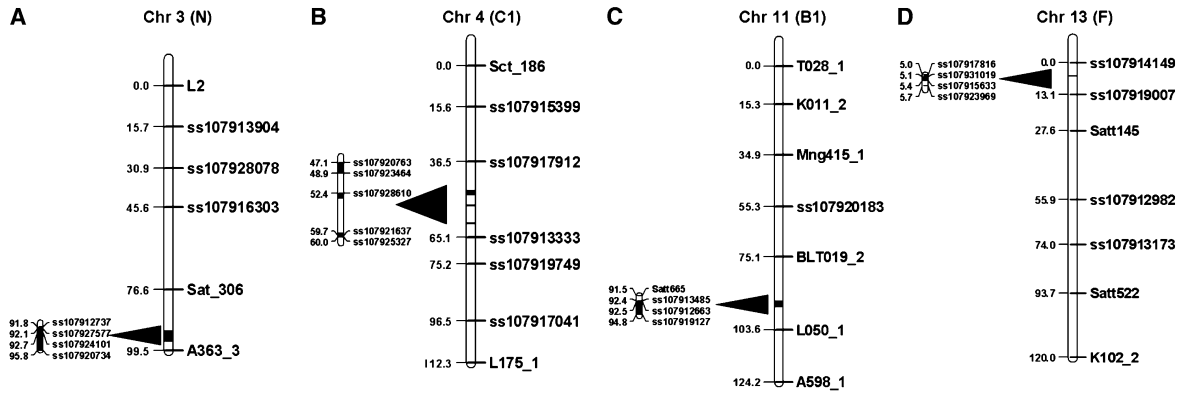
QTL mapping was done by first forming a linkage map with the 390 single-nucleotide polymorphism (SNP) markers that were polymorphic between the parents LD00-2817P and LDX01-1-65. Using this map, QTLs associated with MAMP-induced ROS production and MAMP-induced gene expression were positioned through interval mapping and a logarithm of odds (LOD) threshold corresponding to an experiment-wide threshold of 0.05 defined for each phenotype. One QTL for total ROS production was mapped and the LOD peak was at the SNP marker ss107920734 on chromosome 3 (Fig. 6A). This QTL was significant at a LOD of 2.83 and explained 12.6% of the phenotypic variation for ROS production. The allele from LD00-2817P conferred greater ROS production. Interestingly, genes encoding NADPH oxidase respiratory burst oxidase homolog (RBOH; main ROS producer), MAP kinase kinase, and enzymes involved in the biosynthesis of secondary metabolites (Supplemental Table S14) were localized within 1 cM of this marker.

eQTLs may act either in cis, when mapped within 5 cM of the gene used for the expression analysis, or in

trans, when mapped to a chromosome different from the gene used to measure expression (Swanson-Wagner et al., 2009). Also, it is possible that the expression of a gene may be controlled by multiple QTLs (Potokina et al., 2008; Swanson-Wagner et al., 2009). According to these criteria, two cis-eQTLs and a third trans-eQTL were mapped in the LD00-2817P by LDX01-1-65 population. From the expression of the gene Glyma01g43420 (localized on chromosome 1; WRKY 12 transcription factor), one trans-eQTL was mapped near marker ss107931019 on chromosome 13 and explained 12.8% of the phenotypic variation, and had a LOD of 2.87 (Fig. 6D). The allele from LD00-2817P was associated with low expression levels of this MTI-responsive gene. One of the cis-eQTLs, from the expression of the gene Glyma11g19650 (Argonaute5), mapped near ss107919127 on chromosome 11 (Fig. 6C) and explained 16% of the gene expression variation, and had a LOD of 4.12. Similar to the trans-eQTL, the allele from LD00-2817P was associated with lower gene expression. Interestingly, analysis of the expression of the gene Glyma04g28420 (S-receptor kinase-like protein) identified a major effect cis-eQTL region on chromosome 4 and close to marker ss107928610 (Fig. 6B) and explained 49.4% of the phenotypic variation, and had a LOD of 14.19 ( $P < 0.05$ ). The allele from LD00-2817P was associated with a higher level of gene expression, as compared to the response of the LDX01-1-65 allele. Most of the genes localized in the trans-eQTL and in the major effect cis-eQTL (LOD 14.19) encode for different receptor-like kinases (including LRR receptors), transcription factors (MYB, WRKY, and AP2-EREBP), protein modification enzymes, and transporters (zinc and calcium transporters). The cis-eQTL, in contrast, localized to the chromosome 11 encodes for either transporters (phosphorus and nitrate) or enzymes involved in secondary metabolism. However, all QTLs identified in this work map to regions where genes involved in the biosynthesis of hormones are located (Supplemental Table S14).

#### Broad Pathogen Resistance Triggered by MTI Is Inherited

Individual F3 lines were tested for their response to either *Psg* or *S. sclerotiorum* infection to determine whether the response of the two parental lines to pathogen attack is heritable. We selected two progeny lines that had a weak MTI response (based on the expression of the 25 MAMP-responsive genes used for the QTL analysis; Supplemental Fig. S3; Supplemental Table S16) and two other lines that had a strong MTI response. Representative data for one set of paired lines (one line with weak MTI and other with strong MTI) is shown in Figure 7, and information about the second pair of lines is shown in the Supplemental Figure S4. A significantly lower level of *Psg* colonization after MAMP treatment was found in the lines with the stronger MTI response (Fig. 7A). In contrast, in the lines selected based on their lower MAMP-triggered



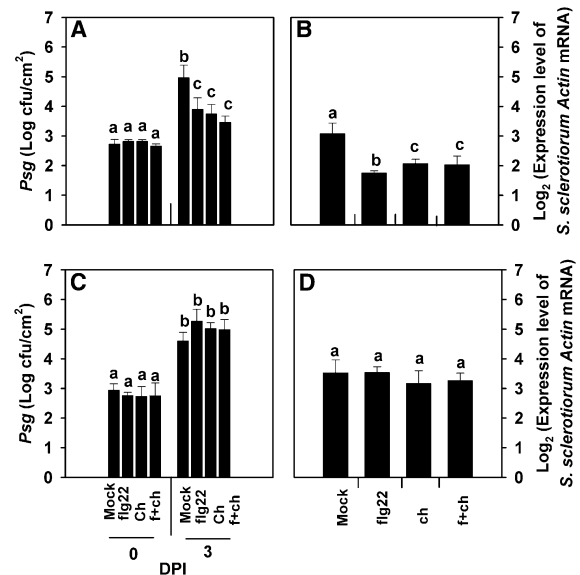
**Figure 6.** Localization of QTLs providing variation of MTI responses. QTLs were mapped in a population of 97 F3 lines developed from a cross between the contrasting genotypes LD00-2817P and LDX01-1-65. Major effect QTLs are shown in B and C, and the minor effect QTLs are shown in A and D. B and C are cis-QTLs, whereas D is the trans-eQTL. A is the QTL associated to the oxidative burst. Arrows indicate the localization of the markers associated to each QTL. The relative position of the mapped markers is given in cM; marker names and LOD values are given in Supplemental Table S13.

gene expression, no significant differences were found between the MAMP and mock treatments, indicating a weak MTI response (Fig. 7C). In the case of *S. sclerotiorum*-infected plants, the lines with a stronger MTI response had significantly less infection than the parental lines, about 50% less (Figs. 7B and 5, B and D). In contrast, no significant differences were observed after *S. sclerotiorum* inoculation in the line showing a weak MTI based on MAMP-triggered gene expression (Fig. 7D). These results suggest variation in MTI response is positively associated with the resistance level to these plant pathogens and is heritable.

## DISCUSSION

Recognition of various MAMPs by their corresponding PRR receptors elicits a variety of biochemical and transcriptional responses that lead to plant-pathogen resistance (Katagiri and Tsuda, 2010; Segonzac and Zipfel, 2011). For example, MAMP-triggered oxidative burst, which is a hallmark of the early events of MTI, has been implicated in lignin production and in signal transduction during programmed cell death (Asai and Yoshioka, 2008). MAMP-triggered oxidative burst appears to play an important role in plant disease resistance. For example, Arabidopsis, tobacco (*Nicotiana tabacum*), and rice plants with NADPH oxidase (the main ROS producer during the oxidative burst) either silenced or inhibited by diphenyleneiodonium exhibited a significant increase in susceptibility to different pathogens (Torres et al., 2002; Yoshioka et al., 2003; Chi et al., 2009). In contrast, an enhanced oxidative burst through the application of different MAMPs leads to increased resistance to pathogen infection (Mersmann et al., 2010). Here, we showed that MAMP treatment of selected soybean genotypes results in a significant

oxidative burst that varies by genotype. There is a rough correlation in the level of MAMP-triggered oxidative burst and the level of pathogen resistance after MAMP treatment. The results of this study provide support that the oxidative burst in soybean plays a role in resistance to two different pathogens.



**Figure 7.** Variation of the pathogen resistance levels in MAMP-treated contrasting F3 lines, defined based on higher (A and B) or lower (C and D) expression of MAMP-responsive genes. Bacterial population (*Psg*; A and C) and *S. sclerotiorum actin* mRNA (B and D) transcripts were determined as described. Within each diagram, values sharing the common letters are not significantly different at FDR < 0.05 by ANOVA.



To analyze genotypic variation in the MAMP response of different soybean genotypes (selected based on their MAMP-triggered oxidative burst and the availability of mapping populations), transcriptome analyses were performed on four soybean genotypes either mock treated or MAMP treated for 30 min, using a newly developed Affymetrix soybean whole-genome transcript array. The results showed significant variation in their transcriptional responses. MAMP-responsive genes map primarily into classes that consist of receptor kinases, protein modification enzymes, transcription factors, proteins involved in calcium regulation, and phytohormones. Similar results were observed in the transcriptional response of Arabidopsis and rice plants treated with different MAMPs (Fujiwara et al., 2004; Navarro et al., 2004; Zipfel et al., 2004; Wan et al., 2008). Consistent with the variation in MAMP-triggered oxidative burst among the four soybean genotypes, the transcriptional analysis also reflects genotypic variation. For example, the genotype LDX01-1-65, which was classified based on its oxidative burst as having a weaker MTI and also lower resistance to cyst nematode than LD00-2817P (Kim et al., 2011), expressed relatively few genes encoding for receptor-like kinases and transcription factors and, in general, showed a lower fold change ( $\leq 2$ ) in gene expression after MAMP addition. In contrast, those genotypes exhibiting a strong MAMP-triggered oxidative burst (e.g. LD00-2817P) also showed a strong transcriptional response to MAMP treatment. Analyses of Arabidopsis plants lacking genes belonging to these two functional categories, like PPRs (e.g. flg, chitin receptors), WRKY transcription factors, and different protein kinases, demonstrated that many are key players in the regulation of MTI and also important for plant disease resistance (Boudsocq et al., 2010; Segonzac and Zipfel, 2011).

It is well established that hormone levels also can strongly affect MTI. For example, variation in the levels of salicylic acid, jasmonic acid, ethylene, and BRs can influence plant disease resistance (Vert, 2008; Verhage et al., 2010). For example, ethylene and BR are important for FLS2 (flg receptor) accumulation and activation of BAK1 (a coactivator triggered by flg22 and EF-Tu), respectively. A defect in perception or accumulation of these hormones can negatively affect MTI and plant resistance (Boutrot et al., 2010; Oh et al., 2010; Jaillais et al., 2011). Consistent with these findings, MAMP treatment of soybean genotypes that exhibit a weak MTI resulted in little change in the expression of genes involved in hormone biosynthesis.

MTI represents a potential method to enhance soybean resistance to a broad array of pathogens and pests. For example, 24 h of treatment with flg22, chitin, or EF-Tu significantly reduced Arabidopsis colonization by *P. syringae* DC3000, *Alternaria brassicicola*, or *Piriformospora indica* (Zipfel et al., 2004; Wan et al., 2008; Jacobs et al., 2011). Several authors have pointed to the potential to harness MTI to protect crop plants from biotic stress (Ingvarsson and Street, 2011). By way of example, Lacombe and colleagues (2010) recently dem-

onstrated that transfer of the EFR1 PRR receptor of the MAMP Elf-18, from Arabidopsis to either tomato or tobacco plants resulted in a significant increase in resistance to bacterial pathogens. However, a more direct approach would be to harness the endogenous MTI pathways in crop plants to elevate resistance to biotic stress through selection and breeding for stronger MTI response. Our current study is a step toward this goal, because breeding for enhanced MTI requires that genotypic variation in the MAMP-triggered responses be heritable with controlling genes accurately mapped to enable marker-assisted selection.

In this study, MTI appeared to be a quantitative trait. Therefore, we sought to identify specific QTLs that regulate soybean MTI by using either total ROS production or gene expression as a phenotype. This approach identified four QTLs, one related to total ROS production and the three others related to gene expression of MAMP-responsive genes. However, none of these QTLs colocalized with previous QTLs associated to cyst nematode resistance identified with this same mapping population (Kim et al., 2011). Interestingly, most of the genes localized in the trans-eQTL, as well as in the major effect cis-eQTL (chromosomes 13 and 4), are predicted to encode different receptor-like kinases (including LRR receptors), transcription factors, or different transporters (like zinc transporters). Hence, they may regulate MTI either by direct effects on intracellular regulation or by modifying membrane-localized signaling events (e.g. ion flux). By contrast, the QTL associated with total ROS production contains genes that encode for the NADPH oxidase RBOH, as well as enzymes involved in the biosynthesis of secondary metabolites. Because RBOH can directly modulate ROS levels during the oxidative burst (Asai and Yoshioka, 2008), this is probably the gene underlying this QTL. The minor effect cis-eQTL contains genes related to transport and secondary metabolism only.

In conclusion, results of this study demonstrate that the MTI responses among the four soybean genotypes varied depending on MAMP treatment and pathogen. Results also indicated that MTI variation among soybean genotypes was quantitatively inherited. Likewise, we propose that it may be possible to directly select for lines with stronger MTI through the development of markers that identify key QTLs that control plant innate immunity.

## MATERIALS AND METHODS

### Plant Material and MAMP Treatment

This study was conducted by testing plants of four soybean (*Glycine max*) genotypes with contrasting genes for resistance to soybean cyst nematode (LD00-2817P [Diers et al., 2010] and LDX01-1-65 [Diers et al., 2005]) or sudden disease syndrome caused by *Fusarium virguliforme* (Ripley [de Farias Neto et al., 2007] and EF59 [Lightfoot et al., 2005]), as well as the 97 F3:4 (here after called as F3) lines from a cross between LD00-2817P (Diers et al., 2010) and LDX01-1-65. The lines from the LD00-2817P by LDX01-1-65 cross are the same lines that were used to test combinations of soybean cyst nematode (*Heterodera glycines* Ichinohe) resistance genes from PI 437654, PI 88788, and Peking that



were bred into LD00-2817 and from *Glycine soja* PI 468916 that were bred into LDX01-1-65 (Kim et al., 2011). Seeds were surface sterilized (30% bleach, rinsed several times with sterile, double-distilled water), planted in soil-less medium (SogeMix SM-2 general purpose growing medium), and germinated in growth chambers (190  $\mu\text{mol}$  photosynthetically active radiation  $\text{m}^{-2} \text{s}^{-1}$ , 23°C, 60% relative humidity, and 16-h photoperiod). Trifoliolate leaves from 3-week-old plants were detached and then vacuum infiltrated with double-distilled water for 2 min. Water-infiltrated trifoliolate leaves from five different plants of each genotype were pooled and cut into approximately 1-cm<sup>2</sup> slices. Equal amount of leaf slices (approximately 30 slices) were transferred into two different petri dishes and then floated overnight on autoclaved double-distilled water. Water was removed from both petri dishes and was replaced with 5 mL of MAMP solution (1  $\mu\text{M}$  flg22; Genescript) and 50  $\mu\text{g}$  of crab shell chitin (called in this study as chitin; Sigma-Aldrich), or 5 mL of mock solution (autoclaved double-distilled water plus equivalent amount of dimethyl sulfoxide [DMSO; Fisher Scientific]). DMSO was included since it was contained in the solution used to dissolve the flg22 peptide. After a 30-min treatment, mock- and MAMP-treated leaf slices were harvested into different tubes and immediately frozen in liquid nitrogen. Samples were stored at  $-80^\circ\text{C}$  until use. All procedures described above were performed under dark conditions.

### Oxidative Burst Measurement in Soybean Leaf Discs

ROS production by leaf tissue was assayed by hydrogen peroxide-dependent luminescence of luminol (Keppler et al., 1989). Under dark conditions, four 4-mm leaf discs per treatment from soybean water-infiltrated trifoliolate leaves were floated for 16 h in 200  $\mu\text{L}$  of autoclaved double-distilled water in a 48-well plate. Continuing under dark conditions, water was removed and replaced with 150  $\mu\text{L}$  of MAMP reaction buffer (20  $\mu\text{M}$  luminol [Sigma-Aldrich], 1  $\mu\text{g}$  horseradish peroxidase [Sigma-Aldrich], 1  $\mu\text{M}$  flg22 [Genescript], 50  $\mu\text{g}$  chitin [Sigma-Aldrich], or combination of both MAMPs) or mock reaction buffer (20  $\mu\text{M}$  luminol, 1  $\mu\text{g}$  horseradish peroxidase, and 2  $\mu\text{L}$  of DMSO [Fisher Scientific]). Luminescence was immediately recorded over 30 min in a CCD photon-counting camera (Photek 216). The luminescence intensity of individual wells was calculated using Photek IFS32 software (Photek 216). The luminescence intensity data from MAMP-treated leaf discs were normalized using intensity values from mock-treated plants. Parental lines had five biological replicates with three technical replicates each, whereas three biological and technical replicates were performed for the F3 lines.

### RNA Extraction, DNase Treatment, and qRT-PCR

Total RNA was purified from 0.5 g of MAMP- or mock-treated leaves by using Trizol reagent (Invitrogen), according to the manufacturer's instructions, and subsequently purified using chloroform extraction. Genomic DNA (gDNA) was removed from the purified RNA by using Turbo DNase (Ambion) following the manufacturer's instructions. Two micrograms of gDNA-free RNA were used to synthesize cDNA as described by Libault et al. (2010). The lack of gDNA contamination was verified by qPCR by using two different primers able to amplify gDNA.

qRT-PCR was performed as reported by Libault et al. (2008) using three housekeeping genes (*con6*, *con16*, and *con4*) to normalize the expression levels of the analyzed genes. Primer design was performed as described by Libault et al. (2010). Primer sequences are reported in Supplemental Tables S1 and S2. Gene expression levels were calculated according to the following equation:  $E = P_{\text{EFF}}^{(-\Delta\text{CT})}$ , where  $P_{\text{EFF}}$  is the primer efficiency calculated using LinRegPCR (Ramakers et al., 2003), and  $\Delta\text{CT}$  is  $\text{CT}_{\text{Housekeeping}}$  minus  $\text{CT}_{\text{Gene}}$  where  $\text{CT}_{\text{Housekeeping}}$  is the geometric median from the cycle threshold (CT) of three different housekeeping genes and  $\text{CT}_{\text{Gene}}$  is the CT from each gene. The gene expression levels were used to calculate the ratio between MAMP- and mock-treated leaves, and then the ratios were Log<sub>2</sub> transformed to obtain the fold change. All the experiments were performed in 384-well plates. Parental lines had three biological replicates, whereas two biological replicates were performed for the F3:4 lines.

### Affymetrix Soybean Whole Sense Transcript (Wild-Type) Exon-Array Design

A custom whole-genome Affymetrix Soybean Whole Sense orientation transcript exon array (Soybean WT array) was designed from the published

soybean genome sequence (see Supplemental Materials and Methods S1). This array includes all the soybean genes, including both high- and low-confidence gene models, reported in Schmutz et al. (2010). To design probes for this array, sequences from the first draft assembly of the soybean genome (Schmutz et al., 2010) were used and downloaded from the Department of Energy-Joint Genome Institute Web site (phytozome: <http://phytozome.net>). The probe selection algorithm was developed by Affymetrix, and probes were selected and designed to span each exon of the predicted gene models, whenever possible. Each probe set has 11 different 25mers probes printed onto the Soybean WT array. In total, the Soybean WT array contains over 1,000,000 probe sets, some of them designed for background normalization purposes. An association probe set file is provided in Supplemental Table S15.

### Affymetrix Soybean Whole Sense Transcript (Wild-Type) Exon-Array Hybridization and Analysis

Total RNA from both MAMP- and mock-treated leaves was used for Affymetrix Soybean WT array analysis following the manufacturer's protocols (Affymetrix). Briefly, 500 ng of gDNA-free total RNA were used to produce fragmented and biotin-labeled cDNA using the Ambion wild-type expression kit (Applied Biosystems) and the Affymetrix wild-type terminal labeling and hybridization kit (Affymetrix). Soybean WT array hybridizations were performed at the University of Missouri (<http://biotech.rnet.missouri.edu/dnacore/>), following standard Affymetrix procedures. The arrays were scanned with a GeneChip 7G plus high-resolution scanner. Both the gene level and the exon level expression values were summarized by Robust Multi-Array Average algorithm by using Expressionist Refiner 6.1 (GeneData). Briefly, the signal intensity cell files from scanner were analyzed by a workflow with GC background subtraction, quantile normalization, and median polish. The gene expression values from different samples were further used for gene differential statistical analysis. The exon level expression values were used in the microarray detection of alternative splicing (MiDAS)  $P$  value calculation. This MiDAS  $P$  value determines the gene splicing. To ensure the reliable exons in above MiDAS test, the detection above background  $P$  values were calculated for each probe set by comparing perfect match probes to members of the background probes with the same GC content. A cutoff of detection above background  $P$  value of 0.05 was used for filtering and determining the exon present/absent. The full microarray datasets generated from the 24 Genechip used in this study were deposited in the Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>) under the accession number GSE32642.

### Identification of Differentially Expressed Genes

A linear mixed-effects model was applied to the normalized log-scale (base 2) expression measures separately for each gene using the bioconductor software R/mmanova (1.16.0 version). Each linear mixed model includes the genotype effect, treatment effect, the interaction between the genotype and treatment effects, and the biological replicate effect, where the biological replicate effect is considered to be random effect. The variance-shrinkage  $F_s$  tests (Cui et al., 2005) were conducted using R/mmanova to identify differentially expressed genes between the treatment and control groups for each of the four genotypes.  $P$  values for the variance-shrinkage  $F_s$  tests were obtained through 100 permutations via the Matest statement in the R/mmanova software, and a  $q$  value (Storey, 2002) was then computed for each  $P$  value to produce lists of differentially expressed genes with estimated FDRs of 5%. Among these significantly differentially expressed genes, genes with fold change above 2 were considered for functional classification through PFAM, KOG, and PANTHER domain predictions. Additionally, MapMan gene functional classification system was used (Thimm et al., 2004; Usadel et al., 2009). For MapMan analysis, an updated soybean mapping file (Yang et al., 2010), which allows to survey all the functional categories included in the software MapMan, was used.

### Selection of Candidate Genes for QTL Analysis

A linear mixed model, as described above, was applied to the normalized GeneChip data from LD00-2817P and LDX01-1-65. The permutation  $F_s$  test (as described in the above section) was applied to conduct three comparisons: the comparison between the MAMP treatment and mock treatment for the two genotypes, and the comparison of fold changes between the LD00-2817P and

LDX01-1-65. The intersection union principle (Berger, 1982) was utilized to identify genes that are statistically significantly up-regulated ( $P$  value  $\leq 0.05$ ) with fold change above 1.5 in both genotypes (LD00-2817P and LDX01-1-65) but has statistically significant larger fold change in the line with strong MTI (LD00-2817P) than in the line with weak MTI (LDX01-1-65;  $P$  value  $< 0.05$ ) with at least 2 times higher fold change in LD00-2817P. A total of 45 genes were identified with this criterion. Based on this list and their biological relevance in MTI, 25 genes from this analysis were selected to analyze their expression by qRT-PCR, as described above. Ratios between MAMP/mock were calculated to determine whether the MAMP treatment induced the expected response in both parental lines. After the logarithm transformation (base 2), the qRT-PCR data from the 99 treated plants (two parents and 97 F3 lines) were tested for normality using both the Lilliefors test and the Jarque-Bera test in Matlab software and normality plots were used as visual aid to identify genes that are normally distributed across the 99 plants. Genes that were declared as nonsignificant for both tests, which indicates no violation against the normality assumption, and have a close-to-straight-line normality plot were selected for QTL studies.

### QTL Mapping Analysis

The population of F3 lines was evaluated with the Illumina GoldenGate 1,536 Universal Soy Linkage Panel 1.0 (USLP 1.0; Hyten et al., 2010) as described by Kim et al. (2011). The population segregated for 390 SNP markers from the USLP 1.0 and a linkage map was formed using these markers with Joinmap 3.0 (Van Ooijen and Voorrips, 2001). QTLs controlling both ROS and the expression level of each gene assayed in the population through qRT-PCR were mapped through interval mapping with the program MapQTL 4.0 (Van Ooijen and Maliepaard, 1996; Van Ooijen, 2000). Genome-wide significance thresholds were determined by permutation tests in MapQTL 4.0.

### Fungal Infection Assay

Cultures of *Sclerotinia sclerotiorum* were started 48-h prior to inoculation by subculturing actively growing edges of fungal colonies from stock cultures onto potato dextrose agar (DIFCO). For inoculation, 4-mm-diameter agar plugs with growing mycelium were cut from the edges of colonies using a cork borer.

Forty-eight hours before inoculation, trifoliolate leaves from 3-week-old soybean plants were detached and floated for 16 h in 20 mL of water in a petri dish. Solution was removed and the trifoliolate leaves were treated with 0.01% Silwet L-77 for 20 min and immediately rinsed several times with double-distilled water. The trifoliolate leaves were then treated for 24 h with 1  $\mu$ M flg22, 50  $\mu$ g chitin, 1  $\mu$ M flg22 plus 50  $\mu$ g chitin, or DMSO (as mock). Trifoliolate leaves were transferred into a petri dish (one per trifoliolate) that contained moistened Whatman paper. One agar plug per leaflet of each trifoliolate (three per trifoliolate in total) was placed in each petri dish. Petri dishes were sealed with Parafilm and then placed in a growth chamber with the same environmental conditions described above. Agar plugs were removed from the leaves 24 h after inoculation. Two days after inoculation, individual leaves from each trifoliolate leaf were harvested and immediately frozen in liquid nitrogen and stored until used. *S. sclerotiorum* infection levels were determined by the abundance of *S. sclerotiorum* actin (Kasza et al., 2004) mRNA via qRT-PCR and following the conditions described in the section "RNA Extraction, DNase Treatment, and qRT-PCR." The experiment was repeated four times, each at different dates and with new inoculum, to obtain four biological replicates. Expression level of the *S. sclerotiorum* marker gene was logarithm base 2 transformed and then was applied to an ANOVA model with the treatment effect. The genotype effect and their interaction were estimated using the SAS software PROC GLM procedure. Forty pairwise comparisons were conducted using  $t$  tests to confirm the statistical significance of the differences among the four treatments for each genotype and the difference among the two parental lines as well as the four selected F3:4 lines for each treatment. The 40  $P$  values were transformed to  $q$  value (Storey, 2002) to control the FDR level at 0.05.

### Bacterial Infection Assay

Stocks of *Pseudomonas syringae* pv *glycinea* (*Psg*) were maintained on NYG agar medium (5 g/L bacto peptone, 3 g/L yeast extract, 20% [w/v] glycerol, 1.5% bacto agar) plates supplemented with rifampicin (50  $\mu$ g/mL) and as 30% glycerol stock at  $-80^{\circ}\text{C}$ . Forty-eight hours before infection, *Psg* was grown

into NYG agar rifampicin (50  $\mu$ g/mL) plates at  $30^{\circ}\text{C}$ . Just prior to the plant infection, a small amount of bacteria were scrapped from the plate and dissolved in 1 mL of 10 mM  $\text{MgCl}_2$ , and the optical density at 600 nm was determined. From this bacterial solution, serial dilutions in 10 mM  $\text{MgCl}_2$  were made to reach a bacterial concentration of approximately  $10^5$  CFU/mL. The CFU levels were determined by plating of serial dilutions onto NYG agar plates.

Forty-eight hours before inoculation, fifteen 4-mm diameter leaf discs from trifoliolate leaves of 3-week-old soybean plants were floated for 16 h in 2 mL of autoclaved double-distilled water in a 12-well plate. The solution was removed, and the leaf discs were treated for 24 h with 1  $\mu$ M flg22, 50  $\mu$ g chitin mixture, 1  $\mu$ M flg22 plus 50  $\mu$ g chitin mixture, or DMSO (as mock). All these treatments were supplemented with 0.01% Silwet L-77. Solutions were removed, and the leaf discs were treated for another 3 h with a *Psg* cell suspension at a concentration of  $10^5$  CFU/mL. The bacterial suspension was removed, and the leaf discs were rinsed several times with autoclaved double-distilled water and then floated for 3 d on water in the same 12-well plate. The quantification of bacterial colonization at 0 and 3 d post inoculation was determined by pooling and grinding four leaf discs in 10 mM  $\text{MgCl}_2$  and then performing different serial dilutions with 10 mM  $\text{MgCl}_2$  until  $10^{-4}$  and  $10^{-6}$  cells  $\text{mL}^{-1}$  were reached. These dilutions were sprayed into NYG agar rifampicin (50  $\mu$ g/mL) plates, and then incubated at  $30^{\circ}\text{C}$  for 2 d. The number of colonies was expressed as CFU/cm<sup>2</sup>. This experiment was repeated four times with the same conclusion. CFU data were logarithm base 10 transformed and then applied to an ANOVA model with the treatment effect. The genotype effect, the day effect, and their interactions were estimated using SAS proc glm procedure. Two hundred and fifty two pairwise comparisons were conducted using  $t$  tests to check the statistical significance of the differences among the four treatments for each genotype on each day, the differences among two parental lines, as well as the four selected F3 lines for each treatment on each day and the differences among the days for each combination of treatments and genotypes. The 252  $P$  values were transformed to  $q$  value (Storey, 2002) to control the FDR level at 0.05.

### Supplemental Data

The following materials are available in the online version of this article.

**Supplemental Figure S1.** MapMan overview of receptors-like kinase in MAMP-treated leaves from LD00-2817P, LDX01-1-65, Ripley, and EF59 genotypes.

**Supplemental Figure S2.** *Sclerotinia* infection in whole plants.

**Supplemental Figure S3.** Expression level of four genes from the 25 genes used for QTL analysis.

**Supplemental Figure S4.** Variation of the pathogen resistance levels in MAMP-treated contrasting F3 lines.

**Supplemental Figure S5.** Illustration of the association between probe id and the predicted gene model.

**Supplemental Table S1.** Primers used to validate the Soybean WT array data by qRT-PCR.

**Supplemental Table S2.** Primer used for eQTL analysis.

**Supplemental Table S3.** Genes differentially regulated in LD00-2817P after 30 min of MAMP treatment (1  $\mu$ M flg22 + 50  $\mu$ g/mL chitin mixture).

**Supplemental Table S4.** Genes differentially regulated in LDX01-1-65 after 30 min of MAMP treatment (1  $\mu$ M flg22 + 50  $\mu$ g/mL chitin mixture).

**Supplemental Table S5.** Genes differentially regulated in Ripley after 30 min of MAMP treatment (1  $\mu$ M flg22 + 50  $\mu$ g/mL chitin mixture).

**Supplemental Table S6.** Genes differentially regulated in EF59 after 30 min of MAMP treatment (1  $\mu$ M flg22 + 50  $\mu$ g/mL chitin mixture).

**Supplemental Table S7.** Common genes significantly up-regulated in LD00-2817P, LDX01-1-65, Ripley, and EF59 after 30 min of MAMP treatment (1  $\mu$ M flg22 + 50  $\mu$ g/mL chitin mixture).

**Supplemental Table S8.** Unique genes up-regulated in LD00-2817P after 30 min of MAMP treatment (1  $\mu$ M flg22 + 50  $\mu$ g/mL chitin mixture).

**Supplemental Table S9.** Unique genes up-regulated in LDX01-1-65 after 30 min of MAMP treatment (1  $\mu$ M flg22 + 50  $\mu$ g/mL chitin mixture).

- Supplemental Table S10.** Unique genes up-regulated in Ripley after 30 min of MAMP treatment (1  $\mu$ M flg22 + 50  $\mu$ g/mL chitin mixture).
- Supplemental Table S11.** Unique genes up-regulated in EF59 after 30 min of MAMP treatment (1  $\mu$ M flg22 + 50  $\mu$ g/mL chitin mixture).
- Supplemental Table S12.** Verification of the Soybean WT array results by qRT-PCR of randomly selected genes.
- Supplemental Table S13.** Marker name, chromosome, and LOD values associated from the QTLs associated to soybean MTL.
- Supplemental Table S14.** Genes localized at 1 cM in each QTL.
- Supplemental Table S15.** Association probe set file can be downloaded from <http://seedgenenetwork.net/annotate#soybeanWT>.
- Supplemental Table S16.** Expression level of 25 different genes tested in the F3:4 mapping population.
- Supplemental Results S1.** Alternative splicing events in LD00-2817P, LD00-1-65, Ripley, and EF59 genotypes.
- Supplemental Materials and Methods S1.** Design of the Soybean WT array.
- Wang D** (2005) Registration of 'LDX01-1-65' soybean. *Crop Sci* **45**: 1671–1672
- Diers BW, Cary T, Thomas D, Colgrove A, Niblack T** (2010) Registration of LD00-2817P soybean germplasm line with resistance to soybean cyst nematode from PI 437654. *J Plant Reg* **4**: 141–144
- Forsyth A, Mansfield JW, Grabov N, de Torres M, Sinapidou E, Grant MR** (2010) Genetic dissection of basal resistance to *Pseudomonas syringae* pv. *phaseolicola* in accessions of *Arabidopsis*. *Mol Plant Microbiol Interact* **12**: 1545–1552
- Fujiwara S, Tanaka N, Kaneda T, Takayama S, Isogai A, Che FS** (2004) Rice cDNA microarray-based gene expression profiling of the response to flagellin perception in cultured rice cells. *Mol Plant Microbe Interact* **9**: 986–998
- Göhre V, Robatzek S** (2008) Breaking the barriers: microbial effector molecules subvert plant immunity. *Annu Rev Phytopathol* **46**: 189–215
- Gómez-Gómez L, Felix G, Boller T** (1999) A single locus determines sensitivity to bacterial flagellin in *Arabidopsis thaliana*. *Plant J* **18**: 277–284
- Hartman GL, West E, Herman T** (2011) Crops that feed the world 2: soybean-worldwide production, use, and constraints caused by pathogens and pests. *Food Security* **3**: 5–17
- Hyten DL, Choi IY, Song Q, Specht JE, Carter TE, Shoemaker RC, Hwang EY, Matukumalli LK, Cregan PB** (2010) A high density integrated genetic linkage map of soybean and the development of a 1,536 Universal Soy Linkage Panel for QTL mapping. *Crop Sci* **50**: 960–968
- Ingvarsson PK, Street NR** (2011) Association genetics of complex traits in plants. *New Phytol* **189**: 902–922
- Jacobs S, Zechmann B, Molitor A, Trujillo M, Petutschnig E, Likpa V, Kogel KH, Schäfer P** (2011) Broad-spectrum suppression of innate immunity is required for colonization of *Arabidopsis* roots by the fungus *Piriformospora indica*. *Plant Physiol* **156**: 726–740
- Jailais Y, Belkhadir Y, Balsemão-Pires E, Dangl JL, Chory J** (2011) Extracellular leucine-rich repeats as a platform for receptor/coreceptor complex formation. *Proc Natl Acad Sci USA* **108**: 8503–8507
- Jones JD, Dangl JL** (2006) The plant immune system. *Nature* **444**: 323–329
- Kasza Z, Vagvölgyci C, Fèvre M, Cotton P** (2004) Molecular characterization and in planta detection of *Sclerotinia sclerotiorum* endopolygalacturonase genes. *Curr Microbiol* **48**: 208–213
- Katagiri F, Tsuda K** (2010) Understanding the plant immune system. *Mol Plant Microbe Interact* **23**: 1531–1536
- Kepler LD, Baker CJ, Atkinson MM** (1989) Active oxygen production during a bacteria-induced hypersensitive reaction in tobacco suspension cells. *Phytopathology* **79**: 974–978
- Kim M, Hyten DL, Niblack TL, Diers BW** (2011) Stacking resistance alleles from wild and domestic soybean sources improves soybean cyst nematode resistance. *Crop Sci* **51**: 934–934
- Lacombe S, Rougon-Cardoso A, Sherwood E, Peeters N, Dahlbeck D, van Esse HP, Smoker M, Rallapalli G, Thomma BPHJ, Staskawicz B, et al** (2010) Interfamily transfer of a plant pattern-recognition receptor confers broad-spectrum bacterial resistance. *Nat Biotechnol* **28**: 365–369
- Libault M, Farmer A, Brechenmacher L, Drnevich J, Langley RJ, Bilgin DD, Radwan O, Neece DJ, Clough SJ, May GD, et al** (2010) Complete transcriptome of the soybean root hair cell, a single-cell model, and its alteration in response to *Bradyrhizobium japonicum* infection. *Plant Physiol* **152**: 541–552
- Libault M, Thibivilliers S, Bilgin DD, Radwan O, Benitez M, Clough SJ, Stacey G** (2008) Identification of four soybean reference genes for gene expression normalization. *Plant Genome* **1**: 44–54
- Lightfoot DA, Njiti VN, Gibson PT, Kassem MA, Iqbal JM, Meksem K** (2005) Registration of Essex x Forrest recombinant inbred line (RIL) mapping population. *Crop Sci* **45**: 1678–1681
- Luna E, Pastor V, Robert J, Flors V, Mauch-Mani B, Ton J** (2011) Callose deposition: a multifaceted plant defense response. *Mol Plant Microbe Interact* **24**: 183–193
- Lygin AV, Hill CB, Zernova OV, Crull L, Widholm JM, Hartman GL, Lozovaya VV** (2010) Response of soybean pathogens to glyceollin. *Phytopathology* **100**: 897–903
- Mersmann S, Bourdais G, Rietz S, Robatzek S** (2010) Ethylene signaling regulates accumulation of the FLS2 receptor and is required for the oxidative burst contributing to plant immunity. *Plant Physiol* **154**: 391–400
- Navarro L, Zipfel C, Rowland O, Keller I, Robatzek S, Boller T, Jones JDG** (2004) The transcriptional innate immune response to flg22: interplay

## ACKNOWLEDGMENTS

The authors wish to thank the following people: (1) Dr. Stella Kantartzis, from Southern Illinois University Carbondale, for providing seeds of the soybean genotype EF59; (2) Dr. Peter Tipton and Melody Kroll, from the University of Missouri, for editing corrections; (3) Dr. Mingyi Zhou, from the DNA Core facility at the University of Missouri, for Genechip processing; (4) Dr. Maritza Zavaleta-Pastor and Erika Matanolis, from the University of Missouri, for technical support on the pathogenesis assays; (5) Dr. Gene Tanimoto, Christopher Davies, Stan Trask, Brant Wong, and others at Affymetrix for designing and making the custom Soybean WT array; and (6) Dr. Kiwamu Tanaka, from the University of Missouri, for discussions.

Received July 12, 2011; accepted September 29, 2011; published September 30, 2011.

## LITERATURE CITED

- Ahmad S, Van Hulten M, Martin J, Pieterse CM, Van Wees SCM, Ton J** (2011) Genetic dissection of basal defence responsiveness in accessions of *Arabidopsis thaliana*. *Plant Cell Environ* **34**: 1191–1206
- Alvarez ME, Nota F, Cambiagno DA** (2010) Epigenetic control of plant immunity. *Mol Plant Pathol* **11**: 563–576
- Asai S, Yoshioka H** (2008) The role of radical burst via MAPK signaling in plant immunity. *Plant Signal Behav* **3**: 920–922
- Berger RL** (1982) Multiparameter hypothesis testing and acceptance sampling. *Technometrics* **24**: 295–300
- Boller T, Felix G** (2009) A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annu Rev Plant Biol* **60**: 379–406
- Boudsocq M, Willmann MR, McCormack M, Lee H, Shan L, He P, Bush J, Cheng SH, Sheen J** (2010) Differential innate immune signalling via Ca<sup>2+</sup> sensor protein kinases. *Nature* **464**: 418–422
- Boutrot F, Segonzac C, Chang KN, Qiao H, Ecker JR, Zipfel C, Rathjen JP** (2010) Direct transcriptional control of the *Arabidopsis* immune receptor FLS2 by the ethylene-dependent transcription factors EIN3 and EIL1. *Proc Natl Acad Sci USA* **107**: 14502–14507
- Chi MH, Park SY, Kim S, Lee YH** (2009) A novel pathogenicity gene is required in the rice blast fungus to suppress the basal defenses of the host. *PLoS Pathog* **5**: e1000401
- Cui X, Hwang JT, Qiu J, Blades NJ, Churchill GA** (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**: 59–75
- de Farias Neto AL, Hashmi AFR, Schmidt M, Carlson SR, Hartman GL, Li S, Nelson RL, Diers BW** (2007) Mapping and confirmation of a new sudden death syndrome resistance QTL on linkage group D2 from the soybean genotypes PI 567374 and 'Ripley'. *Mol Breed* **20**: 53–62
- Diers BW, Arelli PR, Carlson SR, Fehr WR, Kabelka EA, Shoemaker RC,**

- and overlap with Avr gene-dependent defense responses and bacterial pathogenesis. *Plant Physiol* **135**: 1113–1128
- Oh MH, Wang X, Wu X, Zhao Y, Clouse SD, Huber SC** (2010) Autophosphorylation of Tyr-610 in the receptor kinase BAK1 plays a role in brassinosteroid signaling and basal defense gene expression. *Proc Natl Acad Sci USA* **107**: 17827–17832
- Poland JA, Balint-Kurti PJ, Wisser RJ, Pratt RC, Nelson RJ** (2008) Shades of gray: the world of quantitative disease resistance. *Trends Plant Sci* **14**: 21–29
- Potokina E, Druka A, Luo Z, Wise R, Waugh R, Kearsey M** (2008) Gene expression quantitative trait locus analysis of 16,000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. *Plant J* **53**: 90–101
- Ramakers C, Ruijter JM, Deprez RH, Moorman AF** (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci Lett* **339**: 62–66
- Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al** (2010) Genome sequence of the palaeopolyploid soybean. *Nature* **463**: 178–183
- Segonzac C, Zipfel C** (2011) Activation of plant pattern-recognition receptors by bacteria. *Curr Opin Microbiol* **14**: 54–61
- Storey JD** (2002) A direct approach to false discovery rates. *J R Statist Soc B* **64**: 479–498
- Swanson-Wagner RA, DeCook R, Jia Y, Bancroft T, Ji T, Zhao X, Nettleton D, Schnable PS** (2009) Paternal dominance of trans-eQTL influences gene expression patterns in maize hybrids. *Science* **326**: 1118–1120
- Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller 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
- Thomma BPHJ, Nürnberger T, Joosten MHAJ** (2011) Of PAMPs and effectors: the blurred PTI-ETI dichotomy. *Plant Cell* **23**: 4–15
- Torres MA, Dangl JL, Jones JDG** (2002) Arabidopsis gp91phox homologues AtrbohD and AtrbohF are required for accumulation of reactive oxygen intermediates in the plant defense response. *Proc Natl Acad Sci USA* **99**: 517–522
- USDA** (2010) National Agricultural Statistics Service. [http://www.nass.usda.gov/Data\\_and\\_Statistics/index.asp](http://www.nass.usda.gov/Data_and_Statistics/index.asp) (June 25, 2011)
- Usadel B, Poree F, Nagel A, Lohse M, Czedik-Eysenberg A, Stitt M** (2009) A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, maize. *Plant Cell Environ* **32**: 1211–1229
- Van Ooijen JW** (2000) MapQTL Version 4.0: User Friendly Power in QTL Mapping. Addendum to the Manual of Version 3.0. Plant Research International, Wageningen, The Netherlands
- Van Ooijen JW, Maliepaard C** (1996) MapQTL Version 3.0: Software for the Calculation of QTL Positions on Genetic Maps. CPRO-DLO, Wageningen, The Netherlands
- Van Ooijen JW, Voorrips RE** (2001) Join Map 3.0, Software for the Calculation of Genetic Maps. Plant Research International, Wageningen, The Netherlands
- Verhage A, van Wees SCM, Pieterse CMJ** (2010) Plant immunity: it's the hormones talking, but what do they say? *Plant Physiol* **154**: 536–540
- Vert G** (2008) Plant signaling: brassinosteroids, immunity and effectors are back! *Curr Biol* **18**: 963–965
- Vuong TD, Diers BW, Hartman GL** (2008) Identification of QTL for resistance to sclerotinia stem rot in soybean plant introduction 194639. *Crop Sci* **48**: 2209–2214
- Wan J, Zhang XC, Neece D, Ramonell KM, Clough S, Kim SY, Stacey MG, Stacey G** (2008) A LysM receptor-like kinase plays a critical role in chitin signaling and fungal resistance in *Arabidopsis*. *Plant Cell* **20**: 471–481
- Yang SS, Valdés-López O, Xu WW, Bucciarelli B, Gronwald JW, Hernández G, Vance CP** (2010) Transcript profiling of common bean (*Phaseolus vulgaris* L.) using the GeneChip Soybean Genome Array: optimizing analysis by masking biased probes. *BMC Plant Biol* **10**: 85
- Yoshioka H, Numata N, Nakajima K, Katou S, Kawakita K, Rowland O, Jones JDG, Doke N** (2003) *Nicotiana benthamiana* gp91phox homologs NbrbohA and NbrbohB participate in H<sub>2</sub>O<sub>2</sub> accumulation and resistance to *Phytophthora infestans*. *Plant Cell* **15**: 706–718
- Zhang W, Gao S, Zhou X, Chellappan P, Chen Z, Zhou X, Zhang X, Fromuth N, Coutino G, Coffey M, et al** (2011) Bacteria-responsive microRNAs regulate plant innate immunity by modulating plant hormone networks. *Plant Mol Biol* **75**: 93–105
- Zipfel C** (2008) Pattern-recognition receptors in plant innate immunity. *Curr Opin Immunol* **20**: 10–16
- Zipfel C** (2009) Early molecular events in PAMP-triggered immunity. *Curr Opin Plant Biol* **12**: 414–420
- Zipfel C, Robatzek S** (2010) Pathogen-associated molecular pattern-triggered immunity: veni, vidi...? *Plant Physiol* **154**: 551–554
- Zipfel C, Robatzek S, Navarro L, Oakeley EJ, Jones JDG, Felix G, Boller T** (2004) Bacterial disease resistance in *Arabidopsis* through flagellin perception. *Nature* **428**: 764–767

## APPENDIX C

Ohr, H., Bui, A.Q., Le, B.H., Fischer, R.L., and Choi, Y. (2007). **Identification of putative Arabidopsis DEMETER target genes by GeneChip analysis.** *Biochem Biophys Res Commun* 364, 856–860.

# Identification of putative *Arabidopsis* DEMETER target genes by GeneChip analysis

Hyonhwa Ohr <sup>a</sup>, Anhthu Q. Bui <sup>b</sup>, Brandon H. Le <sup>b</sup>, Robert L. Fischer <sup>c</sup>, Yeonhee Choi <sup>a,\*</sup>

<sup>a</sup> Department of Biological Sciences, Seoul National University, Seoul 151-742, Republic of Korea

<sup>b</sup> Department of Molecular, Cell, and Developmental Biology, University of California—Los Angeles, Los Angeles, CA 90095, USA

<sup>c</sup> Department of Plant and Microbial Biology, University of California—Berkeley, Berkeley, CA 94720, USA

Received 9 October 2007

Available online 25 October 2007

---

## Abstract

The *Arabidopsis* DEMETER (DME) DNA glycosylase is required for the maternal allele expression of imprinted Polycomb group (*MEDEA* and *FIS2*) and transcription factor (*FWA*) genes in the endosperm. Expression of DME in the central cell, not in pollen or stamen, establishes gene imprinting by hypomethylating maternal alleles. However, little is known about other genes regulated by DME. To identify putative DME target genes, we generated *CaMV:DME* plants which ectopically express *DME* in pollen and stamens. Comparison of mRNA profiles revealed 94 genes induced by ectopic DME expression in both stamen and pollen. Gene ontology analysis identified three molecular functions enriched in the DME-inducible RNA list: DNA or RNA binding, kinase activity, and transcription factor activity. Semi-quantitative RT-PCR verified the candidate genes identified by GeneChip analysis. The putative target genes identified in this study will provide insights into the regulatory mechanism of DME DNA glycosylase and the functions of DNA demethylation.

© 2007 Elsevier Inc. All rights reserved.

**Keywords:** Gene imprinting; DNA demethylation; DNA glycosylase; GeneChip analysis

---

Double fertilization is a unique characteristic of flowering plants. Fusion of an egg cell with a sperm cell gives rise to the diploid embryo and the second fusion of a diploid central cell with another sperm cell gives rise to the triploid endosperm. The endosperm is the functional homologue of the mammalian placenta and provides nutrients to the developing embryo.

Gene imprinting refers to the differential allele expression of paternal and maternal alleles depending on parental origin. Imprinting occurs in the placenta of mammals [1] and in the endosperm of plants [2,3]. *MEDEA* (*MEA*) is a Polycomb group (PcG) gene imprinted in *Arabidopsis* [4–6]. The maternal *MEA* allele is expressed while the paternal allele is silenced [7]. There are two mechanisms regulating *MEA*

imprinting. Demethylation by the DEMETER (DME) DNA glycosylase activates expression of the maternal *MEA* allele [8,9] in the central cell of the female gametophyte. Silencing of the paternal *MEA* allele is not directly controlled by DNA methylation. Instead, PcG proteins, including maternally expressed *MEA*, repress expression of the paternal *MEA* allele in the endosperm [9]. Thus, *MEA* is a self-imprinted gene [9–11]. DME demethylation occurs by the base excision DNA repair process [9,12]. DME first excises 5-methylcytosine by its glycosylase activity, and then the AP lyase activity of DME nicks the DNA strand. AP endonuclease cleavage generates a 3' hydroxyl, DNA polymerase replaces the excised 5-methylcytosine with cytosine, and finally DNA ligase seals the nick. By excising 5-methylcytosine, DME prevents CpG hypermethylation of its target genes, thus activating their gene expression. DME is specifically expressed in the central cell

---

\* Corresponding author. Fax: +82 2 871 4445.  
E-mail address: yhc@snu.ac.kr (Y. Choi).

of the female gametophyte before fertilization [8]. Its temporal and spatial expression is essential for the establishment of *MEA* imprinting and seed viability.

In mammals, there are approximately 100 imprinted genes with approximately half expressing the maternal allele and half expressing the paternal allele (<http://www.geneimprint.com>). In *Arabidopsis*, only four genes imprinted in the endosperm have been discovered [7,13–15]. *MEA*, *FWA*, and *FIS2* are expressed maternally and are controlled by DME [8,9,13,15]. *PHERESI* (*PHE1*) is expressed paternally and PcG proteins including *MEA*, which is activated by DME, regulate its maternal silencing [14,16]. Therefore, DME is a key regulator of genomic imprinting in flowering plants.

Considering the number of imprinted genes in animals, it is possible that there are more imprinted genes in *Arabidopsis* that are controlled by DME. To test this hypothesis, we used an RNA profiling approach to search for other genes regulated by DME. We generated transgenic plants that ectopically express DME under the control of the *CaMV* promoter in pollen and stamen [17]. Since the endogenous *DME* gene is not expressed in the wild type male reproductive organs (pollen and stamen), ectopic DME expression reveals the expression of novel target genes. We extracted poly(A)<sup>+</sup> RNA from pollen and stamens harvested from *CaMV:DME* transgenic plants and wild type control plants. cRNAs were labeled and hybridized to Affymetrix *Arabidopsis* ATH1 GeneChip arrays. We compared and analyzed RNA profiles from both transgenic and wild type pollen and stamen and detected candidate DME-inducible genes, which were confirmed by semi-quantitative RT-PCR experiments. These DME-inducible genes shed light on the mechanism of DME-mediated demethylation and the role of gene imprinting in seed biology.

## Materials and methods

*Plant materials, transgenic plant isolation, and growth condition.* *CaMV:DME-4* transgenic plants were generated as described [17] and wild type plants (Columbia-*gl* ecotype) were used in this experiment. Transgenic plants contained the *NPTII* gene, which confers resistance to kanamycin. Plants were grown in standard greenhouse conditions (16 h light/8 h dark).

*Pollen collection and RNA extraction.* Pollen grains and stamens were collected from stage 13 and 14 flowers and processed as described [17,18]. Total RNAs were extracted using TRIzol (Invitrogen) as described [7]. Poly(A)<sup>+</sup> RNA was selected from total RNA by using the Dynabeads mRNA Purification kit (DynaL Inc.) according to procedures provided by the manufacturer.

*Preparation of the probes, hybridization, and scanning.* Double-stranded cDNAs were generated using Superscript Choice system (Invitrogen) and were purified by phenol/chloroform extraction. Biotin-labeled cRNA probes were synthesized using BioArray High Yield RNA Transcript Labeling Kit (T7) (Enzo Life Sciences) and purified using RNeasy mini spin columns (Qiagen). Concentration of cRNA probe was determined using a UV-spectrophotometer and the size range of synthesized cRNA was determined by fractionating cRNAs on a 1.3% agarose/formaldehyde gel. cRNAs were fragmented with fragmentation buffer (40 mM Tris-acetate, pH 8.1, 100 mM KOAc, and 30 mM MgOAc) and cRNA size was

determined on a 1.3% agarose/formaldehyde gel. Fragmented biotin-labeled cRNAs were hybridized to the *Arabidopsis* ATH1 GeneChip Array (Affymetrix) for 16 h according to manufacturer's protocol (Affymetrix). The arrays were stained with streptavidin phycoerythrin (SAPE, Molecular probes), goat IgG (Sigma), and anti-streptavidin biotinylated antibody. GeneChip arrays were scanned and analyzed using Affymetrix Microarray Analysis Suite (MAS) 5.0 (Santa Clara, CA). All probe sets on the array were scaled globally to a target intensity of 500 using MAS 5.0 default parameters. All GeneChip data reported in this paper have been deposited in the GEO database (<http://www.ncbi.nlm.nih.gov/geo>) as series GSE9408.

*RT-PCR analysis.* Reverse transcription was done with Retro Transcript (Ambion) following instructions provided by the manufacturer. Primers used in these experiments are in [Supplementary Materials](#).

## Results and discussion

### *DME expression in CaMV:DME pollen*

We previously generated transgenic plants bearing a transgene (*CaMV:DME*) with the 6.8 kb full-length *DME* cDNA ligated to the cauliflower mosaic virus promoter (*CaMV*) [17]. In multiple *CaMV:DME* transgenic lines, DME was ectopically expressed in stamen and pollen, and we detected expression of known DME target genes, *MEA* [17] and *FWA* (data not shown). This suggests that DME expression is both necessary and sufficient to activate target gene expression, and ectopic DME expression in stamen or pollen can induce DME target genes that are normally silenced in the male reproductive organs [17].

### *Transcriptional profiling of CaMV:DME and wild type*

To gain a global view of the transcription network that is regulated by DME, we used Affymetrix GeneChip arrays to identify novel genes up-regulated by ectopic DME expression in *CaMV:DME* pollen and stamen compared to the wild type counterparts. Pollen grains were harvested from *CaMV:DME* and control wild type flowers at stages 13 (open flowers) and 14 (self-pollinating) [19]. We also collected stamens at the same stages from wild type and *CaMV:DME* plants. Stamens included filaments, and anthers containing mature and immature pollen grains. Poly(A)<sup>+</sup> RNAs were extracted from the above four independent tissue samples, and their biotin-labeled cRNA probes were synthesized from double-stranded cDNAs generated from poly(A)<sup>+</sup> RNAs. Biotinylated cRNA probes were hybridized to Affymetrix *Arabidopsis* ATH1 GeneChip arrays and data were analyzed using the GeneSpring program. Scatter plot analysis shows that wild type and *CaMV:DME* RNA profiles from pollen are well correlated ( $r = 0.99$ ) (Fig. 1A). Similar results were seen when wild type and *CaMV:DME* stamen RNAs were compared ( $r = 0.95$ ) (Fig. 1B). The high correlation coefficients between wild type and *CaMV:DME* pollen and stamen indicate that ectopic expression of DME induces a small number of genes.

From the transcriptional profiling data, 6392 and 7281 RNAs were detected in wild type and *CaMV:DME* pollens, respectively (Fig. 2A). These numbers are in close



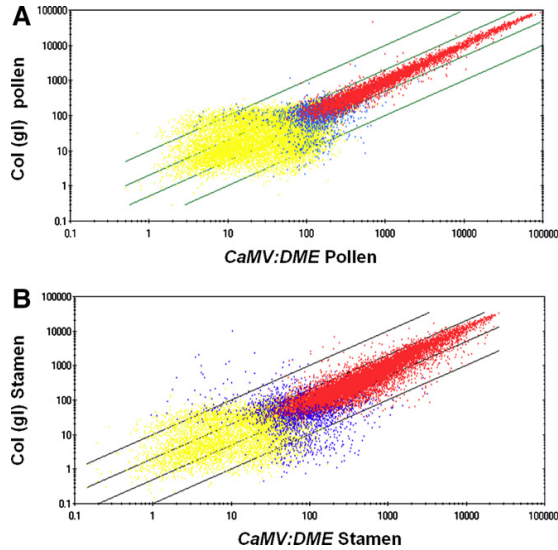


Fig. 1. Scatter plots of wild type and *CaMV:DME* pollen and stamen. Scatter plot analysis of wild type and *CaMV:DME* pollen (A) and stamen (B) was generated using MAS 5.0 of all probe sets represented on the Affymetrix *Arabidopsis* ATH1 GeneChip array. A red dot indicates RNA with a detection call of present by MAS 5.0 in both wild type and *CaMV:DME* pollen or stamen. A yellow dot indicates RNA with a detection call of absent in both wild type and *CaMV:DME* pollen or stamen. Blue dots indicate different detection calls by MAS 5.0 for wild type and *CaMV:DME* pollen and stamen. (For interpretation of the references in color in this figure legend, the reader is referred to the web version of this article.)

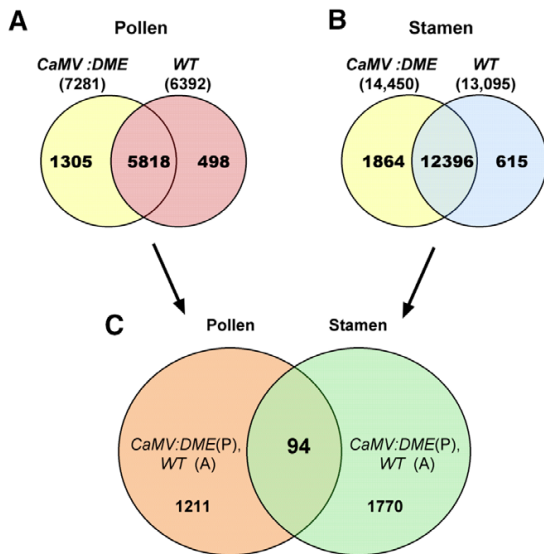


Fig. 2. Analysis of RNA profiles. Venn diagrams of genes expressed in wild type and *CaMV:DME* pollen (A), and wild type and *CaMV:DME* stamen (B). Ninety-four genes show DME-dependent expression in both *CaMV:DME* stamen and pollen (C).

agreement with RNA profiling data of pollen obtained by others [20,21]. RNAs (1305) were detected in *CaMV:DME* pollen and not in wild type control pollen. Similarly, 13,095 and 14,450 RNAs were detected in wild type and *CaMV:DME* stamens, respectively (Fig. 2B). These numbers are in close agreement with RNA profiling data from stamens obtained by others [22]. RNAs (1864) were detected in *CaMV:DME* but not in wild type stamens. We found 94 genes with DME-inducible expression in both *CaMV:DME* stamen and pollen when comparing *CaMV:DME* pollen- and stamen-specific RNAs (Fig. 2C). Thus, only a small number of DME-inducible genes were detected in both pollen and stamen.

These 94 genes were sorted into different functional categories (Supplementary Table 1). There are 4 genes related to cell growth and division, 5 genes to cell structure, 3 genes to disease and defense, 3 genes for energy, 2 genes for intracellular traffic, 9 genes for metabolism, 4 genes for post-transcription, 7 genes for protein destination and storage, 1 gene for protein synthesis, 3 genes for secondary metabolism, 10 genes for signal transduction, 14 genes for transcription, 4 genes for transporter, 2 genes for transposon, and 23 genes unclassified (Supplementary Table 1).

We did not detect expression of *MEA* or *FWA* in the RNAs isolated from *CaMV:DME* pollen or stamen. It is possible that the level of *MEA* and *FWA* RNA was too low to be detected by the *Arabidopsis* whole genome Affymetrix probe array [23].

We also sorted the 94 genes into gene ontologies using the bulk retrieval tool at TAIR (<http://www.arabidopsis.org/tools/bulk/go>) (Fig. 3). The overall transcription profile is similar to the whole *Arabidopsis* genome profile [24] suggesting DME might regulate genes with a broad range of functional categories. Gene ontology analysis identified three molecular functions enriched in the DME-inducible RNA list compared to the whole *Arabidopsis* genome: DNA or RNA binding (11.1% versus 5.4%), kinase activity (8.3% versus 5.0%), and transcription factor activity (8.3% versus 5.0%). Although DME induces a wide range of genes, this enrichment suggests that target genes may share common regulatory pathways or be involved in similar functions such as DNA or RNA binding, kinase activity, and transcription factor activity.

In *Arabidopsis*, a number of genes are demethylated by ROS1, DML2, and DML3 [25,26]. ROS1 (REPRESSOR OF SILENCING 1) is a DME homolog that maintains the transcriptionally active states of a *RD29A::LUCIFERASE* (*RD29A::LUC*) reporter gene and the endogenous *RD29A* gene [12,27,28]. DML2 (DEMETER-LIKE 2) and DML3 (DEMETER-LIKE 3) are two other DME-like genes in *Arabidopsis*. Genes regulated by ROS1 have been identified by comparing RNA profiles from wild type and *ros1* mutant plants [26]. Genes demethylated by ROS1, DML2, and DML3 have been revealed by comparing wild type and *ros1*, *dml2*, and *dml3* triple mutant genome-tiling microarrays [25]. We found little overlap between our DME target genes and these published results. Thus,



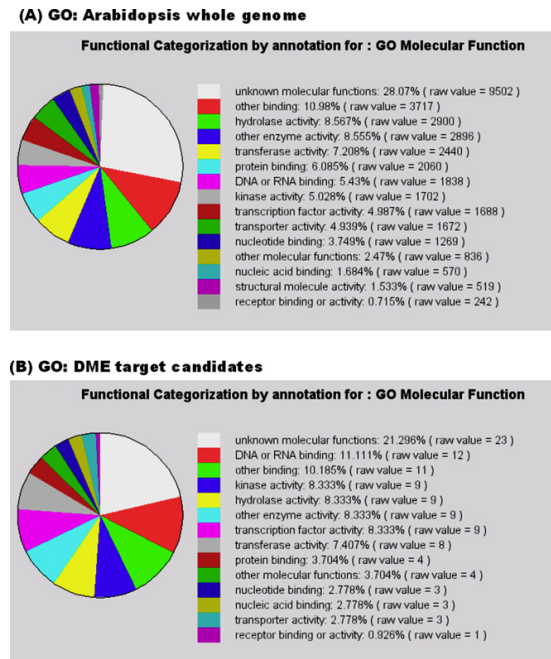


Fig. 3. Functional categorization of wild type *Arabidopsis* whole genome (A) and genes induced by DME (B). Ninety-four genes were sorted out to the functional categories classified in the databases of gene ontology at TAIR. DME-dependent downstream genes show similar distribution of molecular function with the *Arabidopsis* whole genome.

although DME and DMLs have similar DNA demethylase activity [9,12,25], their putative target genes are very distinctive based on the data from microarray and genome-tiling analyses.

#### Semi-quantitative RT-PCR for the possible candidate genes

We chose six of the 94 genes that represent a wide range of biological processes for independent validation. Semi-quantitative RT-PCR was performed to measure the RNA level of putative DME-downstream genes from wild type and *CaMV:DME* pollen tissue. We found that RT-PCR results of these six genes were consistent with the microarray gene profiling results (Fig. 4). That is, expression was not detected in control wild type pollen and was induced in *CaMV:DME* pollen. The list of representative candidate DME target genes includes: *AT1G60930* which encodes a DEAD/DEAH box domain RecQ family DNA helicase involved in DNA replication, recombination, RNA metabolism, and gene silencing [29]; *AT1G80960* which encodes a F-box-related protein with leucine-rich repeat domain; *AT2G18050* that encodes a linker Histone H1-3, involved in DNA binding and chromatin structure that is induced by dehydration and ABA [30]; *AT4G01780*, encoding a XH/XS domain-containing protein; *AT4G23800*, a high mobility group (HMG) family

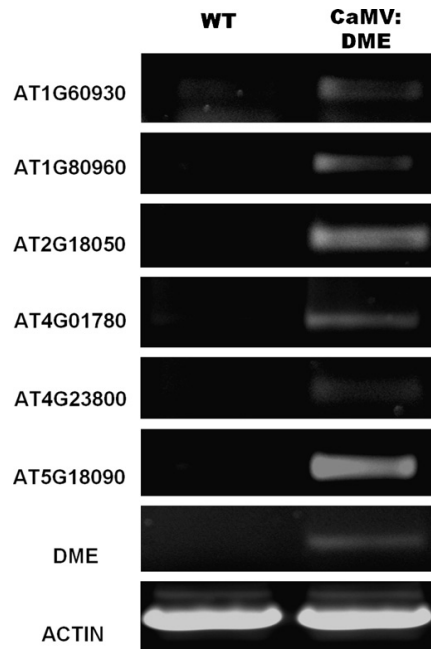


Fig. 4. Semi-quantitative RT-PCR validation of putative DME target genes. Expression patterns of the possible DME-downstream genes were confirmed by semi-quantitative RT-PCR from independently isolated pollen tissue.

protein and a non-histone component of chromatin that is involved in DNA binding inducing DNA-dependent transcription, replication, and repair mechanism [31]; and *AT5G18090*, encoding a transcription factor B3 family protein. Taken together, these results suggest that DME, either directly or indirectly, regulates expression of the above genes in addition to *MEA*, *FWA*, and *FIS2*. Since their induction in expression is in response to ectopic expression of DME in tissues where DME is not normally active, they might be bona fide target genes of DME.

Whether DME changes the methylated status of each downstream gene, their biological function, and whether they are imprinted genes remain to be assessed. These future studies will provide a more global picture of gene imprinting and will help us to understand the mechanism whereby DME activates gene expression by DNA demethylation.

#### Acknowledgments

We would like to thank Professor Bob Goldberg (UCLA) in whose laboratory the GeneChip experiments were carried out. We also thank Professor John Harada (UC Davis) for critical comments on this manuscript. This work was supported by grants from the second stage of the Brain Korea 21 project, Korea Research Foundation (KRF-2005-070-C00129) and BioGreen21 Program

(20050401034633), Rural Development Administration, Republic of Korea to Y. Choi. This work was also supported by grants to R.L.F. from NIH (GM069415) and the USDA (2005-02355), and Ceres, Inc.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2007.10.092.

## References

- [1] A.L. Fowden, C. Sibley, W. Reik, M. Constancia, Imprinted genes, placental development and fetal growth, *Horm. Res.* 65 (2006) 50–58.
- [2] M. Gehring, Y. Choi, R.L. Fischer, Imprinting and seed development, *Plant Cell* 16 (2004), doi:10.1105/tpc.017988.
- [3] R.J. Scott, M. Spielman, Genomic imprinting in plants and mammals: how life history constrains convergence, *Cytogenet. Genome Res.* 113 (2006) 53–67.
- [4] U. Grossniklaus, J.P. Vielle-Calzada, M.A. Hoepfner, W.B. Gagliano, Maternal control of embryogenesis by *MEDEA*, a polycomb group gene in *Arabidopsis*, *Science* 280 (1998) 446–450.
- [5] T. Kiyosue, N. Ohad, R. Yadegari, M. Hannon, J. Dinneny, D. Wells, A. Katz, L. Margossian, J. Harada, R.B. Goldberg, R.L. Fischer, Control of fertilization-independent endosperm development by the *MEDEA* polycomb gene in *Arabidopsis*, *Proc. Natl. Acad. Sci. USA* 96 (1999) 4186–4191.
- [6] M. Luo, P. Bilodeau, A. Koltunow, E.S. Dennis, W.J. Peacock, A.M. Chaudhury, Genes controlling fertilization-independent seed development in *Arabidopsis thaliana*, *Proc. Natl. Acad. Sci. USA* 96 (1999) 296–301.
- [7] T. Kinoshita, R. Yadegari, J.J. Harada, R.B. Goldberg, R.L. Fischer, Imprinting of the *MEDEA* polycomb gene in the *Arabidopsis* endosperm, *Plant Cell* 11 (1999) 1945–1952.
- [8] Y. Choi, M. Gehring, L. Johnson, M. Hannon, J.J. Harada, R.B. Goldberg, S.E. Jacobsen, R.L. Fischer, DEMETER, a DNA Glycosylase domain protein, is required for endosperm gene imprinting and seed viability in *Arabidopsis*, *Cell* 110 (2002) 33–42.
- [9] M. Gehring, J.H. Huh, T.F. Hsieh, J. Penterman, Y. Choi, J.J. Harada, R.B. Goldberg, R.L. Fischer, DEMETER DNA glycosylase establishes *MEDEA* polycomb gene self-imprinting by allele-specific demethylation, *Cell* 124 (2006) 495–506.
- [10] C. Baroux, V. Gagliardini, D.R. Page, U. Grossniklaus, Dynamic regulatory interactions of Polycomb group genes: MEDEA autoregulation is required for imprinted gene expression in *Arabidopsis*, *Genes Dev.* 20 (2006) 1081–1086.
- [11] P.E. Jullien, A. Katz, M. Oliva, N. Ohad, F. Berger, Polycomb group complexes self-regulate imprinting of the Polycomb group gene *MEDEA* in *Arabidopsis*, *Curr. Biol.* 16 (2006) 486–492.
- [12] T. Morales-Ruiz, A.P. Ortega-Galisteo, M.I. Ponferrada-Marin, R.R. Martinez-Macias, R.R. Ariza, T. Roldan-Arjona, *DEMETER* and *REPRESSOR OF SILENCING1* encode 5-methylcytosine DNA glycosylases, *Proc. Natl. Acad. Sci. USA* 103 (2006) 6853–6858.
- [13] T. Kinoshita, A. Miura, Y. Choi, Y. Kinoshita, X. Cao, S.E. Jacobsen, R.L. Fischer, T. Kakutani, One-way control of *FWA* imprinting in *Arabidopsis* endosperm by DNA methylation, *Science* 303 (2004) 521–523.
- [14] C. Kohler, D.R. Page, V. Gagliardini, U. Grossniklaus, The *Arabidopsis thaliana* MEDEA Polycomb group protein controls expression of *PHERESI* by parental imprinting, *Nat. Genet.* 37 (2005) 28–30.
- [15] P.E. Jullien, T. Kinoshita, N. Ohad, F. Berger, Maintenance of DNA methylation during the *Arabidopsis* life cycle is essential for parental imprinting, *Plant Cell* 18 (2006) 1360–1372.
- [16] G. Makarevich, O. Leroy, U. Akinci, D. Schubert, O. Clarenz, J. Goodrich, U. Grossniklaus, C. Kohler, Different Polycomb group complexes regulate common target genes in *Arabidopsis*, *EMBO Rep.* 7 (2006) 947–952.
- [17] Y. Choi, J.J. Harada, R.B. Goldberg, R.L. Fischer, An invariant aspartic acid in the DNA glycosylase domain of DEMETER is necessary for transcriptional activation of the imprinted *MEDEA* gene, *Proc. Natl. Acad. Sci. USA* 101 (2004) 7481–7486.
- [18] D. Preuss, B. Lemieux, G. Yen, R.W. Davis, A conditional sterile mutation eliminates surface components from *Arabidopsis* pollen and disrupts cell signaling during fertilization, *Genes Dev.* 7 (1993) 974–985.
- [19] J.L. Bowman, S.G. Mansfield, Embryogenesis, in: J. Bowman (Ed.), *Arabidopsis: An Atlas of Morphology and Development*, Springer-Verlag, New York, 1994, pp. 351–361.
- [20] D. Honys, D. Twell, Transcriptome analysis of haploid male gametophyte development in *Arabidopsis*, *Genome Biol.* 5 (2004) R85.
- [21] C. Pina, F. Pinto, J.A. Feijo, J.D. Becker, Gene family analysis of the *Arabidopsis* pollen transcriptome reveals biological implications for cell growth, division control, and gene expression regulation, *Plant Physiol.* 138 (2005) 744–756.
- [22] A. Mandaokar, B. Thines, S. Byongchul, B.M. Lange, G. Choi, Y.J. Koo, Y.J. Yoo, Y.D. Choi, G. Choi, J. Browse, Transcriptional regulators of stamen development in *Arabidopsis* identified by transcriptional profiling, *Plant J.* 46 (2006) 984–1008.
- [23] J.C. Redman, B.J. Haas, G. Tanimoto, C.D. Town, Development and evaluation of an *Arabidopsis* whole genome Affymetrix probe array, *Plant J.* 38 (2004) 545–561.
- [24] T.Z. Berardini, S. Mundodi, R. Reiser, E. Huala, M. Garcia-Hernandez, P. Zhang, L.M. Mueller, J. Yoon, A. Doyle, G. Lander, N. Moseyko, D. Yoo, I. Xu, B. Zoeckler, M. Montoya, N. Miller, D. Weems, S.Y. Rhee, Functional annotation of the *Arabidopsis* genome using controlled vocabularies, *Plant Physiol.* 135 (2004) 1–11.
- [25] J. Penterman, D. Zilberman, J.H. Huh, T. Ballinger, S. Henikoff, R.L. Fischer, DNA demethylation in the *Arabidopsis* genome, *Proc. Natl. Acad. Sci. USA* 104 (2007) 6752–6757.
- [26] J. Zhu, A. Kapoor, V.V. Sridhar, F. Agius, J.K. Zhu, The DNA glycosylase/lyase ROS1 functions in pruning DNA methylation patterns in *Arabidopsis*, *Curr. Biol.* 17 (2007) 54–59.
- [27] Z. Gong, T. Morales-Ruiz, R.R. Ariza, T. Roldan-Arjona, L. David, J.J. Zhu, *ROS1*, a repressor of transcriptional gene silencing in *Arabidopsis*, encodes a DNA Glycosylase/Lyase, *Cell* 111 (2002) 803–814.
- [28] F. Agius, A. Kapoor, J.K. Zhu, Role of the *Arabidopsis* DNA glycosylase/lyase ROS1 in active DNA demethylation, *Proc. Natl. Acad. Sci. USA* 103 (2006) 11796–11801.
- [29] F. Hartung, H. Plchova, H. Puchta, Molecular characterisation of RecQ homologues in *Arabidopsis thaliana*, *Nucleic Acid Res.* 28 (2000) 4275–4282.
- [30] R. Ascenzi, J.S. Gantt, Molecular genetic analysis of the drought-inducible linker histone variant in *Arabidopsis thaliana*, *Plant Mol. Biol.* 41 (1999) 159–169.
- [31] J.L. Riechmann, J. Heard, G. Martin, L. Reuber, C. Jiang, J. Keddie, L. Adam, O. Pineda, O.J. Ratcliffe, R.R. Samaha, R. Creelman, M. Pilgrim, P. Broun, J.Z. Zhang, D. Ghandehari, B.K. Sherman, G. Yu, *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes, *Science* 290 (2000) 2105–2110.

## APPENDIX D

Braybrook, S.A., Stone, S.L., Park, S., Bui, A.Q., Le, B.H., Fischer, R.L., Goldberg, R.B. and Harada, J.J. (2006). **Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis.** PNAS 103, 3468-3473.

# Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis

Siobhan A. Braybrook<sup>\*†</sup>, Sandra L. Stone<sup>\*</sup>, Soomin Park<sup>\*</sup>, Anhthu Q. Bui<sup>‡</sup>, Brandon H. Le<sup>‡</sup>, Robert L. Fischer<sup>§</sup>, Robert B. Goldberg<sup>\*¶</sup>, and John J. Harada<sup>\*¶||</sup>

<sup>\*</sup>Section of Plant Biology, College of Biological Sciences, University of California, Davis, CA 95616; <sup>†</sup>Graduate Program in Plant Biology and <sup>‡</sup>Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, CA 90024; and <sup>§</sup>Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720

Contributed by Robert B. Goldberg, December 30, 2005

The B3 domain protein LEAFY COTYLEDON2 (LEC2) is required for several aspects of embryogenesis, including the maturation phase, and is sufficient to induce somatic embryo development in vegetative cells. Here, we demonstrate that LEC2 directly controls a transcriptional program involved in the maturation phase of seed development. Induction of LEC2 activity in seedlings causes rapid accumulation of RNAs normally present primarily during the maturation phase. Several RNAs encode proteins with known roles in maturation processes, including seed-storage and lipid-body proteins. Clustering analyses identified other LEC2-induced RNAs not previously shown to be involved in the maturation phase. We show further that genes encoding these maturation RNAs all possess in their 5' flanking regions RY motifs, DNA elements bound by other closely related B3 domain transcription factors. Our finding that recombinant LEC2 specifically binds RY motifs from the 5' flanking regions of LEC2-induced genes provides strong evidence that these genes represent transcriptional targets of LEC2. Although these LEC2-induced RNAs accumulate primarily during the maturation phase, we show that a subset, including *AGL15* and *IAA30*, accumulate in seeds containing zygotes. We discuss how identification of LEC2 target genes provides a potential link between the roles of LEC2 in the maturation phase and in the induction of somatic embryogenesis.

*Arabidopsis* | B3 domain

Embryogenesis in higher plants can be divided conceptually into two distinct phases. Early in embryogenesis, during the morphogenesis phase, the basic body plan of the plant is established with regional specification of apical–basal and radial domains from which morphological structures derive, fixation of polarity from specification of the shoot–root axis, and formation of embryonic tissue and organ systems (1–3). The morphogenesis phase is followed temporally by the maturation phase, although the two phases can overlap (4, 5). During the maturation phase, embryo cell-division rates decline markedly, embryo cells acquire the ability to withstand desiccation, and embryo cell growth occurs, with the accumulation of storage reserves that comprise lipids and proteins in *Arabidopsis* (6, 7). At the end of the maturation phase, the embryo becomes quiescent metabolically as the seed desiccates.

The maturation phase can be viewed as an interruption of an ancestral life cycle, as occurs in lower plants, in which there are no periods of maturation or dormancy separating the end of embryogenesis and the beginning of postembryonic development (4). Evolution of this unique mode of embryogenesis has enabled higher plants to make seeds. The ability to make seeds has provided tremendous selective advantages that, in part, account for the success of the angiosperms (8, 9). Little is known at a mechanistic level about the processes by which the maturation phase has been integrated into the higher plant life cycle.

LEAFY COTYLEDON2 (LEC2), along with ABA INSENSITIVE3 (ABI3), and FUSCA3 (FUS3), have been implicated to be major regulators of the maturation phase (reviewed in ref. 5). The LEC2 protein contains a DNA-binding B3 domain that is most closely related to that of FUS3 and ABI3 (10–12). The *lec2* mutation causes localized defects in embryo filling, seed protein accumulation, and desiccation tolerance (12, 13). *LEC2* expression is normally limited primarily to seed development, although *LEC2* RNA may be present at very low levels at other stages of the life cycle (12). Ectopic expression of *LEC2* causes accumulation of seed storage lipids and proteins in vegetative organs (ref. 14, and S.L.S., S. L. Paula, L. W. Kwong, J. E. Meuser, J. Pelletier, R.L.F., R.B.G., and J.J.H., unpublished work). The function of LEC2 is not limited to the maturation phase. The *lec2* mutation causes defects during the morphogenesis phase, and ectopic *LEC2* expression induces somatic embryo formation from vegetative cells (12, 13). Furthermore, the *lec2* mutation severely compromises the ability of *Arabidopsis* explants to form somatic embryos (15). These observations suggest that LEC2 plays several roles during embryogenesis, indicating that it is a central regulator of embryo development.

To gain insight into the role of LEC2 in embryogenesis, we have identified genes regulated by the LEC2 transcription factor. We show that a subset of genes activated by LEC2 is expressed predominantly during the maturation phase, and several have known roles in maturation processes. Several of these maturation genes are also expressed early in embryogenesis. We also show that all of these genes possess a common DNA motif that is bound by LEC2, providing strong evidence that these genes are regulated directly by LEC2. The identity of LEC2 target genes suggests a link between the ability of LEC2 to induce maturation processes and somatic embryogenesis.

## Results

### Ectopic LEC2 Activity Induces Changes in Seedling RNA Populations.

To obtain insight into the specific role of LEC2 in embryogenesis and the mechanisms by which it causes developmental abnormalities when expressed ectopically (refs. 12 and 14, and S.L.S., S. L. Paula, L. W. Kwong, J. E. Meuser, J. Pelletier, R.L.F., R.B.G., and J.J.H., unpublished work), we identified genes

Conflict of interest statement: No conflicts declared.

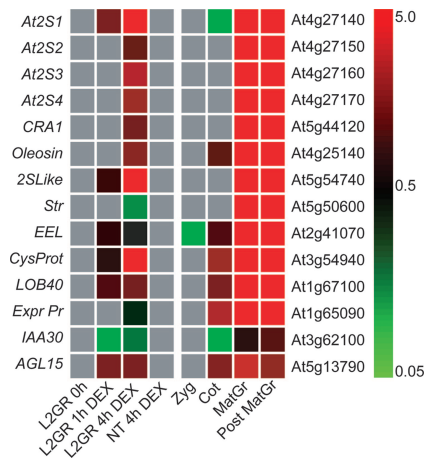
Abbreviations: Dex, dexamethasone; GUS,  $\beta$ -glucuronidase; GR, glucocorticoid receptor; LEC2, LEAFY COTYLEDON2; qRT-PCR, quantitative RT-PCR.

Data deposition: Data for the microarray experiments reported in this paper have been deposited in the Gene Expression Omnibus database, [www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo) (accession no. GSE3959).

<sup>†</sup>To whom correspondence may be addressed. E-mail: bobg@ucla.edu.

<sup>||</sup>To whom correspondence may be addressed at: Section of Plant Biology, College of Biological Sciences, University of California, One Shields Avenue, Davis, CA 95616. E-mail: [jjharada@ucdavis.edu](mailto:jjharada@ucdavis.edu).

© 2006 by The National Academy of Sciences of the USA

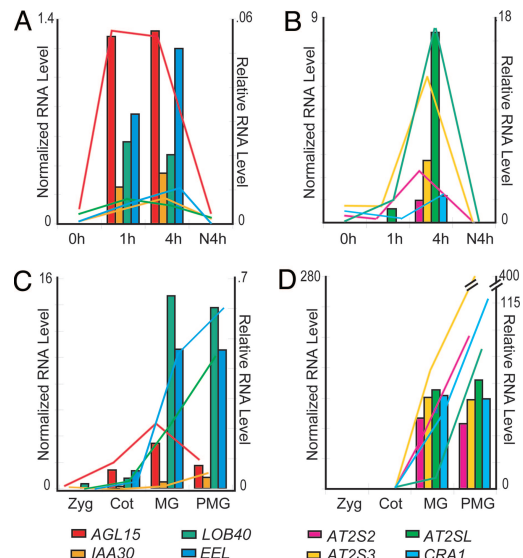


**Fig. 1.** Profiles of RNAs induced by LEC2 that are prevalent during the maturation phase. Representation of mean normalized expression data for 14 LEC2-induced RNAs in LEC2 induction experiments (Left) and during seed development (Right). Gene names and AGI loci are listed. Color scale shows relative RNA levels, with gray representing RNAs not present. L2GR, seedlings treated with Dex for the indicated period; NT, nontransgenic seedlings; Zyg, seeds containing zygotes; Cot, embryos at the cotyledon stage; MatGr, mature green stage; Post MatGr, postmature green stage; Str, steroleosin; CysProt, cysteine proteinase; Expr Pr, expressed protein.

regulated by the LEC2 transcription factor. An inducible form of LEC2 consisting of LEC2 fused with the steroid-binding domain of the glucocorticoid receptor (LEC2-GR) (16) was used. It has been shown that LEC2 activity could be induced by treating plants containing the *35S::LEC2-GR* chimeric gene with the steroid-hormone analogue dexamethasone (Dex) (ref. 14, and S.L.S., S. L. Paula, L. W. Kwong, J. E. Meuser, J. Pelletier, R.L.F., R.B.G., and J.J.H., unpublished work).

To identify genes regulated by LEC2, we isolated RNA from *35S::LEC2-GR* seedlings grown for 8 days that were treated with Dex for either 1 or 4 hours (1-h or 4-h Dex). Nontransgenic 8-day seedlings do not contain LEC2 RNA detectable by RT-PCR, suggesting that the endogenous gene is not active at this stage (12). As negative controls, RNAs from *35S::LEC2-GR* seedlings that were not treated with Dex (0-h Dex) and from nontransgenic plants treated with Dex for 4 h (NT Dex) were used. RNAs were hybridized with ATH1 GeneChip DNA microarrays that contain probes corresponding to  $\approx 24,000$  *Arabidopsis* genes. Specific RNAs were considered present in a population if they were judged by statistical algorithms (see *Materials and Methods*) to be present in both biological replicates of a treatment. RNAs were considered not present if they were absent or not present in both biological replicates. It is possible that some RNAs designated not present may be detectable in these treatments.

The following criteria were used to define RNAs induced by LEC2 activity: (i) RNAs present only in the 1-h Dex treatment (and not in any other treatment); (ii) RNAs present only in the 4-h Dex treatment; and (iii) RNAs present in both the 1-h and 4-h Dex treatments but not in any other treatments. As listed in Table 2 and summarized in Fig. 5, which are published as supporting information on the PNAS web site, the levels of 718 RNAs were altered positively by induction of LEC2 activity by using these criteria. Because LEC2 gene expression is limited predominantly to embryogenesis (12), biologically relevant target genes are expected to be expressed in seeds. Therefore, we used results from a series of DNA microarray experiments with ATH1 GeneChips that profiled RNA populations during *Ara-*



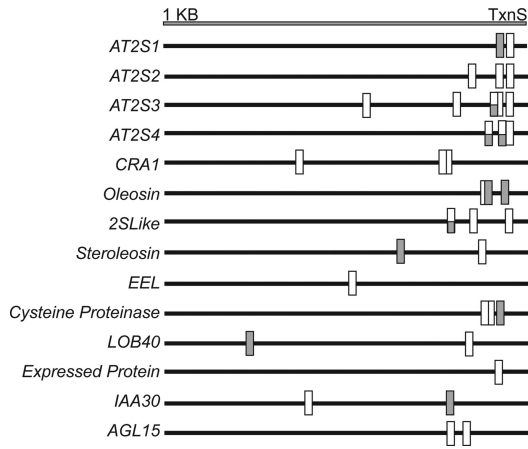
**Fig. 2.** qPCR data validate DNA microarray results and reveal an early role for several LEC2-induced RNAs. Normalized RNA levels from DNA microarray experiments (bars) and relative RNA levels from qPCR experiments (lines) in LEC2 induction experiments (A and B) and during seed development (C and D). (A and C) *AGL15*, *IAA30*, *LOB40*, and *EEL*. (B and D) *AT2S2*, *AT2S3*, *AT2SL*, and *CRA1*.

*bidsopsis* seed development to search for LEC2-induced RNAs present in seeds containing zygotes 24 h after pollination, cotyledon-stage embryos 7–8 days after pollination (DAP), mature green-stage embryos 13–14 DAP, and postmature green-stage embryos 18 to 19 DAP ([www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE680](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE680), and B.H.L., unpublished results). Four hundred twenty LEC2-induced RNAs encoding proteins with divergent functions were present at one or more stages of seed development. We conclude that induction of LEC2 activity in seedlings activated the expression of a number of genes normally expressed during seed development.

**RNAs Involved in Maturation Processes Are Induced by Ectopic LEC2 Activity.** We found that RNAs encoding seed proteins that include 2S and 12S storage proteins and oleosin, the major lipid-body protein, constituted a prominent group of LEC2-induced RNAs that accumulated in seeds. During seed development, seed protein RNAs accumulate, specifically during the maturation phase, indicating that LEC2 regulates genes involved in maturation processes.

To identify other RNAs involved in processes that occur during the maturation phase, we used statistical clustering methods to group RNAs based on their profiles in the LEC2 induction experiments and in seeds containing zygotes, cotyledon-stage, mature green-stage, and postmature green-stage embryos. The latter two stages of seed development represent the maturation phase. We found that 88 RNAs were present in clusters containing seed protein RNAs using either *K* means (17), self-organizing maps (18) or hierarchical (19) clustering methods (see *Materials and Methods*), and 21 RNAs clustered together with all three methods. A summary of RNA accumulation patterns for a subset of these RNAs is shown in Fig. 1. We validated the DNA microarray results for 14 of 17 RNAs by using quantitative RT-PCR (qRT-PCR) as represented in Fig. 2 and detailed in Table 3, which is published as supporting information





**Fig. 3.** LEC2 target genes contain upstream RY motifs. Representation of the location of RY motifs (CATGCA) present within 1 kb of the transcription start site of 14 LEC2 target genes. White boxes, RY motifs on sense strand; gray boxes, RY motifs on antisense strand with respect to the gene.

on the PNAS web site. qRT-PCR experiments also showed that most of these RNAs accumulated primarily, although not exclusively, during the maturation phase (Fig. 2 and Table 3).

In addition to the 2S and 12S storage proteins At2S1–4 and CRA1, respectively, and oleosin, this group also included RNAs encoding a 2S-like protein, steroleosin, and ENHANCED EM LEVEL (EEL), proteins known or postulated to play roles in maturation processes (20, 21). Detection of these RNAs indicated that the clustering methods were efficient in identifying LEC2-induced RNAs with roles in the maturation phase. In addition, we found that the other RNAs in this group encoding cysteine proteinase, LATERAL ORGAN BOUNDARY40 (LOB40), IAA30, AGAMOUS-LIKE15 (AGL15), and the expressed protein AT1G65090 exhibited similar profiles, suggesting that they may also have roles in the maturation phase. We conclude that LEC2 rapidly activates genes involved in maturation processes and designate this group as maturation RNAs.

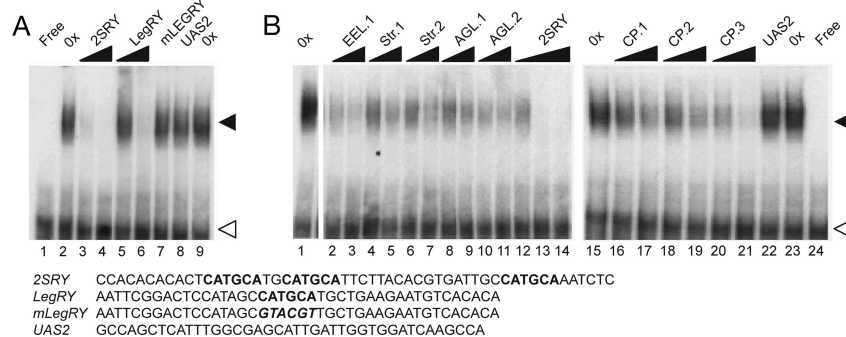
The sensitive qRT-PCR analyses showed that several of the

maturation RNAs, AGL15, IAA30, EEL, and LOB40, were detected in seeds containing zygotes. This result opened up the possibility that these maturation RNAs also play roles at the earliest stage of embryogenesis. The potential significance of this result will be discussed.

**LEC2 Binds with RY Motifs Upstream of Genes Encoding LEC2-Induced Maturation RNAs.** Genes encoding maturation RNAs are potential targets directly regulated by the LEC2 transcription factor. To address this possibility, we asked whether these genes share a common DNA sequence in their 5' flanking regions that may serve as a LEC2-binding site. Two pattern-recognition algorithms, Multiple Em for Motif Elicitation (<http://meme.sdsc.edu/meme/intro.html>) (22) and MotifSampler (<http://homes.esat.kuleuven.be/~thijs/Work/MotifSampler.html>) (23) identified a motif that was significantly enriched within 1 kb of the transcription start site in the 5' flanking regions of the maturation genes. As diagrammed in Fig. 3, this CATGCA DNA sequence, the RY motif (24), was present at least once within 1 kb of the 5' flanking regions of the 14 LEC2-induced maturation genes (see Table 4, which is published as supporting information on the PNAS web site), although no discernible positional or strand-specific pattern was observed. Thus, the RY motif is a candidate to serve as a LEC2-binding site required for the activation of LEC2 target genes.

DNA-binding studies provided additional evidence that LEC2 directly regulates maturation genes. Electrophoretic mobility-shift assays, shown in Fig. 4A, indicate that recombinant LEC2 fused with glutathione *S*-transferase bound the 2SRY oligonucleotide containing three closely spaced RY motifs from the 5' flanking region of the *At2S3* gene (lanes 2 and 9). Competition experiments showed that unlabeled 2SRY oligonucleotide (lanes 3 and 4) and an oligonucleotide containing a RY motif derived from the 5' flanking region of the legumin storage protein gene (lanes 5 and 6) (25) competed for LEC2 binding. By contrast, a legumin oligonucleotide with a mutated RY motif (lane 7) and an oligonucleotide lacking a RY motif (lane 8) did not compete for binding. Together, these results show that LEC2 binds specifically with the RY motif.

We next investigated whether LEC2 binds RY motifs located in the 5' flanking regions of other LEC2-induced maturation genes. Electrophoretic mobility-shift assays were performed in which oligonucleotides containing RY motifs from the upstream regions of representative maturation genes were used as competitors of LEC2 binding with the 2SRY probe. As shown in Fig.



**Fig. 4.** LEC2 binds specifically to RY motifs upstream of LEC2 target genes. (A) Binding of the 2SRY probe DNA oligonucleotides with recombinant LEC2 in the presence and absence of the indicated unlabeled competitor DNA oligonucleotides. LegRY contains a RY motif, whereas the RY motif in mLegRY is mutated. UAS2 does not contain a RY motif. Free, no added LEC2 protein; 0x, no added competitor. (B) DNA oligonucleotides containing RY motifs upstream of LEC2 target genes compete for binding of the 2SRY probe with LEC2. Competitors: EEL; Str, steroleosin; AGL, AGL15; CP, cysteine proteinase. Competition experiments were performed with 500- and 2,000-fold molar excesses of competitor, except for the 2SRY self-competition experiments in B in which 100-, 500-, and 2,000-fold molar excesses were used. Competition experiments with oligonucleotides lacking a RY motif were done with a 2,000-fold molar excess.

**Table 1. LEC2 activates a gene with an upstream RY motif in planta**

Construct	GUS/LUX*
35S:GUS	2.99 ± 0.06
NAPIN:GUS	<0.002
NAPIN:GUS + 35S:LEC2	0.286 ± 0.06

\*GUS constructs were cobombarded with the 35S:Luciferase gene. GUS activity was normalized to luciferase activity.

4B, every oligonucleotide tested that contained a RY motif competed for binding of the 2SR Y probe, indicating that they were bound by recombinant LEC2. Oligonucleotides containing three different RY motifs upstream of the cysteine proteinase gene competed to different extents (lanes 18–23). Differences in binding affinity were observed between RY motifs derived from different genes, suggesting that oligonucleotides containing RY motifs did not compete equally for 2SR Y binding and that DNA sequences flanking the RY motif may influence binding. Our findings that induction of LEC2 activated genes expressed primarily during the maturation phase and that LEC2 bound specific sequences in the 5' flanking regions of these genes strongly suggest that LEC2 transcriptionally regulates genes involved in maturation processes.

**LEC2 Activates a Gene Containing an Upstream RY Motif in Planta.** The ability of LEC2 to bind to the RY motif *in vitro* prompted us to determine whether LEC2 could activate a promoter containing a RY motif in planta. We used biolistic bombardment of *Brassica napus* leaves (26) with 35S:LEC2 and a reporter gene encoding  $\beta$ -glucuronidase (GUS) fused with the 5' flanking region from the *B. napus* napin gene that has three RY motifs (NAPIN:GUS) (27). A 35S:Luciferase gene was used to normalize for the efficiency of bombardment. As shown in Table 1,  $\beta$ -glucuronidase activity was not detected in leaves bombarded with the NAPIN:GUS gene alone. However, we did detect GUS activity in leaves cobombarded with NAPIN:GUS and 35S:LEC2, suggesting that LEC2 activates the NAPIN promoter. Together, these results suggest that the LEC2 transcription factor activates genes containing upstream RY motifs in plants.

## Discussion

**LEC2 Directly Activates Genes Involved in Maturation Processes.** We provide evidence that LEC2 confers maturation characteristics to vegetative plants by directly activating genes involved in maturation processes. We and others have shown that ectopic LEC2 activity causes postembryonic organs to assume characteristics of maturing embryos that include accumulation of seed storage reserves (refs. 12 and 14, and S.L.S., S. L. Paula, L. W. Kwong, J. E. Meuser, J. Pelletier, R.L.F., R.B.G., and J.J.H., unpublished work). Because LEC2 induces somatic embryogenesis (12), it was possible that the onset of maturation processes resulted from progression of somatic embryos through the morphogenesis and maturation phases of embryogenesis. However, our demonstration that induction of LEC2 causes rapid increases in the expression of many maturation genes indicates that LEC2 can activate maturation processes directly.

Many LEC2-induced RNAs are directly involved in storage reserve accumulation, a hallmark of the maturation phase. These RNAs include 2S and 12S storage proteins, oleosin, and steroleosin, an abundant oil body protein that shares similarity with the anchoring hydrophobic domain of oleosin and contains a sterol-binding domain (21, 28). Another LEC2-induced maturation RNA encodes a regulator of reserve accumulation, EEL. This bZIP transcription factor works in cooperation with ABI3

and ABI5 to regulate expression of the *Em* gene that encodes an abundant seed protein (20). Clustering analysis also identified RNAs encoding AGL15, IAA30, LOB40, a cysteine proteinase, and the expressed protein AT1G65090. Although these RNAs have no known role in maturation and insertional mutations in the corresponding genes did not obviously affect seed development (ref. 29, and S.L.S. and J.J.H., unpublished results), their induction by LEC2 and accumulation patterns during seed development suggest their involvement in maturation processes. Thus, integrating RNA profiles from the LEC2 induction experiments with those during seed development has permitted discovery of genes with potential roles in the maturation phase.

Our results suggest that these maturation genes are direct targets of the LEC2 transcription factor for several reasons. First, the levels of these RNAs increased within 1 or 4 h of induction of LEC2 activity (Fig. 1). Second, LEC2 binds with RY motifs that are located in the 5' flanking regions of all induced maturation genes (Figs. 3 and 4). Third, LEC2 activates a reporter gene with a RY motif in its 5' flanking region in planta (Table 1). Taken together, these studies identify a network of genes regulated transcriptionally by LEC2 that play roles in the maturation phase. Consistent with this conclusion, others have suggested that LEC2 activates the *At2S3* and *S3* oleosin genes transcriptionally (14, 30). Of the 718 genes encoding LEC2-induced RNAs and the 420 genes encoding LEC2-induced RNAs present during seed development (Table 2), 41% and 40%, respectively, contain RY motifs within 1 kb of their 5' flanking regions. Thus, it is likely that many induced RNAs represent genes activated indirectly by LEC2. The protein synthesis inhibitor cycloheximide is often used to avoid identifying RNAs regulated indirectly by an induced transcription factor. However, we avoided this approach, because seed protein RNA levels were elevated in nontransgenic seedlings treated with cycloheximide (R.W. Kwong and J.J.H., unpublished results).

The closely related B3 domain transcription factors LEC2, FUS3, and ABI3 have been implicated to play major roles in controlling gene expression during the maturation phase. Loss-of-function mutations in each gene cause defects in seed protein RNA accumulation, although to different degrees (12–14, 30–35). Ectopic expression of each gene causes accumulation of seed protein RNAs in vegetative organs, although ABA is required or enhances seed protein RNA accumulation in plants overexpressing *ABI3* and *FUS3*, respectively (Fig. 1) (33, 36, 37). Our demonstration that LEC2 binds with the same DNA element bound by FUS3 and ABI3 (30, 38–40), the RY motif, provides a partial explanation for similarities in the gain-of-function phenotypes. Consistent with this observation, LEC2, ABI3, and FUS3 share identical or conserved amino acid residues at positions in the B3 domain implicated as being responsible for DNA-binding specificity based on the solution structure of the B3 domain protein RAV1 (41). Thus, all three transcription factors bind RY motifs through their B3 domains and activate maturation-specific genes, although we note that ABI3 activity is influenced by an additional domain that binds ABA response elements (40).

Although the precise regulatory relationship between these transcription factors remains to be determined, our results indicate that LEC2 is sufficient to activate maturation genes in seedlings, a stage at which *LEC2* is not normally expressed. If LEC2 works in concert with other factors to activate the maturation genes, these other factors must be present in seedlings. We note, however, that 298 RNAs up-regulated by LEC2 in induction experiments were not detected during seed development (Table 2). Some of these 298 RNAs may be present in seeds but were not detected in the microarray experiments because of the limited sensitivity of detection or because the RNAs are present at seed stages that were not analyzed. Alternatively, LEC2-mediated transcriptional activation of these

genes may be inhibited specifically during seed development by other mechanisms, such as chromatin conformation or negatively acting transcription factors.

**Roles for LEC2 Target Genes in Somatic Embryogenesis.** A striking effect of ectopic *LEC2* expression is the induction of somatic embryo development (12). Although this study has focused largely on *LEC2*-induced RNAs that accumulate primarily during the maturation phase, it has provided clues about the mechanisms by which *LEC2* induces somatic embryogenesis. *Arabidopsis* somatic embryos can be formed from maturation-phase zygotic embryos treated with the synthetic auxin, 2,4-D (42). Although auxin is generally required to induce somatic embryogenesis (43), most tissues, including morphogenesis-phase zygotic embryos, do not give rise to somatic embryos in response to auxin treatment (42). A simple interpretation of these observations is that auxin is the induction signal for somatic embryogenesis and that specific tissues, such as those from maturation-phase zygotic embryos, are competent to respond to this signal and enter embryonic pathways.

Several *LEC2*-induced RNAs, but particularly *AGL15* and *IAA30*, accumulate predominantly during the maturation phase and are also detected in seeds containing zygotes (Fig. 2; and Table 3), opening up the possibility that these RNAs play roles at the earliest stages of embryo development. Expression of *AGL15* is correlated with embryogenesis in that *AGL15* RNA is detected in all tissues tested undergoing zygotic or somatic embryogenesis (44–46). Although a loss-of-function mutant allele of *AGL15* does not provide insight into its function, ectopic expression of *AGL15* affects embryonic programs (47). Specifically, cultured *35S:AGL15* zygotic embryos produce secondary embryonic calli that express maturation genes. Furthermore, ectopic *AGL15* expression enhances the competency of shoot apical meristems to undergo somatic embryogenesis in response to auxin treatment. Thus, *LEC2*, by inducing *AGL15* expression, may cause tissues to become competent to respond to the auxin signal and initiate somatic embryogenesis.

Another *LEC2*-induced RNA, *IAA30*, encodes one of a family of auxin signaling proteins (48) that is enriched in the quiescent center of *Arabidopsis* root apical meristems (49). Thus, *IAA30* may play one of two distinct roles in the induction of somatic embryogenesis. First, *IAA30* may confer competency for somatic embryogenesis. This hypothesis is based on *IAA30* expression in root stem cells, the quiescent center. Somatic embryos can form from enlarged shoot apical meristems of seedlings (50), suggesting that meristematic cells have competence to undergo somatic embryogenesis. Second, *IAA30* may be involved in inducing somatic embryogenesis through its role in auxin signaling. Auxin signals the initiation of somatic embryogenesis in cultured cells. *IAA30* may alter plant responses to auxin or cause an increase in free auxin levels, thereby initiating somatic embryogenesis. Therefore, *LEC2* may promote somatic embryo development by affecting both induction and competence. We have recently shown that induction of *LEC2* activity causes a rapid increase in the activity on an auxin-induced promoter (S.L.S., S. L. Paula, L. W. Kwong, J. E. Meuser, J. Pelletier, R.L.F., R.B.G., and J.J.H., unpublished work), suggesting that it triggers induction of somatic embryogenesis through auxin signaling. *LEC2* may also induce somatic embryo development by increasing tissue competency to undergo somatic embryogenesis through *AGL15* and, perhaps, *IAA30*.

### Conclusion

These observations are relevant to the role of *LEC2* in zygotic embryogenesis. Our studies provide evidence that *LEC2* is a major regulator of the maturation phase, directly activating

genes that are intimately involved in maturation processes. *LEC2* likely acts in concert with other transcription factors, including *ABI3* and *FUS3*, to regulate the maturation phase. We also show that *LEC2* activates genes such as *AGL15* and *IAA30* that are expressed in seeds containing zygotes. Given that *LEC2* is expressed at the earliest stage of seed development tested, well before the maturation phase, and remains active until midmaturation phase (12), our results are consistent with a model in which *LEC2* acts at the earliest stages of embryogenesis to activate genes that induce zygotes to undergo embryonic development. However, it is not clear why other maturation genes are not activated early in embryogenesis with the onset of *LEC2* activity. Additional studies are needed to define other mechanisms, such as changes in chromatin conformation or transcriptional repression, that control the expression of maturation genes during seed development.

### Materials and Methods

**Plant Materials.** Growth of *Arabidopsis thaliana* (L.) Heyn ecotype Wassilewskija was described in ref. 51. Plants containing *35S:LEC2-GR* that encodes *LEC2* fused at its carboxyl terminus with the steroid-binding domain of the glucocorticoid receptor (16) were generated as described by S.L.S., S. L. Paula, L. W. Kwong, J. E. Meuser, J. Pelletier, R.L.F., R.B.G., and J.J.H. (unpublished work).

**LEC2-GR Induction Experiment.** Seedlings were grown on medium for 8 days as described in ref. 51. Seedlings containing the *35S:LEC2-GR* transgene were either frozen in liquid nitrogen immediately (0-h Dex) or placed in liquid medium (52) containing 30  $\mu$ M Dex (dexamethasone-21-acetate in ethanol, Sigma) for either 1 h or 4 h with gentle shaking. Nontransgenic seedlings were treated similarly with Dex for 4 h. Treated seedlings were frozen in liquid nitrogen and stored at  $-80^{\circ}\text{C}$ . Two biological replicates were prepared for each treatment.

**GeneChip Hybridization and Data Analysis.** cRNA was prepared and hybridized with Affymetrix ATH1 GeneChips as described in the Gene Expression Omnibus accession GSE680 ([www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE680](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE680)). One GeneChip was hybridized for each biological replicate. Microarray suite software package (MAS 5.0, Affymetrix) (53) was used to assess probe set signals and to generate present/absent calls for each RNA. Each data set was then normalized to the 50th percentile signal by using the GENESPRING data analysis platform (Silicon Genetics), and a mean normalized expression value was determined for each treatment from the replicate values. Data for the microarray experiments has been deposited in the Gene Expression Omnibus ([www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo)) under accession no. GSE3959.

**Statistical Clustering.** Normalized expression data for 718 *LEC2*-induced RNAs, irrespective of their present or absent calls, were clustered by using three methods (GENESPRING data analysis platform): *K*-means (parameters: 5 clusters, 100 iterations, Pearson correlation) (17), self-organizing maps (parameters: 4 clusters, 220,000 iterations, neighborhood radius = 2.0) (18), and hierarchical (parameter: Pearson correlation) (19). All RNAs that clustered with storage protein RNAs were examined, and present and absent calls were used to eliminate RNAs without reliable values. A second clustering analysis that identified *AGL15* was done by using a subset of the 718 RNAs that encoded transcription factors to identify RNAs present at lower levels that clustered with *EEL*.

**qRT-PCR Analysis.** Real-time, qRT-PCR was performed as described in ref. 55. Primer sets were derived from Affymetrix



probe sets for each RNA (see Table 5, which is published as supporting information on the PNAS web site).

**Electrophoretic Mobility-Shift Assays.** Recombinant GST-LEC2 was generated by using the LEC2 cDNA cloned into pGEX-KG (54). Fusion protein produced in *Escherichia coli* BL21 cells was partially purified by glutathione-affinity chromatography.

DNA-binding reactions were done by using 0.1  $\mu$ g of total protein (7 ng of GST-LEC2) incubated with 1.6 fmol of [<sup>32</sup>P]-labeled 2SRY probe in binding buffer (25 mM Hepes, pH 8.0, 1 mM DTT, 200  $\mu$ g/ml BSA, 75 mM KCl, and 10% Glycerol). For competition assays, unlabeled competitor was incubated with protein for 10 min at room temperature before the addition of labeled probe. After the addition of labeled

probe, reactions were incubated for 20 min at room temperature. Binding reactions were fractionated at 4°C by 5% polyacrylamide-gel electrophoresis.

**Particle-Bombardment Experiments.** Particle-bombardment experiments were conducted as described in ref. 26, except that *NAPIN:GUS* (27) was used as the reporter plasmid, and *35S:LEC2* (12) was used as the effector plasmid.

We acknowledge use of the University of California, Los Angeles, Plant Genomics Center for the Affymetrix GeneChip hybridization experiments, and we thank Makoto Murase (Mitsubishi Chemical) for the gift of the *NAPIN:GUS* plasmid and Julie Pelletier and Raymond Kwong for technical assistance with this project. This work was supported by grants from Ceres, Inc. (Thousand Oaks, CA), and the National Science Foundation.

- Berleth, T. & Chatfield, S. (2002) in *The Arabidopsis Book*, eds. Somerville, C. R. & Meyerowitz, E. M. (Am. Soc. Plant Biologists, Rockville, MD), 10.1199/tab.0009, www.aspb.org/publications/arabidopsis.
- Jurgens, G. (2001) *EMBO J.* **20**, 3609–3616.
- Laux, T., Wurschum, T. & Breuninger, H. (2004) *Plant Cell* **16**, S190–S202.
- Harada, J. J. (1997) in *Advances in Cellular and Molecular Biology of Plants, Volume 4, Cellular and Molecular Biology of Seed Development*, eds. Larkins, B. A. & Vasi, I. K. (Kluwer, Dordrecht, The Netherlands), pp. 545–592.
- Vicente-Carbajosa, J. & Carbonero, P. (2005) *Int. J. Dev. Biol.* **49**, 645–651.
- Wobus, U. & Weber, H. (1999) *Curr. Opin. Plant Biol.* **2**, 33–38.
- Baud, S., Boutin, J.-P., Miquel, M., Lepiniec, L. & Rochat, C. (2002) *Plant Physiol. Biochem.* **40**, 151–160.
- Steeves, T. A. (1983) *Can. J. Bot.* **61**, 3550–3560.
- Walbot, V. (1978) in *Dormancy and Developmental Arrest*, ed. Clutter, M. E. (Academic, New York), pp. 113–166.
- Giraudat, J., Hauge, B. M., Valon, C., Smalle, J., Parcy, F. & Goodman, H. M. (1992) *Plant Cell* **4**, 1251–1261.
- Luerssen, H., Kirik, V., Herrmann, P. & Misera, S. (1998) *Plant J.* **15**, 755–764.
- Stone, S. L., Kwong, L. W., Yee, K. M., Pelletier, J., Lepiniec, L., Fischer, R. L., Goldberg, R. B. & Harada, J. J. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 11806–11811.
- Meinke, D. W., Franzmann, L. H., Nickle, T. C. & Yeung, E. C. (1994) *Plant Cell* **6**, 1049–1064.
- Santos Mendoza, M., Dubreucq, B., Miquel, M., Caboche, M. & Lepiniec, L. (2005) *FEBS Lett.* **579**, 4666–4670.
- Gaj, M., Zhang, S., Harada, J. & Lemaux, P. (2005) *Planta* **222**, 977–988.
- Sablowski, R. W. M. & Meyerowitz, E. M. (1998) *Cell* **92**, 93–103.
- Hartigan, J. A. & Wong, M. A. (1979) *Appl. Statist.* **28**, 100–108.
- Kohonen, T. (2000) *Self-Organizing Maps* (Springer, Berlin).
- Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868.
- Bensmihen, S., Rippha, S., Lambert, G., Jublot, D., Pautot, V., Granier, F., Giraudat, J., Parcy, F. (2002) *Plant Cell* **14**, 1391–1403.
- Jolivet, P., Roux, E., d'Andrea, S., Davanture, M., Negroni, L., Zivy, M. & Chardot, T. (2004) *Plant Physiol. Biochem.* **42**, 501–509.
- Bailey, T. L., Elkan, C. (1994) in *Second International Conference on Intelligent Systems for Molecular Biology* (AAAI Press), pp. 28–36.
- Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P., Moreau, Y. (2002) *J. Comp. Biol.* **9**, 447–464.
- Dickinson, C., Evans, R. & Nielsen, N. (1988) *Nucleic Acids Res.* **16**, 371.
- Baumlein, H., Nagy, I., Villarroel, R., Inze, D. & Wobus, U. (1992) *Plant J.* **2**, 233–239.
- Zhang, J. Z., Santes, C. M., Engel, M. L., Gasser, C. S. & Harada, J. J. (1996) *Plant Physiol.* **110**, 1069–1079.
- Josefsson, L., Lenman, M., Ericson, M. & Rask, L. (1987) *J. Biol. Chem.* **262**, 12196–12201.
- Lin, L.-J., Tai, S. S. K., Peng, C.-C. & Tzen, J. T. C. (2002) *Plant Physiol.* **128**, 1200–1211.
- Wang, H., Caruso, L. V., Downie, A. B. & Perry, S. E. (2004) *Plant Cell* **16**, 1206–1219.
- Kroj, T., Savino, G., Valon, C., Giraudat, J. & Parcy, F. (2003) *Development (Cambridge, U.K.)* **130**, 6065–6073.
- Baumlein, H., Misera, S., Luerben, H., Kollé, K., Horstmann, C., Wobus, U. & Muller, A. J. (1994) *Plant J.* **6**, 379–387.
- Nambara, E., Keith, K., McCourt, P. & Naito, S. (1995) *Development (Cambridge, U.K.)* **121**, 629–636.
- Parcy, F., Valon, C., Raynal, M., Gaubier-Comella, P., Delseny, M. & Giraudat, J. (1994) *Plant Cell* **6**, 1567–1582.
- Parcy, F., Valon, C., Kohara, A., Misera, S. & Giraudat, J. (1997) *Plant Cell* **9**, 1265–1277.
- Keith, K., Kraml, M., Dengler, N. G. & McCourt, P. (1994) *Plant Cell* **6**, 589–600.
- Gazzarrini, S., Tsuchiya, Y., Lumba, S., Okamoto, M. & McCourt, P. (2004) *Dev. Cell* **7**, 373–385.
- Kagaya, Y., Okuda, R., Ban, A., Toyoshima, R., Tsutsumida, K., Usui, H., Yamamoto, A. & Hattori, T. (2005) *Plant Cell Physiol.* **46**, 300–311.
- Reidt, W., Wohlfarth, T., Ellerstrom, M., Czihal, A., Tewes, A., Ezcurra, I., Rask, L. & Baumlein, H. (2000) *Plant J.* **21**, 401–408.
- Monke, G., Altschmied, L., Tewes, A., Reidt, W., Mock, H.-P., Baumlein, H. & Conrad, U. (2004) *Planta* **219**, 158–166.
- Ezcurra, I., Wycliffe, P., Nehlin, L., Ellerstrom, M. & Rask, L. (2000) *Plant J.* **24**, 57–66.
- Yamasaki, K., Kigawa, T., Inoue, M., Tateno, M., Yamasaki, T., Yabuki, T., Aoki, M., Seki, E., Matsuda, T., Tomo, Y., et al. (2004) *Plant Cell* **16**, 3448–3459.
- Gaj, M. D. (2001) *Plant Cell Tiss. Org. Cult.* **64**, 39–46.
- Zimmerman, J. L. (1993) *Plant Cell* **5**, 1411–1423.
- Heck, G. R., Perry, S. E., Nichols, K. W. & Fernandez, D. E. (1995) *Plant Cell* **7**, 1271–1282.
- Perry, S. E., Nichols, K. W. & Fernandez, D. E. (1996) *Plant Cell* **8**, 1977–1989.
- Perry, S. E., Lehti, M. D. & Fernandez, D. E. (1999) *Plant Physiol.* **120**, 121–130.
- Harding, E. W., Tang, W., Nichols, K. W., Fernandez, D. E. & Perry, S. E. (2003) *Plant Physiol.* **133**, 653–663.
- Liscum, E. & Reed, J. W. (2002) *Plant Mol. Biol.* **49**, 387–400.
- Nawy, T., Lee, J.-Y., Colinas, J., Wang, J. Y., Thongrod, S. C., Malamy, J. E., Birnbaum, K. & Benfey, P. N. (2005) *Plant Cell* **17**, 1908–1925.
- Mordhorst, A. P., Voerman, K. J., Hartog, M. V., Meijer, E. A., van Went, J., Koornneef, M. & de Vries, S. C. (1998) *Genetics* **149**, 549–563.
- West, M. A. L., Matsudaira Yee, K. L., Danao, J., Zimmerman, J. L., Fischer, R. L., Goldberg, R. B. & Harada, J. J. (1994) *Plant Cell* **6**, 1731–1745.
- Olsen, L. J., Ettinger, W. F., Damsz, B., Matsudaira, K., Webb, M. A. & Harada, J. J. (1993) *Plant Cell* **5**, 941–952.
- Hubbell, E., Liu, W.-M. & Mei, R. (2002) *Bioinformatics* **18**, 1585–1592.
- Guan, K. L. & Dixon, J. E. (1991) *Anal. Biochem.* **192**, 262–267.
- Yamagishi, K., Nagata, N., Yee, K. M., Braybrook, S. A., Pelletier, J., Fujioka, S., Yoshida, S., Fischer, R. L., Goldberg, R. B. & Harada, J. J. (2005) *Plant Physiol.* **139**, 163–173.

## APPENDIX E

Sanders, P.M., Bui, A.Q., **Le, B.H.**, Goldberg, R.B. (2005). **Differentiation and degeneration of cells that play a major role in tobacco anther dehiscence.** Sex Plant Reprod, 17, 219-241

Paul M. Sanders · Anhtu Q. Bui · Brandon H. Le  
Robert B. Goldberg

## Differentiation and degeneration of cells that play a major role in tobacco anther dehiscence

Received: 14 July 2004 / Accepted: 22 August 2004 / Published online: 19 October 2004  
© Springer-Verlag 2004

**Abstract** Dehiscence is the terminal step in anther development that releases pollen grains from the wall of each theca at a specific site between the two locules. In tobacco, two groups of cells—the circular cell cluster and the stomium—are required for anther dehiscence and define the position at which pollen is released. The processes responsible for the differentiation of the circular cell cluster and the stomium from cells in specific anther regions are unknown. Nor is it understood what initiates the programmed degeneration of these cell types that ultimately is responsible for pollen release from the anther. We characterized stomium and circular cell cluster differentiation and degeneration using both light and transmission electron microscopy throughout anther development, from the emergence of stamen primordia to anther dehiscence at flower opening. We observed that histological changes within primordium L1 and L2 cells destined to become the stomium and circular cell cluster occur at the same time after the differentiation of surrounding locule regions. Sub-epidermal cells that differentiate into the circular cell cluster divide, enlarge, and generate vacuoles with calcium oxalate crystals prior to any detectable changes in pre-stomium epidermal cells. Differentiation and division of cells that generate the stomium occur after cell degeneration initiates in the circular cell cluster. Prior to dehiscence, the stomium consists of a small set of cytoplasmically dense cells that are easily distinguished from their larger, highly vacuolate epidermal neighbors. Plasmodesmata connections within and between cells of

the stomium and circular cell cluster were observed at different developmental stages, suggesting that these cells communicate with each other. Circular cell cluster and stomium cell death is programmed developmentally and occurs at different times. Degeneration of the circular cell cluster occurs first, contributes to the formation of a bilocular anther, and generates the site of anther wall breakage. The stomium cell death process is complete at flower opening and provides an opening for pollen release from each theca. We used laser capture microdissection and real-time quantitative reverse-transcription polymerase chain reactions to demonstrate that stomium cells can be isolated from developing anthers and studied for the presence of specific mRNAs. Our data suggest that a cascade of unique gene expression events throughout anther development is required for the dehiscence program, and that the differentiation of the stomium and circular cell cluster in the interlocular region of the anther probably involves cell signaling processes.

**Keywords** Tobacco · Anther dehiscence · Stomium · Circular cell cluster · Laser capture microdissection

### Introduction

Dehiscence is the process that results in release of pollen grains from the anther at flower opening (Keijzer 1987; Bonner and Dickinson 1989; Goldberg et al. 1993; Beals and Goldberg 1997; Scott et al. 2004). In most flowering plants, the anther wall breaks along the lateral side of each anther half, or theca, within an indentation formed between the two locules (Fig. 1)—a region referred to as either the anther notch (Goldberg et al. 1995; Beals and Goldberg 1997; Sanders et al. 2000) or the stomial groove (D'Arcy 1996). In tobacco and other solanaceous plants (D'Arcy et al. 1996), two specialized cell types are found within the notch region: the stomium and the circular cell cluster (Fig. 1; Koltunow et al. 1990;

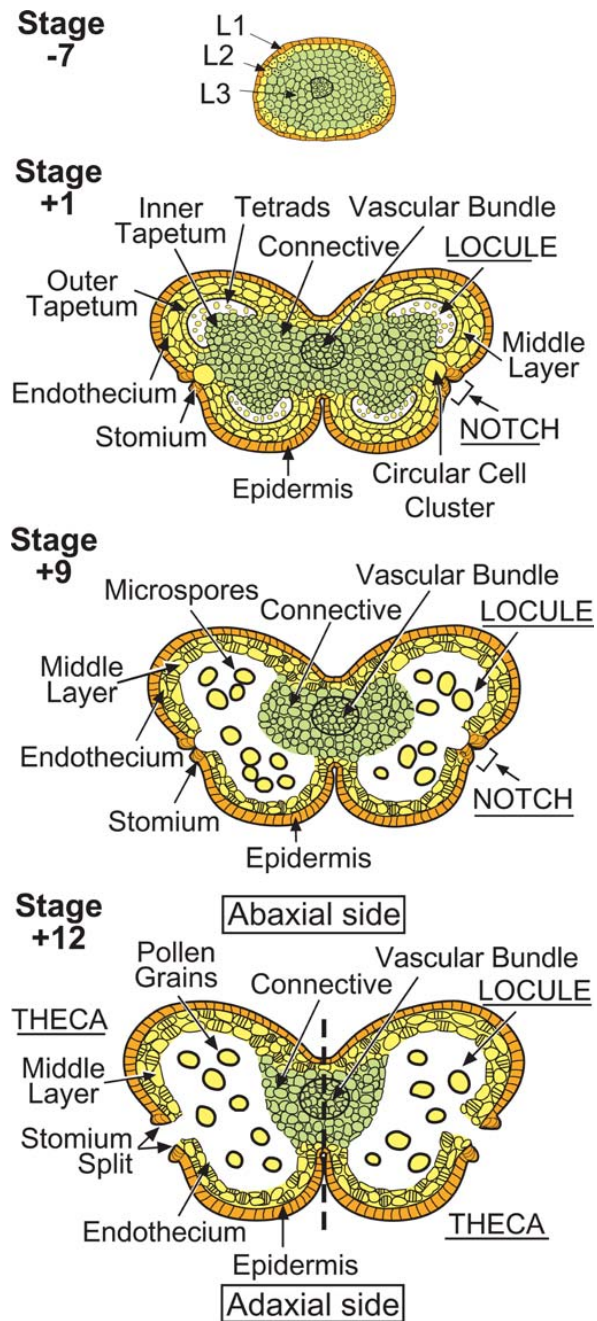
P. M. Sanders · A. Q. Bui · B. H. Le · R. B. Goldberg (✉)  
Department of Molecular, Cell,  
and Developmental Biology,  
University of California,  
Los Angeles, CA 90095–1606, USA  
E-mail: bobg@ucla.edu  
Tel.: +1-310-8259093  
Fax: +1-310-8258201

*Present address:* P. M. Sanders  
AgriGenesis Biosciences, One Fox Street,  
P.O. Box 50, Auckland, New Zealand

**Fig. 1** Schematic representation of tobacco anther development based upon histological studies at the light microscope level (Satina and Blakeslee 1941; Joshi et al. 1967; Koltunow et al. 1990). The stages of anther development were previously described in Koltunow et al. (1990; phase 1, stages -7 to -1; phase 2, stage +1 to +12). The colors depict cells derived from the L1, L2 and L3 primordial layers at different stages of anther development (modified from Goldberg et al. 1993, 1995). In the tobacco anther, the inner middle layer is crushed during meiosis. The outer middle layer contributes to the anther wall, expands, and acquires fibrous bands similar to the endothecium. At stage +12, the dotted line dissects the anther in half to indicate the two theca. The two cell-types of the anther notch, the stomium and the circular cell cluster, span the length of the anther and are longitudinal columns of cells. The adaxial side of the anther is towards the center of the flower and faces the pistil. The abaxial side of the anther is outwards from the center of the flower and faces the petals

Goldberg et al. 1993, 1995; Beals and Goldberg 1997). The stomium is a specialized set of epidermal cells that degenerate and break at flower opening to allow pollen grains to be released (Fig. 1). In contrast, the circular cell cluster consists of highly specialized sub-epidermal cells that accumulate calcium oxalate crystals (Horner and Wagner 1980, 1992; Trull et al. 1991; D'Arcy et al. 1996; Iwano et al. 2004) and participate in the cell-death process that ultimately unites both locules of each theca into one large pollen chamber (Fig. 1; Koltunow et al. 1990; Goldberg et al. 1993, 1995; Beals and Goldberg 1997). The circular cell cluster has also been referred to as the intersporangial septum (Bonner and Dickinson 1989), hypodermal stomium (Horner and Wagner 1992), and oxalate package (D'Arcy et al. 1996). Recently, it has been shown that the calcium stored within the circular cell cluster becomes associated with pollen grains after circular cell cluster degeneration and facilitates the pollination process (Iwano et al. 2004). Anthers of most non-solanaceous plants (D'Arcy et al. 1996), including *Arabidopsis thaliana* (Sanders et al. 1999, 2000), do not have a circular cell cluster. However, specialized septum cells in the notch region of these anthers function analogously in dehiscence and, after their degeneration, unite the two locules of each theca into a confluent chamber (Venkatesh 1957; D'Arcy et al. 1996; Sanders et al. 2000). Neither the specification events that position the stomium and circular cell cluster in the territory between the developing anther locules, nor the mechanisms that control and coordinate their differentiation and degeneration, are known.

Tobacco anther development, including the dehiscence program, occurs in two phases (Table 1; Koltunow et al. 1990; Goldberg et al. 1993). In phase 1 (Table 1, stages -7 to -1; Koltunow et al. 1990), differentiation events establish the four locules and the stereotyped pattern of cell types that originate from the L1, L2, and L3 cell layers of the primordium (Fig. 1, stage -7) and that are present in the mature anther (Fig. 1, stage +1; Satina and Blakeslee 1941; Goldberg et al. 1993; Hill and Malmberg 1996). These include stomium and circular cell cluster formation in the notch



region of the expanding anther (Fig. 1, stage +1). In addition, microspore mother cells within the locules undergo meiosis to generate haploid microspores (Fig. 1, stage +1). During phase 2 (Table 1, stages +1 to +12; Koltunow et al. 1990), filament elongation occurs, anther enlargement takes place, pollen grains differentiate from microspores in each locule, fibrous bands appear in endothelial and connective cells, and cell degeneration events within the connective, circular cell

**Table 1** Major events within the notch region during tobacco anther development. Stages of tobacco anther development were taken from Koltunow et al. (1990). Major events were taken from our previous studies of tobacco anther development (Koltunow et al. 1990; Beals and Goldberg 1997) and from the observations presented here. *CCC* circular cell cluster, *MMC* cells microspore mother cells. Deep refers to periclinal cell divisions that increase the number of layers of cells seen in transverse sections (Fig. 5A). Events in italics are associated with the dehiscence program

Stage	Anther events	Circular cell cluster <sup>a</sup>	Stomium
Phase 1			
-7	Rounded primordia; tissue differentiation initiated	Cannot distinguish	Cannot distinguish
-6	Intense mitotic activity in four corners of primordia; invagination of inner side	Cannot distinguish	Cannot distinguish
-5	Wall layers, including endothecium and tapetum formed; connective established	Distinguish pre-CCC cells by position; single layer of L2 cells in future notch region	Distinguish pre-stomium cells by position in L1 layer between developing locules
-4	Tapetum and locules distinct; middle layer crushed; vacuoles formed in L3 connective cells; <i>stomium</i> and <i>CCC</i> specification occur	No vacuoles seen in pre-CCC cells; slight differential staining in paraplasm sections	No vacuoles seen in pre-stomium cells; difficult to distinguish from other L1 cells; slight differential staining of pre-stomium cells in paraplasm sections
-3	Meiosis begins; callose deposition between MMCs; <i>stomium</i> and <i>CCC</i> specification occur	Identify CCC cells by morphology; dense cytoplasm with few small vesicles, no large vacuoles as present in L2 and L3 neighbors; initial divisions occur; CCC is 2 cells deep in transverse section	Identify stomium cells by morphology; dense cytoplasm and absence of the large vacuole present in neighboring epidermal cells
-2	Meiosis in progress; tapetum large and multinucleate; <i>CCC</i> differentiation begins; thick callose walls between MMCs	Distinct differential staining in paraplasm sections; continued cell divisions with various angles of division planes; multiple small vesicles	Distinct differential staining in paraplasm sections; single epidermal layer; small vesicles in stomium cells
-1	Meiosis in progress; <i>CCC</i> differentiation and division occur	Similar to stage -2	Similar to stage -2
Phase 2			
+1	Meiosis complete; microspores in tetrads; <i>CCC</i> differentiation complete; <i>stomium</i> differentiation begins; all sporophytic tissue formed	12-14 cells in a circular cluster 2-4 cells deep; cells expanding; small vesicles begin to aggregate	Cytoplasmically dense cells; vesicles present, but no vacuole; large nuclear to cytoplasmic ratio; a single layer within the epidermis
+2	Microspores separate; <i>CCC</i> degeneration begins; <i>stomium</i> differentiation and division occur	Cells have expanded; vesicles aggregate to create large vacuole; deposition of calcium oxalate crystals; initiation of cell degeneration	Cell divisions initiated; periclinal cell division creates 2-cell structure across epidermis; vesicles aggregate to form large vacuole; cells remain small
+3	Tapetum shrinks; fibrous lignin bands appear in expanded endothecium and middle wall layers; <i>stomium</i> division occurs; pollen grains begin to differentiate	Advanced cell degeneration; no cell contents other than calcium crystals; cell walls degenerate	Periclinal cell division generates a stomium three cells deep within the width of an epidermal cell; occasional anticlinal division generates a stomium 4 cells across
+4	<i>CCC</i> adjacent to <i>stomium</i> degenerated; tapetum degenerates; <i>stomium</i> differentiation complete	Space previously occupied by circular cell cluster now only contains calcium crystals; no cellular features or cell walls	Stomium differentiation complete; multi-tiered structure with 9-12 cells that is 3 cells deep
+5 to +11	<i>Fibrous bands intensify</i> in endothecium and middle layer; connective degeneration occurs; anther becomes bilocular; pollen grains form	<i>CCC</i> no longer present and a "hole" is left in its place within the anther	No major changes observed
+12	<i>Stomium degenerates and breaks; walls flip open, pollen released and dehiscence occurs</i>		Stomium cell degeneration
+13	Senescence		

<sup>a</sup> In solanaceous plants, CCC also known as intersporangial septum (e.g., tomato, Bonner and Dickinson 1989), hypodermal stomium (e.g., sweet pepper; Horner and Wagner 1992), or oxalate package (D'Arcy et al. 1996)



cluster, and stomium lead to pollen release at flower opening (Fig. 1, stages +9 and +12; Goldberg et al. 1993, 1995; Beals and Goldberg 1997). We showed previously that the tobacco *TA56 thiol endopeptidase* gene is a marker for cell degeneration events that occur within the connective, circular cell cluster, and stomium (Koltunow et al. 1990; Goldberg et al. 1995; Beals and Goldberg 1997). The mechanisms responsible for switching the anther from a cell differentiation program (phase 1) to a cell degeneration and death program (phase 2) are not known.

Previous studies in our laboratory with transgenic tobacco plants showed that a functional stomium is required for anther dehiscence (Beals and Goldberg 1997). Targeted ablation of either the circular cell cluster and stomium or the stomium alone using a cytotoxic *barnase* gene driven by cell-specific promoters generates anthers that do not dehisce (Beals and Goldberg 1997). This indicates that dehiscence is not simply a mechanical process, but involves specific, exquisitely timed, cellular events. In addition, we (Sanders et al. 1999) and others (Dawson et al. 1993, 1999; Park et al. 1996) identified a large number of *A. thaliana* male-sterile dehiscence mutants, including those that either fail to dehisce (Dawson et al. 1993, 1999; Sanders et al. 1999; Steiner-Lange et al. 2003) or are defective in the timing of anther dehiscence (Sanders et al. 1999, 2000; Stintzi and Browse 2000; Ishiguro et al. 2001; Park et al. 2002; Von Malek et al. 2002). Analogous mutants have been found in other plant species (Kaul 1988). One *A. thaliana* non-dehiscence mutant, *ms35*, has a defect in the *MYB26* transcription factor gene and lacks endothelial cell fibrous bands, indicating the importance of these cells in anther dehiscence (Dawson et al. 1999; Steiner-Lange et al. 2003). In contrast, several late-dehiscence *A. thaliana* mutants, such as *DELAYED DEHISCENCE1 (DDE1)/OPR3* (Sanders et al. 2000; Stintzi and Browse 2000), *DELAYED DEHISCENCE2 (DDE2)/ALLENE OXIDASE SYNTHASE* (Sanders et al. 1999; Park et al. 2002; Von Malek et al. 2002), and *DEFECTIVE IN ANther DEHISCENCE1 (DAD1)* (Ishiguro et al. 2001), have defects in jasmonic acid (JA) biosynthesis, indicating that this hormone is involved in coordinating the timing of stomium breakage with flower development and opening. Other hormones, such as ethylene, auxin, and gibberellic acid (GA), play a role in dehiscence, because dehiscence mutant phenocopies can be induced by either blocking or over-expressing genes involved in hormone activity (Murray et al. 2003; Rieu et al. 2003; Cecchetti et al. 2004; Achard et al. 2004). The differentiation of specialized cell types required for anther dehiscence and the cell-degeneration processes that ultimately lead to pollen release at flower opening suggest that unique gene sets are required to program these events during anther development. What these genes are and how they are regulated remain to be determined.

The stomium and circular cell cluster provide a novel system to study the differentiation and cell-death processes that are required for anther dehiscence. As a first

step, we investigated the cellular and morphological events that occur in these cells throughout tobacco anther development at the level of the transmission electron microscope (TEM). We addressed three main questions: (1) when do cells that give rise to the circular cell cluster and stomium become specified within the anther primordium? (2) what primordium cell layer differentiates into the circular cell cluster? and (3) how are events leading to circular cell cluster and stomium formation and degeneration coordinated? We found that (1) differentiation events leading to circular cell cluster and stomium formation within the notch region occur after locule development begins, (2) the circular cell cluster is derived from founder cells in the primordium L2 layer that are contiguous to L1 cells destined to become the stomium, (3) circular cell cluster differentiation and degeneration occur before analogous events in the stomium take place, and (4) plasmodesmata connections occur between cells of the stomium and circular cell cluster. In addition, we demonstrate that laser capture microdissection (LCM) can be used to isolate stomium cells from differentiating anthers, and to detect individual stomium mRNAs. We propose that cell signaling plays a major role in specifying the circular cell cluster and stomium within the notch region during anther development.

---

## Materials and methods

### Growth of plants

Tobacco plants (*Nicotiana tabacum* cv. Samsun) were grown in the greenhouse under natural light conditions (Goldberg et al. 1978). Floral stages (phase 1, stages -7 to -1; phase 2, stages +1 to +12) used to follow anther development were described by Koltunow et al. (1990).

### Light microscopy of anther sections

Anthers were dissected from staged floral buds and fixed in glutaraldehyde as described by Cox and Goldberg (1988). The fixed anthers were dehydrated, embedded in paraffin, and sliced into 10  $\mu$ m sections (Cox and Goldberg 1988). Sections were stained with 0.05% toluidine blue and photographed with Kodak Gold 100 film (ISO 100/21) using bright-field microscopy in an Olympus compound microscope (Model BH2, Olympus, Lake Success, N.Y.).

### Transmission electron microscopy of anther sections

Staged anthers were hand-dissected and transverse sections (~1.0 mm) were fixed (2.0% glutaraldehyde, 0.05 M sodium phosphate pH 7.2, 0.1% tannic acid) for 2 h at room temperature, and then rinsed four times in 0.05 M sodium phosphate pH 7.2 (each rinse 15 min).

The anther sections were treated with 1% osmium tetroxide (in 0.05 M sodium phosphate pH 7.2) for 2 h at room temperature followed by dehydration in a graded ethanol series (10%, 20%, 35%, 50%, 70%, 85%, 95%, each for 30 min and three times in 100% ethanol, each for 1 h). The addition of 0.1% tannic acid in the fixative was taken from Botha et al. (1993) to enhance preservation of membranes and plasmodesmata structures. The anther sections were embedded in Spurr's epoxy resin (Spurr 1969; Ted Pella, Redding, Calif.) and sectioned using an ultramicrotome (Sorvall Model MT-600, Dupont, Wilmington, Del.). Sections of 1  $\mu$ m (stained with 1.0% toluidine blue at 42°C for 1–2 h) were examined to determine the region of the anther for further analysis and then 80 nm ultra-thin sections were prepared for TEM. These sections were placed on formvar-coated grids and stained with uranyl acetate and lead citrate. The anther sections were observed in a JEOL electron microscope 100CX II (JEOL, Peabody, Mass.) at 80 kV.

#### Transmission electron micrographs and figure preparation

The electron micrographs of anther notch regions and plasmodesmata were taken with Kodak EM film no. 4489. The electron micrographs of the tobacco anther notch were taken at a magnification of 1,900 $\times$ . At later stages of development, several individual electron micrographs were required to encompass the area of interest. For example, at anther stage  $-5$ , the composite photograph of the notch region was comprised of six negatives, whereas at anther stage  $+4$  the composite notch region photograph was comprised of 32 negatives. Individual photographs were joined together to create a composite image that was then digitally scanned (600 dpi) into a computer, either from the original image or from a copy negative for large format originals. The images were manipulated digitally with Adobe Photoshop (Adobe Systems, San Jose, Calif.) to enhance contrast and to remove the outlines of individual photographs. The electron micrographs of plasmodesmata were taken at a magnification of 29,000 $\times$ . The plasmodesmata images presented in Figure 8 were digital scans (600 dpi) of the TEM micrographs that were captured at 300% of their original size.

#### LCM of stomium cells

Stage  $+6$  anthers were trimmed to  $\sim 4$  mm and processed for LCM according to Kerk et al. (2003), using ethanol:acetic acid fixative. Fixed anthers were embedded in paraffin (Paraplast-plus, Fisher Scientific, Pittsburgh, Pa.) according to procedures used for in situ hybridization experiments in our laboratory (Cox and Goldberg 1988). Anthers were sliced into 10  $\mu$ m transverse sections (Reichert Jung 820-II Histocut Rotary

Microtome), floated in water onto penfoil slides (Leica Microsystems, Bannockburn, Ill.), dried overnight at 42°C on a slide warmer (Fisher Scientific), and stored at room temperature until used. Prior to LCM, anther sections were de-paraffinized in xylene (two changes of 2 min each) and air-dried for 1 h. Approximately 45 stomium regions (400–500 cells) were captured from unstained anther sections using a Leica AS LMD Microdissection System (Leica Microsystems).

#### Real-time quantitative reverse transcription-polymerase chain reaction

LCM-captured stomium RNA was isolated using a PicoPure RNA Isolation Kit (Arcturus, Mountain View, Calif.), treated with RNase-free DNase I (Ambion, Austin, Tex.), purified using RNeasy Plant Mini Kit (Qiagen, Valencia, Calif.), and eluted into 15  $\mu$ l RNase-free water. Complementary DNA (cDNA) was synthesized in a 20  $\mu$ l reaction with an iScript cDNA Synthesis Kit (Bio-Rad, Hercules, Calif.), using all of the stomium RNA as a template. One-fortieth of the cDNA volume (0.5  $\mu$ l) was amplified by quantitative polymerase chain reaction (qPCR) in a 25  $\mu$ l reaction volume using iQ SYBR Green Supermix and an iCycler iQ Real-Time PCR Detection System (Bio-Rad). The following primers were used: TA56 Fw 5'-gctttggtacttaggcttggtgagagt-3'; TA56 Rv 5'-cttgctcttgacaggagtaacagcac-3'; TA20 Fw 5'-ctgccatgaaattgaatctacaatg-3' TA20 Rv 5'-cgaagtaagtagaaaggatggagggtg-3' (Koltunow et al. 1990; Beals and Goldberg 1997).

---

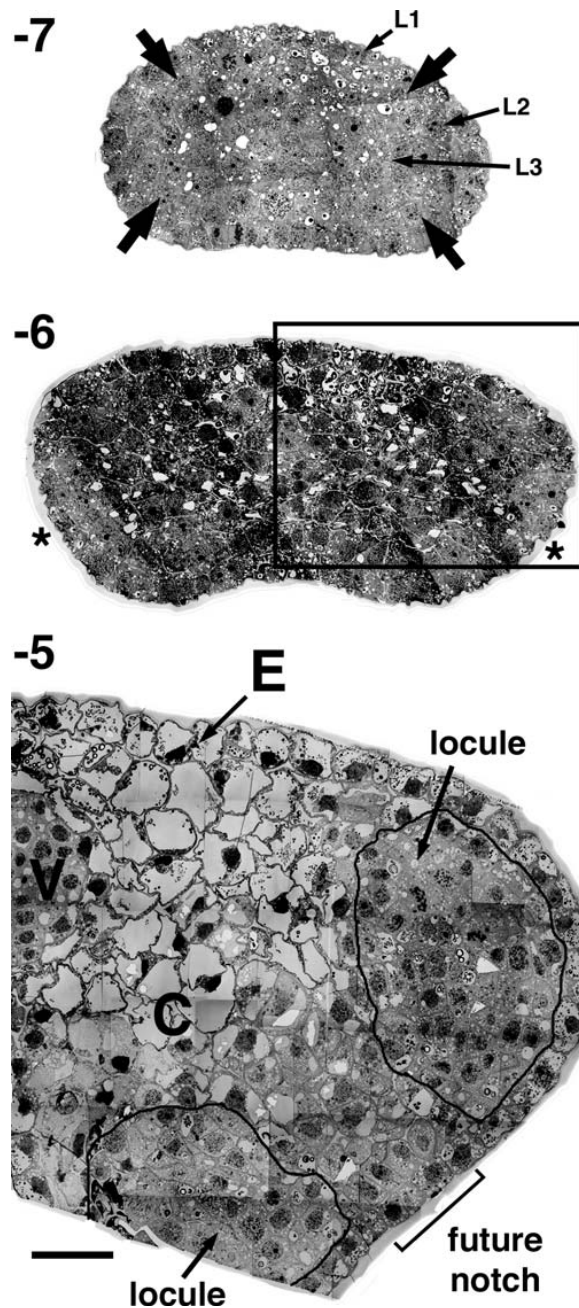
## Results

Locule differentiation occurs prior to the visible appearance of the circular cell cluster and stomium in the notch region

We studied the development of cells within the anther notch using TEM. We focused on circular cell cluster and stomium development to characterize their individual roles in establishing the site of wall breakage in anther dehiscence. Our studies were carried out at a low TEM magnification (1,900 $\times$ ), which afforded a greater resolution of the cells and cellular events than could be observed in either Paraplast or plastic sections with the light microscope. Using TEM, we characterized events in the anther notch from stages  $-7$  to  $+5$  and  $+9$  to  $+12$  of anther development (Fig. 1; Tables 1, 2). Our goal was to visualize events that (1) generate the differentiated circular cell cluster and stomium, and (2) create the site for dehiscence along the anther wall.

Figure 2 shows the initial changes that occurred within the stamen primordium to generate a four-locule anther. In transverse section, the primordium was a uniform collection of cells (Fig. 2, stage  $-7$ ). By stage  $-6$ , cell divisions changed the round primordium into an

**Fig. 2** Establishment of anther shape from the stamen primordia. Anther sections from stages -7, -6, and -5 were fixed, embedded in Spurr's epoxy resin, sliced into ultra-thin sections and prepared for transmission electron microscopy (TEM) as described in Materials and methods. TEM micrographs were taken at a magnification of 1,900 $\times$ . The images of stages -7 and -6 are complete transverse sections of stamen primordia. The boxed area in stage -6 represents the region of the developing anther shown for stage -5. Stage -7, complete transverse section of a stamen primordium; *arrows* sites of archesporial cell differentiation, whose cell lineages will generate the four reproductive locules. Stage -6, complete transverse section of a stamen primordia; *asterisks* site of future anther notch regions, one of which is shown in the stage -5 partial transverse section. Stage -5, partial transverse section of a developing anther; *cells outlined in black* regions of developing locules. The designation of cells contained within the black border is based on their histology. These cells show mitotic activity and have not yet expanded or acquired large vacuoles. In a stage -5 anther, the cell-types contained within the L2-derived anther locules are the endothecium, middle layers, tapetum and sporogenous cells. The future notch is indicated between the two developing anther locules. *C* Connective; *E* epidermis; *Locule* region derived from the L2-archesporial lineage that will create the anther locule; *L1*, *L2*, *L3* the three cell layers present within the stamen primordia; *V* vascular region. Bar 30  $\mu$ m



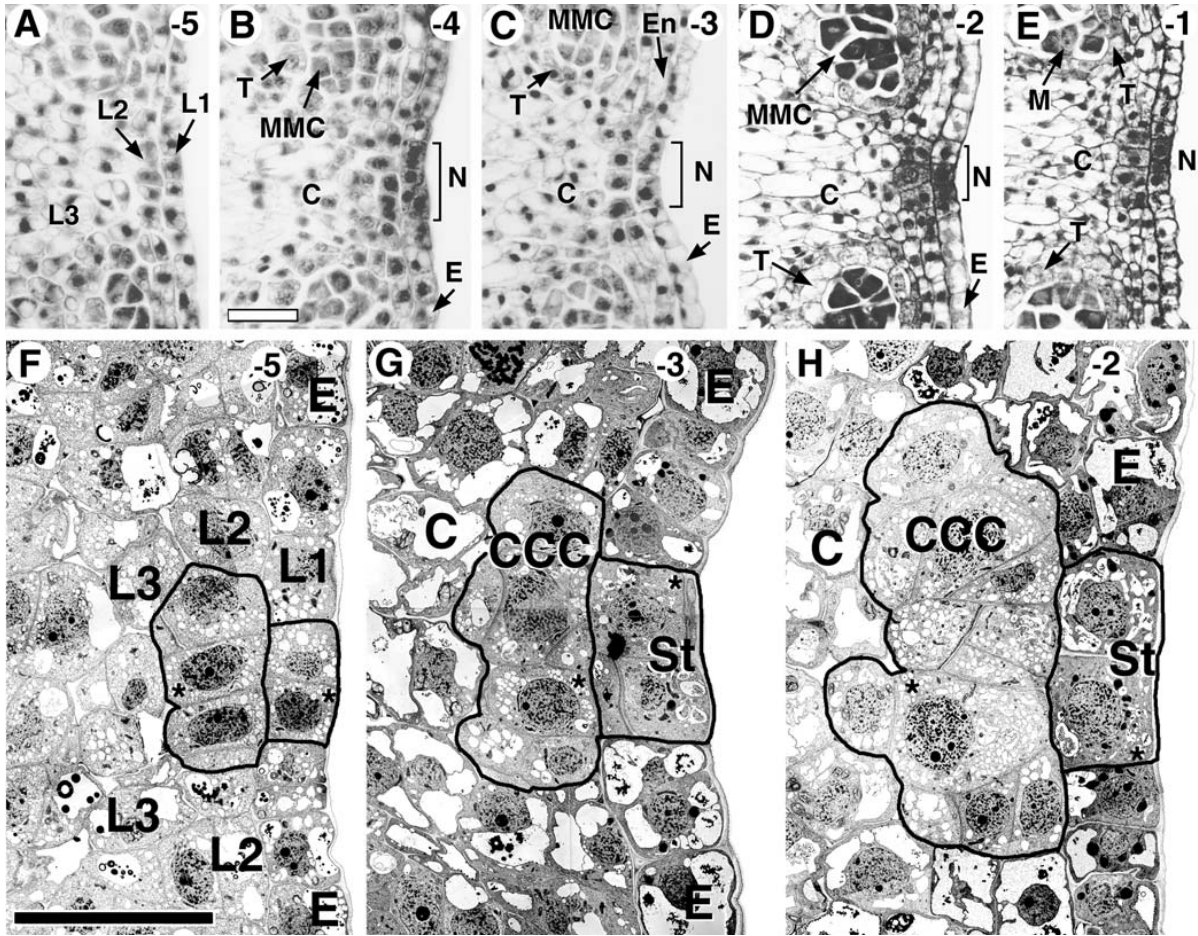
elongated structure (Fig. 2). At stages -7 and -6, the cells of the anther were relatively small in size and highly cytoplasmic with prominent nuclei. The stereotyped four-locule pattern of the anther began to emerge after stage -6 (Fig. 2, stage -6). At stage -5, the differentiation events within the future locule territories were apparent (Fig. 2, stage -5). Mitotic activity was observed in the four corners of the developing anther (Fig. 2, stage -5). In addition, cells of the developing connective and central region of the anther had acquired large vacuoles and expanded in size. Cells destined to become the vascular bundle could be identified and were round in shape and cytoplasmically dense with prominent nuclei (Fig. 2, stage -5). In contrast, we did not observe any distinctive cell types or mitotic activity in the region between the developing microsporangia (Fig. 2, stages -7 to -5). Epidermal and sub-epidermal cells in this region were indistinguishable from their adjacent neighbors (Fig. 2, stages -7 to -5), i.e., cells destined to become the circular cell cluster and stomium were not yet apparent within L1, L2, or L3 cell layers of the pre-notch region. Together, these results indicate that the initial developmental events in the anther primordium, including the differentiation of cells that generate the locules, occur prior to any visible cellular events within the site of the future anther notch (Figs. 1; 2; Table 1).

The circular cell cluster is derived from cells of the primordium L2 layer

We examined both bright-field photographs and TEM micrographs of sections throughout phase 1 of anther development to determine when the circular cell cluster and stomium cells became specified within the notch

region (Table 1). Figure 3 shows the developing notch region from stages -5 to -1. We aligned these photographs so that (1) cells within the developing notch are in the center, and (2) the same area of each anther is shown in both the bright-field photographs (Fig. 3A-E) and TEM micrographs (Fig. 3F-H). Light microscopy allowed us to obtain an overview of notch-region development in parallel with other developmental events that occurred within the anther (e.g., locule formation).





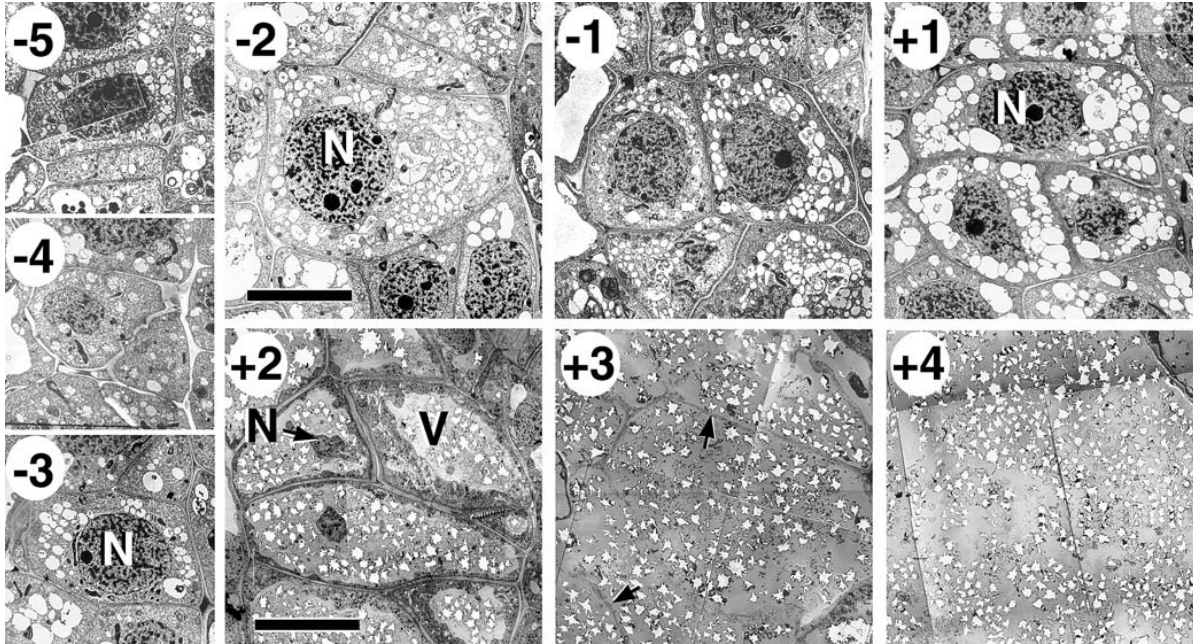
**Fig. 3A–H** Development of the anther notch at stages  $-5$  to  $-1$ . Anthers were fixed and embedded in both paraffin and Spurr's epoxy resin as described in Materials and methods. Paraffin-embedded anthers were sliced into  $10\ \mu\text{m}$  transverse sections, stained with toluidine blue and the notch regions were photographed by bright-field microscopy. Spurr's embedded anthers were sliced into ultra-thin sections and prepared for TEM. Anther developmental stages are shown in the top right corner of each photograph. The stomium and the circular cell cluster span the length of the anther and are a long column of cells of which these photographs represent a single layer. In TEM micrographs, the cells of the stomium and the circular cell cluster are highlighted by a *black border*. **A–E** Bright-field photographs of the developing anther notch. **A** Stage  $-5$ , **B** stage  $-4$ , **C** stage  $-3$ , **D** stage  $-2$ , **E** stage  $-1$ . **F–H** TEM micrographs of the developing anther notch. **F** Stage  $-5$ , **G** stage  $-3$ , **H** stage  $-2$ . The stage  $-5$  notch is a consecutive section from the anther shown in Fig. 2. The circular cell cluster and stomium cells marked with an *asterisk* are shown in greater detail in Figs. 4 and 7. **C** Connective; **CCC** circular cell cluster; **E** epidermis; **En** endothecium; **L1**, **L2**, **L3** the three cell layers present within the stamen primordia; **M** meiocyte; **MMC** microspore mother cells; **N** notch; **St** stomium; **T** tapetum. **Bars** **B**  $20\ \mu\text{m}$  (also the scale for **A**, **C**, **D**, **E**), **F**  $30\ \mu\text{m}$  (scale for **G**, **H**)

In contrast, TEM enabled us to obtain a detailed characterization of the developing notch region at the cellular level. We were able to determine which cells within

the notch became the circular cell cluster and stomium by visualizing events within specific cell layers at later stages (e.g., stages  $-3$  to  $-2$ ) and then tracing the cell lineages backwards to earlier stages (e.g., stage  $-5$ ).

At the level of the light microscope, circular cell cluster and stomium cells within the pre-notch region could not be distinguished at stages  $-5$  to  $-3$  (Fig. 3A–C). In contrast, cells within developing locule territories (e.g., microspore mother cells, tapetum) could be identified by their shape, differential staining, and mitotic activity (Fig. 3A–C). Epidermal and connective cells were also identified at this time of development (Fig. 3A–C). By stages  $-2$  and  $-1$ , L1 and L2 cells within the interlocular region could be distinguished from their neighbors by differential staining (Fig. 3D–E; data not shown), suggesting that notch-region cells destined to become the circular cell cluster and stomium have been specified by this stage of anther development.

At the resolution of the TEM, changes within cells destined to become the circular cell cluster were detected earlier than with the light microscope. At stage  $-5$ , cells within the future notch region were relatively uniform and could be distinguished only by their location within



**Fig. 4** Cells of the circular cell cluster during tobacco anther development. Stage  $-5$  to  $+4$  anther sections were fixed, embedded in Spurr's epoxy resin, sliced into ultra-thin sections and prepared for TEM as described in Materials and methods. The images presented here are close-ups of the circular cell cluster cells marked with an *asterisk* in Figs. 3 and 6. The nuclei at stage  $+2$  are distorted and shrunken. The white star-like shapes at stages  $+2$  to  $+4$  represent calcium oxalate druse crystals. The *arrows* in stage  $+3$  indicate cell wall remnants of the degenerating circular cell cluster. *N* Nucleus, *V* vacuole. *Bars* Stage  $-2$   $10\ \mu\text{m}$  (scale for stages  $-5$  to  $+1$ ), stage  $+2$   $20\ \mu\text{m}$  (scale for stages  $+2$  to  $+4$ )

the primordium L1, L2, and L3 layers (Figs. 2, stage  $-5$ ; 3F). The L1 and L2 cells within the future notch region were relatively uniform in appearance, cytoplasmically rich, and contained small vesicles and/or vacuoles (Fig. 3F). In contrast, L2 cells outside of the pre-notch region in the corners of the primordium had divided to generate the developing locules (Figs. 2, stage  $-5$ ; 3A). In addition, L3 cells flanking the future notch region were more vacuolate in appearance than neighboring L2 and L1 cells (Figs. 2, stage  $-5$ ; 3F). This also was the case for epidermal cells contiguous to the L1 cells of the future notch site (Figs. 2, stage  $-5$ ; 3F). Within the notch, pre-circular cell cluster and stomium cells were visualized as a single layer of L2 and L1 cells, respectively (Fig. 3F).

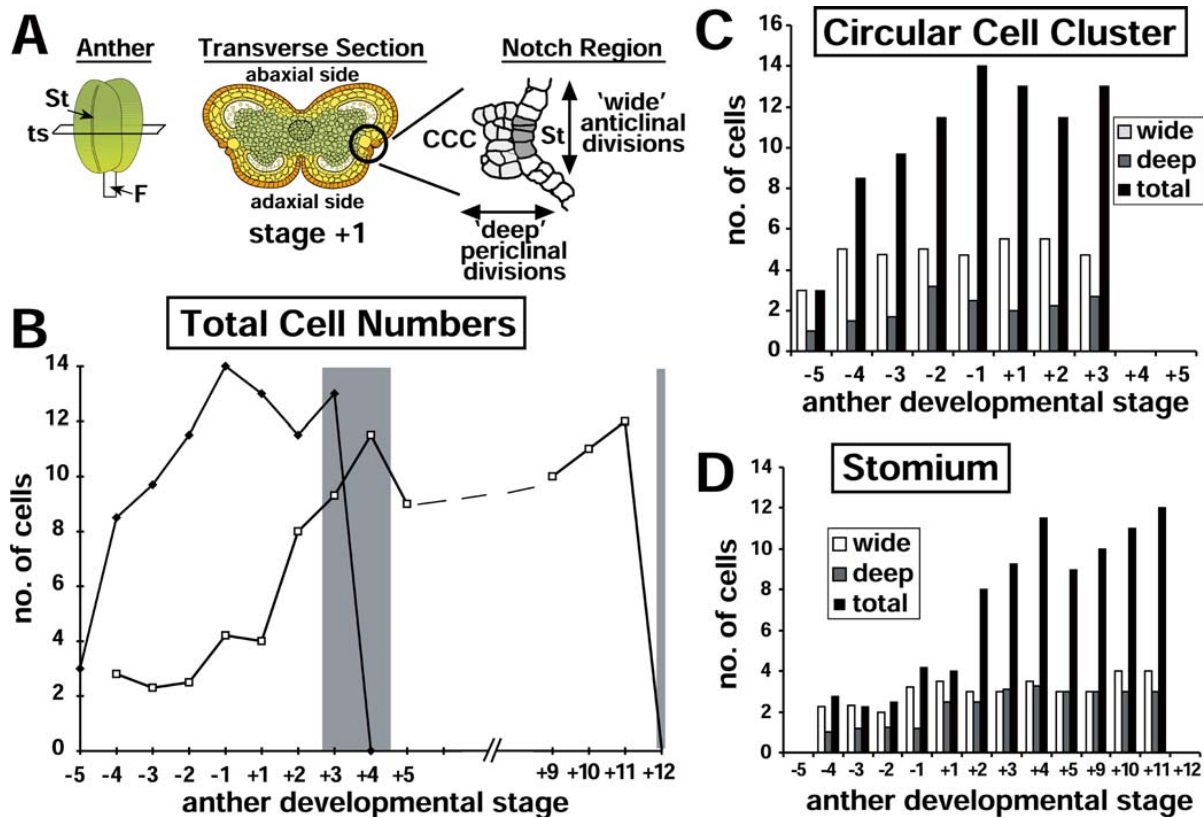
By stages  $-3$  and  $-2$ , the circular cell cluster was identified as a group of distinctively-shaped L2-derived cells between the L1 epidermis and the L3-derived connective layers (Fig. 3G, H). Cells within the L2 layer destined to become the circular cell cluster divided, expanded, and accumulated numerous small vesicles (Fig. 3F–H). In contrast, L1 epidermal cells that bordered the developing circular cell cluster (i.e., pre-stomium cells) did not divide and remained cytoplasmically

dense (Fig. 3F–H). Cells surrounding the pre-notch region accumulated large vacuoles and were visibly distinct compared with pre-circular cell cluster and pre-stomium cells (Figs. 2, stage  $-5$ ; 3G, H). By stage  $-2$ , the TEM micrographs showed that the circular cell cluster formed a group of histologically distinct cells within the developing anther notch (Fig. 3H)—the same stage as when differential staining was observed within this area in the Paraplast sections (Fig. 3D).

Figure 4 shows TEM close-ups of individual circular cell cluster cells at different stages of anther development. During stages  $-5$  to  $-2$ , cells of the circular cell cluster increased significantly in size and accumulated numerous small vesicles (Fig. 4; Table 1). Figure 5A shows that a large increase in the number of cells contained within the circular cell cluster occurred during stages  $-5$  to  $-1$ . The number of cells within the circular cell cluster reached a maximum of  $\sim 14$  by stage  $-2$  in a typical transverse  $10\ \mu\text{m}$  section (Fig. 5A–C). This increase was due primarily to periclinal cell divisions that increased the number of cell layers (Figs. 3; 5A–C). Together, these results show that the circular cell cluster is specified from L2 cells within the pre-notch region and that circular cell differentiation events can be visualized as early as stage  $-3$ , prior to any changes in contiguous pre-stomium cells in the L1 layer.

Circular cell cluster differentiation and degeneration occur before analogous stomium events take place

We studied the developing notch region from stages  $+1$  to  $+4$  in both the light microscope and TEM to characterize developmental changes within the circular cell

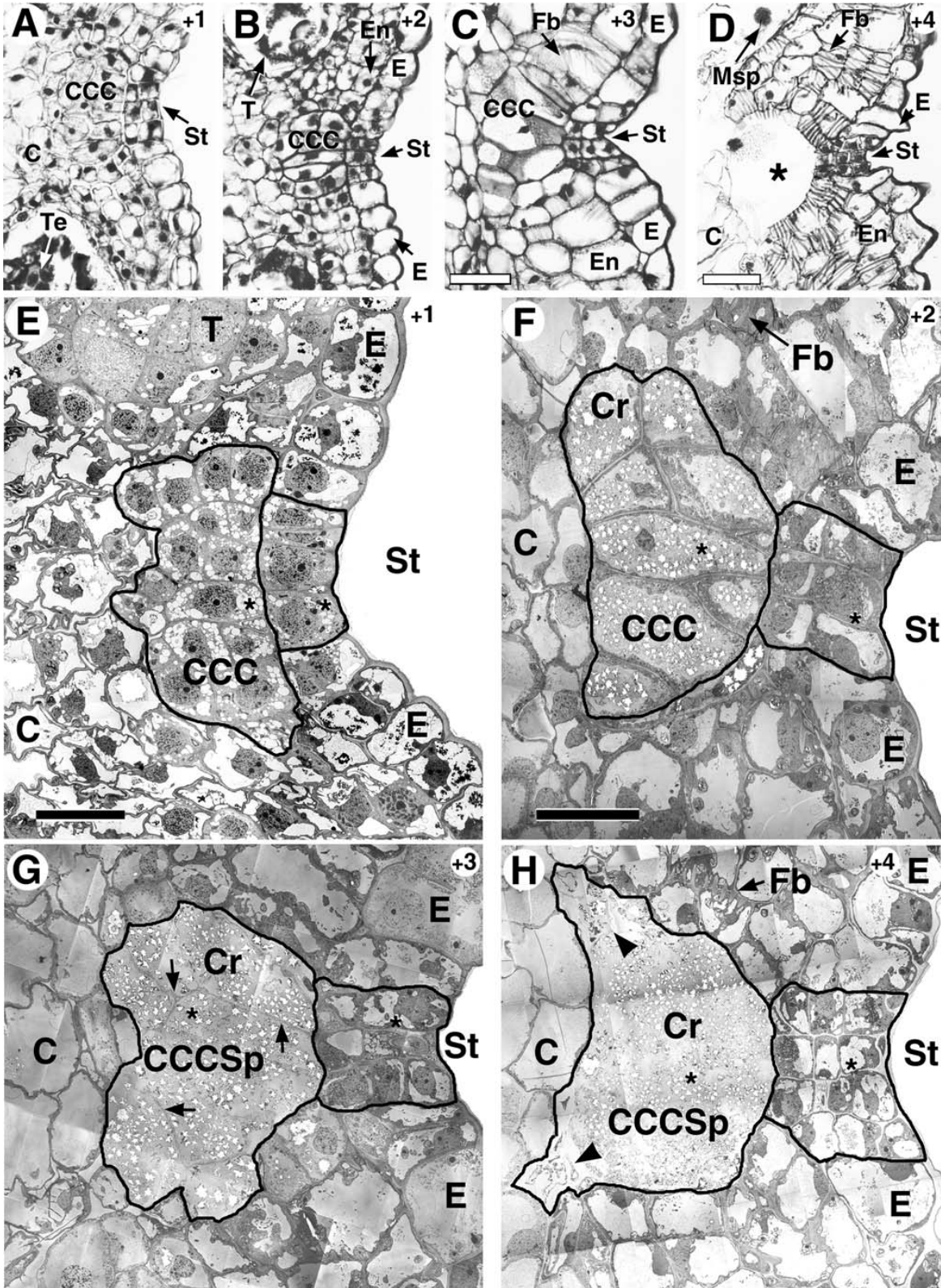


**Fig. 5A–D** Number of cells in the circular cell cluster and stomium during anther development. The number of cells in the circular cell cluster and stomium during anther development were counted from transverse sections visualized in the TEM (Figs. 3, 6, 9; data not shown). The total number of circular cell cluster and stomium cells in the entire anther is larger. Anther developmental stages are from Koltunow et al. (1990); phase 1, stages  $-7$  to  $-1$  and phase 2, stages  $+1$  to  $+12$ . **A** Cartoon representation of sectioning used to view notch region cell types. Transverse sections through the anther generate a butterfly shape. TEM analysis focused on the notch region of each transverse section. Periclinal cell divisions generated an increase in the width (*deep*) of the circular cell cluster and stomium as observed in transverse section. Anticlinal cell divisions generated an increase across (*wide*; adaxial to abaxial) the circular cell cluster and stomium as visualized in transverse section. **B** Total number of cells in the circular cell cluster and stomium at different stages of anther development. *Closed diamonds* Circular cell cluster, *open squares* stomium. Cell numbers were not counted between stages  $+6$  to  $+8$  (*dashed line*). *Shaded areas* Periods of cell-death: stages  $+3$  to  $+4$  for the circular cell cluster and stage  $+12$  for the stomium. Cells in one to four sections were counted depending upon the stage. **C** Number of circular cell cluster cells in transverse sections at different stages of anther development. **D** Number of stomium cells in transverse sections at different stages of anther development. *Bars* Average number of cells in counted transverse sections

cluster and to identify when stomium differentiation began (Figs. 4; 6; Table 1). The bright-field photographs presented in Figure 6A, B showed that by stages  $+1$  and  $+2$  the circular cell cluster had formed into the circular arrangement of sub-epidermal cells for which it was

named (Koltunow et al. 1990); i.e., circular cell cluster differentiation had occurred. From stages  $+2$  to  $+4$  major changes were observed in the notch region at the level of the light microscope (Fig. 6B–D; Table 1). Circular cell cluster degeneration took place, the stomium differentiated from contiguous L1 layer cells, fibrous bands were deposited in the endothecium and middle layer, and the epidermal cells neighboring the developing stomium expanded and contained a single, large vacuole (Fig. 6B–D; Table 1; data not shown).

TEM micrographs of the developing notch region from stages  $+1$  to  $+4$  are shown in Figure 6E–H, and close-ups of individual circular cell cluster cells at these stages are shown in Figure 4. At stage  $+1$ , the cells of the circular cell cluster had a distinctive round shape and lacked the large vacuole seen in surrounding cells (Fig. 6A, E). The circular cell cluster was two to three cells deep (or across) in a typical section due to periclinal cell divisions (Figs. 5A; 6E). The small vesicles that accumulated during stages  $-5$  to  $-2$  (Figs. 3F–H; 4) had begun to aggregate (Figs. 4; 6E). By stage  $+2$ , the cells of the circular cell cluster underwent further expansion and contained a single large vacuole (Figs. 4; 6B,F). The circular cell cluster size increase was caused by cell enlargement and not by cell division, as the number of cells did not increase after stage  $-1$  (Fig. 5B,C). The vacuole present in cells of the stage  $+2$  circular cell cluster was formed by aggregation of small vesicles that







**Fig. 6A–H** Circular cell cluster degeneration and stomium differentiation. Anther sections from stages +1 to +4 were fixed and embedded in either Paraplast or Spurr's epoxy resin as described in Materials and methods. Paraffin-embedded anthers were sliced into 10  $\mu\text{m}$  transverse sections, stained with toluidine blue, and notch regions were photographed by bright-field microscopy. Spurr's embedded anthers were sliced into ultra-thin sections and prepared for TEM. Anther developmental stages are shown in the top right corner of each photograph. In the TEM micrographs, the cells of the stomium and circular cell cluster are highlighted by a *black border*. **A–D** Bright-field photographs of the developing anther notch region. **A** Stage +1, **B** stage +2, **C** stage +3, **D** stage +4. In **D** the *asterisk* represents the space previously occupied by the circular cell cluster. **E–H** TEM micrographs of the developing anther notch region. **E** Stage +1, **F** stage +2, **G** stage +3, **H** stage +4. *Black borders* (**E–H**) surround the circular cell cluster and stomium. In **G**, *arrows* indicate cell wall remnants from circular cell cluster degeneration and in **H**, *arrowheads* indicate initial stages of connective cell degeneration. The *black border* in **H** surrounding the space left by the degenerated circular cell cluster includes neighboring connective cells that are degenerating (*arrowheads*). The circular cell cluster and stomium cells marked with an *asterisk* are shown in greater detail in Figs. 4 and 7. *C* Connective, *CCC* circular cell cluster, *CCCSp* space generated by degeneration of circular cell cluster cells, *Cr* calcium oxalate crystals, *E* epidermis, *En* endothecium, *Fb* fibrous bands, *Msp* microspore, *St* stomium, *T* tapetum, *Te* tetrad. *Bars* **C** 20  $\mu\text{m}$  (scale for **A**, **B**), **D** 25  $\mu\text{m}$ , **E** 30  $\mu\text{m}$  (scale for **G**, **H**), **F** 30  $\mu\text{m}$

had accumulated earlier (Figs. 3; 4; 6E,F). The cytoplasm within these cells was compressed against the cell wall and the nucleus appeared shrunken and distorted (Figs. 4; 6F). In addition, calcium oxalate druse crystals (Horner and Wagner 1980; D'Arcy et al. 1996) were present within the large vacuoles of the stage +2 circular cell cluster and were visualized as “white-star” shapes in TEM micrographs (Figs. 4; 6F). The “white-stars” represented holes left behind after sectioning, as the crystals were not sectioned. The presence of calcium oxalate crystals contributed to the speckled appearance of the circular cell cluster cells in the light microscope (Fig. 6C).

The creation of a single large vacuole, the accumulation of calcium oxalate crystals, and the shrinking of the nucleus in cells of the circular cell cluster at stage +2 marked the beginning of the circular cell cluster degeneration program (Fig. 6C,D,F,H). By stage +3, all that remained of the circular cell cluster was vacuole contents and cell wall remnants (Figs. 4; 6G). The druse crystals and the vacuole contents in which they were embedded remained in the space occupied previously by the circular cell cluster (Figs. 4; 6G–H). By stage +4, the wall remnants were completely degenerated and cells of the circular cell cluster were no longer present, leaving a large circular hole in the anther (Figs. 4; 6D,H; Table 1).

Connective cells contiguous to the circular cell cluster and the tapetum also began to degenerate during this period (stages +2 to +4) resulting in a bilocular anther (Figs. 1; 6D,F,H; Table 1). In addition, the notch region became more pronounced as a consequence of locule territory enlargement, the endothecium expanded, and lignified fibrous bands (Manning 1996) were deposited within both connective and endothecial cells (Fig. 6C,D,F,H; Table 1). Together, these results show

that circular cell cluster differentiation occurred between stages –3 and +1 of anther development and that cells of the circular cell cluster underwent a programmed cell death process and degenerated by stage +4 (Fig. 6D,H; Table 1).

Stomium cells are specified early in the development of the notch region

When do the L1 cells within the anther primordium give rise to the stomium and how do events that occur within the stomium compare with those observed in the circular cell cluster during anther development? At stage –5, L1-layer primordium cells in the future notch region were indistinguishable from each other in both the light microscope (Fig. 3A) and the TEM (Figs. 2 stage –5; 3F). These included the L1 cells immediately adjacent to L2 cells that gave rise to the circular cell cluster and their L1 neighbors (Fig. 3A,F). At stage –3, in contrast, L1 cells destined to become the stomium were cytoplasmically dense and did not contain large vacuoles like their neighboring L1 epidermal cells (Fig. 3C,G). Typically, two or three pre-stomium L1 cells were observed in TEM transverse sections at stages –3 and –2, and these cells were in contact with the differentiating circular cell cluster (Figs. 3G,H; 5D). At the light microscope level, L1-layer pre-stomium cells stained more intensely with toluidine blue than did their epidermal neighbors at stages –2 and –1, similar to those in the developing circular cell cluster (Fig. 3D,E; data not shown). Together, these data show that a small number of L1-layer primordium cells become specified as early as stage –3 to follow a stomium differentiation pathway, and that these cells can be distinguished from neighboring epidermal cells by their staining properties, dense cytoplasm, absence of prominent vacuoles, and contact with the differentiating circular cell cluster.

Development of a multi-layered stomium occurs after the circular cell cluster degenerates

The pre-stomium cells within the L1 layer did not undergo any detectable changes during stages –3 to –1 of anther development at the level of either the light or electron microscope (Figs. 3; 7). Close-ups of individual pre-stomium cells indicate that they had a high nuclear to cytoplasmic volume and only small vesicles or vacuoles (Fig. 7). Nor did they increase significantly in number during this period, in contrast with cells of the circular cell cluster (Fig. 5B,D).

From stages +1 to +4, however, the pre-stomium cells underwent several divisions, including two periclinal divisions and an occasional anticlinal division (Figs. 5B,D; 6E–H). At stage +2, one periclinal division generated a developing stomium that was three cells across and two cells deep in a typical section—approximately the width of a single epidermal cell (Figs. 5D;

6F). At stage +3, a second periclinal division generated a stomium that was three cells across and three cells deep (Figs. 5D; 6G). In some sections, we observed four rows of stomium cells, suggesting that an anticlinal cell division had occurred (Fig. 6E). Thus, by stage +4 the mature stomium was a group of approximately 9–12 small cells that were positioned within the epidermal layer of the anther (Figs. 5B,D; 6 C,D,G,H; Table 1). The multi-tiered stomium was clearly distinguished from contiguous epidermal cells that were larger and more highly vacuolate (Fig. 6 C,D,G,H). Individual stomium cells remained histologically similar during stages +3 to +5 (Figs. 6G,H; 7), but expanded and acquired prominent vacuoles, structures not apparent in earlier stages of stomium cell development (Fig. 7, stages –5 to –1). Stomium cell division events occurred after the circular cell cluster stopped dividing (Figs. 5B; 6; Table 1). By the time a multi-tiered stomium formed, the circular cell cluster had degenerated (Figs. 5B; 6C,D,G,H; Table 1). Together, development of the stomium and the death of the circular cell cluster by stage +4 established the future site for anther wall breakage at dehiscence.

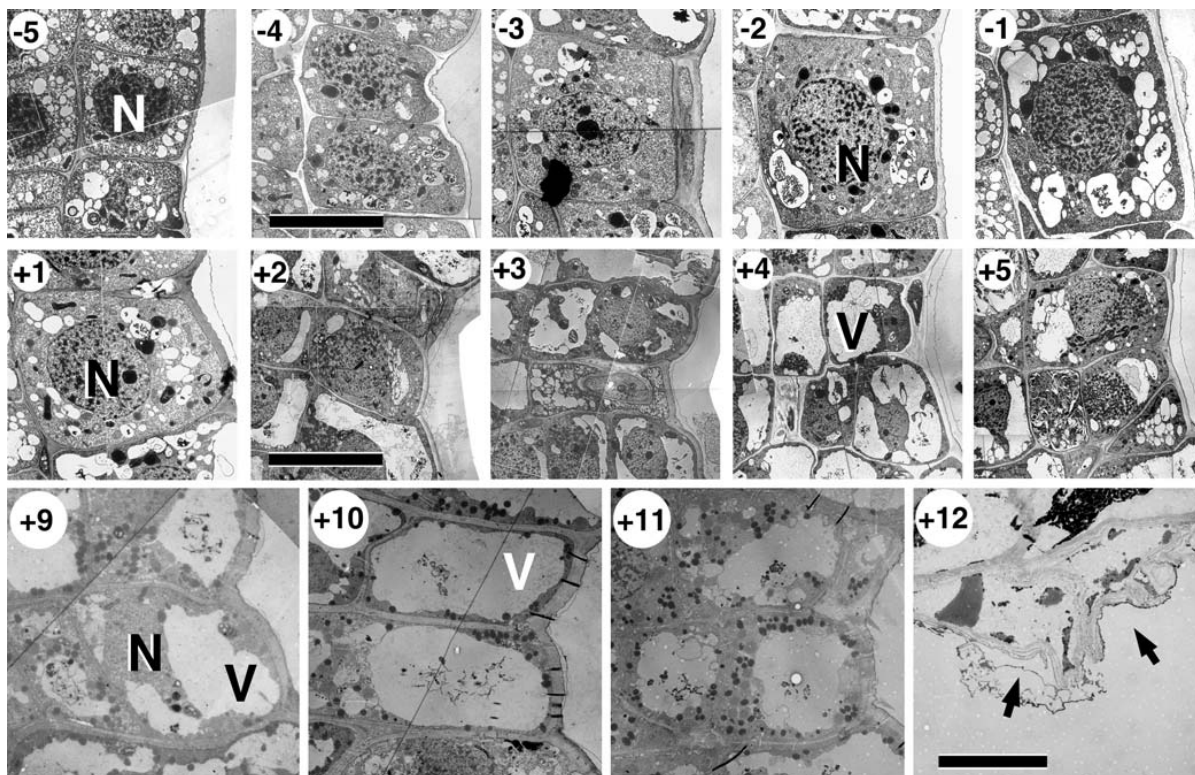
Plasmodesmata connections occur between cells within the notch region

We examined the boundaries between cells of the developing notch region to determine whether there

were plasmodesmata connections, and, if so, what cells were connected by these cytoplasmic channels. The presence of plasmodesmata might provide clues to possible interactions between notch-region cell types (Haywood et al. 2002). We focused our efforts on stages –4 to +2, when both the circular cell cluster and stomium became specified and differentiated (Figs. 3; 6). To identify plasmodesmata, we used a higher TEM magnification (~29,000×) than that used to visualize notch region development (1,900×; Figs. 2; 3; 4; 6; 7).

Plasmodesmata were observed within the notch region at all developmental stages examined (Fig. 8). Connections occurred between similar cell types (e.g., stomium-stomium; Fig. 8, st-st) and different cell types (e.g., stomium-circular cell cluster; Fig. 8, st-ccc). Both

**Fig. 7** Stomium cells during tobacco anther development. Anther sections from stages –5 to +5 and stages +9 to +12 were fixed, embedded in Spurr's epoxy resin, sliced into ultra-thin sections and prepared for the TEM as described in Materials and methods. The images presented here are close-ups of stomium cells marked with an *asterisk* in Figs. 3, 6, 9, and 10A. Stage +10 is from a consecutive transverse section from the same anther shown in Fig. 9. At stages +9 to +11, small dark lipid vesicles are seen to accumulate. The *black arrows* in stage +12 indicate the degenerated stomium cells after dehiscence and anther opening. *N* Nucleus, *V* vacuole. *Bars* Stage –4 10 μm (scale for stages –5 to +1), stage +2 20 μm (scale for stages +2 to +5), stage +12 10 μm (scale for stages +9 to +12)



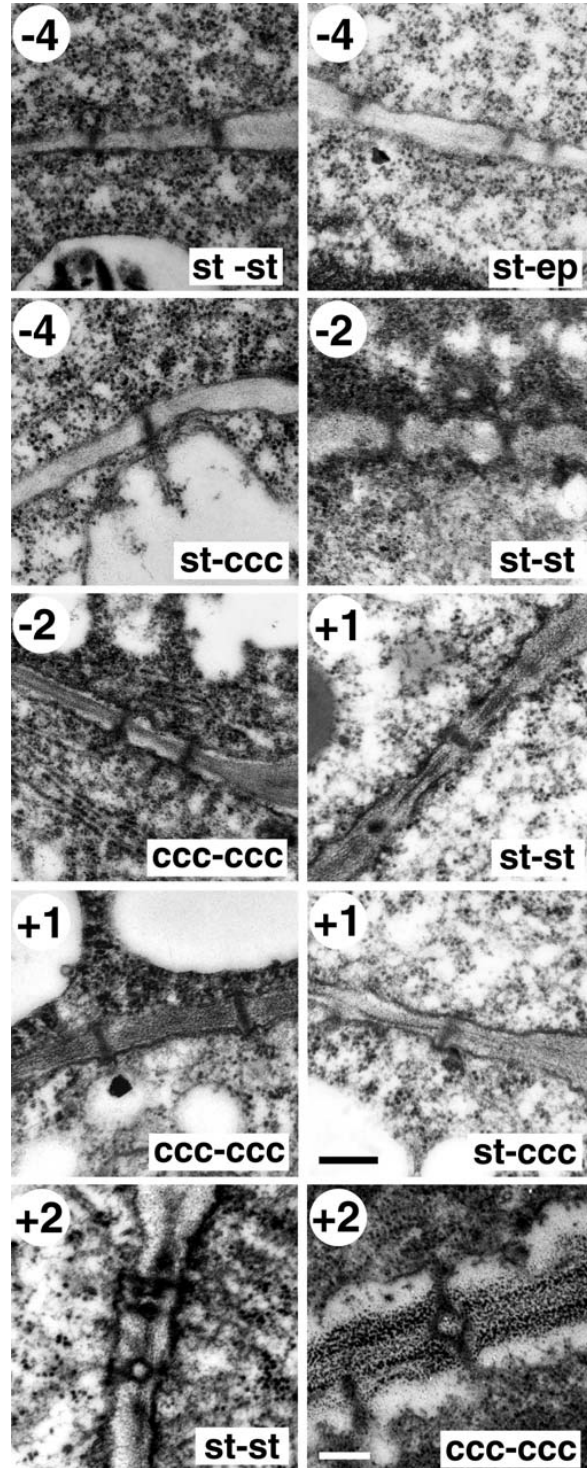
**Fig. 8** Plasmodesmata in the anther notch region during anther development. Anther sections from stages  $-4$ ,  $-2$ ,  $+1$ , and  $+2$  were fixed, embedded in Spurr's epoxy resin, sliced into ultra-thin sections and prepared for the TEM as described in Materials and methods. Primary plasmodesmata identified between the cells of the developing notch are shown for stages  $-4$ ,  $-2$  and  $+1$ . Secondary plasmodesmata are shown for stage  $+2$ . The labels refer to the borders between two cell-types: *st-st* between two stomium cells, *st-ep* between stomium and epidermal cells, *st-ccc* between stomium and circular cell cluster cells, *ccc-ccc* between two circular cell cluster cells. *Bars* Stage  $+1$  200 nm (scale for stages  $-4$  and  $-2$ ), stage  $+2$  200 nm

primary plasmodesmata (Fig. 8, stages  $-4$  to  $+1$ ) and secondary plasmodesmata (Fig. 8, stage  $+2$ ) were detected. No differences were observed between the types of cells that were connected by plasmodesmata at different stages (data not shown). Nor were there detectable differences in the number of plasmodesmata connections observed between different cell types (data not shown). The only apparent change in plasmodesmata was visualized at the time of circular cell cluster degeneration—plasmodesmata connections between the circular cell cluster and other cell types disappeared (data not shown). Together, these results indicate that the entire notch region is connected by cytoplasmic channels and that plasmodesmata occur between the circular cell cluster and stomium when these cells become specified (Fig. 8, stage  $-4$ ).

Stomium cell death establishes the site for pollen release at dehiscence

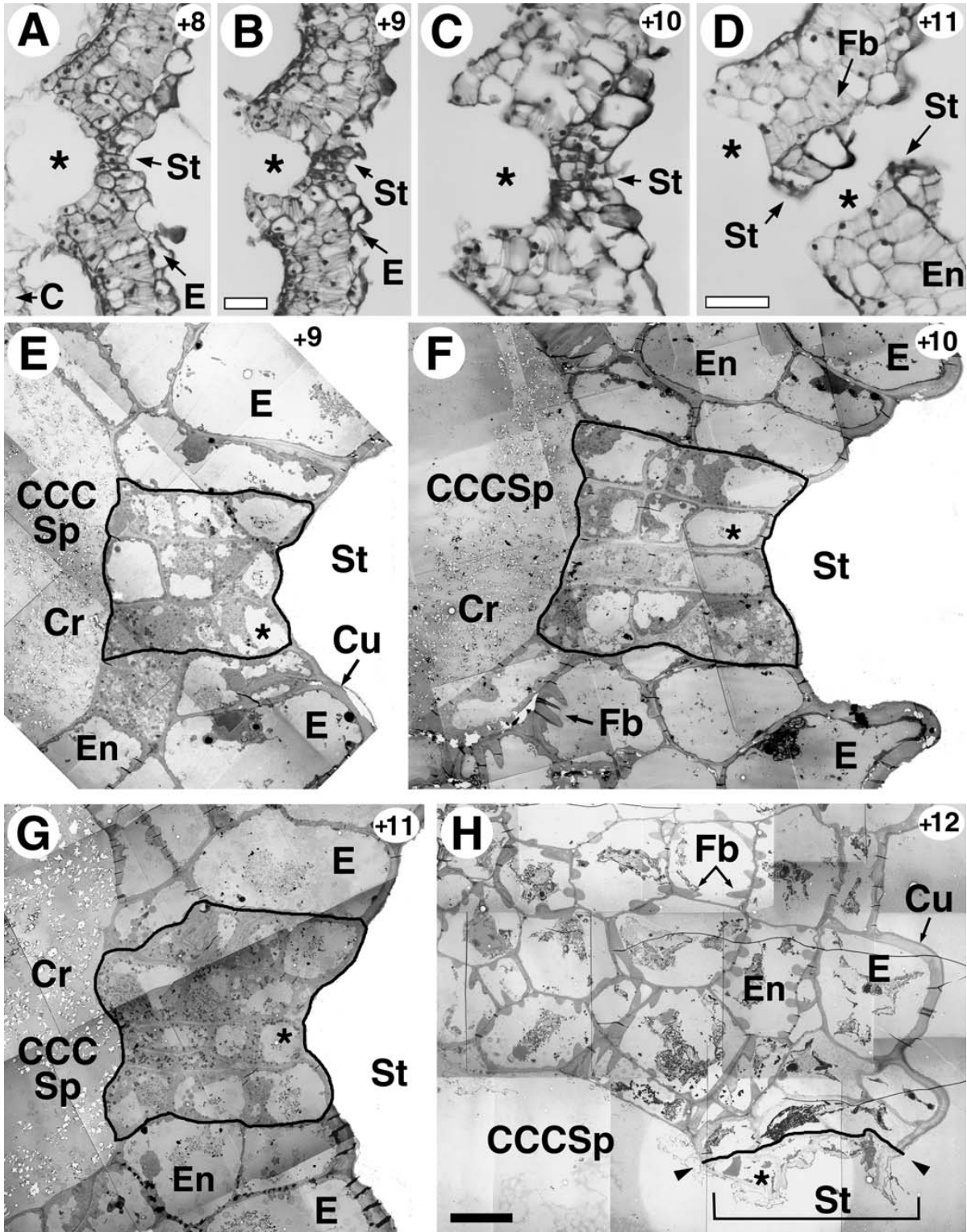
We examined the notch region at the terminal stages of anther development (stages  $+8$  to  $+12$ ; Table 1) in both the light microscope (Fig. 9A–D) and TEM (Figs. 9E–G; 10) to visualize cellular events that occurred within the stomium in the period prior to dehiscence (stages  $+8$  to  $+11$ ) and during dehiscence (stage  $+12$ ). From stages  $+8$  to  $+11$  the stomium consisted of a multi-tiered set of cells flanked by calcium-oxalate-filled space left over from the degenerated circular cell cluster (Fig. 9A–G), similar to that observed earlier in anther development (Fig. 6H, stage  $+4$ ). The stomium did not increase in cell number after stage  $+4$  (Fig. 5B,D), and remained the narrowest site within the anther wall (Fig. 9A–C). Stomium cells were markedly smaller in size and contained prominent nuclei in comparison with neighboring epidermal and endothelial cells, which expanded significantly during this period and contained large vacuoles (Fig. 9A–G). In addition, fibrous bands became more numerous in both the endothecium and connective (Fig. 9A–D,F,H). The only visible change in the stomium cells from stages  $+9$  to  $+11$  was an increase in the number of lipid vesicles and an enlargement of vacuole size due to coalescence of smaller vacuoles (Figs. 7; 9E–G). During this period the stomium consisted of intact cells, i.e., stomium cell degeneration had not yet occurred.

We compared a mechanically-sheared stage  $+11$  notch region (Fig. 10B) with one that dehisced at stage



$+12$  (Figs. 7; 9H; 10A) to determine if stomium cell death played a role in the dehiscence process. Fixed, stage  $+11$  anthers often break at the stomium when









**Fig. 9A–H** Late events in stomium development. Stage +8 to +12 anthers were fixed and embedded in either Paraplast or Spurr's epoxy resin as described in Materials and methods. Paraffin-embedded anthers were sliced into 10  $\mu\text{m}$  transverse sections, stained with toluidine blue, and the notch regions were photographed by bright-field microscopy. Spurr's embedded anthers were sliced into ultra-thin sections and prepared for TEM. **A–D** Bright-field photographs of the developing anther notch. **A** stage +8, **B** stage +9, **C** stage +10, **D** stage +11. The *asterisks* in **A–D** represent the space previously occupied by circular cell cluster cells (see Fig. 6). The break at the stomium in the stage +11 anther in **D** was caused by mechanical shearing during sectioning, but demonstrates the biological process seen at anther dehiscence, stage +12. The space occupied previously by the circular cell cluster in **D** divided into two regions during sectioning and is marked by the *asterisks*. **E–H** TEM micrographs of the developing anther notch. **E** Stage +9, **F** stage +10, **G** stage +11, **H** stage +12. In **E–H** the stomium cells marked with an *asterisk* are shown in greater detail in Fig. 7. The *black border* in **E–G** outlines the stomium cells. The *black border* and *arrowheads* in **H** delineate the stomium from neighboring cells in the Stage +12 dehiscid anther. *C* Connective, *CCCSp* space created by degeneration of the circular cell cluster, *Cr* calcium oxalate crystals, *Cu* cuticle, *E* epidermis, *En* endothecium, *Fb* fibrous bands, *St* stomium. *Bars* **B** 25  $\mu\text{m}$  (scale for **A**), **D** 25  $\mu\text{m}$  (scale for **C**), **H** 15  $\mu\text{m}$  (scale for **E**, **F**, **G**)

sectioned, as illustrated by the bright-field section shown in Figure 9D. The weakness of the stage +11 stomium probably reflects cellular changes (e.g., degeneration events) that occur in preparation for dehiscence at stage +12 (Figs. 1; 10). The TEM analysis of a sheared stage +11 stomium showed that both sides of the anther wall were in close proximity to each other and that the cells were intact (Fig. 10B). In addition, calcium oxalate crystals were still present in the torn stage +11 notch region (Fig. 10B). This was in marked contrast to the stomium at dehiscence (Figs. 9H; 10A). Figure 10A shows a dehiscid stage +12 anther with both sides of the wall positioned across the breakage point, the top half of which is shown in Figure 9H. All cells within the dehiscid notch region (e.g., connective, endothecium) showed that the initial stages of senescence had begun—cells were distorted and their cytoplasmic contents had pulled-away from cell membranes and walls (Figs. 9H; 10A). In addition, the calcium oxalate crystals had disappeared (compare Figs. 9H and 10A with Fig. 9E–G). In contrast with intact epidermal and endothelial cells at stage +12, the stomium appeared to have crumpled, collapsed, and degenerated (Figs. 7; 9H; 10A). Together, these results show that stomium cells remained intact within the anther wall up to stage +11, and that they underwent a cell-death program at stage +12 similar to one that occurred earlier within the circular cell cluster (Fig. 6E–H). Stomium degeneration allowed the anther wall to break and release pollen at flower opening (Fig. 1).

LCM can be used to detect specific stomium mRNAs

We used LCM to determine whether we could isolate stomium cells and detect the presence of specific stomi-

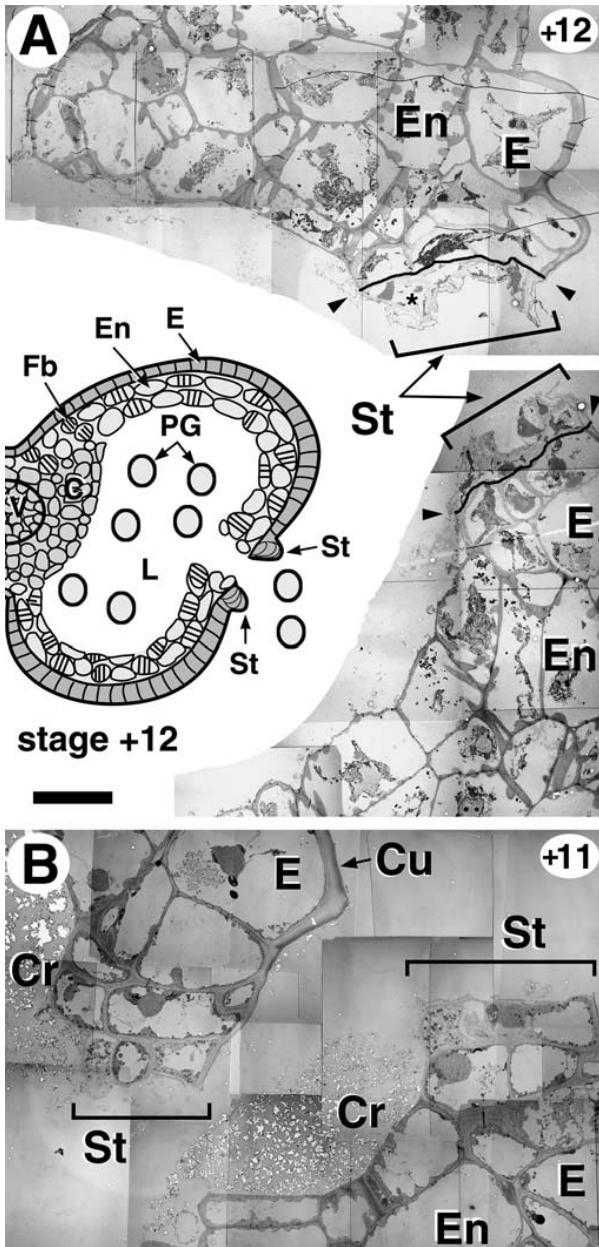
um mRNAs (Kerk et al. 2003). LCM is a powerful tool for using genomic approaches (Asano et al. 2002; Kerk et al. 2003; Nakazano et al. 2003) to identify the genes and proteins that function within the circular cell cluster and stomium during anther development and that are critical for the dehiscence process. Figure 11 shows a transverse section of a stage +6 anther notch region before LCM (Fig. 11A) and after LCM (Fig. 11B). We were able to capture the entire cluster of stomium cells from the notch region and separate it from the rest of the anther using LCM (Fig. 11B).

We isolated RNA from 45 captured stomium cell clusters containing a total of approximately 450 cells (Fig. 5D), and used real-time quantitative reverse transcription-PCR (qRT-PCR) to detect the presence of TA56 and TA20 mRNAs that we had shown previously by in situ hybridization to be localized within the stomium (Koltunow et al. 1990; Beals and Goldberg 1997). TA56 encodes a thiol endopeptidase (Beals and Goldberg 1997), while TA20 encodes a protein of unknown function (Goldberg et al. 1993; Beals and Goldberg 1997). Real-time qRT-PCR detected the presence of both sequences in stomium RNA, and indicated that TA20 mRNA was approximately 4-fold more prevalent than TA56 mRNA (Fig. 11C). Relative to internal rRNA standards, we estimated that TA20 and TA56 mRNAs represented approximately 1.4% and 0.4% of the stage +6 stomium mRNA population, or about 7,000 and 2,000 molecules per cell, respectively (Goldberg et al. 1978). These mRNA prevalences are consistent with those expected for mRNAs that can be detected within specific cell types using in situ hybridization procedures (Cox and Goldberg 1988). Together, these data show that LCM can be used successfully to capture stomium cells from the notch region of the anther and identify specific mRNAs.

---

## Discussion

We characterized the differentiation and degeneration of the circular cell cluster and stomium during tobacco anther development using both light microscopy and TEM. These cell types form within the notch region of the anther and are required for dehiscence and release of pollen grains at flower opening (Fig. 1). Our major findings are summarized in Table 1 and a schematic representation of the cellular events that occur within the notch region at the level of the TEM is shown in Figure 12. Our results show that the stomium and circular cell cluster are specified early in phase 1 of anther development, following the differentiation of territories leading to locule formation. The circular cell cluster differentiates from L2 cells in the territory between the developing locules prior to the initiation of stomium differentiation and becomes a 12–14 cell specialized tissue containing calcium-oxalate crystals (Figs. 12; 13; Bonner and Dickinson 1989; Horner and Wagner 1992; D'Arcy et al. 1996). Thus, the L2 initials that give rise to



**Fig. 10A–B** Stomium cell death and anther dehiscence. Anther sections from stages +11 and +12 were fixed and embedded in Spurr's epoxy resin as described in Materials and methods. Spurr's embedded anthers were sliced into ultra-thin sections and prepared for TEM. Anther developmental stages are shown in the top right corner. **A** Stage +12. Complete notch region of the split stomium shown in Fig. 9H. After breakage at the stomium, the two halves of the anther wall separate from each other. These two halves have been placed in close proximity to each other for this figure. The asterisk indicates the stomium cell highlighted in Fig. 7. The cartoon inset of a stage +12 anther theca represents the role of stomium breakage in pollen release. The black border and the arrowheads delineate the stomium cells from the other cells of the dehiscing anther. Brackets highlight the two halves of the stomium that split at dehiscence. **B** Split notch region from a stage +11 anther that was mechanically sheared prior to fixation. Brackets highlight the two halves of the mechanically split stomium. *C* Connective, *Cr* calcium oxalate crystals, *Cu* cuticle, *E* epidermis, *En* endothecium, *Fb* Fibrous bands, *L* locule region in a bilocular anther, *PG* pollen grain, *St* stomium, *V* vascular region. Bar A 20  $\mu$ m (scale for the TEM micrographs in both **A** and **B**)

bilocular anther, while stomium degeneration occurs just prior to flower opening and provides a longitudinal slit for pollen release along each theca of the anther (Figs. 1, 12). These degeneration events occur independently of phase 2 developmental processes that take place within the locules (e.g., pollen differentiation, tapetum degeneration; Fig. 1), because male-sterile mutants obtained from either genetic screens (Kaul 1988; Dawson et al. 1993; Sanders et al. 1999) or the targeted ablation of the tapetum with cytotoxic genes (e.g., *TA29/barnase*, *TA29/diphtheria toxin A*; Koltunow et al. 1990; Mariani et al. 1990) undergo normal dehiscence.

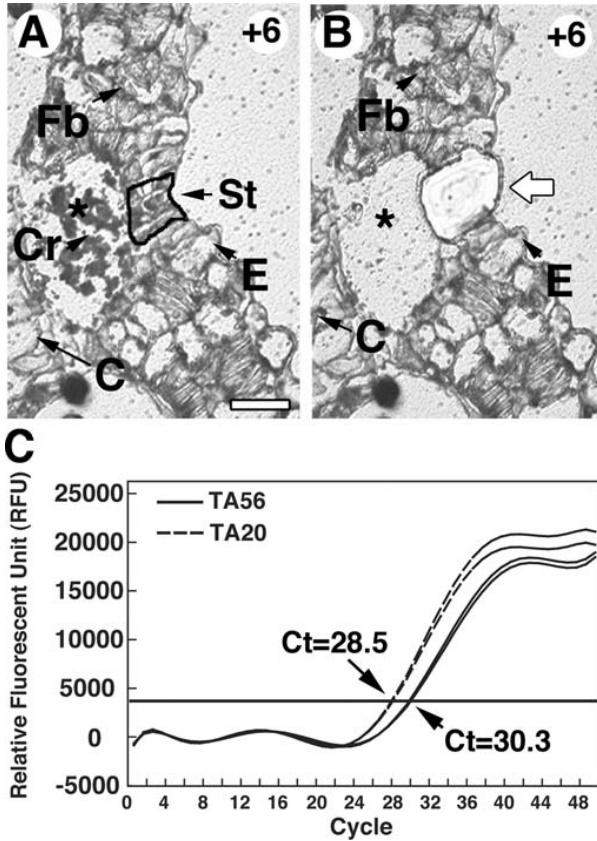
The TA56 thiol endopeptidase mRNA accumulation pattern reflects the sequential degeneration of the circular cell cluster and stomium—TA56 mRNA accumulates first in the circular cell cluster and then in the stomium (Fig. 11), and these events are under precise transcriptional control (Koltunow et al. 1990; Beals and Goldberg 1997). Other hydrolytic enzymes (e.g., cellulase) have been shown to be present in anthers just prior to dehiscence (del Campillo and Lewis 1992; Lashbrook et al. 1994; Neelam and Sexton 1995). In striking contrast with results obtained from tapetal cell ablation, targeted ablation of either the circular cell cluster and stomium or the stomium alone with a *TA56/barnase* gene late in anther development leads to anthers that fail to dehisce (Beals and Goldberg 1997). Collectively, these findings indicate that temporally-regulated cellular processes involving specific gene sets are required for anther dehiscence.

Coordinated events within several different cell types are required for anther dehiscence

In addition to the circular cell cluster and stomium, the endothecium and connective both play a major role in anther dehiscence (Keijzer 1987; Bonner and Dickinson 1989; Manning 1996). Like the circular cell cluster, the

the circular cell cluster are within a different region of the primordium than those that generate the archesporial cells (Figs. 12; 13). The stomium, in contrast, is specified from L1 cells contiguous to the circular-cell-cluster initials, and differentiates into a multi-tiered 9–12 cell structure after the circular cell cluster has begun to degenerate (Figs. 12; 13; Table 1).

Both the circular cell cluster and the stomium undergo a cell-death program and degenerate during phase 2 of anther development, although the timing differs (Figs. 1; 12; Table 1). The circular cell cluster degenerates first, contributing to the formation of a



**Fig. 11A–C** Using laser capture microdissection (LCM) to isolate stomium cells and detect specific mRNAs. Stage +6 anthers were fixed, embedded in Paraplast, sliced into 10  $\mu\text{m}$  transverse sections, and de-paraffinized with two washes of xylene before LCM (see Materials and methods). **A, B** Bright-field photographs of stage +6 anther notch region before and after LCM. **A** Notch region before LCM; *black outline* stomium cells marked for LCM. **B** Notch region after LCM; *asterisks* space previously occupied by the degenerated circular cell cluster (see Fig. 6), *large white arrow* hole left in the anther section after LCM of stomium cells. The calcium oxalate crystals present in **A** were scattered and lost during LCM. **C** Real-time quantitative reverse transcription-polymerase chain reaction (qRT-PCR) analysis of stomium mRNAs using TA56 and TA20 primers as outlined in Materials and methods. *Horizontal line* Ct, or PCR threshold value, calibrated relative to known amounts of plasmid DNA standards (data not shown). Curves show replica qRT-PCR reactions for each primer pair; *Ct* average value (standard deviation < 0.2 Ct). The amount of PCR product varies according to the function  $2^{\Delta\text{Ct}}$ , where  $\Delta\text{Ct}$  is the absolute difference in Ct values. A  $\Delta\text{Ct}$  of 1 represents a 2-fold difference in the amount of PCR product. *C* Connective tissue, *Cr* calcium oxalate crystals, *E* epidermis, *Fb* fibrous bands, *P* pollen, *St* stomium. *Bar A* 50  $\mu\text{m}$  (scale for **B**)

endothecium is derived from L2 layer cells of the anther primordium, but from a different territory (Fig. 13). Endothelial cells generate wall thickenings, or fibrous bands, during phase 2 of anther development (Figs. 1; 9D,F; 13; Table 1) Endothelial wall thickenings are composed of lignin and have been proposed to serve as a “spring” that flips the anther wall layers apart after

stomium breakage (Keijzer 1987; Bonner and Dickinson 1989; Manning 1996). Fibrous bands within the endothecium are required for dehiscence, because an *A. thaliana* male-sterile mutant (*ms55*) that lacks endothelial wall thickenings due to a defect in the *MYB26* gene fails to dehisce (Dawson et al. 1999; Steiner-Lange et al. 2003). The connective, derived from the L3 primordium cell layer, also generates wall thickenings during phase 2 of tobacco anther development (Figs. 6D, H; 9H; 10A; 13). Degeneration of the connective, in addition to the circular cell cluster, establishes a confluent pollen chamber within each theca allowing the pollen grains to exit through a single opening (Fig. 1; Koltunow et al. 1990; Beals and Goldberg 1997). We showed previously that the *A. thaliana non-dehiscence1* mutant fails to undergo dehiscence as a consequence of premature connective cell death, indicating the importance of these cells, either directly or indirectly, in pollen release (Sanders et al. 1999).

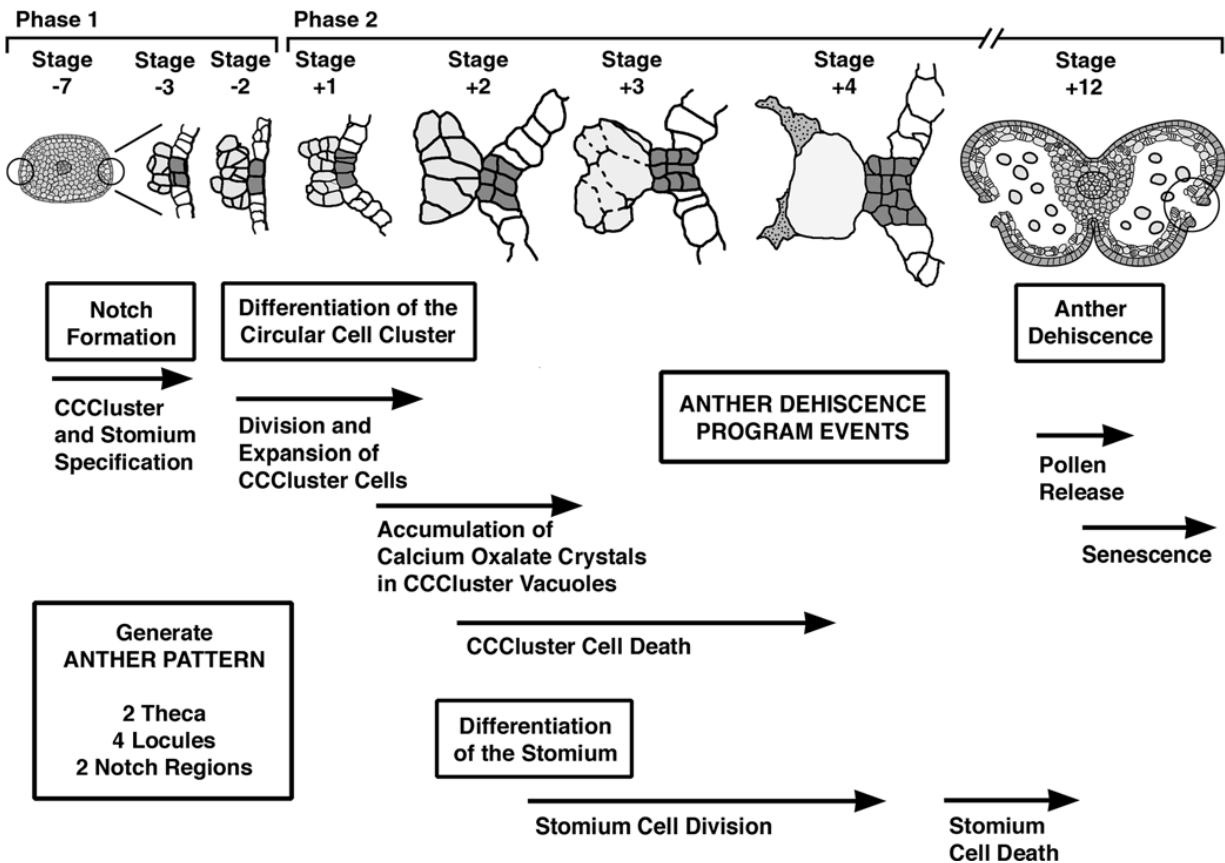
The differentiation and specialization of diverse anther cell types required for dehiscence and pollen release are highly coordinated events that are timed precisely during floral development and scheduled to be completed when the flower opens (Fig. 13). The dehiscence program begins with the specification of cell types within the anther primordium (e.g., circular cell cluster, stomium, endothecium) and ends with breakage of the stomium within the anther wall and pollen release at flower opening (Fig. 13, Table 1). Thus, dehiscence requires a continuum of programmed events during both phase 1 and phase 2 of anther development. How cells required for dehiscence become specified within the L1, L2, and L3 layers of the anther primordium, and what genes and cellular processes guide their specialization during anther development remain major unanswered questions.

The anther dehiscence program is similar in tobacco and *A. thaliana*

Experiments with the anthers of related solanaceous plants (e.g., tomato, Bonner and Dickinson 1989; sweet pepper, Horner and Wagner 1980, 1992) have shown that similar events occur—a calcium-oxalate-filled circular cell cluster differentiates and degenerates, followed by the formation of a multi-tiered stomium that breaks at flower opening. The presence of a circular cell cluster filled with calcium oxalate druse crystals is a feature of solanaceous anthers (D’Arcy et al. 1996), and has been shown recently to enhance the pollination process by supplying calcium ions (Iwano et al. 2004).

Recently, we characterized anther development and the dehiscence program in *A. thaliana* (Sanders et al. 1999, 2000). A comparison of notch-region development and dehiscence in tobacco and *A. thaliana* is summarized in Table 2 and shows the remarkable conservation of precisely timed developmental events leading to pollen release in these divergent plant species. In *A. thaliana*, a small number of septum cells (two or three) differentiates





**Fig. 12** Summary of events that occur during tobacco anther notch-region development. Notch-region cartoons were drawn from the TEM images shown in Figs. 3 and 6. Slight variation in notch-region development occurs within an anther (i.e., position of transverse section) and from anther to anther. The cells outlined in the cartoons (stages -3 to +4) represent the stomium (dark shading), circular cell cluster (light shading) and epidermis (white). At stage +4 the initial degeneration of connective cells is shown as a speckled region. The stage -7 primordium and stage +12 anther cartoons were taken from Fig. 1. *CCcluster* Circular cell cluster

in the notch region from sub-epidermal cells and degenerates midway through phase 2, creating a bilocular anther (Table 2; Sanders et al. 2000). The connective does not degenerate in *A. thaliana*, but does accumulate fibrous bands (Table 2; Sanders et al. 1999). The septum functions similarly to the circular cell cluster, but is simpler in structure, does not accumulate calcium oxalate crystals, and resides at the intersection of two round locules, in contrast with the U-shaped pollen sacs in tobacco (Fig. 1). Following septum degeneration, the stomium becomes visible within the notch region and degenerates just as the flower opens leading to pollen release (Table 2; Sanders et al. 1999, 2000). The *A. thaliana* stomium is also simpler than that observed in tobacco, consisting of only three cells that reside within a single epidermal layer of the notch region (Table 2; Sanders et al. 2000). Cellular events leading to the differentiation and degeneration of

the circular cell cluster/septum and stomium are conserved, because a *TA56/GUS* reporter gene shows the same notch-region transcriptional specificity in *A. thaliana* as it does in tobacco. That is, the *TA56/GUS* gene is transcribed specifically within septum and stomium during phase 2 of anther development (Sanders 2000; P.M. Sanders and R.B. Goldberg, unpublished results). This suggests that the regulatory networks responsible for activating genes required for carrying out specialized dehiscence functions within the circular cell cluster/septum and stomium are ancient and evolved with the emergence of flowering plants.

What controls the specification of the notch region in the territory between the two locules?

How does the notch region form within the territory between the two locules, and how do the L1 and L2 cells within this territory become specified to follow stomium and circular cell cluster differentiation pathways? Figure 14 presents two alternative mechanisms by which these specification events might occur. Mechanism 1 proposes that the notch region forms autonomously during early phase 1 of anther development and that the L1 and L2 cells within this territory are pre-programmed to follow stomium and circular cell cluster pathways

**Table 2** Comparison of anther notch region development and the dehiscence program in tobacco and *Arabidopsis thaliana*

Tobacco <sup>a</sup>		<i>A.thaliana</i> <sup>a</sup>	
Stage <sup>b</sup> -7 to -5	Major events in notch development Anther pattern is generated from the stamen primordia. U-Shaped locules are derived from archesporial cell lineages. Expansion of the two locules in each theca creates the anther notch region	Stage <sup>c</sup> 1-5	Major events in notch development Anther pattern is generated from the stamen primordia. Round locules are derived from archesporial cell lineages. Expansion of the 2 locules in each theca creates the anther notch region
-5 to -1	CCC and stomium are specified within the notch region. CCC differentiates prior to the stomium	6-9	Septum and stomium are specified within the notch region <sup>d</sup>
+1	Meiosis complete within locules and tetrads present. End of phase 1 and beginning of phase 2	7	Meiosis complete within locules and tetrads present. End of phase 1 and beginning of phase 2
+1 to +2	CCC differentiation and division occur and calcium oxalate crystals accumulate	10	Septum cells present within notch region. Calcium oxalate crystals do not form
+2 to +4	CCC degeneration occurs. Stomium differentiation and division generate a multi-tiered structure with 9-12 cells	11	Expansion of endothecium cells occurs. Fibrous bands deposited in endothecium and connective cells
+3 to +5	Expansion of the endothecium and connective and deposition of fibrous bands. Connective degeneration leads to a bilocular anther	12	Septum cells degenerate and lead to a bilocular anther. A 3-celled, single-layered stomium is observed. Connective does not degenerate
+12	Degeneration of stomium creates break in anther wall and dehiscence/pollen release occurs. End of phase 2	13	Degeneration of stomium creates break in anther wall and dehiscence/pollen release occurs. End of phase 2

<sup>a</sup>Tobacco and *A. thaliana* have anthers with a 4 locule structure consisting of 2 locules per theca and a stomium between the 2 locules (1 theca → 2 locules → 1 stomium; see Fig. 1)

<sup>b</sup>Tobacco anther developmental stages were taken from Koltunow et al. (1990). Phase 1, stages -7 to -1; Phase 2 stages +1 to +12. Table 1 details the major events in tobacco anther development and notch region development

<sup>c</sup>*A. thaliana* anther developmental stages taken from Sanders et al. (1999). Phase 1, stages 1-8; Phase 2 stages 7-13

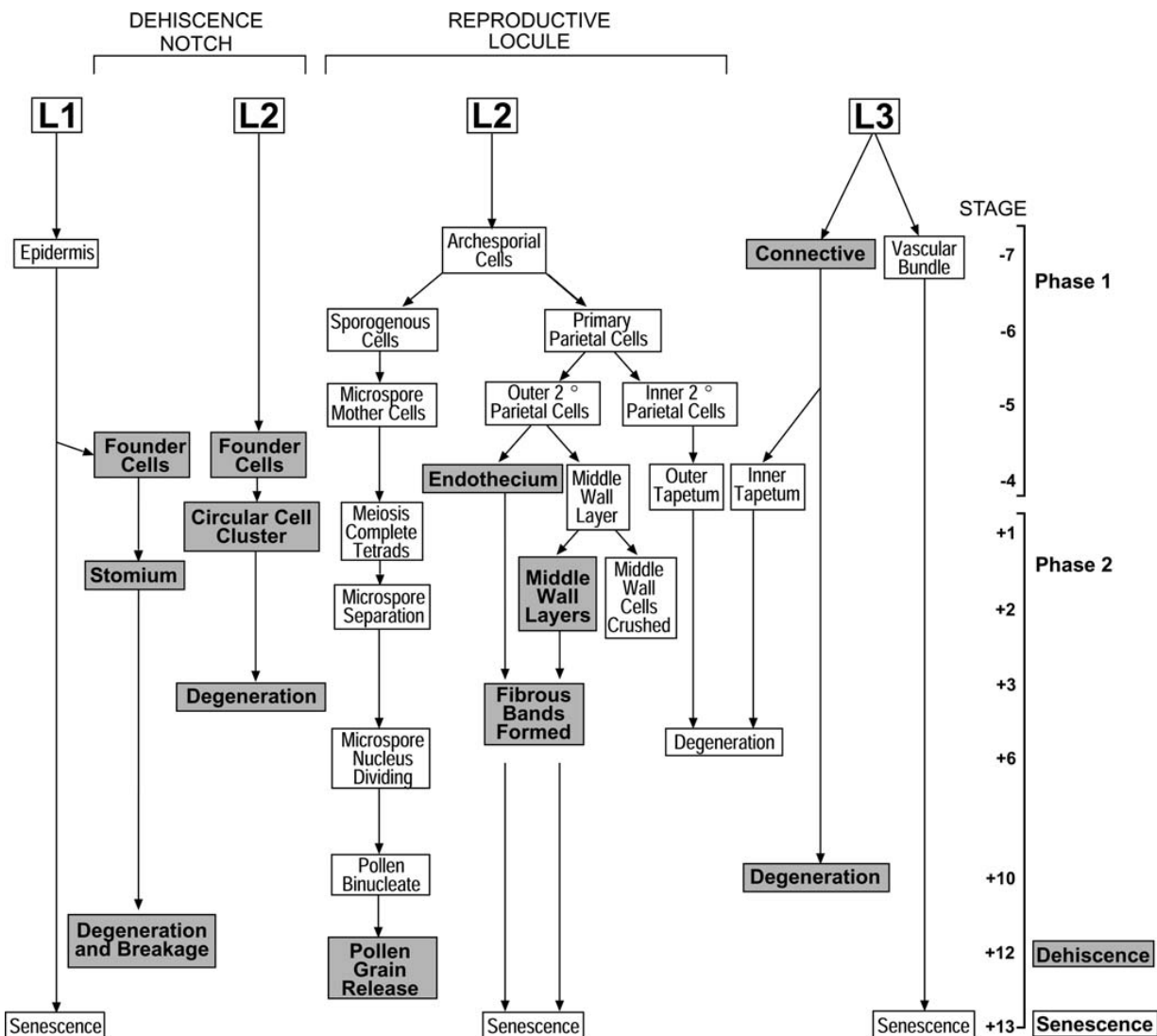
<sup>d</sup>Septum cells in *A. thaliana* analogous in position and function to CCC of tobacco, although septum cells do not accumulate calcium oxalate crystals. The anther developmental stage(s) at which septum and stomium cells differentiate in *A. thaliana* is not known; however, septum and stomium cells are present at stages 10 and 12, respectively (Sanders et al. 1999, 2000). Septum differentiation and degeneration occur before similar events take place in the stomium (Sanders et al. 1999, 2000), analogous to what occurs in tobacco for the CCC and stomium (Table 1)

(Davidson 1991). Mechanism 2, in contrast, hypothesizes that notch region specification is dependent upon signals generated by the differentiating locules. These signals are transmitted obliquely from each locule and, at their intersection, enable stomium and circular cell cluster initials to be specified from receptive L1 and L2 cells (Fig. 14), i.e., stomium and circular cell cluster L1 and L2 founder cells are conditionally specified as a consequence of their position between the two differentiating locules (Davidson 1991). One corollary to mechanism 2 is that L1 or L2 cells on the periphery of the anther primordium should have the potential to differentiate into a stomium or circular cell cluster, respectively, provided the relevant signals are present. Mechanism 1 predicts that altering the number and/or position of locules within each theca should have no effect on notch region development. In contrast, mechanism 2 predicts that alterations in the pattern of locules within the anther should affect notch-region development and dehiscence.

We used mutant *A. thaliana aintegumenta* and *ettin* anthers to show that notch region development and dehiscence are dependent upon the formation of two

contiguous locules (Sanders 2000; P.M. Sanders, Y. Mizukami, R.L. Fischer, and R.B. Goldberg, unpublished observations). Mutations in either the *AINTEGUMENTA* (*ANT*) gene that encodes an AP2-class transcription factor (Elliot et al. 1996; Klucher et al. 1996) or the *ETTIN* (*ETT*) gene that encodes an auxin-response-element transcription factor (Sessions et al. 1997) affect both the position and number of locules within the anther (Sessions et al. 1997). We observed that septum and stomium cell differentiation occurs only at the intersection between two locules, regardless of where these locules are positioned within the anther (Sanders 2000; P.M. Sanders, Y. Mizukami, R.L. Fischer, and R.B. Goldberg, unpublished observations). Notch region formation and dehiscence do not occur within either an anther or theca containing only a single locule (Sanders 2000; P.M. Sanders, Y. Mizukami, R.L. Fischer, and R.B. Goldberg, unpublished observations).

Experiments with *A. thaliana* fertility mutants that have a defect in the *SPOROCTELESS/NOZZLE* MADS-box transcription factor gene (Schiefthaler et al. 1999; Yang et al. 1999) support this conclusion. Archesporial cells do not differentiate in *sporocyteless/nozzle*



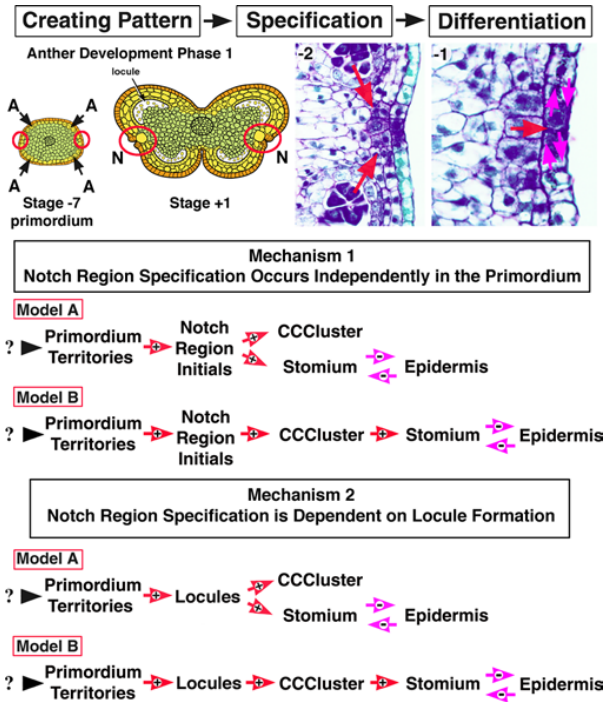
**Fig. 13** Cell lineages participating in tobacco anther dehiscence. Stages at which dehiscence events occur within specific cell types were taken from the results reported in this paper (Table 1). Stages at which other developmental events occur were taken from our previous studies (Koltunow et al. 1990). Cells derived from L1, L2, and L3 primordium layers were based upon the experiments presented here (circular cell cluster, stomium), those reported previously from our laboratory (Koltunow et al. 1990), and the histological studies of Satina and Blakeslee (1941) and Joshi et al. (1967). Middle wall-layer lineage is based upon Davis (1966). Cells of the inner tapetum can originate from both the L2 and L3 layers (Hill and Malmberg 1996). The **bold lettering** and **shaded boxes** represent cell types and events that are part of the anther dehiscence program. Adapted from Goldberg et al. (1993, 1995)

cells are present (Yang et al. 1999). This is consistent with the studies reported here, which show that histological changes within the notch region become apparent only after the locules have begun to differentiate (Fig. 3). Together, these results suggest strongly that the differentiation of the notch region is dependent upon signals generated by two locules within each theca (Fig. 14, mechanism 2). What these signals are, which cells produce them, and how they interact with L1 and L2 cells in the interlocular region to induce stomium and circular cell cluster/septum specification events remain to be determined.

anther primordia, and, as a consequence, fail to differentiate pollen-containing locules and associated wall layers (e.g., tapetum, endothecium). *Sporocytiless/nozzle* anthers also do not contain septum and stomium cells within their rudimentary notch region, even though differentiated epidermal and connective-like sub-epidermal

Do the circular cell cluster and stomium differentiate independently of each other within the notch region?

Do cells of the circular cell cluster and stomium interact with each other during anther development and what



**Fig. 14** Models for the specification and differentiation of the circular cell cluster and stomium. The cartoons for stage -7 and +1 anthers were taken from Fig. 1. The red circles at stage -7 highlight the region in the primordium that will become the notch region, as seen at stage +1. The arrows at stage -7 designate the territories of the four archesporial cell regions that will develop into the locules, as seen in stage +1. The bright-field photographs show the notch region of stage -2 and stage -1 anthers and were taken from Fig. 3. Arrows with a + sign indicate positive interactions, while arrows with a - sign indicate negative interactions. Individual arrows are conceptual and do not reflect the possible number of interactions, molecules, and/or pathways responsible for each step. *A* Archesporial cell territories, *CCCluster* circular cell cluster, *N* notch

defines the boundary of the notch region? Two models for stomium and circular cell cluster differentiation are presented in Figure 14, both of which are dependent upon signaling for the initial specification of the notch region (mechanism 2). In model A, the stomium and circular cell cluster differentiate independently of each other after their L1 and L2 founder cells are specified by a common signaling mechanism. Alternatively, model B proposes that stomium differentiation is dependent upon signals generated by contiguous cells of the circular cell cluster. This model assumes that circular cell cluster development is required before stomium differentiation and division can occur (Fig. 12), and that pre-stomium initials require signals in addition to those obtained from the locules in order to undergo a differentiation pathway. Both models propose that negative interactions (e.g., between the stomium and epidermal neighbors) limit the size of the notch region and define its boundary (Fig. 14; Larkin et al. 2003). Signals between the circular cell cluster and stomium or the stomium and epidermal

cells could be transmitted by plasmodesmata connections (Fig. 8; Haywood et al. 2002), more conventional ligand/receptor pathways (Larkin et al. 2003), or both.

At the present time we do not know which model is correct or whether interactions occur between stomium and adjacent epidermal cells. The circular cell cluster differentiates and divides from stages -4 to +2, whereas analogous stomium events occur later, during stages +1 to +4, while the circular cell cluster is degenerating (Fig. 12; Table 1). Stomium differentiation is completed after the circular cell cluster has disappeared, and stomium cell death occurs in the absence of a circular cell cluster (Fig. 12; Table 1). Thus, the circular cell cluster is not required for cellular processes that occur within the stomium during most of phase 2 of anther development (Fig. 12, stages +4 to +12; Table 1). These observations are consistent with the hypothesis that the circular cell cluster and stomium differentiate and function independently of each other after they are specified within the primordium interlocular region (Fig. 14, model B).

One way to test these models is to use targeted cell ablation studies (Mariani et al. 1990; Koltunow et al. 1990; Goldberg et al. 1995; Beals and Goldberg 1997) to selectively eliminate either the circular cell cluster or the stomium during phase 1 of anther development (Fig. 1). For example, if model B is correct, then destruction of the circular cell cluster will have no effect on stomium differentiation. Conversely, ablation of the stomium will not affect circular cell cluster formation. The absence of the stomium, however, might induce contiguous epidermal cells to enter a stomium differentiation pathway due to the elimination of negative signals (Fig. 14). The ability to use LCM methods (Kerk et al. 2003) to isolate cells from the circular cell cluster and stomium at any stage in their development opens the door for identifying cell-specific genes and promoters that can be used drive the ablation process (Fig. 11).

What coordinates cellular events within the circular cell cluster and stomium with other developmental processes within the flower?

The developmental processes that occur within the notch region must be coordinated with other events that occur within the anther (Goldberg et al. 1993). For example, pollen grain formation, tapetal cell degeneration, expansion of wall and epidermal layers, deposition of fibrous bands in the connective and endothecium, and connective cell degeneration need to be synchronized with the differentiation and degeneration of the circular cell cluster and stomium (Fig. 13; Table 1). In addition, anther development and dehiscence must be timed with events that occur within other floral organs (i.e., sepals, petals, pistil) so that successful pollination and fertilization can take place when the flower opens.

Clues to what controls the timing of anther dehiscence come from studies with late-dehiscence mutants and genes involved in hormone activity. *A. thaliana*



mutants that are defective in either JA biosynthesis (e.g., *dde1/opr3*, *dde2/allene oxidase synthase*, *dad1*) or perception (e.g., *coi1*) are male sterile and have anthers that dehisce too late for successful pollination to occur (Feys et al. 1994; McConn and Browse 1996; Sanders et al. 2000; Stintzi and Browse 2000; Ishiguro et al. 2001; Park et al. 2002; Von Malek et al. 2002). For example, *dde1/opr3* plants lack 12-oxophytodienoate reductase, an enzyme in the JA biosynthetic pathway (Sanders et al. 2000; Stintzi and Browse 2000). The stomium and septum differentiate normally in *dde1/opr3* anthers, but stomium degeneration is delayed, indicating that JA is an important signal for controlling the timing of stomium breakage and anther dehiscence (Sanders et al. 2000). Other hormones, such as ethylene, auxin, and GA, have also been shown to play a role in anther dehiscence. Inhibiting the *etr1-1 ethylene receptor* gene (Rieu et al. 2003) or increasing auxin sensitivity by *Agrobacterium tumefaciens rolB* gene over-expression (Cecchetti et al. 2004) generates a late-dehiscence phenotype in transgenic tobacco plants analogous to that observed in *dde1/opr3* anthers (Sanders et al. 2000; Stintzi and Browse 2000). In addition, overexpressing the *HvGAMYB* gene in barley, a GA-induced transcriptional regulator, leads to anthers that fail to dehisce (Murray et al. 2003), perhaps acting in cooperation with a microRNA (Achard et al. 2004). Together, these studies indicate that a complex set of hormonal interactions is required, in part, to coordinate events within the flower leading to anther dehiscence. How this occurs, as well as the molecular processes and genes that control the differentiation and degeneration of cells required for anther dehiscence, remains to be determined.

**Acknowledgements** We thank Birgitta Sjostrand for her expertise and help with the TEM studies, Sharon Hue Tu for assistance taking the TEM pictures, and Layla Maloff for teaching us how to use the Leica Laser Capture Microdissection System. We also thank Lynne Olson and Reed Hutchinson at the OID Digital Imaging Facility at the UCLA Center for Health Sciences for help in the preparation of figures. This work was funded by NSF and HHMI grants to R.B.G.

## References

Achard P, Herr A, Baulcombe DC, Harberd NP (2004) Modulation of floral development by a gibberellin-regulated microRNA. *Development* 131:3357–3365

Asano T, Masumura T, Kusano H, Kikuchi S, Kurita A, Shimata H, Kadowaki K-i (2002) Construction of a specialized cDNA library from plant cells isolated by laser capture microdissection: toward comprehensive analysis of the genes expressed in the rice phloem. *Plant J* 32:401–408

Beals TP, Goldberg RB (1997) A novel cell ablation strategy blocks tobacco anther dehiscence. *Plant Cell* 9:1527–1545

Bonner LJ, Dickinson HG (1989) Anther dehiscence in *Lycopersicon esculentum* Mill. I. Structural aspects. *New Phytol* 113:97–115

Botha CEJ, Hartley BJ, Cross RHM (1993) The ultrastructure and computer-enhanced digital image analysis of plasmodesmata at the Kranz mesophyll-bundle sheath interface of *Themeda triandra* var. *imberbis* (Retz) A. Camus in conventionally-fixed leaf blades. *Ann Bot* 72:255–261

Campillo E del, Lewis LN (1992) Occurrence of 9.5 cellulase and other hydrolases in flower reproductive organs undergoing major cell wall disruption. *Plant Physiol* 99:1015–1020

Cecchetti V, Pomponi M, Altamura MM, Pezzotti M, Marsilio S, D'Angeli SD, Tornielli GB, Constantino P, Cardarelli M (2004) Expression of *rolB* in tobacco flowers affects the coordinated processes of anther dehiscence and style elongation. *Plant J* 38:512–525

Cox KH, Goldberg RB (1988) Analysis of gene expression. In: Shaw CH (ed) *Plant molecular biology: a practical approach*. IRL Press, Oxford, pp 1–34

D'Arcy WG (1996) Anthers and stamens and what they do. In: D'Arcy WG, Keating RC (eds) *The anther: form, function and phylogeny*. Cambridge University Press, Cambridge, pp 1–24

D'Arcy WG, Keating RC, Buchmann SL (1996) The calcium oxalate package or so-called resorption tissue in some angiosperm anthers. In: D'Arcy WG, Keating RC (eds) *The anther: form, function and phylogeny*. Cambridge University Press, Cambridge, pp 159–191

Davidson EH (1991) Spatial mechanisms of gene regulation in metazoan embryos. *Development* 113:1–26

Davis GL (1966) *Systematic embryology of the angiosperms*. Wiley, New York

Dawson J, Wilson ZA, Aarts MGM, Braithwaite AF, Briarty LG, Mulligan BJ (1993) Microspore and pollen development in six male-sterile mutants of *Arabidopsis thaliana*. *Can J Bot* 71:629–638

Dawson J, Sözen E, Vizir I, Waeyenberge SV, Wilson ZA, Mulligan BJ (1999) Characterization and genetic mapping of a mutation (*ms35*) which prevents anther dehiscence in *Arabidopsis thaliana* by affecting secondary wall thickening in the endothecium. *New Phytol* 144:213–222

Elliott RC, Betzner AS, Huttner E, Oakes MP, Tucker WQ, Gerentes D, Perez P, Smyth DR (1996) *AINTEGUMENTA*, an *APET-ALA2*-like gene of *Arabidopsis* with pleiotropic roles in ovule development and floral organ growth. *Plant Cell* 8:155–168

Feys BJ, Benedetti CE, Penfold CN, Turner JG (1994) *Arabidopsis* mutants selected for resistance to the phytotoxin coronatine are male sterile, insensitive to methyl jasmonate, and resistant to a bacterial pathogen. *Plant Cell* 6:751–759

Goldberg RB, Hoschek G, Kamalay JC, Timberlake WE (1978) Sequence complexity of nuclear and polysomal RNA in leaves of the tobacco plant. *Cell* 14:123–131

Goldberg RB, Beals TP, Sanders PM (1993) Anther development: basic principles and practical applications. *Plant Cell* 5:1217–1229

Goldberg RB, Sanders PM, Beals TP (1995) A novel cell-ablation strategy for studying plant development. *Phil Trans R Soc Lond B* 350:5–17

Haywood V, Kragler F, Lucas WJ (2002) Plasmodesmata: pathways for protein and ribonucleoprotein signaling. *Plant Cell* 14 [Suppl]:S303–S325

Hill JP, Malmberg RL (1996) Timing of morphological and histological development in premeiotic anthers of *Nicotiana tabacum* cv. Xanthi (Solanaceae). *Am J Bot* 83:285–295

Horner HT, Wagner BL (1980) The association of druse crystals with the developing stomium of *Capsicum annuum* (Solanaceae) anthers. *Am J Bot* 67:1347–1360

Horner HT, Wagner BL (1992) Association of four different calcium crystals in the anther connective tissue and hypodermal stomium of *Capsicum annuum* (Solanaceae) during microsporangogenesis. *Am J Bot* 79:531–541

Ishiguro S, Kawai-Oda A, Ueda J, Nishida I, Okada K (2001) The *DEFECTIVE IN ANTER DEHISCENCE1* gene encodes a novel phospholipase A1 catalyzing the initial step of jasmonic acid biosynthesis, which synchronizes pollen maturation, anther dehiscence, and flow opening in *Arabidopsis*. *Plant Cell* 13:2191–2209

Iwano M, Tetsuyuki E, Shiba H, Takayama S, Isogai A (2004) Calcium crystals in the anther of *Petunia*: the existence and biological significance in the pollination process. *Plant Cell Physiol* 45:40–47

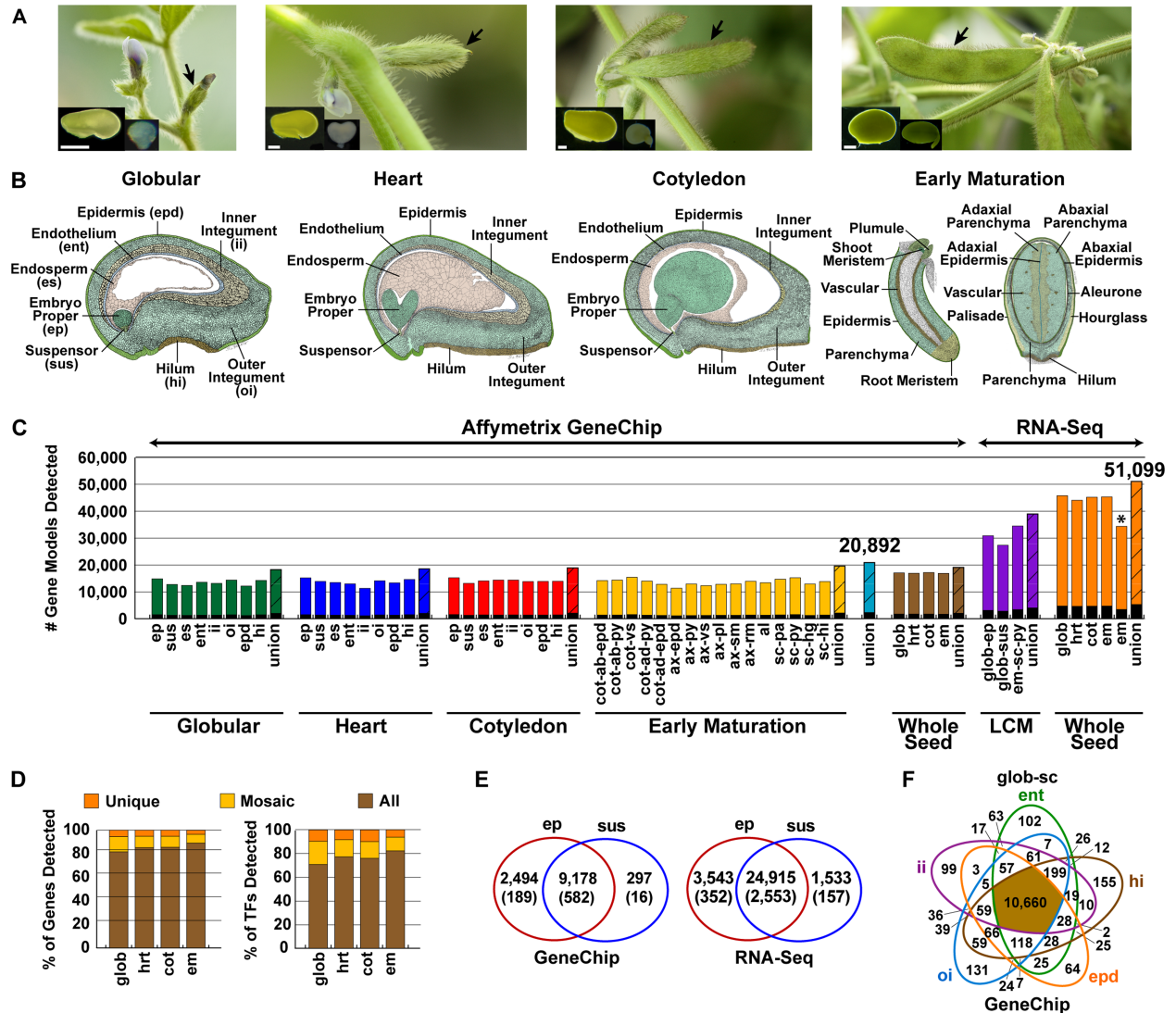


- Joshi PC, Wadhvani AM, Johri BM (1967) Morphological and embryological studies of *Gossypium*. Proc Natl Inst Sci India 33:37–93
- Kaul MLH (1988) Male sterility in higher plants. Springer, Berlin Heidelberg New York
- Keijzer CJ (1987) The process of anther dehiscence and pollen dispersal. I. The opening mechanism of longitudinally dehiscing anthers. New Phytol 105:487–498
- Kerk NM, Ceserani T, Tausta SL, Sussex IM, Nelson TM (2003) Laser capture microdissection of cells from plant tissues. Plant Physiol 132:27–35
- Klucher KM, Chow H, Reiser L, Fischer RL (1996) The *AINTEGUMENTA* gene of *Arabidopsis* required for ovule and female gametophyte development is related to the floral homeotic gene *APETALA2*. Plant Cell 8:137–153
- Koltunow AM, Truettner J, Cox KH, Wallroth M, Goldberg RB (1990) Different temporal and spatial gene expression patterns occur during anther development. Plant Cell 2:1201–1224
- Larkin JC, Brown ML, Schiefelbein J (2003) How do cells know what they want to be when they grow up? Lessons from epidermal patterning in *Arabidopsis*. Annu Rev Plant Biol 54:403–430
- Lashbrook CC, Gonzalez-Bosch C, Bennett AB (1994) Two divergent endo- $\beta$ -1,4-glucanase genes exhibit overlapping expression in ripening fruit and abscising flowers. Plant Cell 6:1485–1493
- Manning JC (1996) Diversity of endothelial patterns in the angiosperms. In: D'Arcy WG, Keating RC (eds) The anther: form, function and phylogeny. Cambridge University Press, Cambridge, pp 136–158
- Mariani C, De Beuckeleer M, Truettner J, Leemans J, Goldberg RB (1990) Induction of male sterility in plants by a chimeric ribonuclease gene. Nature 347:737–741
- McConn M, Browse J (1996) The critical requirement for linolenic acid is pollen development, not photosynthesis, in an *Arabidopsis* mutant. Plant Cell 8:403–416
- Murray F, Kalla R, Jacobsen J, Gubler F (2003) A role for HvGMYB in anther development. Plant J 33:481–491
- Nakazano M, Qui F, Borsuk L, Schnable PS (2003) Laser capture microdissection, a tool for the global analysis of gene expression in specific plant cell types: identification of genes expressed differentially in epidermal cells or vascular cells. Plant Cell 15:1–15
- Neelam A, Sexton R (1995) Cellulase (endo  $\beta$ -1,4 glucanase) and cell wall breakdown during anther development in sweet pea (*Lathyrus odoratus* L.): isolation and characterization of partial cDNA clones. J Plant Physiol 146:622–628
- Park J-H, Halitschke R, Kim HB, Baldwin IT, Feldmann KA, Feyereisen R (2002) A knock-out mutation in allele oxidase synthase results in male sterility and defective wound signal transduction in *Arabidopsis* due to a block in jasmonic acid biosynthesis. Plant J 31:1–12
- Park SK, Yoon YH, Kim BC, Hwang YH, Chung IK, Nam HG, Kim DU (1996) Pollen of a male-sterile mutant of *Arabidopsis thaliana* isolated from a T-DNA insertion pool is not effectively released from the anther locule. Plant Cell Physiol 37:580–585
- Rieu I, Wolters-Arts M, Derksen J, Mariani C, Weterings K (2003) Ethylene regulates the timing of anther dehiscence in tobacco. Planta 217:131–137
- Sanders PM (2000) Anther dehiscence: genetic and molecular characterization. Ph.D. Dissertation, University of California, Los Angeles, ISBN 0–599–61025–5, UMI Abstract ATT 9957828 (<http://wwwlib.umi.com/dissertations/fullcit/9957828>)
- Sanders PM, Bui AQ, Weterings K, McIntire KN, Hsu YC, Lee PY, Truong MT, Beals TP, Goldberg RB (1999) Anther developmental defects in *Arabidopsis thaliana* male-sterile mutants. Sex Plant Reprod 11:297–322
- Sanders PM, Lee PY, Bieggen C, Boone JD, Beals TP, Weiler EW, Goldberg RB (2000) The *Arabidopsis* *DELAYED DEHISCENCE1* gene encodes an enzyme in the jasmonic acid synthesis pathway. Plant Cell 12:1041–1062
- Satina S, Blakeslee AF (1941) Periclinal chimeras in *Datura stramonium* in relation to development of leaf and flower. Am J Bot 28:862–871
- Schieffhale U, Balasubramanian S, Sieber P, Chevalier D, Wisman E, Schneitz K (1999) Molecular analysis of *NOZZLE*, a gene involved in pattern formation and early sporogenesis during sex organ development in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 96:11664–11669
- Scott RJ, Spielman M, Dickinson HJ (2004) Stamen structure and function. Plant Cell 16 [Suppl]:S46–S60
- Sessions A, Nemhauser JL, McColl A, Roe JL, Feldmann KA, Zambryski PC (1997) *ETTIN* patterns the *Arabidopsis* floral meristem and reproductive organs. Development 124:4481–4491
- Spurr AH (1969) A low viscosity epoxy resin embedding medium for electron microscopy. J Ultrastruct Res 26:31–43
- Steiner-Lange S, Unte US, Eckstein L, Yang C, Wilson ZA, Schmeizer E, Dekker K, Saedler H (2003) Disruption of *Arabidopsis thaliana* *MYB26* results in male sterility due to non-dehiscence anthers. Plant J 34:519–528
- Stintzi A, Browse J (2000) The *Arabidopsis* male sterile mutant, *opr3*, lacks the 12-oxophytodienoic acid reductase required for jasmonate biosynthesis. Proc Natl Acad Sci USA 97:10625–10630
- Trull MC, Holaway BL, Friedman WE, Malmberg RL (1991) Developmentally regulated antigen associated with calcium crystals in tobacco anthers. Planta 186:13–16
- Venkatesh CS (1957) The form, structure, and special ways of dehiscence of anthers of *Cassia* 3. Subgenus *Senna*. Phytomorphology 7:253–273
- Von Malek B, van der Graaff E, Schneitz K, Keller B (2002) The *Arabidopsis* male-sterile mutant *dde2-2* is defective in the *ALLENE OXIDASE SYNTHASE* gene encoding one of the key enzymes of the jasmonic acid biosynthesis pathway. Planta 216:187–192
- Yang WC, Ye D, Xu J, Sundaresan V (1999) The *SPORO-CYTELESS* gene of *Arabidopsis* is required for initiation of sporogenesis and encodes a novel nuclear protein. Genes Dev 13:2108–2117

## APPENDIX F

An Atlas of Gene Activity in Soybean Seed Regions, Compartments, and Tissues During  
Development

This project is a collaborative effort between our lab and John Harada's at UC Davis as part of the NSF Plant Genome Program grant to identify all the genes required to "make a soybean seed." This project involved profiling gene activity in every tissue, region, and compartment of a soybean seed from fertilization to maturation, using LCM coupled with soybean GeneChip arrays. For this project, with the help of Anhtu Bui and Javier Wagmaister, I annotated > 37,000 features on the soybean GeneChip array (see appendix I), classifying each probe set feature into functional categories, like I did for the *Arabidopsis* GeneChip arrays. I established computational approaches to analyze these transcriptome datasets. Additionally, I generated RNA-Seq data from whole seeds and embryo regions to compare against the GeneChip datasets. The results from this project is currently being prepared for a manuscript where I will be a co-author with the senior author on the project, Jungim Hur. Furthermore, I designed and helped create the Seed Gene Network website that currently holds all these datasets and provides visualization of the datasets across development. This Appendix F contains a figure from the manuscript in preparation summarizing the project results and screenshots of the Seed Gene Network website detailing the data analysis and visualization tools create for this project.



**Figure F-1. Genes active in compartments, regions, and tissues throughout soybean seed development.** (A) Developmental stages of soybean seeds used in the study. Soybean plants with pods (arrows) containing seeds with globular (glob), heart (hrt), cotyledon (cot), and early-maturation (em) stage embryos in the insets are shown. Scale bars indicate 50 microns (glob, hrt, and cot stage) and 100 microns (em stage) in length. (B) Drawings of longitudinal soybean seeds containing glob, hrt, and cot stage embryos. Longitudinal and cross section of em stage embryo axis and seeds containing embryo cotyledons, aleurone, and seed coat layer, respectively. All of studied seed parts are listed for each developmental stage. Drawings were traced from bright-field photograph of two microns plastic sections. (C) Number of genes

detected in the seed parts throughout soybean development by GeneChip and RNA-Seq analysis. The black portion of each bar indicates the number of gene models associated with detected transcription factor (TF) mRNAs. For GeneChip analysis, the number of gene models detected indicates the number of gene models associated with detected mRNAs (SI Materials and Methods), and these numbers are represented in Figs. S1 and S2. For RNA-Seq, the number of the gene models detected indicates gene models associated with uniquely mapped reads. Union (colored stripe bar) refers to the union of mRNAs detected at each developmental stage or throughout seed development. \* indicates the number of gene models detected from amplified em stage seed total RNA. (D-F) Regulation of gene activity in different parts of the seed during development. Detected genes by GeneChip were filtered to identify compartment-, region-, or tissue-specific genes and shared genes by two or more seed parts using the criteria described by Le et al. {Le, 2010 #2}. (D) Genes, including TFs, detected in different parts of the seed at every developmental stage by GeneChip. In each bar, the orange, yellow, and brown portion indicate genes specific to different part of the seed, shared by mosaic combinations of seed parts, and detected in all the seed parts at each developmental stage, respectively. (E) Venn diagram of mRNAs detected in glob-ep and glob-sus using GeneChip and RNA-Seq. Number of gene models detected in glob-ep (red oval) and glob-sus (blue oval) by GeneChip and RNA-Seq are shown. Parentheses indicate the number of TF gene models. (F) Venn diagram of mRNAs detected in seed coat tissues at glob stage using GeneChip. Number of genes specific to each glob coat tissue at stage seed, shared by mosaic combinations, and detected in all the seed coat tissues are shown.

Supported by: 

# GENE NETWORKS IN SEED DEVELOPMENT

*Identifying all the genes and gene networks required to "make a seed"*

Home   Soybean GeneChip Experiments   Arabidopsis GeneChip Experiments   GeneChip Annotations   Sequencing   RNAi   People   Data & Resources   Genome Browsers

## Welcome to Gene Networks in Seed Development Website!

**ABOUT THE PROJECT**

This NSF-funded project is a collaborative effort between the **Goldberg laboratory** at UCLA and the **Harada laboratory** at UCSD to understand what are all the genes required to make a soybean seed. We used soybean and *Arabidopsis* Affymetrix GeneChips, Laser Capture Microdissection (LCM), and next-generation high-throughput sequencing technologies to profile the mRNA sets present in different seed regions and compartments throughout development. Our long-term goal is to understand the genes and regulatory networks required to make a seed. [Click here to learn more about this project and what has been accomplished.](#)

**GENECHIP EXPERIMENTS**

To date, we have profiled the mRNA sets present in 71 soybean and *Arabidopsis* seed compartments from preglobular- to early maturation-stage seeds. All GeneChip data are stored in this web-based database. Under the **Soybean GeneChip Experiments** and **Arabidopsis GeneChip Experiments** sections on top, we created the built-in analysis tools to allow the user to not only browse the database by probe identification, gene ontology, and functional category, but also compare gene activity in different seed compartments during development.

**RNAi**




*Soybean*  
*(Glycine max)*




*Arabidopsis*  
*thaliana*

**Figure F-2. Seed Gene Network website (<http://seedgenenetwork.net/>).** This website serves as a portal for the dissemination of data generated from the NSF-funded project. Data including microarrays, mRNA-Seq, smRNA-Seq, and BS-DNA-Seq generated to study gene regulation during seed development are all integrated in this website.

Supported by: 

# GENE NETWORKS IN SEED DEVELOPMENT

*Identifying all the genes and gene networks required to "make a seed"*

[Home](#)
[Soybean GeneChip Experiments](#)
[Arabidopsis GeneChip Experiments](#)
[GeneChip Annotations](#)
[Sequencing](#)
[RNAi](#)
[People](#)
[Data & Resources](#)

Genome Browsers

## Browse Soybean mRNAs Profiling Database

[More Tools](#)
[Analyze](#)
[Blast](#)

Use the search form below to browse and search the gene expression profile of your gene of interest, then view the seed expression patterns using the following options:

- Type in the Probe Set Identifier, *Arabidopsis* Genome Initiative (AGI) locus ID, Predicted Gene Model ID (Please separate multiple IDs by ';'. No Space.), **OR**
- Type in Gene Ontology ID (Multiple IDs are not acceptable), **OR**
- Select Functional Category, **OR**
- Type in the keyword related to the gene of interest.

Those options can be combined to make your selection more specific. Once you set up your searching criteria, click **Submit Query** to start the analysis.

Please be patient and do not cancel the process prematurely. A list of genes matching the criteria you set up on this page will be returned in the "Search Results" page. Click on the probe set identifier to view the gene expression profile. Also you can download the list of genes by clicking the button the top of "Search Results" page as a text-format file.

**Browse by Probe Set or Gene ID**

Probe Set Identifier:  (e.g. Gma.11106.1.S1\_a\_at)

AGI Locus ID  (e.g. AT1G21970)

Predicted Gene Model ID  (e.g. Glyma14g40590)

**Browse by Gene Ontology Term ID**

GO: Biological Processes  (e.g. GO:0006355)

GO: Cellular Components  (e.g. GO:0005634)


GO: Molecular Function  (e.g. GO:0003677)

**Browse by Functional Categories or Keywords**

Functional Category:

Description/Keyword:  (e.g. transcription, CCAAT)

**Figure F-3. Seed Gene Network website -- Search Form.** GeneChip data can be searched by probe set ID, gene ontology terms, or functional categories, providing multiple ways to explore the GeneChip datasets.

Supported by: 

# GENE NETWORKS IN SEED DEVELOPMENT

*Identifying all the genes and gene networks required to "make a seed"*

[Home](#)  
 [Soybean GeneChip Experiments](#)  
 [Arabidopsis GeneChip Experiments](#)  
 [GeneChip Annotations](#)  
 [Sequencing](#)  
 [RNAi](#)  
 [People](#)  
 [Data & Resources](#)

Genome Browsers

## Search results

Showing results where probe name contains *Gma.11106.1.S1\_a\_at*, is in project *Soybean Array* (1 total matches).


Click the button below download GeneChip annotation, data, or both for your search results in text format.

[Download GeneChip Annotation](#)  
 [Download GeneChip Data](#)  
 [Download GeneChip Annotation and Data](#)

Probe Set	Project	Gene Model	AGI Locus	Functional Category	Description
1 Gma.11106.1.S1_a_at	Soybean Array	Glyma06g02990, Glyma04g02980	AT3G54340	Transcription	AP3 (APETALA 3); DNA binding / transcription factor

**Figure F-4. Seed Gene Network website -- Search Results Output.** The output displays a table containing the probe set ID and associated annotations. Clicking on the probe set ID will direct the user to the probe set data page containing more annotation information and visual display of the day (see **Figure F-5 and F-6**).



Supported by: 

# GENE NETWORKS IN SEED DEVELOPMENT

*Identifying all the genes and gene networks required to "make a seed"*

Home Soybean GeneChip Experiments Arabidopsis GeneChip Experiments GeneChip Annotations Sequencing RNAi People Data & Resources

Genome Browsers

## GeneChip Expression Profile for Probe Set - Gma.11106.1.S1\_a\_at

Expression Profile Summary | Download GeneChip Data | Probe Set Description | View in Genome Browser


The images below indicates showing where mRNAs related to this probe set accumulated in seeds throughout development. The color corresponds to the average signal intensity of biological replicates (shown in the right).

The plot below displays the average signal intensity of biological replicates for all seed tissues/compartments studied in this project.

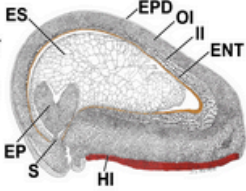
- Click the images to enlarge. To save the images and plot, PC users right-click on the image and choose 'save image as'; Mac users hold 'ctrl' key and click on the image, then choose 'save image as'.
- To view the description of this probe set, click "Probe Set Description" on the top.
- Click "Download GeneChip Data" on the top to get the signal intensities of all GeneChip experiments for this probe set.
- Also, you can view where the probe set aligns in the genome and 454 ESTs that map to corresponding gene using "Genome Browser".

Present (P)	Absent (A)
Insufficient (INS)	Avg. Signal < 500
Avg. Signal 500-5,000	Avg. Signal 5,000-10,000
Avg. Signal >10,000	

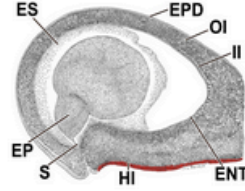
**Globular Stage**



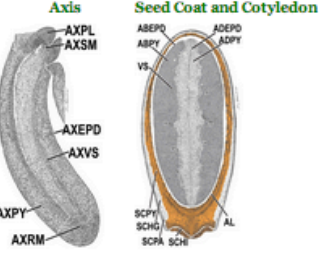
**Heart Stage**



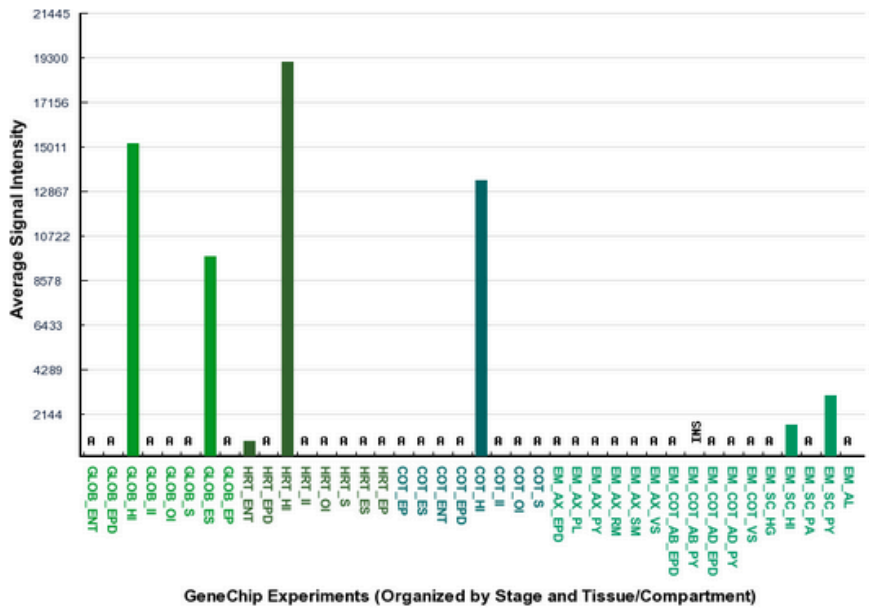
**Cotyledon Stage**



**Early Maturation Stage**



**Figure F-5. Seed Gene Network website -- Heat map visualization of the GeneChip datasets.** This page shows visualization of the GeneChip signal intensity for the selected probe set in different seed regions, compartments, and tissues displayed as a heat map overlaying hand drawn images of soybean seed compartments across development. Users can download the GeneChip data specific for this probe set including annotation information.



**Figure F-6. Seed Gene Network website -- Bar plot visualization of the GeneChip datasets.** An alternative view of the GeneChip signal intensity for the select probe set in different seed regions, compartments, and tissues across development displayed as bar plots.

## APPENDIX G

Gene Activity In Different Regions of a Globular Stage SRB Embryo By EST Sequencing

This project was part of a DOE grant awarded to Bob Goldberg to study gene activity in the embryo proper and suspensor regions of a globular stage embryo. Embryo proper and suspensor regions were hand dissected from the globular stage SRB embryo, taking advantage of SRB giant embryo. Sanger EST sequencing and 454 NGS was used profile gene activity in these embryo regions. My role for this project involved isolating DNA from plasmid clones and analysis of EST sequence data using BLAST, establishing a web-based relational database for the processing and storage of sequence data, carried out *in-situ* hybridization to study the mRNA localization pattern within the embryo region. Part of this work is published in the *Plant Physiology* review (Chapter one). Additionally, I was also involved in sequencing and annotating the SRB genome using 454 and Illumina NGS technologies.

Embryogenesis in higher plants begin with a double fertilization event, one sperm cell fertilizing the egg cell to form the zygote, the other sperm cell fertilizing the central cell to form the endosperm (**Figure G-1**). The zygote undergoes an asymmetric division, giving rise to a small apical cell and a large basal cell. The small apical cell will divide to form the embryo proper and the whole plant, whereas the large basal cell divides to form the suspensor, a terminally differentiated structure that supports and nourishes the embryo proper and degenerates later during development. Little is known about the mechanism of asymmetrical division of the zygote as well as the function of the suspensor and embryo proper during early embryogenesis. The study of early embryogenesis using model plants such as *Arabidopsis* is difficult due to the small size of an *Arabidopsis* seed and the enclosure of the embryo within the protective seed coat. To circumvent these limitations, we chose to use *Phaseolus coccineus* (aka Scarlet Runner Bean (SRB)) with its large seed and embryo. The SRB globular stage embryo is ~100 times larger than *Arabidopsis*, enabling hand micro-dissection and capturing of embryo proper and suspensor region (**Figure G-2**).

Our lab developed a strategy to identify genes active during early plant embryo development (**Figure G-3**). Using SRB as a system, we hand dissected embryo proper and suspensor region from 6-7 days after pollination (DAP) embryos, created a cDNA library, and carried out EST sequencing. Selected genes (e.g. transcription factors) identified from sequencing were verified by in-situ hybridization assay and real-time qPCR.

Using the strategy outlined in **Figure G-3**, cDNA libraries from globular stage suspensor and embryo proper regions were generated by using the SMART Cloning kit (Clontech). Double stranded cDNAs were digested with SfiI enzyme, size selected over a Sepharyl 400 column, and fragments > 0.4Kb were selected. Digested cDNAs were directionally ligated with SfiI-digested lambda arms (TripleEx2) to generate the libraries and single pass Sanger sequencing using Big Dye Terminator kit. Approximately 20,000 ESTs were generated using the Sanger sequencing method and another 380,000 ESTs obtained through 454 sequencing.

Approximately 400,000 ESTs have been sequenced to-date 305,363 and 85,776 ESTs from the suspensor and embryo proper libraries, respectively. All the sequence information are stored in a public database at <http://estdb.biology.ucla.edu/PcEST> (**Figure G-4**). **Figure G-4** shows a summary of the ESTs organized into functional categories. EST sequences were blasted against the NCBI non-redundant database and blast hit with e-value < e-04 was considered significant. Blast description was used to assign the ESTs into functional categories. More than 30% of the ESTs sequenced from the embryo proper library code for proteins involved in protein synthesis in comparison to 18% from the suspensor library. This is not surprising due to the fact that the embryo proper at the globular stage is still dividing and transcriptionally active whereas the suspensor have been fully differentiated. On the other hand, the suspensor have more than twice the number of ESTs encoding enzymes involved in gibberellic acid (GA) biosynthesis. This confirms previous works that showed the enrichment of GAs within the suspensor compared to the embryo proper.

In-situ hybridization analysis were carried out to determine the mRNA localization of selected ESTs (**Figure G-5**). Several ESTs showed mRNA localization primarily in the suspensor (G564, PCEP03567, GA 20-oxidase) and embryo proper (PCEP01711), including transcription factors. These results indicated that there are regional localization of mRNAs in the embryo proper and suspensor during early embryogenesis and that mRNAs with similar localization pattern might form a network of genes active specifically in each embryo region.

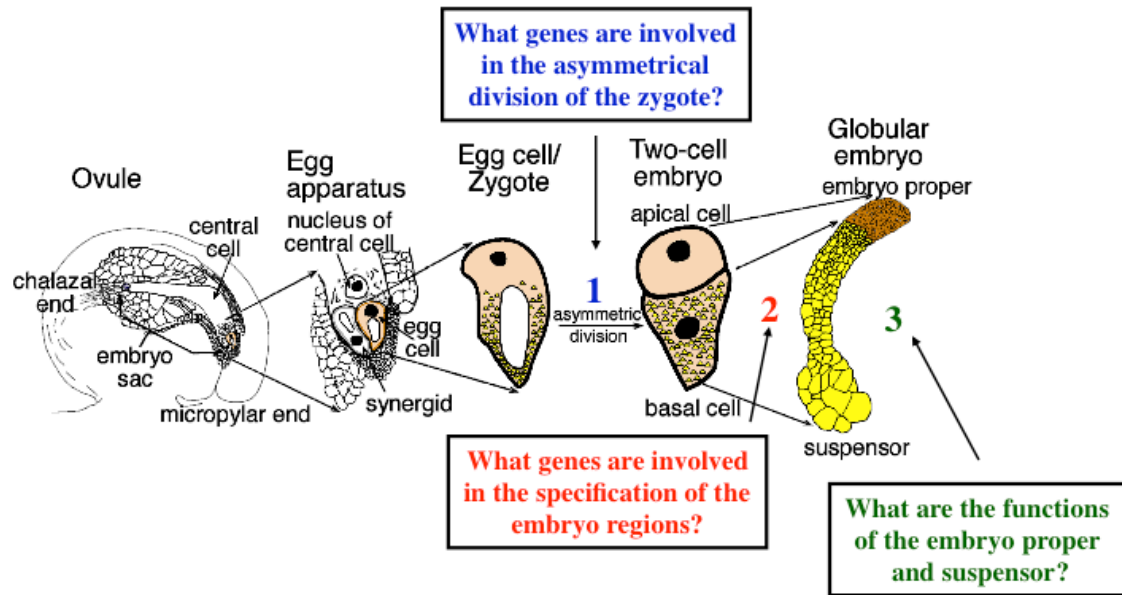


Figure G-1. Key questions in early embryo development.

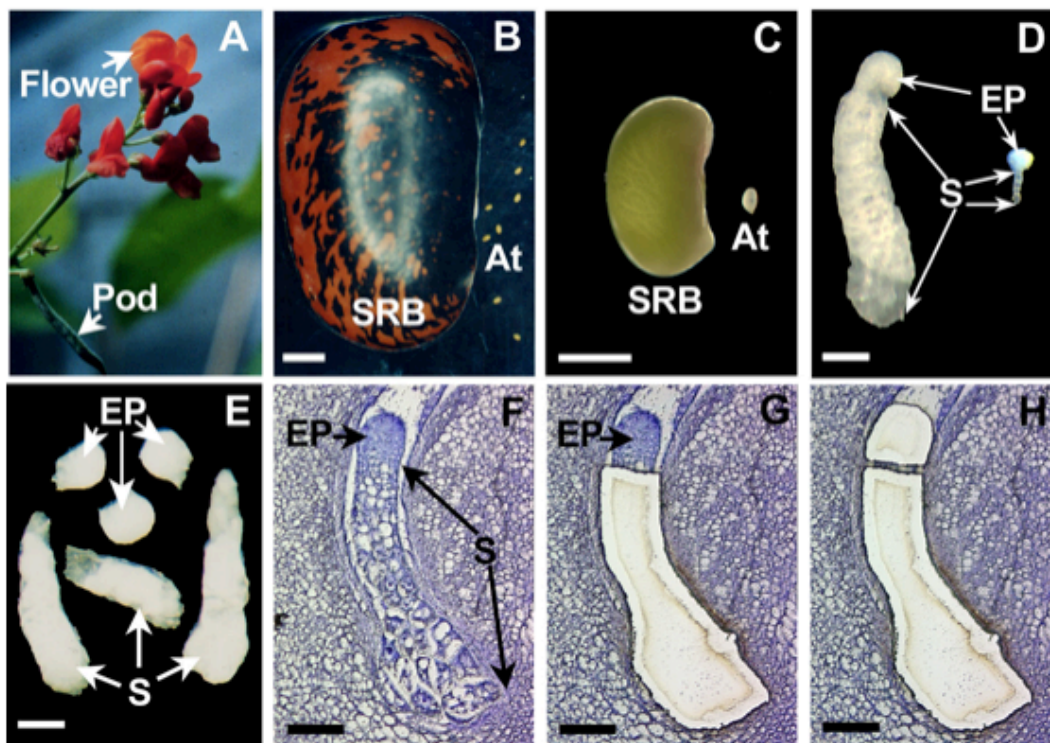
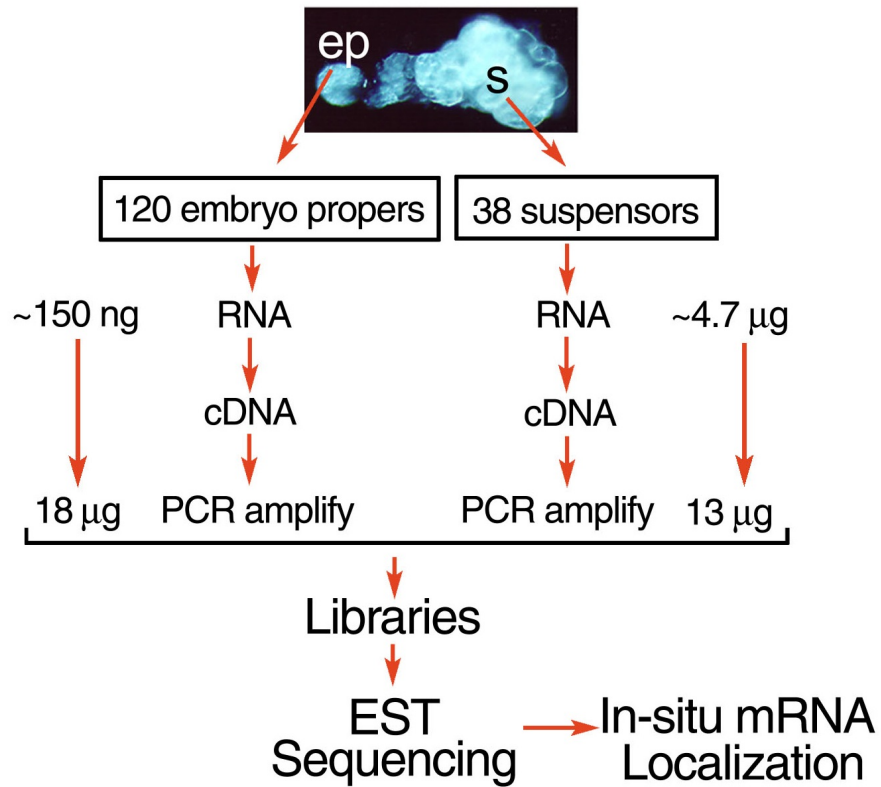


Figure G-2. Scarlet runner bean as a model system to study early embryo development.

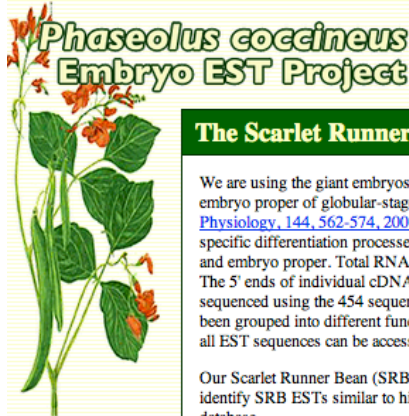
Scale bars: 1mm for panels B and C; 100 um for panels D-H, At, Arabidopsis thaliana, EP, embryo proper, S, suspensor, SRB, scarlet runner bean.

# Scarlet Runner Bean



**Figure G-3. Strategy for discovering genes active during early plant embryo development.** ep, embryo proper; s, suspensor.





Home

Search

### The Scarlet Runner Bean (*Phaseolus coccineus*) EST Project

We are using the giant embryos of the Scarlet Runner Bean (*Phaseolus coccineus*) to identify genes that are active in the suspensor and embryo proper of globular-stage embryos shortly after fertilization ([Weterings et al., Plant Cell, 13, 2409-2425, 2001](#); [Le et al., Plant Physiology, 144, 562-574, 2007](#); [Kawashima et al., PNAS, 106, 3627-3632, 2009](#)). Our long-term goal is to understand the region-specific differentiation processes that occur during early embryo development and how genes are activated specifically in the suspensor and embryo proper. Total RNAs isolated from hand-dissected suspensor and embryo proper were used to construct cDNA libraries. The 5' ends of individual cDNA clones from each library were sequenced using the Sanger sequencing method whereas random cDNA fragments were sequenced using the 454 sequencing technology. To date, we have sequenced 305,363 suspensor and 85,776 embryo proper ESTs. These ESTs have been grouped into different functional categories based on BLAST searches and organized into an EST relational database on our lab server. Moreover, all EST sequences can be accessed directly from NCBI ([GenBank Accession Series CA896559 to CA916678 and GD289845 to GD660862](#)).



Our Scarlet Runner Bean (SRB) globular-stage suspensor and embryo proper EST database is open to the scientific community. Therefore, anyone can identify SRB ESTs similar to his/her experimental DNA sequence(s) by performing BLASTN and/or TBLASTX searches against sequences in our EST database.

[Click here](#) to browse or BLAST your sequences against the Scarlet Runner Bean EST database.

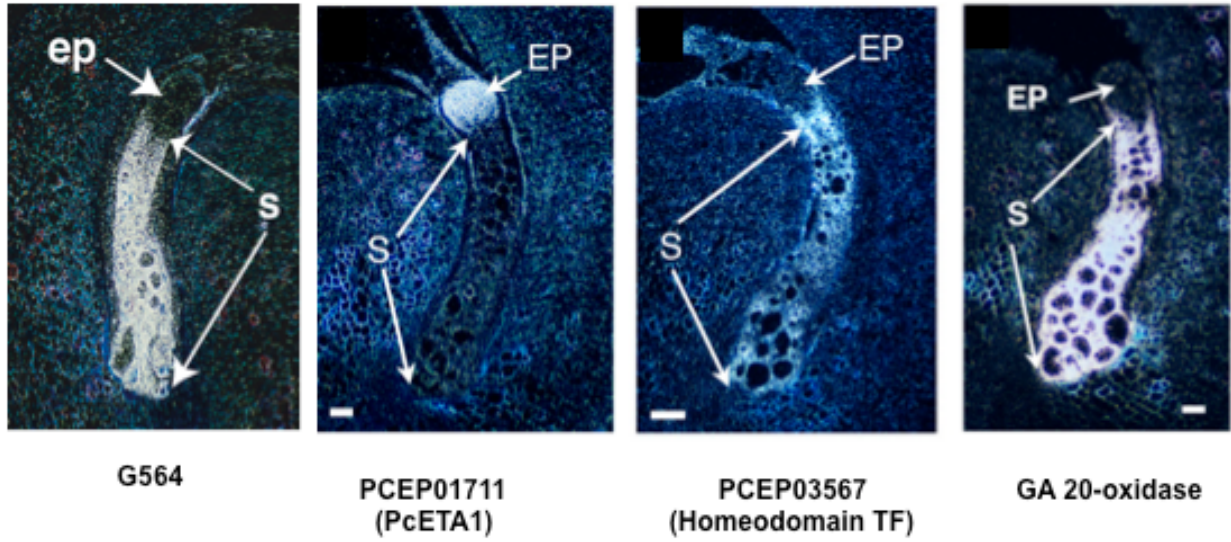
#### Summary of Embryo EST Project

*This table includes the number of ESTs identified in each functional category from the suspensor and embryo proper library. Functional categories were assigned based on BLAST homologies to other genes. The e-value cut-off to consider if a blast hit was significant is 1e-04.*

*Abbreviations: EP - Embryo Proper; S - Suspensor*

FUNCTION	EP	S
Metabolism	3,451	13,839
Energy	2,658	9,598
Cell Growth/Division	492	1,503
Transcription	1,385	4,254
Post-transcription	1,170	3,126
Protein Synthesis	15,074	31,710
Protein Destination and Storage	3,442	12,324
Transporters	2,027	6,878
Intracellular Traffic	851	3,475
Cell Structure	2,410	7,411
Signal Transduction	1,021	3,611
Disease/Defense	1,591	12,605
Secondary Metabolism	2,092	12,194
Transposons	21	52
Unknown or Unclassified Proteins	5,762	23,202
No Significant Homology to Public Databases	42,329	159,581
Total ESTs	85,776	305,363

**Figure G-4. Phaseolus coccineus embryo EST website (<http://estdb.biology.ucla.edu/PcEST>).** This website contains a summary of all 400,000 ESTs grouped by functional categories assigned using the ESTDB sequence analysis server (Appendix H). DNA sequence for each ESTs along with annotations is accessible here.

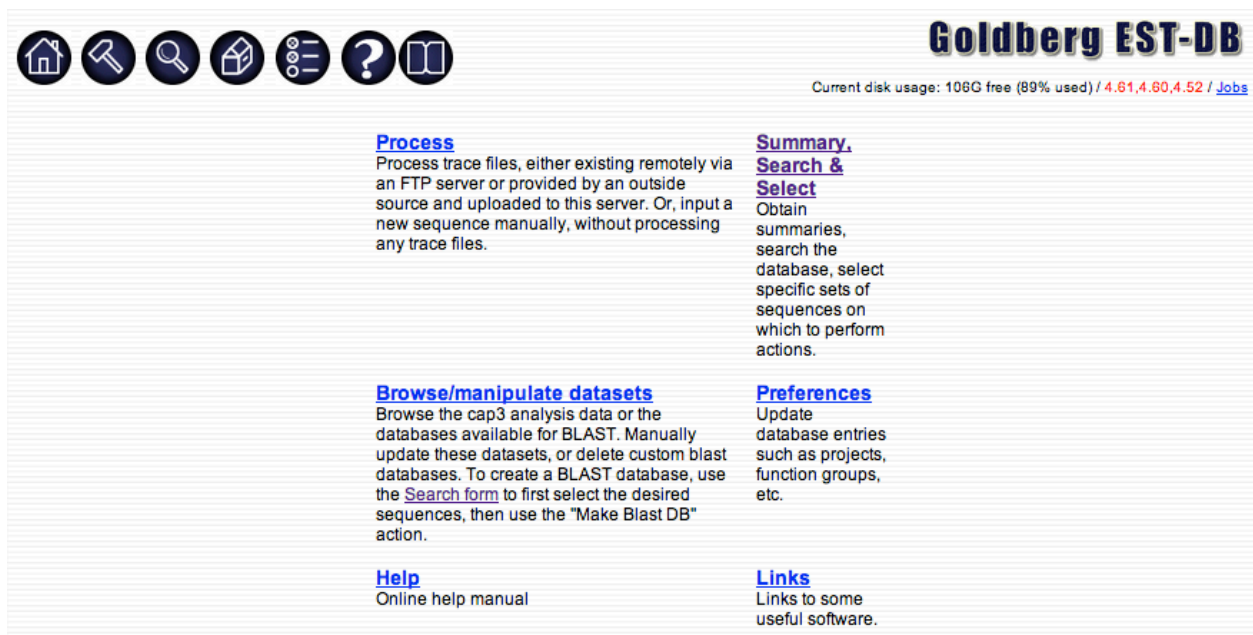


**Figure G-5. mRNA localization of selected molecular makers by in-situ hybridization analysis.** Exposure time varies from one day to one month for the different markers. Scale bar = 100  $\mu$ m. EP, embryo proper; S suspensor.

## APPENDIX H

ESTDB - A Web-Based Relational Database For Analyzing And Storing DNA Sequences

This project involved the development of a web-based relational database for the high-throughput analysis of Sanger sequences from the SRB EST sequencing project (Appendix G) with Anhthu Bui, Harry Hahn, and Bob Goldberg. Several function of this web-based sequence analysis toolkit includes batch download of sequence data directly from the sequencing facility, batch BLAST of EST sequences against publicly available databases (e.g. NCBI non-redundant database), storing BLAST results and sequence annotations, and assembly of ESTs into clusters using the CAP3 contig assembly program. I worked closely with Harry Hahn, our system administrator, to develop and add the necessary tools needed for sequence analysis. Below are screenshots from the web site highlighting the functionality built-in for sequence analysis (**Figures H1 to H7**).



**Goldberg EST-DB**

Current disk usage: 106G free (89% used) / 4.61.4.60.4.52 / [Jobs](#)

**Process**  
Process trace files, either existing remotely via an FTP server or provided by an outside source and uploaded to this server. Or, input a new sequence manually, without processing any trace files.

**Summary, Search & Select**  
Obtain summaries, search the database, select specific sets of sequences on which to perform actions.

**Browse/manipulate datasets**  
Browse the cap3 analysis data or the databases available for BLAST. Manually update these datasets, or delete custom blast databases. To create a BLAST database, use the [Search form](#) to first select the desired sequences, then use the "Make Blast DB" action.

**Preferences**  
Update database entries such as projects, function groups, etc.

**Help**  
Online help manual

**Links**  
Links to some useful software.

**Figure H-1. ESTDB Sequence Analysis website (<http://estdb.biology.ucla.edu/~goldberg>).**

The “Process” link will bring up a series of forms that will guide the user through the sequence processing steps. The “Summary, Search & Select” link allow users to browse through the database or view a summary of the ESTs. The “Browse/manipulate datasets” link allow users to generate/delete custom BLAST databases for local BLAST searches as well as generating/deleting CAP3 contig assembly runs. The “Preferences” link allow the user to modify customizable annotation parameters (e.g., functional categories, TF family).



## Process files from UCLA sequencing facility

In order to process files from the UCLA sequencing facility, you must first download the files from [WebSeq](#). The file, containing all of the sequences you wish to process should be called "webseq.zip". Please use the following to upload your webseq.zip to estdb for processing by clicking the "Browse" button, then navigating to the location of webseq.zip on your computer. Please note that the upload and unzipping process make take a bit of time. Please be patient and do not cancel the process prematurely.

You may also use this form to upload a single (non-zipped) sequence file.

No file chosen

## Process 454 file

In order to process 454 sequencing file, you must enter both sequence file and primer file.

please enter 454 file

No file chosen

## Process local files

Select this item if you wish to process files not located at UCLA's sequencing facility. These files should have been uploaded to [estdb.biology.ucla.edu](http://estdb.biology.ucla.edu) via ftp prior to selecting this option.

## Create new sequence manually

Use the following form if you would like to create a new sequence entry for the database manually, rather than processing a trace file using one of the options above. After you submit the form below, you will be taken to the sequence annotation page for the new sequence.

**Please keep in mind the sequence naming conventions used! For example, suspensor sequences are of the form PCS#####.**

: Sequence name

**Figure H-2. ESTDB Sequence Analysis website -- Sequence Processing.** This form initiates the processing of Sanger data with sequence chromatogram, 454 sequence only files, or local files on the user's desktop. Subsequent pages will have step by step guide for processing batch EST sequences.



## Summaries

View Summary:

## Search Database

Use the following form to select sequences based on various criteria.

1351039 sequences, averaging 182 bases in length, with a std dev of 101.2768053290759048

Name	from <input type="text"/> to <input type="text"/>
	contains <input type="text"/>
Date	after <input type="text"/> before <input type="text"/> (eg 12-31-2000)
Length	less than <input type="text"/> greater than <input type="text"/>
Session	<input type="text"/>
Projects	<input type="text" value="Choose"/>
Vectors	<input type="text" value="Choose"/>
Primers	<input type="text" value="Choose"/>
Origins	<input type="text" value="Choose"/>
Function Groups	<input type="text" value="Choose"/>
Homologous	<input type="text"/>
References	<input type="text"/>
Arabidopsis hit?	<input type="radio"/> Yes <input type="radio"/> No <input checked="" type="radio"/> Doesn't matter
Poly-A tail?	<input type="radio"/> Yes <input type="radio"/> No <input checked="" type="radio"/> Doesn't matter
Accession #	<input type="text"/>
DBESTID	<input type="text"/>

View

Show  hits  
(selecting "ALL" may result in queries that take a long time)

**Figure H-3. ESTDB Sequence Analysis website -- Summary and Search Form.** This portal allow users to get a quick summary of the sequence datasets in the database based on project or search through the database using a variety of searchable parameters (e.g. functional categories, EST ID, sequence lengths, etc).



## Summary

Project :

Links take you to  view, showing  at a time

Category	Type	# sequences	Avg len	Std Dev
<b>Projects</b>	<a href="#">Suspensor</a>	4379	430	140.950608978216
<b>Vectors</b>	<a href="#">pCR2.1</a>	81	394	159.673338413149
	<a href="#">pTriplEx2</a>	4298	431	140.504791742903
<b>Primers</b>	<a href="#">5' TriplEx</a>	4291	431	140.462135558357
	<a href="#">M13F</a>	3	389	220.599939558771
	<a href="#">M13R</a>	68	413	156.227841624606
	<a href="#">n.d.</a>	5	400	178.880686492422
	<a href="#">TZ</a>	12	290	136.891131261970
<b>Function groups</b>	<a href="#">01.2 Cytochrome P450</a>	19	460	141.944557169134
	<a href="#">01 Metabolism</a>	282	472	130.233078424544
	<a href="#">02 Energy</a>	219	445	131.108510075694
	<a href="#">03 Cell Growth/Division</a>	49	422	139.094707250255
	<a href="#">04.1.01 AP2/EREBP</a>	4	459	141.315250415516
	<a href="#">04.1.02 Basic Helix-Loop-Helix (bHLH)</a>	3	389	157.608163917144
	<a href="#">04.1.03 Basic Leucine Zipper (bZIP)</a>	6	439	243.314953643763
	<a href="#">04.1.04 CCAAT</a>	2	579	196.575685169860
	<a href="#">04.1.05 General</a>	8	456	77.4208120782593113
	<a href="#">04.1.07 Homeodomain</a>	3	588	43.6157463920237565
	<a href="#">04.1.08 MADS-Box</a>	6	375	185.356683181373
	<a href="#">04.1.09 MYB</a>	4	428	196.515266582523
	<a href="#">04.1.10 NAC Domain</a>	3	421	111.446549221290
	<a href="#">04.1.11 WRKY</a>	3	479	58.7565599174537560
	<a href="#">04.1.12 Zinc Finger-Others</a>	3	380	111.383721132549
	<a href="#">04.1.13 Unclassified</a>	20	336	153.131616935443
	<a href="#">04.1.14 B3 Domain</a>	3	268	122.021855965779
	<a href="#">04.1.15 Aux/IAA</a>	9	387	153.582044668133
	<a href="#">04.1.17 Alfin-like</a>	1	281	
	<a href="#">04.1.18 ARF</a>	1	574	
	<a href="#">04.1.19 Zf-C2C2</a>	2	429	5.6568542494923802
	<a href="#">04.1.20 Zf-C3H</a>	3	535	27.6103845198384250
<a href="#">04.1.21 CPP</a>	4	417	91.1884495609687035	
<a href="#">04.1.22 E2F-DP</a>	1	553		

**Figure H-4. ESTDB Sequence Analysis website -- Project Summary.** The summary page provides a brief summary of all sequences based on project. For each project, ESTs are further divided by additional categories including functional categories with summaries of total ESTs and EST length.





## Edit Sequence Record for

Hide BLAST/PFAM data | [\(Jump to blast result summary\)](#) | [\(Jump to local blast\)](#) | [\(Jump to pfam result summary\)](#)

Perform action:

```
CTGTTGGNCCTCNCNTTAAAGGAGGCTAACAACTTTCTGNGGC
CATTAAAGCTGAAGGCTCCTTTGGGTGGTTGAAGAAGAAGAG
AANCCCTTTATGTTGAGGGAGGTGATGCTGGAACCCGGGAGA
ATTACATCAATGAGCTAATCAGGAGAATGAATTGAGCTGGC
ACTGTTCAATCTTTGACAAGCTGAATCACTATTTCTTGNTC
GCTTTTTAGATCAATATCAGACAATATTTGCATCTGAAATTT
GAGAATAGCNCATTTATGACTTCGATAAATCTTTTAAAGTTTG
AATGTTTTTCGTTTATAAGTATGTTCCGGTATTTTAAATTT
AAGGNATTGCTAGTCATATGCTCCCTTTAATGATAAGATTT
CTTTTGGAAATTTGAATTTTCTTTTGGGAGTCGAAAAA
AAAAAAAAAAAA
```

Length : 441 # Ns : 7  
 Date entered : 09/29/1998 Date modified : 12/10/2002  
 Session : S00101-200

Project	Suspensor
Vector	pCR2.1
Primer	M13R
Sequenced from	5' end
cDNA size	1200
Function group	05 Protein Synthesis
Most homologous	60S ribosomal protein L7
Origin	Plant
Arabidopsis hit?	<input type="radio"/> Yes <input checked="" type="radio"/> No
poly-A tail?	<input checked="" type="radio"/> Yes <input type="radio"/> No
Accession #	CA899889
DBESTID	16529363

Part of the contig containing PCS166, PCS4162, PCEP3500, PCS5016, PCS5177.  
 AB (10-31-01)

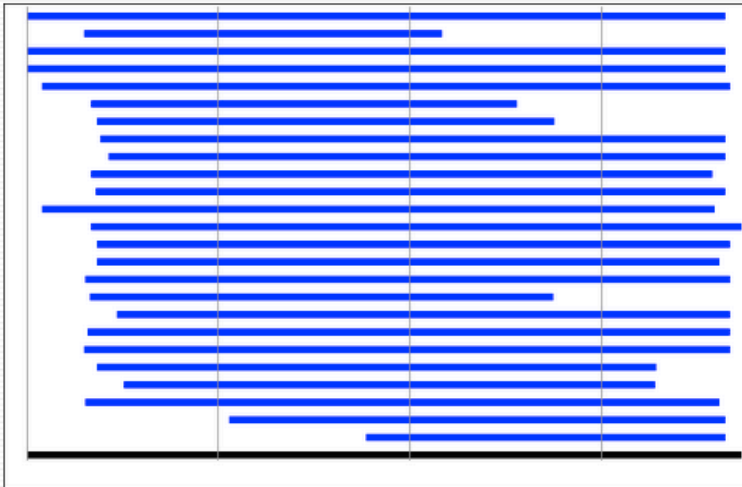
References

### Blast results

Best evalule: 2e-17

Blast run	Hits (click to see alignment)	E-value	NCBI record
<b>blastx -d nr -b 10 -v 10</b> • Run on 07/25/2002 • <a href="#">See full results</a> • <a href="#">Discard this run</a>	<a href="#">ribosomal protein L7 [Solanum tuberosum]</a>	2e-17	<a href="#">see NCBI record</a>
	<a href="#">(NM_126186) putative ribosomal protein L7 ...</a>	2e-17	<a href="#">see NCBI record</a>
	<a href="#">(NM_112204) ribosomal protein, putative [A...</a>	2e-17	<a href="#">see NCBI record</a>
	<a href="#">(AY052724) A12g44120/F6E13.25 [Arabidopsis L...</a>	2e-17	<a href="#">see NCBI record</a>
	<a href="#">(NM_129975) 60S ribosomal protein L7 [Arab...</a>	2e-17	<a href="#">see NCBI record</a>
	<a href="#">ribosomal protein L7 - human &gt;gi 1335288 emb CA...</a>	2e-13	<a href="#">see NCBI record</a>
	<a href="#">60S RIBOSOMAL PROTEIN L7 &gt;gi 19483861 g...</a>	2e-13	<a href="#">see NCBI record</a>
	<a href="#">(M85235) ribosomal protein [Mus musculus]</a>	2e-13	<a href="#">see NCBI record</a>
	<a href="#">(XM_087652) similar to 60S ribosomal prote...</a>	2e-13	<a href="#">see NCBI record</a>
	<a href="#">(L16558) ribosomal protein L7 [Homo sapiens]</a>	2e-13	<a href="#">see NCBI record</a>

**Figure H-5. ESTDB Sequence Analysis website -- Sequence Record Page.** The sequence record page is unique for each EST based on the sequence name. This page stores sequence data, chromatogram, and allows user to manually annotate sequence information. Users can run local BLAST against a remote copy of the NCBI non-redundant database and all BLAST run are stored for each EST.



[See boxshaded alignment](#)

### Contig 1009 from set SuspXSuspensor\_Ceres\_6-11-02

- PCSC08832-
  - PCS05280\_3'- is in PCSC08832-
  - PCSC08938- is in PCSC08832-
  - PCSC08922- is in PCSC08832-
- PCSC09254-
  - PCSC09131- is in PCSC09254-
  - PCSC05735- is in PCSC09254-
  - PCSC07668- is in PCSC09254-
  - PCSC00738X5- is in PCSC09254-
  - PCSC10280- is in PCSC09254-
  - PCSC07965- is in PCSC09254-
  - PCSC08379- is in PCSC07965-
  - PCSC11520- is in PCSC09254-
- PCSC09042-
  - PCS02978\_3'- is in PCSC09042-
  - PCS03000\_3'- is in PCS02978\_3'-
  - PCS01382X5- is in PCSC09042-
  - PCSC08724- is in PCS01382X5-
  - PCS01317X5\_3'- is in PCS01382X5-
  - PCS05947- is in PCS01382X5-
  - PCS01757X5- is in PCSC09042-
  - PCS03354- is in PCSC09042-
  - PCSC08726- is in PCS03354-
  - PCSC08429- is in PCS03354-
  - PCS00322X5\_3'- is in PCS03354-

```

PCSC08832-      . . . . .
                TTTTTTTTTTTTTTTTTTTTCTGAGTTGAAATTGAAATTGAAAGACTTTAT
PCSC05280_3'-      TGAGTTGAAATTGAAATTGAAAGACTTTAT
PCSC08938-      TTTTTTTTTTTTTTTTTTTTCTGAGTTGAAATTGAAATTGAAAGACTTTAT
PCSC08922-      TTTTTTTTTTTTTTTTTTTTCTGAGTTGAAATTGAAATTGAAAGACTTTAT
PCSC09254-      TTTTTTTTTTTTTTTTTTTTCTGAGTTGAAATTGAAATTGAAAGACTTTAT
PCSC09131-      TTTGAAATTGAAATTGAAAGACTTTAT
PCSC05735-      TTTTGGAAATTGAAAGACTTTAT
PCSC07668-      ATTGAAATTGAAAGACTTTAT
PCSC00738X5-      AATTTGAAAGACTTTAT
PCSC10280-      TTGAAATTGAAATTGAAAGACTTTAT
PCSC07965-      GAAATTGAAATTGAAAGACTTTAT
PCSC11520-      TTTTTTTTTTTTTTTTTTATTTTTTGAATTGAAATTGAAAGACTTTAT

```

Figure H-6. ESTDB Sequence Analysis website -- Contig Assembly Summary. EST sequences were assembled into contigs using the CAP3 contig assembly program. An image overview showing overlap of each EST in the contig. Below the image, a list of all the ESTs associated with the contig, the order of the EST alignment, and a multiple sequence alignment view of ESTs in the contig.



```

PCSC08832- TTTTTTTTTTTTTTTTTTTTTTTTTTTTTCTGAGTTGAAA TTGAAATTGAAAGACTTTAT
PCSC05280_3'-----TGAGTTGAAA TTGAAATTGAAAGACTTTAT
PCSC08938- TTTTTTTTTTTTTTTTTTTTTTTTTTTTCNGAGTTGAAA TTGAAATTGAAAGACTTTAT
PCSC08922- -TTTTTTTTTTTTTTTTTTTTTTTTTTTTCTGAGTTGAAA TTGAAATTGAAAGACTTTAT
PCSC09254- -----TTTTTTTTTTTTTTTTTTTTTTTTTTGAAA TTGAAATTGAAAGACTTTAT
PCSC09131- -----TTGAAA TTGAAATTGAAAGACTTTAT
PCSC05735- -----TTT TTGAAATTGAAAGACTTTAT
PCSC07668- -----ATTGAAATTGAAAGACTTTAT
PCSC0738X5- -----AAATTGAAAGACTTTAT
PCSC10280- -----TTGAAA TTGAAATTGAAAGACTTTAT
PCSC07965- -----GAAA TTGAAATTGAAAGACTTTAT
PCSC11520- -----TTTTTTTTTTTTTTTTTTTTTATTTTTTTGAAA TTGAAATTGAAAGACTTTAT
PCSC09042- -----TTGAAA TTGAAATTGAAAGACTTTAT
PCSC02978_3'-----TTT TTGAAATTGAAAGACTTTAT
PCSC03000_3'-----TTT TTGAAATTGAAAGACTTTAT
PCSC01382X5- -----GAGTTGAAA TTGAAATTGAAAGACTTTAT
PCSC01317X5_3'-----GTTGAAA TTGAAATTGAAAGACTTTAT
PCSC05947- -----TGAAAGACTTTAT
PCSC01757X5- -----AGTTGAAA TTGAAATTGAAAGACTTTAT
PCSC03354- -----TTTTTTGAAA TTGAAATTGAAAGACTTTAT
PCSC08726- -----AAA TTGAAATTGAAAGACTTTAT
PCSC08429- -----AGACTTTAT
PCSC00322X5_3'-----TTTTTGAAA TTGAAATTGAAAGACTTTAT
PCSC08379- -----
PCSC08724- -----
consensus t t g a a a t t g a a a g a c t t t a t

PCSC08832- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC05280_3' TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC08938- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC08922- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC09254- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC09131- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC05735- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC07668- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC0738X5- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC10280- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC07965- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC11520- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC09042- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC02978_3' TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC03000_3' TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC01382X5- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC01317X5_3' TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC05947- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC01757X5- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC03354- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC08726- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC08429- TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC00322X5_3' TTATATTTTGTCCCTTTGATTAGGAAGATACAGGAGAA TTATTTATGCAGCAACATAGA
PCSC08379- -----TGCAGCAACATAGA
PCSC08724- -----
consensus t t a t a t t t t g t c c c t t g a t t a g g a a g a t a c a g g a g a a t t a t t t a t g c a g c a a c a t a g a

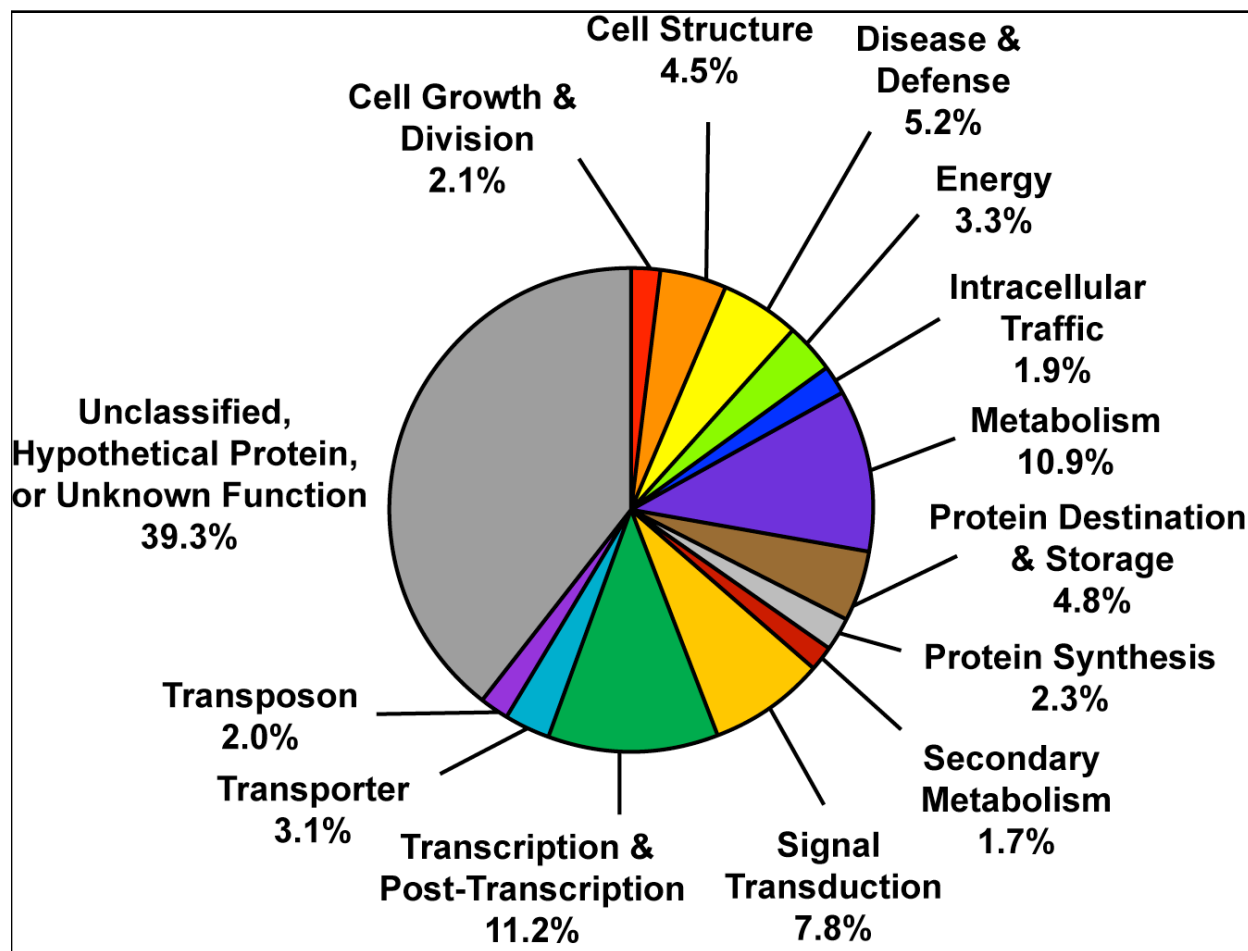
```

Figure H-7. ESTDB Sequence Analysis website -- Multiple Sequence Alignment View. A boxshade view of ESTs assembled using CAP3 and the consensus sequence generated from the assembly.

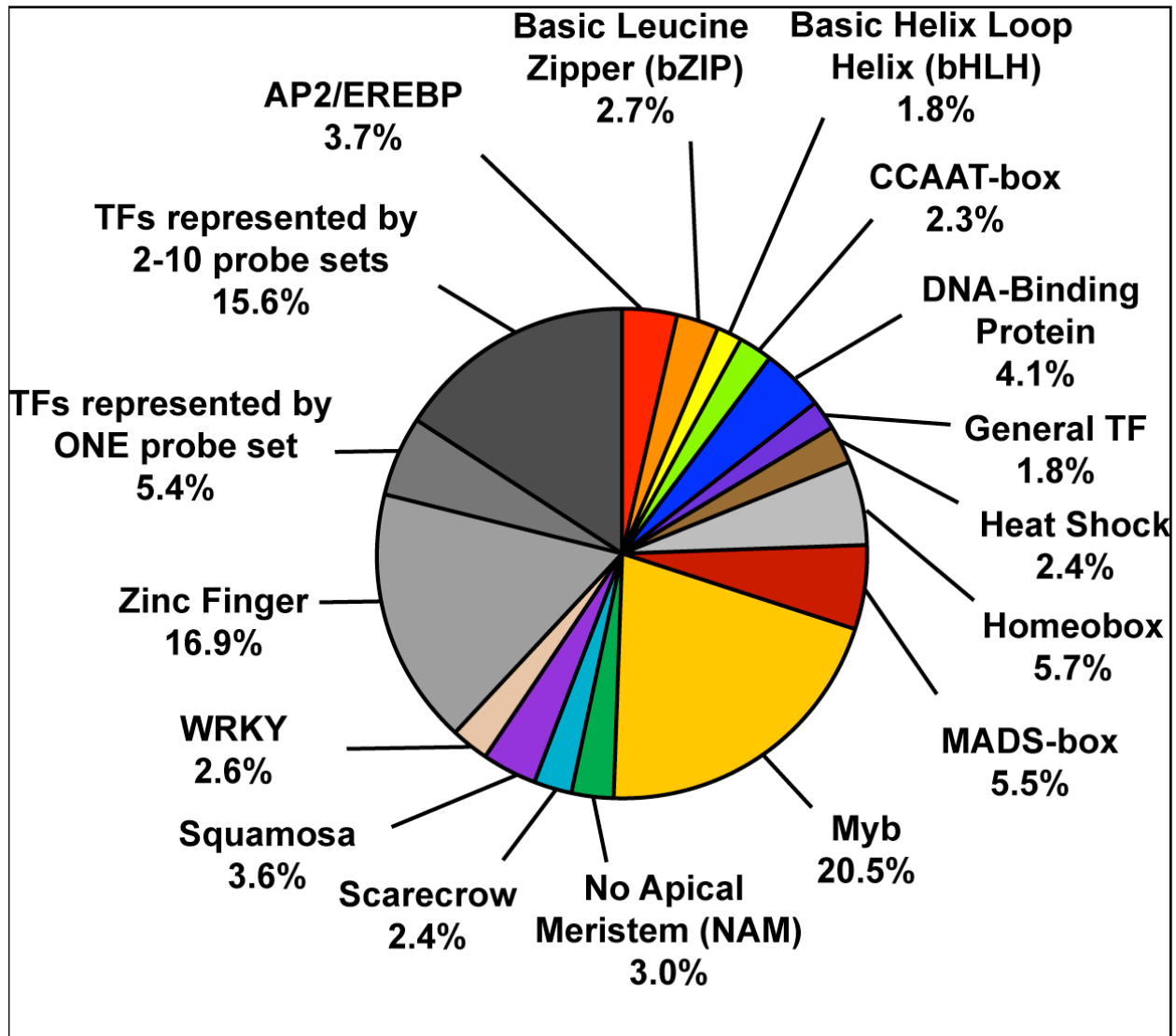
## APPENDIX I

### Functional Annotation of Genes Represented on the *Arabidopsis* and Soybean GeneChip Arrays

This project involved the annotation of every gene feature on the *Arabidopsis* and Soybean GeneChip arrays and the categorization of these features into functional categories including the identification and classification of TF genes based on TF gene family. These annotations and functional categories guided the interpretation and analysis of the GeneChip datasets. The *Arabidopsis* and soybean GeneChip arrays were annotated with the help from Anhthu Bui and Javier Wagmaister. My role included carrying out BLAST analysis of each array sequence against the public sequence databases, assigning annotations and functional categories to each unique sequence feature, and categorizing TFs into TF families. **Figures I-1 to I-6** describes the distribution of features assigned into functional categories on the AtGenome1 (8K) and ATH1-121501 (22K) *Arabidopsis* GeneChip arrays and the Soybean Genome GeneChip array. These annotations and summaries are available through our lab web site for *Arabidopsis* (<http://seedgenenetwork.net/presentation#arabidopsis>) and soybean (<http://seedgenenetwork.net/presentation#soybeanIVT>).



**Figure I-1. Distribution of 8,247 features on the *Arabidopsis* AtGenome1 (8K) GeneChip array into major functional categories.** Sequences were obtained from the Torrey Mesa Research Institute (TMRI) website ([http://www.tmri.org/gene\\_exp/index.html](http://www.tmri.org/gene_exp/index.html)). Sequences were searched against the NCBI non-redundant database using the BLASTX program and annotations were manually assigned based on the BLAST results.



**Figure I-2. Distribution of transcription factor families on the *Arabidopsis* AtGenome1 (8K) GeneChip array.** Sequences encoding transcription factors (TFs) were further classified into major TF families based on sequence annotation and known DNA binding domain. There are 704 features annotated as TFs and assigned into TF families.

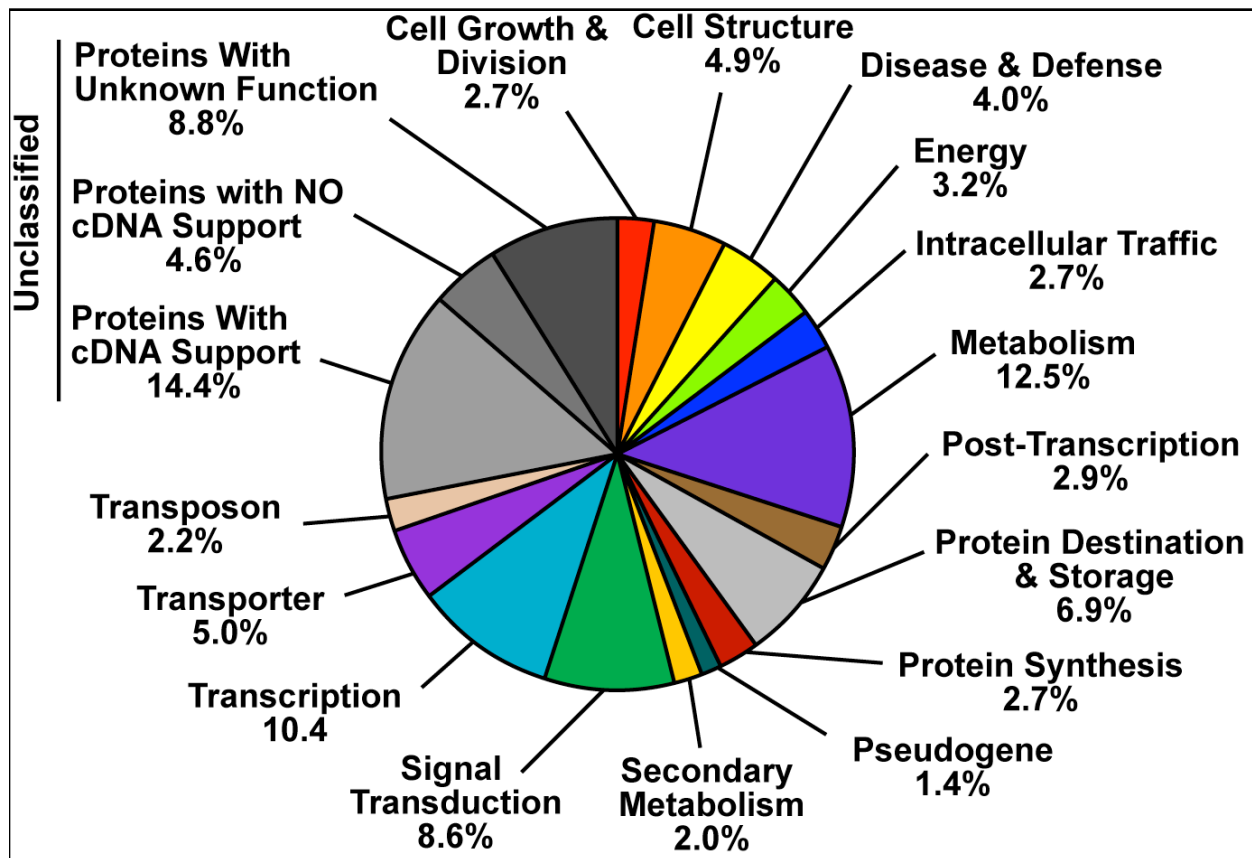
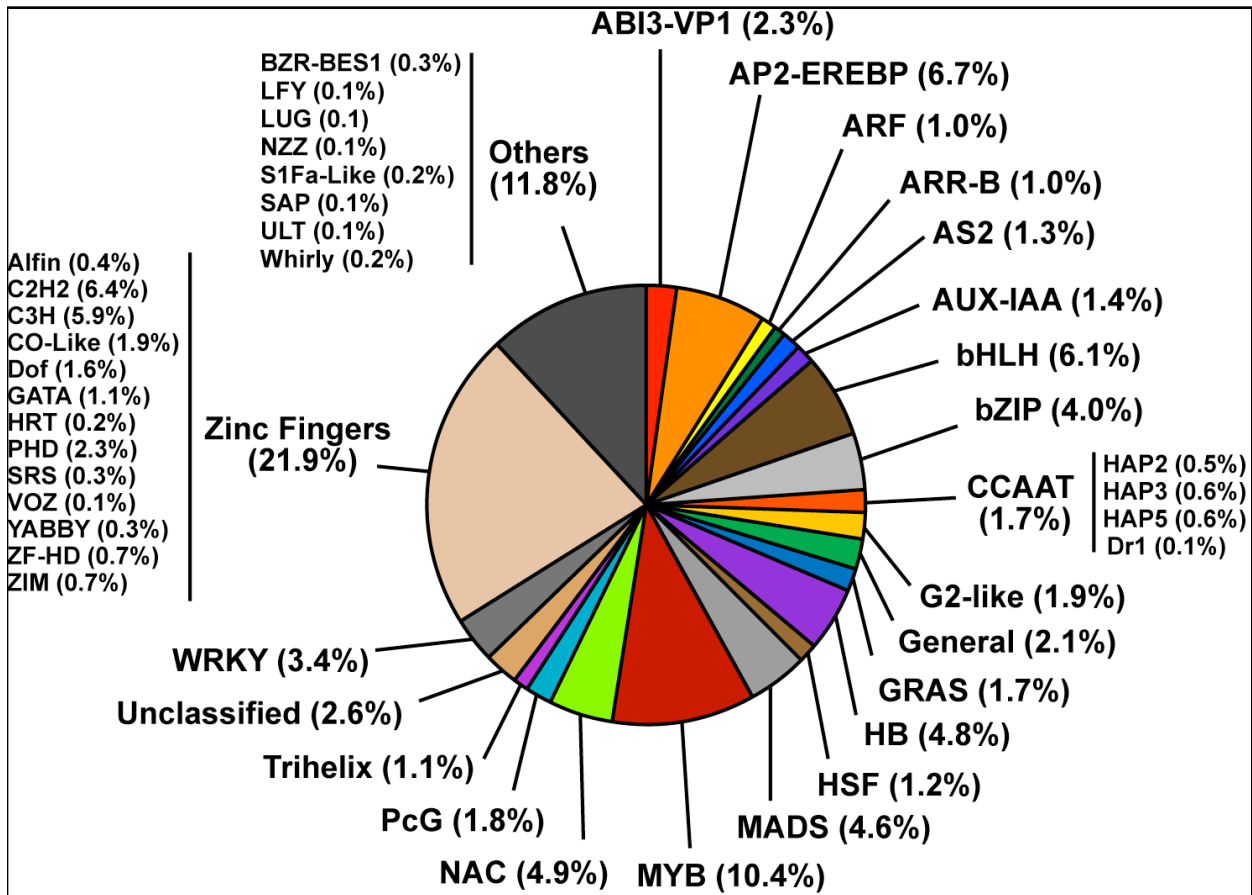


Figure I-3. Distribution of 22,747 features on the *Arabidopsis* ATH1-121501 (22K) GeneChip array into major functional categories. Sequences were obtained from the Affymetrix website (<http://www.affmetrix.com//support/technical/byproduct.affx?product=arab>). Sequences were BLASTed like in Figure I-1 and functional categories were manually assigned for each feature.





**Figure I-4. Distribution of transcription factor families represented on the *Arabidopsis* ATH1-121501 (22K) GeneChip array.** Sequences were assigned into TF families using the methods described in Figure I-2. There are 1,484 features annotated as TF and assigned into TF families.

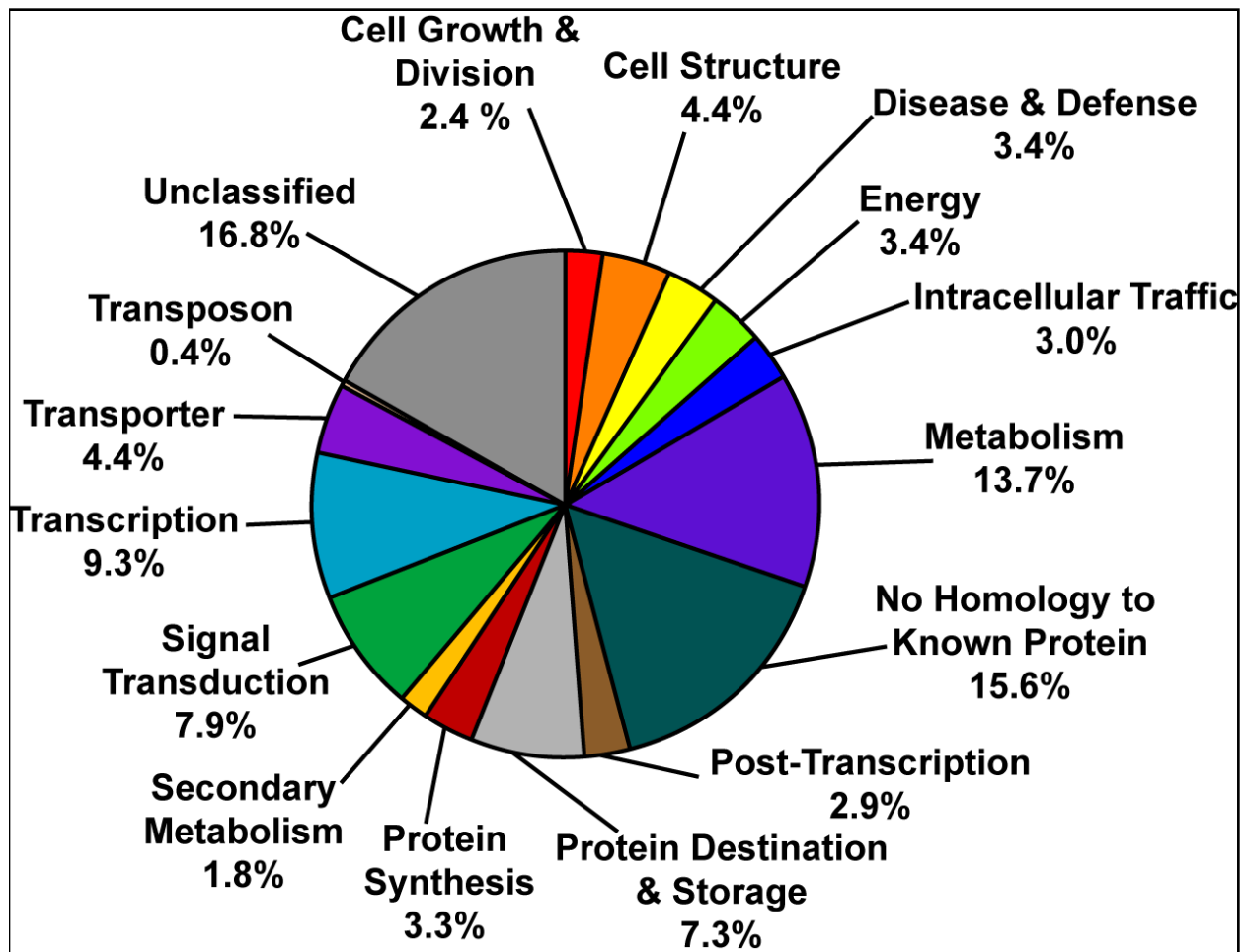


Figure I-5. Distribution of 37,593 features on the Soybean Genome GeneChip array into functional categories. Some sequence annotations and functional categories were assigned based on homology to sequences on the *Arabidopsis* ATH1-121501 (22K) GeneChip array to expedite the annotation process. The remaining sequences were manually annotated as before.

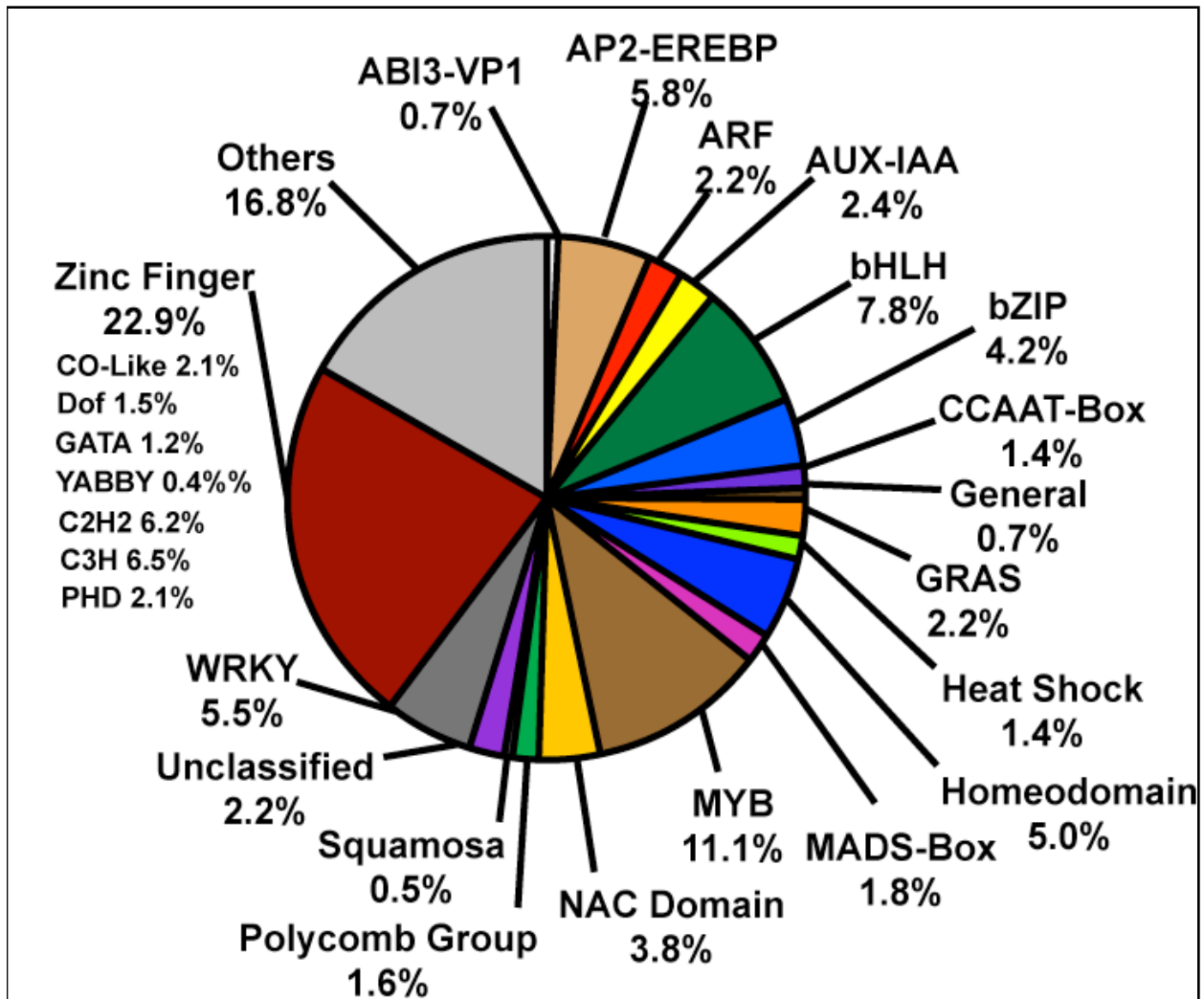


Figure I-6. Assignment of 2,832 features on the Soybean Genome GeneChip array into transcription factor families. Transcription factor families were assigned based on information from the Soybean transcription factor database (Soybean TFDB - <http://plantfdb.cbi.pku.edu.cn/web/index.php?sp=gm>).

## APPENDIX J

Design And Characterization of the Affymetrix Soybean Whole Genome Transcript Array

## **Design of an Affymetrix Soybean Whole Transcript Sense array**

### **Background**

The Soybean Whole Transcript (WT) Sense array interrogates all the genes in the genome. The first generation Affymetrix Soybean Genome array was designed by the Soybean Consortium using publicly available soybean full-length cDNAs and ESTs. The Soybean Genome array consists of 37,000 probe sets interrogating ~ 25,000 distinct genes/transcripts. The release of the whole genome sequence of soybean (available at [Phytozome.net](http://Phytozome.net)) allowed the creation of an array that can survey all the genes (both high and low confidence gene models) in the genome.

### **Design**

The design of the Soybean WT array is different from the Soybean Genome array. For the Soybean Genome array, probes were selected to correspond to the 3' end of the transcript or cDNA. However, for the Soybean WT array, probes were selected to span every exon of the predicted gene models/transcripts, if possible. This approach allows for the interrogation of the transcript (from 5' to 3') and can help determine exon usage in different splice variants, alternative polyadenylation sites, etc. For information regarding this array design, please check out other references from Affymetrix ([http://media.affymetrix.com:80/support/technical/technotes/gene\\_1\\_0\\_st\\_technote.pdf](http://media.affymetrix.com:80/support/technical/technotes/gene_1_0_st_technote.pdf)).

### **Sequence Data**

All sequence data used to design probes on the array were obtained from the Department of Energy - Joint Genome Institute (DOE-JGI) web site (phytozome: <http://phytozome.net>). Probes were designed from the first draft assembly of the soybean genome1 (version 1.0). The probe selection algorithm was developed by Christopher Davies and Brant Wong at Affymetrix.

## **Probe Selection**

Probes were selected to interrogate one transcript only, although some probes might map to multiple transcripts (if no unique probes can be obtained for that exon region). Each probe are 25-mer in length. There are 1,221,261 probes designed to target 66,195 gene models within the soybean genome. Given that the soybean genome undergone several genome duplication events, not all the ~ 1.2 million probes will uniquely hybridize to a single gene. Each gene is targeted to be represented by ~ 25 probes although most genes will have less than 25 probes.

## **Source Files**

Unlike the Soybean Genome array where a probeset ID is assigned to each cDNA or transcript, each probe, exon, and gene model included in the Soybean WT array has a distinct Affymetrix identifier. To simplify downstream analysis of data generated from this array, a probe association file was created that listed and outlined the gene model, exon, associated probes, and the relevant probe sequence (<http://seedgenenetwork.net/presentation#soybeanWT> ). An illustration showing the association of the probes to the predicted gene model is shown in **Figure J-1**.

## **Contributors to the Design of the Array**

The Soybean WT Genome array was designed in a collaboration with the Goldberg Lab and Affymetrix with advice and suggestions from other members of the soybean community, including Randy Shoemaker.

### **Goldberg Lab**

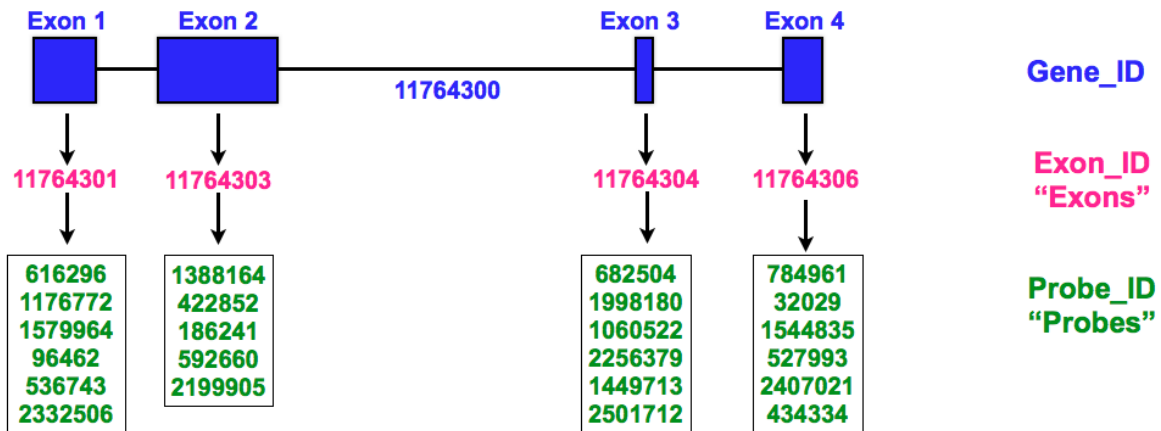
Bob Goldberg, Brandon Le, Chen Cheng, Min Chen, and Anhthu Bui

### **Affymetrix**

Gene Tanimoto, Christopher Davies, Stan Trask, Brant Wong, Eric Schell, Xue Mei Zhou, and Patricia Chan

**A**

Gene_Model	Gene_ID	Genomic_Start	Genomic_End	Exon_ID	Probe_ID	Probe_Start	Probe_End	% GC	Probe_Seq_On_Array
Glyma01g00320	11764300	116300	127990	11764301	96462	116394	116418	11	GCAACATCACATATAGGACTTAGGG
Glyma01g00320	11764300	116300	127990	11764301	536743	116410	116434	12	GACTTAGGGCTAGCGTCTTTATCAC
Glyma01g00320	11764300	116300	127990	11764301	616296	116414	116438	11	TAGGGCTAGCGTCTTTATCACAATC

**B**

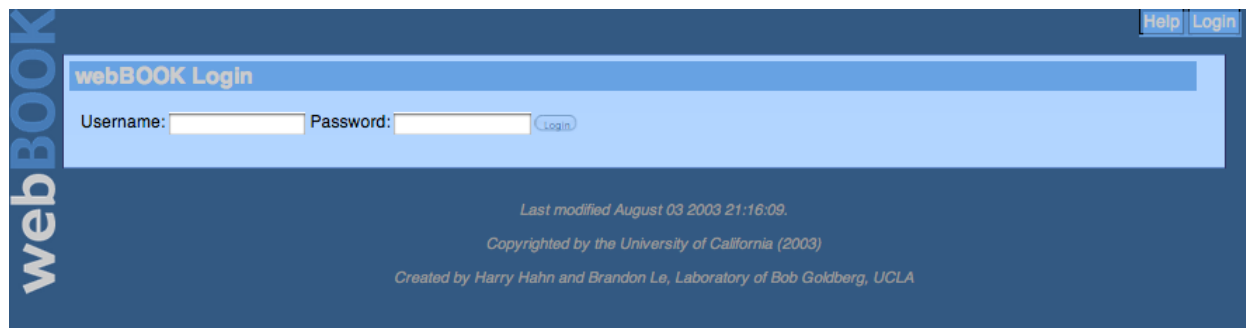
**Figure J-1. Illustration of the association between probe id and the predicted gene model.** (A) A snapshot of information in the probe association table highlighting the association of Affymetrix assigned identifiers (gene\_id, exon\_id, probe\_id) with the predicted gene models. (B) A cartoon showing the organization of the Affymetrix identifiers (gene\_id, exon\_id, probe\_id) for the gene model, Glyma01g00320.

## APPENDIX K

Webbook - A Web-Based Lab Notebook As An Undergraduate Teaching Tool



This project involved the development of a web-based database that serves as an electronic lab notebook for teaching undergraduates how to do research. The tool was developed as part of Bob Goldberg's undergraduate lab course, Honors Collegium 70AL - Gene Discovery Laboratory. This web database stores students data entry and house protocols for carrying out research at the bench. Students uploaded data (e.g., gel images) are stored for each experiment and is easily accessible. The webbook was programmed by Harry Hahn with original input design provided by myself, Anhthu Bui, and Bob Goldberg. My role was providing leadership in the design, maintenance, and implementation of the site. Below are several screenshots showing the functionality of the webbook (**Figures K-1 to K-5**).



**Figure K-1. Webbook website** (<http://estdb.biology.ucla.edu/webbbok>). Each user is assigned a username and password to add personalization of each data entry.

webBOOK

[Home](#)
[Admin](#)
[Projects](#)
[Stocks](#)
[Protocols](#)
[Calendar](#)
[Browse](#)
[Help](#)
[Contact](#)
[Logout](#)

## Projects

- I. [In-situ Hybridization Analysis of Homeodomain Transcription Factor](#) — [Edit] — *ble*
  - 1. [Experiment Linearization of Plasmid DNA for Run-off Transcription](#) — [Edit]
    - » [Create a new experiment for project \*In-situ Hybridization Analysis of Homeodomain Transcription Factor\*](#)
- II. [Searching for a knockout of the SEL-1 ortholog in Arabidopsis](#) — [Edit] — *ble*
  - » [Create a new experiment for project \*Searching for a knockout of the SEL-1 ortholog in Arabidopsis\*](#)
- III. [Search for Knock-out in AT2G32370](#) — [Edit] — *emcali14*
  - 1. [Experiment Determining the size of At2G32370](#) — [Edit]
  - 2. [Experiment Superpool identification of At2G32370](#) — [Edit]
  - 3. [Experiment DNA blot and preparation for hybridization](#) — [Edit]
  - 4. [Experiment Prehybridizing the blot](#) — [Edit]
  - 5. [Experiment Generation of Radioactive probes using random prime labeling kit](#) — [Edit]
  - 6. [Experiment Hybridization of the probe to the blot and autoradiography exposure of the blots](#) — [Edit]
  - 7. [Experiment Reamplification and Gel Electrophoresis of the Reverse amplified superpools 15, 16, 17 and a control](#) — [Edit]
  - 8. [Experiment Second round of Identifying Line with At2G32370 knockout-9 DNA subpools from superpool 16 from Maddison Facility](#) — [Edit]
  - 9. [Experiment Confirmation of seed pool with T-DNA insert](#) — [Edit]
  - 10. [Experiment Gel purification of seed pool CSJ3451 for sequencing](#) — [Edit]
  - 11. [Experiment Screen CSJ 3451 for T-DNA insert in At2g32370 \(Plants 1-48\)](#) — [Edit]
  - 12. [Experiment Screen CSJ 3451 for T-DNA insert in At2g32370 \(plants 92-96\)](#) — [Edit]
  - » [Create a new experiment for project \*Search for Knock-out in AT2G32370\*](#)
- IV. [Knocking out the Atthyreceptor Gene](#) — [Edit] — *cristycross*
  - » [Create a new experiment for project \*Knocking out the Atthyreceptor Gene\*](#)
- V. [Searching for a knockout in arabidopsis for At5g52820](#) — [Edit] — *jtimpson*
  - » [Create a new experiment for project \*Searching for a knockout in arabidopsis for At5g52820\*](#)
- VI. [Search for Knockout Punch in Arabidopsis At1g58100](#) — [Edit] — *nimaj*
  - » [Create a new experiment for project \*Search for Knockout Punch in Arabidopsis At1g58100\*](#)
- VII. [Searching for a knockout in the PPan-like gene of arabidopsis](#) — [Edit] — *lcadams*
  - 1. [Experiment Fractionation of PPan PCR products by Gel Electrophoresis](#) — [Edit]
    - » [Create a new experiment for project \*Searching for a knockout in the PPan-like gene of arabidopsis\*](#)
- VIII. [Uncovering the mystery of the ATEY2 gene.](#) — [Edit] — *star7bs*
  - » [Create a new experiment for project \*Uncovering the mystery of the ATEY2 gene.\*](#)

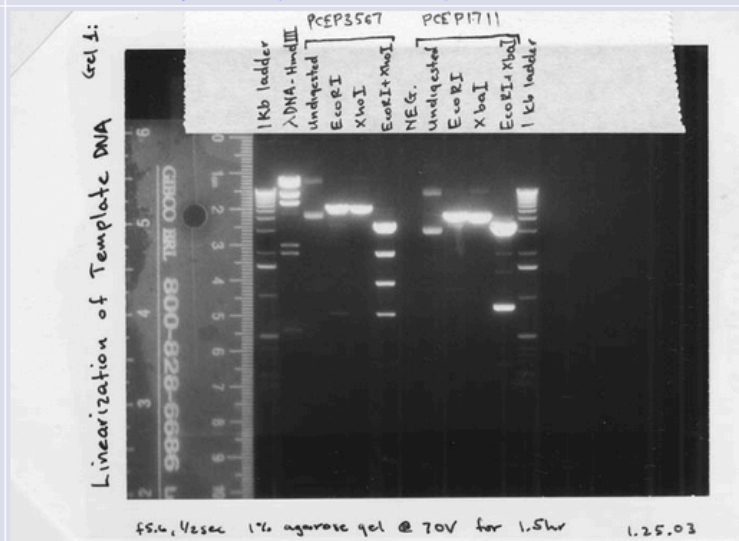
**Figure K-2. Webbook website -- Project Page.** The project page provides a bird's eye view of all the projects generated by the user. Administrators can view all users' projects. From this page, the user can click on a link to view the project info. Each project can have a series of attached experiments.

## Experiments

Experiment *Linearization of Plasmid DNA for Run-off Transcription* — [Edit]

<b>Created:</b>	2003-04-20 23:31:04
<b>Last modified:</b>	2006-04-18 20:25:21
<b>Goal</b>	To linearize plasmid DNA corresponding to PcETA1 (PCEP1711) and Homeodomain Transcription Factor (PCEP3567).
<b>Background</b>	Previous work done by Yuping Bi showed that the PCEP3567 plasmid DNA can be digested with EcoRI and XhoI. These enzymes, when combined, will release the cDNA from the plasmid. The cDNA insert size is - 2 Kb. Furthermore, Yuping Bi showed that the PcETA1 (PCEP1711) cDNA can be release from the vector using EcoRI and XbaI restriction enzymes. The corresponding cDNA insert size is - 0.9 Kb. For a reference of the data, please check Yuping Bi blue book (SRB27 - 9/16/00 and 3/8/01).
<b>Approach</b>	Digest the plasmid DNA templates using restriction enzyme that will cut the plasmid DNA only once at the 5' end and 3' end of the insert in the plasmid.
<b>Controls</b>	Digest 1 ug of lambda DNA using HindIII restriction enzyme as a control for complete digestion.
<b>Results</b>	
<b>Discussion</b>	The experiment did not work!!! I forgot to incubate the DNA at 65oC for 10 minutes to denature and separate the cohesive ends of the lambda phage DNA. Because I didn't do so, I obtained a very large band - 27 Kb. In the reactions with two enzymes, we think that the incubation period was too long and some of the enzymes might have star activity (non-specific digestion of DNA).
<b>Next</b>	I will repeat the experiment.
<b>Stocks</b>	
<b>Protocols</b>	· <a href="#">Radioactive In-situ Hybridization (Scarlet Runner Bean)</a>

Figure 1. Linearization of DNA Template



Digested DNA were loaded on a 1% agarose gel and ran at 70V for 1.5hr. Polaroid was taken with an aperture of f5.6 for 1/2 seconds.

**Figure K-3. Webbook website -- Experiment Page.** An example of an experiment page showing information about the experiment and an attached gel image from the experiment. This page includes background information, protocol used to carry out the experiment, and a discussion of the results.

Home Admin Projects Stocks Protocols Calendar Browse Help Contact Logout

**webBOOK**

**Protocols**

Select a category of protocols to display:

- [HC70AL S09 APPENDIX ONE](#) — [Edit]
- [HC70AL S09 APPENDIX TWO](#) — [Edit]
- [HC70AL S09 EXPERIMENT EIGHT](#) — [Edit]
- [HC70AL S09 EXPERIMENT FIVE](#) — [Edit]
- [HC70AL S09 EXPERIMENT FOUR](#) — [Edit]
- [HC70AL S09 EXPERIMENT NINE](#) — [Edit]
- [HC70AL S09 EXPERIMENT ONE](#) — [Edit]
- [HC70AL S09 EXPERIMENT SEVEN](#) — [Edit]
- [HC70AL S09 EXPERIMENT SIX](#) — [Edit]
- [HC70AL S09 EXPERIMENT THREE](#) — [Edit]
- [HC70AL S09 EXPERIMENT TWO](#) — [Edit]
- [HC70AL S11 APPENDIX ONE](#) — [Edit]
- [HC70AL S11 APPENDIX TWO](#) — [Edit]
- [HC70AL S11 EXPERIMENT 1 - INTRODUCTION TO GENERAL MOLECULAR BIOLOGY TECHNIQUES](#) — [Edit]
- [HC70AL S11 EXPERIMENT 2 - SCREENING SALK T-DNA MUTAGENESIS LINES \(GENE ONE\)](#) — [Edit]
- [HC70AL S11 EXPERIMENT 3 - RNA ISOLATION AND RT-PCR ANALYSIS \(GENE ONE\)](#) — [Edit]
- [HC70AL S11 EXPERIMENT 4 - IDENTIFYING FEATURES OF MUTANT SEEDS USING NOMARSKI MICROSCOPY \(GENE ONE\)](#) — [Edit]
- [HC70AL S11 EXPERIMENT 5 - SCREENING SALK T-DNA MUTAGENESIS LINES \(GENE TWO\)](#) — [Edit]
- [HC70AL S11 EXPERIMENT 6 - RNA ISOLATION AND RT-PCR ANALYSIS \(GENE TWO\)](#) — [Edit]
- [HC70AL S11 EXPERIMENT 7 - IDENTIFYING FEATURES OF MUTANT SEEDS USING NOMARSKI MICROSCOPY \(GENE TWO\)](#) — [Edit]
- [HC70AL S11 EXPERIMENT 8 - AMPLIFYING & CLONING A GENE UPSTREAM REGION \(GENE TWO\)](#) — [Edit]
- [Knockout Guidelines](#) — [Edit]
- [Plant Layout Chart](#) — [Edit]

---

**Create protocol**

Protocol Name:

Category:

File: (?)  No file chosen

File type:

**Figure K-4. Webbook website -- Protocols Page.** This page stores all lab protocols and organize them into groups such as DNA, RNA, GeneChip, LCM, etc. The Knockout protocols list experiments for the HC70AL lab course. Students can click on the link to view a pdf of the protocol.

webBOOK

Home Admin Projects Stocks Protocols Calendar Browse Help Contact Logout

### RNA Stocks

- [lec1-1 Floral Bud RNA](#) — [Edit]
- [lec1-1 leaf RNA](#) — [Edit]
- [lec1-1 roots RNA](#) — [Edit]
- [lec1-1 stem RNA](#) — [Edit]
- [RNA Soybean Maturation-A Axis-epidermis-1](#) — [Edit]
- [RNA Soybean Maturation-A Axis-epidermis-2](#) — [Edit]
- [RNA Soybean Maturation-A Axis-parenchyma-1](#) — [Edit]
- [RNA Soybean Maturation-A Axis-parenchyma-2](#) — [Edit]
- [RNA Soybean Maturation-A Axis-Plumule-1](#) — [Edit]
- [RNA Soybean Maturation-A Axis-Plumule-2](#) — [Edit]
- [RNA Soybean Maturation-A Axis-Shoot Meristem-1](#) — [Edit]
- [RNA Soybean Maturation-A Axis-Shoot Meristem-2](#) — [Edit]
- [RNA Soybean Maturation-A Axis-vascular-1](#) — [Edit]
- [RNA Soybean Maturation-A Axis-vascular-2](#) — [Edit]
- [RNA Soybean Maturation-A Cotyledon ABaxial epidermis-1](#) — [Edit]
- [RNA Soybean Maturation-A Cotyledon ABaxial epidermis-2](#) — [Edit]
- [RNA Soybean Maturation-A Cotyledon ABaxial parenchyma-1](#) — [Edit]
- [RNA Soybean Maturation-A Cotyledon ABaxial parenchyma-2](#) — [Edit]
- [RNA Soybean Maturation-A Cotyledon ADaxial epidermis-1](#) — [Edit]
- [RNA Soybean Maturation-A Cotyledon ADaxial epidermis-2](#) — [Edit]
- [RNA Soybean Maturation-A Cotyledon ADaxial parenchyma-1](#) — [Edit]
- [RNA Soybean Maturation-A Cotyledon ADaxial parenchyma-2](#) — [Edit]
- [RNA Soybean Maturation-A Cotyledon vascular-1](#) — [Edit]
- [RNA Soybean Maturation-A Cotyledon vascular-2](#) — [Edit]
- [RNA Soybean Maturation-A Root meristem-1](#) — [Edit]
- [RNA Soybean Maturation-A Root meristem-2](#) — [Edit]
- [RNA Soybean Maturation-A seed coat parenchyma-1](#) — [Edit]
- [RNA Soybean Maturation-A seed coat parenchyma-2](#) — [Edit]
- [RNA Soybean Maturation-C Aleurone-1](#) — [Edit]
- [RNA Soybean Maturation-C Aleurone-2](#) — [Edit]
- [RNA Soybean Maturation-C Axis-epidermis-1](#) — [Edit]
- [RNA Soybean Maturation-C Axis-epidermis-2](#) — [Edit]
- [RNA Soybean Maturation-C Axis-parenchyma-1](#) — [Edit]
- [RNA Soybean Maturation-C Axis-parenchyma-2](#) — [Edit]
- [RNA Soybean Maturation-C Axis Plumule-1](#) — [Edit]
- [RNA Soybean Maturation-C Axis Plumule-2](#) — [Edit]
- [RNA Soybean Maturation-C Axis-Root Meristem-1](#) — [Edit]
- [RNA Soybean Maturation-C Axis-Root Meristem-2](#) — [Edit]
- [RNA Soybean Maturation-C Axis-Shoot Meristem-1](#) — [Edit]
- [RNA Soybean Maturation-C Axis-Shoot Meristem-2](#) — [Edit]
- [RNA Soybean Maturation-C Axis-vascular-1](#) — [Edit]
- [RNA Soybean Maturation-C Axis-vascular-2](#) — [Edit]

**Figure K-5. Webbook website -- Lab Stocks.** This page provides a central repository for all stocks generated in the lab including RNA, DNA, primers, seeds. For each stock, the stock concentration and storage location are recorded along with the stock quality, if applicable. For primer stocks, the sequence for each primer is recorded along with initial test results of each primer pair.

## APPENDIX L

Design and Implementation of Web-Based Relational Databases

Most of the projects that I have been involved with resulted in a mountain of data (e.g. transcriptomes, methylomes). To make these data available to the general research community as part of the NSF-funded project, I worked closely with Harry Hahn, Weihong Yan, Min Chen, Anhthu Bui, and Bob Goldberg to create, design, and implement these web-based databases in addition to our lab website. Below is a compilation of web sites and databases that I have been involved in designing or creating for our projects including screenshots from each website showing the design and functionality of each web site (**Figures L-1 to L-8**).

## Goldberg Lab

<http://www.mcdb.ucla.edu/Research/Goldberg>

I designed and created the lab web site to highlight the research being carried out in the lab as well as Bob Goldberg's unique teaching techniques and courses.


<a href="#">News</a>	<a href="#">Teaching Website</a>	<a href="#">Research</a>	<a href="#">Publications</a>	<a href="#">Seed Institute</a>	<a href="#">Lab Members</a>	<a href="#">Lab Pictures</a>	<a href="#">Lab Movies</a>	<a href="#">Video Projects</a>	<a href="#">Links</a>
----------------------	----------------------------------	--------------------------	------------------------------	--------------------------------	-----------------------------	------------------------------	----------------------------	--------------------------------	-----------------------

### Welcome to the Goldberg Laboratory

#### BOB GOLDBERG, UCLA

*Learn about what goes on in the Goldberg Laboratory*

**Bob Goldberg**




Department of Molecular, Cell,  
Developmental Biology  
Terasaki Life Sciences Bldg 4121  
Office: (310) 825-9093  
Lab: (310) 825-3270  
Email: [bobg@ucla.edu](mailto:bobg@ucla.edu)



Choose from one of the above links to learn more about Bob Goldberg and his lab

**Meet the Lab Members**



[Meet the Goldberg Lab Members](#)

[Check Out Lab Pictures](#)

Figure L-1. The Goldberg Lab website (<http://www.mcdb.ucla.edu/Research/Goldberg>).



## Seed Gene Network

<http://seedgenenetwork.net>

I designed and assisted with the creation of this website as a portal for the dissemination of data generated from the NSF-funded project to profile gene activity in every compartment, tissue, and cell types across different stages of soybean seed development. This website is constantly updated to include integration of next-generation sequencing data (mRNA-Seq, smRNA-Seq, BS-DNA-Seq) to examine gene regulation during seed development from a systems biology view. The website was created by Harry Hahn and is maintained by Weihong Yan, Min Chen, and myself. This website hosts data generated from projects in Chapter Four and appendices A, F, I, and J. See appendix F for more detailed information about this web site.

The screenshot shows the homepage of the Seed Gene Network website. At the top, there is a green banner with the text "GENE NETWORKS IN SEED DEVELOPMENT" and the subtitle "Identifying all the genes and gene networks required to 'make a seed'". To the right of the banner, it says "Supported by:" followed by the National Science Foundation logo. Below the banner is a navigation menu with links: Home, Soybean GeneChip Experiments, Arabidopsis GeneChip Experiments, GeneChip Annotations, Sequencing, RNAi, People, Data & Resources, and Genome Browsers. The main content area features a heading "Welcome to Gene Networks in Seed Development Website!". Below this is a section titled "ABOUT THE PROJECT" which describes the NSF-funded collaborative effort between the Goldberg laboratory at UCLA and the Harada laboratory at UCD. It mentions the use of soybean and Arabidopsis Affymetrix GeneChips, Laser Capture Microdissection (LCM), and next-generation high-throughput sequencing technologies. A link is provided to learn more about the project and what has been accomplished. Below this is a section titled "GENECHIP EXPERIMENTS" which states that mRNA sets have been profiled in 71 soybean and Arabidopsis seed compartments from preglobular- to early maturation-stage seeds. It mentions that all GeneChip data are stored in a web-based database and that built-in analysis tools allow users to browse the database by probe identification, gene ontology, and functional category, and also compare gene activity in different seed compartments during development. To the left of the "GENECHIP EXPERIMENTS" section is an illustration of a soybean plant labeled "Soybean (Glycine max)". To the right is an illustration of an Arabidopsis plant labeled "Arabidopsis thaliana". There is also a small image of a GeneChip array.

Figure L-2. Seed Gene Network (<http://seedgenenetwork.net>).

## Lab Webbook

<http://estdb.biology.ucla.edu/webbook>

I designed and assisted with the creation of this website to serve as a central repository of lab protocols, lab stock information (e.g. DNA, RNA, seeds) as well as lab experimental notes and data. This website was also used to teach undergraduate students how to do science as part of Bob Goldberg's HC70AL - Gene Discovery Laboratory course (see Appendix K for more details).

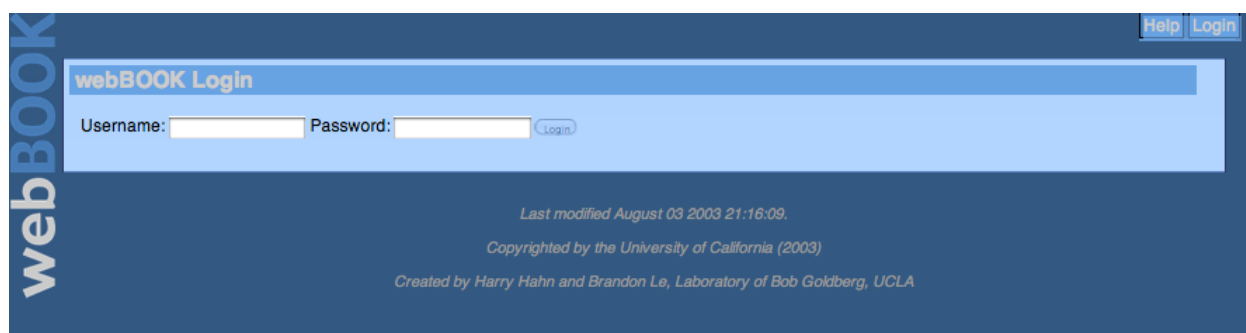


Figure L-3. Webbook (<http://estdb.biology.ucla.edu/webbook>).

## ESTDB

<http://estdb.biology.ucla.edu/~goldberg>

I designed and implemented this web-based DNA sequence analysis tool for high-throughput semi-automated annotation and analysis of EST sequences incorporating tools for DNA sequence analysis (BLAST - sequence alignment, CAP3, contig assembly program, PFAM - protein family database, PHRED - base calling program for original sequence trace files). The website was programmed by Harry Hahn, with input from myself, Anhthu Bui, and Bob Goldberg. This website was developed as part of the SRB EST sequencing project (see Appendix G for more details).

**Goldberg EST-DB**

Current disk usage: 106G free (89% used) / 4.61.4.60.4.52 / Jobs

**Process**  
Process trace files, either existing remotely via an FTP server or provided by an outside source and uploaded to this server. Or, input a new sequence manually, without processing any trace files.

**Summary, Search & Select**  
Obtain summaries, search the database, select specific sets of sequences on which to perform actions.

**Browse/manipulate datasets**  
Browse the cap3 analysis data or the databases available for BLAST. Manually update these datasets, or delete custom blast databases. To create a BLAST database, use the [Search form](#) to first select the desired sequences, then use the "Make Blast DB" action.

**Preferences**  
Update database entries such as projects, function groups, etc.

**Help**  
Online help manual


**Links**  
Links to some useful software.

Figure L-4. ESTDB Sequence Analysis website (<http://estdb.biology.ucla.edu/~goldberg>).

*Arabidopsis Seed Development GeneChip*

<http://estdb.biology.ucla.edu/genechip>

This website host all GeneChip data generated from profiling *Arabidopsis* seed development from before fertilization through maturation and other times in the plant life cycle (see Chapter two). The website have several features including the ability to browse the datasets and to carry out comparative analysis between different datasets (see below). The website was programmed by Harry Hahn with design and directions from myself, Anhthu Bui, and Bob Goldberg.



**Arabidopsis thaliana  
Genechip Project**

Home About Browse Blast People Links

Description Experiments Procedure Annotation

**The Goldberg Lab Arabidopsis thaliana Genechip Database**

The Goldberg Lab *Arabidopsis thaliana* GeneChip Database is a search engine that can be used by the scientific community to search for the expression profile of any gene throughout the Arabidopsis life cycle.

**Profiling Gene Activity During Arabidopsis Seed Development and the Entire Life Cycle**

Seed contains a complex collection of cells, tissues, and organs with distinct developmental fates. We are interested in identifying the genes and regulatory networks that play important roles in seed development. Utilizing a genomics approach, we profiled mRNAs active during Arabidopsis seed development using whole-genome Affymetrix GeneChips ATH1-121501. Specifically, we profiled mRNAs active during seed development at 24 hour post-fertilization, 3-4 days after pollination (DAP), 7-8 DAP, 13-14 DAP, and 18-19 DAP. These gene sets were compared with those active in pre-fertilization ovules, seedlings 3 days after imbibition, leaves, roots, stems, and floral buds of the mature plant. Collectively, we have carried out whole-genome analysis of genes active during the entire *Arabidopsis* life cycle.

**Citation**

If you use GeneChip data from this web site, please cite this web site and indicate that the data were taken from the experiments carried out in the laboratories of Bob Goldberg (UCLA) and John Harada (UC Davis) by Brandon Le (UCLA), Anhthu Bui (UCLA), and Julie Pelletier (UC Davis). Please contact Brandon Le (BLE at UCLA dot EDU) if you have any questions about these experiments or this web site.

All of the data in this web site will soon be submitted for publication - **Identification of Seed-Specific Transcription Factors From a Global Analysis of Gene Activity During the Arabidopsis Life Cycle** - Brandon H. Le, Anhthu Q. Bui, Javier A. Wagmaister, Julie Pelletier, Linda Kwong, Zixing Fang, Steve Horvath, Gary N. Drews, Robert L. Fischer, Jack K. Okamoto, John J. Harada, and Robert B. Goldberg.

**Figure L-5. *Arabidopsis thaliana* GeneChip Project website (<http://estdb.biology.ucla.edu/genechip>).** This site accompanies the PNAS paper (Le et al. 2010) on using GeneChip to profile gene activity throughout *Arabidopsis* seed development and other time in the plant life cycle. All the data from the paper is accessible here and can be browsed on a probe set by probe set basis.

**Arabidopsis thaliana Genechip Project**

Home About Browse Blast People Links

**Annotation filtering criteria**

Probe Set Identifier:  (ex 260854\_at)

GeneChip Array:

AGI Locus ID  (ex ATIG21970)

GO: Biological Processes  (ex GO:6355)

GO: Cellular Components  (ex GO:5634)

GO: Molecular Function  (ex GO:3677)

Functional category:

Description/Keyword:  (ex transcription, CCAAT)

**Data filtering criteria**

Experiment

- ATH1-121501/WT C24/Leaf/Vegetative
- ATH1-121501/WT Ws-0/Cellularized Endosperm/Linear Cotyledon Stage
- ATH1-121501/WT Ws-0/Chalazal Endosperm/Globular Stage
- ATH1-121501/WT Ws-0/Chalazal Endosperm/Heart Stage
- ATH1-121501/WT Ws-0/Chalazal Endosperm/Linear Cotyledon Stage
- ATH1-121501/WT Ws-0/Chalazal Endosperm/Mature Green Stage

Detection Call \*

Signal Value

\* A=Absent in all replicates, M=Marginal in all replicates, P (all)=Present in all replicates, P (majority)=Present in more than half of the replicates.

**Figure L-6. Arabidopsis thaliana GeneChip Project website -- Search Form.** This form allows the user to search through the GeneChip datasets by several parameter including the probe set identifier, functional category, or keyword. The user can also filter the datasets based on GeneChip detection call and/or signal intensity between multiple datasets.

Phaseolus coccineus ESTs (*PcESTs*)

<http://estdb.biology.ucla.edu/PcEST>

This website was developed to provide to the research community a collection of ESTs identified from different regions (embryo proper and suspensor) of a SRB globular stage embryo (see Appendix G). All the information on this website was summarized and annotated by myself and Anhthu Bui. The website have functions to browse the dataset or using BLAST to identify ESTs with homology to a user sequence of interest (see below). The website was programmed by Harry Hahn with design, data, and summaries provided by myself, Anhthu Bui, and Bob Goldberg.

# Phaseolus coccineus Embryo EST Project



Home

Search

## The Scarlet Runner Bean (*Phaseolus coccineus*) EST Project

We are using the giant embryos of the Scarlet Runner Bean (*Phaseolus coccineus*) to identify genes that are active in the suspensor and embryo proper of globular-stage embryos shortly after fertilization ([Weterings et al., Plant Cell, 13, 2409-2425, 2001](#); [Le et al., Plant Physiology, 144, 562-574, 2007](#); [Kawashima et al., PNAS, 106, 3627-3632, 2009](#)). Our long-term goal is to understand the region-specific differentiation processes that occur during early embryo development and how genes are activated specifically in the suspensor and embryo proper. Total RNAs isolated from hand-dissected suspensor and embryo proper were used to construct cDNA libraries. The 5' ends of individual cDNA clones from each library were sequenced using the Sanger sequencing method whereas random cDNA fragments were sequenced using the 454 sequencing technology. To date, we have sequenced 305,363 suspensor and 85,776 embryo proper ESTs. These ESTs have been grouped into different functional categories based on BLAST searches and organized into an EST relational database on our lab server. Moreover, all EST sequences can be accessed directly from NCBI ([GenBank Accession Series CA896559 to CA916678 and GD289845 to GD660862](#)).



Our Scarlet Runner Bean (SRB) globular-stage suspensor and embryo proper EST database is open to the scientific community. Therefore, anyone can identify SRB ESTs similar to his/her experimental DNA sequence(s) by performing BLASTN and/or TBLASTX searches against sequences in our EST database.

[Click here](#) to browse or BLAST your sequences against the Scarlet Runner Bean EST database.

### Summary of Embryo EST Project

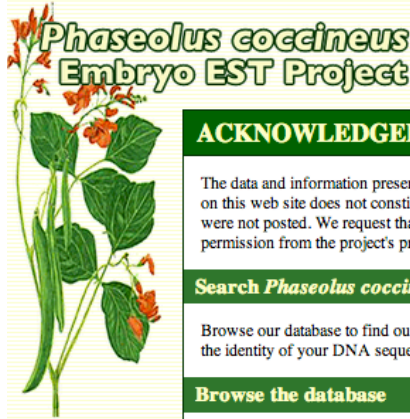
This table includes the number of ESTs identified in each functional category from the suspensor and embryo proper library. Functional categories were assigned based on BLAST homologies to other genes. The e-value cut-off to consider if a blast hit was significant is 1e-04.

Abbreviations: EP - Embryo Proper; S - Suspensor

FUNCTION	EP	S
Metabolism	3,451	13,839
Energy	2,658	9,598
Cell Growth/Division	492	1,503
Transcription	1,385	4,254
Post-transcription	1,170	3,126
Protein Synthesis	15,074	31,710
Protein Destination and Storage	3,442	12,324
Transporters	2,027	6,878
Intracellular Traffic	851	3,475
Cell Structure	2,410	7,411
Signal Transduction	1,021	3,611
Disease/Defense	1,591	12,605
Secondary Metabolism	2,092	12,194
Transposons	21	52
Unknown or Unclassified Proteins	5,762	23,202
No Significant Homology to Public Databases	42,329	159,581
Total ESTs	85,776	305,363

**Figure L-7. *Phaseolus coccineus* ESTs (PcEST) website (<http://estdb.biology.ucla.edu/PcEST>).** A summary of more than 400,000 ESTs from the embryo proper (EP) and suspensor (S) region of globular stage SRB embryo. This site contains all EST sequence data and annotations from the EST project in Appendix G and all EST sequence data were processed using the ESTDB Sequence Analysis website (see Appendix H).





[Home](#)

[Search](#)

### ACKNOWLEDGEMENT OF INFORMATION AND DATA USE AND DISCLAIMER

The data and information presented in this Seed Gene web site are provided to the scientific community as a resource to advance seed research. The data on this web site does not constitute a scientific publication and were posted as soon as they were checked for quality and accuracy. Problem experiments were not posted. We request that the information and data presented in this web site not be published prior to our publication and/or without a written permission from the project's principle investigator. Bob Goldberg (bobg@ucla.edu).

### Search *Phaseolus coccineus* EST Database

Browse our database to find out which mRNAs are present in the suspensor and embryo proper regions or run search against our database to determine the identity of your DNA sequence(s).

### Browse the database

Explore mRNAs found in the suspensor and/or embryo proper by using the form below. Browse mRNAs identified in the *Phaseolus coccineus* EST project categorized by functional category in the suspensor and/or embryo proper or browse putative identification (e.g. cytochrome P450, protein kinase, etc...).

Embryo region:	<input type="text" value="All"/>
Functional category:	<input type="text" value="Any"/>
Putative identification:	<input type="text"/>
<input type="button" value="Search"/> <input type="button" value="Reset"/>	

### Search the database

If you know the *Phaseolus coccineus* ID, GenBank Accession Number, or Genbank dbEST ID, you can search using the form below.

Phaseolus coccineus EST ID	<input type="text"/>
Genbank accession:	<input type="text"/>
dbEST_ID	<input type="text"/>
<input type="button" value="Search"/> <input type="button" value="Reset"/>	

### Run BLAST against database

You can BLAST your DNA sequence against the suspensor and/or embryo proper databases by (1) entering the name of your DNA sequence in the "NAME OF YOUR SEQUENCE" field, (2) selecting a database using the pull-down menu, (3) selecting the type of BLAST run, (4) entering your DNA sequence and (5) pressing the "RUN BLAST" button.

Name of your sequence:	<input type="text"/>
Database to BLAST against:	<input type="text" value="All"/>
BLAST Program (?)	<input type="text" value="blastn"/>
Your sequence:	<input type="text"/>
<input type="button" value="Run BLAST"/> <input type="button" value="Reset"/>	

**Figure L-8. *Phaseolus coccineus* ESTs (PcEST) website -- Search Form.** This page allows user to search the EST database based on several fields including embryo region, functional categories, and annotations. ESTs can be further access by EST ID, Genbank Accession, or dbEST\_ID. Alternatively, users can enter a sequence of interest and BLAST against our ESTs to find similar sequences in the SRB embryo regions.