Behind the Scenes: Planning a Multispecies Microarray Experiment

Naomi Altman, Jim Leebens-Mack, Laura Zahn, André Chanderbali, Donglan Tian, Lillian Werner, Hong Ma, and Claude dePamphilis

The Floral Genome Project (FGP) is an ambitious multi-institutional research program using genomic approaches to better understand the evolution and diversity of flowering plants (angiosperms), particularly focusing on genes active in the developing floral buds.

To achieve this objective, a key goal was to develop a microarray system that can be used to compare the activity of related genes both within and between species, particularly genes involved in the reproductive tissues. This requires substantial information about the genomics of the target species—13 species of flowering plants holding key positions in the angiosperm evolutionary tree and two nonflowering seed plants (gymnosperms). Five of these species have been selected for microarray gene expression studies.

Basic Biology

The basics of gene expression and microarray technology are described in the article beginning on Page 4 and pursued in more detail in the articles staring on pages 16 and 40. The discussion in these articles supposes that the genes in the organism are known already and that a microarray system is available. (By microarray system, we mean chips that can be purchased or printed with known genomic material.) However, at the start of the Floral Genome Project, little genomic information and no microarray system were available for any of the species in the study, and whole genome sequences for most of the FGP species will not

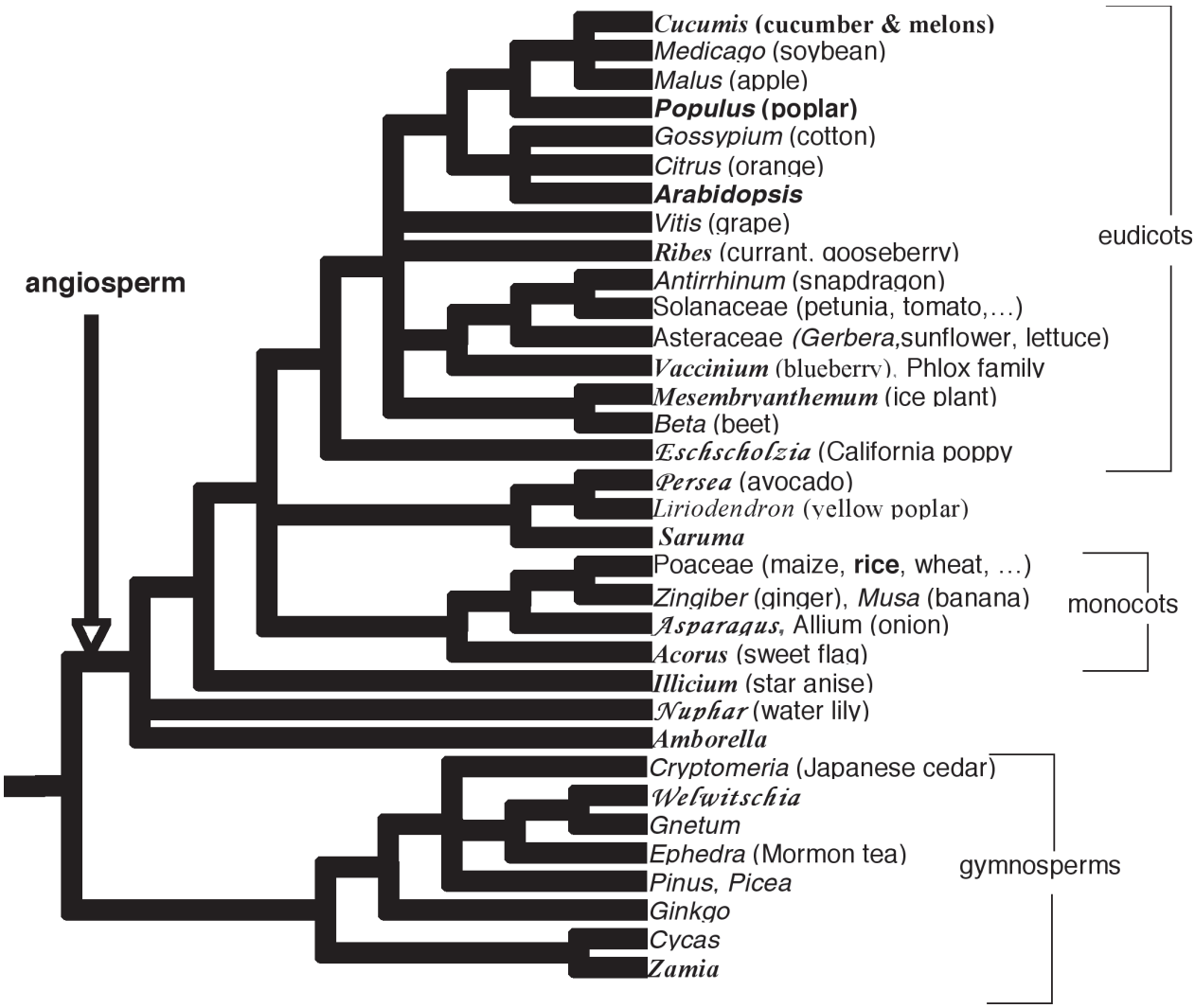


Figure 1. Phylogenetic tree of flowering plants and gymnosperm relatives, showing the plant species targeted for cDNA library building and EST sequencing in the Floral Genome Project (FGP). FGP species occupy important positions on the angiosperm evolutionary tree, including the earliest living branches of monocots, eudicots, and the basal-most angiosperms. FGP plants were targeted to identify genes expressed in floral development by isolation of 10,000 EST sequences (*bold italic*) or 2,000 EST sequences (*italics*). These ESTs were used to examine species using microarrays (*script*). Other plants (names in black) are targets of other ongoing plant genome initiatives, including full genome sequences for *Arabidopsis*, poplar, and rice (bold).

be available in the foreseeable future. Furthermore, while the microarray technology is used frequently to understand expression profiles for one or many processes in a single species, one important objective of our study was to compare gene expression profiles for individual floral genes across multiple species. Because the binding of cDNA in the samples to the probes or spots on the microarray chips is highly specific, only cDNA from very closely related species can be hybridized effectively to the microarray chips. The FGP species were selected specifically to represent key taxonomic positions in the evolutionary tree of plants (see Figure 1) and are not closely related. Therefore, a microarray system must be developed for each species for which a microarray study is planned. Both the microarray chips and the experiments must be designed to facilitate cross-species comparisons.

During the process of gene expression, each gene that is active in a tissue is transcribed into mRNA. The number of copies of mRNA produced by the gene is a direct measure of the level of gene expression. Determining the expression level of genes in a particular tissue, or the ratio of expression in two or more tissues (or under two or more conditions) is the end goal of a microarray study. Knowledge of the mRNAs present in a tissue also can be used as a starting point for acquiring genomic information about a species.

The starting point for obtaining genomic information for the Floral Genome Project species was the use of mRNA to construct a cDNA library for each species. In this process, mRNA is extracted from tissues, reverse transcription is used to convert the mRNA into

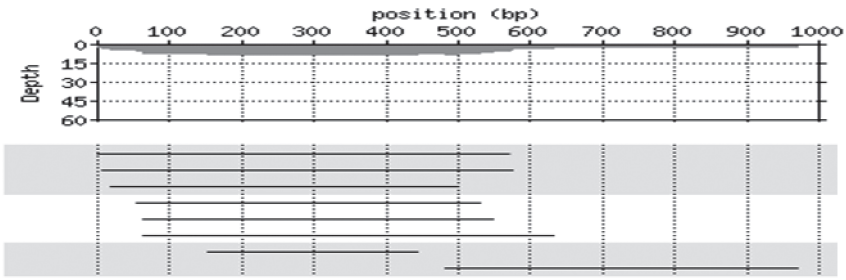
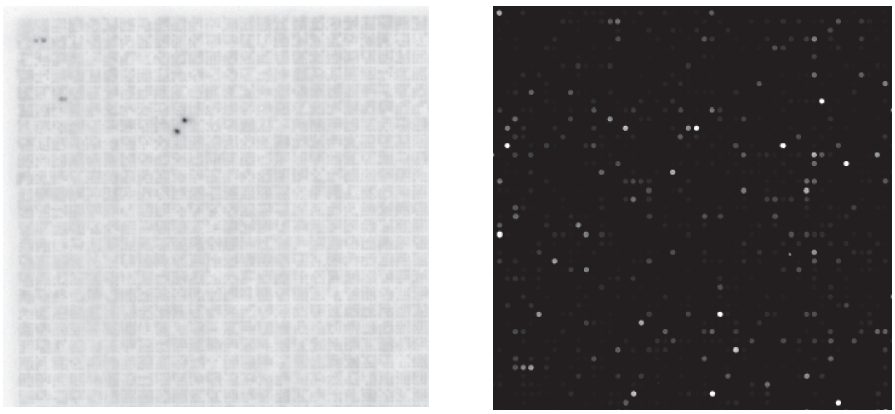


Figure 2. **Eight ESTs representing clones from the same poppy gene. The diagram at the top shows the assembled unigenes, with the depth representing the number of overlaps at each position. The lower diagram shows the positions of the individual ESTs.**

cDNA (which does not have the introns and is more stable than mRNA), and each cDNA molecule is individually cloned into bacteria cells. Thousands of bacterial cell lines, each containing copies of an individual cDNA, can be stored in freezers as small cultures (clones) in sets of 96-well or 384-well microtiter plates. Highly expressing genes, which produce many mRNA copies, may be cloned into many of the cell lines, while lowly expressing genes will be present in very few of the cells lines—or may not be captured at all. Until the cDNA is sequenced, it is not known which genes are maintained in each cell line. The bacterial cells can be amplified, thus replicating each cDNA with each bacterial cell division. High throughput methods have been developed for extracting the cDNAs from the bacterial cultures in the microtiter plates.

Expressed sequence tags (ESTs), which identify 500–750 base pairs of the cDNA using high-throughput sequencing are usually sufficient to infer the identity of a gene or place the gene in one of more than 10,000 gene families known to exist in plants. Highly expressed genes may be sequenced many times in a few thousand EST sequences, and computer algorithms have been developed to assemble overlapping ESTs into "contigs" or "unigenes" (see Figure 2). Rare mRNAs with low expression but still present in the cDNA library may not be captured in the sequencing at all. Therefore, EST and unigene sets typically will be biased toward genes with high levels of expression. In addition, unigenes assembled for more highly expressed genes usually will be closer to being complete. Although sequencing errors are inevitable, average error rates can be assessed based on the quality of the raw data, with low-quality data typically being removed before further analysis.

Because the FGP is focused on genes expressed during flower development, a cDNA library was created from mRNAs extracted from very young flower buds. Clearly, a cDNA library created from a single tissue will be biased in gene content compared to all the genes in a complete genome. There will be an overrepresentation of genes involved in the development and function of that tissue and underrepresentation of genes involved in other tissues and processes.



a) **b)**

Figure 3. **Gene expression arrays. a) A single-gene probe hybridized to just a few cDNA spots on a macroarray. The cDNAs were chemically fixed after bacterial clones were printed on the nylon filter. b) A gene expression microarray printed at high density on a coated glass slide. The probes are labeled with fluorescent dye. The gray scale represents the intensity of fluorescence of the red dye. The cover of this issue depicts a combined "false-color" image produced from a two-channel microarray. Spots are colored on a scale of red through green, with yellow meaning equal expression in both channels. The color indicates the degree of differential expression, while the brightness indicates the hybridization intensity.**

For example, we expect our cDNA library to be deficient in genes preferentially expressed in root tissues.

One of our concerns was to obtain good coverage of the set of genes expressed in developing flower buds. Sequencing is expensive, so in selecting clones to sequence, we would like to deplete the sample of clones that already have been sampled and enrich for clones with genes related to those that have been shown to be involved in floral development in *Arabidopsis*, the species for which the genomics of floral development has been studied most intensively.

Because the number of unique genes in a cDNA library is finite, the frequency of new genes found in additional sets of randomly picked clones will drop as the total number of sequenced clones increases. We therefore used a technology called a macroarray (see Figure 3) to screen the cDNA library for clones with genes similar to known *Arabidopsis* genes of particular interest. A few thousand bacterial clones containing cDNA are spotted directly onto a nylon filter and allowed to grow overnight. The cDNA from each clone is fixed on the filter, and the resulting macroarray can be probed with fragments from the genes of interest. The fragments will hybridize to the spots corresponding to cDNA clones with sequences similar to the fragment sequence. These clones can then be sequenced, and the sequences clustered into unigenes as described above.

In total, we sequenced about 10,000 randomly chosen cDNAs, as many as 20 or so targeted genes from the cDNA libraries for each of nine highly sequenced FGP species, and 2,000 or slightly more for the other FGP species (see Figure 1). In each of the highly sequenced species, these ESTs clustered into just more than 6,000 unigenes suitable for designing probes for the species-specific microarray chips.

Constructing the Microarray Chips

Given the plethora of choices of microarray print technologies, our first problem was to select a method appropriate for our needs. Comparisons of microarray technologies show that correspondence among the patterns of gene expression estimated using these different technologies is high, but each technology comes with its own limitations and biases.

Basically, there are two methods for creating the material printed on a microarray. On a cDNA chip, cDNA extracted from the cDNA library are spotted on a glass slide using a robotic printer. This allows an array to be built even before the sequence of each cDNA is known. On an oligonucleotide ("oligo") chip, the gene sequence of the cDNA is used to create highly specific oligos, which are arrayed on the glass slide (described in the article starting on Page 4). There are a number of technologies for printing oligo chips, some of which fabricate the oligos in liquid that is then spotted on the array just like cDNA and others that fabricate the oligos directly on the chip.

While much of the literature on the analysis of microarrays assumes the material printed on the chip is known, there are many opportunities for errors in generating the chip. On cDNA chips, two common sources of error are mistakes in labeling the microtiter plates containing the cDNAs (including simply rotating the plates by 180o when placing them in the robot) and contamination of the cDNA library by *E. coli* that have a different cDNA than the one originally placed in the well. On oligo chips, three common sources of error include sequencing error in the original ESTs, using parts of genomic sequence that are not actually part of the coding regions of a gene (such as introns or part of the *E. coli* chromosome), and fabricating the oligo from the wrong strand of the cDNA.

Another problem can arise in species that are not fully sequenced. Gene, and even whole genome duplication events, are relatively frequent in plant evolution. Therefore, many genes have long segments that are very similar. Sets of related genes with high levels of sequence similarity due to shared ancestry are called gene families. Oligos printed from these similar segments of the gene may hybridize to mRNA from more than one gene. When the species is not fully sequenced, we do not know whether the partially sequenced genes we observe in our ESTs have closely related (and therefore similar) gene family members that were not sequenced. We can, however, make some guesses about gene family size and the stretches of the gene that are highly similar among gene family members based on information from the three plant species with fully sequenced genomes and EST sequences from many plant species.

A further consideration in selecting a microarray print technology is cost. Generating a chip using a spotting robot is inexpensive once the material to be spotted is available. Extracting cDNA or fabricating oligos for spotting is both expensive and time-consuming, but, once it is done, hundreds of chips can be printed with little additional cost. On the other hand, technologies that fabricate oligos on the chip surface are expensive on a per-chip basis, but have much lower fixed cost if an experiment requires only a few chips.

Finally, there is the question of the accuracy of the technology. cDNAs have lengths and compositions that greatly vary. As a result, the rates at which the sample cDNAs hybridize to the matching strand printed on the chip varies considerably from gene to gene. Further, spotting technology introduces noise. Oligos usually are constructed to have all the same length, and computer scientists—together with biochemists—have created algorithms to select oligos that are gene specific and have a high probability of nearly uniform hybridization rates. Fabrication of the oligos directly on the chip appears to introduce much less noise than spotting. However, even the best oligo selection methods sometimes pick oligos that fail to hybridize.

In the 18-month period during which we were deliberating over choice of the microarray technology, the cost of fabrication of the oligos directly on the chip dropped precipitously. At the same time, we were developing the experimental design for the microarray study that would most efficiently detect differences in gene expression among the eight organs we wanted to compare (leaves, small flower buds, medium flower buds, fruits and sepals, petal, stamen, and carpels from mature flower buds [see Figure 6]). We determined we would need no more than 40 chips for each species, and possibly as few as 16. While other laboratories might want to use our chips, we did not want to handle the logistics of creating a warehouse of spotting material. Hence, we decided to use a two-color system for which the oligos are fabricated on the chip surface. Once we designed the chips,

the company that produces them could distribute copies to other labs for use in additional experiments.

Having chosen a microarray technology, we had to determine which oligos should be printed on the microarray chip. The usual computer algorithms for selecting oligos typically compare all gene sequences in a fully sequenced genome to avoid cross-hybridization with similar gene family members. However, we know we have sequenced only a fraction of most gene families in our species. We collaborated with bioinformaticians at The University of North Carolina at Charlotte to develop a pipeline that designed oligos with nearly uniform hybridization rates for each unigene in our EST library while avoiding stretches of sequence that might be conserved among closely related gene family members. Because we did not have complete genome sequences for our species, we compared our unigene sequences to the complete *Arabidopsis* and rice gene sets in order to predict which of our unigenes were likely to have similar (but not sampled) gene family members and identify the portions of the sequence (domains) likely to be similar among gene family members. As a test of our design strategy, in some gene families, we intentionally selected a few oligos from these highly similar domains to see if they might show higher hybridization levels, or less organ-specificity due to contributions from other members of the gene family.

This design process produced custom microarray chips for each of the species included in our study, with about 10,000 spots (oligos) per chip representing about 6,000 unigenes, some of which were unique to a particular species and some of which could be identified as close relatives to genes in other species.

The Pilot Experiment

Whenever a large and expensive experiment is to be done, it is a good idea to run a smaller and less-expensive pilot experiment. The pilot can be used to train the experimenters, determine potential problems, and fine-tune the experimental protocol.

We decided to run two pilot experiments: a small pilot experiment using *Arabidopsis* leaf and flower tissue with commercially available *Arabidopsis* chips and another using California poppy leaf and flower tissue using the chips we designed for poppy—one of our 15 study species. We decided to use *Arabidopsis* in the first pilot experiment because many microarray experiments have been done with this species (with resulting data available at *www.arabidopsis.org*) and because the company that manufactured our chips also produces a microarray chip for *Arabidopsis*, which is mass-produced at relatively low cost and has been checked rigorously for reliability. This allowed us to familiarize ourselves with the experimental procedures and verify that we were able to obtain good measures of gene expression by comparing the results of our pilot experiment to the results of other experiments. We decided to do a second pilot with California poppy to ensure the protocols we developed for selecting oligos for the arrays and sample preparation were working properly.

For the *Arabidopsis* pilot, we had a number of other objectives, including selection of labeling kit, comparison of results between the two laboratories doing the experiments, and comparison of replicates taken from different plants (biological replicates) with replicates taken from the same tissue samples (technical replicates). Three kits were available for labeling the RNA. We wondered if the kit that uses the smallest quantity of RNA was adequate, as some of the tissues in the experiment are small and difficult to collect. Half of the microarray experiments are being performed at Pennsylvania State University, while the other half are being performed at the University of Florida. This means plants are grown under different environmental conditions, the experiments are done by different individuals, and the scanning is done on different scanners with different sensitivities and image analysis software. We wanted to compare results at the two locations when the same species was used.

Table 1—Maximum variance of pairwise contrasts for several designs with 16 microarrays and eight tissues. The error variance for gene *i* is $\sigma^2_{i\varepsilon}$

correlation	reference	2 identical loops	interwoven loop
0.75	$1.75\,\sigma^2_{i\varepsilon}$	$1.217\,\sigma^2_{i\varepsilon}$	$0.875\,\sigma^2_{i\varepsilon}$
0.50	$1.50\,\sigma^2_{i\varepsilon}$	$0.857\,\sigma^2_{i\varepsilon}$	$0.750\,\sigma^2_{i\varepsilon}$
0.25	$1.25\,\sigma^2_{i\varepsilon}$	$0.645\,\sigma^2_{i\varepsilon}$	$0.625\,\sigma^2_{i\varepsilon}$

The good news from this pilot was that our best set of array experiments gave results that were qualitatively similar for the samples processed at both universities, and both were qualitatively similar to the results produced by a collaborating project at Penn State. "Qualitatively similar" means most of the genes we found to be higher in leaves than buds were higher at both sites and in the other project, and most genes we found to be lower in leaves also were consistent. Our experiment had, at most, three replicates (independently grown trays of plants) per treatment, while the previous experiments had two replicates and used Affymetrix GeneChips©. We did not expect to have perfect agreement for genes with only moderate or low changes in expression.

Another piece of good news was that we were able to obtain reliable results without using large quantities of RNA for our samples. This was an important finding as we knew we would have limited tissue for some of the floral organs we planned to use in later experiments.

Most importantly, we found we had good reproducibility of results at each site (Pennsylvania State University and the University of Florida) from chips that were hybridized on the same day—whether they were replicates from the same tray of plants or came from independently grown plants—but poor reproducibility from arrays hybridized on different days or sites. Adding a protectorant that reduced dye degradation over time appeared to reduce

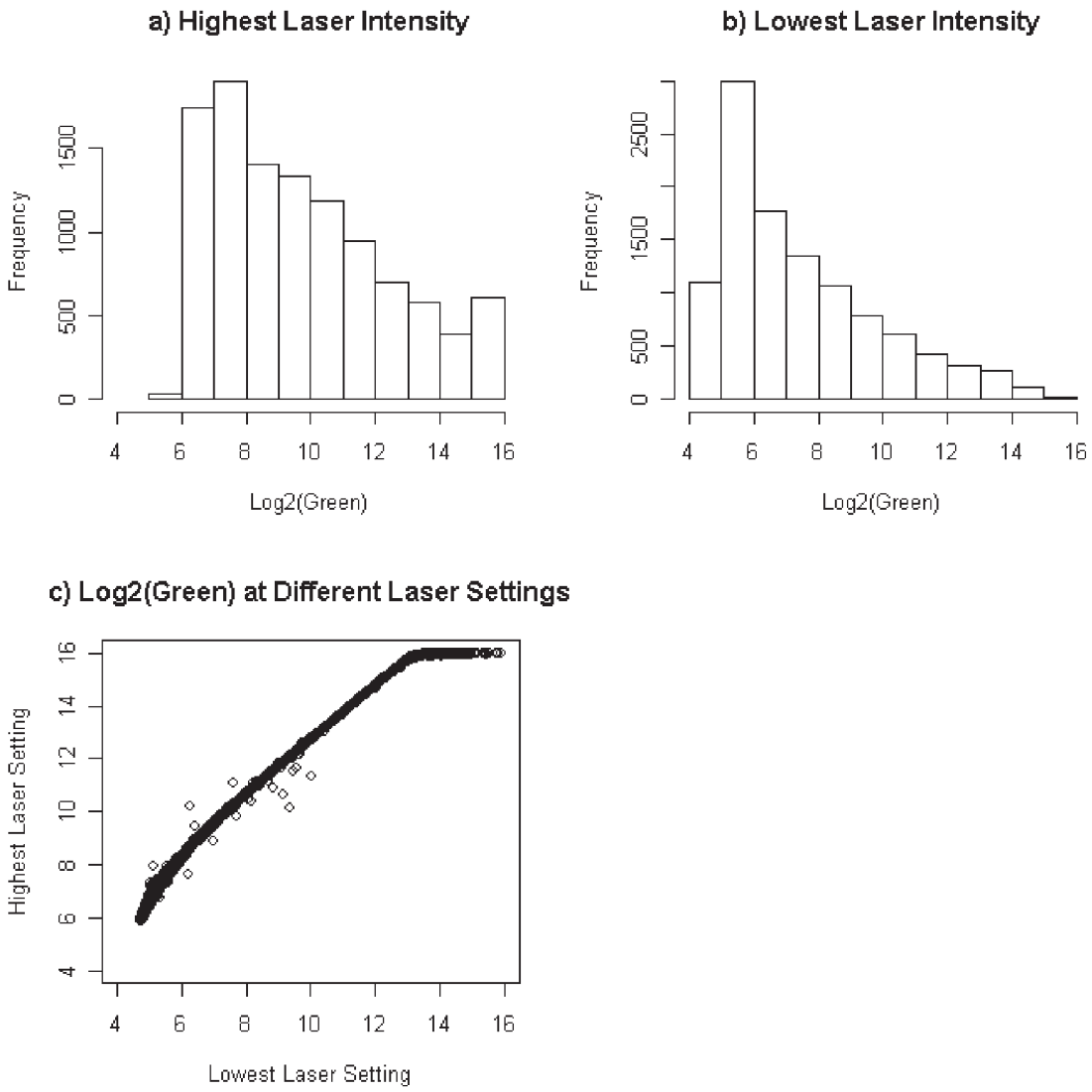


Figure 4. **One microarray was scanned at four laser intensities. In this plot, the highest and lowest intensity scan of the same array in the green channel is shown. a) The highest laser intensity has an excess of data near the upper boundary of 2^{16}. b) The lowest laser intensity has an excess of data in the lowest bin of 2^4. c) A plot of the highest versus lowest settings demonstrates that, at the highest laser setting, many points are recorded at maximum intensity, while, at the lowest laser setting, there does not appear to be a loss of data at the low end.**

this problem. However, the investigators running the experiments decided that, for each species, they should try to run the entire experiment in a single day. We also determined chips could be stored for at least several days, if stored in a nitrogen atmosphere at low temperature, and could be rescanned with minimal data loss. This was an important technical improvement, as it allowed us to store arrays until we were sure all the data were usable.

An important source of variability appeared to be due to differences in the scanners available at each site. We learned that other investigators have scanned chips multiple times at different scanner settings and claimed that combining the data across settings improved sensitivity at the highest and lowest levels of expression. However, we found that by using a simple exploratory technique, we could use several scans to pick a single setting at each, which is sensitive throughout the range of data and comparable across sites. This helps avoid combining the data from the several scans. This is illustrated in Figure 4, which shows the $\log_2$ (Green) readings for one of the arrays at the highest and lowest of the four laser intensities tested.

To understand Figure 4, you must know that the highest number that can be recorded by the scanner is 2^{16} intensity units. The scanner shines a laser on the spots, which fluoresce at an intensity proportional to both the intensity of the laser and the quantity of hybridized material. If the laser intensity is high, then high-intensity spots will fluoresce more strongly than can be recorded by

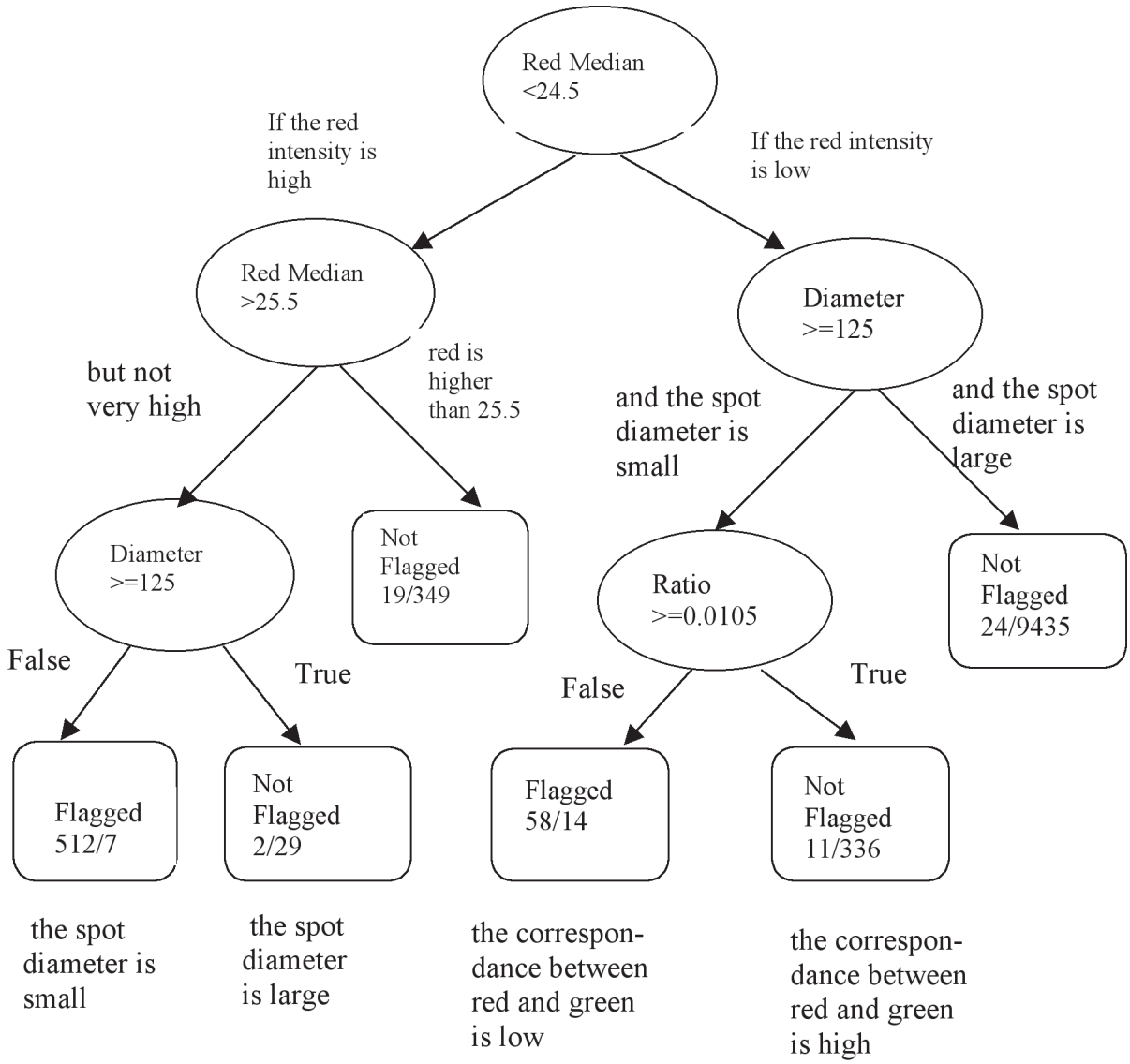


Figure 5. **A decision tree for flagging "bad" spots. The rectangles represent the final decision to flag or not. The numbers in the rectangles are x/y, where x is the number of spots flagged by the investigators and y is the number of spots not flagged. Red median is the median fluorescence of the red sample. Ratio is a measure of the increase in red fluorescence as a function of green fluorescence. Diameter is the diameter of the bright area of the spot. We see that small spots with low red fluorescence are generally flagged by the investigators. Two spots were flagged by the scanner software; both were classified with the large group of unflagged spots in the middle right of the figure.**

the scanner, and the resulting intensities will be truncated at the maximum of 2^{16}. But at this high intensity, spots with low, but nonzero, hybridization will fluoresce strongly enough to be recorded. At a lower laser intensity, the high spots will fluoresce less and be recorded at a level less than 2^{16}, but the fluorescence of the low-intensity spots may be too weak to be detected by the scanner with this setting. In Figure 4, the two histograms show the shift in intensity units when the scanner is reset and the truncation at 2^{16} for the high setting. In the scatterplot, the y-axis is the $\log_2$ (fluorescence) for the highest setting of the laser. The truncation of the high-intensity spots can be seen by the pile-up of data at $\log_2(y)=16$. In fact, all settings of the laser except the least intense show evidence of high-end truncation, but there does not appear to be a loss of sensitivity at the low end, even for the least intense laser setting. We concluded that, at one site, the default scanner setting was adequate, while, at the other, the lowest laser setting was the most sensitive to the range of the data and should be used.

An image of an array can be seen on the cover of this issue. Scratches, material sticking to the chip (the likely cause of the blob in the lower right corner) can be seen, along with irregularly shaped or very small spots. The scanner software recognizes some of these bad spots, but

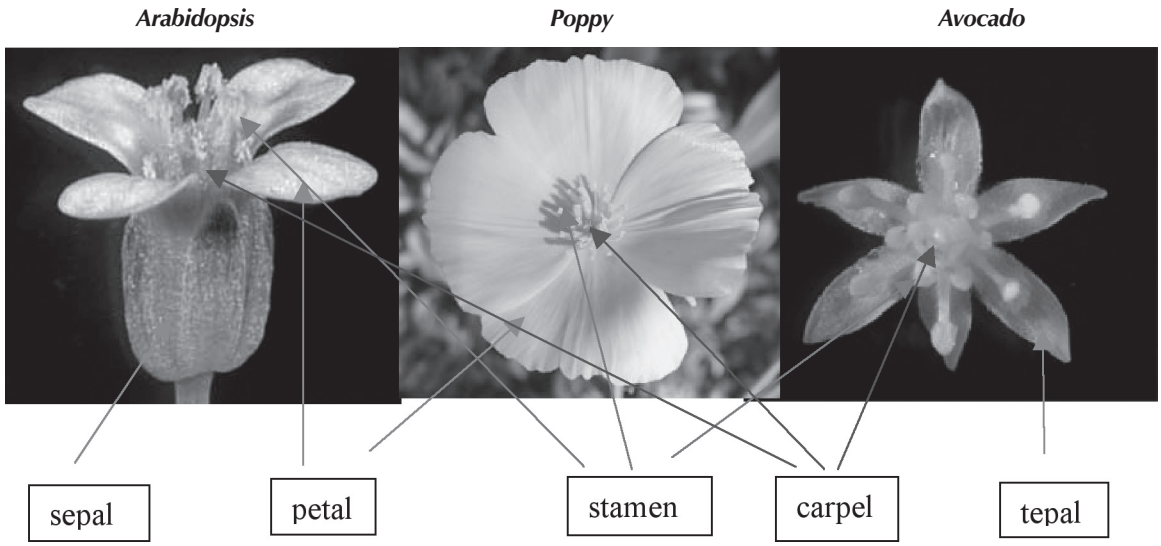


Figure 6. **Flowers of three angiosperms: *Arabidopsis*, poppy, and avocado. The structures at the base of the *Arabidopsis* flower are sepals; the four flat structures are petals. The sepals of the poppy are hidden behind the four large petals. The avocado flower does not have sepals or petals. The six petal-like structures are an undifferentiated organ called tepals.**

not all. A common practice in the analysis of two-channel microarrays is to look at the chip images and manually flag spots that appear to have poor data (e.g., Wise, Hardin, and Hoopes in the article starting on Page 40). Flagged spots generally are considered to be missing data and not used in subsequent analyses. We found that manually flagging can take up to 45 minutes per array on a chip, which is a substantial time investment. We therefore used recursive partitioning on the various spot summaries produced by the scanner to see if we could use a statistical algorithm to mimic the flagging done by the investigators. We found this worked well at recovering the decisions of the investigator. The flagging algorithm is displayed in Figure 5.

The decision rules in the flagging algorithm revealed the investigators were primarily flagging low-intensity spots (many of which can be seen on the lower right corner of the image on the cover), and therefore losing the important information that some genes are not expressed in some tissues. For our study, these should not be considered missing data, but rather important information that some genes have very specific expression patterns. For the remaining experiments, we used all the data.

The pilot experiment also indicated we printed the wrong strand of the double-stranded DNA. Although many of our ESTs represent well-known genes and the orientation of the coding strand is obvious, some ESTs are from regions of the gene that are not translated into proteins. The orientation of these untranslated regions (UTRs) and of portions of protein coding genes that are not known from other organisms is much less certain. Because the oligos are fabricated on each array, and the arrays are produced in small quantities, we were able to alert our supplier and correct orientation errors before running our main experiment. The ability of the user to improve the array designs at a lower cost with each successive experiment is an advantage of on-the-chip oligo synthesis.

Designing the Microarray Experiments

We planned to use the microarray experiments to compare organs or tissues—fruit, carpel, stamen, petal, sepal, leaf, and whole buds—in each species at two stages: small and medium. Because we wanted to compare gene expression across several species, we needed to sample these tissues at the same developmental stage in each species. This is particularly challenging because we selected species with very different flower morphologies. Careful developmental studies are needed to establish the structures that are most comparable in different species. Figure 6 shows pictures of two of the five species for which we are running microarray experiments (California poppy and avocado) and the model species, *Arabidopsis*, showing the differences in floral organ arrangement.

Our budget allowed us to purchase about 16 arrays per species, as well as a few extras to run one or two pilot arrays or use as back-ups. As two samples can be placed on each microarray, this allowed 32 RNA samples on 16 arrays for each species.

The hybridization intensities for the two samples hybridized to the same array are much more highly correlated than samples hybridized to different arrays. As well, there are reports in the microarray literature that some genes have an affinity for one or the other of the two dyes, making it imperative to ensure each tissue has an equal number of samples labeled with each dye. Because of this, we needed to determine which samples should be paired on the array and, on each array, which samples are labeled with which dye.

A commonly used graphic for microarray studies is the arrow diagram, as displayed in Figure 7. Each arrow is one microarray, with the red sample at the tail and the green sample at the head.

One common experimental design uses a reference sample on each chip—often a mixture of RNA from all the organs in the study. In this "reference

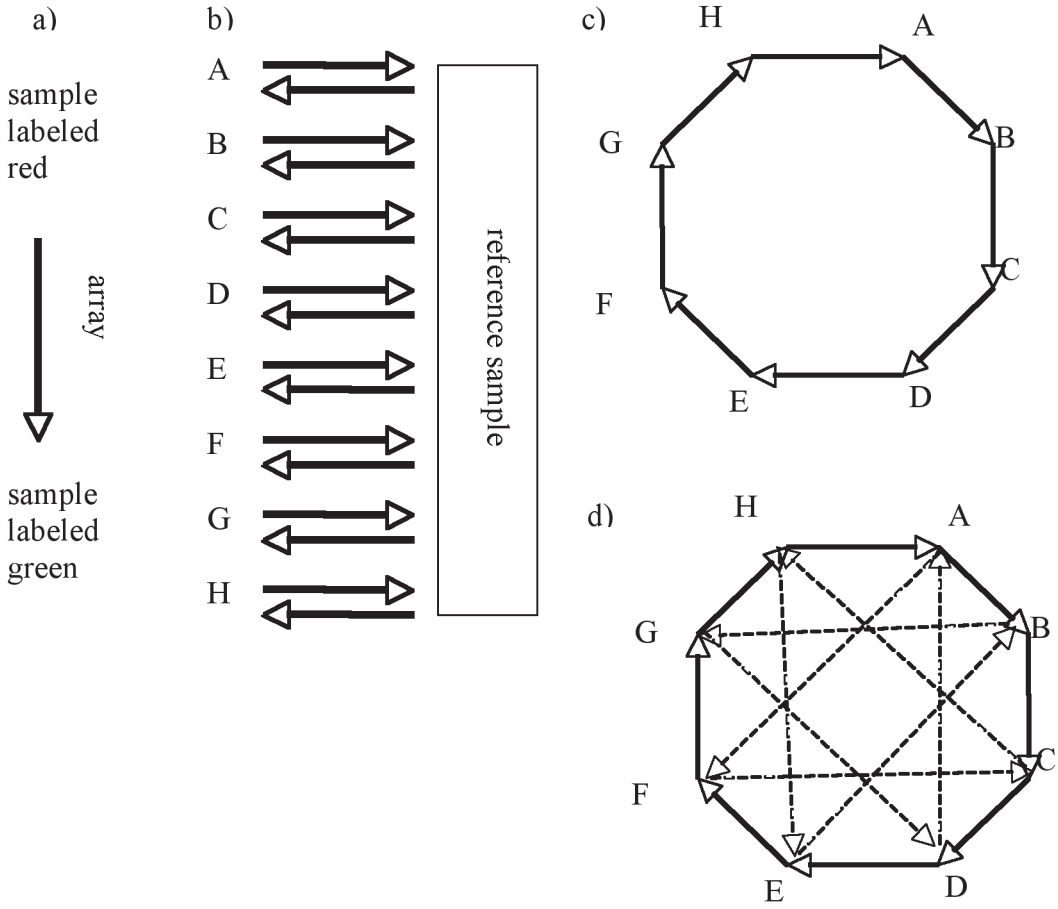


Figure 7. **Two possible designs for a microarray study with eight tissue types and 16 microarrays. Each arrow represents one microarray. The head of the arrow represents the green channel; the tail represents the red. The letters represent the type of tissue sample that is hybridized to the array. a) Using an arrow to represent a microarray. b) A dye-swap reference design with two samples per tissue. c) A loop design with two samples per tissue on eight arrays. This can be run twice to obtain 16 arrays with four samples per tissue. d) A double loop design with four samples per tissue.**

design," the unit of study is the log ratio of the tissue sample to the reference sample for each spot (e.g., Wise, Hardin, and Hoopes in the article starting on Page 40). In a reference design, we really have only one sample for each tissue on each array, as the second sample is the reference sample. With 16 microarrays, this means we could have two replicates per tissue, one labeled with each dye. The design is shown in Figure 7b.

Another common design is a loop design. In this design, there is no reference sample. Instead, each tissue appears twice in the loop, labeled once with each color. This is shown in Figure 7c. A loop design with eight tissues uses eight microarrays, so we could run two full loops with four replicate samples for each organ for the same effort and cost as two replicate samples in the reference design.

In choosing among the available arrangements, we concentrated on obtaining the maximum precision for comparisons between tissues, using standard ANOVA computations for the variance of a comparison. We assumed the samples on the same array are correlated. For several values of this correlation, Table 1 shows the maximum variance for a pairwise comparison between two tissues for three designs with eight tissues and 16 microarrays: the reference design, two copies of the loop design in Figure 7c, and the interwoven loop design in Figure 7d. The low power of the reference design is due to the fact that half the samples are the reference sample, used only for comparison purposes. We did consider using this design with leaf RNA as the reference sample, as leaf is the tissue that is most comparable across species. However, because the hybridization ratio is the unit of comparison, spots with no hybridization for the reference sample cannot be used. For this reason, RNA from leaf—the tissue most different from the source of the ESTs—is not suitable for use as the reference sample. Because a mixed tissue reference would not assist us in cross-species comparisons and the reference design is less powerful for within-species comparisons, we chose to use the interwoven loop design.

One criticism of loop design is that it is difficult to add a new tissue to the loop. However, for one of our species,

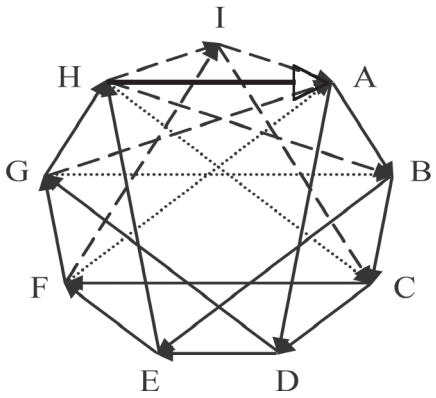


Figure 8. **A new tissue (I) can be added to a double loop design by removing four arrays and adding six. This maintains the balance of the original design. Arrays that have been added are denoted by dashed arrows; deleted arrays are denoted by dotted lines.**

we recently decided to add root tissue to the experiment. This was readily done by adding six new arrays, as shown in Figure 8.

An important issue for all experiments is clear understanding of the role of experimental replication. When M. K. Kerr and G. A. Churchill introduced loop designs in the *Biostatistics* article "Experimental Design for Gene Expression Microarrays," they labeled each RNA sample with both dyes so each sample appeared on two arrays in each loop. Their experiment did not have biological replication within tissue. If only one plant is sampled, then conclusions can be drawn about only the particular individual used in the experiment. Usually, more general conclusions are desired, and, for this, biological replicates (independent plant samples) are necessary. To obtain sufficient RNA for analysis from some of the tissues, we needed to pool tissue from several plants. Hence, we created four independent pools of plants and performed four independent RNA extractions and labelings for each tissue.

Results

So far, we have completed the poppy microarray experiment and have done one of the loops in the avocado experiment. Unfortunately, there are no published data for the same floral organs in other species for comparison. However, a recently completed *Arabidopsis* experiment includes some roughly comparable tissues, including leaf, silique (the seed pod, which is comparable to fruit), anther (part of the stamen), young inflorescence (roughly comparable to young bud), and stage 12 flower (a more mature floral stage than medium bud). Poppy and *Arabidopsis* have the same flower organs. However, avocado has tepals instead of sepals and petals. Biological hypotheses about the development of floral organs state that the outer layers of the tepal may be similar to sepals, while the inner layers may be similar to petals, although the similarity between outer and inner tepals is higher than the similarity of sepals and petals.

To match genes across species, we used a sequence matching program called BLAST. This program allowed us to submit a sequence from our unigene sets and find the most similar match to any other gene(s). Since *Arabidopsis* is the most well-studied plant, we used BLAST to match all of our poppy and avocado unigenes to *Arabidopsis* genes and return the closest match with a threshold for the lowest acceptable degree of similarity. Not all poppy and avocado unigenes have a close match to an *Arabidopsis* gene. A basic question we wished to investigate in our cross-species comparisons is whether genes with the highest degree of sequence similarity in two species exhibit similar expression patterns.

Figure 9 shows representative results for genes in poppy and avocado with close relatives in *Arabidopsis*. Figure 9a shows *PISTILLATA*, a gene known to affect the development of the petals and stamens in *Arabidopsis* plants. *PISTILLATA* and *PISTILLATA*-like genes were found to have the same pattern of expression in all three species. It is particularly interesting that the expression of this gene in the inner tepal of avocado is similar to its expression in petals in the other two species, but the expression in the outer tepal differs substantially from both sepal and petal. Because *Arabidopsis*, poppy, and avocado represent distantly related flowering plants, similar expression of *PISTILLATA*-like genes may reflect conserved function in the large group of other plant species in which they are included. Hopefully in the future, these biological hypotheses can be tested using experiments that disrupt or overexpress the *PISTILLATA* gene in species of interest.

Figure 9b shows a gene in poppy, the function of which is unknown, that has close relatives in both *Arabidopsis* and avocado. Although this gene has high expression in the poppy stamen, it shows little specificity to stamen in the other two species. We propose that this gene may have acquired a new function in the poppy stamen that is lacking or operates at a lower level in the other two species.

Figure 9c shows another gene with high expression in poppy stamen that has a close relative in *Arabidopsis* with unknown function. This gene has high expression in the anther of *Arabidopsis*, and, hence, may be an interesting target for our future studies. The relative in avocado has much hybridization expression in all the tissues in the study, which appears to indicate low expression in these tissues.

Conclusions

A search of the *Current Index to Statistics* shows the first statistical paper with the term *microarray* in its name or keywords appeared in 2001, with 48 papers published by 2002 and 175 (not including book reviews and discussions of primary articles) as of mid-2006. Google Scholar (*http://scholar.google.com*) reports more than 172,000 articles with the term microarray. Clearly, there is a growth

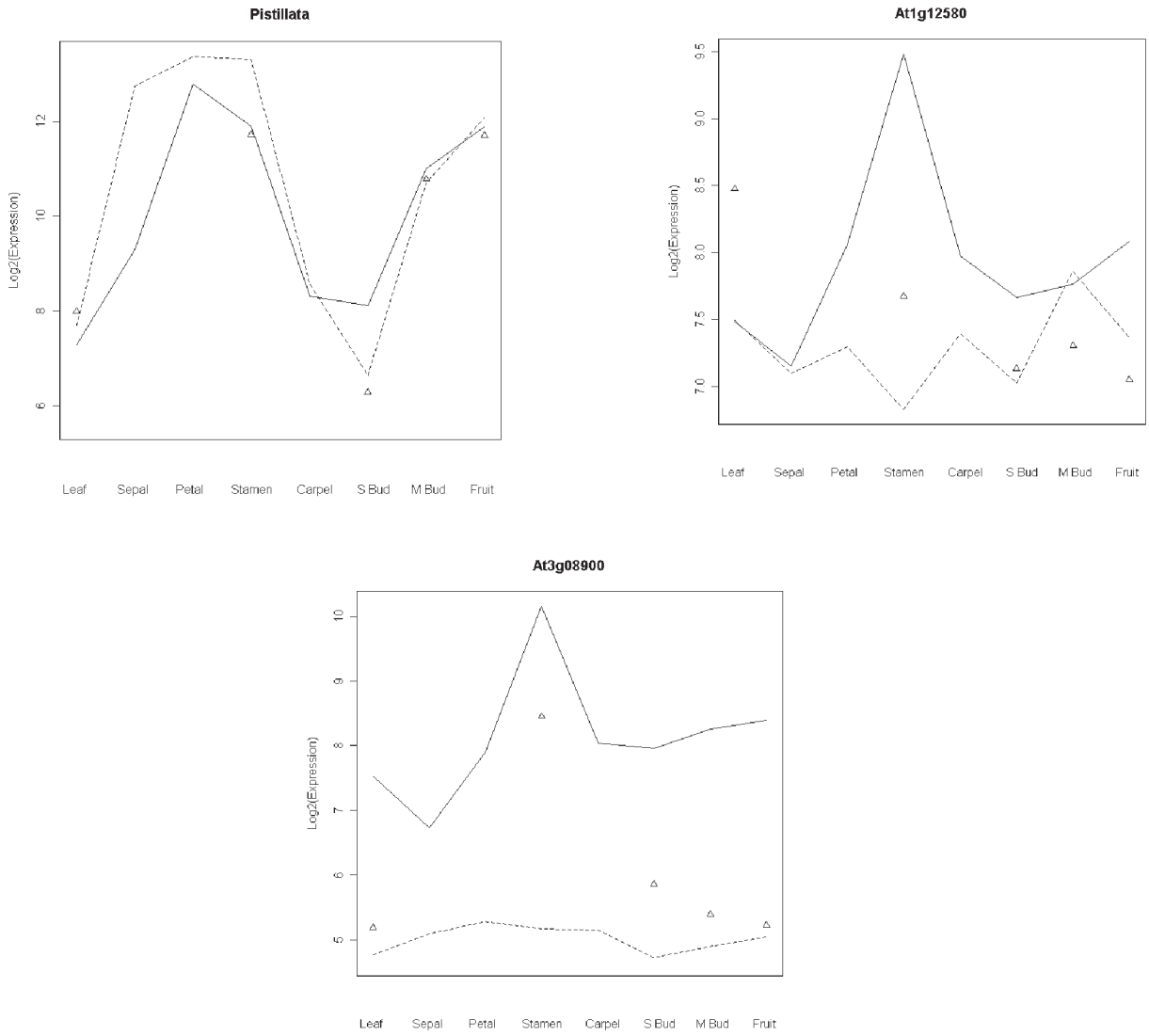


Figure 9. **In each figure, the expression levels are joined by a solid line in poppy, a dashed line in avocado, and triangles in *Arabidopsis*. *PISTILLATA* is a gene known to be involved in floral development in *Arabidopsis*. It has similar patterns of expression in all three species investigated, except in the sepal, which is not present in the avocado flower. The gene At1g2580 has an unknown function in *Arabidopsis*. It is highly expressed in the poppy stamen, but not in the stamens of the other two species. The gene At3g08900 also has unknown function. It is highly expressed in the stamens of both poppy and *Arabidopsis*, but has much lower expression in other tissues. It does not appear to express in any of the avocado floral tissues.**

industry in the use of microarrays for biological discovery and diagnostic testing, and a parallel, if sparser, growth in the involvement of research statisticians in this industry. A brief perusal of the current biological literature also shows rapid evolution of new array-like technologies, including array chips that include all the genomic material from the organism ("tiling" arrays), arrays chips designed to determine how genes are regulated by proteins ("chromatin immunoprecipitation" arrays), and chips to measure protein content (protein arrays).

The statistical literature to date has focused on the analysis of the data after the chips have been scanned. While it is difficult to determine the content of a paper from the title and keywords—and some papers cover multiple topics—a rough count is that the majority of papers fall into five categories: preprocessing (8%), including image analysis and normalization; differential expression analysis (40%) to determine which genes have different expression under different conditions or tissues; sample classification (9%) to distinguish, for example, between normal and cancerous tissues; and gene clustering (7%) to determine which genes have similar expression patterns. There are also a number of introductory or overview

papers. Only 14 of the statistical papers included the word *"design"* in the name or keywords, and none of these considered the selection of genomic material to be printed on the arrays. Another way to look at this is that the statistical literature treats the microarray as a static measurement instrument and is engaged in the design and analysis of experiments using this instrument.

Increasingly, however, design of the measurement instrument (i.e., the spots on the array) will be an activity requiring strong statistical input. This is partly because the design of a microarray will depend heavily on the organism and partly because microarrays are being used for an increasingly complex set of purposes. Microarrays designed to measure gene expression in a single species are not adequate for measuring the regulation of expression by protein binding or other cell mechanisms and may not be adequate for comparing with other species or for detecting important genotypes within the species. The issues involved are on the boundary of statistics and biology and require either cross-training or close collaboration among biologists and statisticians.

One of the objectives of the Floral Genome Project was to develop genomics resources, including microarray chips, to assist in understanding floral development across all lineages of flowering plants. By the end of the project, we will have cDNA libraries and EST data for 13 angiosperms and two gymnosperms, as well as microarray chips for four of the angiosperms and one gymnosperm. Our goal is to determine, both within and between species, which genes are involved in floral development, how the evolution of gene families impacts floral evolution and floral form, and how elements of the cellular biology regulate gene expression. Our desire to perform a new kind of experiment that was focused on cross-species comparisons of the flowering process in plants without complete genome sequences and the need to stay within a budget led us to a series of design choices that were guided by statistical and biological considerations. The analyses will be extremely data-rich, involving thousands of genes for each species (and an increasing number of species as we are able to add information from other plant genomics projects exploring related issues).

References

Albert, V.A.; Soltis, D.E.; Carlson, J.E.; Farmerie, W.G.; Wall, P.K.; Ilut, D.C.; Mueller, L.A.; Landherr, L.L.; Hu, Y.; Buzgo, M.; Kim, S.; Yoo, M.-J.; Frohlich, M.W.; Perl-Treves, R.; Schlarbaum, S.E.; Bliss, B.; Zhang, X.; Tanksley, S.; Oppenheimer, D.G.; Soltis, P.S.; Ma, H.; dePamphilis, C.W.; and Leebens-Mack, J.H. (2005). "Floral Gene Resources from Basal Angiosperms for Comparative Genomics Research." *BMC Plant Biology*, *www.biomedcentral.com/1471-2229/5/5*.

Altman, N.S. and Hua, J. (2006). "Extending the Loop Design for Two-Channel Microarray Experiments." (submitted to *Biostatistics*).

Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; and Lipman, D.J. (1997). "Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs." *Nucleic Acids Res.,* 25: 3389-3402.

Blanc, G. and Wolfe, K.H. (2004). "Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distribution of Duplicate Genes." *Plant Cell,* 16: 1667-1678.

Breiman, L.; Friedman, J.H.; Olshen, R.A.; and Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.

Buzgo, M.; Soltis, D.E.; Soltis, P.S.; and Ma, H. (2004). "Towards a Comprehensive Integration of Morphological and Genetic Studies of Floral Development." *Trends Plant Sci.*, 9: 164-173.

Buzgo, M.; Soltis, P.S.; Kim, S.; and Soltis, D.E. (2005). "The Making of a Flower." *The Biologist,* 52(3):149-154.

Dudley, A.M; Aach, J.; Steffen, M.A.; and Church, G.M. (2002). "Measuring Absolute Expression with Microarrays Using a Calibrated Reference Sample and an Extended Signal Intensity Range." *Proc. Nat. Acad. Sci.*, 99: 7554-7559.

Irizarry, R.A.; Warren,D.; Spencer, F.; Biswal, S.; Frank, B.C.; Gabrielson, E.; Garcia, J.G.N.; Geoghegan, J.; Germino, G.; Griffin, C.; Hilmer, S.C.; Hoffman, E.; Jedlicka, A.E.; Kawasaki, E.; Kim, I.F.; Morsberger, L.; Lee, H.; Petersen, D.; Quackenbush, J.; Scott, A.; Wilson, M.; Yang, Y.; Ye, S.Q.; and Yu, W. (2004). "*Multiple Lab Comparison of Microarray Platforms*. The Berkeley Electronic Press, *www.bepress.com/jhubiostat/paper71*.

Kerr, M.K. and Churchill, G.A. (2001). "Experimental Design for Gene Expression Microarrays." *Biostatistics*, 2: 183-201.

Lee, H.-S.; Wang, J.; Tan, L.; Jiang, H.; Black, M.A.; Madlung, A.; Lukens, L.; Pires, J.C.; Comai, L.; Osborn, T.C.; Doerge, R.W.; and Chen, C.J. (2004). "Sensitivity of 70-mer Oligonucleotides and cDNAs for Microarray Analysis of Gene Expression in *Arabidopsis* and its related species. *Plant Biotechnology Journal,* 2: 45-57.

Wang, J-P. Z.; Lindsay, B.G.; Leebens-Mack, J.; Cui, L.; Wall, K.; Miller, W.C.; and dePamphilis, C.W. (2004). "EST Clustering Error Evaluation and Correction." *Bioinformatics,* 20: 2973-84.

Wang, J.Z.; Lindsay, B.G.; Cui, L.; Wall, P.K.; Marion, J; Zhang, J.; and dePamphilis, C.W. (2005). "Gene Capture Prediction and Overlap Estimation in EST Sequencing from One or Multiple Libraries." *BMC Bioinformatics*, 6: 300.

Yuen, T.; Wurmbach, E.; Pfeffer, R.L.; Ebersole, B.J.; and Sealfon, S.C. (2002). "Accuracy and Calibration of Commercial Oligonucleotide and Custom cDNA Microarrays." *Nucleic Acids Research*, 30(10):e48.

Zahn, L.M.; Feng, B.; and Ma, H. (2006). "Beyond the ABC Model: Regulation of Floral Homeotic Genes." In *Developmental Genetics of the Flower (Eds. Soltis, D.E.; Soltis, P.S.; and Leebens-Mack, J.H.), Advances in Botanical Research (Vol. 44)*. Elsevier Limited: London.

Zhang, X.; Zhang, D.; Feng, B.; Altman, N.S.; and Ma, H. (2005). "Genome-wide Expression Profiling and Identification of Gene Activities During Early Flo*wer Development in* Arabidopsis." *Plant Molecular Biology,* 135: 1084-1099.