

# Using Microarray Data with Comparative Sequence Analysis to Infer Evolutionary Relationships

**Naomi Altman**  
**Penn State University**  
Dept. of Statistics  
[naomi@stat.psu.edu](mailto:naomi@stat.psu.edu)

**March 1, 2006**

## The general problem:

We would like to explore the evolutionary origins of an organ or trait by tracking the expression of associated genes in phylogenetically related species.

e.g. The origins of flowers.

The origins of color vision.

The available data are:

gene expression (primarily microarrays)

sequence

## We will assume:

- the species phylogeny is known
- the gene phylogeny is partially known

## Our main concern will be:

- combining the gene expression data with the phylogenetic information

## The context:

Floral Genome Project: objective: understand the genetic origin of flowers

# Why is this problematic?



We need to combine microarray data from several related species.

- 1) tissue equivalence
- 2) microarray equivalence

# Tissue Equivalence



- When are tissues "equivalent" in different species?

e.g. What tissues in chicken, frog and mouse are equivalent to the brain of a 6 month old human fetus?

# Tissue Equivalence

- When are tissues "equivalent" in different species?

e.g. What tissues in chicken, frog and mouse are equivalent to the brain of a 6 month old human fetus?

This problem is worse in plants which have a lot of organ diversity.

e.g. Flowers may be unisexual or bisexual. Are the petals of the 3 types of flowers equivalent?

# Tissue Equivalence



- When are tissues "equivalent" in different species?

We generally try to control the environment in an experiment.

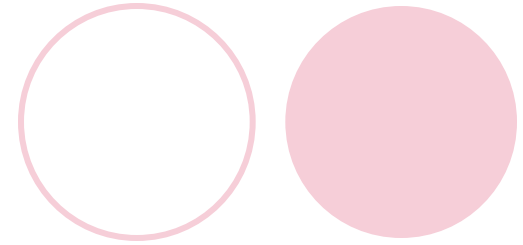
How do you grow a cactus and a water lily in "equivalent" environments?

Fortunately, these are biological problems, not statistical problems (i.e. I only have to raise the issue, not solve it).

# Microarray Equivalence

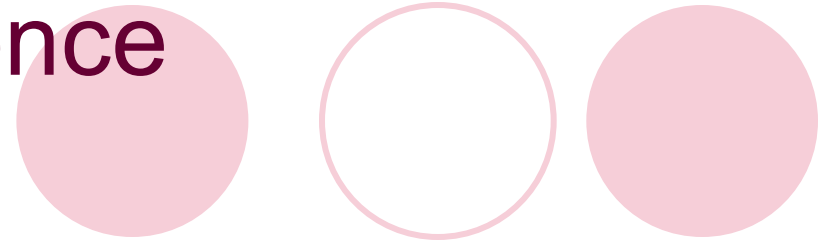
4 problems:

1. Platform effect
2. Study Effect
3. Oligo effect
4. Paralogs/Orthologs



# Microarray Equivalence

Irizarry et al (2004)



sent "identical" RNA samples to 10 labs using 3 commercial platforms

Overlap for "most differentially expressed" genes was 40% - 60%

# Microarray Equivalence

Platform effect

Each species has its own array.

Often these will be derived from ESTs, not whole genomes.

We cannot normalize the arrays together.



# Microarray Equivalence

## Platform effect

We cannot normalize the arrays together.

The Affymetrix Latin Square  
Spike-In Experiment

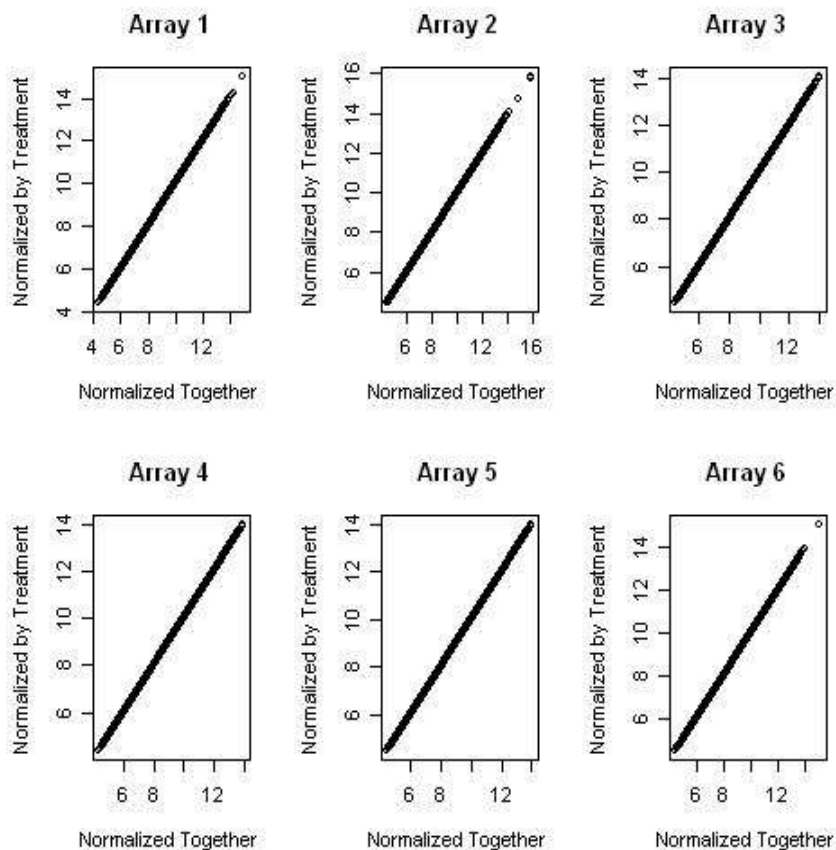
6 arrays, with 3 replicate of 2  
conditions.

Same RNA except for 42 spiked  
in genes.

#significant eBayes t-tests ( $p < .01$ )

Together  
144

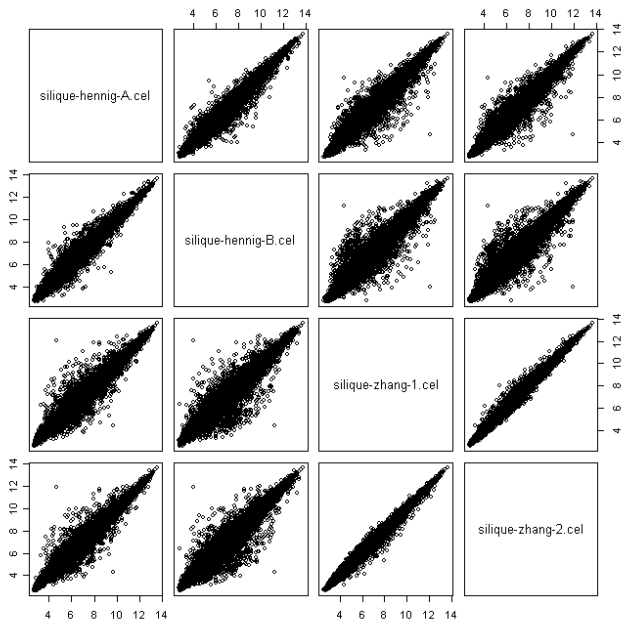
by TRT  
14374



# Microarray Equivalence

Study effect

There is almost always a strong "block" effect due to lab, experiment time, etc.



4 Affy ATH1 arrays, silique

2 from Hennig study, 2 from Zhang study

An eBayes ANOVA finds 2459 genes differentially express ( $p < .01$ ) if we do an ANOVA comparing labs, versus 4 if we just assign 1 array from each lab to "treated".

# Microarray Equivalence

Study effect



There is almost always a strong "block" effect due to lab, experiment time, etc.

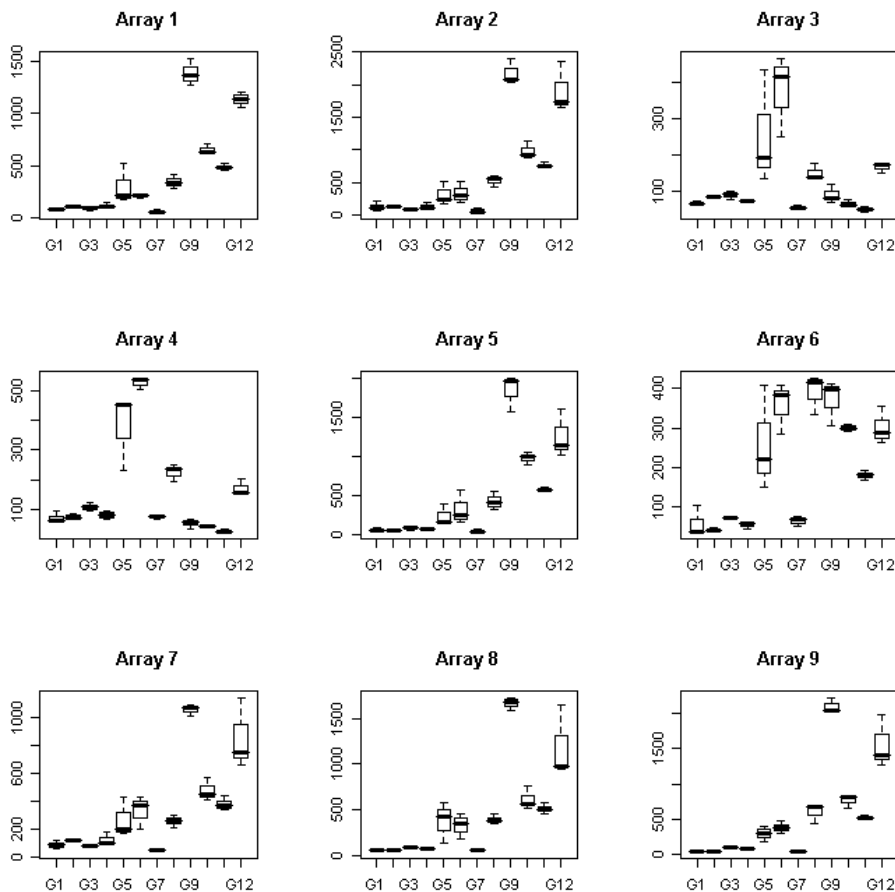
We seldom have the same lab handling different organisms at the same time.

Even the FGP has different labs and personnel handling each species.

# Microarray Equivalence

## Oligo effect

Even for 1 oligo, there are spot to spot effects:

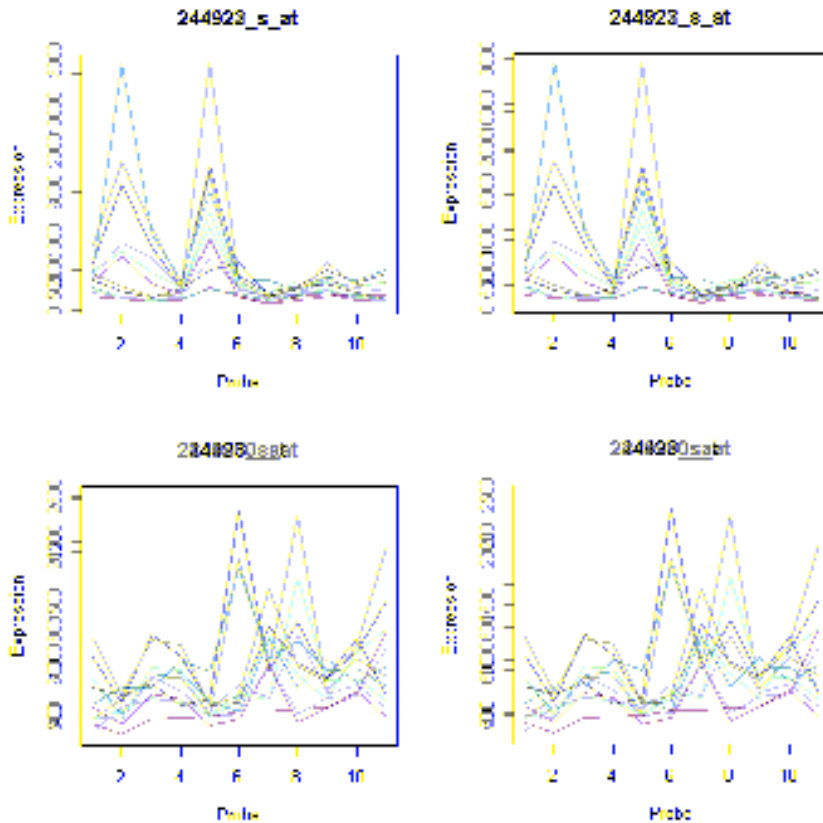


These are 12 spots, 3 replicates per spot after normalization on 9 arrays in the Red channel from a loop design experiment. Some replicates are highly variable. (R interpolates to obtain the boxplots.)

# Microarray Equivalence

Oligo effect

With multiple oligos per gene, there can be a strong oligo effect.



These are 4 perfect match probe sets from 8 Affymetrix arrays. There is a strong oligo effect even from the same gene.

# Microarray Equivalence

Oligo effect



With arrays created for different species, we are unlikely to use the same oligos for homologous genes across the species.

# Microarray Equivalence

## Paralogs/Orthologs

We generally identify orthologous genes by BLAST or other algorithms that score genes by similarity. In genomes that have undergone duplications, there may be many paralogous genes. Paralogs are maintained in the genome by functional differentiation of various sorts. In comparing incompletely sequenced organisms, the "best match" may be the paralog of the ortholog. The ortholog may have been lost.

# Microarray Equivalence

Paralogs/Orthologs



We used TribeMCL to classify the complete genomes of rice and *arabidopsis*.

Clusters of 2, within *arabidopsis* were interpreted as paralogous pairs occurring after divergence from rice.

These were confirmed using phylogenetic analysis. 280 pairs represented on the ATH1 array were selected.

# Microarray Equivalence

Paralogs/Orthologs

We used ANOVA to investigate the gene by tissue interaction in **280** pairs of close paralogs of transcription factors in *arabidopsis*.

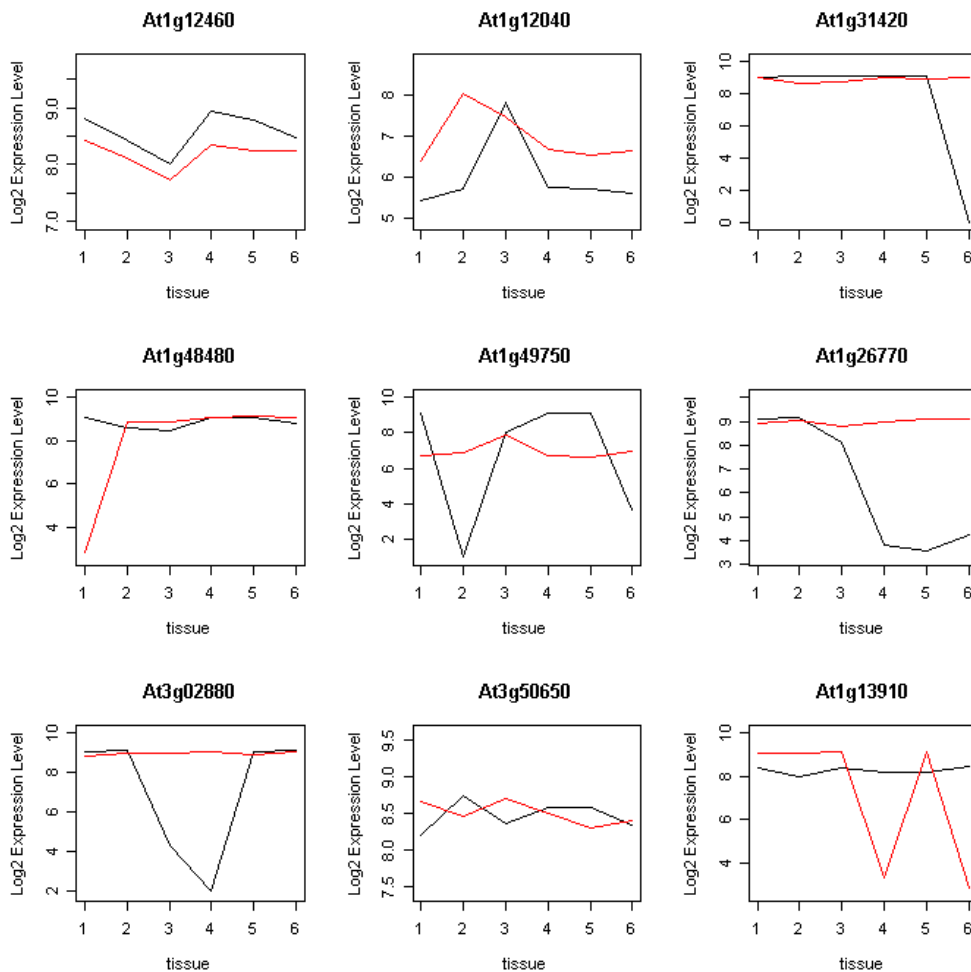
**85%** of the pairs had significant gene by tissue interactions indicating one of:  
non-functioning of one gene  
partitioning of function between the 2 genes  
different function of the 2 genes

Duarte et al, 2005,

# Microarray Equivalence

## Paralogs/Orthologs

Duarte et al, 2005,



Some of the pairs showing some of the expression patterns in 6 tissues

# Floral Genome Project

Currently: microarrays for 3 plant species with 3 more species to follow by summer (and hopefully another 5

arabidopsis

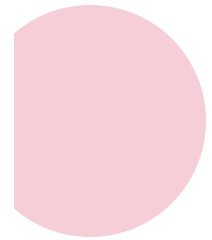
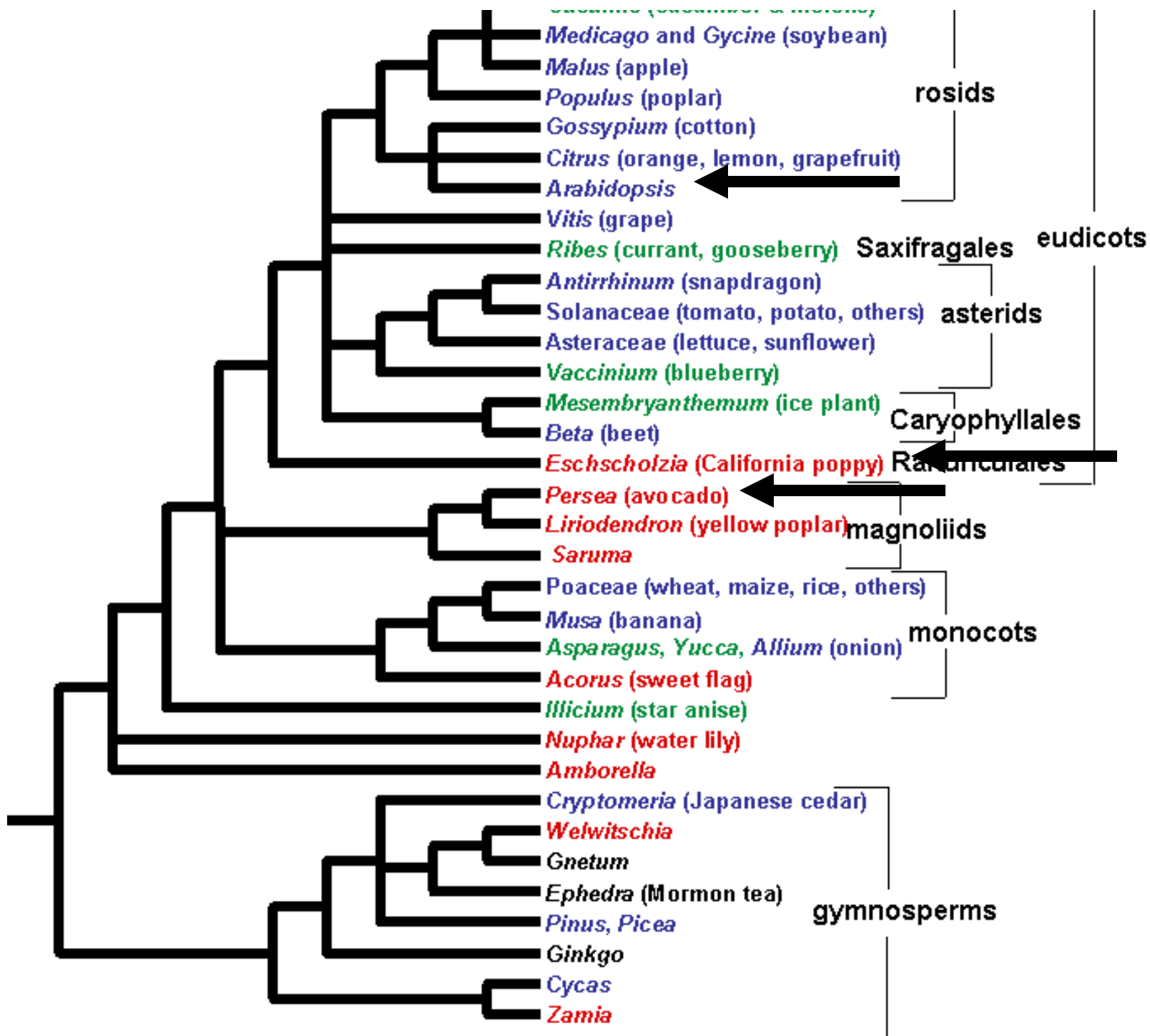


avocado



California  
poppy





# Feasibility of Exploring the Evolution of Gene Expression

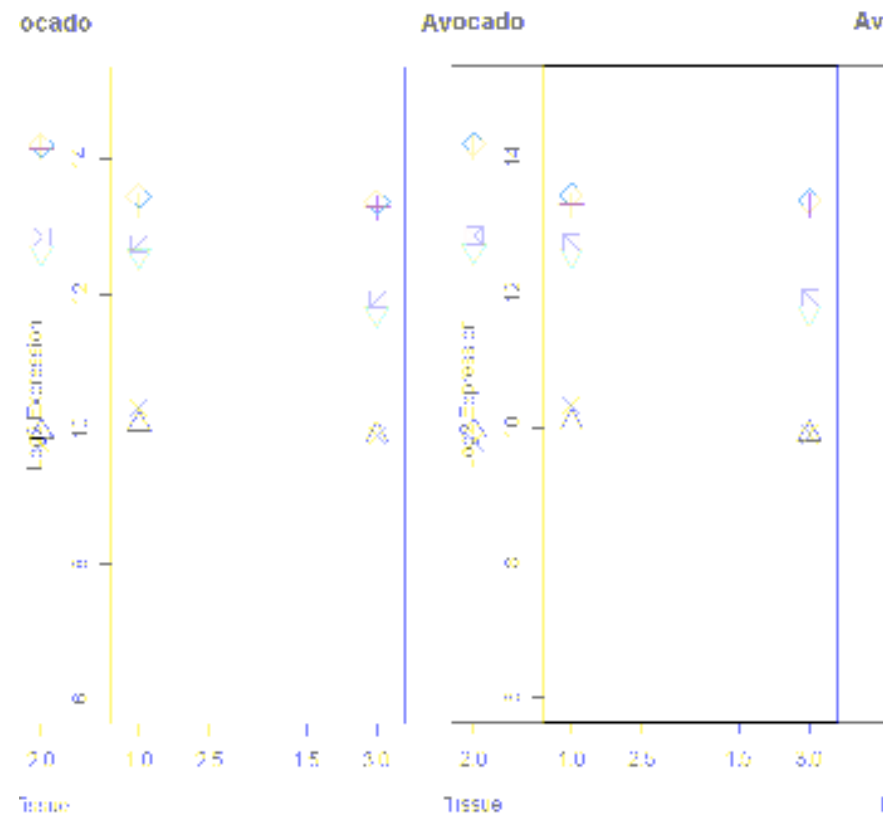
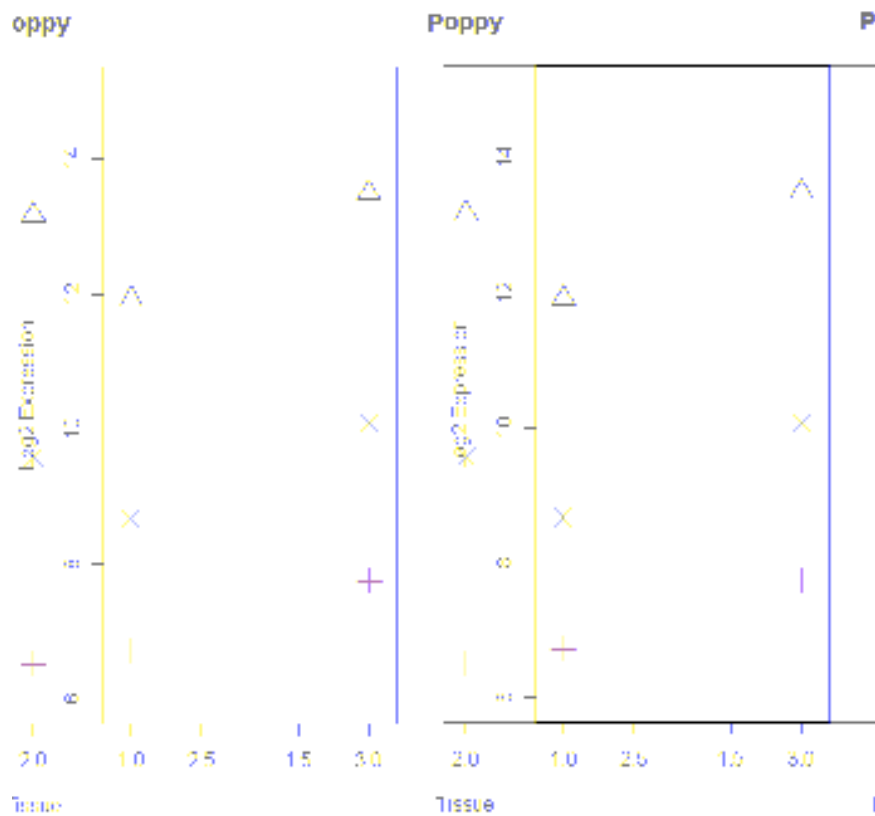
What we have done to date:

- Partially sequence ESTs from species selected for phylogenetic position
- BLAST resulting unigenes against fully sequenced species
- printed microarrays with several oligos per gene or spot replicates
- within species analysis of differential expression
- heuristic matches of expression patterns (gene by gene)

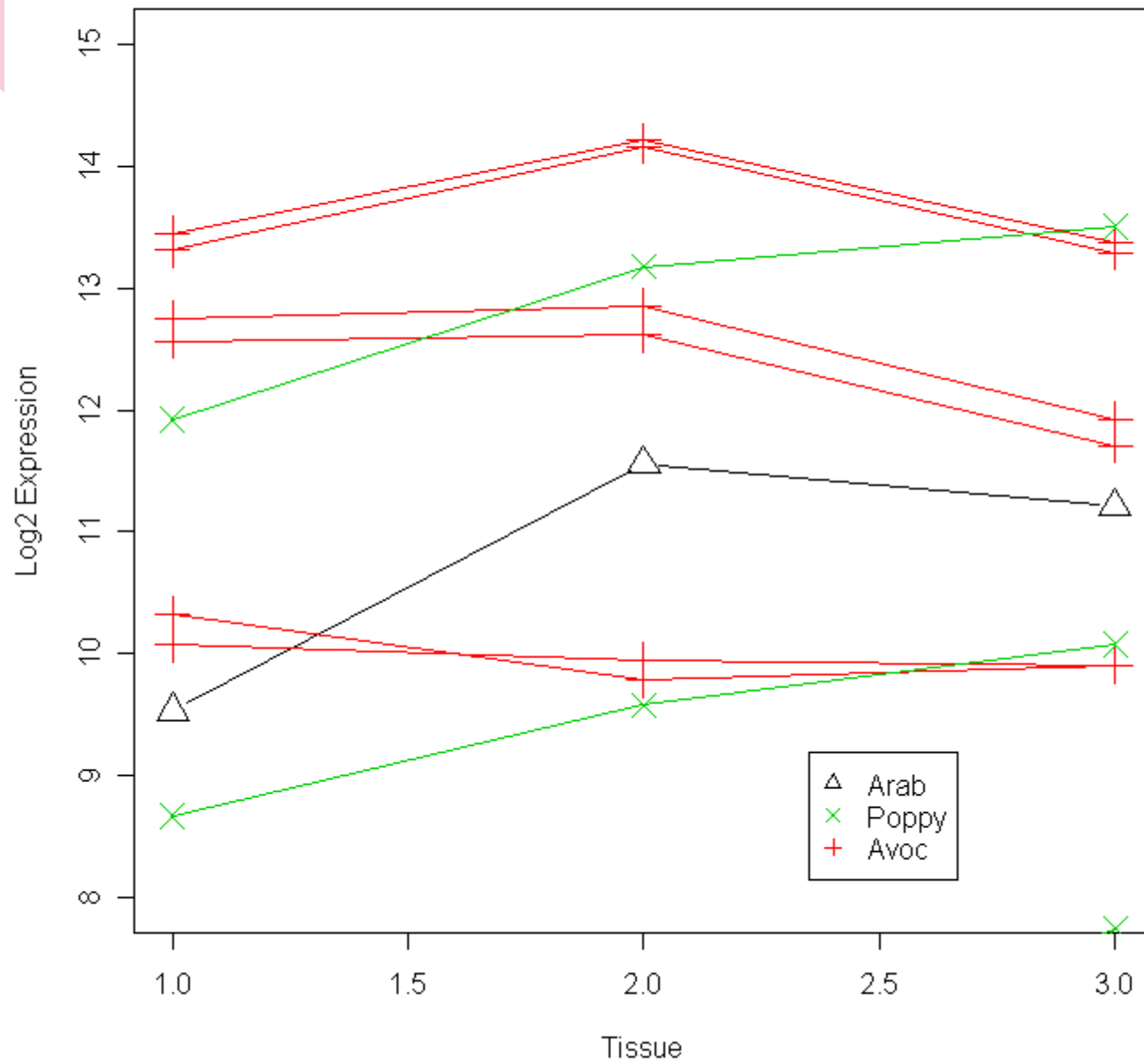
# Expression for Homologs of At1g09210

Poppy: 3 oligos

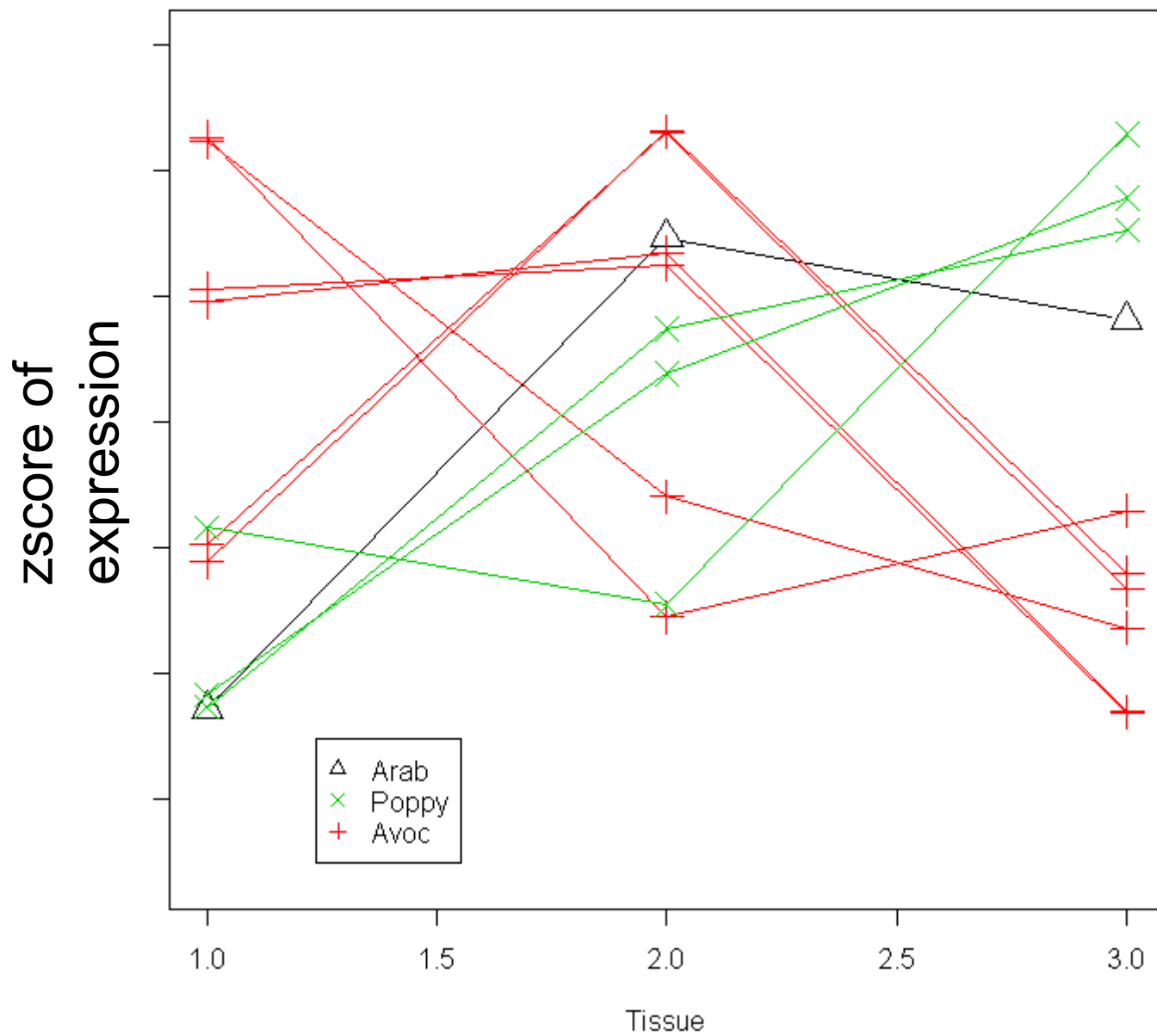
Avocado: 3 oligos  
printed in duplicate



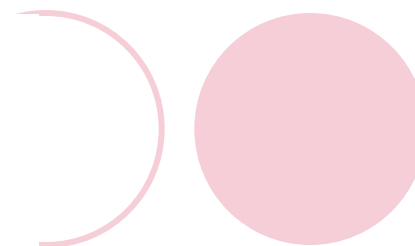
### 3 Species



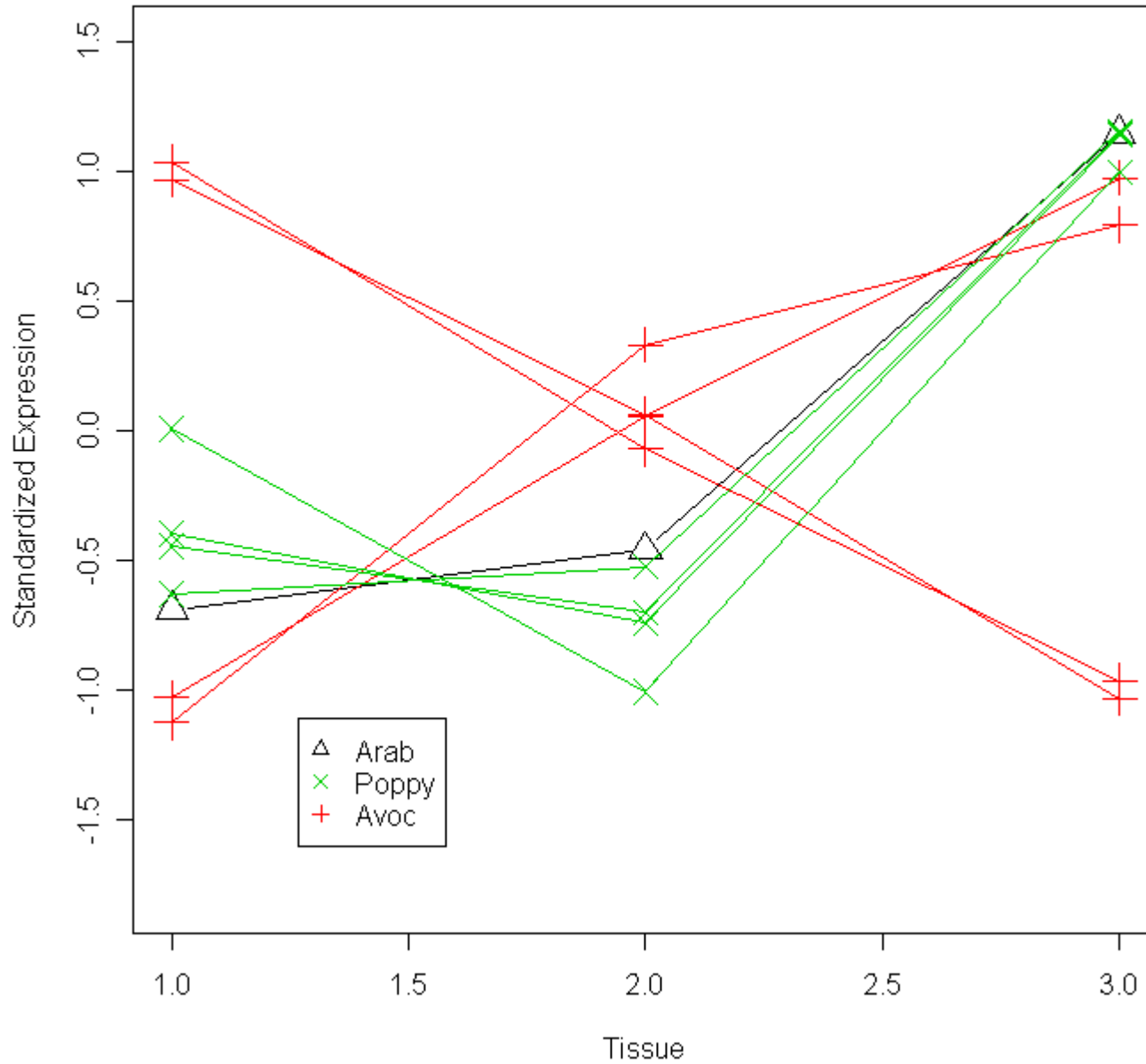
### 3 Species



homologs of  
At1g09210



### 3 Species



homologs of  
At1g14610

(2 oligos per  
species, each  
spotted twice)

# Conclusions



This problem will not be solved by statistics alone.

Components:

1. Microarray oligo design
2. Sequence matches across species
3. Phylogenies to identify paralogs
4. ANOVA-type analysis based on ranks or standardized data or ... which takes into account random effects for arrays, oligos within gene, etc.

# Conclusions



This problem will not be solved by statistics alone.

Normalization will be important, but we have to be careful not to introduce false positives.

Tissue matching will be critical - e.g.

in arabidopsis, we have "silique" which is the seed pod

in avocado, we have the avocado fruit

And many thanks to  
The Floral Genome Project

[fgp.huck.psu.edu](http://fgp.huck.psu.edu)

especially:

Laura Zahn (poppy)

Andre Chanderbali (avocado)

Xiaohong Zhang (arabidopsis)

Jill Duarte (analysis of paralogs)

Kerr Wall (bioinformatics)

Jim Leebens-Mac (project manager and project guru)

Hong Ma (co-PI)

Claude dePamphilis (PI and project grand guru)