Correlated Expression Analysis of Plant Genes using Regression

Roshonda Barner, North Carolina A&T State University

Dr. Ann Loraine, University of North Carolina-Charlotte

Background Information

- A microarray machine measures the mRNA abundance in a gene.
- mRNA reading is called the gene's expression value
- When a gene is active in a role, the expression value will be high.
- When two genes are involved in the same role, or have the same function, they will be highly correlated.

Purpose of Research

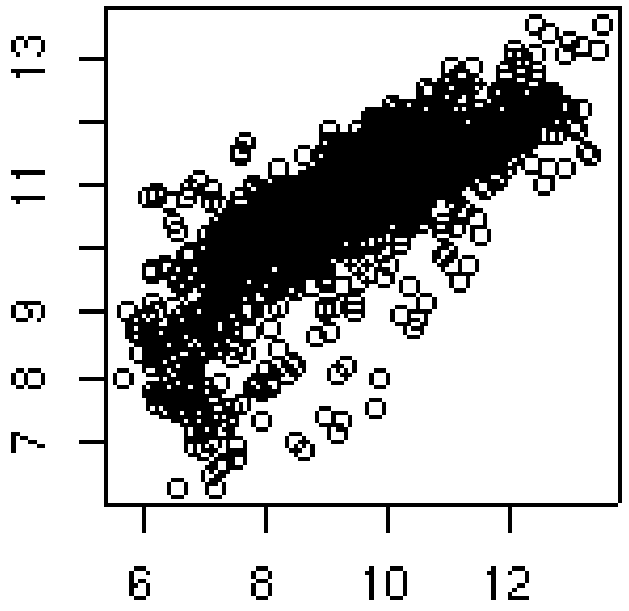
- Biologists' intuition is genes that play similar roles in the cell have a closer relationship in higher expression values.
- I want to find out if this relationship is true.
- If the genes are not more correlated in higher values, how can we filter the expression values to find out how much genes correlate.

Materials and Methods

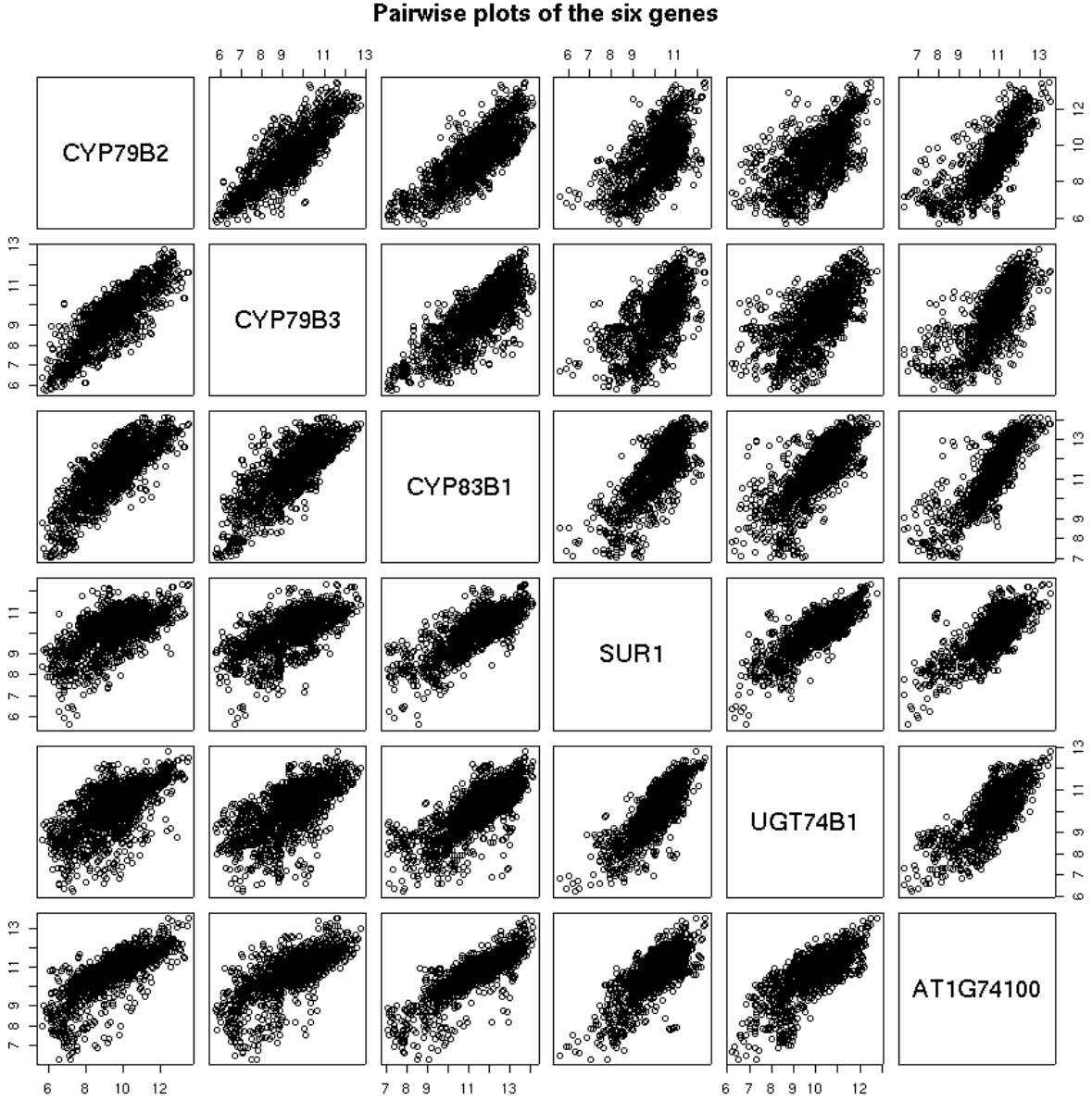
- I want to find out how much the genes' expression values relate to each other.
- To do so I picked six genes and 1776 expression values for each of those genes and ran a pairwise regression of the six genes.

Materials and Methods (continued)

- R - a computer language for statistical computing and graphics.
- Emacs - a text editor, to write the functions that I want to run in R. I used the function "source( )" in R to run what I wrote in Emacs.



- This is a closer look at two of the six genes plotted against each other.
- Notice that in very low expression values the data tends to be very heteroskedastic.


Pairwise plots of the six genes
CYP79B2
CYP79B3
CYP83B1
SUR1
UGT74B1
AT1G74100

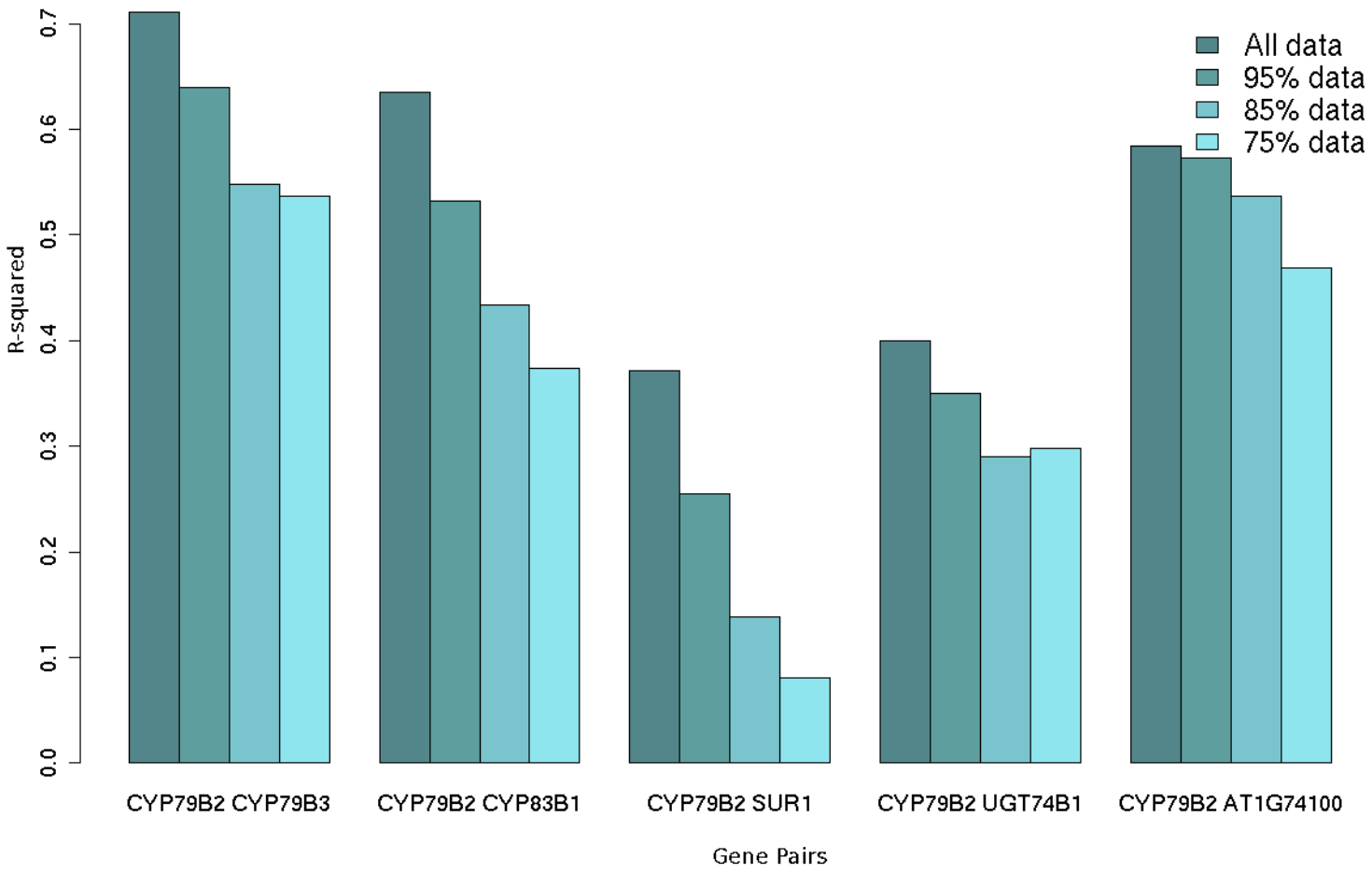
Filtering by Expression Values

- Since the plots seem to show a lot of heteroskedasticity then I will filter out the lower parts of the data and just analyze the upper data and see determine if the r-squared increases from there.
- I want the r-squared to increase because this number tells us what percentage of the variation in y is explained by the regression model. If the number is closer to one then most of the variation in y is explained and if the number is closest to zero then less of the variation in y is explained by the model.

Filtering by Expression Values Results

Gene 1	Gene 2	all	top95	top85	top75
CYP79B2	CYP79B3	0.71	0.64	0.55	0.54
CYP79B2	CYP83B1	0.64	0.53	0.43	0.37
CYP79B2	SUR1	0.37	0.26	0.14	0.08
CYP79B2	UGT74B1	0.4	0.35	0.29	0.3
CYP79B2	AT1G74100	0.58	0.57	0.54	0.47
CYP79B3	CYP83B1	0.63	0.52	0.44	0.38
CYP79B3	SUR1	0.41	0.32	0.24	0.18
CYP79B3	UGT74B1	0.44	0.42	0.34	0.3
CYP79B3	AT1G74100	0.52	0.47	0.38	0.34
CYP83B1	SUR1	0.6	0.49	0.34	0.23
CYP83B1	UGT74B1	0.59	0.56	0.46	0.36
CYP83B1	AT1G74100	0.71	0.65	0.61	0.53
SUR1	UGT74B1	0.69	0.64	0.52	0.41
SUR1	AT1G74100	0.58	0.44	0.26	0.16
UGT74B1	AT1G74100	0.61	0.52	0.39	0.32

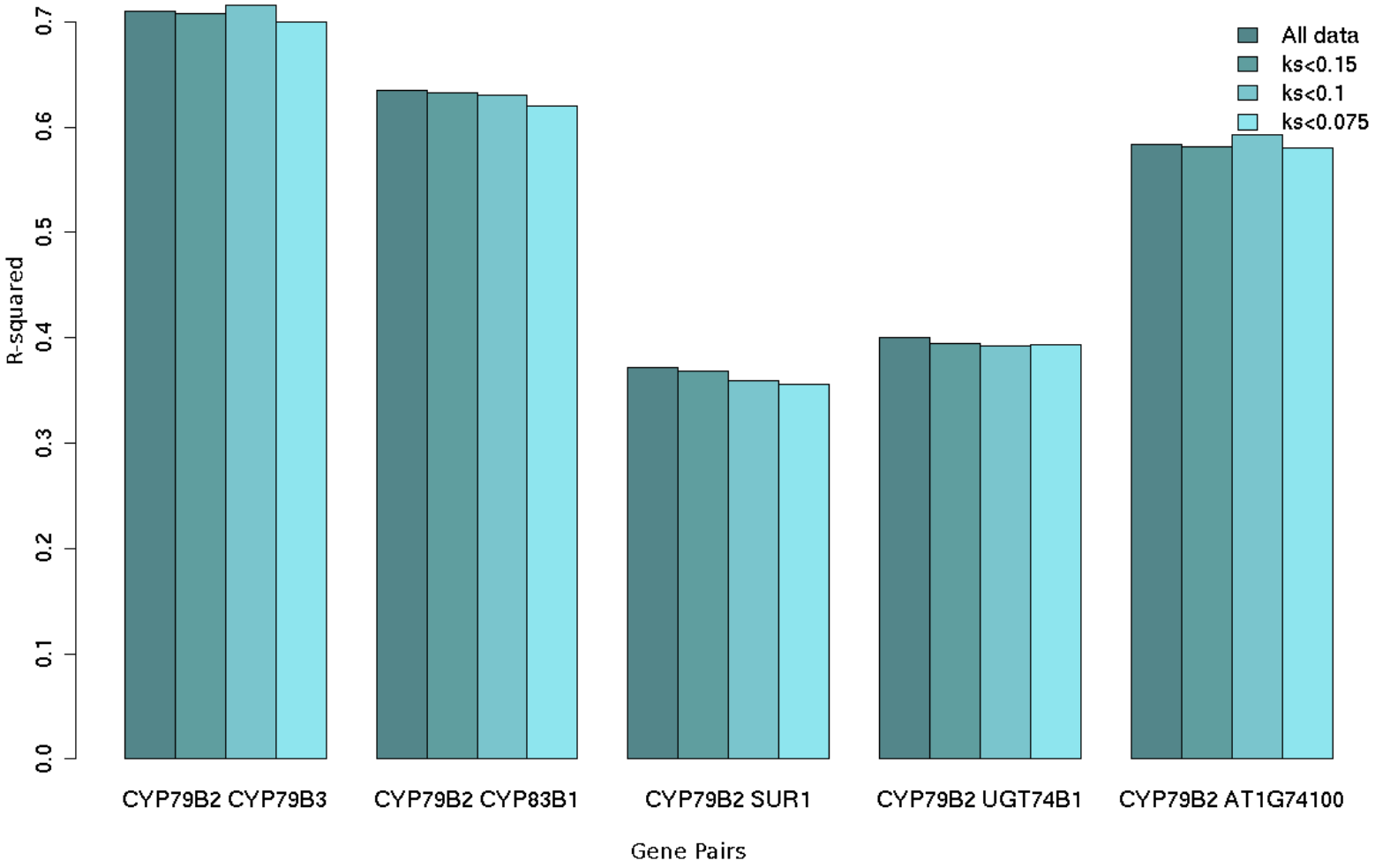
Filtering By Expression Values Results



Filtering By Kolmogorov-Smirnov (K-S) test statistic

- Now I will remove the data by the K-S test statistic.
- The K-S test statistic basically tells us how much of an outlier a gene's expression value is compared to the other expression values of that gene.

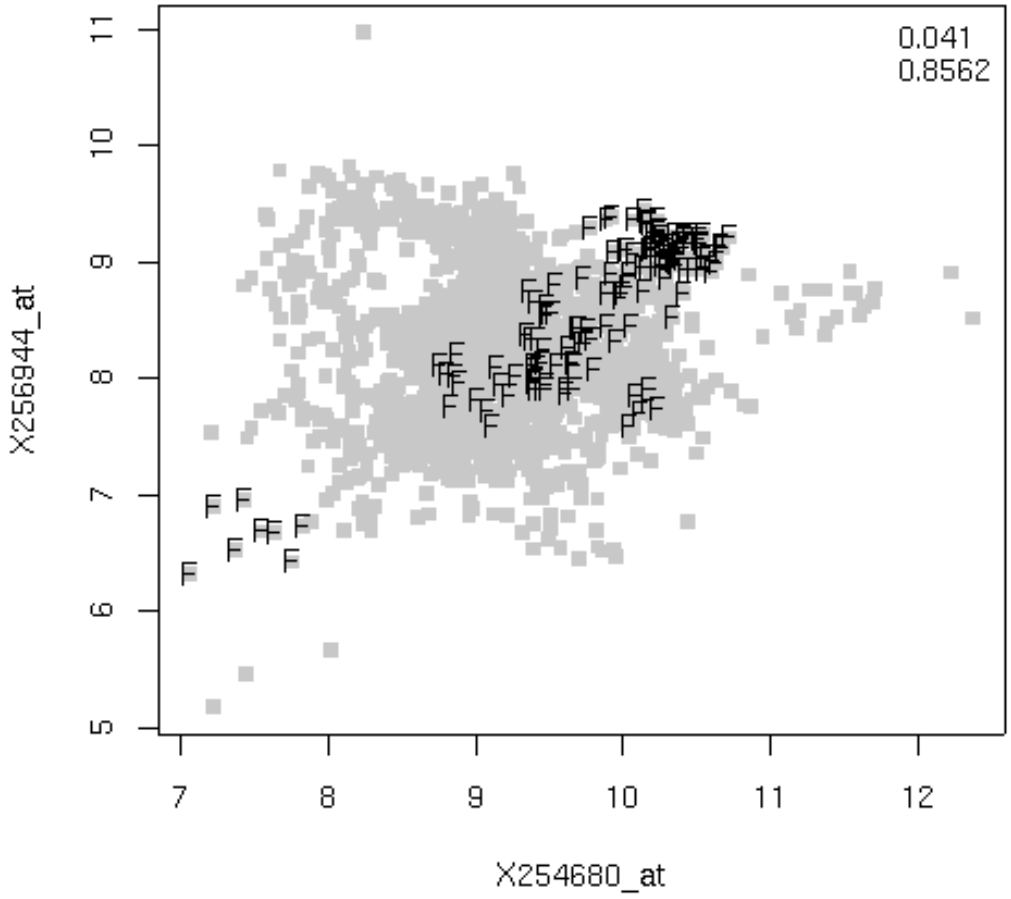
Filtering By Kolmogorov-Smirnov (K-S) test statistic



Filtering by the Sample Type

- For this third method I want to filter by the sample type.
- To execute this I look at 168 genes in the Arabidopsis Thaliana plant and I want to just look at the genes in which the sampling came from flowers.

Filtering by Sample Type Results



Conclusion

- Filtering by expression values seemed to be a not so good approach to finding correlations between genes.
- Filtering by the K-S test statistic did not seem to help or hurt the model.
- Filtering by the sample type was the best filtering method.
  - This makes sense and the reason it does is because if one wants to look at genes involved in flowering he should look at just genes from the flower not all parts.

Acknowlegments

- I would like to thank Ann Loraine of University of North Carolina-Charlotte Bioinformatics Department for helping me perform the research.