Using ANOVA to evaluate differential protein expression

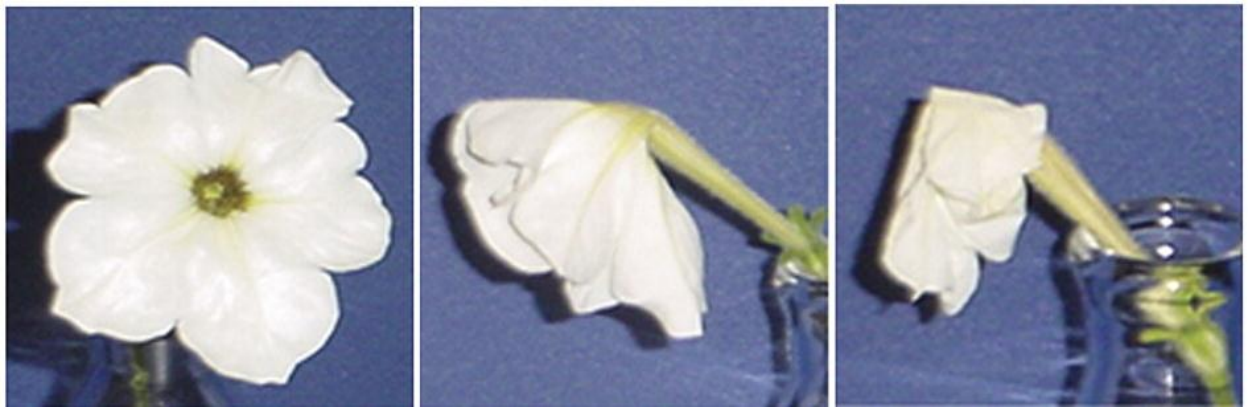
Shaun Broderick

The Ohio State University

OARDC

The research question: Which proteins are differentially expressed during flower senescence?

- Senescence is an active stage of flower development
- Senescence activates the degradation of floral proteins
  - Proteases, nucleases, electron transport proteins, lipoxygenases
- Remobilization of nutrients
  - transmembrane proteins (i.e. pumps)



Why petunias?

Petunias serve as a model organism for studying flower senescence because they have a well-defined senescence pattern

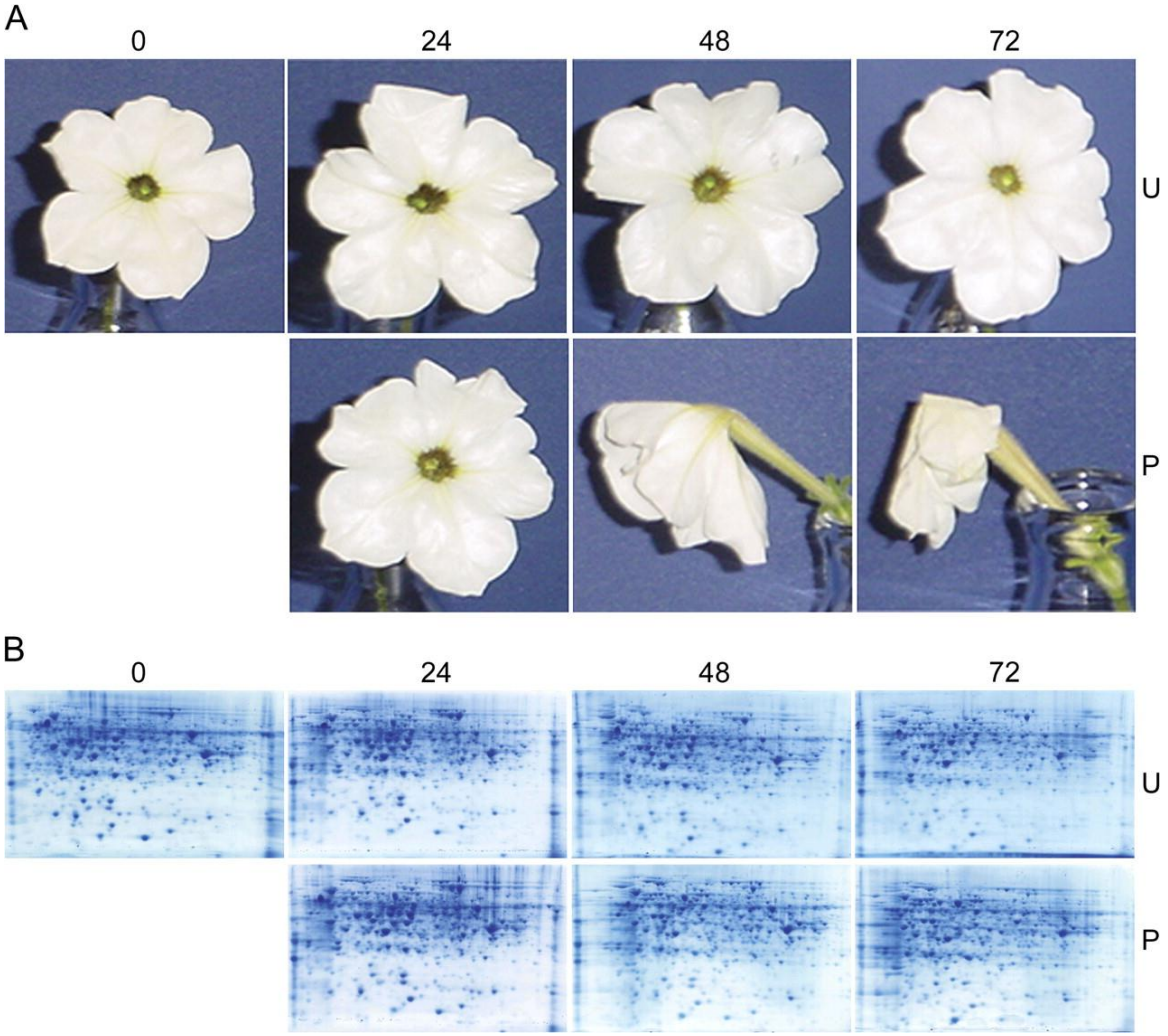


Figure Credit: Bai et al., 2010 © 2010 The Author(s).

Experimental design for protein analysis

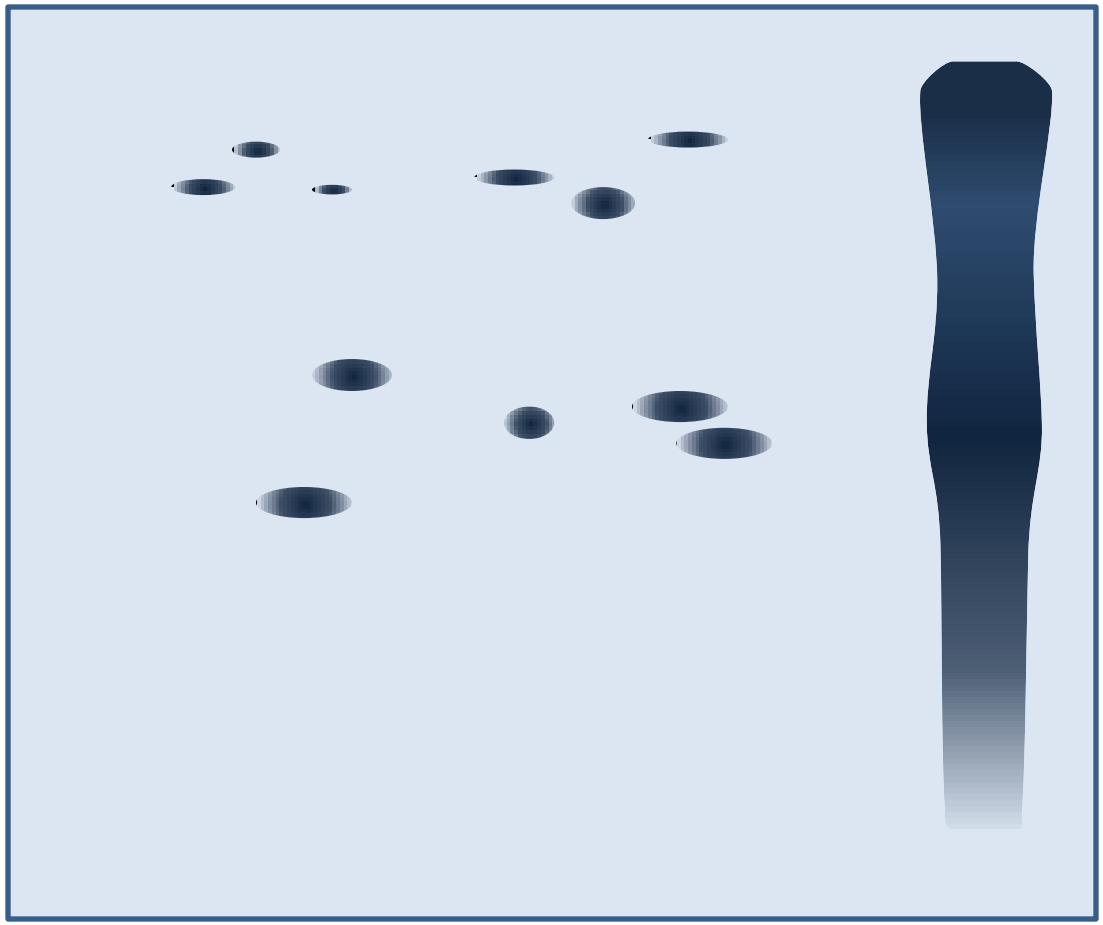
- Treatments: pollinated and unpollinated
- Senescence cycle is completed in 72 hours—corollas will be collected at 0, 24, 48 and 72 hours
- 8 corollas will be pooled from at least 3 plants for each time point
- Three 2D gels will be poured in each rep
- Total soluble protein will be extracted from corollas and separated on 2D gel
- Protein spots will be cut from gel and sequenced

2D-gel electrophoresis: Finding differentially expressed proteins

- Extract all soluble proteins from senescing corollas
- Extract all soluble proteins from non-senescing corollas of the same developmental stage
- Proteins can be separated by their physical properties of net charge and molecular weight

Protein changes visualized by 2D gels

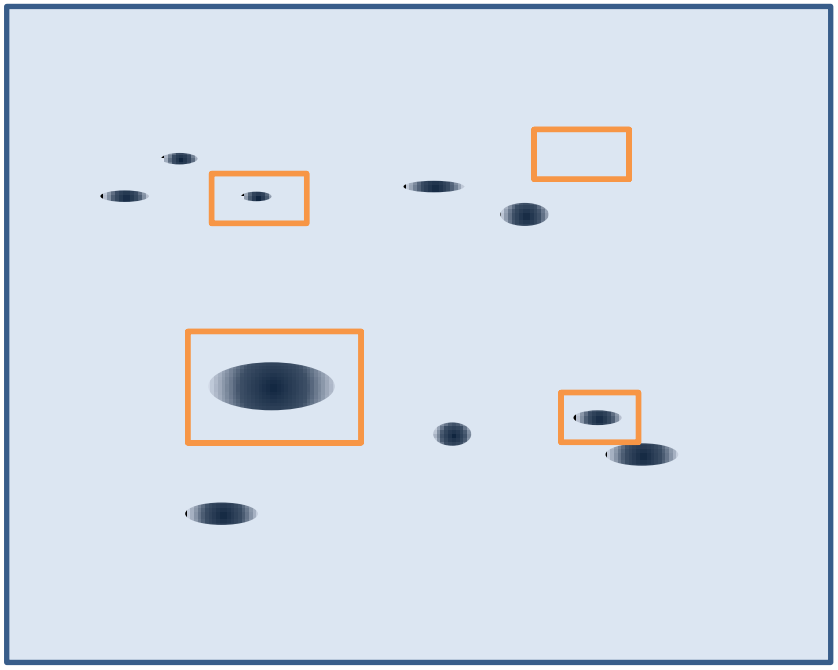
2nd separated by mass
(smaller proteins move faster)



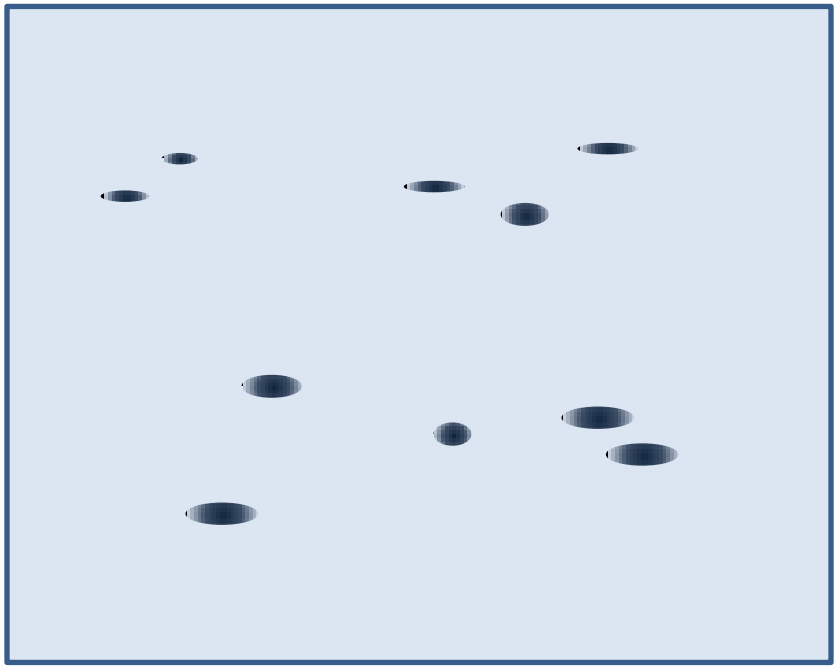
1st separated by their isoelectric point (i.e. charge)

Pollinated vs. unpollinated proteins

Pollinated



Unpollinated



Potential problem: Variation in protein extracts will change protein patterns and amounts;
how do we set an objective cut-off for “differential expression”?

Data analysis overview

- Use SAS to rearrange a data set
- Evaluate distribution of data
- Transform data
- Estimate an experiment-wide cutoff

Analyzing the data

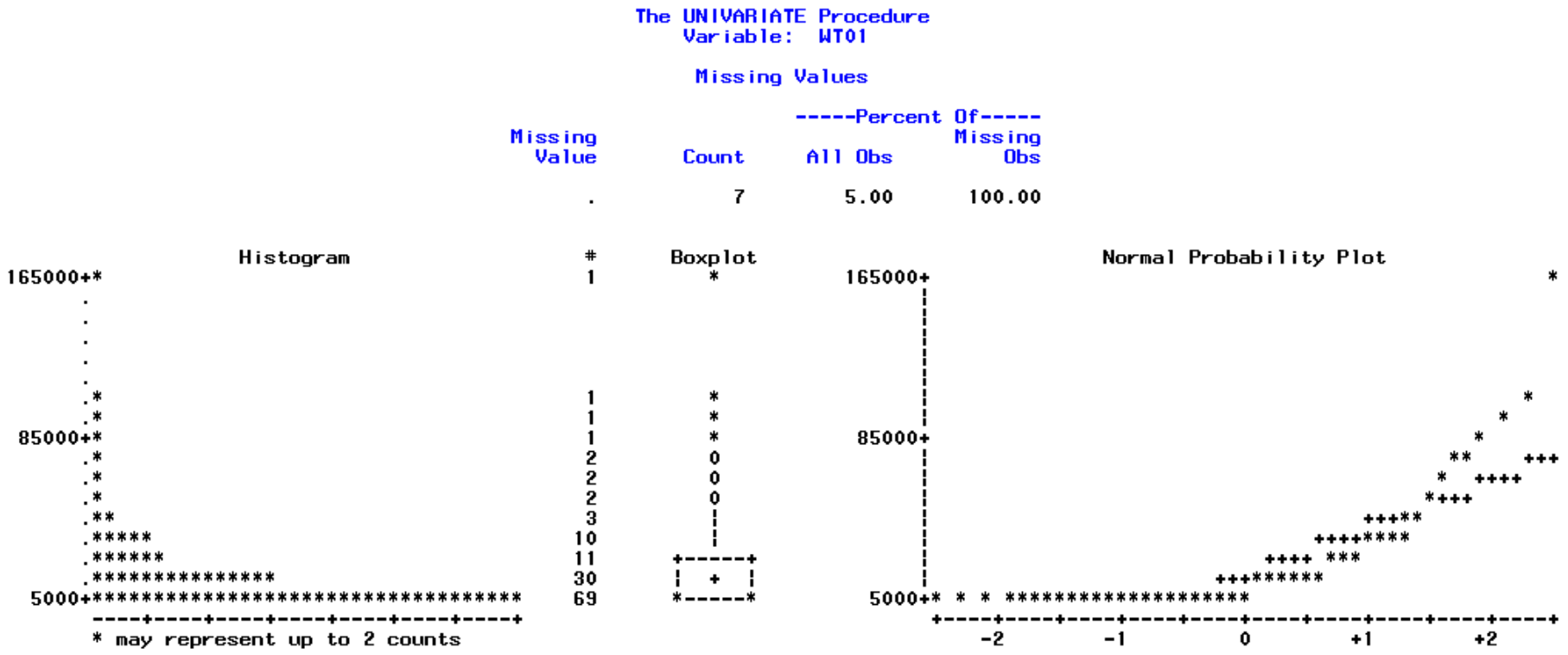
- What analyses would be most appropriate?
- Do the data fit the assumptions for ANOVA (normality)?
- Transform data (either in Excel or SAS)
- How will the data need to be structured?
- What are the main factors in the model?
- Statistically determine the cutoff point (i.e. least significant difference [LSD]) for statistical significance

Are the data normally distributed?

Useful SAS code for visualizing the data

```
options formdlim='~';
data one;
  title 'Bai Protein Expression data - sequenced Proteins';
  infile 'C:\Documents and Settings\joneslab\Desktop\SeqProt.csv'
dlm=','  firstobs=2 ;
  input protein $ SSP WT01 WT02 WT03 WTU241 WTU242 WTU243 WTU481
WTU482 WTU483  WTU721 WTU722 WTU723 HAP241 HAP242 HAP243 HAP481
HAP482 HAP483 HAP721 HAP722 HAP723;
Proc univariate normal plot;
       var WT01 WT02 WT03 WTU241 WTU242 WTU243;
       run;
```

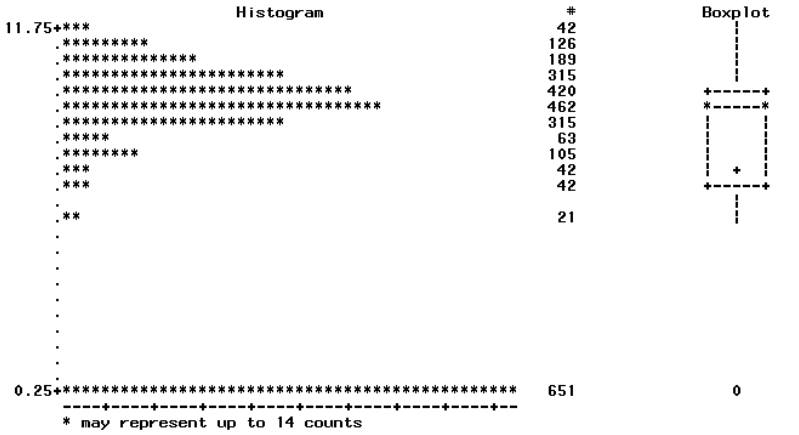
Are the data normally distributed?



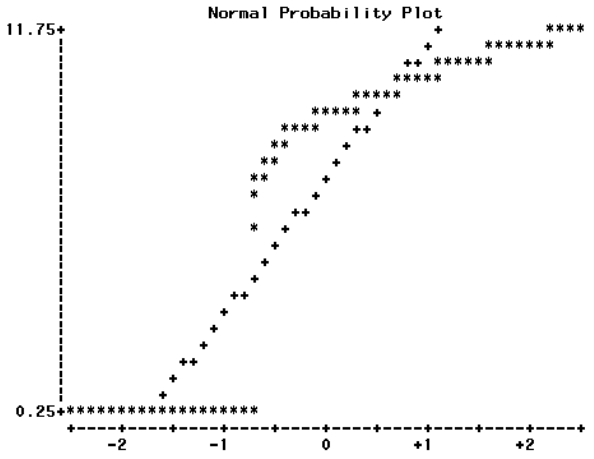
The data are not normally distributed. The histogram should appear as a bell shaped curve turned on its side. The normal probability plot should appear as a straight line. These data require transformation.

Transformed via log + 1

```
options formdlim='~';
data one;
  title 'Bai Protein Expression data - sequenced Proteins';
  infile 'C:\Documents and Settings\joneslab\Desktop\SeqProt.csv' dlm=',' firstobs=2 ;
  input protein $ SSP WT01 WT02 WT03 WTU241 WTU242 WTU243 WTU481 WTU482 WTU483 WTU721 WTU722 WTU723 HAP241 HAP242
HAP243 HAP481 HAP482 HAP483 HAP721 HAP722 HAP723;
lnWT01 = log(WT01 + 1);
lnWT02 = log(WT02 + 1);
lnWT03 = log(WT03 + 1);
lnWTU241 = log(WTU241 + 1);
lnWTU242 = log(WTU242 + 1);
lnWTU243 = log(WTU243 + 1);
lnWTU481 = log(WTU481 + 1);
lnWTU482 = log(WTU482 + 1);
lnWTU483 = log(WTU483 + 1);
lnWTU721 = log(WTU721 + 1);
lnWTU722 = log(WTU722 + 1);
lnWTU723 = log(WTU723 + 1);
lnHAP241 = log(HAP241 +1);
lnHAP242 = log(HAP242 +1);
lnHAP243 = log(HAP243 +1);
lnHAP481 = log(HAP481 +1);
lnHAP482 = log(HAP482 +1);
lnHAP483 = log(HAP483 +1);
lnHAP721 = log(HAP721 + 1);
lnHAP722 = log(HAP722 + 1);
lnHAP723 = log(HAP723 + 1);
Proc univariate normal plot;
            var lnWT01 lnWT02 lnWT03 lnWTU241 lnWTU242 lnWTU243;
            run;
```



Data now fit the ANOVA assumption of normality.



Structuring data for SAS

	A	B	C	D	E	F	G	H
1	Protein #	WT 0 hap 1-1	WT 0 hap 1-2	WT 0 hap 1-3	WT unpo 24 hr 1-1	WT unpo 24 hr 1-2	WT unpo 24 hr 1-3	WT unpo 48 hr 1-1
2	SB36-14	18096.7	.	10560.9	14082.2	13229.9	15154.9	7202.3
3	SB50-9	5454.5	.	2986.6	14318.7	10447.8	12314.8	0
4	SB36-28	5766.4	9154.9	9429.2	6841.6	7067.5	2678.2	9294.2
5	SB36-30	5122.4	3866.7	3008.6	3985.5	3143.8	6611.2	3381.2

```
EXP = lnWT01; treatment = 0; time = 0; Rep = 1; output;
EXP = lnWT02; treatment = 0; time = 0; Rep = 2; output;
EXP = lnWT03; treatment = 0; time = 0; Rep = 3; output;
EXP = lnWTU241; treatment = 1; time = 24; Rep = 1; output;
EXP = lnWTU242; treatment = 1; time = 24; Rep = 2; output;
EXP = lnWTU243; treatment = 1; time = 24; Rep = 3; output;
EXP = lnWTU481; treatment = 1; time = 48; Rep = 1; output;
EXP = lnWTU482; treatment = 1; time = 48; Rep = 2; output;
EXP = lnWTU483; treatment = 1; time = 48; Rep = 3; output;
EXP = lnWTU721; treatment = 1; time = 72; Rep = 1; output;
EXP = lnWTU722; treatment = 1; time = 72; Rep = 2; output;
EXP = lnWTU723; treatment = 1; time = 72; Rep = 3; output;
EXP = lnHAP241; treatment = 2; time = 24; rep = 1; output;
EXP = lnHAP242; treatment = 2; time = 24; rep = 2; output;
EXP = lnHAP243; treatment = 2; time = 24; rep = 3; output;
EXP = lnHAP481; treatment = 2; time = 48; rep = 1; output;
EXP = lnHAP482; treatment = 2; time = 48; rep = 2; output;
```

The baseline for protein quantities is at 0 hours after pollination.

SAS Command: Proc GLM

Main effects (all fixed within the model):

- protein (protein spots on the gel)
- treatment (unpollinated/pollinated)
- time (after pollination)
- rep (3, 2D gels per treatment)

```
Proc glm;
      class protein treatment time rep;
      model EXP = protein treatment time rep rep*protein
      rep*treatment rep*time treat*protein time*protein
      time*treatment;
Proc varcomp method = REML;
      class protein treatment time rep;
      model EXP = protein treatment time rep rep*protein
      rep*treatment rep*time treat*protein time*protein
      time*treatment;
quit;
```

Output of Proc GLM

The GLM Procedure

Dependent Variable: EXP

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	989	36342.17871	36.74639	20.22	<.0001
Error	1926	3499.80085	1.81713		
Corrected Total	2915	39841.97956			

Mean Square Error (MSE)

R-Square	Coeff Var	Root MSE	EXP Mean
0.912158	17.39963	1.348011	7.747356

Source	DF	Type III SS	Mean Square	F Value	Pr > F
protein	139	22324.65836	160.60905	88.39	<.0001
treatment	1	519.20558	519.20558	285.73	<.0001
time	2	108.58773	54.29387	29.88	<.0001
Rep	2	0.35196	0.17598	0.10	0.9077
Protein*Rep	278	33.08495	0.11901	0.07	1.0000
treatment*Rep	2	2.07059	1.03530	0.57	0.5658
time*Rep	4	1.39061	0.34765	0.19	0.9430
Protein*treatment	139	5433.79977	39.09208	21.51	<.0001
Protein*time	278	4550.96421	16.37037	9.01	<.0001
treatment*time	2	216.32514	108.16257	59.52	<.0001

Sources in red were found to be significant. Rep is not significant.

Appropriate cut off

- The data are inherently noisy
- The cut off (or least significant difference, LSD) should be based on the variance of the data, or Mean Square Error and the level of statistical significance
- The LSD can be calculated with $e^{[t \times \sqrt{(MSE/n)}]/2}$, where t is 1.96 for $P = 0.05$, 2.576 for $P = 0.01$, and 3.291 for $P = 0.001$ and n is the number of replicates

Use Excel for determining LSD

This is calculated and given in the SAS output

Choose the appropriate α-level for the experiment; $n = 3$.
LSD = [t*(SQRT(MSE/n))]/2

MSE	LSD	Fold	α-level
1.817	0.762	2.144	p = 0.05
1.817	1.002	2.725	p = 0.01
1.817	1.281	3.599	p = 0.001

The LSD must be "untransformed" using Fold = $(LSD)^e$
This can be done in excel with Fold=power(2.718, LSD)

Why is it a **fold** difference?
A log transformation was necessary to fit normality assumptions. Log data are only interpretable as fold differences

What parameters are used to select an appropriate significance level?
For a more conservative value, you would select a smaller α-level. This reduces the chance of a Type I error

Summary

1. Determine if the data fit the assumptions
2. Transform as needed
3. Restructure the data in SAS
4. Determine the main effects for the experiment
5. Use Proc GLM to determine the MSE
6. Calculate the cutoff (LSD) for statistical significance

References Cited

- Bai, S., B. Willard, L. J. Chapin, M. T. Kinter, D. M. Francis, A. D. Stead, and M. L. Jones. 2010. Proteomic analysis of pollination-induced corolla senescence in petunia. Journal of Experimental Botany 61:1089-1109. (Available online at: http://dx.doi.org/10.1093/jxb/erp373) (verified 20 Oct 2011).

- Kerr, M. K., M. Martin, and G. A. Churchill. 2000. Analysis of variance for gene expression microarray data. Journal of Computational Biology 7:819-837. (Available online at: http://dx.doi.org/10.1089/10665270050514954) (verified 20 Oct 2011).