Lecture #6

Differential expression

Outline

- Differentially expressed genes
- Filtering genes
- Two-sample tests
    - Parametric tests
        - Student’s t-test
        - Welch’s modified t-test
        - Fold change
    - Non-parametric tests
        - Wilcoxon-Mann-Whitney test
- Greater than two-sample test
    - Parametric tests
        - One-factor ANOVA (fixed effects)
        - Two-factor ANOVA (fixed effects)
    - Non-parametric tests
        - Kruskal-Wallis test
- Partial least squares regression
- Gene shaving

Why be concerned with differentially expressed genes?

- Differential expression allows us to form hypotheses about the genes that discriminate one state from another

- Genes that are over/under-expressed in different states can provide:
  - Models specific for tissues, disease, treatments, etc.
  - Markers for disease-state screening
  - Mechanistic analysis
  - Therapeutic targets

General Methodology

- What is the general distribution of the genes?
  - Parametric tests assume that the data follows a specific distribution
  - Non-parametric tests do not make such assumptions
- Can the data be transformed to give a more robust test?
- For each gene, conduct a statistical test
- Calculate the scoring statistic (e.g. test statistic) for each test
- Determine if the scoring statistic exceeds the pre-determined threshold
- Correct the scoring statistic, accounting for the number of statistical tests
  - Multiple testing correction

Gene filtering

- Usually one of the preliminary steps to choosing differentially expressed genes involves reducing the number of genes to begin with

- This will eliminate those genes that either have small/no expression intensity or genes whose expression does not vary across samples

- In Affymetrix data:
  - The A/P calls can be a primary filter
    - e.g retain only those genes with a P call across *n-i* samples, where i can be 1,2…*n*
  - Mean expression intensities that fall below a specified value
  - Low variance across all samples

- In cDNA data:
  - Genes that have expression intensities where the background is larger than the signal
    - Results in negative value for either Cy5 or Cy3 net intensity
  - Low variance across all samples

Student's t-test (two-sample)

- $X_1, \ldots X_m$ are $N(\mu_X, \sigma^2)$ and $Y_1, \ldots Y_n$ are $N(\mu_Y, \sigma^2)$
  - The variances are assumed to be equal, so the pooled variance is calculated as:

$$s^2 = \frac{1}{m+n-2}\left(\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right).$$

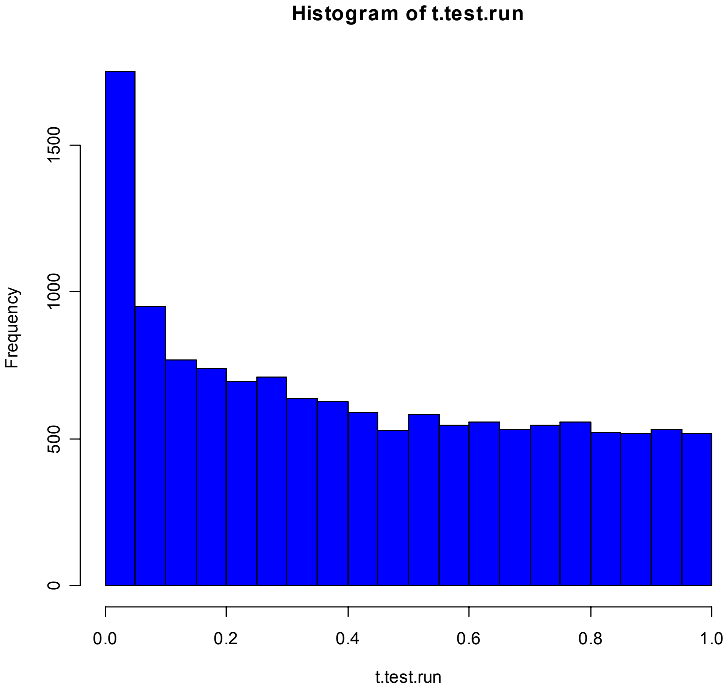
  - The test-statistic for the null, $\mu_X = \mu_Y$, is calculated as:

$$T(X,Y) = \frac{\bar{X} - \bar{Y}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

  - Under the null, $\mu_X = \mu_Y$, the test statistic follows a t_{m+n-2} distribution

Student’s t-test example

- Distribution of p-values for ~8,000 genes from Eisen et al. DLBCL data set



F-test of variances

- Test to determine the homogeneity of variances between two groups
  - Useful for determination of differential expression tests

- s_1^2 and s_2^2 are sample variances with n_1-1 and n_2-1 degrees of freedom
  - Follows an F-distribution with numerator (n_1-1) and denominator (n_2-1)
  - Confidence interval: $F_{df1,df2,\alpha} < s_1^2/s_2^2 < F_{df2,df1,1-\alpha}$
  - Note: $F_{df1,df2,\alpha} = 1/(F_{df2,df1,1-\alpha})$

- This test is for two groups. To test multiple groups, use Bartlett's test (homogeneity of covariance)

- F-test in R:
  - `>var.test(x,y)`

Welch's modified t-test (two-sample)

- $X_1,\ldots X_m$ are $N(\mu_X,\sigma^2{}_X)$ and $Y_1,\ldots Y_n$ are $N(\mu_Y,\sigma^2{}_Y)$
  - The variances are different, so the test-statistic for the null, $\mu_X = \mu_Y$, is calculated as:

$$T(X,Y) = \frac{\bar{X} - \bar{Y}}{\sqrt{s_X^2/m + s_Y^2/n}}.$$

  - Under the null, $\mu_X = \mu_Y$, the degrees of freedom are calculated as:

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{\left(\frac{s_1^2}{m}\right)^2}{m-1} + \frac{\left(\frac{s_2^2}{n}\right)^2}{n-1}}$$

Fold Change

- Significance tests determine differential expression between means as a function of variance

- Fold change is a relative measure of the magnitude of difference between means
  - Variance is not assessed in calculation
  - Common fold change threshold is usually 1.5-3

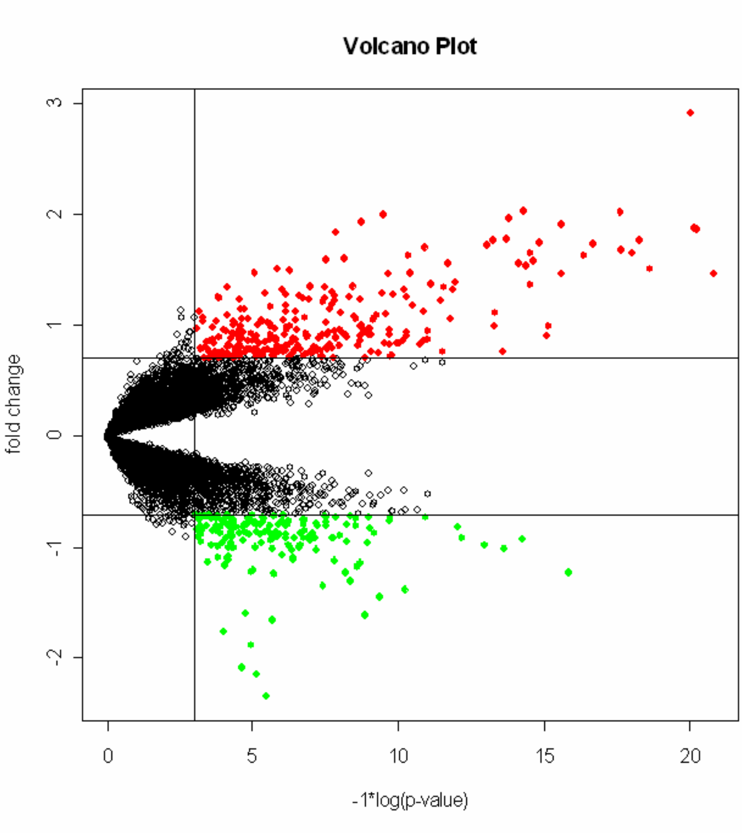
- Linear scale for each gene
  Fold change = mean(X) / mean(Y)
  Value of 1 is indicative of no change

- Log scale for each gene
  Fold change = mean(X) – mean(Y)
  Value of 0 is indicative of no change

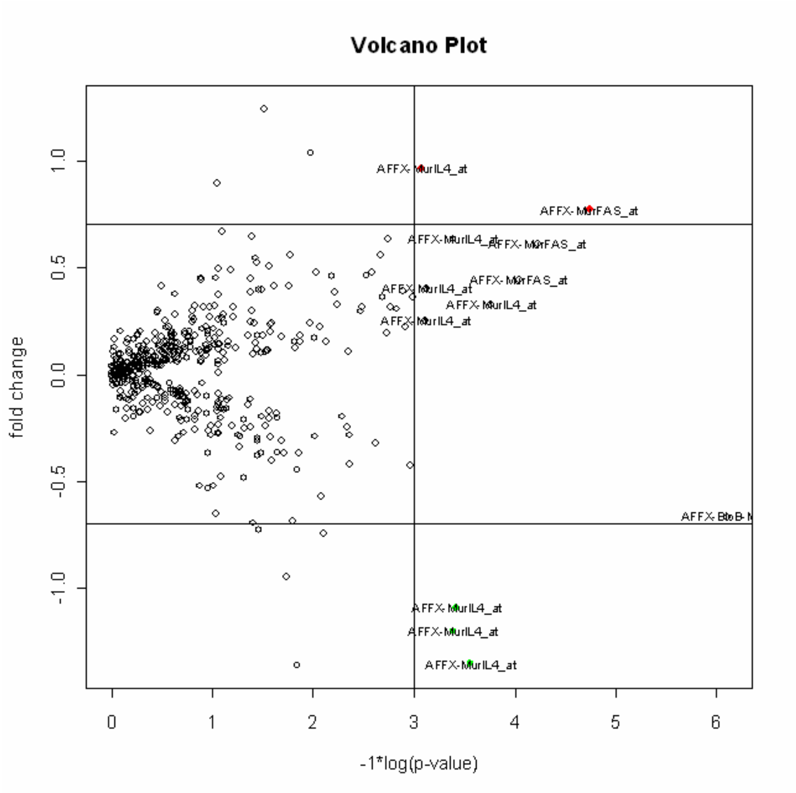
- Remember that two-channel arrays values are intrinsically fold changes due to the two hybridizations (control and treated)
  - *log(R) – log(G)*

- Combination of fold change and p-value provide most significantly differentially expressed genes

Fold vs. p-value plot (volcano)



Fold vs. p-value plot (volcano)



Wilcoxon-Mann-Whitney u-test (two-sample)

- Both samples are combined and the values are ranked in the pooled sample

Value	Group	Rank
20	1	3
30	1	4
15	2	2
60	2	5
10	2	1

- The test statistic is calculated as a function of the sum of ranks in one of the groups

- For large sample sizes, a normal approximation is used

$$Z = [W_1 - n\,(n+m+1)\,/\,2]\,/\,[sqrt(nm\,(n+m+1)/12)] \sim N(0,1)$$

- Depending on ratio of *m/n*, can perform better for very different sample sizes than parametric test

Experimental design basic terminology

- Type of conditions that the experimental units are manipulated by are factors
  - Groups
  - Doses
  - Assay time points
- The different modes of a factor are the factor levels
  - male & female
  - control, mid-level, high-level
  - 0 hrs, 10 hrs, 15 hrs, 25 hrs
- Multiple ANOVA models exist (with corrections), which can be contingent upon different experimental designs and testing parameters
  - We will only concern ourselves with a fixed effects factors, without repeated measures, and near balanced designs

One-factor ANOVA – completely randomized design

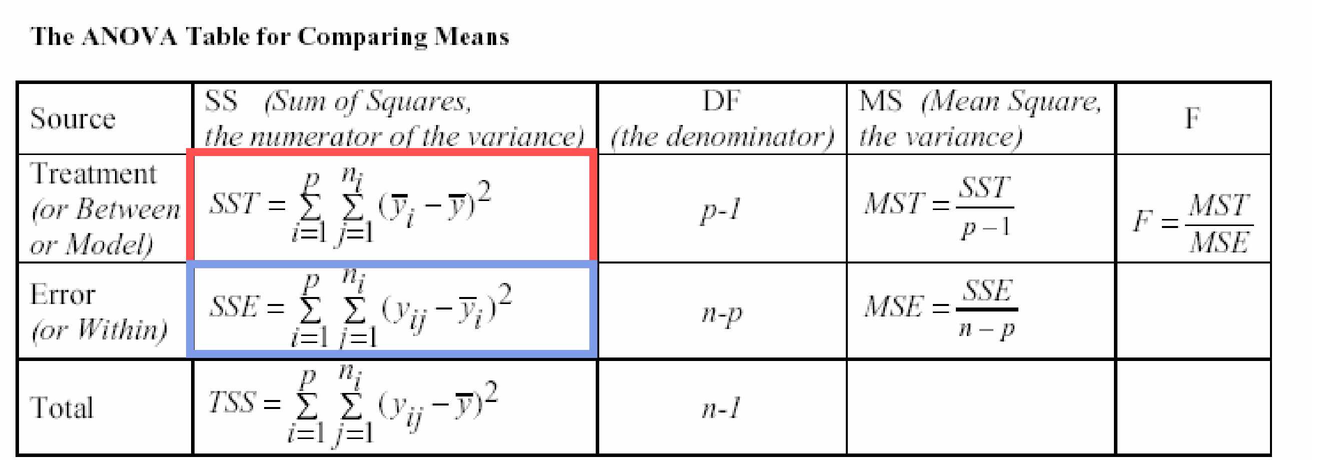
- The completely randomized design consists of independent random sampling from several populations when each population is identified as the population of responses under a particular treatment
  - Randomly sample a population and assign treatments

- What are we testing?
  - Is there any significant difference between the means of each treatment?
  - $y_{ij} = \mu + \beta_j + e_{ij}$

    μ is overall mean; β_j is jth treatment effect; $e_{ij} \sim N(0,\sigma)$
  - H_O: $\beta_1 = \beta_2 = \ldots\ \beta_k = 0$

One-factor ANOVA – completely randomized design

- ANOVA table decomposed



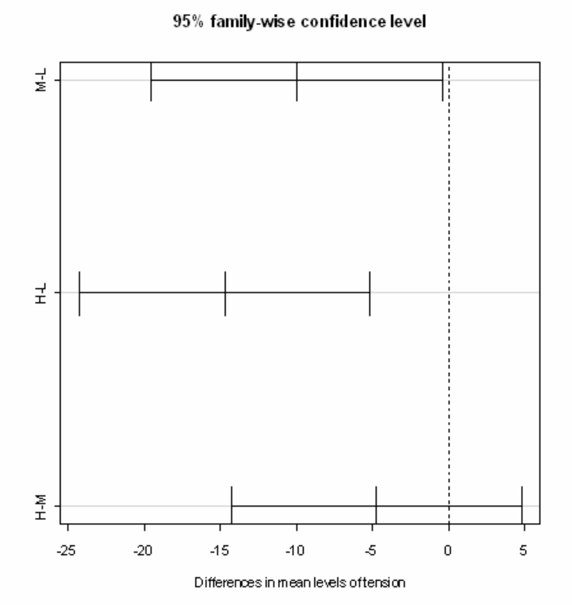
The ANOVA Table for Comparing Means

Source	SS (*Sum of Squares, the numerator of the variance*)	DF (*the denominator*)	MS (*Mean Square, the variance*)	F
Treatment *(or Between or Model)*	$SST = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2$	*p-1*	$MST = \frac{SST}{p-1}$	$F = \frac{MST}{MSE}$
Error *(or Within)*	$SSE = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	*n-p*	$MSE = \frac{SSE}{n-p}$	
Total	$TSS = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$	*n-1*		

- **Sum of squares due to differences in the treatment means**
- **Residuals are deviations reflecting inherent variability in the experimental material and measuring device**
- Reject H_O if F-ratio > F_α(p-1,n-p)

One-factor ANOVA – example

- Yarn breaks data set (during weaving)
  Tension is the factor (3 levels: H, M, L) and breaks is the continuous variable



```
                Df Sum Sq  Mean Sq  F value   Pr(>F)
tension       2 2034.3  1017.1    7.2061   0.001753 **
Residuals    51 7198.6   141.1
```

Two-factor ANOVA – completely randomized design

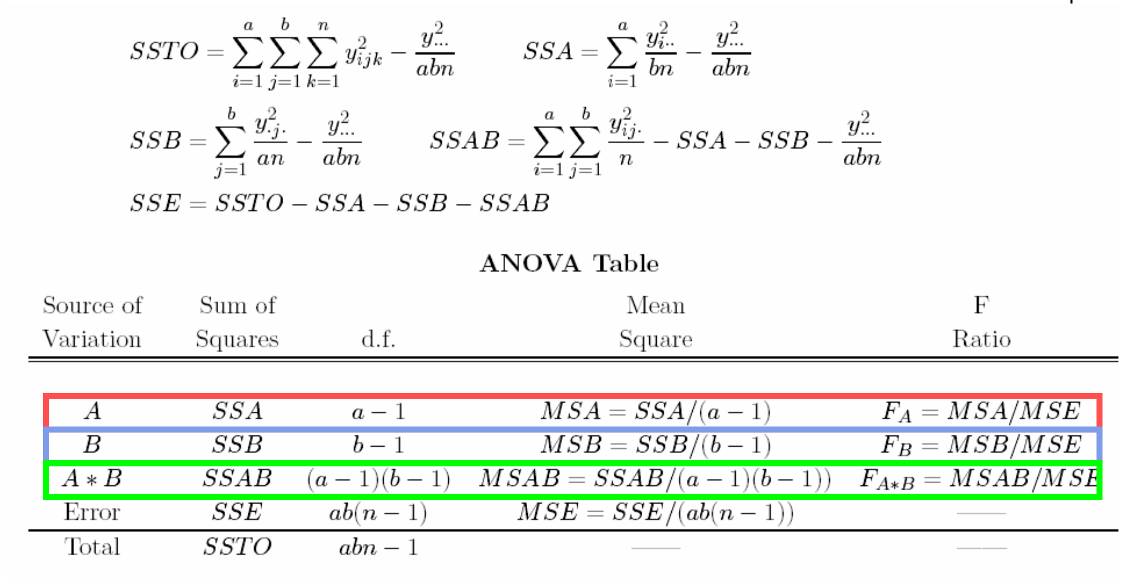
- The completely randomized design consists of independent random sampling from several populations when each population is identified as the population of responses under a particular treatment
  - Randomly sample a population and assign treatments

- What are we testing?
  - What are the effects of factor A, factor B, and the simultaneous effect of the combination of factors A and B on the response of interest?
  - $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$

    μ is overall mean; α_i is ith treatment effect of factor A ; β_j is jth treatment effect of factor B; $(\alpha\beta)_{ij}$ is the interaction term; $e_{ijk} \sim N(0,\sigma)$

Two-factor ANOVA – completely randomized design

- ANOVA table decomposed



$$SSTO = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{\cdots}^2}{abn} \qquad SSA = \sum_{i=1}^{a} \frac{y_{i\cdot\cdot}^2}{bn} - \frac{y_{\cdots}^2}{abn}$$

$$SSB = \sum_{j=1}^{b} \frac{y_{\cdot j\cdot}^2}{an} - \frac{y_{\cdots}^2}{abn} \qquad SSAB = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{y_{ij\cdot}^2}{n} - SSA - SSB - \frac{y_{\cdots}^2}{abn}$$

$$SSE = SSTO - SSA - SSB - SSAB$$

ANOVA Table

Source of Variation	Sum of Squares	d.f.	Mean Square	F Ratio
A	SSA	$a-1$	$MSA = SSA/(a-1)$	$F_A = MSA/MSE$
B	SSB	$b-1$	$MSB = SSB/(b-1)$	$F_B = MSB/MSE$
$A*B$	$SSAB$	$(a-1)(b-1)$	$MSAB = SSAB/(a-1)(b-1))$	$F_{A*B} = MSAB/MSE$
Error	SSE	$ab(n-1)$	$MSE = SSE/(ab(n-1))$	——
Total	$SSTO$	$abn-1$	——	——

- **Test for factor A main effects: reject H_O if $F_A > F_\alpha(a\text{-}1,ab(n\text{-}1))$; $H_O = \alpha_1, \alpha_2 \ldots \alpha_a = 0$**
- **Test for factor B main effects: reject H_O if $F_B > F_\alpha(b\text{-}1,ab(n\text{-}1))$; $H_O = \beta_1, \beta_2 \ldots \beta_b = 0$**

Kruskal-Wallis test for comparing *k* treatments

- Non-parametric analog to the one-way ANOVA

- The *k* samples are combined and the values are ranked in the pooled sample

- The average ranks for individual samples are calculated *(R.bar)*

- The test statistic is then calculated as:

$$KW = \frac{12}{N(N+1)} \sum_{i=1}^{K} n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2$$

- The test is rejected for $KW > x^2_{K-1}$

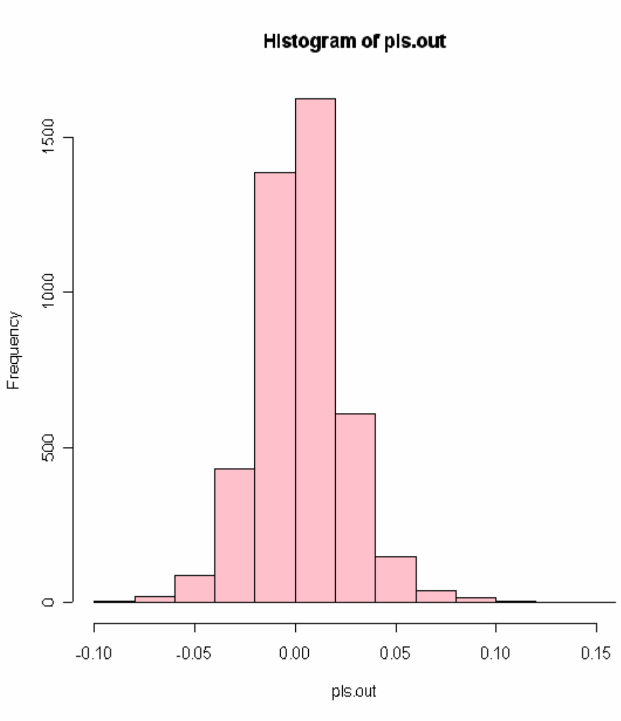
Partial least squares regression (PLS)

- PLS is a multivariate regression method
- Very generally, PLS, like PCA works to maximize the variability of a matrix by calculating linear combinations of the original variables
- However, PCA maximizes this variability between the samples/genes, while PLS relates the data matrix, **X** to a response, **Y**
  - **X** is this example is a matrix of genes by samples
  - **Y** in this example is the expected continuous response or class membership
- PLS is a regression approach, where the predictor variables are weighted according to their ability to predict the response variable

PLS example Spellman et al. yeast data
(cdc15 experiment)

Gene weights are computed, based on the similarity to the response

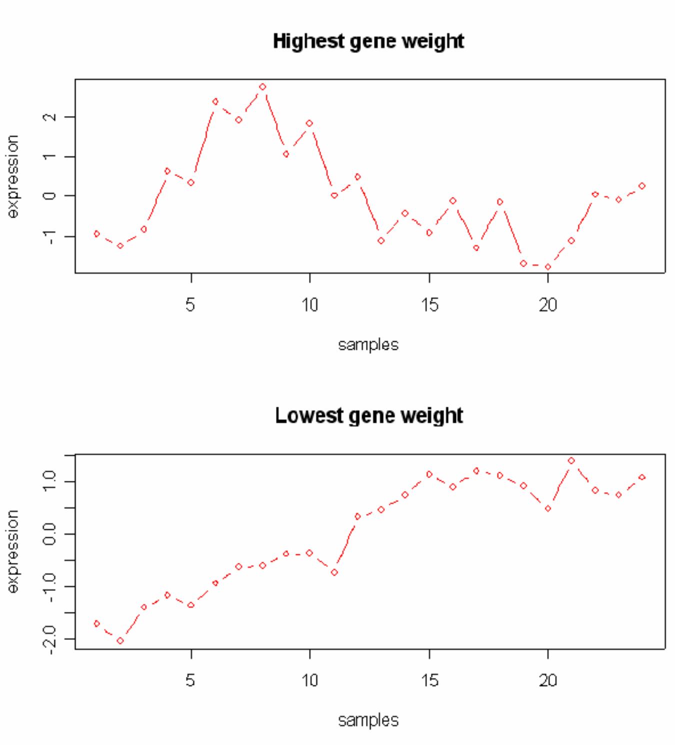
Large positive weights indicate a strong match, while large negative weights indicate a strong opposite match



PLS example Spellman et al. yeast data
(cdc15 experiment)

Response was specified as:

up (1) at first 12 times states and down (0) at next 12 times states



Gene Shaving Gene Selection

- A method for identifying gene subsets with coherent expression relevant measurements (samples)
- Iterative sampling method to "identify groups of genes that optimally separate samples into predefined classes"
- Randomization correction procedure is implemented to protect against determining spurious structure in the data

Primary Gene Shaving Methodology

- Start with an expression matrix *X,* (genes x samples), mean center each gene

- Compute the largest principal component over the genes
  - Linear combination of genes explaining maximal variance

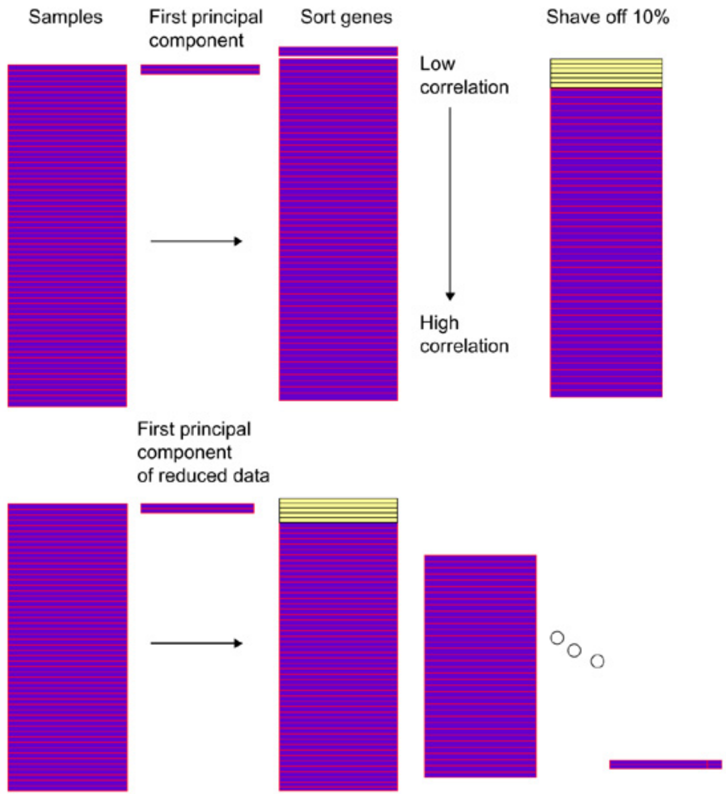
- Calculate the absolute inner-product between the largest principal component and all genes
  - Correlation between largest principal component and gene *k*

- Shave off 10% of the genes with the lowest correlation values

- Repeat procedure until 1 gene remains

- This nested sequence of genes clusters are then evaluated for the optimal cluster size, *k* using a gap statistic

Primary Gene Shaving Methodology


Samples
First principal component
Sort genes
Low correlation
High correlation
Shave off 10%
First principal component of reduced data

Gap Estimate

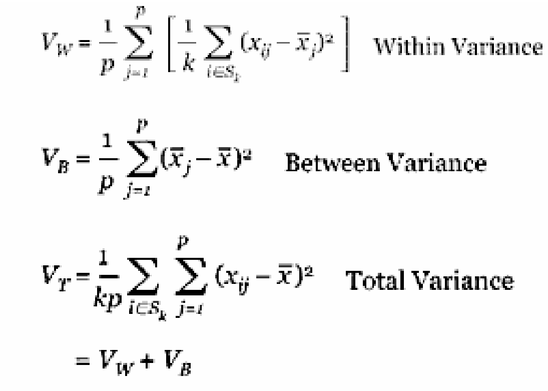
- The first step of the shaving method creates a series of gene clusters, S_k ranging in size from 90% the number of genes to 1

- If this method were applied to random data, many genes would exhibit patterns similar to actual data

- Require a method to calibrate the shaving process to differentiate real patterns from spurious patterns

Gap Estimate – cluster quality measure

- Looking for clusters with high-variance clusters and high coherence between members of the clusters
- Similar method to ANOVA variance components



$$V_W = \frac{1}{p}\sum_{j=1}^{p}\left[\frac{1}{k}\sum_{i\in S_k}(x_{ij}-\bar{x}_j)^2\right]$$ Within Variance

$$V_B = \frac{1}{p}\sum_{j=1}^{p}(\bar{x}_j-\bar{x})^2$$ Between Variance

$$V_T = \frac{1}{kp}\sum_{i\in S_k}\sum_{j=1}^{p}(x_{ij}-\bar{x})^2$$ Total Variance

$$= V_W + V_B$$

Between variance: variance of the mean gene

Within-variance: variability of each gene about the cluster average, also averaged over samples

Gap Estimate – cluster quality measure

- Percent variance explained

$$R^2 = 100\frac{V_B}{V_T} = \frac{\frac{V_B}{V_W}}{1 + \frac{V_B}{V_W}}$$

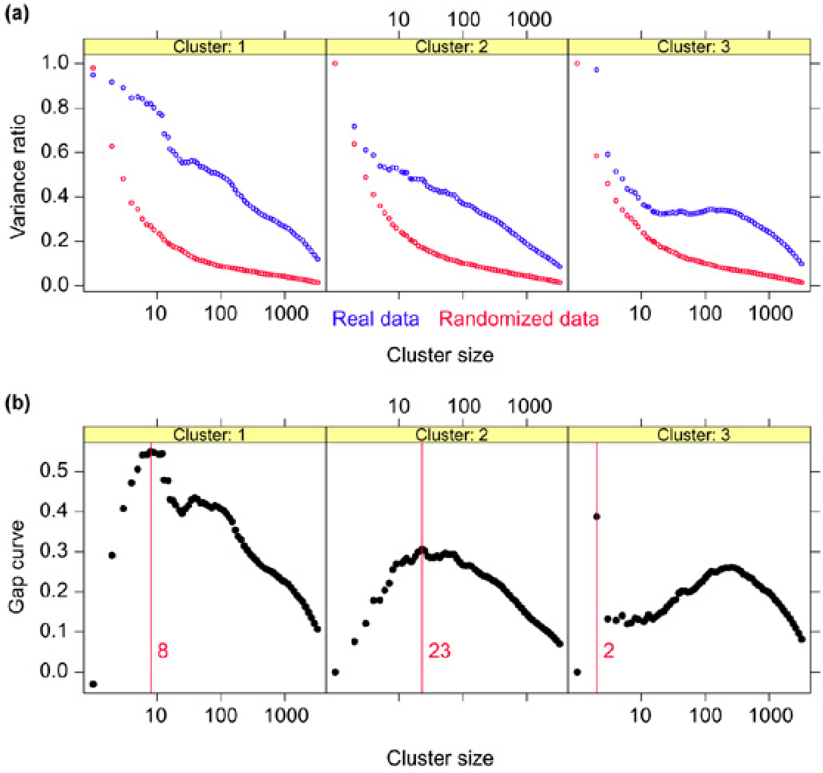
Large R^2 implies tight cluster of coherent genes
D_k is the R^2 measure for the *k*th member of the sequence

- Using a permuted data set, X^{*b}, D_k^{*b} is the R^2 measure for cluster S_k^{*b}
- D.bar$_k^*$ is the average of D_k^{*b} over b permuted random matrices
- The gap function is defined as:

$$\mathrm{Gap}(k) = D_k - \overline{D}_k^*$$

Select the optimal number of genes from the value of *k* producing the largest gap

Variance Plots of Real and Random Data

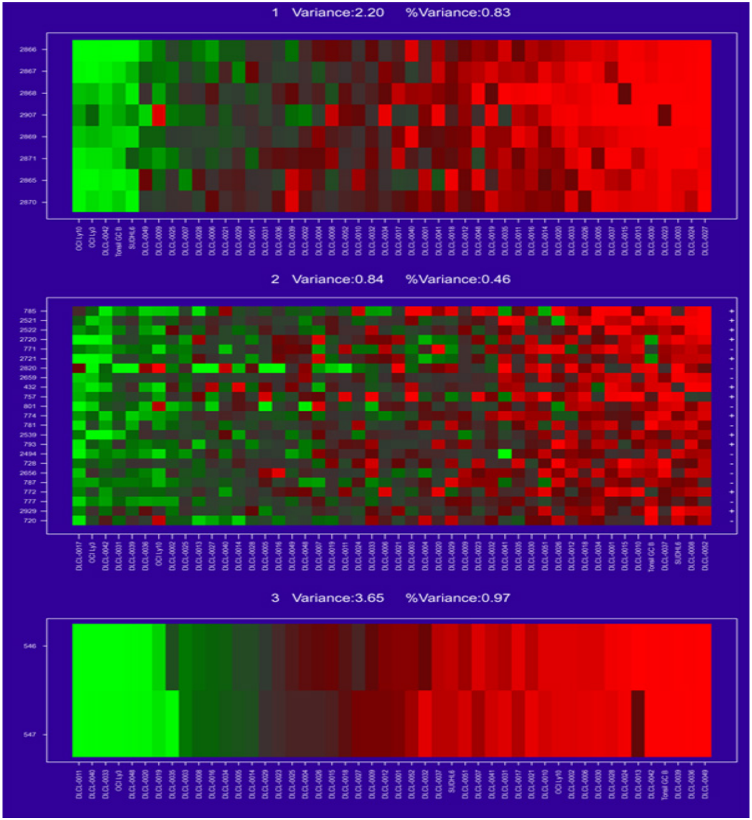


Heat Maps of Top 3 Clusters

8 genes

23 genes

2 genes



1 Variance:2.20 %Variance:0.83

2 Variance:0.84 %Variance:0.46

3 Variance:3.65 %Variance:0.97

References

- Hastie T, Tibshirani R, Eisen M, Alizadeh A, Levy R, Staudt L, Chan W, Botstein D, and Brown P. (2000) ‘Gene shaving’ as a method for identifying distint sets of genes with similar expression patterns. *Genome Biology.* **1:**research0003.1-0003.21

- Pounds S and Morris S. (2003) Estimating the occurrence of FPs and FNs in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*. **19**, 1236- 1242

- Dudoit S, Yang Y, Callow M, and Speed, T. (2000) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Technical Report #578*

- Bhattacharyya, G., Johnson, R. Statistical Concepts and Methods.