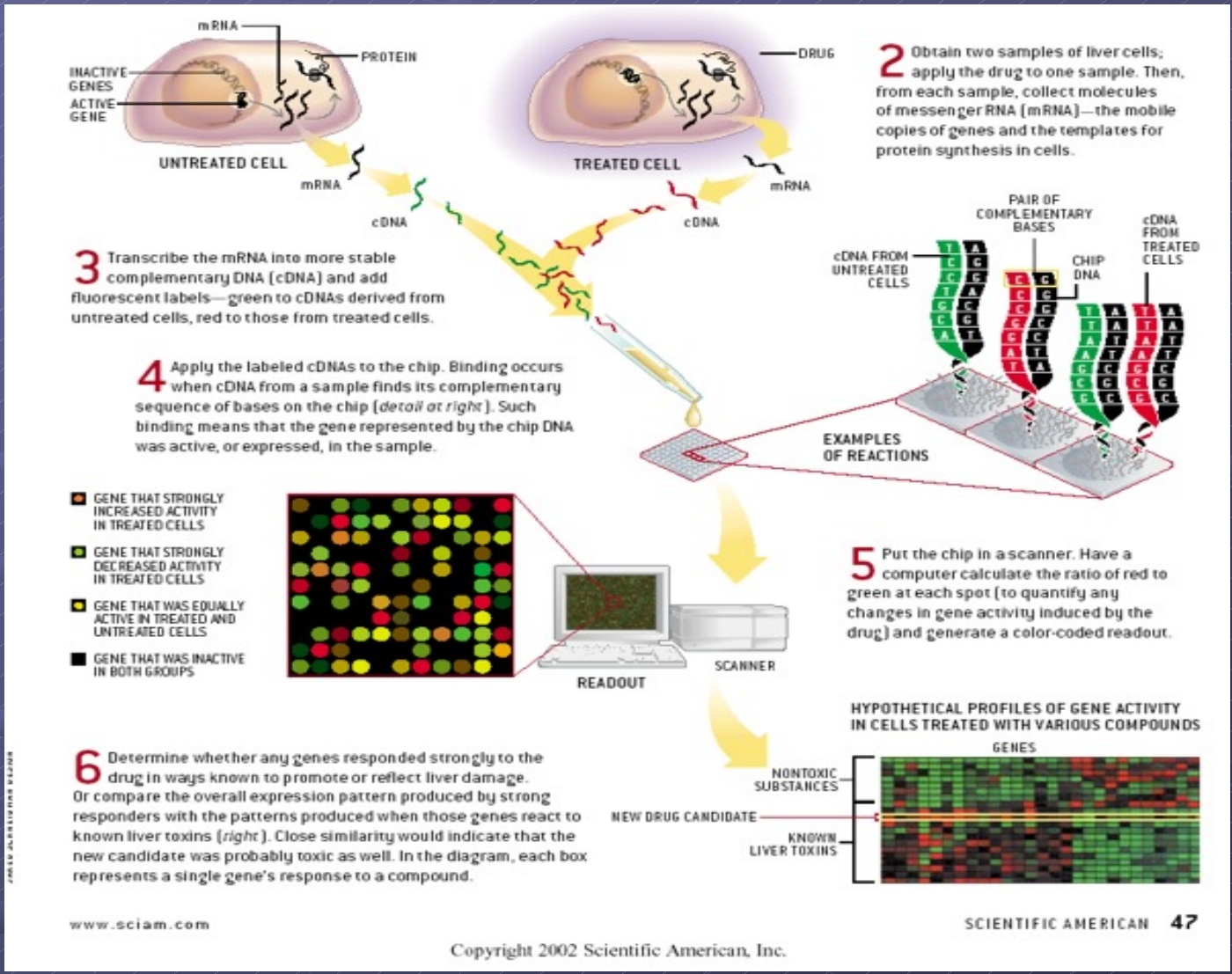
Statistical Tools are Needed for Array Data

Arnold Saxton, UT Animal Science
Michael Langston, UT Computer Science
Brynn Voy, ORNL Life Science Division

Two-color cDNA arrays



Statistical Tools: Design

- Design 1

Array	Red	Green
1	Ctrl	Trt1
2	Ctrl	Trt2
3	Trt1	Ctrl
4	Trt2	Ctrl

Design 2

Array	Red	Green
1	Ctrl	Trt1
2	Trt1	Trt2
3	Trt2	Ctrl
4	Trt2	Trt1

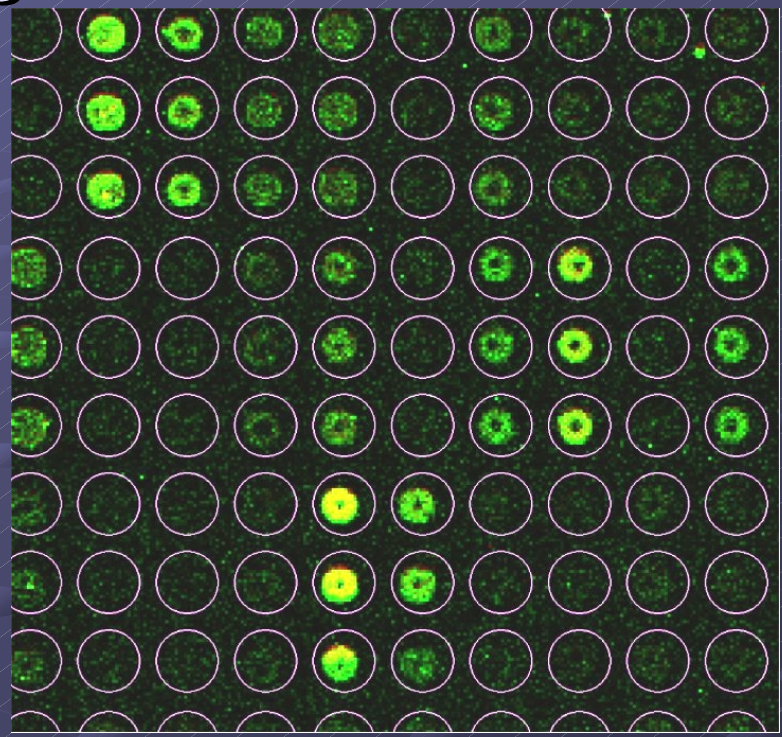
- Control spots, spiked known samples, pooling

Real Application

- 6 strains of mice were exposed to a low dose of X-rays.
- 3.5 hr later, RNA from spleen of exposed and unexposed collected from 4-8 mice/strain.
- 38 microarrays, with exposed/unexposed samples from the same strain always on the same array.

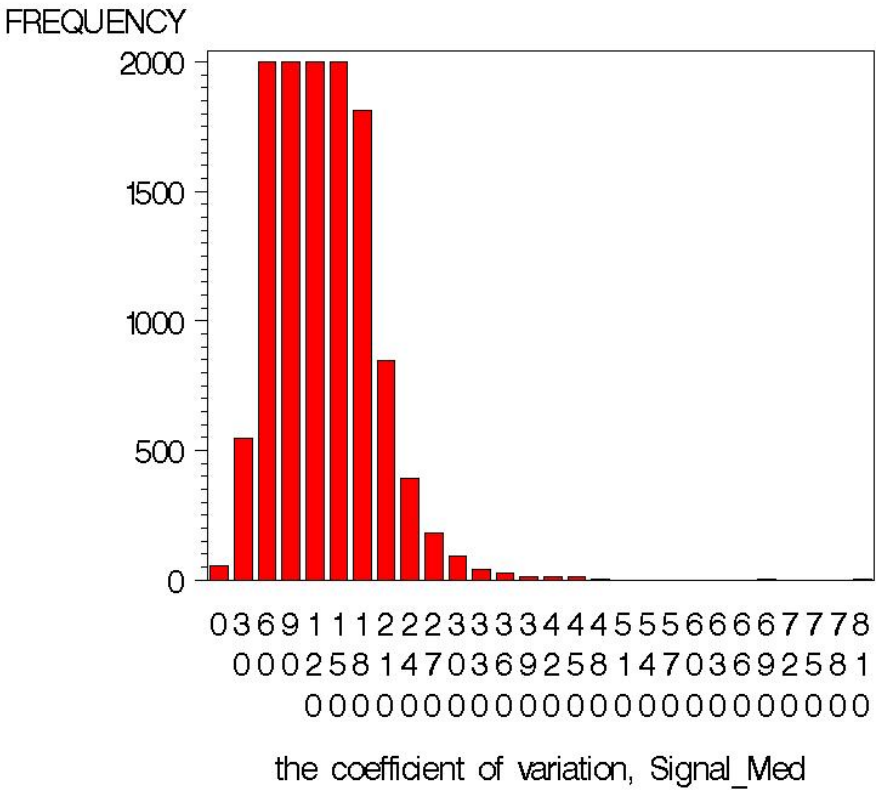
Statistics Needed: Background Correction

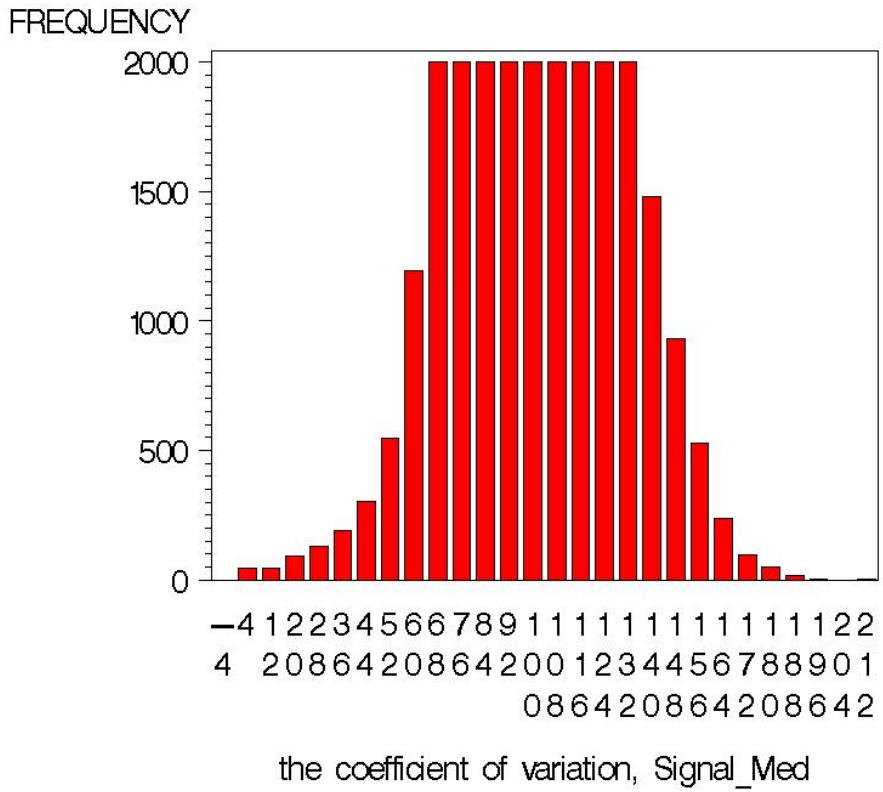
- Ignore Background?
- Subtract off?
- Spatial model?



Statistics Needed: Outliers

- Delete if signal-median>16*MAD std.dev.





Data Removed

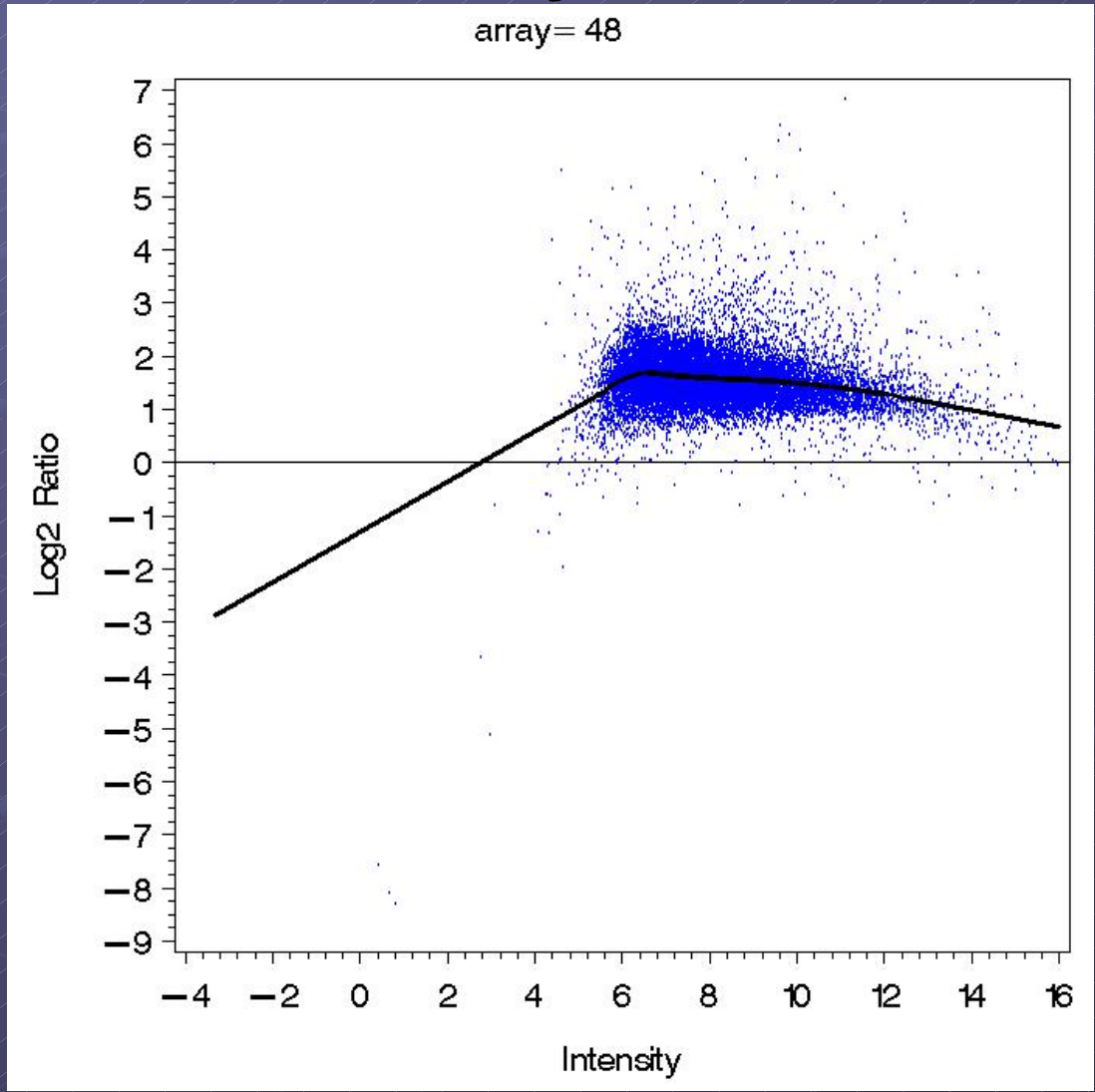
Reason	Frequency	Percent
Background	13465	0.76
Flag=1	5936	0.34
Flag=2	452922	25.65
Flag=3	10404	0.59
Flag=4	4	0.00
Good data	1275628	72.25
Outlier	7273	0.41

Statistics Needed: Intensity Variation

Red/Green ratio plotted against average signal

Loess?

Robust method?



Statistics Needed: Test Treats

- Global model

$$y_{ijk} = u + Array_i + Dye_j + e_{ijk}$$

- Residuals fit by gene

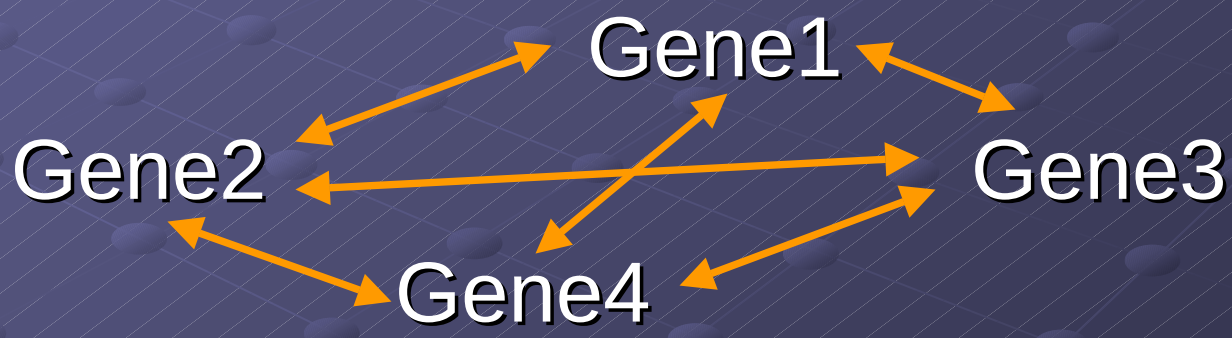
$$e_{ijkl} = u + Array_i + Dye_j + Strain * Dose_{kl} + error_{ijkl}$$

Statistics Needed: Multiple Tests

- 17,000 genes * 66 pairwise tests among 6 strains * 2 doses.
- False Discovery Rate
- q-value
- Bonferonni
- Method to account for correlated genes?

Co-expression: Clique

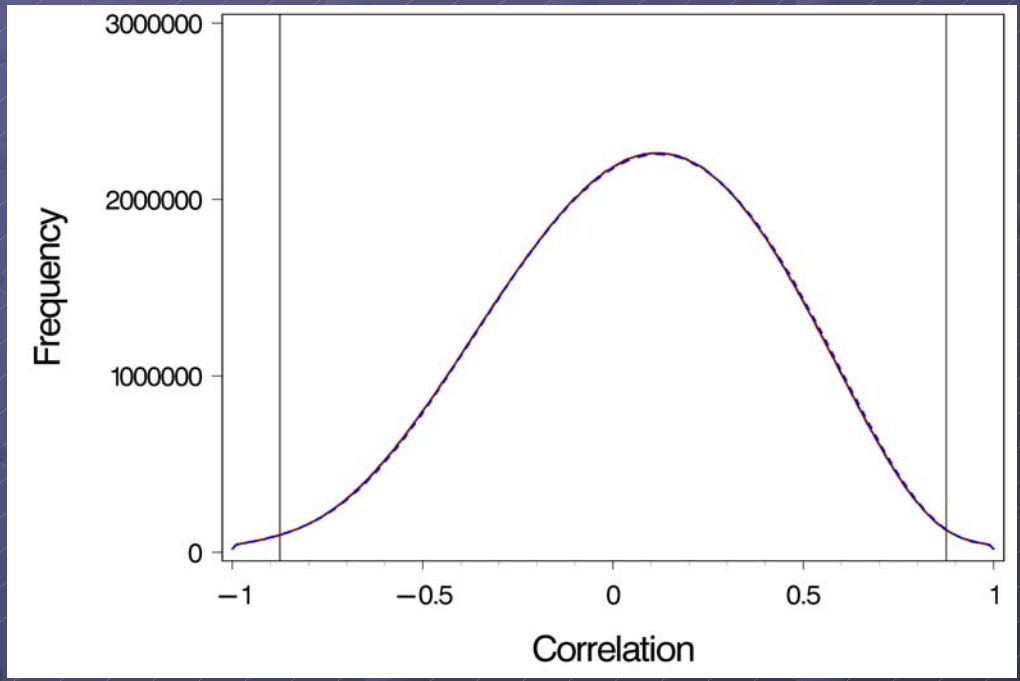
- Objective: Find clusters of genes to help discover function.
- Clique is a completely connected subgraph.



- We compute cliques for both un/exposed

Statistics Needed: Distance

- Pearson, Spearman, Shrinkage, First order partials.....
- 17,000 genes => 100 million correlations

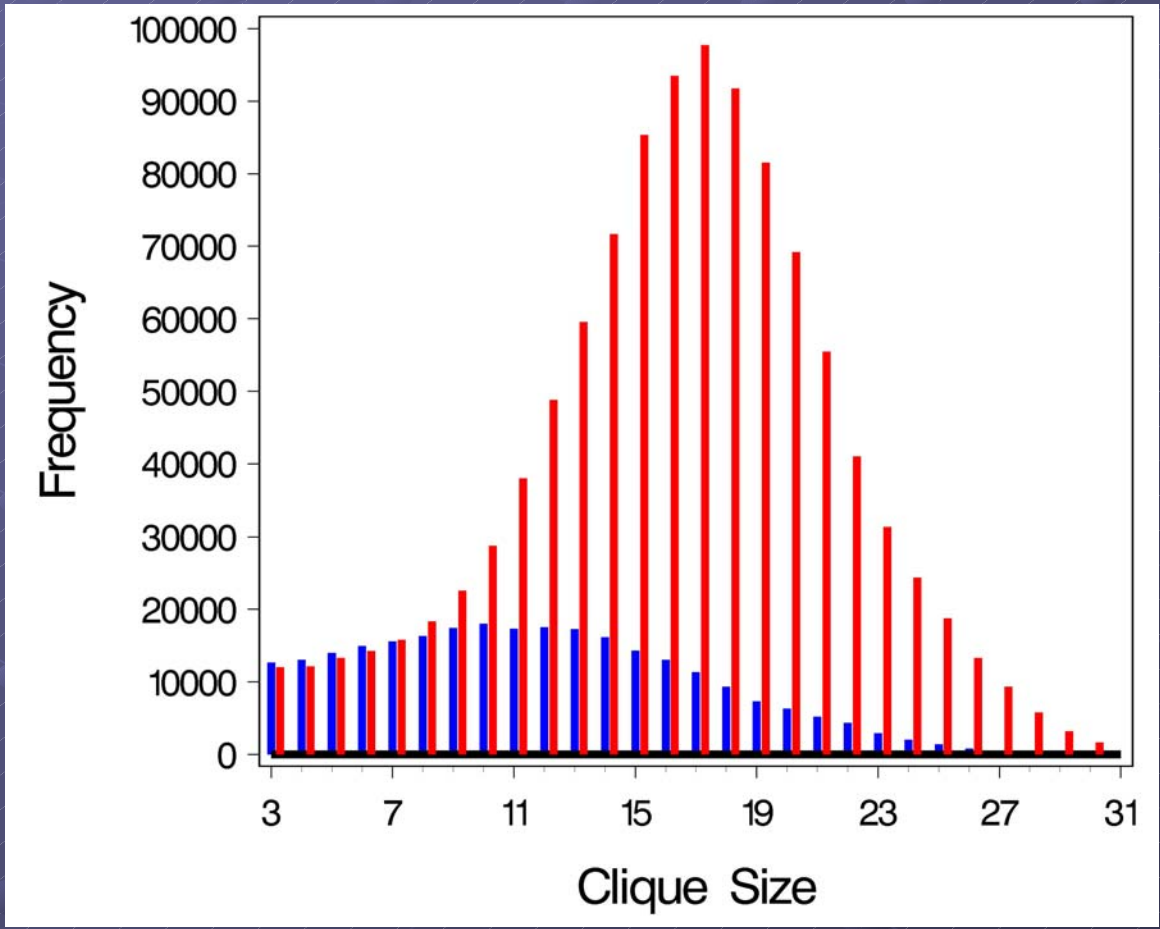


Statistics Needed: Threshold

- Statistical difference from 0, adjusted for multiple testing, gives Threshold=.85
- In our example, only about 150,000 correlations are accepted as "real".

Statistics Needed: Clique P-values

- 268,611 unexposed cliques, 1M exposed.



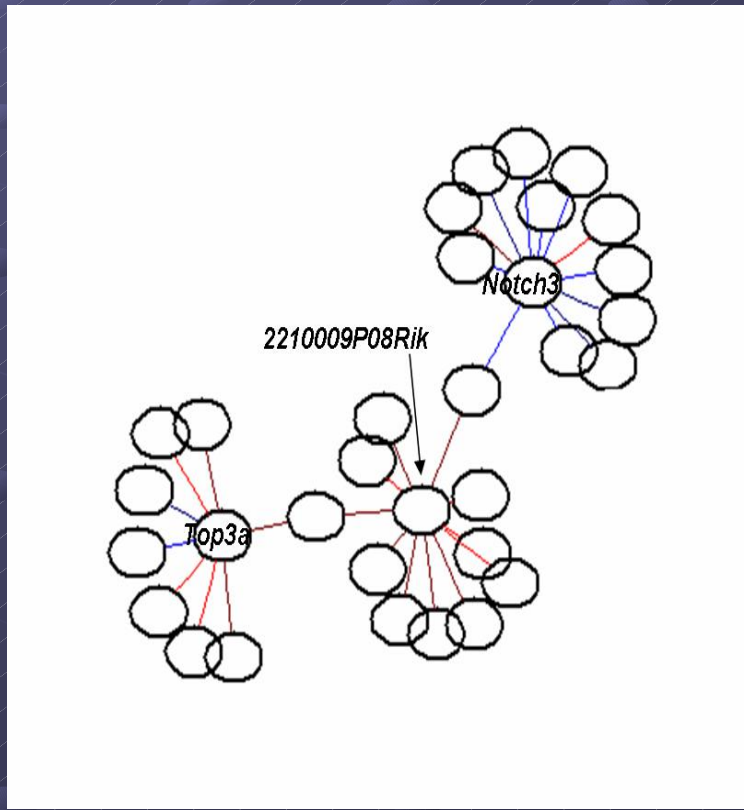
Statistics Needed: Compare Trts

- Did unexposed clique [1,66,1242,77] => exposed clique[1,59,813,12999] happen by chance?
- "Differential correlation":

did unexposed r=.90

=> exposed r=.15

happen by chance?



Conclusions

- Microarrays were first analyzed without statistics: a 2-fold change was "significant".
- Extensive statistics are now used.
- Many improvements are possible.
- Statistics for proteomic data.