

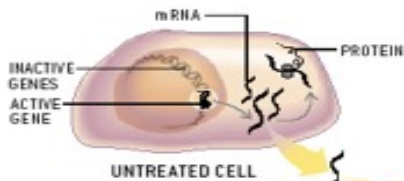
Statistical Tools are Needed for Array Data

Arnold Saxton, UT Animal Science

Michael Langston, UT Computer Science

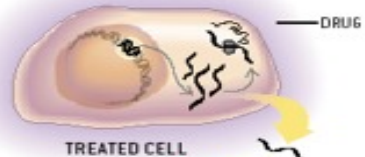
Brynn Voy, ORNL Life Science Division

Two-color cDNA arrays

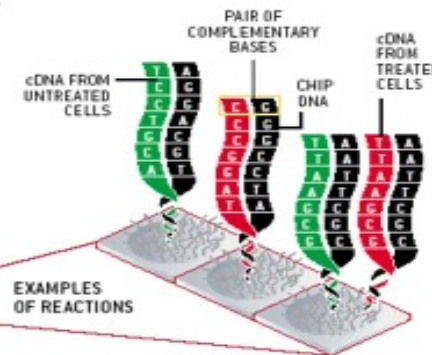


3 Transcribe the mRNA into more stable complementary DNA (cDNA) and add fluorescent labels—green to cDNAs derived from untreated cells, red to those from treated cells.

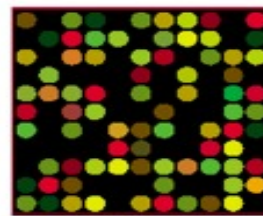
4 Apply the labeled cDNAs to the chip. Binding occurs when cDNA from a sample finds its complementary sequence of bases on the chip (detail at right). Such binding means that the gene represented by the chip DNA was active, or expressed, in the sample.



2 Obtain two samples of liver cells; apply the drug to one sample. Then, from each sample, collect molecules of messenger RNA (mRNA)—the mobile copies of genes and the templates for protein synthesis in cells.

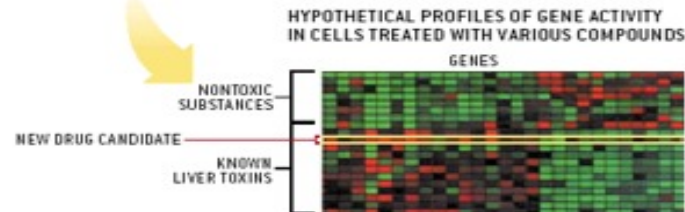


- GENE THAT STRONGLY INCREASED ACTIVITY IN TREATED CELLS
- GENE THAT STRONGLY DECREASED ACTIVITY IN TREATED CELLS
- GENE THAT WAS EQUALLY ACTIVE IN TREATED AND UNTREATED CELLS
- GENE THAT WAS INACTIVE IN BOTH GROUPS



5 Put the chip in a scanner. Have a computer calculate the ratio of red to green at each spot (to quantify any changes in gene activity induced by the drug) and generate a color-coded readout.

6 Determine whether any genes responded strongly to the drug in ways known to promote or reflect liver damage. Or compare the overall expression pattern produced by strong responders with the patterns produced when those genes react to known liver toxins (right). Close similarity would indicate that the new candidate was probably toxic as well. In the diagram, each box represents a single gene's response to a compound.



Statistical Tools: Design

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

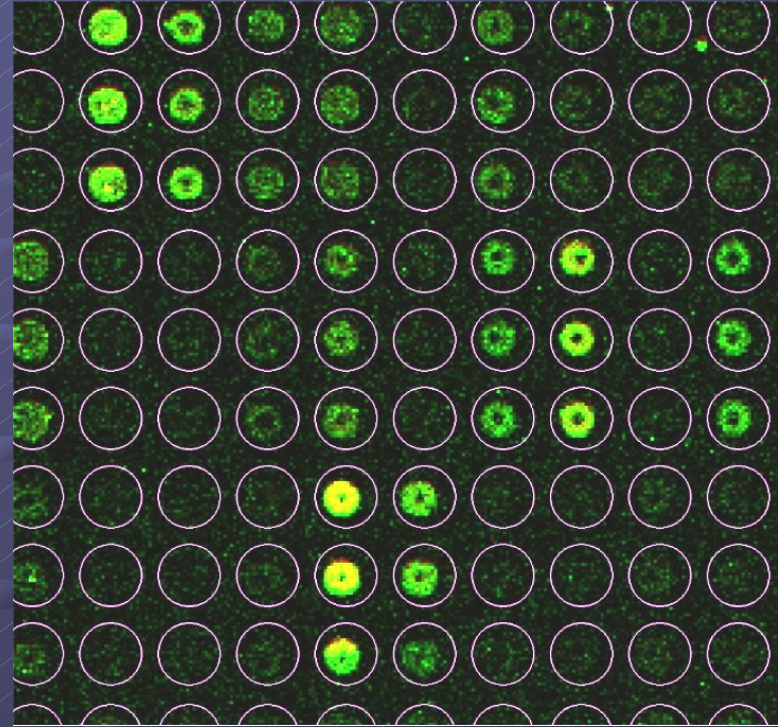
- Control spots, spiked known samples, pooling

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.

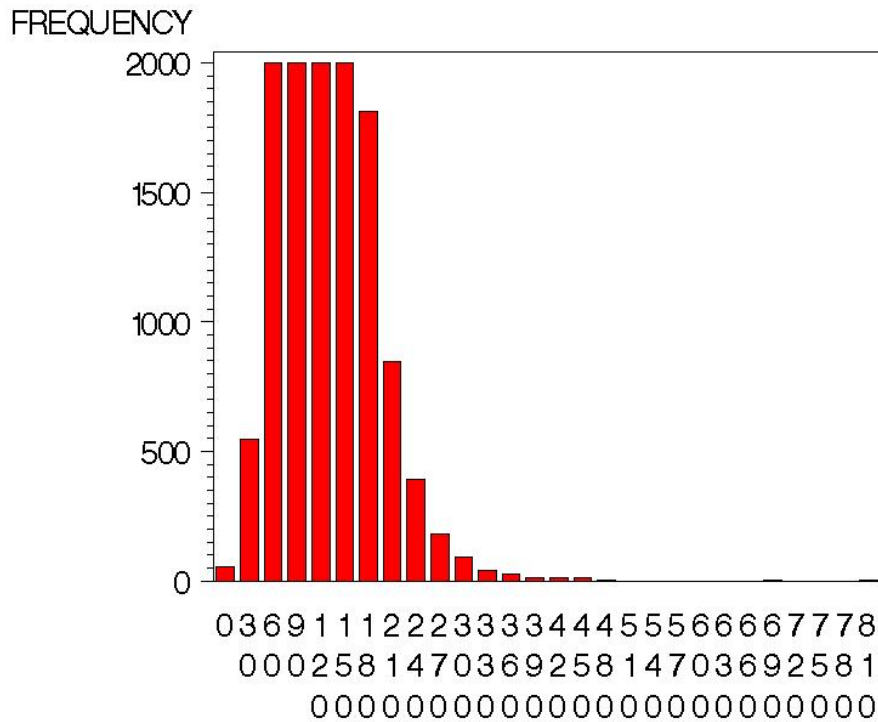
Statistics Needed: Background Correction

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

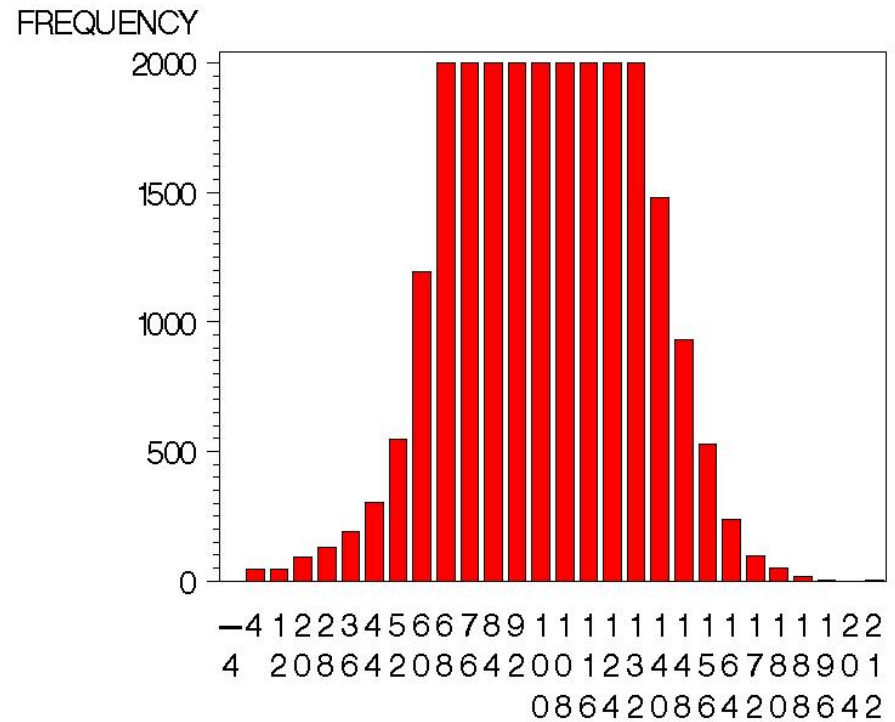


Statistics Needed: Outliers

- Delete if $\text{signal-median} > 16 * \text{MAD std.dev.}$



the coefficient of variation, Signal_Med



the coefficient of variation, Signal_Med

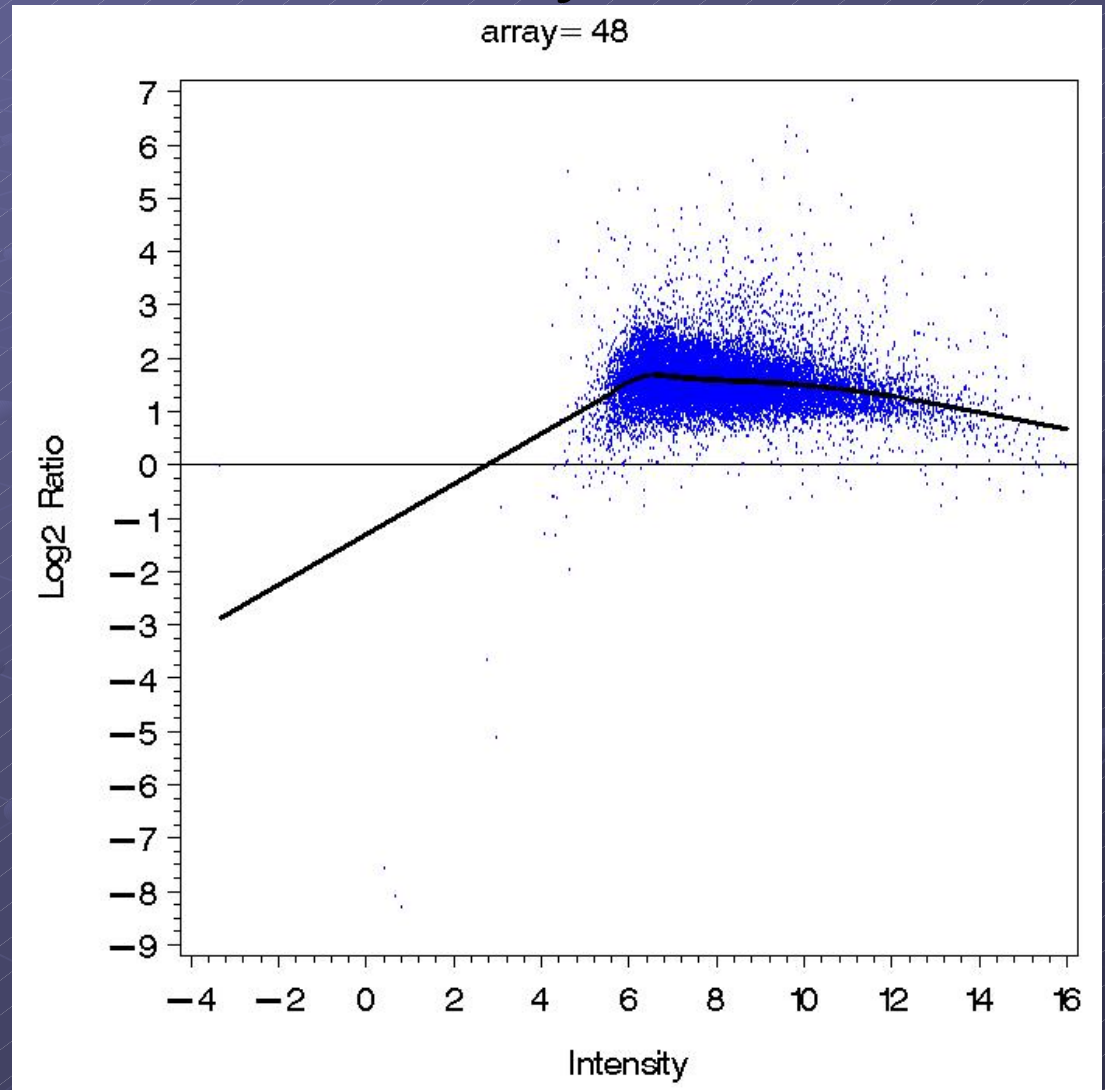
Data Removed

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

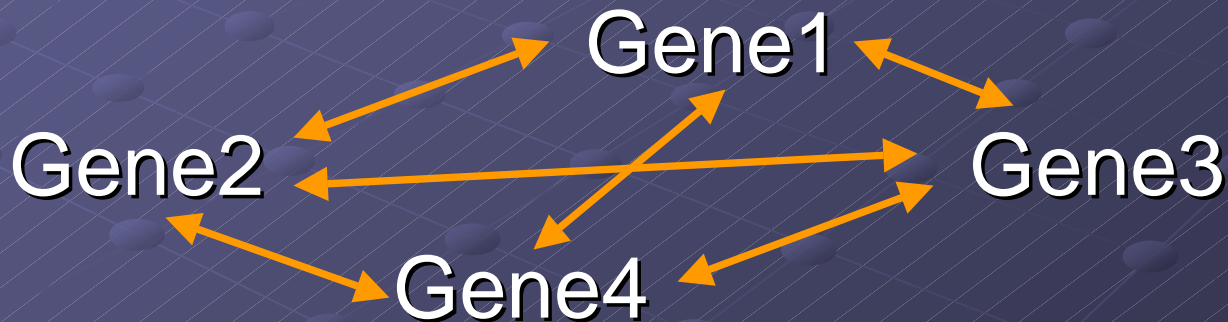
$$e_{ijkl} = u + Array_i + Dye_j + Strain * Dose_{kl} + error_{ijkl}$$

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

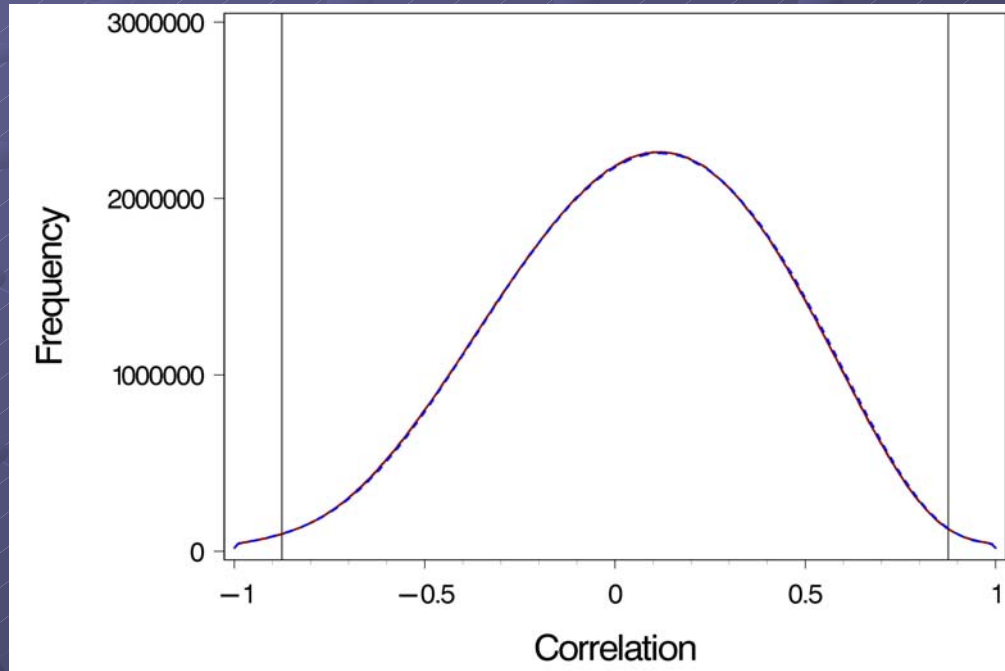
- Objective: Find clusters of genes to help discover function.
- Clique is a completely connected subgraph.



- We compute cliques for both un/exposed

Statistics Needed: Distance

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

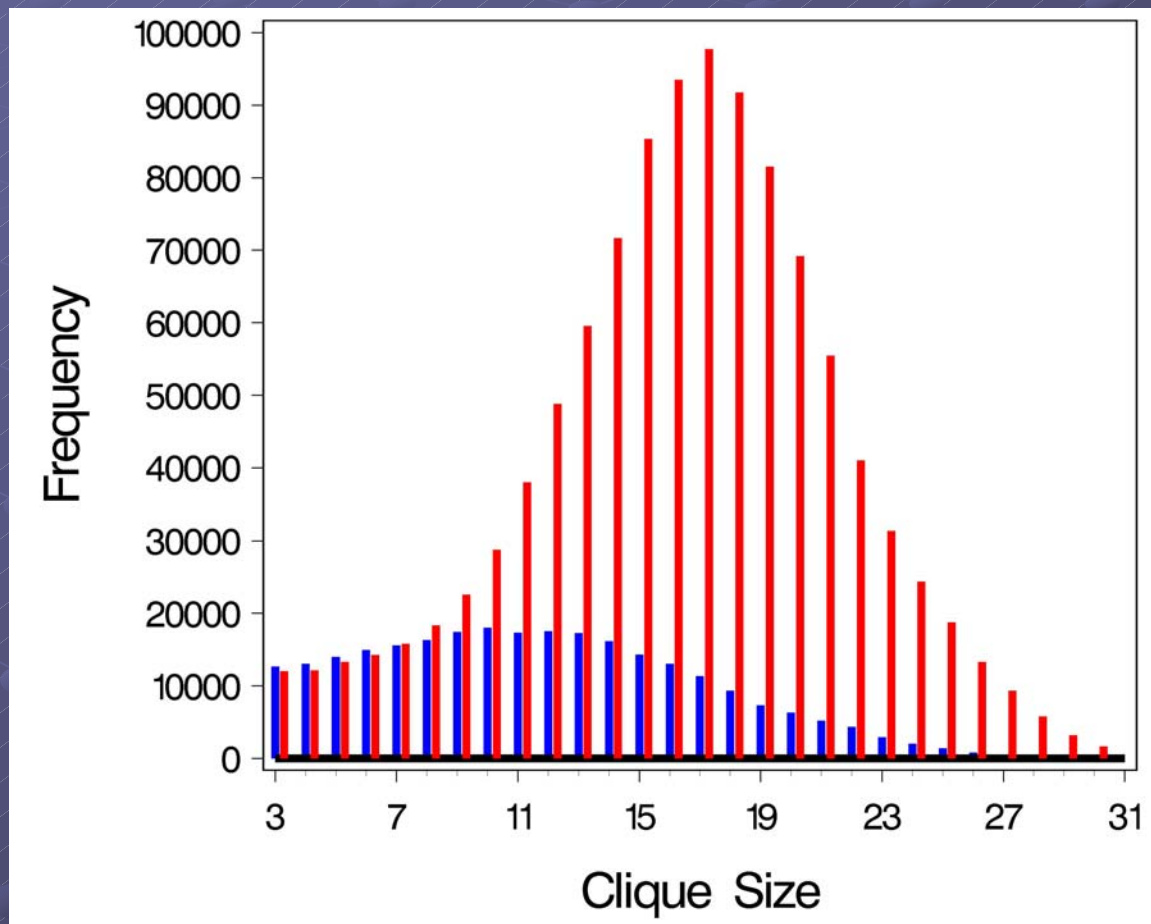


Statistics Needed: Threshold

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

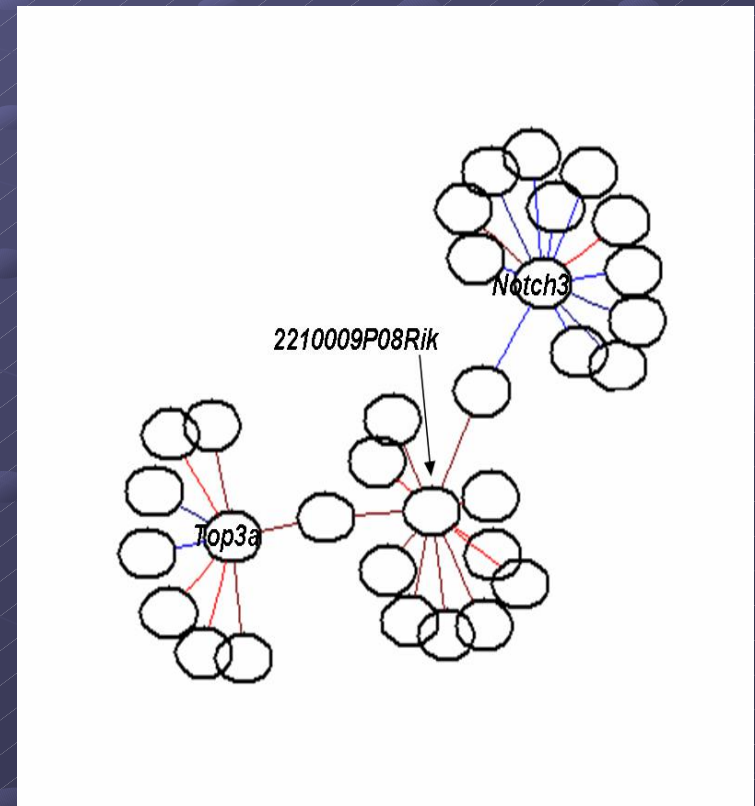
Statistics Needed: Clique P-values

- 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.