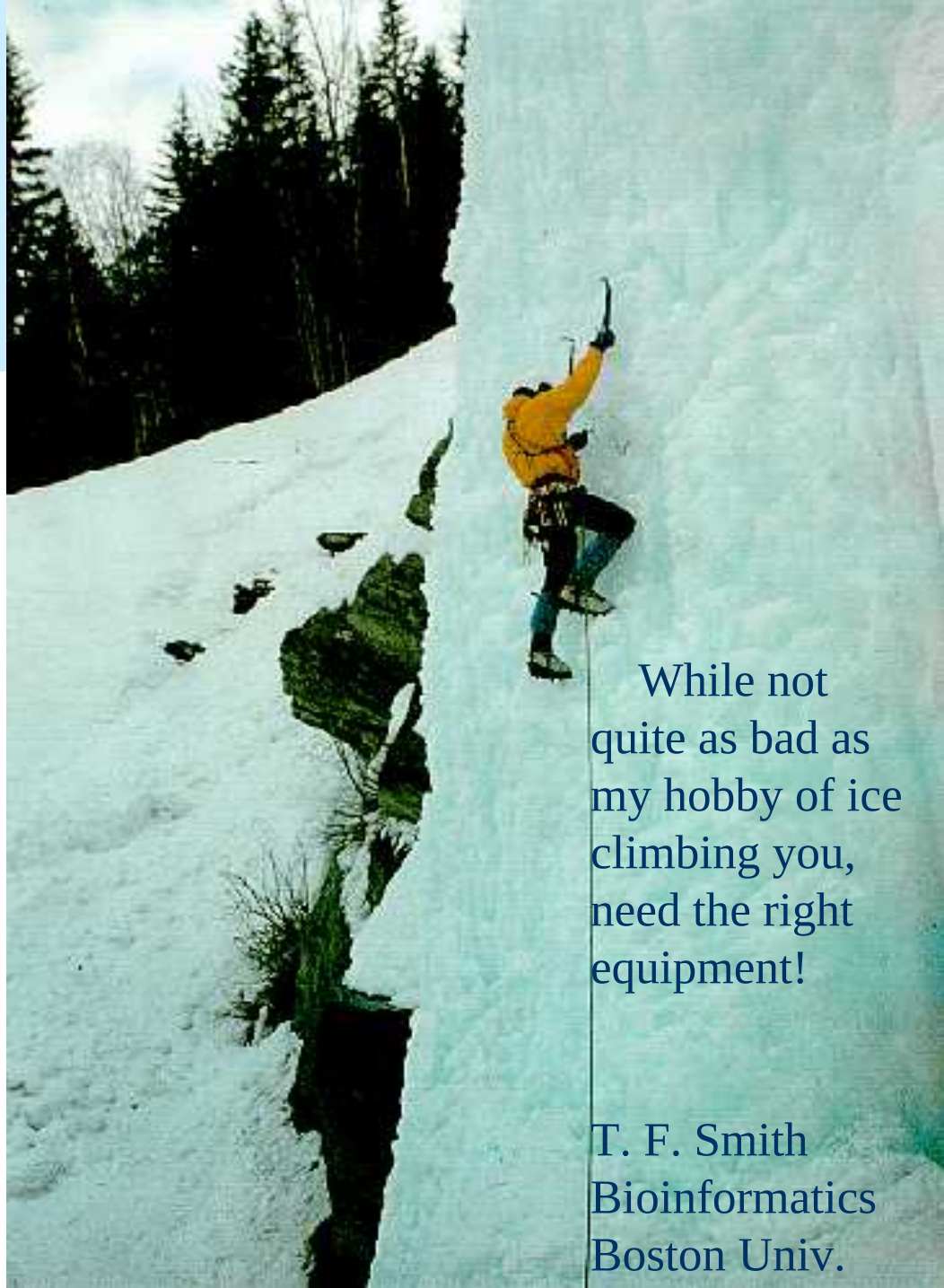
# Microarray analysis challenges.



While not quite as bad as my hobby of ice climbing you, need the right equipment!

T. F. Smith
Bioinformatics
Boston Univ.

# Experimental Design Issues

- Reference and Controls RNA choices
- Separate or pooled
- Number of replicates

    Independent or true replicates

First, is this a simple two treatment/condition comparison?

or

Multi-treatment/condition comparison experiment?

Also is it anticipated that there maybe latter additional data for comparison? If so the this should affect the design!

**A one time diagnostic two sample comparison**:

Analysis requires only the identification of a given subset of genes with changed.

Then any standard normalization and reference might do.

**For multi-sample/treatment comparisons or those with latter additional data for comparison:**

The choice of a control or reference RNA is critical.

Note that under such a design you will have measured the the reference RNA expression profile many more times than any of your query samples and this could provide a rather robust description of the reference sample! And, thus removing much of the noise associated with that sample.

# Reference or Control Comparative Samples

In the simple cases the reference RNA is just from the “untreated” sample.

In time course expression profiles there are at least two choices:

A single “zero time point” sample or,
a set of untreated samples, one for each time point.

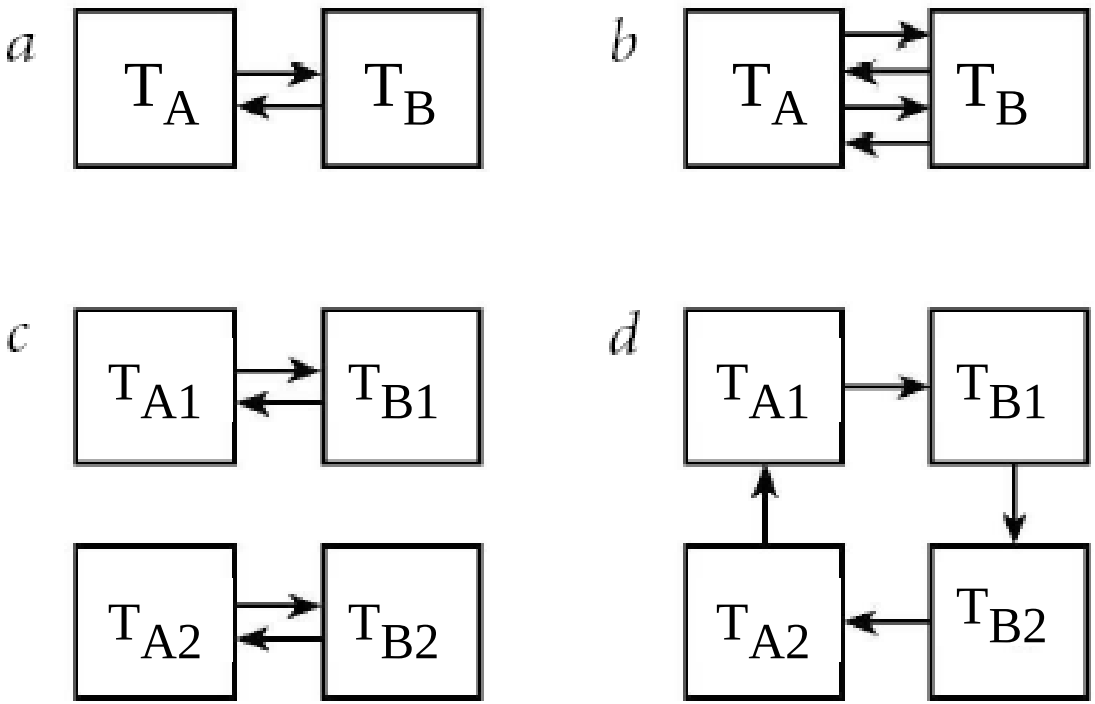
In all cases one needs a reference RNA sample that has few if any nonexpressed genes of interest!

Dividing by zero is a problem!

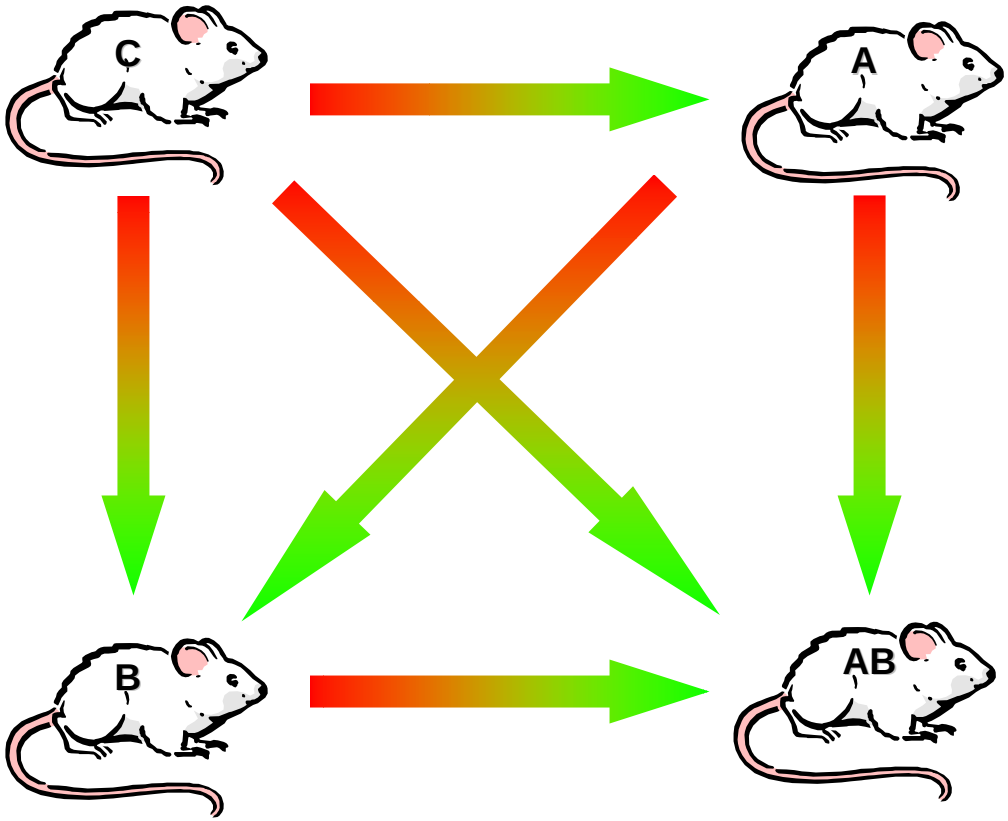
Know your probable sources of variation, and which of those you need to understand for the questions under investigation:

This will influence you design and the numbers of replicates.

# Experimental Design: Direct comparison of samples



# Factorial Designs- 2 x 2



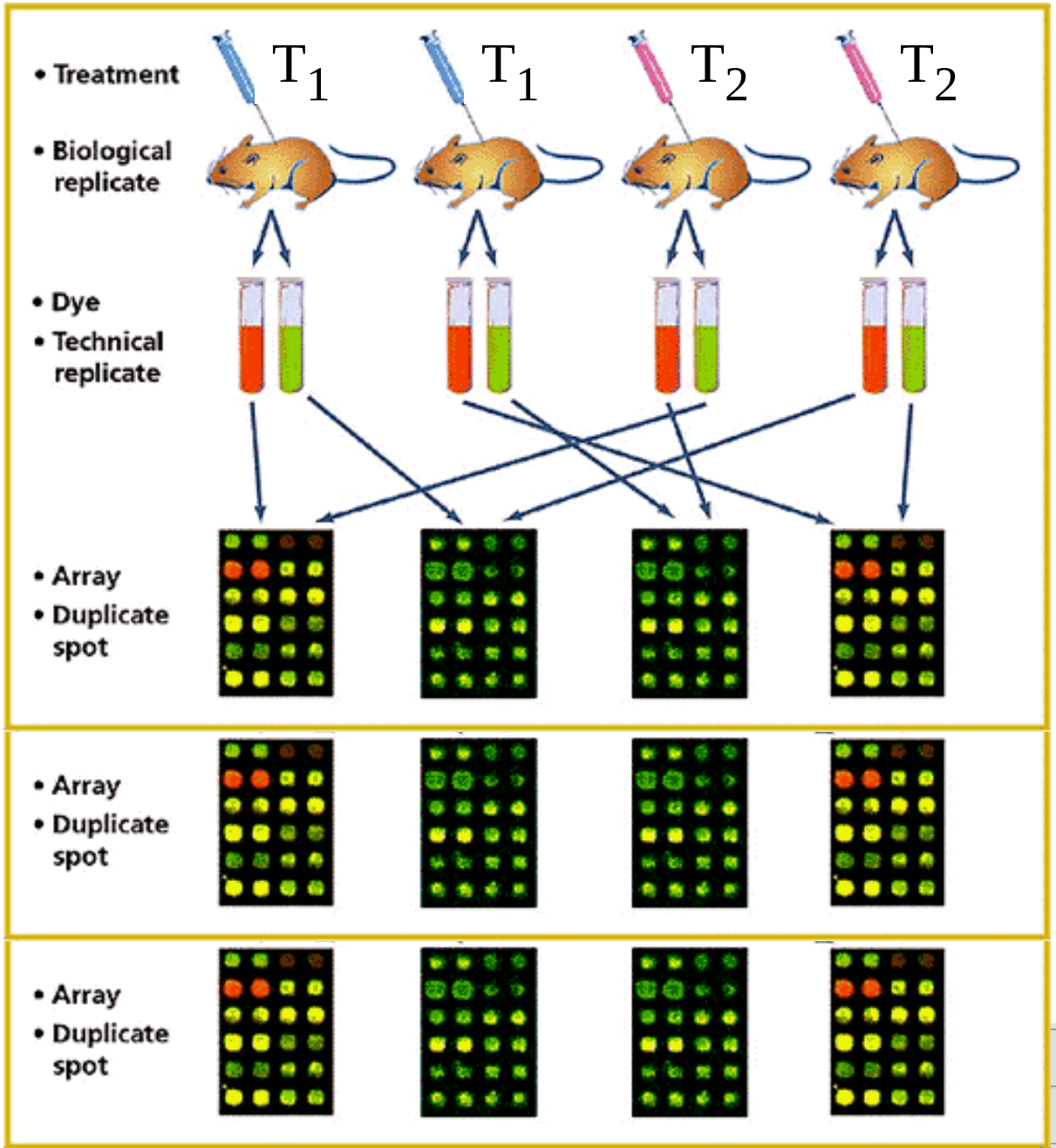

- Effect of treatment $T_A$ is estimated as $\log(T_A/C)$
- Effect of treatment $T_A$ in the presence of treatment $T_B$ is $\log(T_{AB}/T_B)$
- $\log(T_{AB}/T_B)-\log(T_A/C)$ is the *Interaction* between treatments

# Levels of Replication

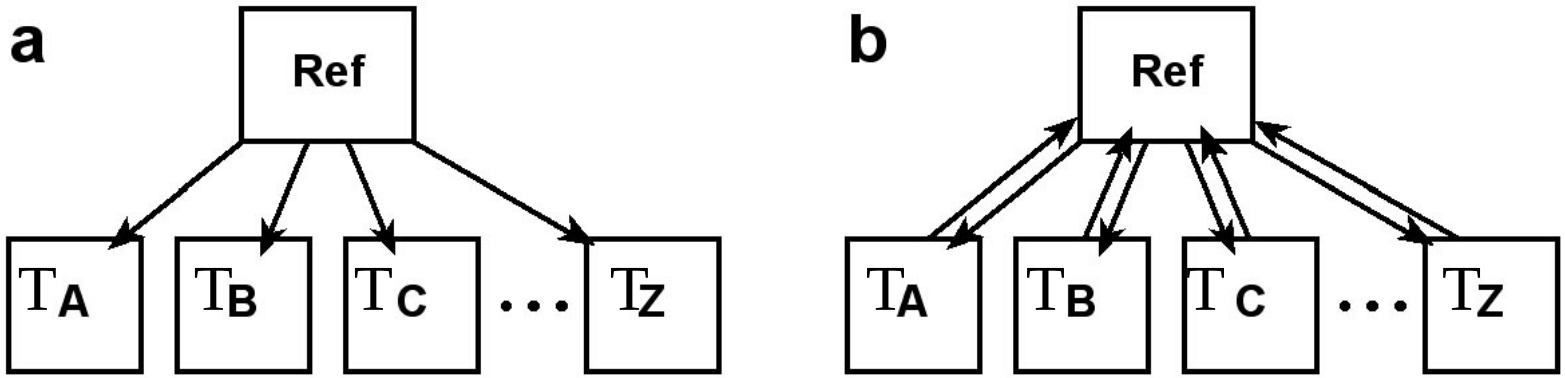
- Multiple arrays.
- Multiple spots containing the same DNA oligo sequence on the same microarray.
- Multiple spots containing different oligo sequences that assay the same gene RNA on the same microarray.
- Multiple spots containing different oligo sequences that assay different RNA products from the same gene on the same microarray.

- Replication of total experiment or
- Replication of just the hybridization step or
- Replication of only some other set of steps.

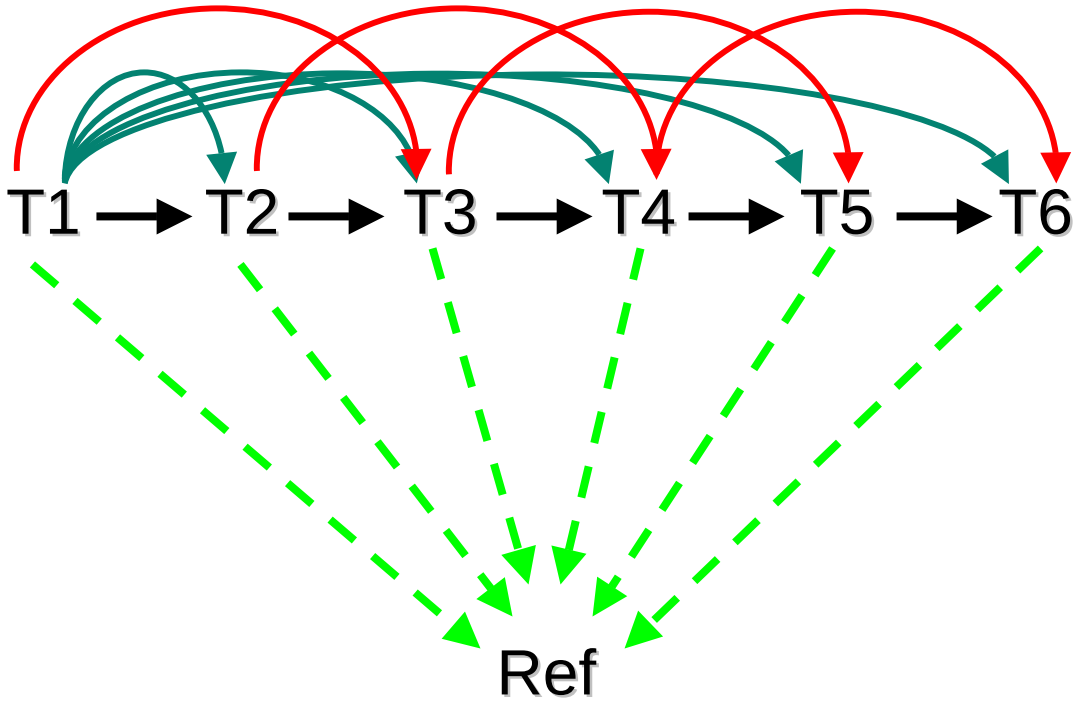
# A typical two condition comparison experiment



# Indirect Multiple Comparison: Reference Sample



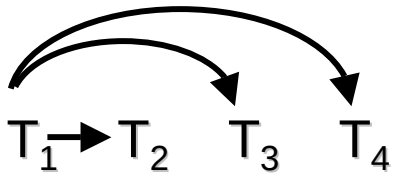
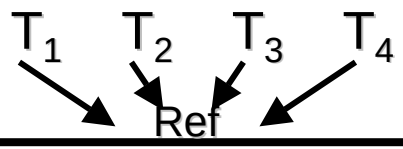
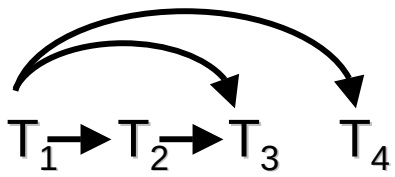
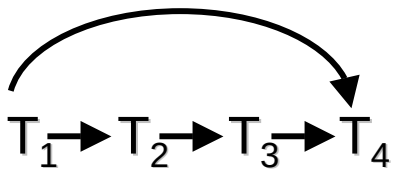
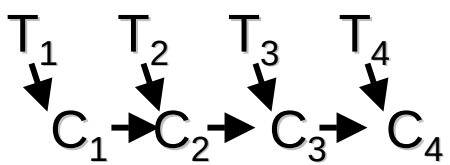
# Time Course Studies




- Possible designs
  - All samples vs common pooled reference
  - All samples vs time zero
  - Direct hybridization between times

pooled ref
compare to *t1*
*t vs t+1*
*t vs t+2*

# Design choices for time course studies

| | | Ave variance |
|---|---|---|
| N=3 |   ($T_1$ as common ref) | 1.5 |
| | $T_1 \rightarrow T_2 \rightarrow T_3 \rightarrow T_4$ (direct hybridization) | 1.67 |
| N=4 |   (common reference) | 2 |
| |   ($T_1$ as common reference + add.) | 1.06 |
| |   (Loop design) | 0.83 |
| N=7 |   timed common ref's) | ~1.0 |

# General rules of thumb when using common references

- More than 4 time points: use common reference design, unless other considerations take precedence (wt vs mutant time course)
  - Common reference designs give *extensibility*, and the ability to make *pair-wise comparisons*
- If possible use mRNA of scientific interest as common reference (control, wt, or time zero)
  - If no common reference is available
    - Universal total RNA set (Stratagene).
    - Untreated time sampled.
    - Pool of all points across time points.

# Design Summary

- Balance Dyes using Dye-Swap or Loops
- Use Independent Biological Replicates (unless one must average out some known biological variation).
- Use Technical Replicates (at all levels at least initially to identify sources of variation).

**After normalization we can combine true replicates.An An absolute minimum of three, 3,** array replicates for each experimental condition is required. The minimum to identify potential "outliers" or inconsistencies! However five is a more realistic minimum.

One normally calculates the log ratios for each gene represented on spot, i of array, k .

$$R_{ik} = \mathrm{Log}_2 [Cy3_{ik} / Cy5_{ik}] = \mathrm{Log}_2 [T_{ik} / C_{ik}]$$

Why Log? It is symmetric for the ratio.

Given normalized data one combined them
Producing an "averaged" ratio,

$$\mathrm{R_i} = \mathrm{Log}_2[ < (R_{ik} * MM ) / M_k >_{\mathrm{ave}}] .$$

Many forms of "average" can be used over the, n, equivalent arrays (see this afternoon's discussions).

## Normalizing microarray data

Sources of systematic variation will affect different microarray experiments to different extents. Thus to compare microarrays, such systematic variation as:

- Differences in labelling efficiency between the two dyes ( Cy3 &Cy5).
- Differences in the power of the two reading lasers.
- Differences in amount of, or quality of, the two RNA samples.
- Spatial biases in ratios across the surface of the microarray.

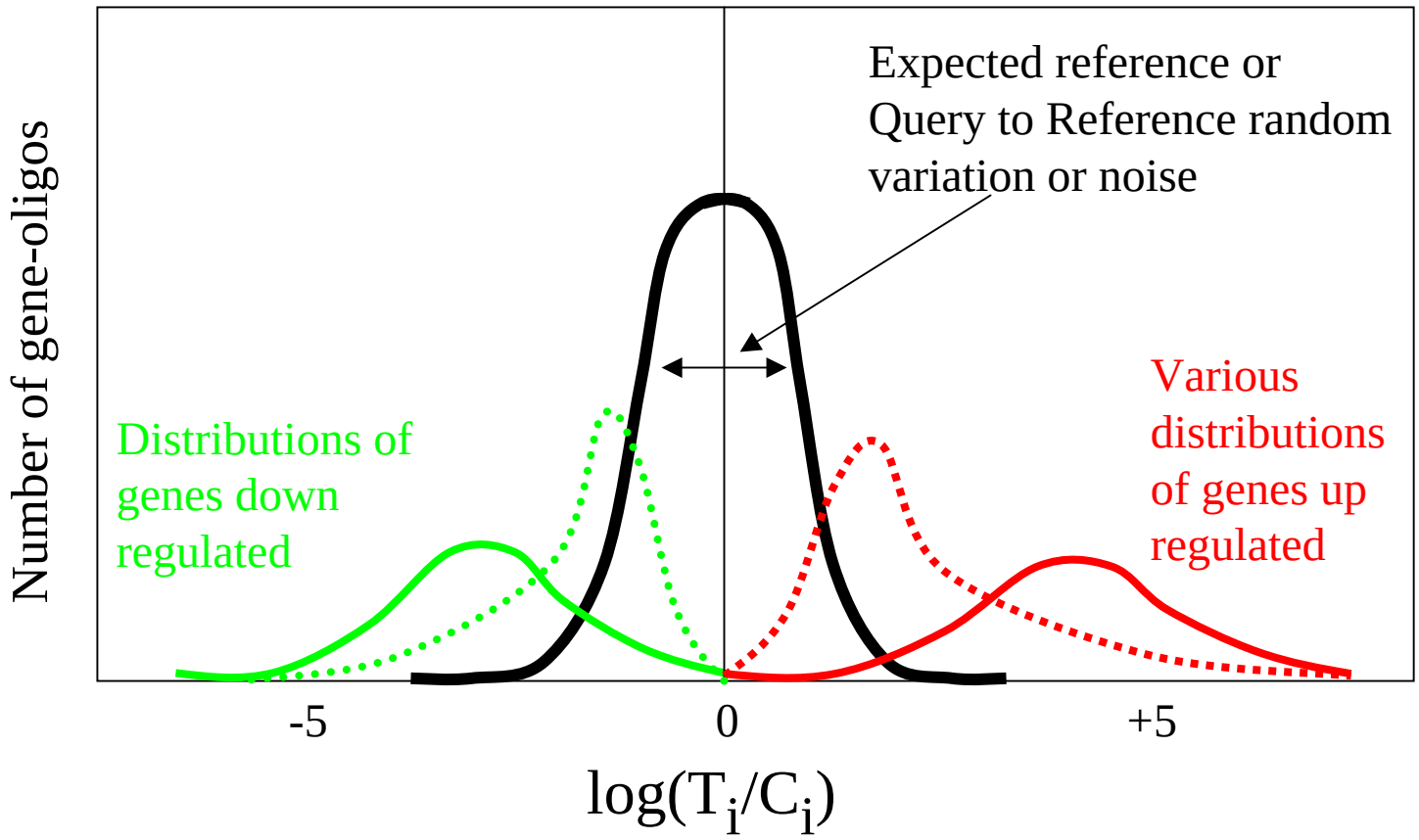
**Normalization will gone over this afternoon.**

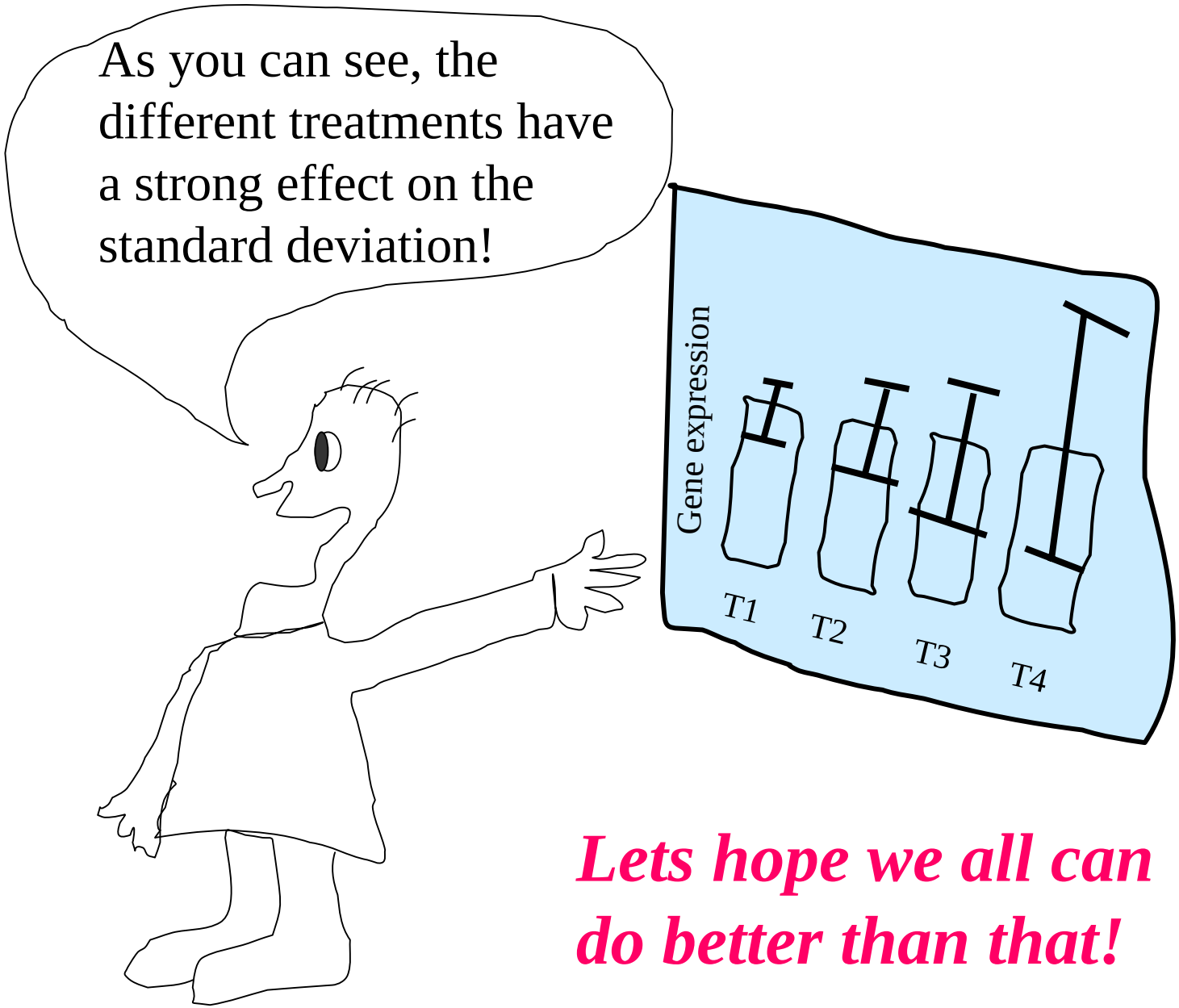
**Other sources arising from the chip construction*:**

- Spatial biases in spot density across the surface of the microarray.
- Spatial biases in oligo quality across the surface of the microarray.

*These have been discussed earlier.

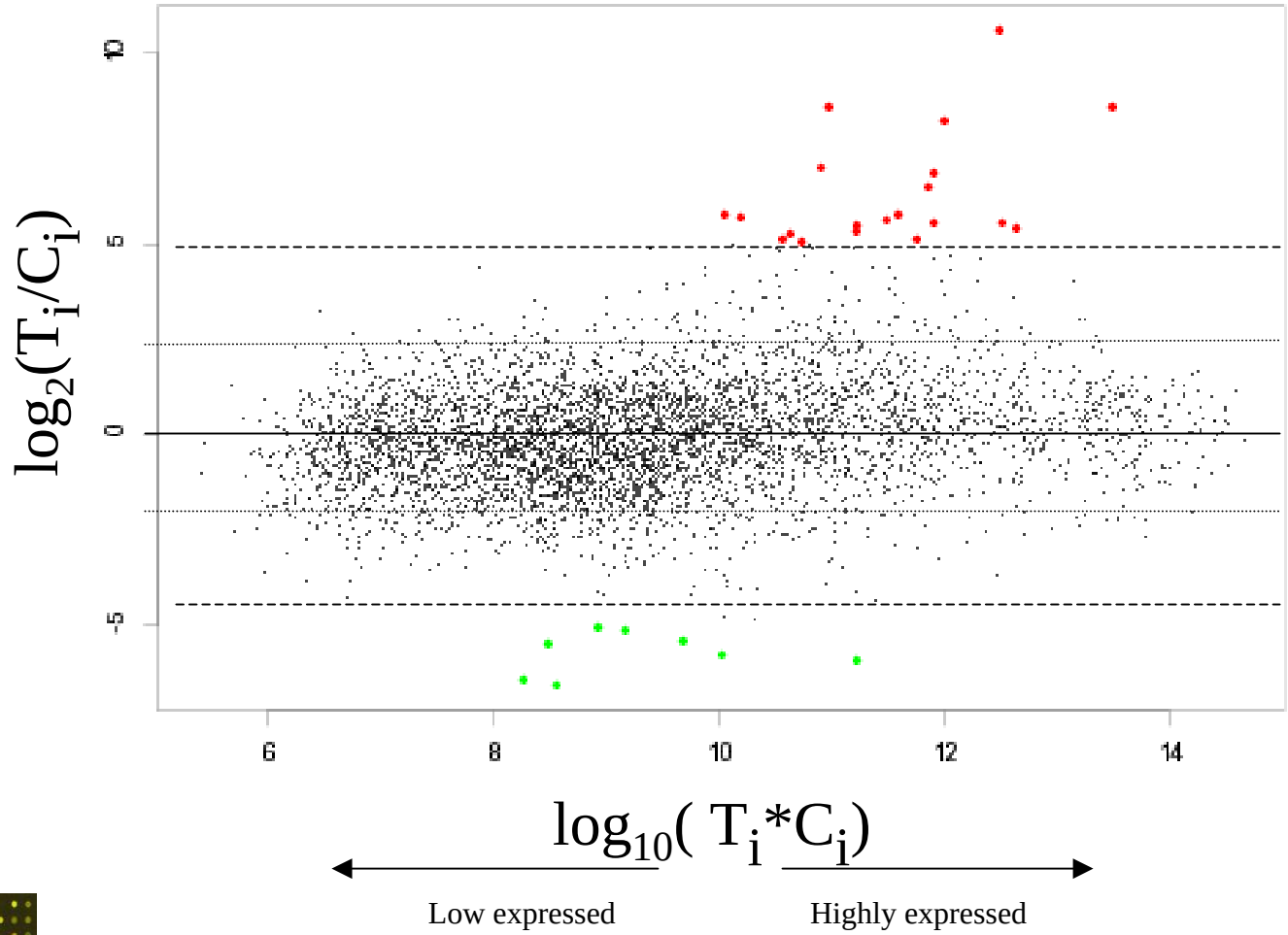
# What we might expect for the distribution of ratio values:






***Lets hope we all can do better than that!***

# Plot of Real Normalized data




Genes expressed up relative to reference by a factor of 32.

Genes expressed down relative to reference by a factor of 1/32.

**These built in controls:**

•Duplicate same gene array elements;

•Common set of “house keeping genes”;

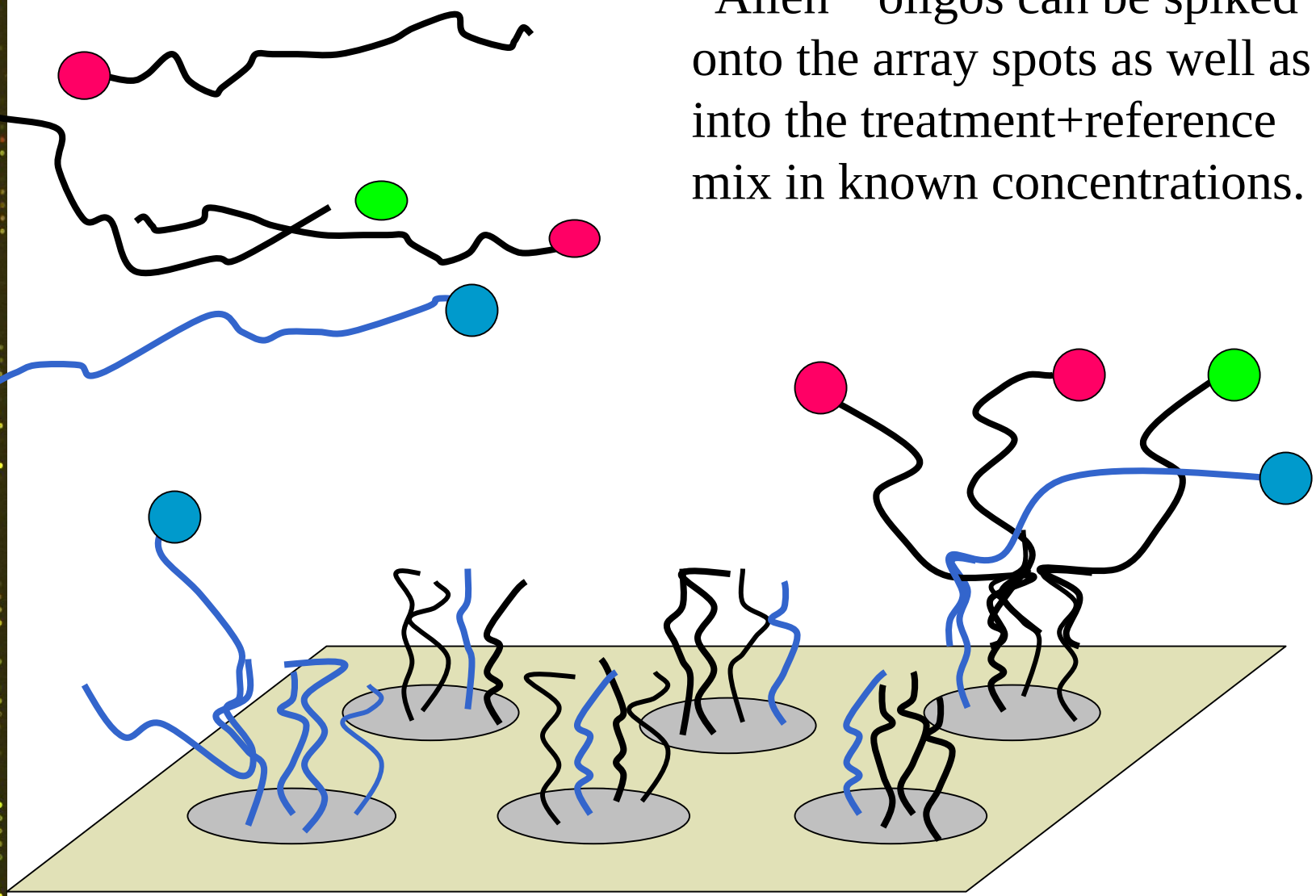
•Foreign gene spots and spikes; and/or

•Alien gene spots and spikes.

Can provide a normalization across a single or across multiple arrays, given the assumption of non-variance from one sample or experiment to another.

The Alien control oligos are designed specifically not to match (hybridize) with either your comparative reference or query RNAes. In addition “alien genes” can be constructed to match multiple alien oligo spots. These will then provide a positive query or reference spiking control. Particularly useful here is labeling by a third color dye.

In a collaboration between the MGH-PGA and Modular Genetic Inc.



"Alien " oligos can be spiked onto the array spots as well as into the treatment+reference mix in known concentrations.

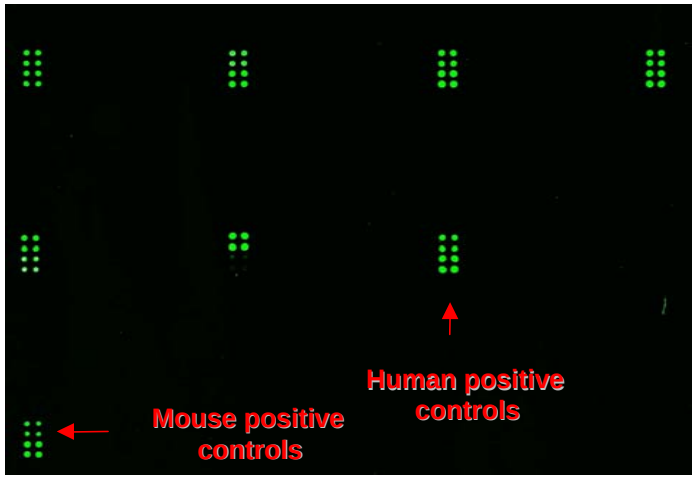
This would allow “normalization” and analysis of each spotted oligo represented gene independently!

The requirements are at least three dyes, a three color reader, and very exact measurements of the alien oligos and of the align gene spikes!

$$R_i = \mathrm{Log}_2 \{(T_i / C_i)([\mathrm{A}_{\mathrm{gene}}\%]/[\mathrm{A}_{\mathrm{oligo}}\%]\}$$

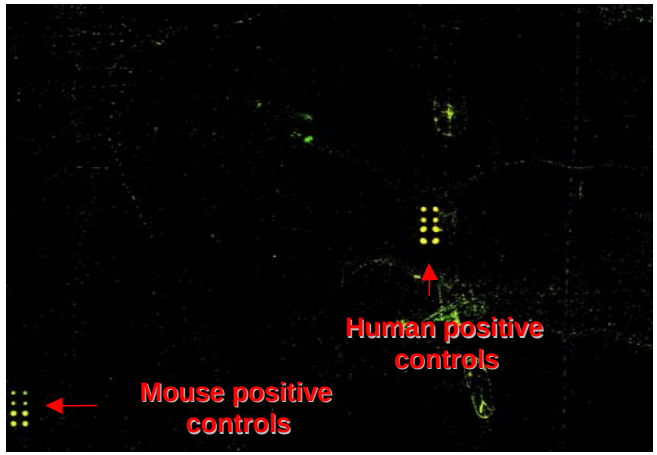
# Alien Oligonucleotides: Test of cross- hybridization

Terminal deoxynucleotidyl transferase Labeling (dCTP-Cy3)




Alien oligos and controls all label using the TdT labeling method

Hybridization with Stratagene's Universal RNA Mouse (Cy3) and Human (Cy5) Sets




No significant cross-hybridization to alien oligonucleotide 70mers

# So you now have an idea as to which genes represented on the chip did something of interest, now what?

First what was the original question?

Was it a simple diagnostic comparison?

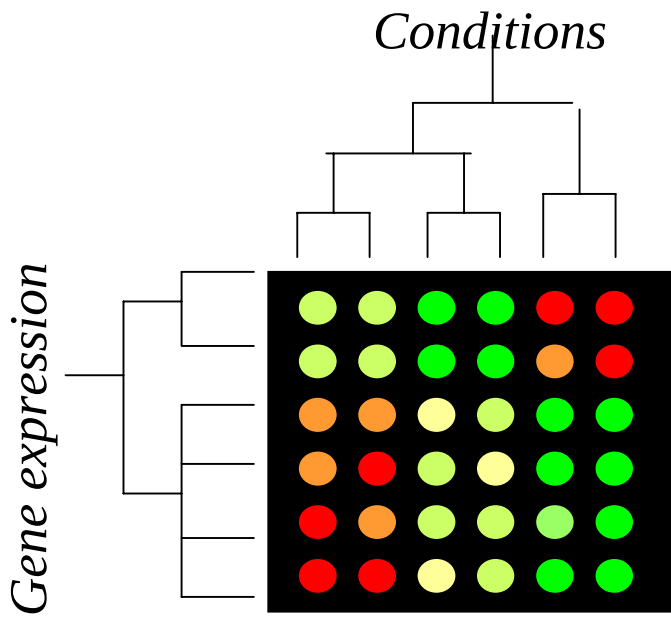
Need verify by qPCR?

Was it a multi treatment or time course?

Need to identify similar behaving genes?

You need to identify and cluster similar behaving genes:

Gene can be clustered by their expression behavior,
by their biochemical functions,
cellular roles, or common regulatory sites,
by the treatments or conditions
and/or
by any combination there of.



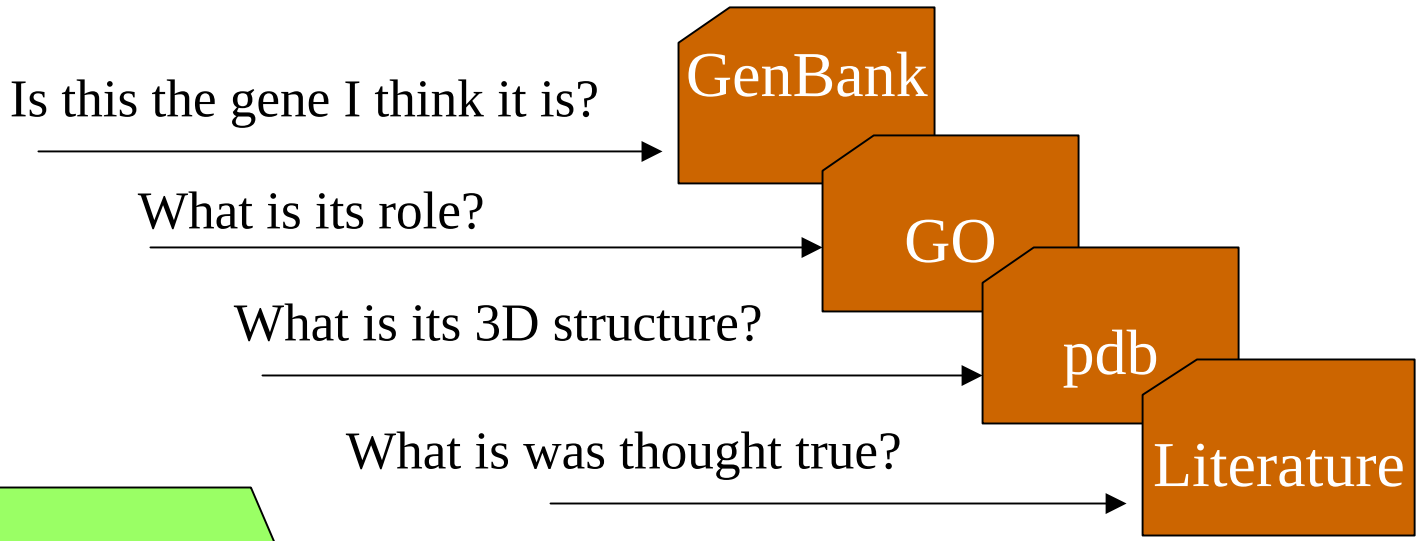

There are many clustering methods based on different assumptions. All however, group by some measure of similarity. (see latter discussions.)

The identification of gene biochemical function, sequence or structural domain family membership, cellular or network role, developmental stage or up stream regulatory elements is at best difficult!

These generally depends on gene annotations and/or literature references. Neither of which are compete, consistent or error free. In addition, we have no truly really reliable algorithms for identifying such things as up stream regulatory sites.

Often, of course, what one hopes to infer from related gene expression profiles is one or more of the above. And this, is even harder.

Thus it is the interpretation of array expression data that is the major challenge. This requires the creation of complex data structures and links to many external databases.




**New knowledge and new questions.**

With whom does it interact? → Yeast 2 hybrid data

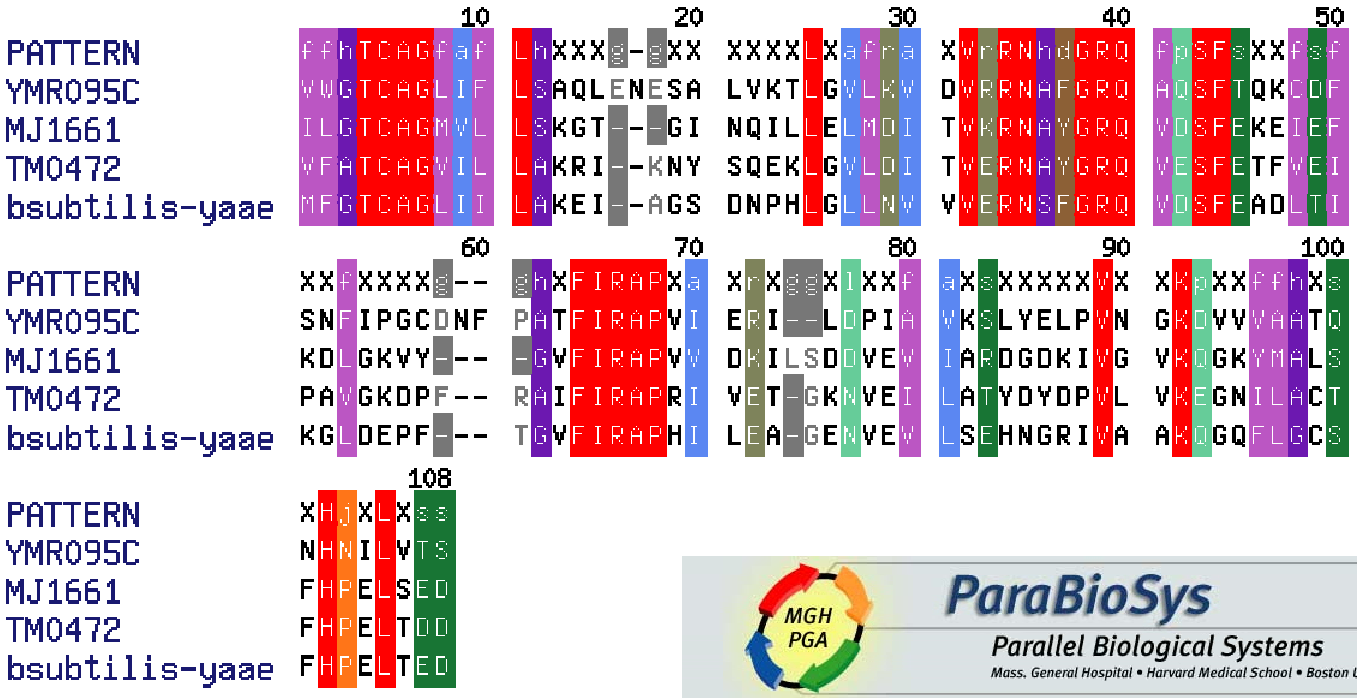
How is it regulated or what does it regulate? ↔ Gene Network models

**Sequence similarities represented as a shared pattern,**

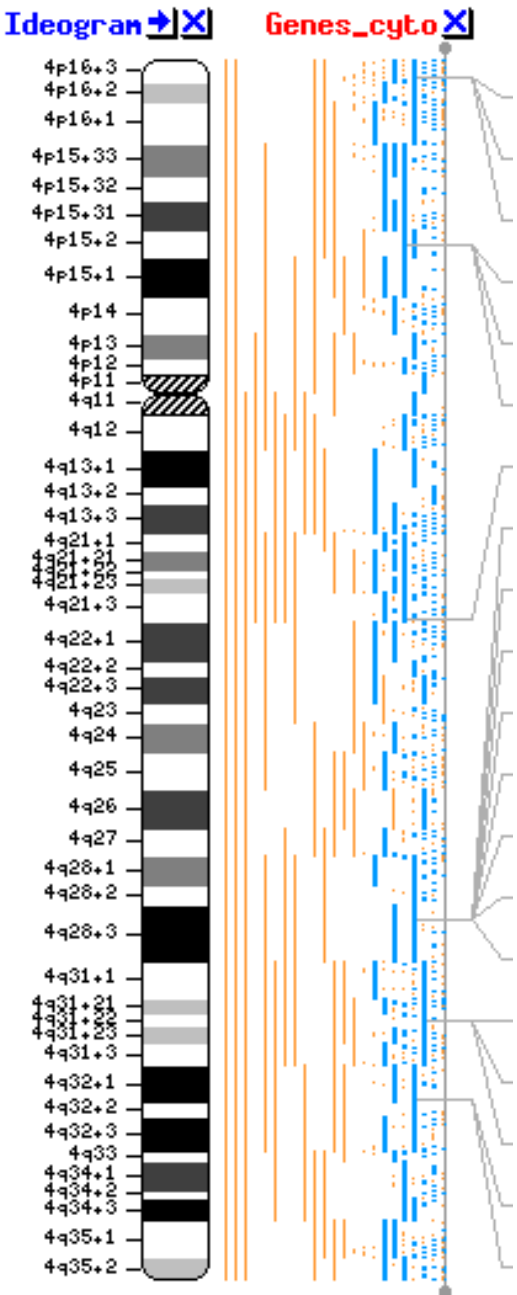
**ffhTCAGfafLhXXXggXXXXXXLXafraXVrRNhdGRQfpSFsXXffXXfXXXXgg**
**hXFIRAPXaXrXggXlXXfaX sXXXXXVXXKpXXffhXsXHjXLXss**

**Is powerful in identifying a functional family, but….**

| | |
|---|---|
| **MJ1661** | **conserved hypothetical {M. jannaschii}** |
| **TM0472** | **amidotransferase, putative {Mycobacterium leprae}** |
| **YMR095C** | **stationary phase induced gene {yeast}** |
| **bsubtilis-yaae** | **similar to hypothetical proteins {Bacillus subtilis}** |

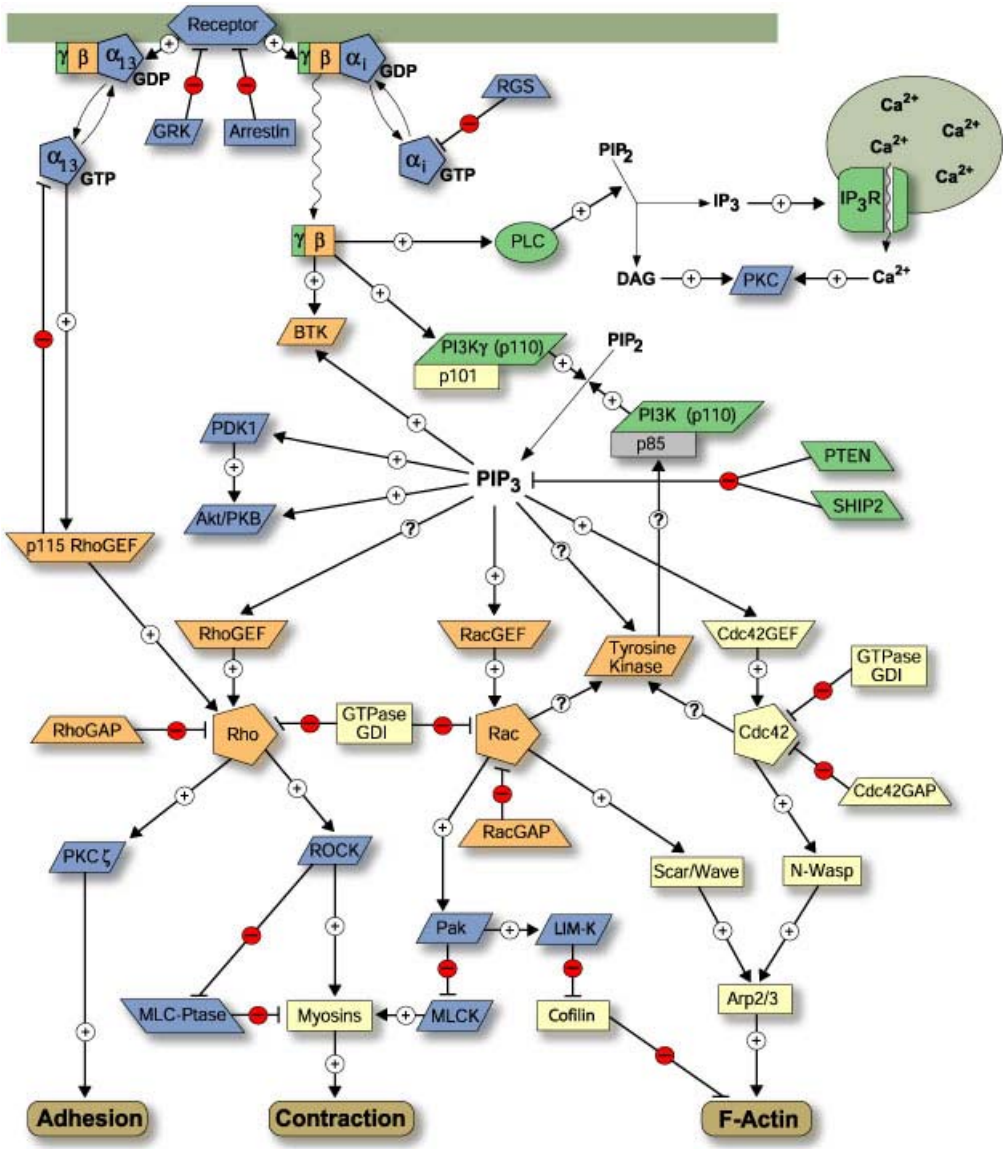


# Example of Homo sapiens chrom-4 gene annotations




| Symbol | cyto | Description |
|---|---|---|
| | 4p16 | LIM domain binding 2 |
| | 4p16 | Ellis-van Creveld-like syndrome |
| | 4p16 | Ksp37 protein |
| | 4p15 | Epilepsy, partial, with pericentral spikes |
| | 4p15 | ubiquitin specific protease 17 |
| | 4p15 | Parkinson Disease (autosomal dominant, Lewy body) 4 |
| | 4q21 | Hyper-IgE syndrome |
| | 4q28 | mastermind-like 3 (Drosophila) |
| | 4q28 | MAD, mothers against decapentaplegic homolog 1 |
| | 4q28 | SET domain-containing protein 7 |
| | 4q28 | RAB33B, member RAS oncogene family |
| | 4q28 | deafness, autosomal dominant 42 |
| | 4q28 | fibrinogen, gamma polypeptide |
| | 4q28 | fibrinogen, B beta polypeptide |
| | 4q28 | fibrinogen, A alpha polypeptide |
| | 4q31 | protocadherin 18 |
| | 4q31 | deafness, autosomal recessive 26 |
| | 4q31 | high-mobility group box 2 |
| | 4q32 | toll-like receptor 2 |
| | 4q32 | hepatitis B virus integration site 6 |

**An example regulatory network for which the Alliance for Cellular Signaling, AfCS*, is collecting vast amounts of gene expression time course profiles.**




*See AfCS at the Nature web site:

http://www.signaling-gateway.org/


Thanks for your kind attention.
Temple F. Smith
Bioinformatics
Boston University
Key West 2001
" Sunset Sail "
Sullivan

# Some reference from Dr. Churchill's group:


•Cui and Churchill(2003), How many mice and how many arrays? Replication in mouse cDNA microarray experiments, submitted to CAMDA '02 proceedings. Posted on 1/14/2003.

•Cui and Churchill(2002), Statistical Tests for Differential Expression in cDNA Microarray Experiments, submitted to Genome Biology. Posted on 12/27/2002.

•Cui, Kerr and Churchill(2002), **Data** Transformation for cDNA Microarray Data, submitted. Posted on 7/25/2002. Supplemental figures for the paper.

•Wu, Kerr and Churchill(2002), MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments, Chapter of The analysis of gene expression data: methods and software, in press, Springer (two color figures are here: Color figure 4 and Color figure 6).

•Cui, Hwang, Qiu, Blades and Churchill (2003), Improved Statistical Tests for Differential Gene Expression by Shrinking Variance Components, submitted. Posted on 10/24/2003.

# Microarray construction issues

•What attributes of a spot should be considered when determining its quality are:
•How close is it to saturation?
•How far above background is its signal?
•How consistent is the measured ratio for each pixel in the spot?
•How large is the spot?

•In addition to a metric of spot quality, there may also be usful metrics of array quality, eg:
•Is there evidence of spatial bias?
•What percentage of spots on the array are considered of good quality?
•What is the overall signal to background like?

The simplest normalization over a chips' two dyes is the total Dye ratio intensity normalization: (this assumes the total labled amount of RNA is approximately the same for each.)

$$N_k = \Sigma_i [Cy5_{ik}] / \Sigma_j [Cy3_{jk}]$$

Then

$$Cy3_{ik} = N_k * Cy3_{ik}$$

While

$$Cy5_{ik} = Cy5_{ik}$$

or vise versa.

Next one normally calculates the log ratios for each gene representing spot, i on array, k .

$$R_{ik} = \mathrm{Log}_2 [Cy3_{ik} / Cy5_{ik}]$$

Why Log? It is symmetric for the ratio.