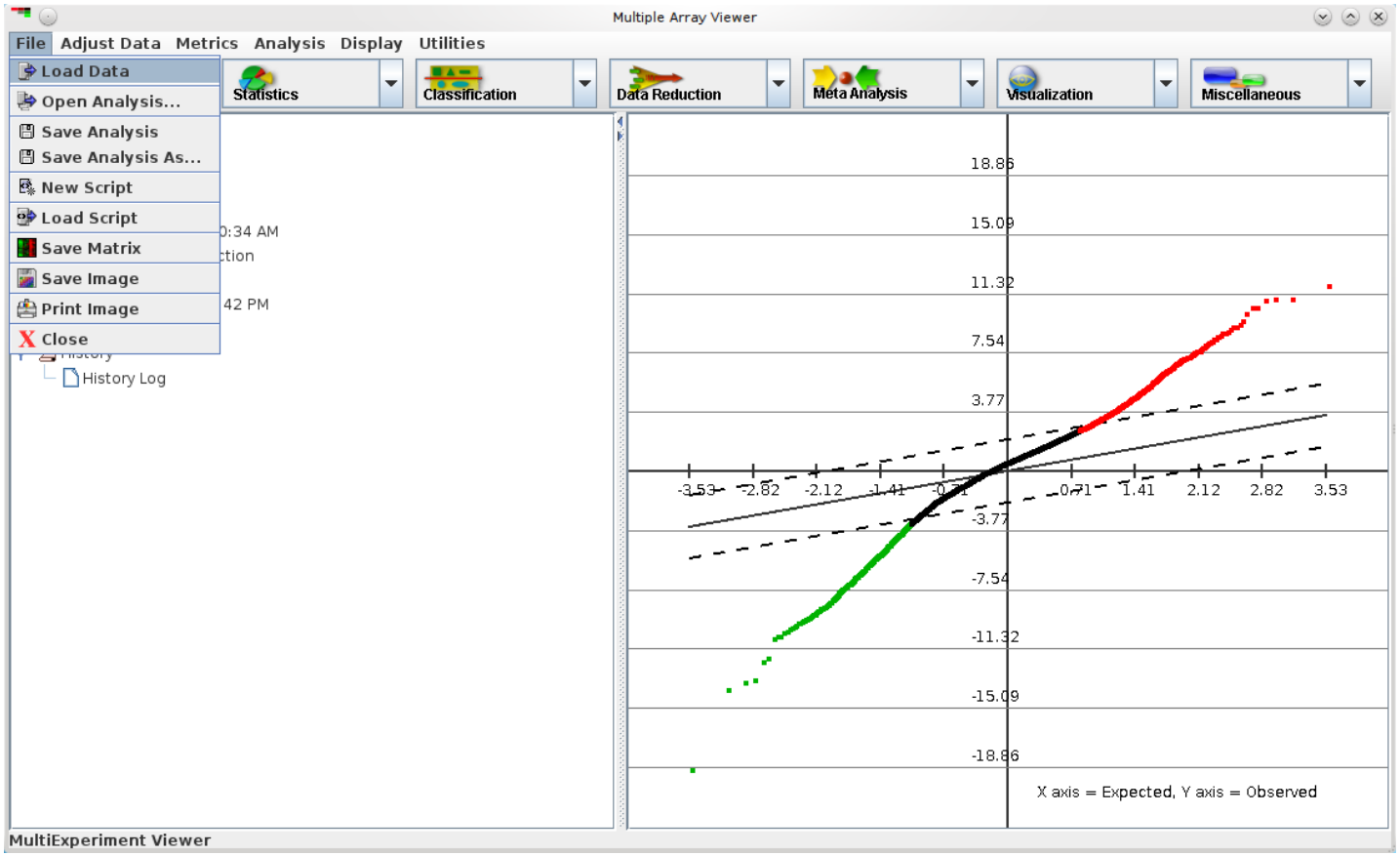
Estimating Relative Expression

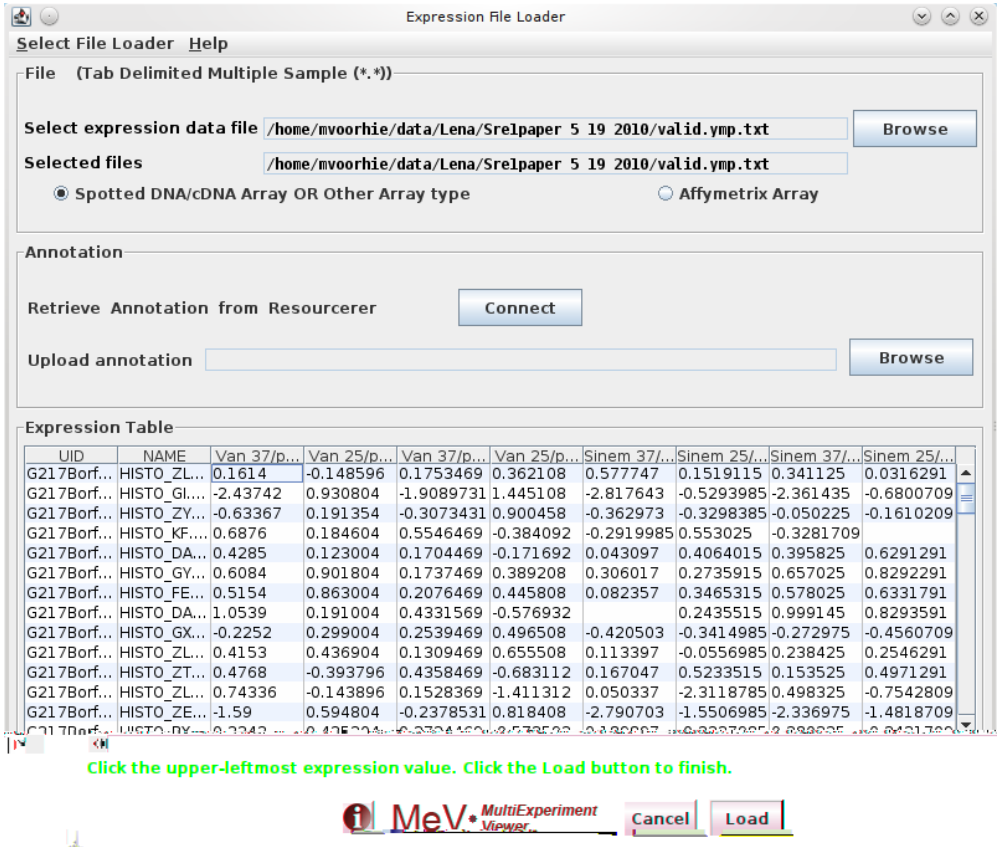
Mark Voorhies

4/6/2011

MeV SAM: Load data

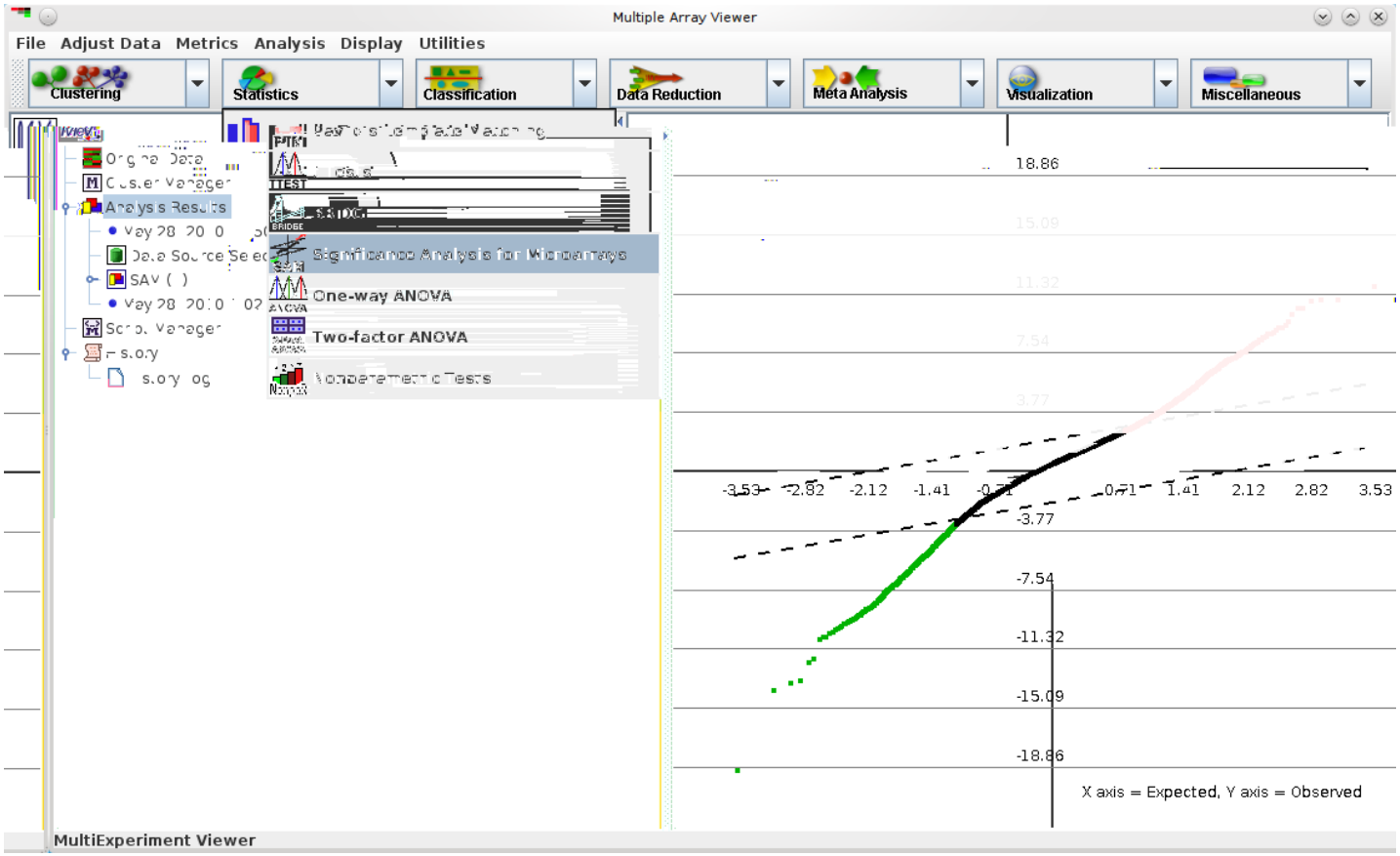


MeV SAM: normalized, log transformed, TDT data

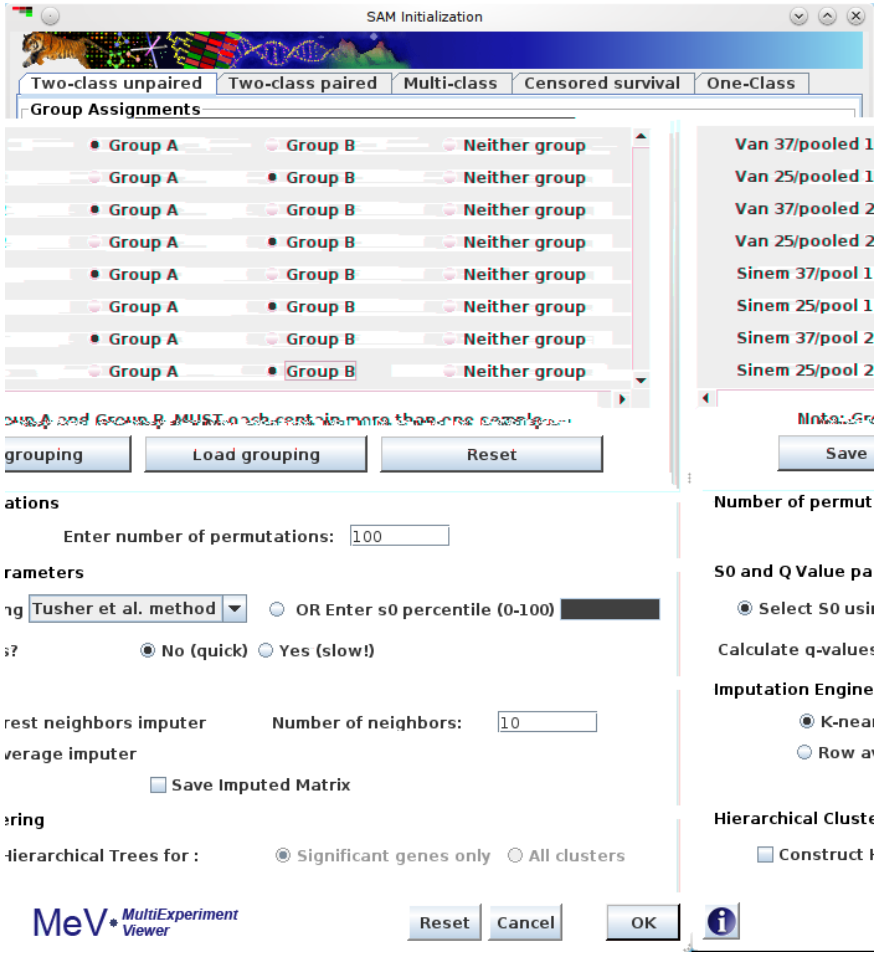


UID	NAME	Van 37/p...	Van 25/p...	Van 37/p...	Van 25/p...	Sinem 37/...	Sinem 25/...	Sinem 37/...	Sinem 25/...
G217Borf...	HISTO_ZL...	0.1614	-0.148596	0.1753469	0.362108	0.577747	0.1519115	0.341125	0.0316291
G217Borf...	HISTO_GI...	-2.43742	0.930804	-1.9089731	1.445108	-2.817643	-0.5293985	-2.361435	-0.6800709
G217Borf...	HISTO_ZY...	-0.63367	0.191354	-0.3073431	0.900458	-0.362973	-0.3298385	-0.050225	-0.1610209
G217Borf...	HISTO_KF...	0.6876	0.184604	0.5546469	-0.384092	-0.2919985	0.553025	-0.3281709	
G217Borf...	HISTO_DA...	0.4285	0.123004	0.1704469	-0.171692	0.043097	0.4064015	0.395825	0.6291291
G217Borf...	HISTO_GY...	0.6084	0.901804	0.1737469	0.389208	0.306017	0.2735915	0.657025	0.8292291
G217Borf...	HISTO_FE...	0.5154	0.863004	0.2076469	0.445808	0.082357	0.3465315	0.578025	0.6331791
G217Borf...	HISTO_DA...	1.0539	0.191004	0.4331569	-0.576932		0.2435515	0.999145	0.8293591
G217Borf...	HISTO_GX...	-0.2252	0.299004	0.2539469	0.496508	-0.420503	-0.3414985	-0.272975	-0.4560709
G217Borf...	HISTO_ZL...	0.4153	0.436904	0.1309469	0.655508	0.113397	-0.0556985	0.238425	0.2546291
G217Borf...	HISTO_ZT...	0.4768	-0.393796	0.4358469	-0.683112	0.167047	0.5233515	0.153525	0.4971291
G217Borf...	HISTO_ZL...	0.74336	-0.143896	0.1528369	-1.411312	0.050337	-2.3118785	0.498325	-0.7542809
G217Borf...	HISTO_ZE...	-1.59	0.594804	-0.2378531	0.818408	-2.790703	-1.5506985	-2.336975	-1.4818709

MeV SAM: Choose SAM

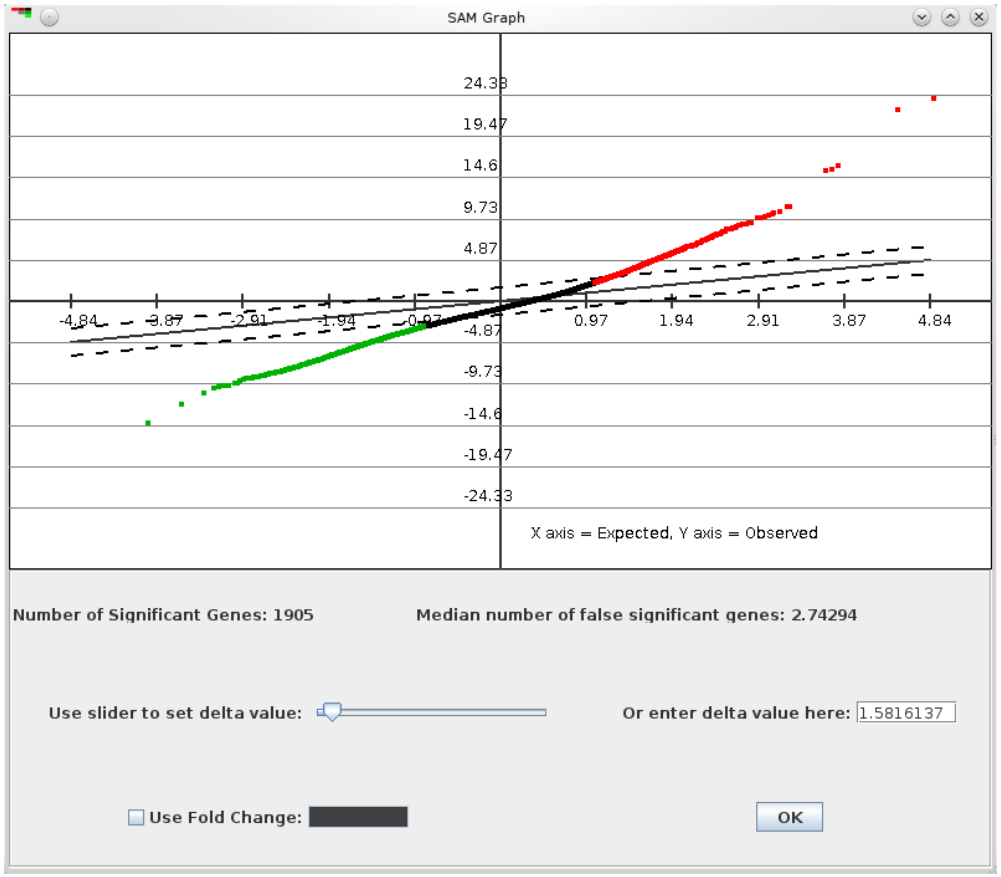


MeV SAM: Describe experiment, choose parameters



MeV SAM: Choose permutations for FDR

MeV SAM: Choose delta



PNAS 98:5116

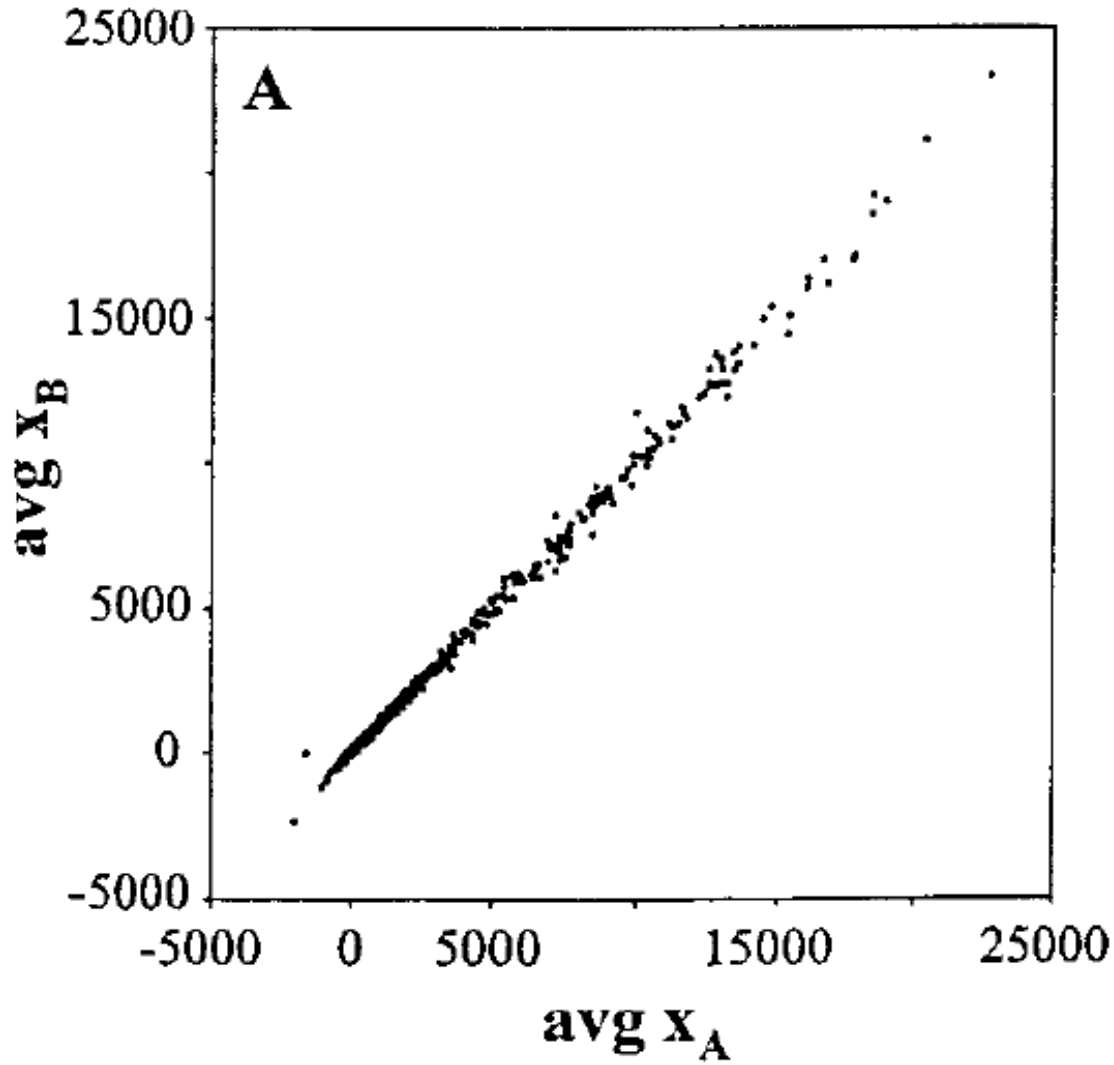
Significance analysis of microarrays applied to the ionizing radiation response

Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu

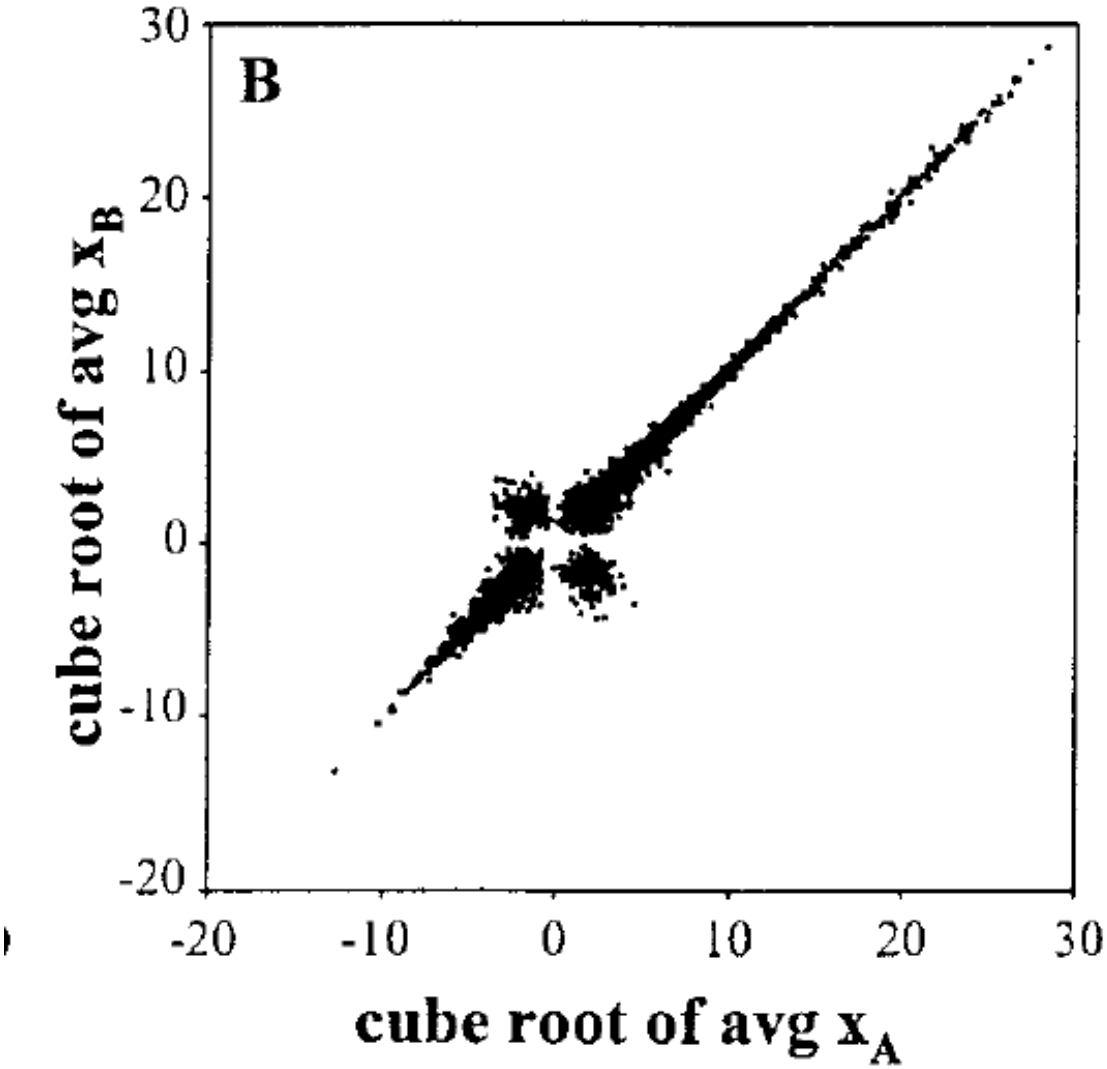
Normalizing Affy Arrays with Technical Replicates

This is very similar to mean normalization for two color arrays.

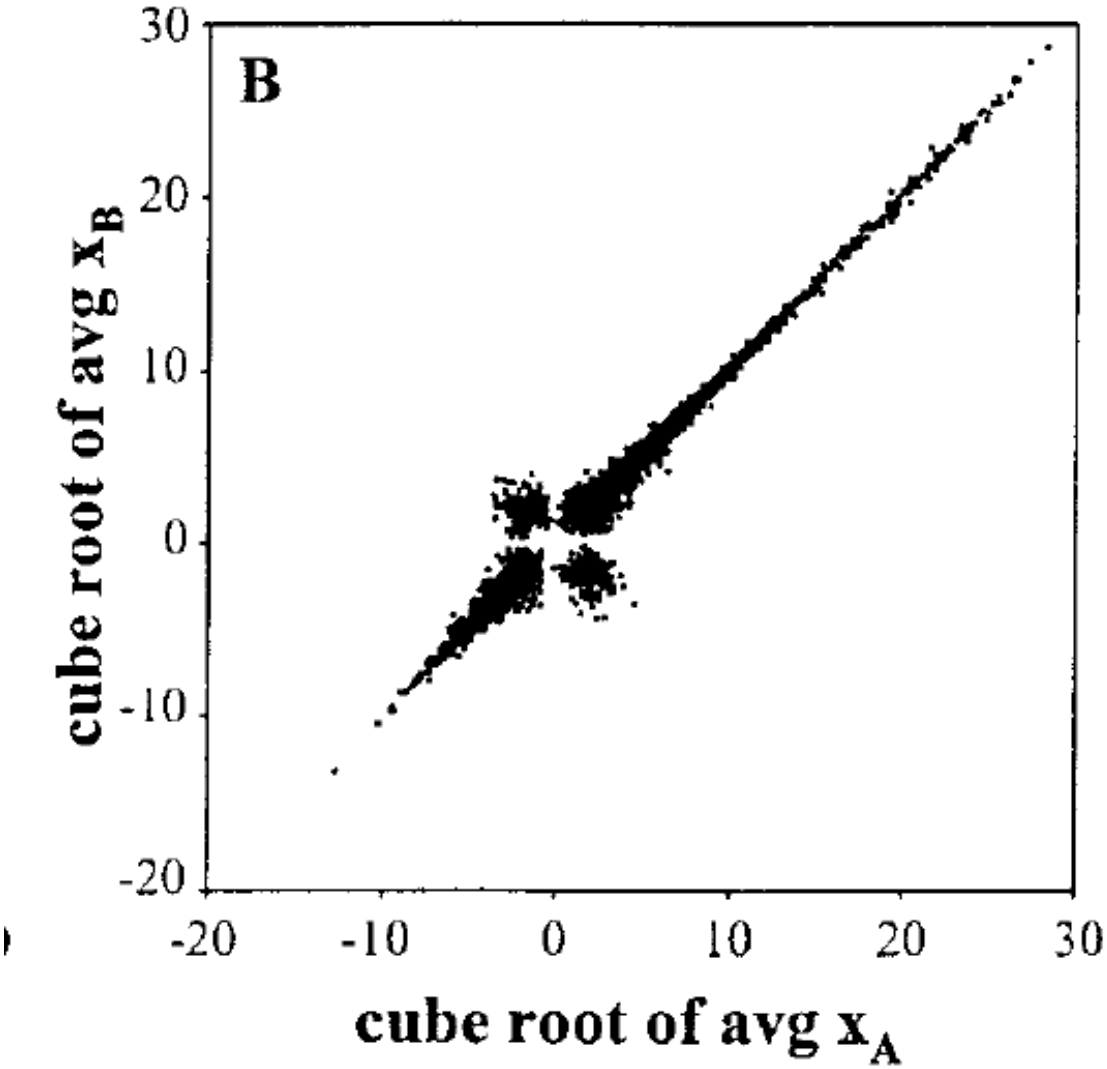
1a: Comparing normalized data



1b: Cube Root Transform

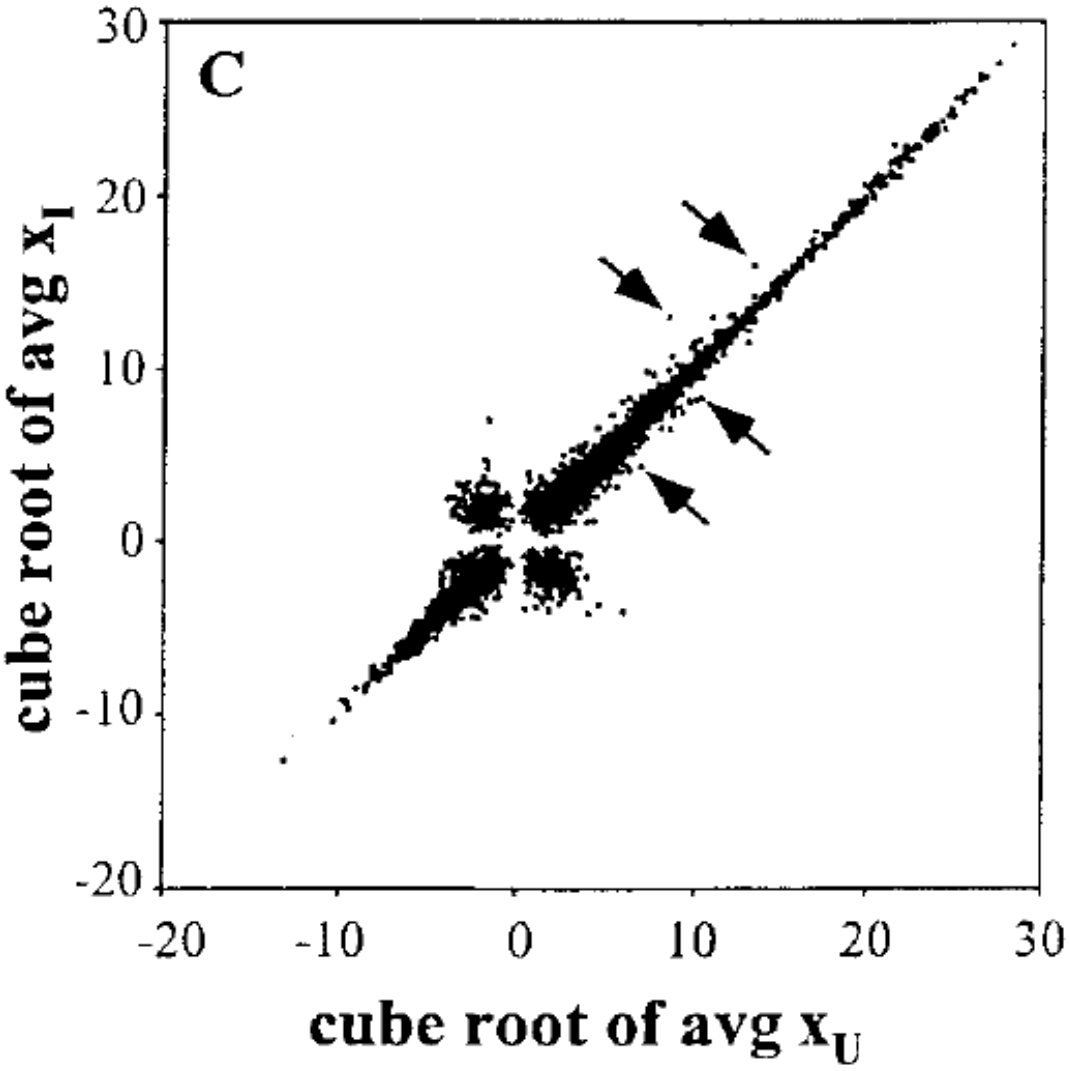


1b: Cube Root Transform

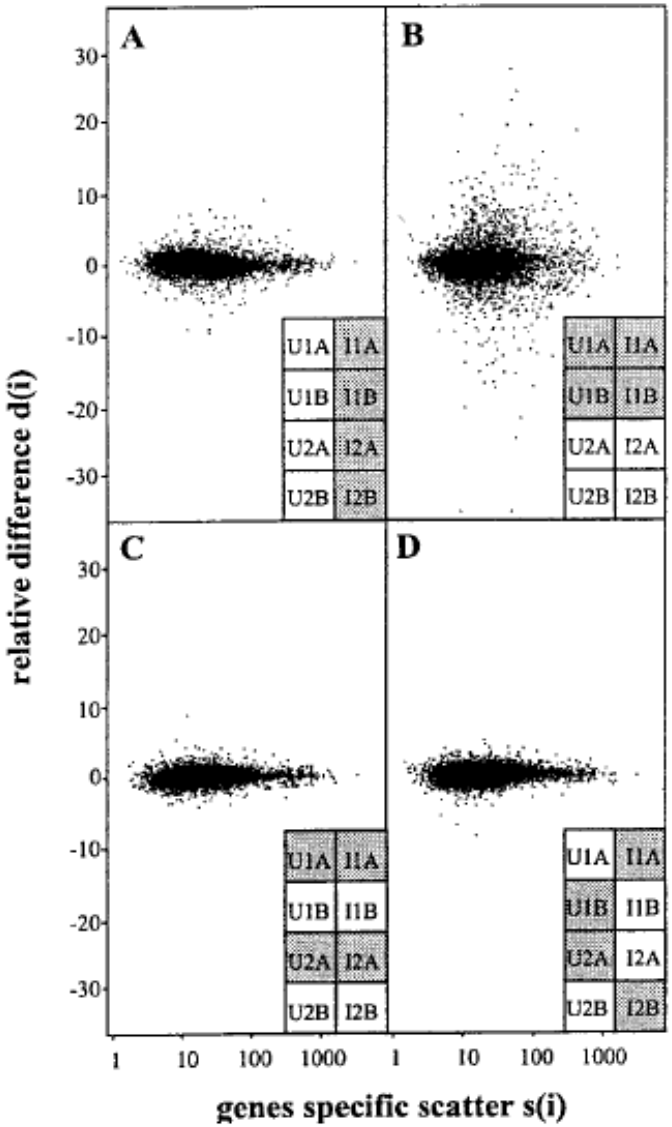


- Motivation is to get good resolution of all of the data
- Problems: weird behavior near zero, compression of error for negative values, not biologically motivated
- Better: filter low intensity data and log transform

1c: Outliers in treatment comparison

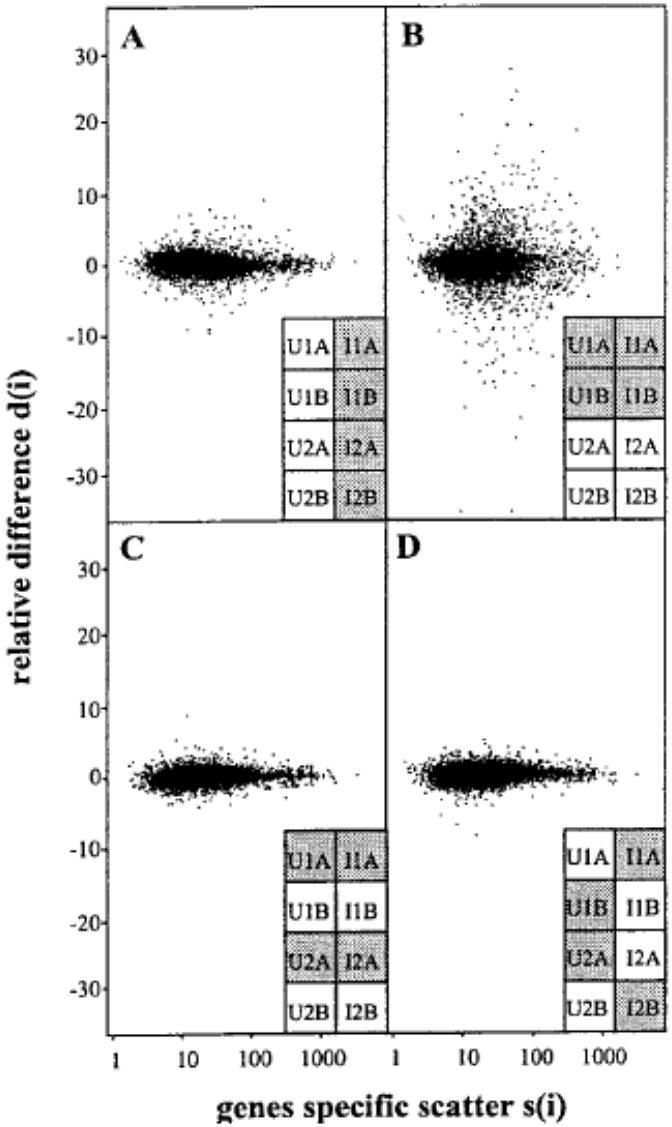


2: d(i) statistic



$$d(i) = \frac{\bar{x_I}(i) - \bar{x_U}(i)}{s(i) + s_0}$$

2: d(i) statistic

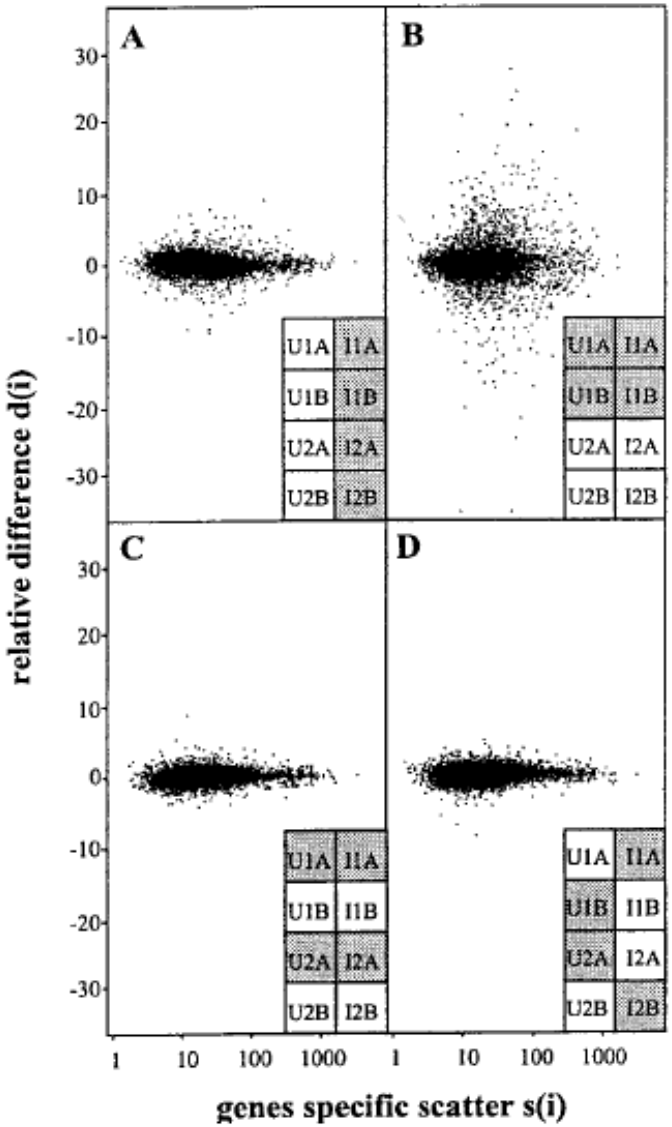


$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

$$s(i) = \sqrt{a \sum_m [x_m(i) - \bar{x}_I(i)]^2 + \sum_n [x_n(i) - \bar{x}_U(i)]^2}$$

$$a = (1/n_I + 1/n_U)/(n_I + n_U - 2)$$

2: d(i) statistic



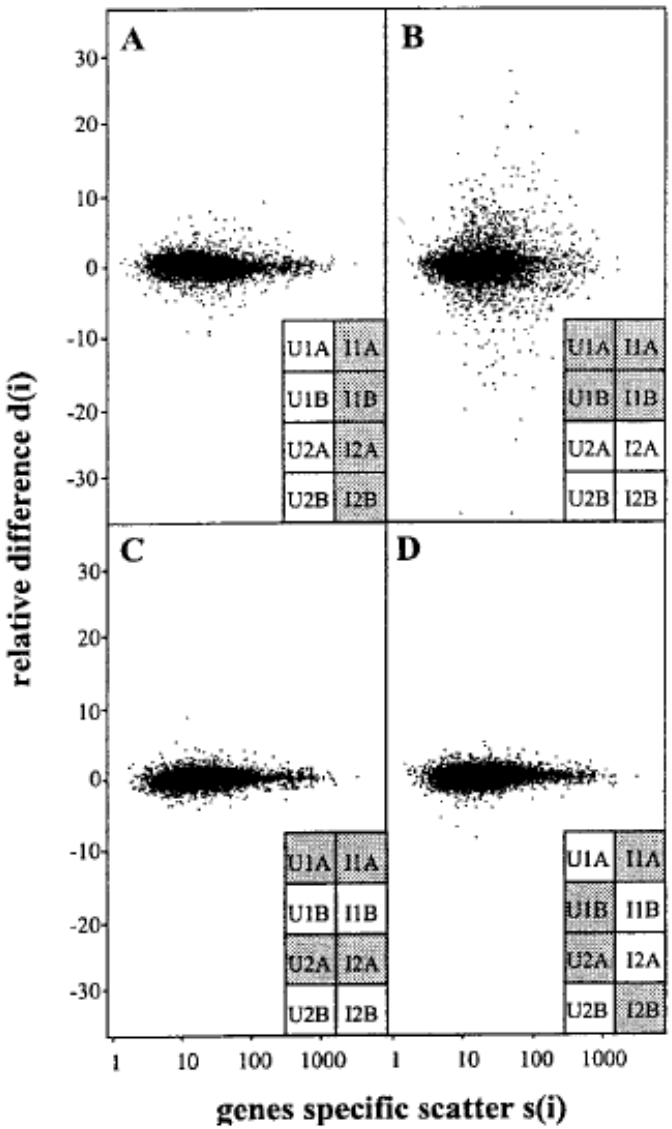
$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

$$s(i) = \sqrt{a \sum_m [x_m(i) - \bar{x}_I(i)]^2 + \sum_n [x_n(i) - \bar{x}_U(i)]^2}$$

$$a = (1/n_I + 1/n_U)/(n_I + n_U - 2)$$

- 1 vs. 2 = biological replicate
- A vs. B = technical replicate
- I vs. U = treatment

2: d(i) statistic



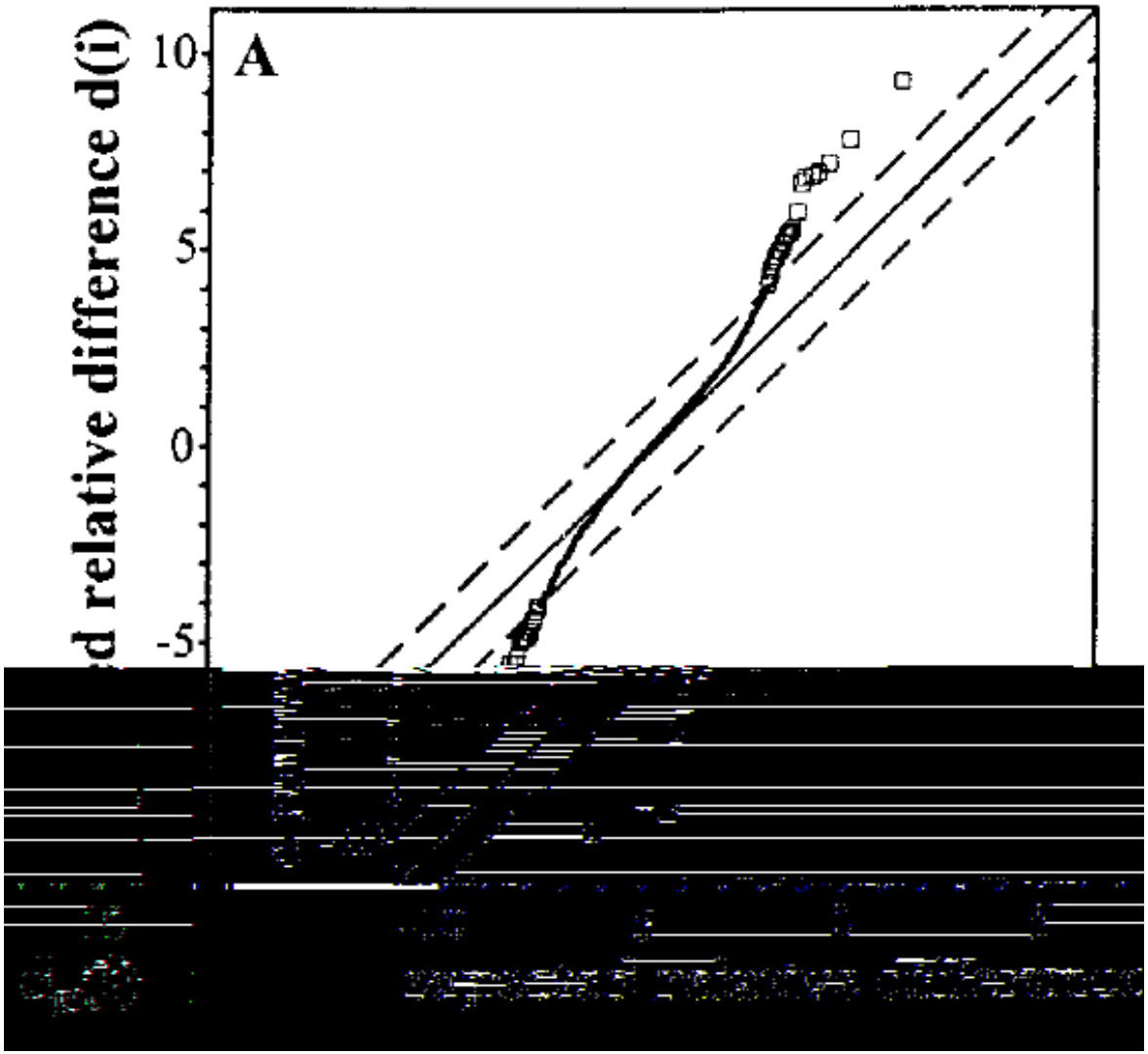
$$d(i) = \frac{\bar{x_I}(i) - \bar{x_U}(i)}{s(i) + s_0}$$

$$s(i) = \sqrt{a \left[\sum_m [x_m(i) - \bar{x_I}(i)]^2 + \sum_n [x_n(i) - \bar{x_U}(i)]^2 \right]}$$

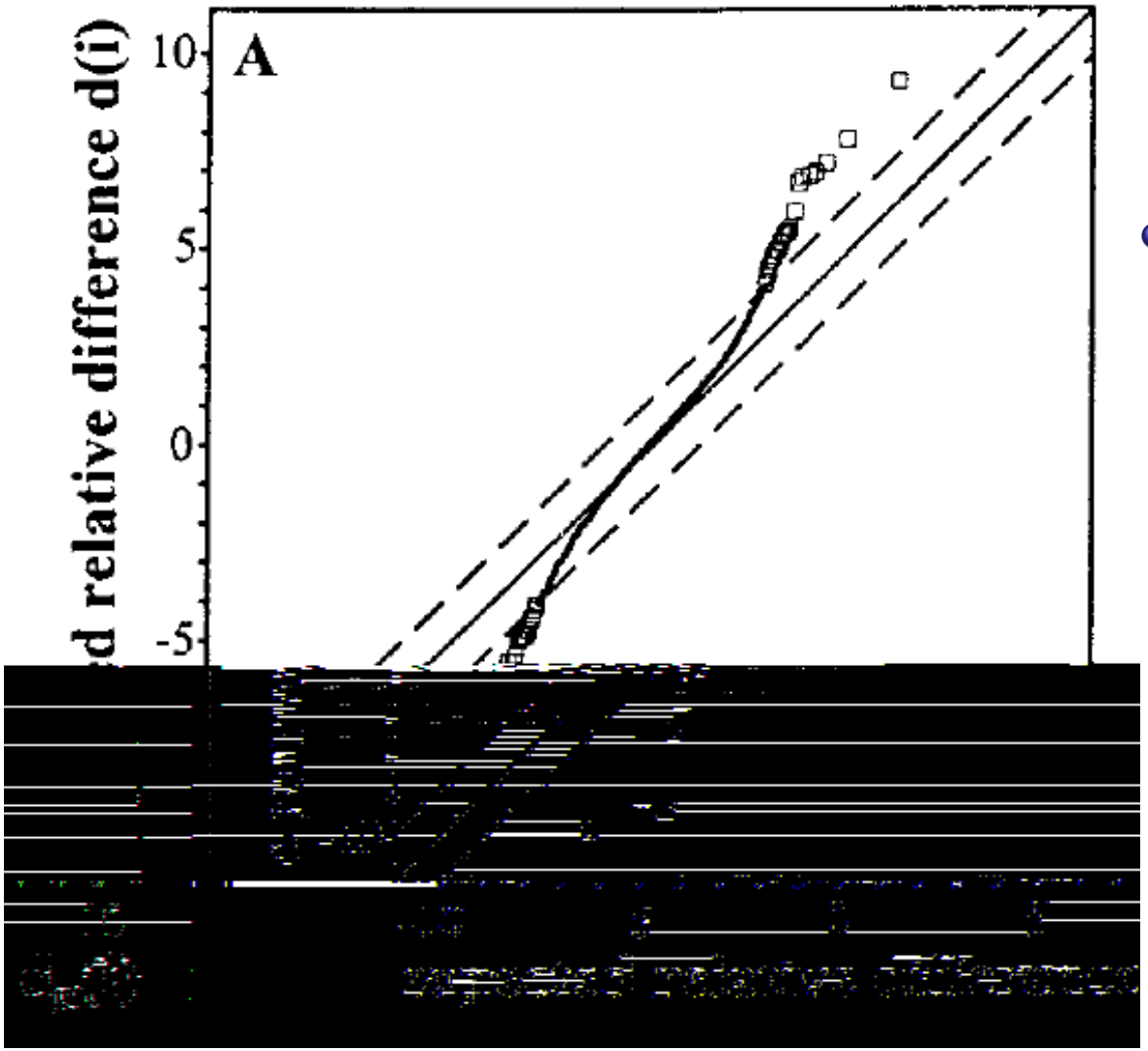
$$a = (1/n_I + 1/n_U)/(n_I + n_U - 2)$$

- 1 vs. 2 = biological replicate
- A vs. B = technical replicate
- I vs. U = treatment
- s_0 forces a minimum variance for the low intensity data

3a: The SAM plot

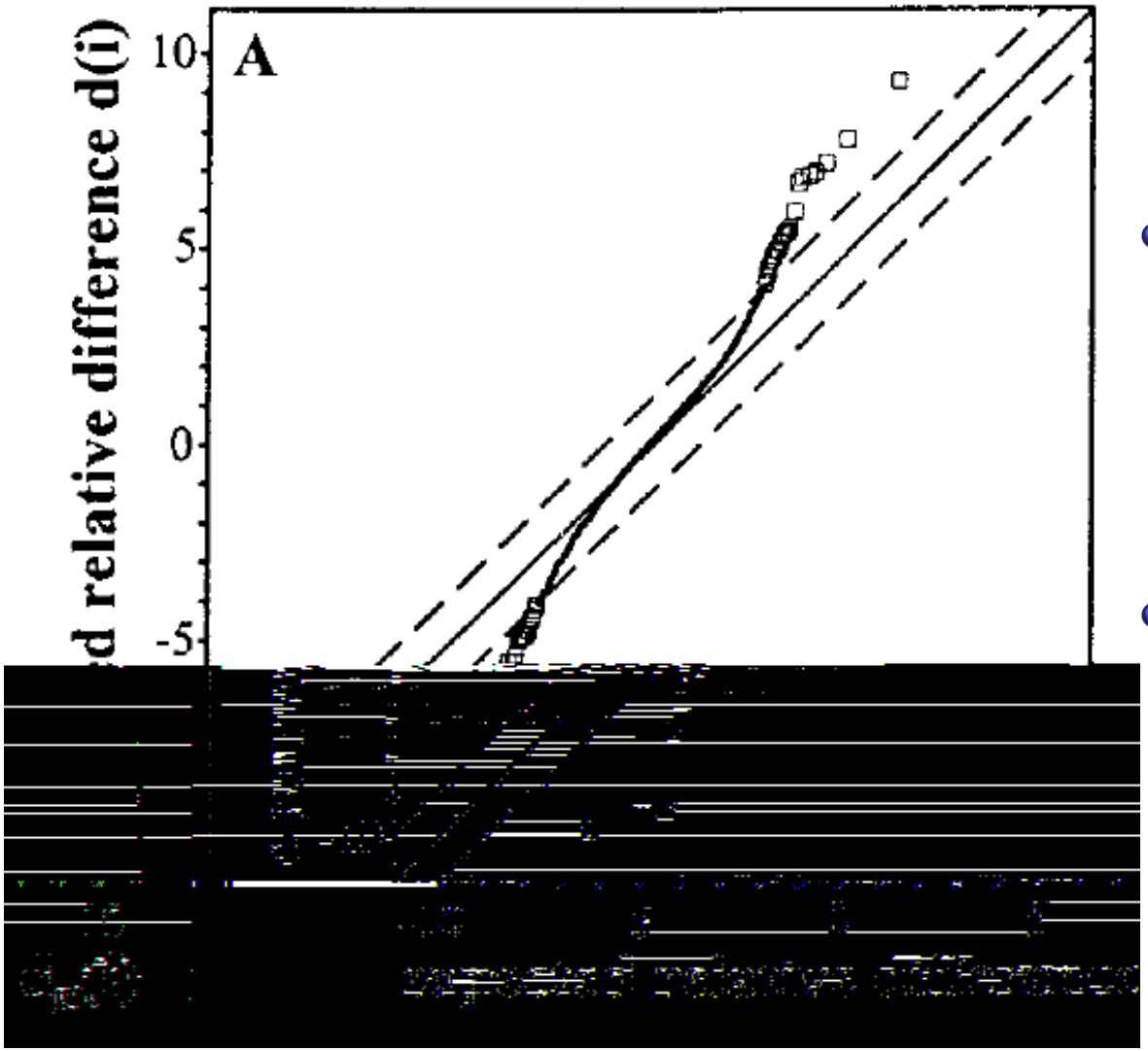


3a: The SAM plot



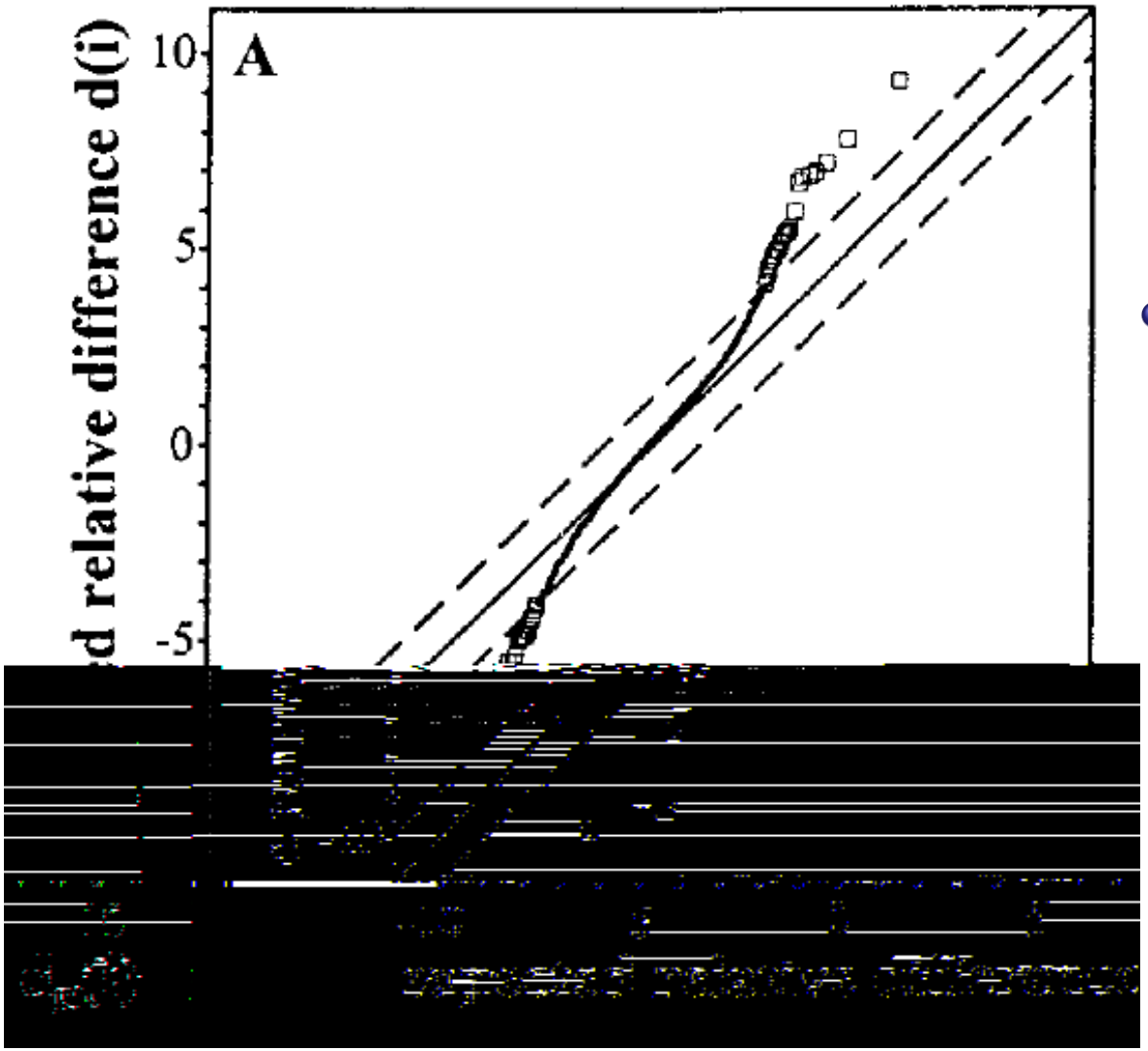
- "Expected" is the average d(i) for all "balanced" permutations of the data.

3a: The SAM plot



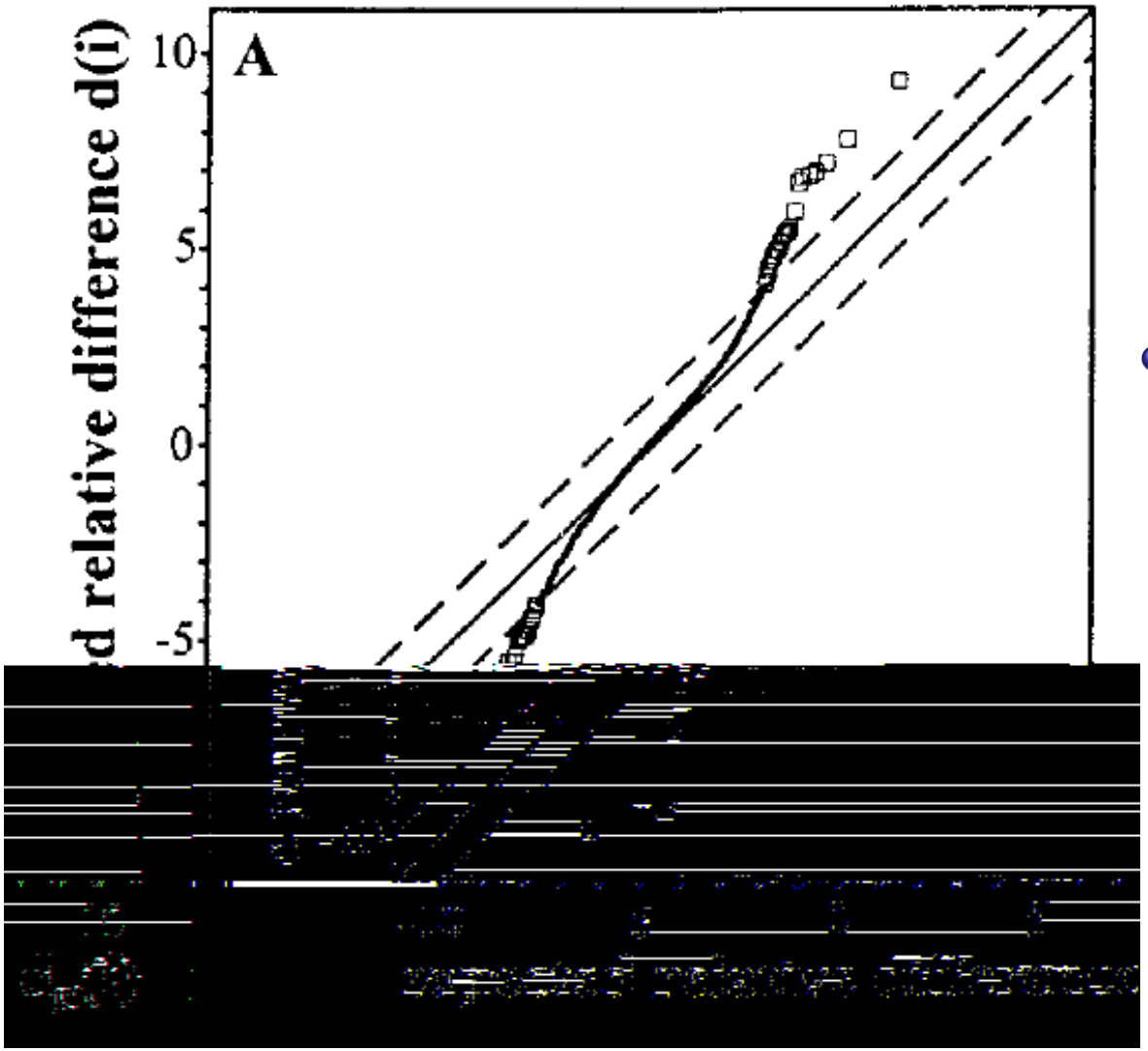
- "Expected" is the average d(i) for all "balanced" permutations of the data.
- "delta" is an offset from the line of best fit, giving two diagonal thresholds.

3a: The SAM plot



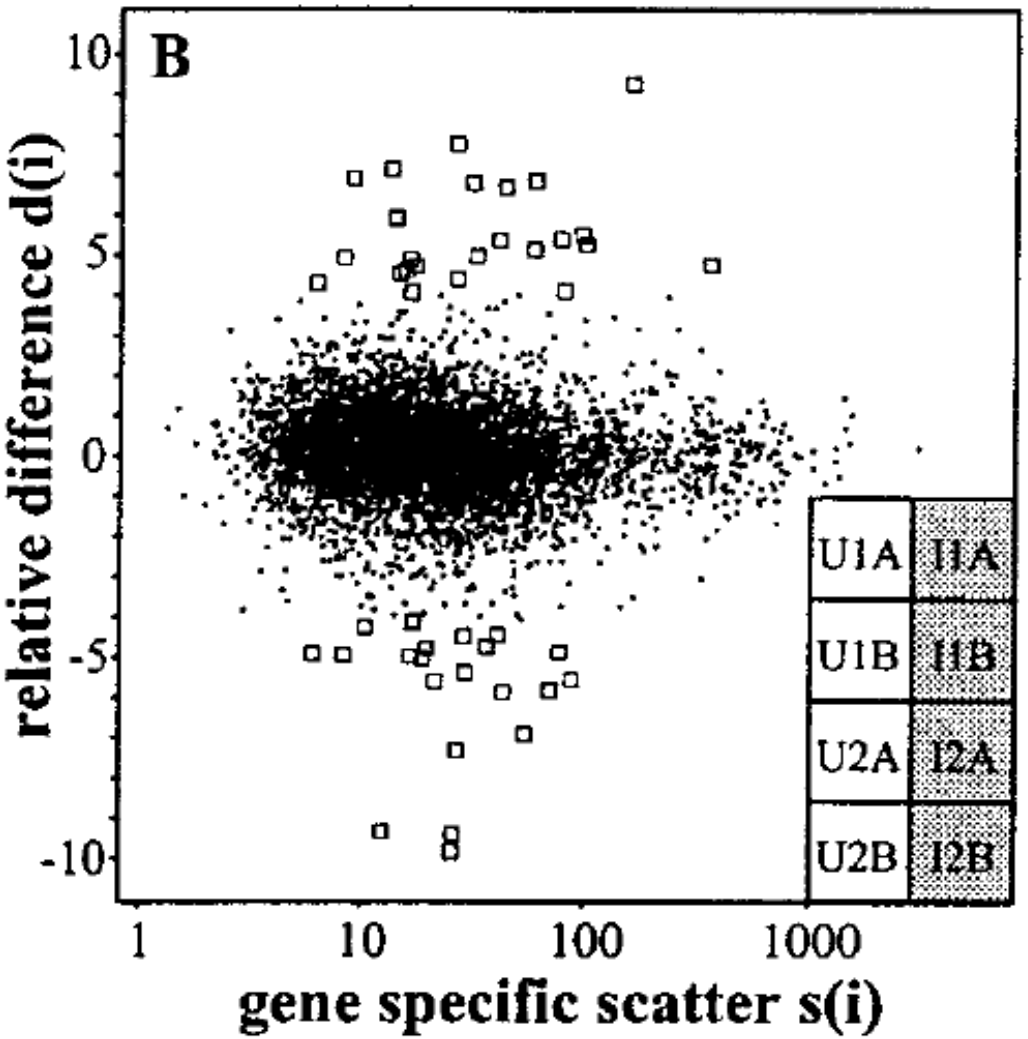
- FDR is calculated by replacing "observed" with each of the balanced permutations in turn (or a random sample for large data sets).

3a: The SAM plot

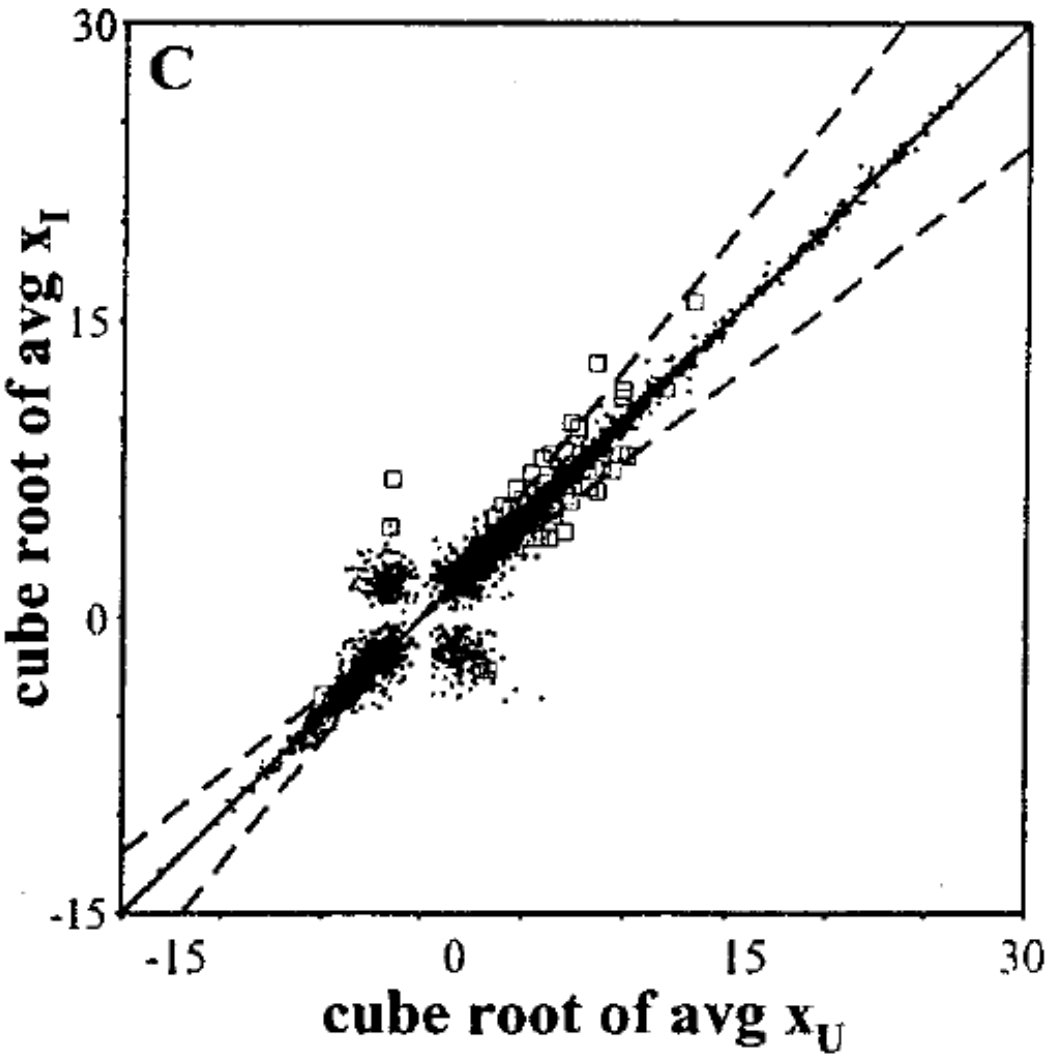


- The plot is monotonic, so diagonal thresholds are also horizontal and vertical thresholds.

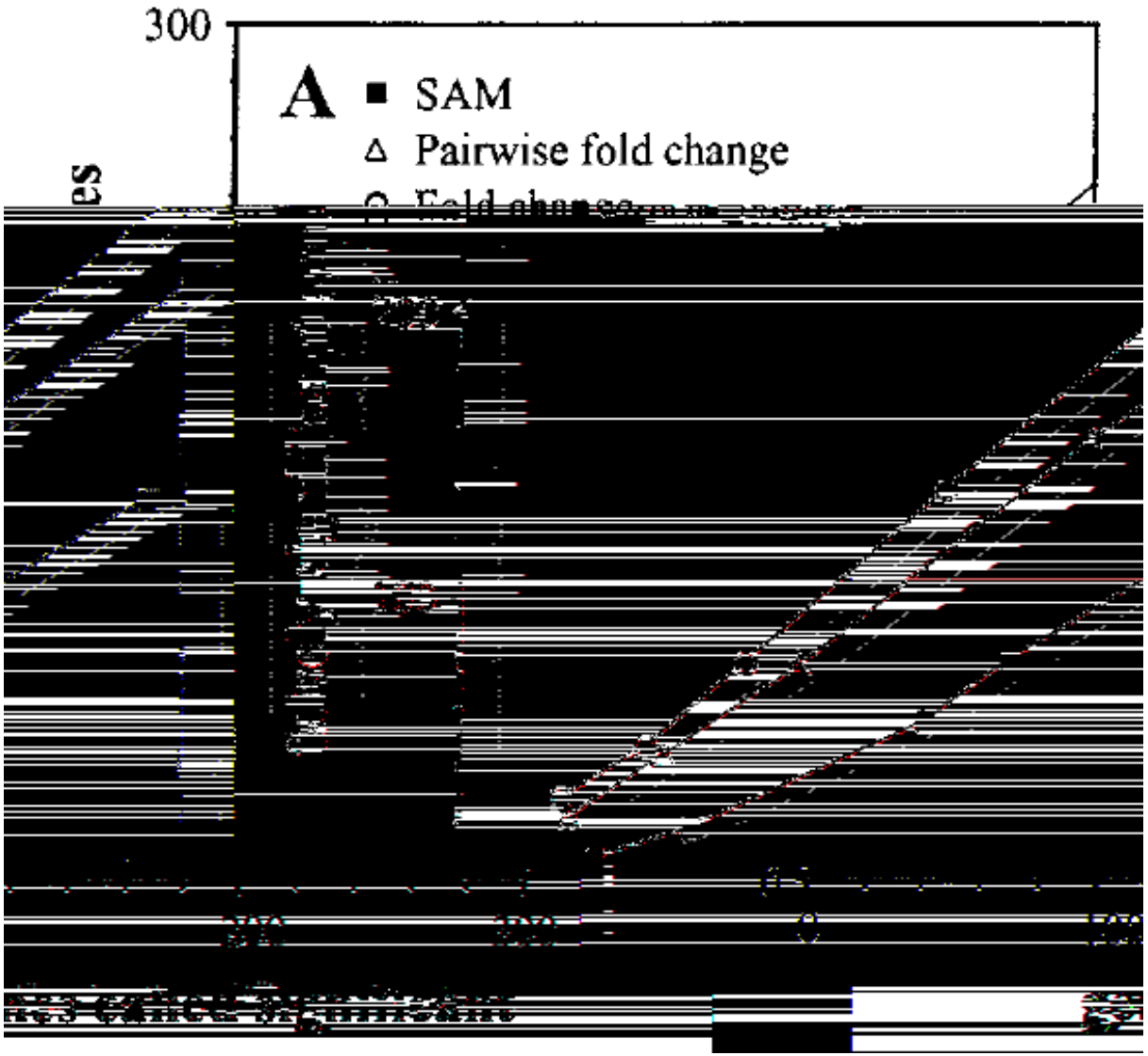
3b: Variance of significant genes



3c: Expression levels of significant genes

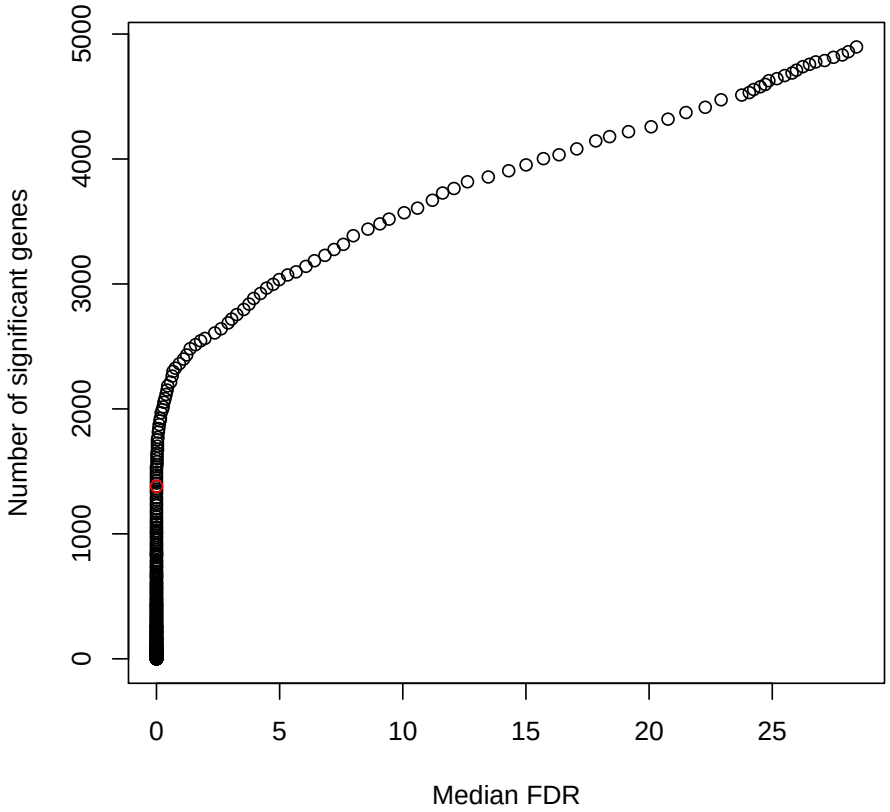


4a: Sensitivity vs. Specificity

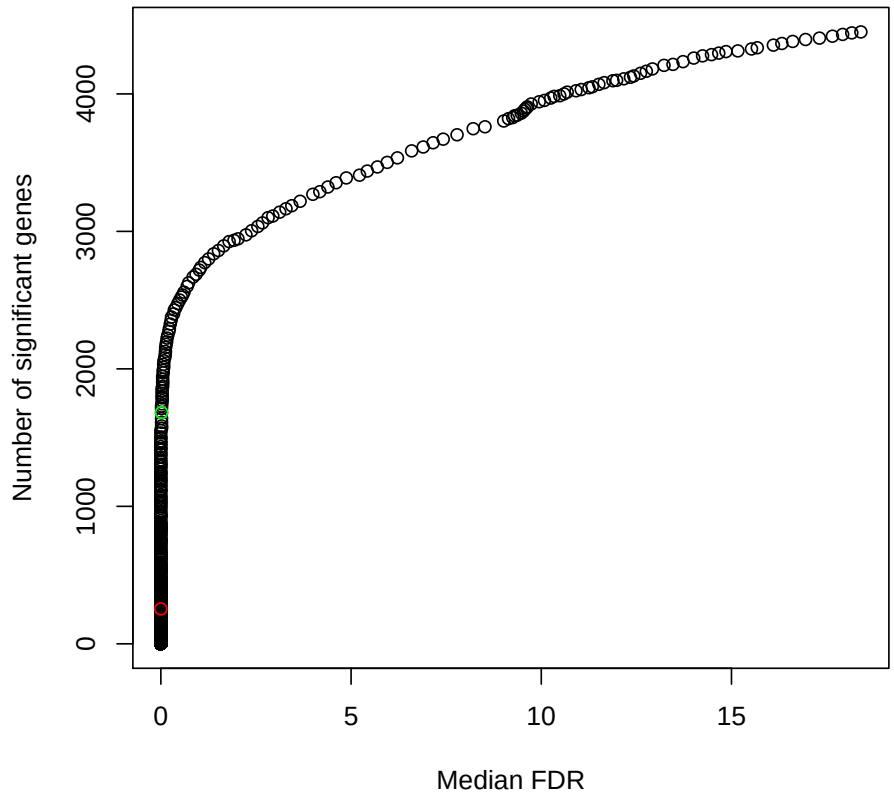


Sensitivity vs. Specificity: Pseudo-ROC plots

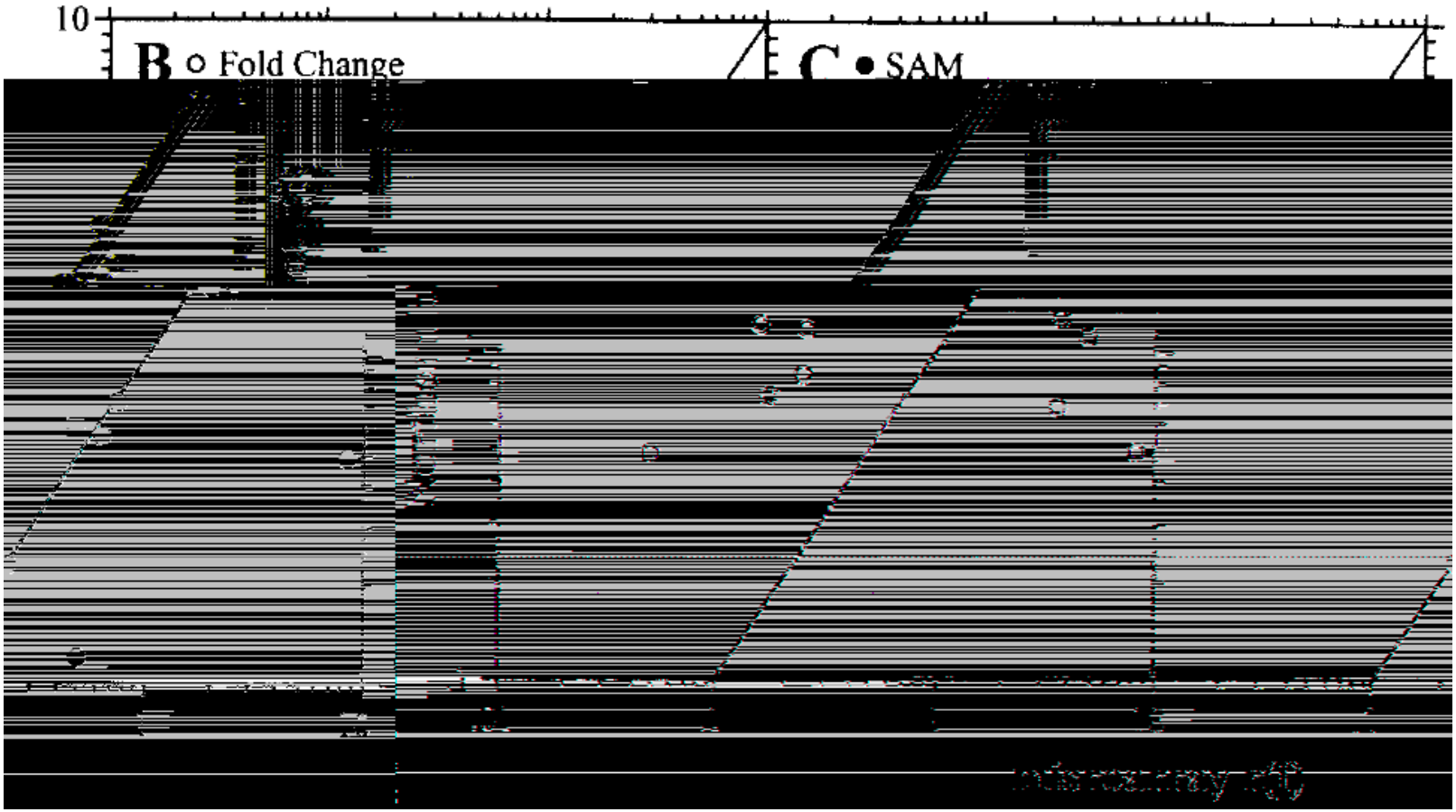
Two-class SAM



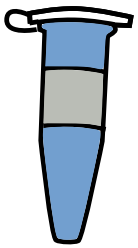
One-class SAM



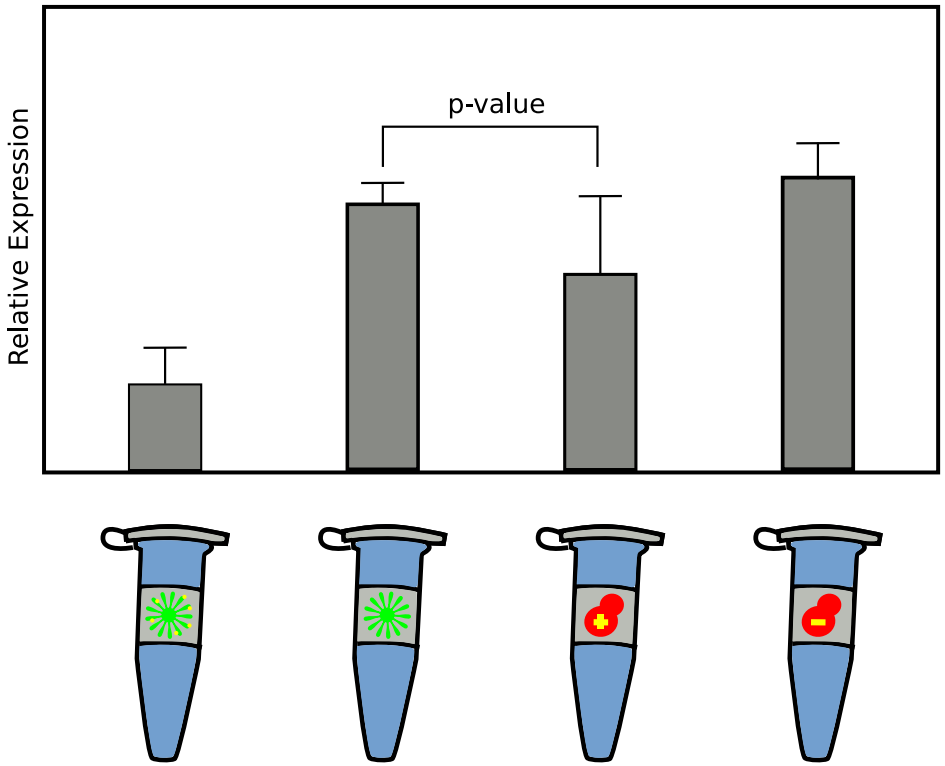
4b and c: Experimental Validation



What does BAGEL do?



What does BAGEL do?



Interactive BAGEL

```
Please type the full name of a text file of microarray ratio results to analyze:

EtOHBAGELDataset.unx
Checking Filename...
Filename OK.

Verifying Input File EtOHBAGELDataset.unx
               Cor1    Cor2    Cor3
Initializing ExpressionNodeNameList...
Assigning names to expression nodes.NewName=Normal.NewName=Normal.NewName=EtOH

Number of Hybs: 3

Press RETURN to verify or q to quit:
Assigning more names to expression nodes...

Please verify that 2 expression nodes are desired,
that all desired nodes are are listed below, and
that each of the following are unique expression nodes.
Normal
EtOH

Press RETURN to verify or q to quit:
Assigning experimental node names to hyb list...
Assigning reference node names to hyb list...
File EtOHBAGELDataset.unx header rows verified.

Current MCMC settings:
(E)rror Model: Additive errors, estimating/constraining Coefficient of Variation terms
(C)onstrained Coefficient of Variation: TRUE
(I)nitial values:
Mu[Normal] := 1.00      Coefficient of Variation[Normal] := 0.2000
Mu[EtOH] := 1.00        Coefficient of Variation[EtOH] := 0.2000
(M)u step size: 0.50
(V)ariance/CV step size: 0.500
(B)urn in, # generations: 20000
```

BMC Bioinformatics 5:54

Resolution of large and small differences in gene expression using models for the Bayesian analysis of gene expression levels and spotted DNA microarrays

Jeff Townsend

Additive Error Models

$$f\left(zij|\mu_i,\sigma_i^2,\mu_j,\sigma_j^2\right)=\frac{\sigma_i^2\mu_j+\sigma_j^2\mu_i z_{ij}}{\sqrt{2\pi}\left(\sigma_i^2+\sigma_j^2 z_{ij}^2\right)^{3/2}}e^{\frac{-\left(\mu_i-\mu_j z_{ij}\right)^2}{2\left(\sigma_i^2+\sigma_j^2 z_{ij}^2\right)}} \quad (1)$$

Additive Error Models

$$f\left(zij|\mu_i,\sigma_i^2,\mu_j,\sigma_j^2\right)=\frac{\sigma_i^2\mu_j+\sigma_j^2\mu_i z_{ij}}{\sqrt{2\pi}\left(\sigma_i^2+\sigma_j^2 z_{ij}^2\right)^{3/2}}e^{\frac{-\left(\mu_i-\mu_j z_{ij}\right)^2}{2\left(\sigma_i^2+\sigma_j^2 z_{ij}^2\right)}} \quad (1)$$

$$f\left(zij|\mu_i,\mu_j,\nu\right)=\frac{\nu^2\mu_i^2\mu_j+\nu^2\mu_j^2\mu_i z_{ij}}{\sqrt{2\pi}\left(\nu^2\mu_i^2+\nu^2\mu_j^2 z_{ij}^2\right)^{3/2}}e^{\frac{-\left(\mu_i-\mu_j z_{ij}\right)^2}{2\left(\nu^2\mu_i^2+\nu^2\mu_j^2 z_{ij}^2\right)}} \quad (2)$$

Multiplicative Error Models

$$f\left(zij|\mu_i,\sigma_i^2,\mu_j,\sigma_j^2\right)=\frac{1}{z_{ij}\sqrt{2\pi\left(\sigma_i^2+\sigma_j^2\right)}}e^{\frac{-\left(\log_e z_{ij}-(\mu_i-\mu_j)\right)^2}{2\left(\sigma_i^2+\sigma_j^2\right)}} \quad (3)$$

Multiplicative Error Models

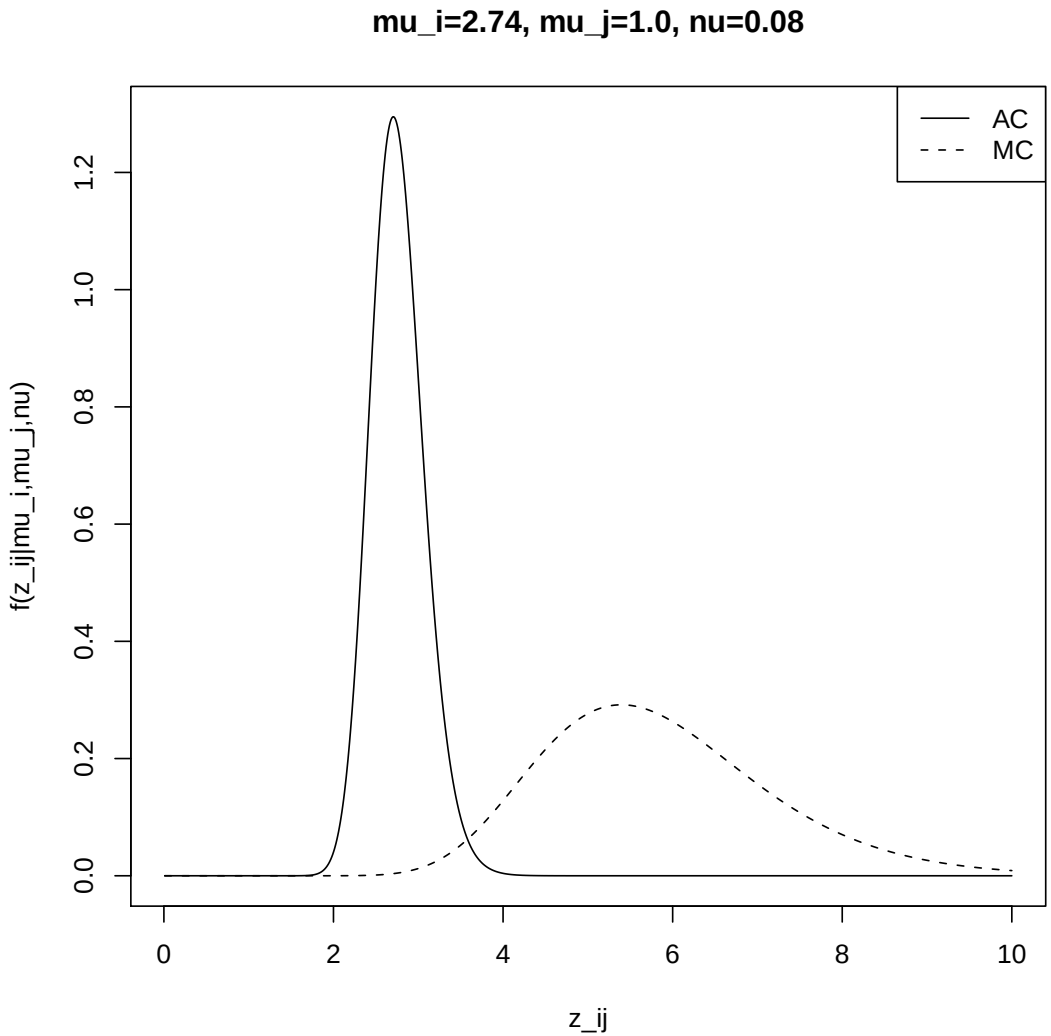
$$f\left(zij|\mu_i,\sigma_i^2,\mu_j,\sigma_j^2\right)=\frac{1}{z_{ij}\sqrt{2\pi\left(\sigma_i^2+\sigma_j^2\right)}}e^{\frac{-\left(\log_e z_{ij}-(\mu_i-\mu_j)\right)^2}{2\left(\sigma_i^2+\sigma_j^2\right)}} \quad (3)$$

$$f\left(zij|\mu_i,\mu_j,\nu\right)=\frac{1}{z_{ij}\sqrt{2\pi\left(\nu^2\mu_i^2+\nu^2\mu_j^2\right)}}e^{\frac{-\left(\log_e z_{ij}-(\mu_i-\mu_j)\right)^2}{2\left(\nu^2\mu_i^2+\nu^2\mu_j^2\right)}} \quad (4)$$

What do the error models look like?

```
# Parameters
mu1 = 2.74
mu2 = 1.0
nu = 0.08

# Domain
x0 <- (1:1000)/100

# Additive error model
ac <- ((nu^2*mu1^2*mu2+nu^2*mu2^2*mu1*x0)/
         (sqrt(2*pi)*(nu^2*mu1^2+nu^2*mu2^2*x0^2)^(3/2)))*
   exp(-(mu1-mu2*x0)^2/2/(nu^2*mu1^2+nu^2*mu2^2*x0^2))

# Multiplicative error model
mc <- 1/(x0*sqrt(2*pi*(nu^2*mu1^2+nu^2*mu2*2)))*
   exp(-(log(x0)-(mu1-mu2))^2/(2*(nu^2*mu1^2+nu^2*mu2^2)))

# Plot
plot(x0,ac,type="l", xlab = "z_ij", ylab = "f(z_ij|mu_i,mu_j,nu)",
      main = "mu_i=2.74, mu_j=1.0, nu=0.08")
points(x0,mc,type="l",lty = 2)
legend("topright", c("AC","MC"), lty = c(1,2))
```

What do the error models look like?



Acceptance Criterion

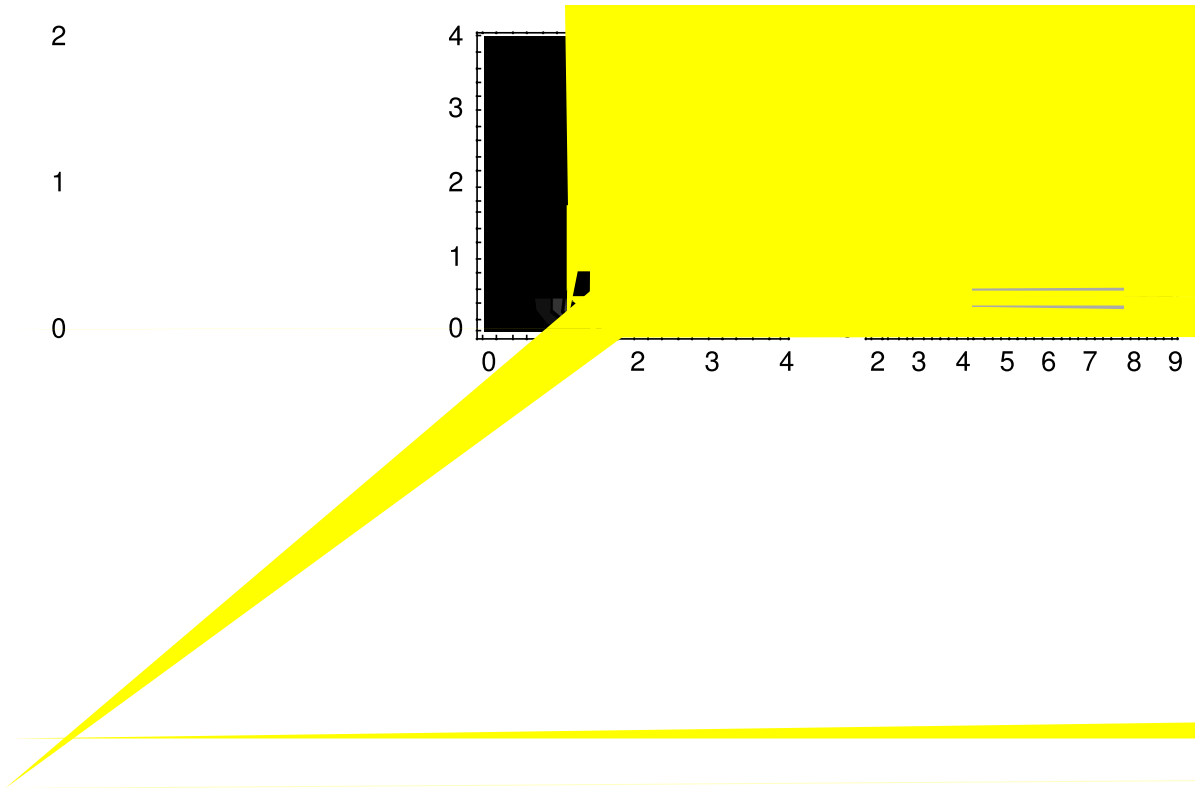
$$h\left(\mu_i, \nu_i, \mu_j, \nu_j | Z\right) = \frac{\left(\prod_{i,j,k}^{n} f\left(z_{ijk} | \mu_i, \nu_i, \mu_j, \nu_j\right)\right) g\left(\mu_i, \nu_i, \mu_j, \nu_j\right)}{\int_{M_l} \int_{V_l} \left(\left(\left(\prod_{i,j,k}^{n} f\left(z_{ijk} | \mu_i, \nu_i, \mu_j, \nu_j\right)\right) g\left(\mu_i, \nu_i, \mu_j, \nu_j\right)\right)\right) d\nu_l d\mu_l} \tag{5}$$

Acceptance Criterion

$$h\left(\mu_i, \nu_i, \mu_j, \nu_j | Z\right) = \frac{\left(\prod_{i,j,k}^{n} f\left(z_{ijk} | \mu_i, \nu_i, \mu_j, \nu_j\right)\right) g\left(\mu_i, \nu_i, \mu_j, \nu_j\right)}{\int_{M_l} \int_{V_l} \left(\left(\left(\prod_{i,j,k}^{n} f\left(z_{ijk} | \mu_i, \nu_i, \mu_j, \nu_j\right)\right) g\left(\mu_i, \nu_i, \mu_j, \nu_j\right)\right)\right) d\nu_l d\mu_l} \tag{5}$$

$$\zeta(0,1) < \frac{\left(\prod_{i,j,k}^{n} f\left(z_{ijk} | \mu_i, \nu_i, \mu_j, \nu_j\right)\right) g\left(\mu_i, \nu_i, \mu_j, \nu_j\right)}{\left(\prod_{i,j,k}^{n} f\left(z_{ijk} | \mu_i', \nu_i', \mu_j', \nu_j'\right)\right) g\left(\mu_i', \nu_i', \mu_j', \nu_j'\right)} \tag{6}$$

Fitting parameters by Markov Chain Monte Carlo (MCMC)



Source: Townsend and Hartl (2002) Genome Biology 3:71

Table 1: Performance on real data

Figure 1: Performance on ratio of truncated Gaussians

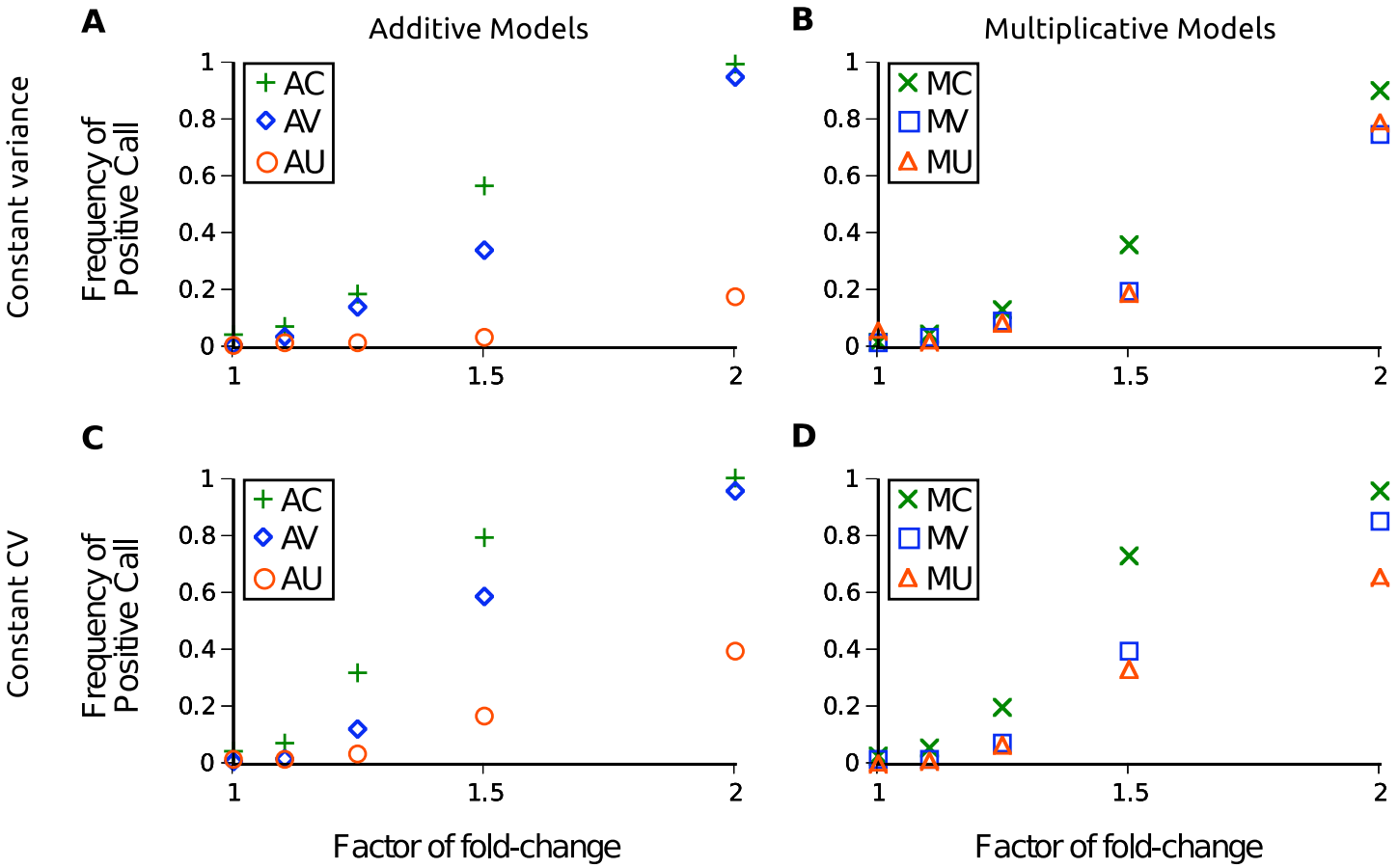


Figure 2: Performance on lognormal distribution

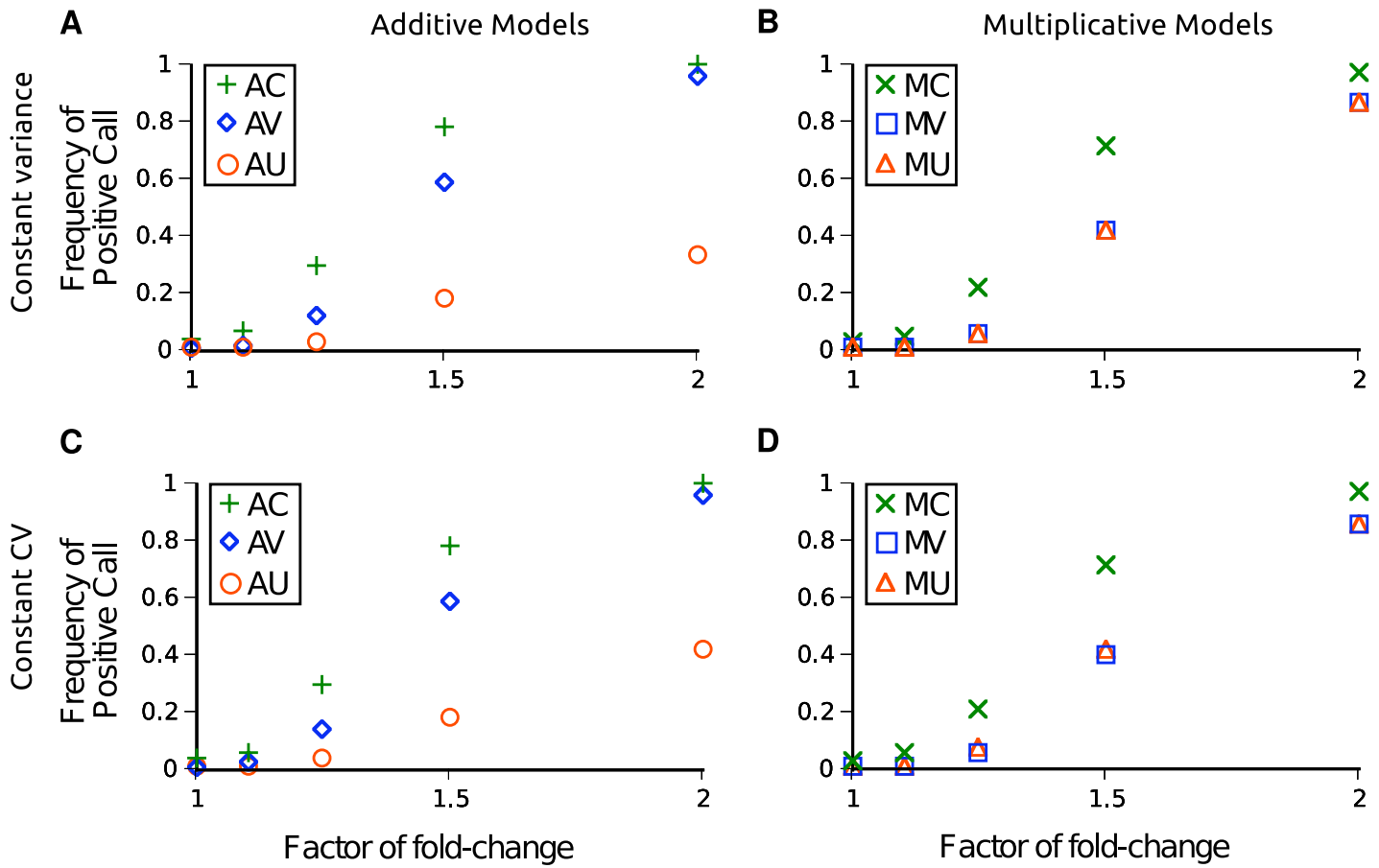


Figure 3: More distributions!

Figure 4: Power calculation for simulated data

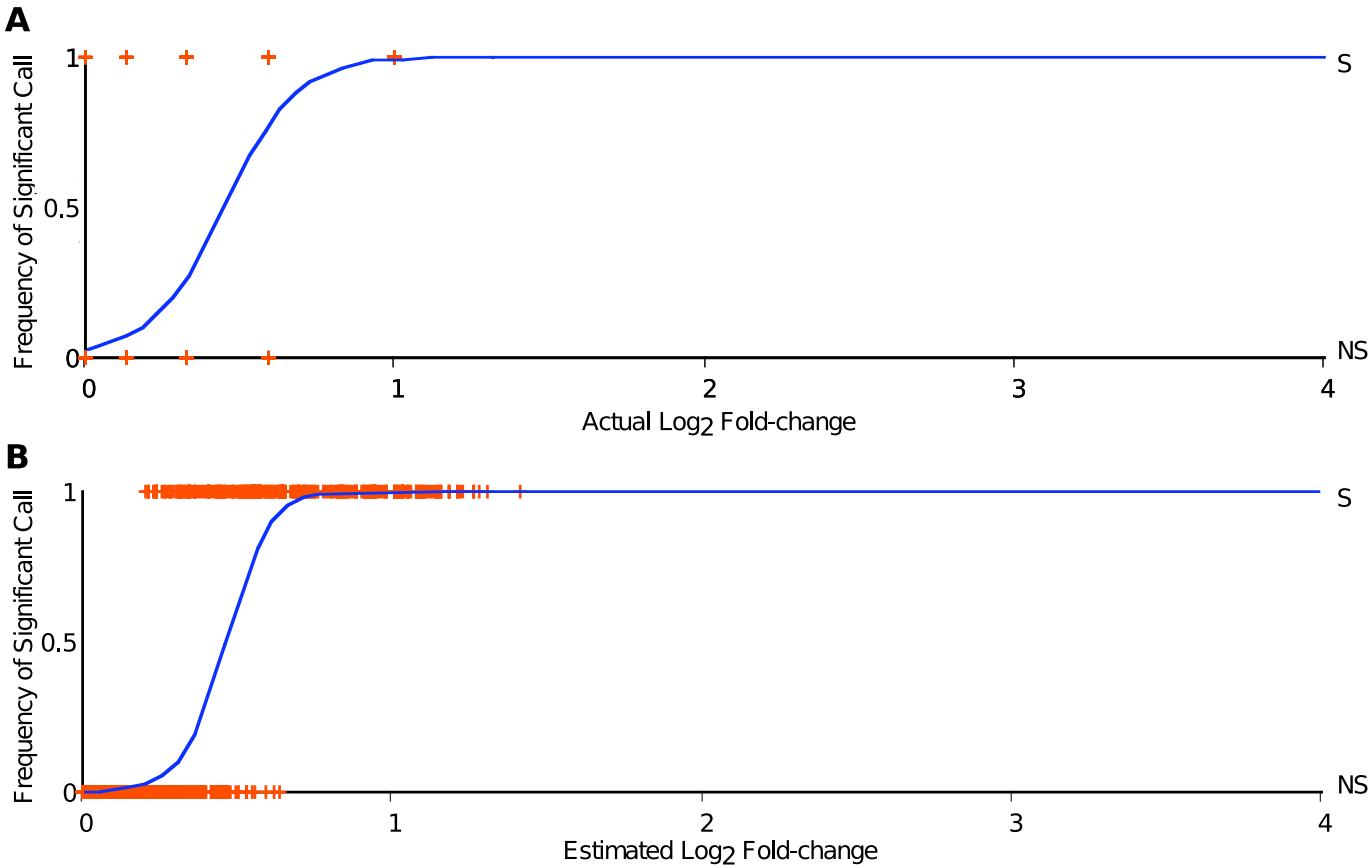
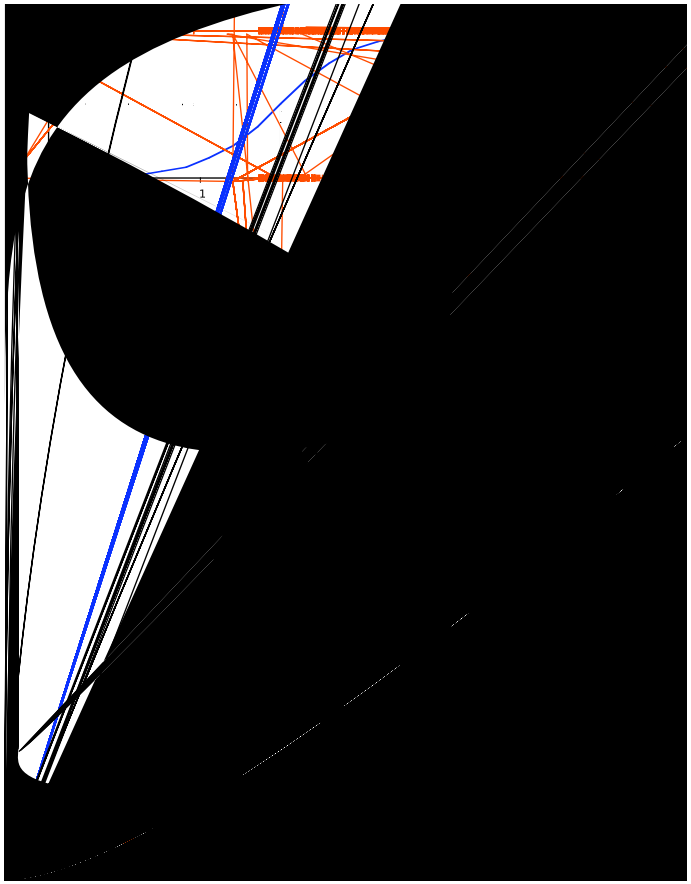
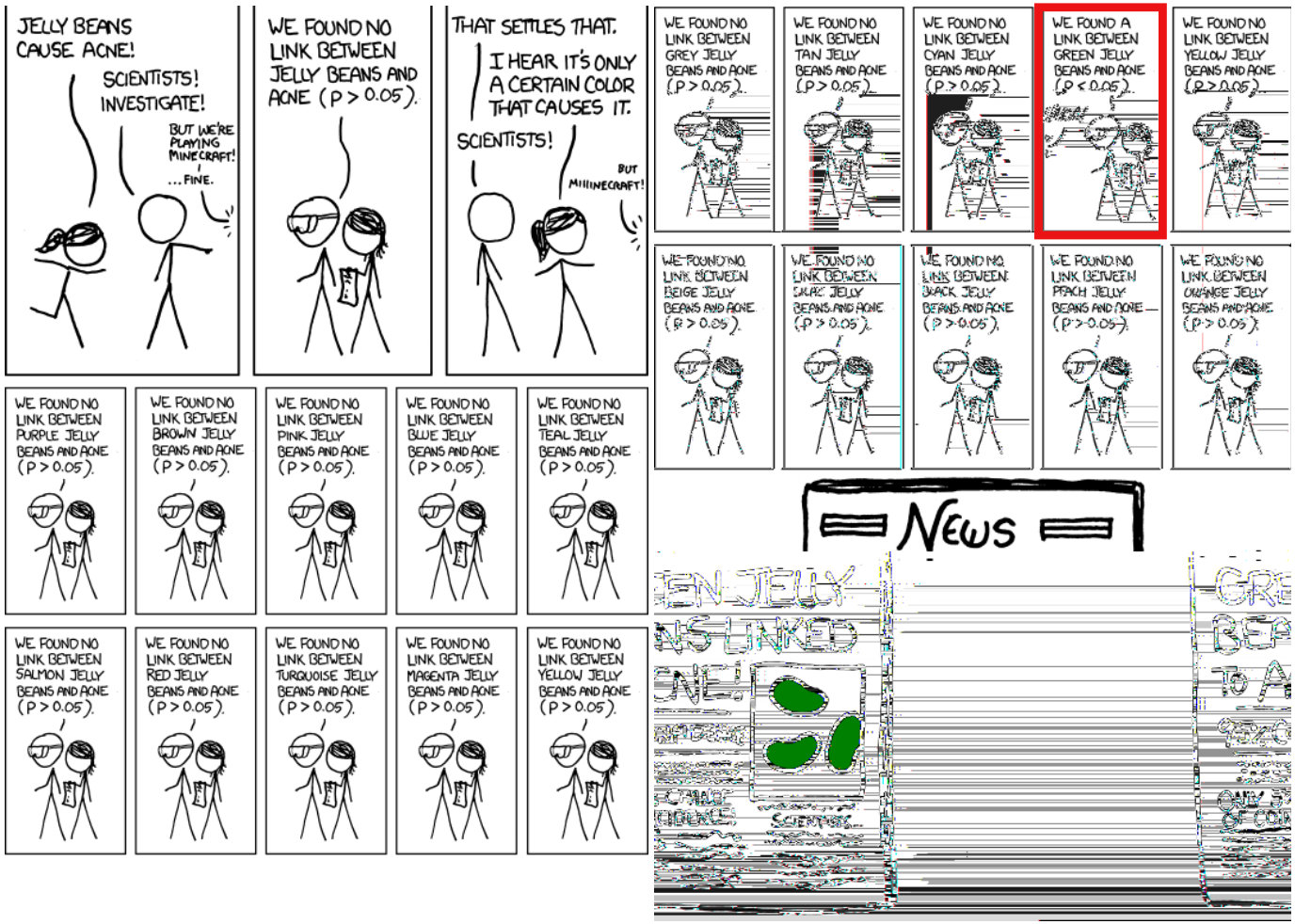


Figure 5: Power calculation for real data



Multiple Hypothesis Testing



http://xkcd.com/882/