# Microarray Data Analysis - II

**FIOCRUZ Bioinformatics Workshop**
**6 June, 2001**

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

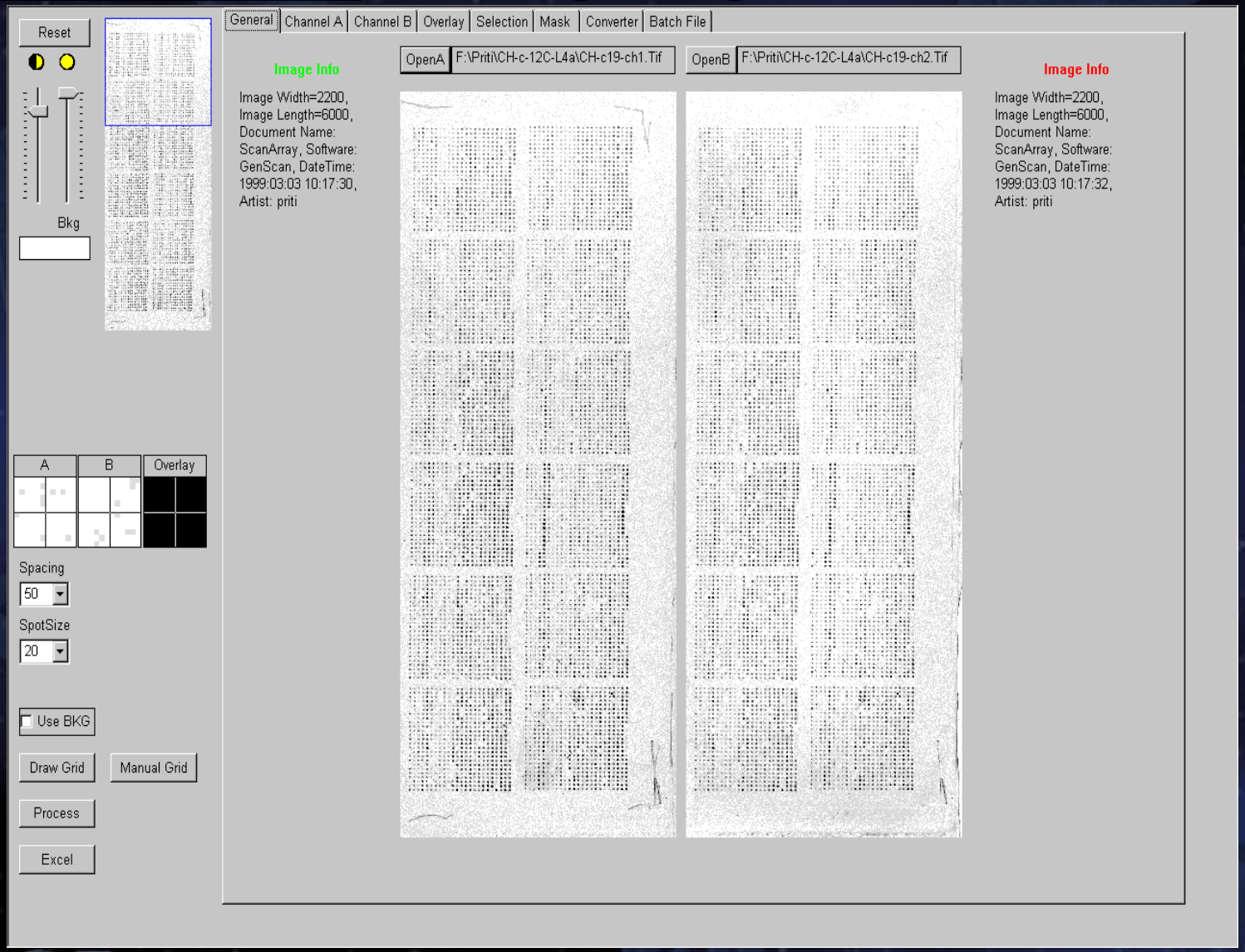
# Challenges in Microarray Data Analysis

- Spot Identification and Quantitation.
- Normalization of data from each experiment.
- Identification of Differentially Expressed Genes
- Identified of genes with correlated patterns of expression.
- Interpretation of data with respect to pathways.
- Literature filtered analysis.

# Image Processing Issues

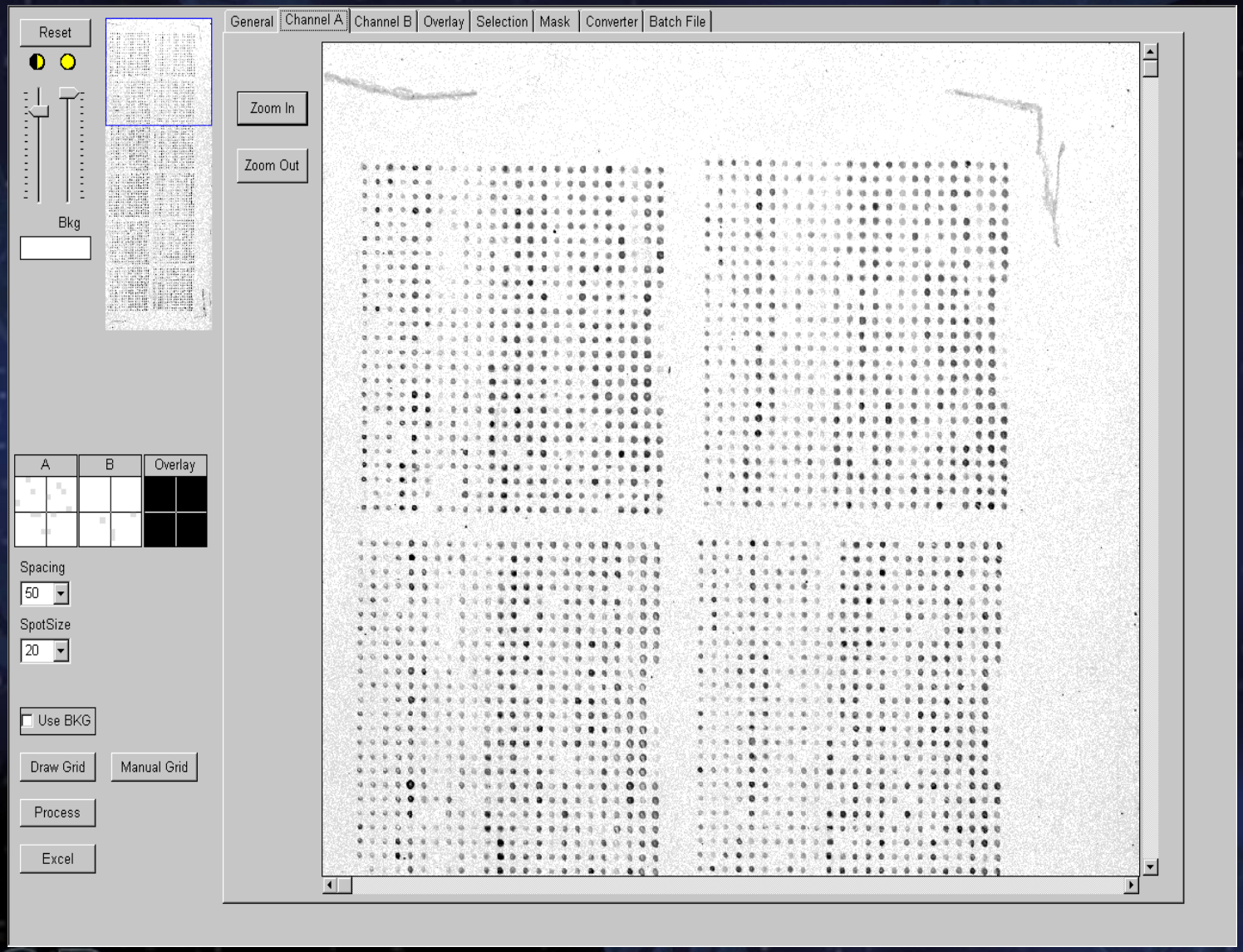
- Spot Finding
- Background Subtraction
- Reproducibility
- Measure - median *vs.* mean (integrated intensity)
- Quality measures

# TIGR Spotfinder Loading Image Data

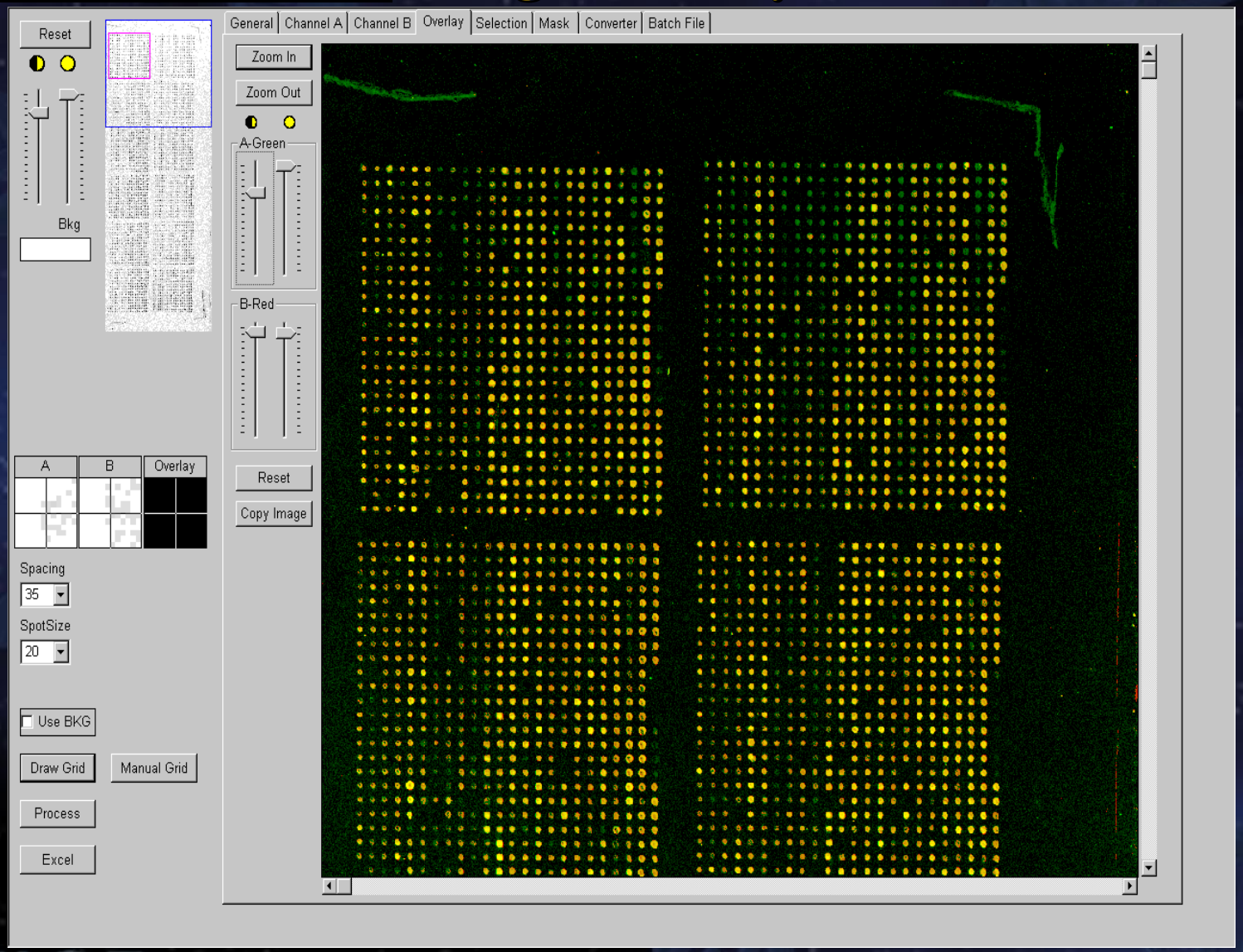



TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

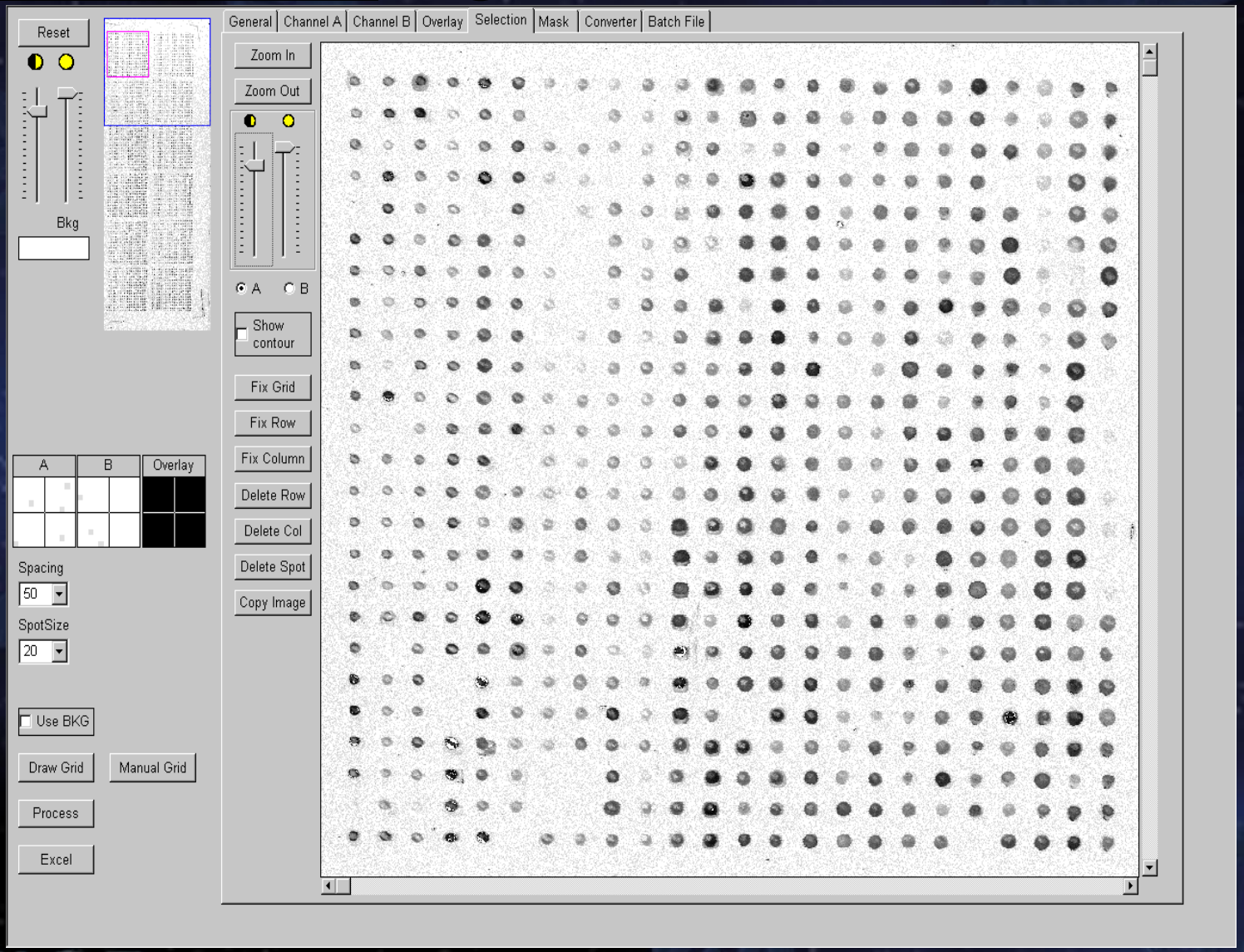
# TIGR Spotfinder Zooming In



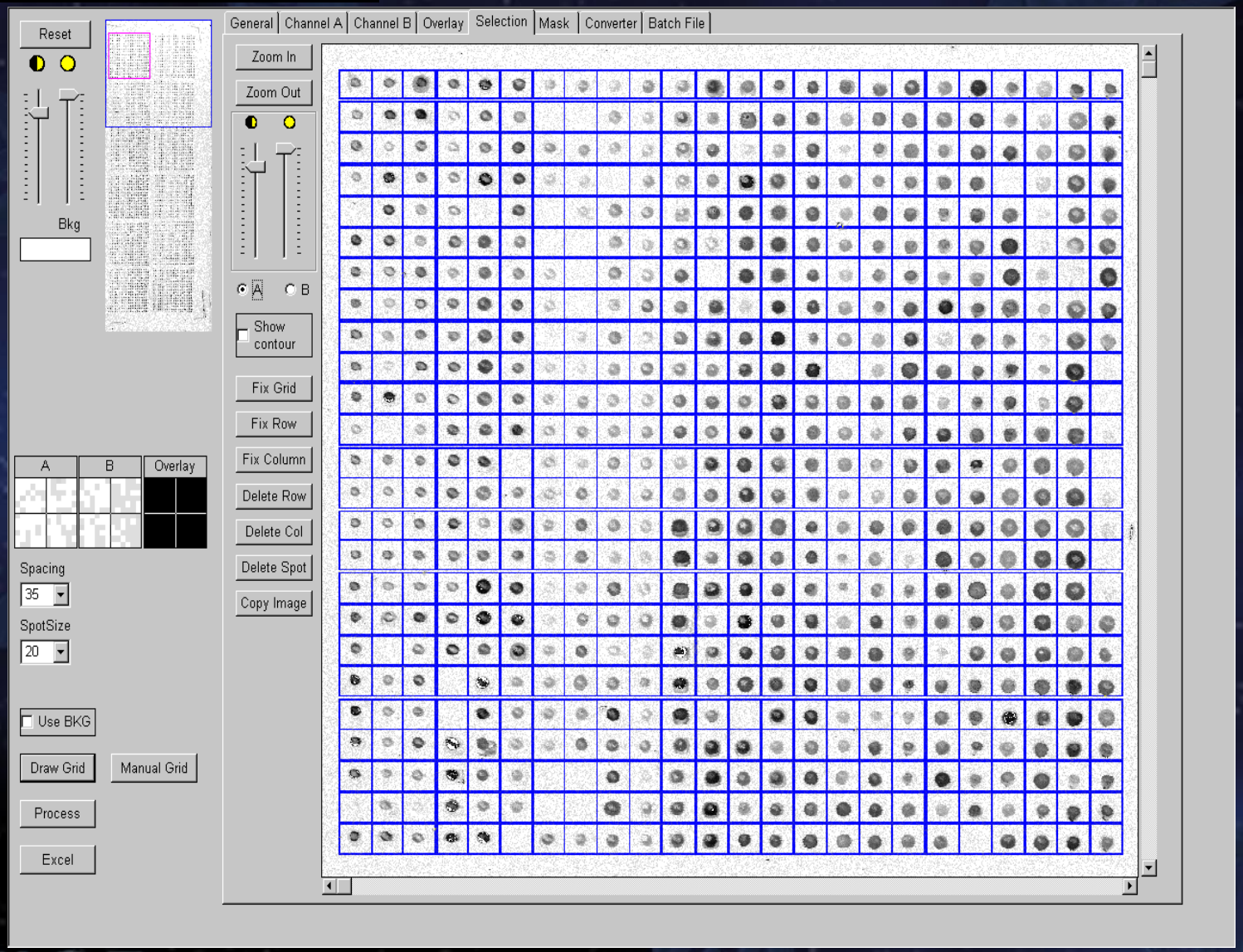
# TIGR Spotfinder Image Overlay



# TIGR Spotfinder Region Selection

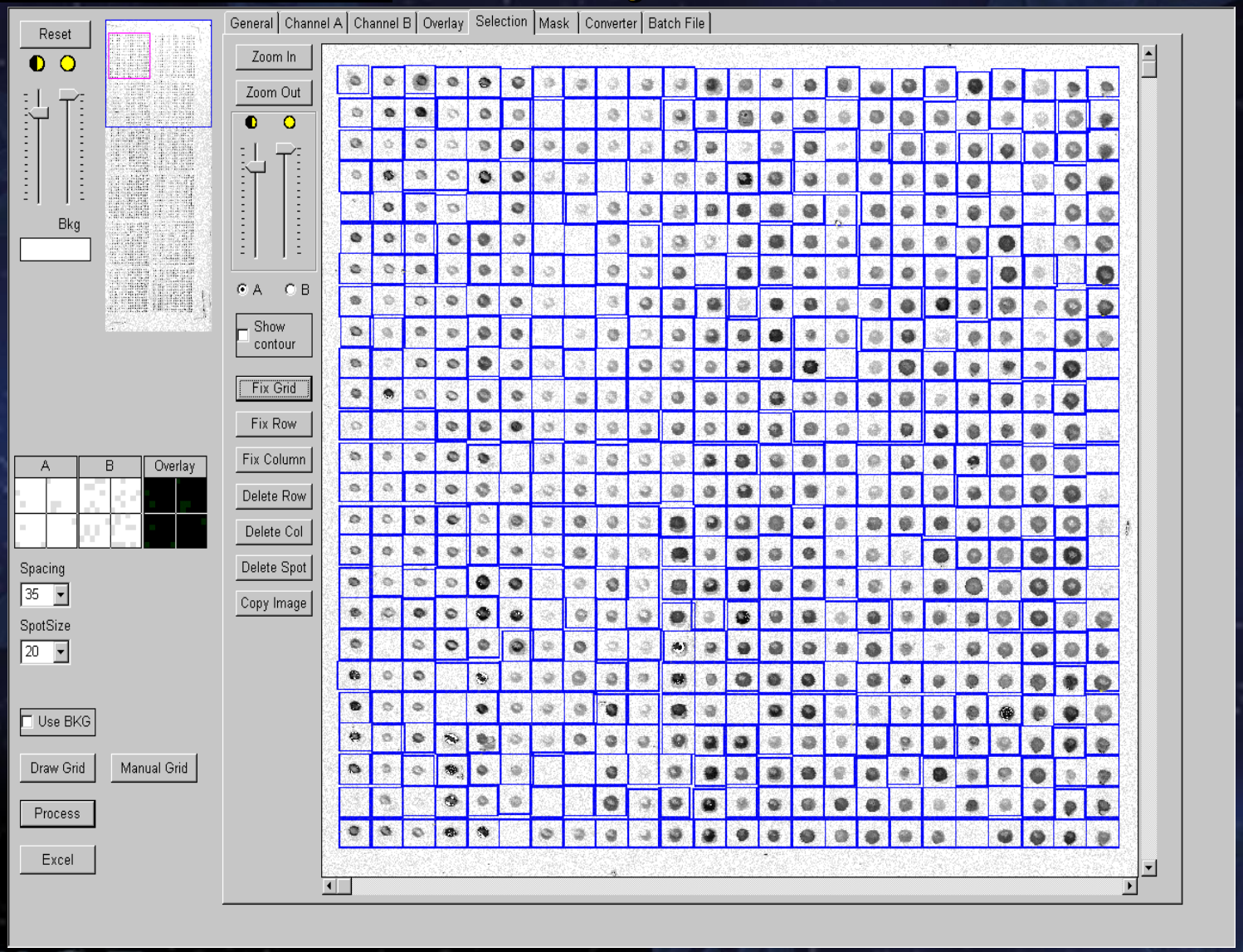


# TIGR Spotfinder Grid Determination

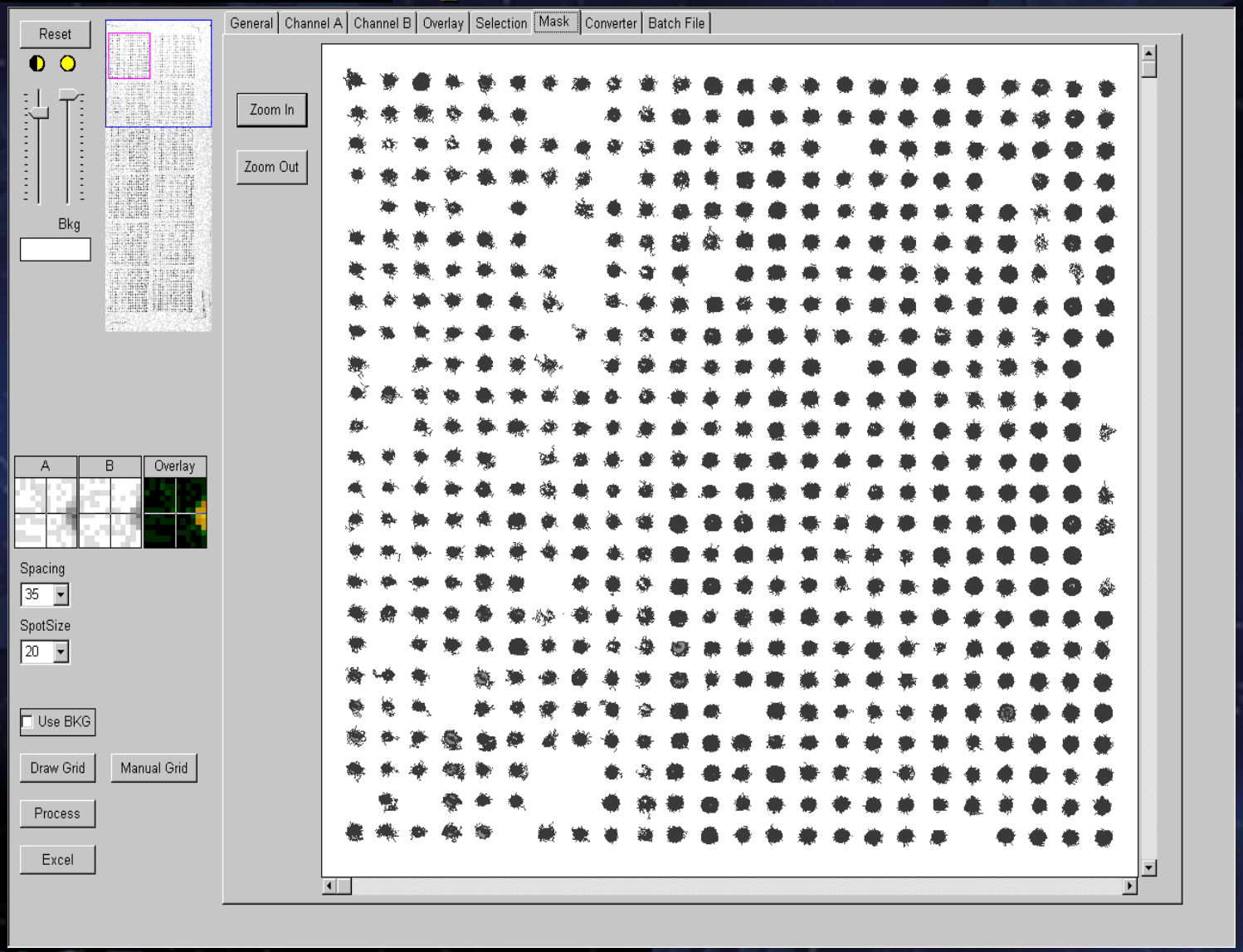



TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

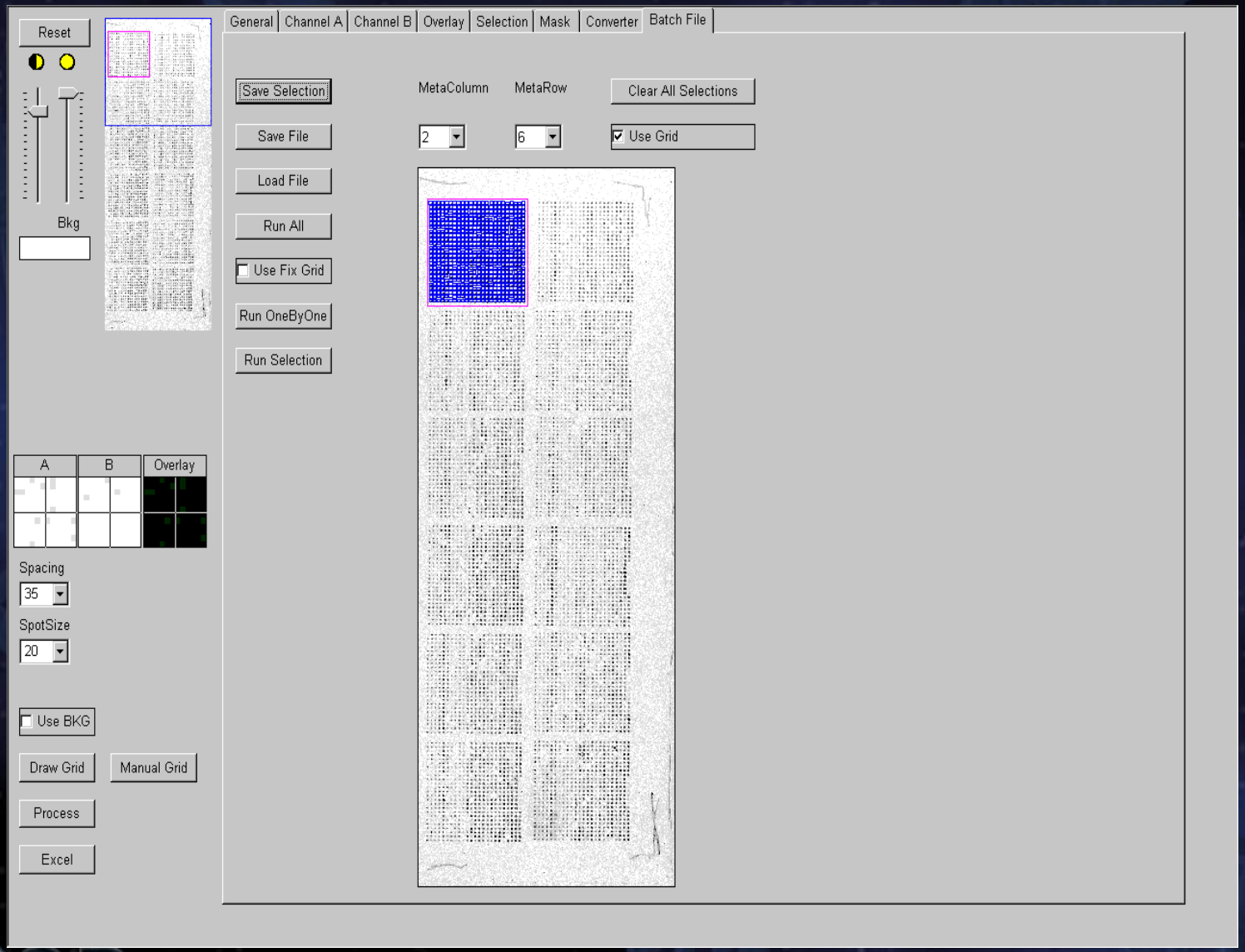
# TIGR Spotfinder Grid Adjustment



# TIGR Spotfinder Spot Determination



# TIGR Spotfinder Batch Mode




TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

# TIGR Spotfinder Data Ouptut to Spreadsheet

Microsoft Excel - Book1

File Edit View Insert Format Tools Data Window Help

Arial 8 100%

A1 = 0.658366180733696

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.6583662 | 0.3200546 | 0 | 1.4709515 | 5.5659516 | 0.8221982 | 0 | 1.2417782 | 15.806612 | 0.6525556 | 2.2853259 | 1.1422822 | 1.5317368 | 0.5656727 | 0.2204492 | 0.6117383 | 0.3719255 |
| 2 | 0.7558224 | 0.4726607 | 0.6765916 | 2.1505848 | 2.4587546 | 0 | 16.717262 | 0 | 0.545069 | 0.2482868 | 0.6166225 | 0.6360926 | 2.7529202 | 0.4951814 | 0.7414343 | 0.7803267 | 0.3516873 |
| 3 | 0.6014479 | 0.6786517 | 0.4542122 | 0.8245431 | 1944 | 0.9568038 | 0 | 1.162365 | 0 | 0.4939302 | 0.5924173 | 0.4673734 | 0.7887445 | 0.6112968 | 0.6759708 | 0.3229271 | 0.4144771 |
| 4 | 0.7138603 | 0.627963 | 0.6439672 | 0.7704063 | 0.8925518 | 0 | 0.4864845 | 0 | 1.5022185 | 1.4716133 | 0.6375733 | 1.8214447 | 1.4876398 | 0.5303082 | 0.9627375 | 0.5498852 | 0.4752789 |
| 5 | 19867 | 0.5250862 | 0.5794229 | 4.6387927 | 1.6258561 | 0.6763154 | 0.6856697 | 0.6019967 | 0.9303186 | 0.7640353 | 4732 | 0.1956629 | 0.6116118 | 0.7179828 | 0.511428 | 0.5777325 | 0.3184048 |
| 6 | 0.5597972 | 0.573433 | 0.5087411 | 1.0288527 | 0.8213453 | 1.088831 | 0.5182413 | 7.3380015 | 0.5439334 | 0.7649303 | 0.6407183 | 0.8737587 | 2.5823746 | 3.0950537 | 1.0833581 | 0.5501509 | 1.4692662 |
| 7 | 0.8181596 | 5.0143941 | 0.4104919 | 0.8590075 | 2.4741482 | 0.511308 | 0.2833126 | 0.7497842 | 0.5266239 | 0.5785949 | 0.5328029 | 0.5844418 | 1.3443897 | 0.6685372 | 0.7394423 | 0.4709574 | 0.3382549 |
| 8 | 0.4960111 | 0.9603443 | 0.6195294 | 0.8830633 | 1.8163842 | 0.7538587 | 3.0679385 | 0.888154 | 0.8022219 | 0.5235721 | 0.4453822 | 0.6102882 | 4.3070269 | 0.5359387 | 1.1485683 | 0.5624667 | 0.5095583 |
| 9 | 1.1002291 | 0.331194 | 0.9343573 | 0.4383423 | 0.5945796 | 2.1851852 | 0 | 1.5584225 | 0.6857057 | 0.3514542 | 0.9283824 | 1.5142216 | 0.5062739 | 2.3805146 | 0.6122436 | 0.4766529 | 1.1163227 |
| 10 | 0.667834 | 0.6013884 | 0.9775251 | 0.5791472 | 0.6932317 | 4.3197004 | 2.2649694 | 0.8664637 | 0.4157382 | 13935 | 2.2690295 | 0.7714797 | 0.4356746 | 1.8112027 | 0.4992847 | 0.3741092 | 0.5494838 |
| 11 | 0.4865241 | 0.6553847 | 0.628634 | 0.3819333 | 0 | 1.0413585 | 2.6687442 | 0.4432519 | 0.410186 | 0.2044601 | 0.7793285 | 18592 | 0.3790219 | 0.805083 | 0.5479772 | 1.7025396 | 0.6865773 |
| 12 | 0.377371 | 0.7351658 | 0.6047932 | 0.6122858 | 27360 | 22.128571 | 0 | 0.6999262 | 1.1087549 | 0.7171094 | 1.023534 | 28160 | 19872 | 41403 | 19633 | 0.71932 | 0.8772932 |
| 13 | 0.3941197 | 0.4684356 | 0.549851 | 0.3378463 | 0.5800944 | 0.4016139 | 0.4345092 | 0.8373468 | 0.5058177 | 0.193656 | 0.6716211 | 0.5191203 | 0.3814872 | 0.4716132 | 0.3442173 | 0.6011677 | 0.7315219 |
| 14 | 0.6270496 | 0.6417714 | 0.6545356 | 0.5588574 | 0.6905183 | 0.9231741 | 0.9723208 | 0.6572631 | 0.5128263 | 0.2494581 | 0 | 0.6706132 | 0.5712125 | 43.367568 | 0.7071647 | 0 | 0.602618 |
| 15 | 0.9472301 | 0.4330814 | 0.5178506 | 0.4985484 | 0.6018318 | 0.5159809 | 0.64819 | 0.3768859 | 0.5054452 | 1.0501037 | 0.6222883 | 0.5850231 | 0.6007415 | 0.9281552 | 0.7076223 | 1.6884949 | 0.9790521 |
| 16 | 0.6911812 | 0.4346258 | 1.1916842 | 0.4309419 | 8.9242692 | 43848 | 0 | 96.033457 | 0.6009259 | 0.7480124 | 0.6310674 | 32015 | 0.6647776 | 0.6107295 | 0.4037619 | 0.4548992 | 0.5031051 |
| 17 | 1.9877921 | 0.8708843 | 0.5843845 | 0.4244362 | 10239 | 0.630209 | 0.8887241 | 0 | 0.5655554 | 0.924516 | 1.1751844 | 0.9742245 | 0.5055666 | 0.675697 | 0.6563043 | 0.6590467 | 0.7881861 |
| 18 | 0.8784821 | 0.762783 | 0.6200871 | 0.4537044 | 0.8391829 | 0.4867697 | 0.4590029 | 15229 | 3.3292708 | 0.6948884 | 0.6198393 | 0.951506 | 0.9830806 | 0.5503178 | 0.297055 | 0.5330157 | 0.6457372 |
| 19 | 0.7443983 | 0.3993024 | 0.2824506 | 9.7648983 | 1.6374832 | 0.4936004 | 4.3733205 | 5.9774657 | 3.660936 | 0.9652109 | 0.7294435 | 1.0556538 | 0.2780486 | 0.5547685 | 0.4831949 | 0.6634127 | 0.8785753 |
| 20 | 0.6150365 | 0.7199398 | 0.3694191 | 3.0180843 | 17.80643 | 0.6281572 | 45.43388 | 2.4153739 | 20010 | 2.0190544 | 0.7906932 | 0.5755896 | 23672 | 21.117735 | 47.738657 | 3.7311394 | 0.561419 |
| 21 | 0.3831165 | 0 | 0.7697642 | 0 | 0 | 0.9106237 | 0.559151 | 0.4493976 | 0.5317561 | 1.9017256 | 0.454886 | 0.6134742 | 0.4145606 | 0.4072714 | 0.6959529 | 0.5330013 | 0.5375891 |
| 22 | 0.7813621 | 0.9994261 | 0.6439033 | 0.4590316 | 3.3381008 | 0.4194966 | 0.661315 | 0.6708735 | 1.2265628 | 6.7601277 | 0.8665072 | 0.4766182 | 0.6210309 | 0.2603647 | 0.4316437 | 0.6079227 | 0.6760341 |
| 23 | 0.9928715 | 0.8081178 | 0.9006058 | 3.9962708 | 0.5796124 | 0.326307 | 0.5219831 | 1.2754642 | 13366 | 0.7130079 | 0.6032611 | 0.4392349 | 0.2545188 | 1.0666558 | 0.4684213 | 0.7079562 | 0.4384588 |
| 24 | 0.6423902 | 0.5016539 | 0.6781809 | 0 | 0 | 0.4935037 | 0.3806448 | 0.4131515 | 0 | 1.7617517 | 0.608026 | 0.5602693 | 1.2712959 | 0.3531592 | 0.8132697 | 0.5016375 | 0.5417107 |
| 25 | 2.7186579 | 0.5461783 | 4.7906256 | 1.3970144 | 1.2894225 | 33492 | 0.5698123 | 0.6633413 | 45537 | 7.7535278 | 0.4481022 | 0.8431524 | 0.5930069 | 0.5466802 | 0.9798421 | 0.5280782 | 0.6872405 |
| 26 | 0.4448915 | 0.2928477 | 1.138561 | 0.872978 | 2.9418556 | 3.2201626 | 0.4299104 | 0.7401343 | 0.8747521 | 0.6191381 | 0.5955113 | 0.7019722 | 0.7324154 | 0.3111703 | 0.6549464 | 0 | 0.5135729 |
| 27 | 0.8660603 | 0.4925402 | 0.5055647 | 18821 | 0 | 0 | 0.4171291 | 0.4299915 | 0.5592586 | 0.367861 | 0.5143132 | 0.95747 | 0.6374241 | 0.5562525 | 0 | 0.6587389 | 0.6611734 |
| 28 | 0.6870497 | 0.5846247 | 0.5644515 | 5.7504524 | 0.7351635 | 2.645988 | 0.2871387 | 0.6415056 | 28739 | 0.6660191 | 0.5158808 | 0.9255612 | 0.5047607 | 0.6047711 | 0.6464061 | 0.7697136 | 0.7831176 |
| 29 | 0.8289175 | 0.5033416 | 0.5644581 | 5.5740416 | 0.6687968 | 1.2334758 | 0 | 0.6071581 | 0.5320682 | 0.9103183 | 0.4253956 | 2.0644427 | 0.3088956 | 0.6550209 | 2.8981733 | 0.3853836 | 0.6782611 |
| 30 | 0.6018762 | 0.7041271 | 0.8834399 | 0.5162058 | 0.5014956 | 3.0018989 | 0.4932798 | 1.2487259 | 0.4648862 | 1.0122426 | 0.3936311 | 0.5598506 | 0.2981063 | 0.5395712 | 1.1278056 | 0.5936225 | 1.2177372 |
| 31 | 1.8010817 | 0.6359243 | 0.5820145 | 1.2524697 | 1.64248 | 19.682836 | 0.6768014 | 0 | 0.8281743 | 0.788474 | 0 | 0.9617461 | 0.468756 | 0.7439989 | 0.4900083 | 0.6258416 | 0.5829106 |
| 32 | 0.5271594 | 0.788637 | 0.8683437 | 0.4918654 | 0.8378943 | 28691 | 0.3787699 | 0.697312 | 0 | 1.0318284 | 7.5672938 | 0.8437288 | 1.162031 | 0.529426 | 0.2146548 | 0.5499871 | 0.3565854 |
| 33 | 1.1121659 | 0.7409447 | 0.6622219 | 0.7243704 | 0.4079886 | 2.7667886 | 1.5015783 | 0.5296397 | 0.4397992 | 0.8804963 | 0.4392882 | 0 | 0.6234124 | 0.7472773 | 0.3598265 | 0.6405565 | 0.9221139 |
| 34 | 0.6449821 | 0.8512406 | 0.4972888 | 3.7825806 | 0.6082631 | 0.5507858 | 0.9195138 | 0.8752735 | 0.664297 | 0.5932464 | 0.3333748 | 0.6606828 | 0.71511 | 0.6435152 | 0.4954949 | 0.6247832 | 0.6587191 |
| 35 | 28964 | 0.9977072 | 0.5916031 | 5.875817 | 2.1456126 | 24.453188 | 0 | 0 | 0.4265595 | 0.7368994 | 2.5250656 | 0.7106283 | 0.5794094 | 0.6161443 | 0.57432 | 0.4170127 | 0.9118404 |
| 36 | 1.1047913 | 1.1000884 | 0.7202179 | 0 | 1.0189481 | 1.0768023 | 0.6247294 | 6.124451 | 8.1567366 | 0.5836243 | 0.611794 | 3.1444078 | 0.6042322 | 18.418329 | 0.4720763 | 0.5798694 | 1.1561933 |
| 37 | 0 | 0.7718797 | 0.91057 | 0.9746901 | 0.6604389 | 0.6350821 | 0.6017641 | 0.729244 | 0.5387656 | 2.3600445 | 0.6795764 | 0.7348712 | 0.3155152 | 0.470978 | 0.4544045 | 0.5517463 | 0.3706317 |
| 38 | 0.7383087 | 0.7188266 | 1.1767919 | 1.2322487 | 0.8297257 | 0 | 0.4339617 | 0.8561166 | 0.3670076 | 0.599923 | 0.873193 | 23288 | 0.6733388 | 0.4580579 | 0.5989658 | 0.4193539 | 0.3398962 |

Integral 1 / Integral 2 / Integral Ratio / Spot Area / Saturation Factor

Ready

Click here to begin | 0 | Y: 0 | pixel value in A: 12 | pixel value in B: 47

Start | Inbox - Microsoft Outlook | TIGR_Spotfinder - 0.0.0 | Microsoft Excel - Book1 | 6:38 PM

**Output includes:**
Integrated Intensity 1,
Integrated Intensity 2,
Ratio, Spot Area,
Saturation,
Mean and Median
Intensities,
Quality Factors

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

# Comparison of Mean, Median, and Mode Ratios

## Mean ratio

| | | | | | |
|---|---|---|---|---|---|
| | 1.012 | 0 | 0.966 | 0.987 | 0.897 |
| | 1.135 | 0 | 1.037 | 1.034 | 1.015 |
| | 1.008 | 0 | 1.058 | 1.008 | 1.058 |
| | 1.079 | 0 | 1.059 | 1.061 | 1.026 |
| | 1.022 | 0 | 1.069 | 1.031 | 1.019 |
| | 1.070 | 0 | 1.032 | 1.024 | 1.139 |
| | 0.986 | 0 | 1.058 | 1.064 | 1.047 |
| | 1.057 | 0 | 0.990 | 1.063 | 1.022 |
| | 0.935 | 0 | 1.105 | 1.069 | 1.079 |
| | 1.094 | 0 | 1.024 | 1.057 | 0.892 |
| | 1.014 | 0 | 1.040 | 0.997 | 1.019 |
| | 0.985 | 0 | 1.005 | 1.067 | 1.035 |
| | 1.011 | 0 | 1.033 | 1.035 | 1.143 |
| | 1.232 | 0 | 0.996 | 1.169 | 1.077 |
| | 0.819 | 0 | 1.085 | 1.118 | 1.039 |
| | 0.942 | 0 | 0.999 | 1.129 | 1.061 |
| average | 1.025 | | 1.035 | 1.057 | 1.036 |
| stdev | 0.092 | | 0.037 | 0.049 | 0.067 |

## Median ratio

| | | | | | |
|---|---|---|---|---|---|
| | 1.000 | 0 | 0.930 | 1.053 | 0.898 |
| | 1.067 | 0 | 1.053 | 1.056 | 1.015 |
| | 1.008 | 0 | 1.042 | 1.056 | 1.003 |
| | 1.098 | 0 | 1.047 | 1.034 | 1.026 |
| | 0.980 | 0 | 0.998 | 1.004 | 0.955 |
| | 1.041 | 0 | 1.040 | 1.045 | 1.074 |
| | 1.030 | 0 | 1.081 | 1.029 | 1.134 |
| | 1.013 | 0 | 0.987 | 1.085 | 1.014 |
| | 0.896 | 0 | 1.025 | 1.007 | 1.079 |
| | 1.026 | 0 | 1.028 | 1.072 | 0.874 |
| | 1.001 | 0 | 1.067 | 1.020 | 1.014 |
| | 0.977 | 0 | 0.928 | 1.094 | 0.979 |
| | 1.061 | 0 | 1.105 | 1.027 | 1.096 |
| | 1.136 | 0 | 0.963 | 1.067 | 1.020 |
| | 0.929 | 0 | 1.033 | 1.083 | 1.062 |
| | 0.877 | 0 | 0.974 | 1.133 | 1.110 |
| average | 1.009 | | 1.019 | 1.054 | 1.022 |
| stdev | 0.068 | | 0.051 | 0.034 | 0.072 |

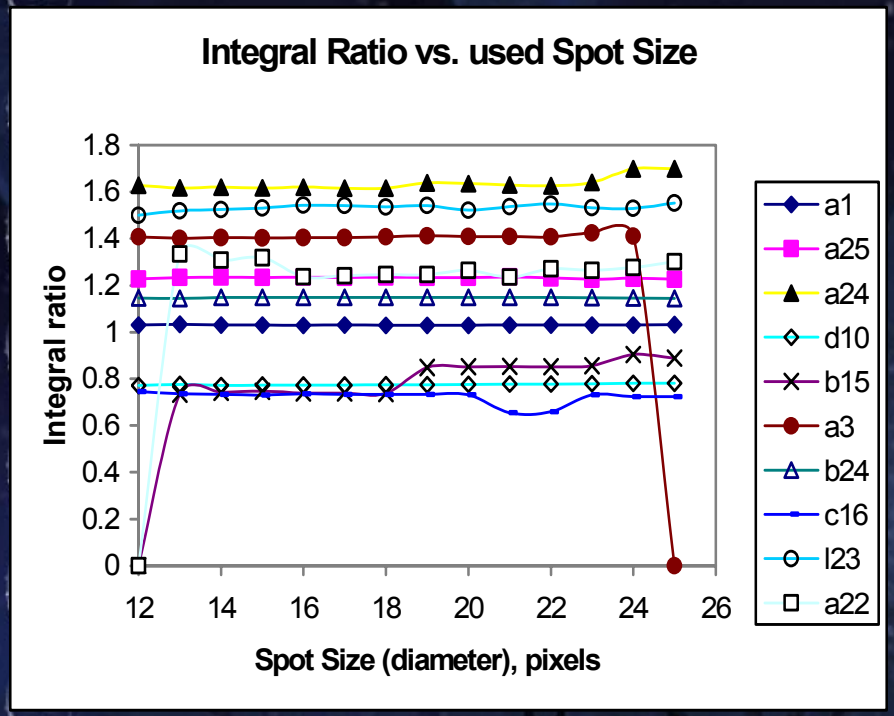
## Mode ratio

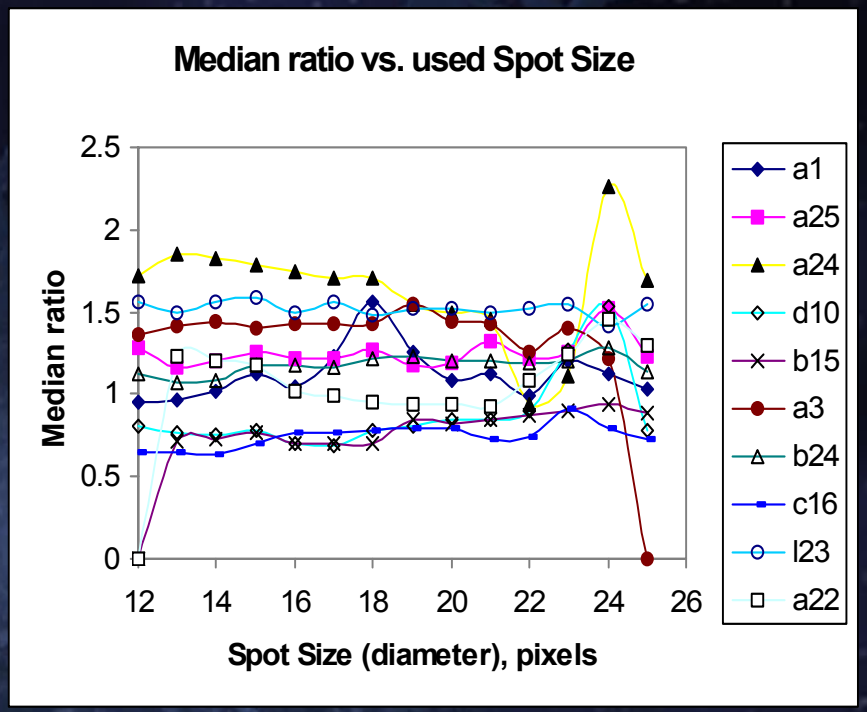
| | | | | | |
|---|---|---|---|---|---|
| | 1.398 | 0 | 1.113 | 0.830 | 0.518 |
| | 1.984 | 0 | 0.554 | 0.906 | 5.721 |
| | 1.536 | 0 | 1.113 | 1.152 | 0.570 |
| | 1.051 | 0 | 2.490 | 1.684 | 1.437 |
| | 2.794 | 0 | 0.976 | 1.544 | 1.651 |
| | 1.095 | 0 | 1.564 | 1.203 | 1.516 |
| | 1.332 | 0 | 1.253 | 0.614 | 94.797 |
| | 1.697 | 0 | 16.921 | 2.039 | 0.788 |
| | 0.550 | 0 | 1.065 | 0.873 | 0.916 |
| | 1.022 | 0 | 0.742 | 0.377 | 0.681 |
| | 1.707 | 0 | 1.714 | 1.411 | 2.167 |
| | 1.784 | 0 | 0.392 | 0.976 | 1.080 |
| | 2.269 | 0 | 0.318 | 1.708 | 0.615 |
| | 1.932 | 0 | 0.604 | 1.331 | 0.579 |
| | 0.052 | 0 | 1.618 | 0.380 | 1.254 |
| | 0.641 | 0 | 0.543 | 1.373 | 1.022 |
| average | 1.428 | | 2.061 | 1.150 | 7.207 |
| stdev | 0.690 | | 4.004 | 0.479 | 23.391 |

**A comparison of Cy3/Cy5 ratios calculated using the mean, median, and mode ratios for control spots that should have a measured ratio of 1 for the 1st, 3rd, 4th, 5th columns.**

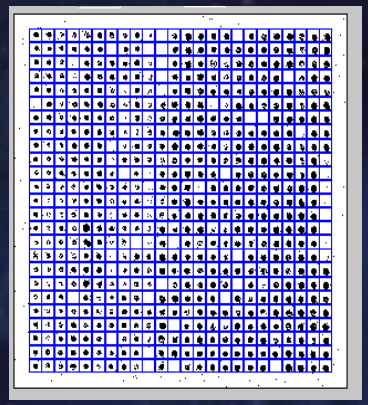
TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

# Integral (Mean) Ratio *vs.* Median Ratio



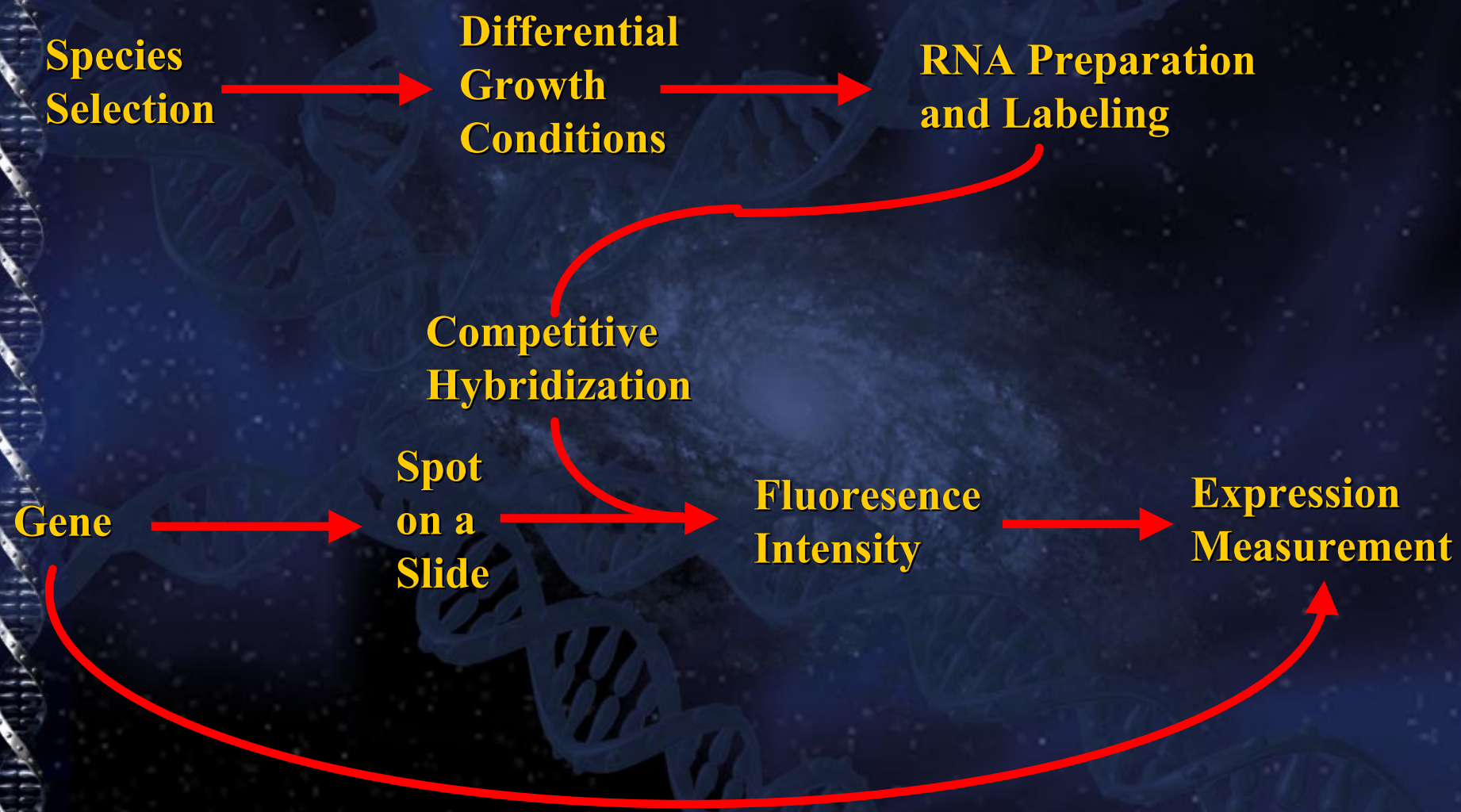




**A comparison of Cy3/Cy5 ratios for various spot sizes using either the integrated intensity or the pixel median. In this case, the actual spot size is approximately 15 pixels in diameter.**



TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

# Microarray Expression Analysis




TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

# Data Analysis Issues

- Presentation
- Multiple Views
- Normalization
- Identification of Differentially Expressed Genes
- Multiple Experiments

# Why Normalize Data?

- Goal is to measure ratios of gene expression levels

  $(\text{ratio})_i = R_i/G_i$

  where $R_i/G_i$ are, respectively , the measured intensities for the *i*th spot.

- In a self-self hybridization, we would expect all ratios to be equal to one:

  $R_i/G_i = 1$ for all *i*. But they may not be.

- Why not?
  - Unequal labeling efficiencies for Cy3/Cy5
  - Noise in the system
  - Differential expression

- Normalization brings (appropriate) ratios back to one.

# Normalization Approaches

- Total Intensity
- Linear Regression
- Ratio statistics described by Chen, Dougherty, & Bittner
  *J. Biomed. Optics* (1997) 2(4) 364-374
- Iterative log(ratio) mean centering

**Any of these using:**

- Entire Data Set
- User-defined Data Set/Controls

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

# Normalization Approaches

## Entire Data Set

- Probe Quantification less important
- No assumption on which genes constitute "housekeeping" set
- Uses all the data
- No independent confirmation

## User-defined Data Set/Controls

- Requires definition of "housekeeping" set
  *or* good added controls
- Requires good RNA quantitation
- Ignores much data

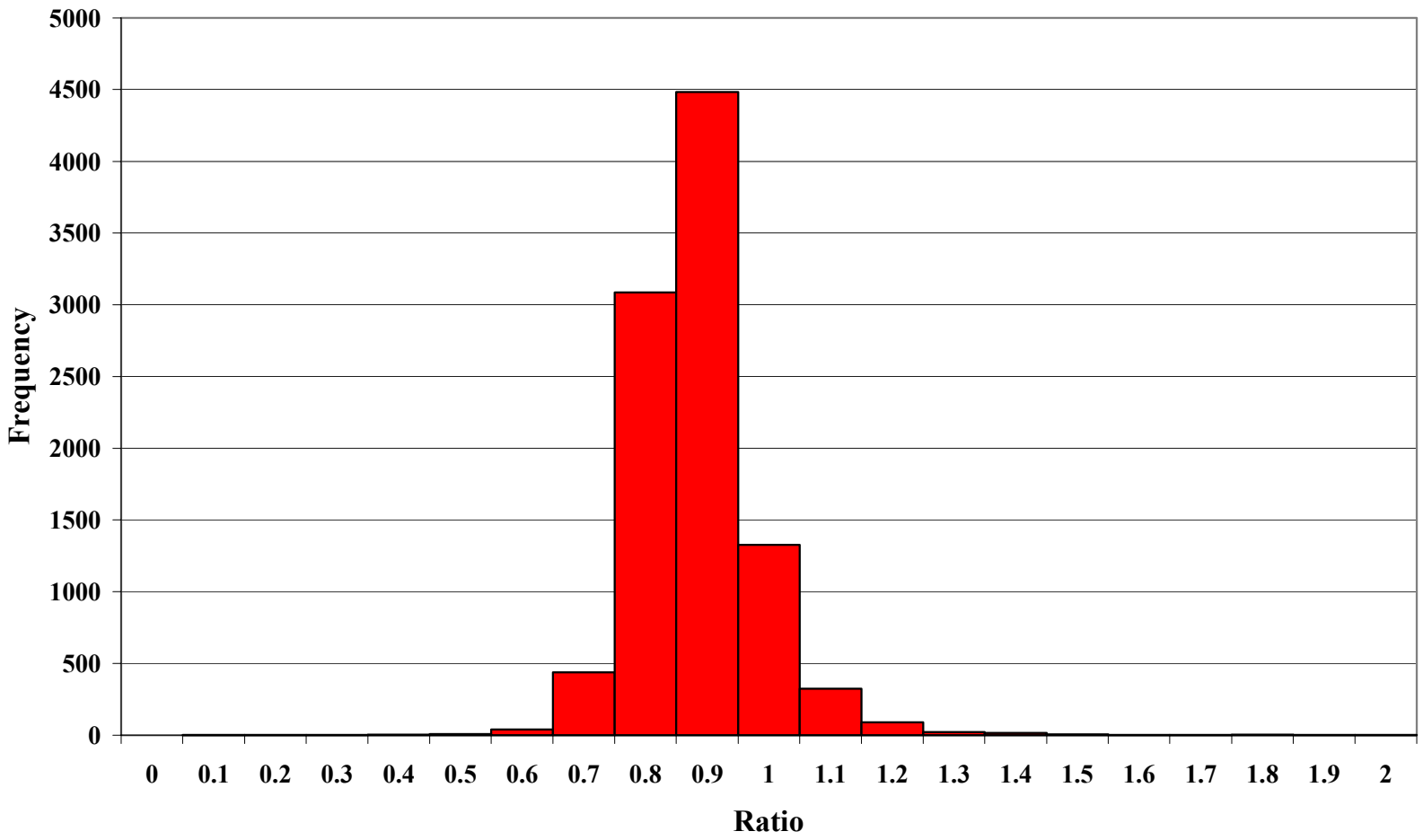
TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

# Normalization Approaches

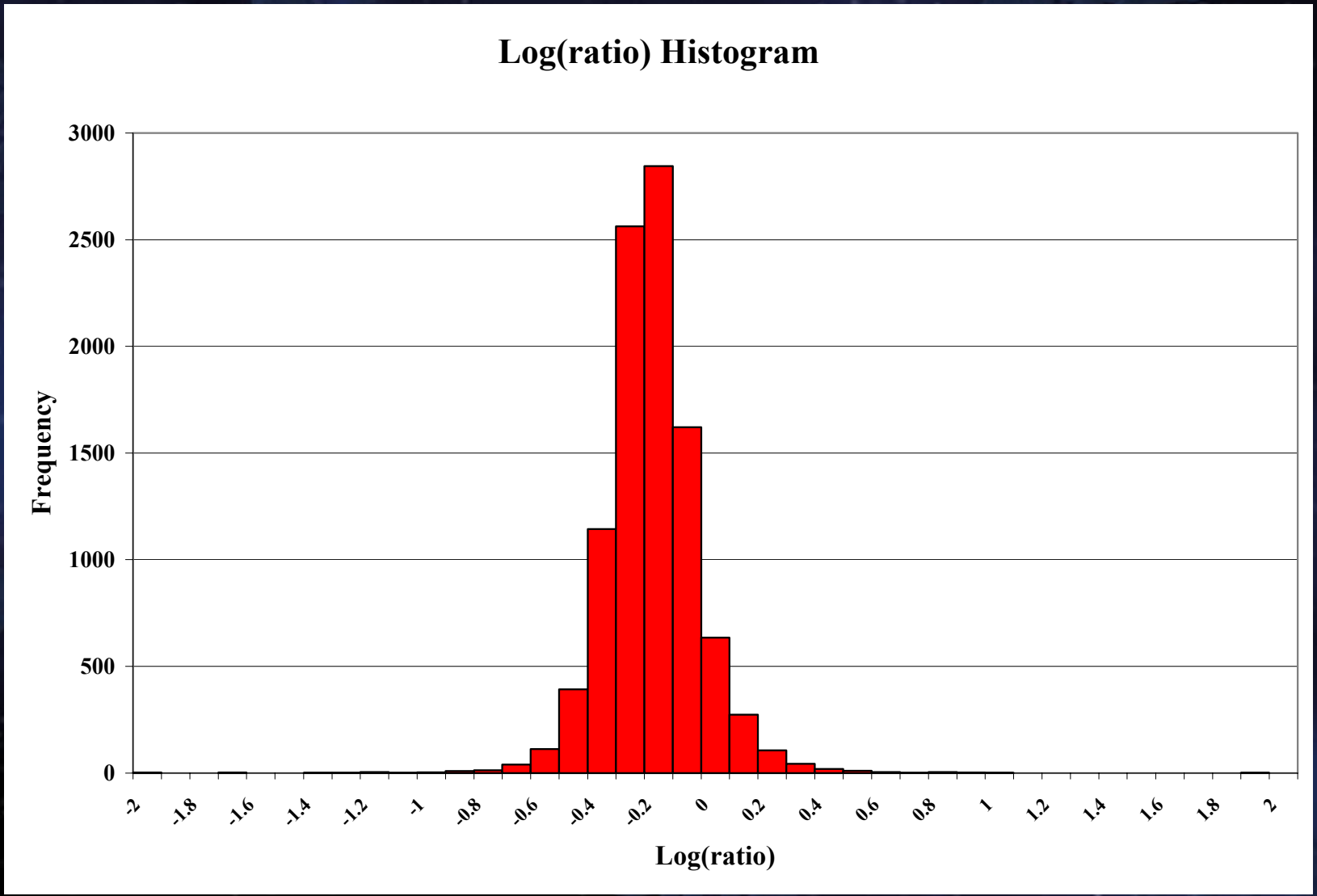
Solution(?)

- Experiment dependent
- Use a combination of techniques
- SMART Experimental design

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH




TIGR
THE INSTITUTE FOR GENOMIC RESEARCH




TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

# Normalization Approaches: Total Intensity

- Assumption: Total RNA (mass) used is same for both samples.
- So, averaged across thousands of genes, total hybridization should be the same

**Normalization Factor:** $$N = \frac{\sum_{k=1}^{N_{array}} R_k}{\sum_{k=1}^{N_{array}} G_k}$$

**Normalization:** $G'_k = NG_k$ and $R'_k = R_k$.

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

# Normalization Approaches: Linear Regression

**Assumption: Total RNA used is constant, some genes expressed with ratio of 1, slope of best fit line normalized to 1**

$$R_k = \beta_0 + \beta_1 G_k + u_k$$

**Normalization Factor:**

$$S(\beta_0, \beta_1) = \sum_{k=1}^{n} u_k^2 = \sum_{i \backslash k=1}^{n} (R_k - \beta_0 - \beta_1 G_k)^2$$

The values of $\beta_0$ and $\beta_1$ that minimize $S(\beta_0, \beta_1)$, $b_0$ and $b_1$, are given by

$$b_1 = \frac{\sum_{k=1}^{n} (R_k - \overline{R})(G_k - \overline{G})}{\sum_{k=1}^{n} (G_k - \overline{G})^2} \quad \text{and} \quad b_0 = \overline{R} - b_1 \overline{G},$$

where $\overline{R} = \frac{\sum R_k}{n}$ and $\overline{G} = \frac{\sum G_k}{n}$.

**Normalization:** $G'_k = \left[\frac{1}{b_1}\right] G_k$ and $R'_k = R_k$.

# Normalization Approaches: Ratio Statistics (1)

**Assumption: Total RNA used is constant, some genes expressed with ratio of 1, variations are functions of the common mean**

$$\sigma_{G_k} = c\mu_{G_k} \text{ and } \sigma_{R_k} = c\mu_{R_k}, \text{ with } \mu_{G_k} = \mu_{R_k} = \mu_k .$$

**Probablilty Density for Ratio $T_k$:** $f_{T_k}(t) \approx \frac{(1+t)\sqrt{1+t^2}}{c(1+t^2)\sqrt{2\pi}} \exp\left[\frac{-(t-1)^2}{2c(1+t^2)}\right]$

This density can be used to calculate the mean, standard deviation and confidence interval limits for the distribution of measured ratio values. As functions of $c$, these parameters can be estimated using a polynomial approximation

$$y = a_3c^3 + a_2{}^2c^2 + a_1c + a_0$$

with constants are chosen appropriately:

μ: $(a_3, a_2, a_1, a_0)$ = (0.364, 1.279, –0.0427, 1.001)

$\sigma$: $(a_3, a_2, a_1, a_0)$ = (–2.805, 2.911, –2.706, 0.979)

lower limit at 95% confidence: $(a_3, a_2, a_1, a_0)$ = (28.644, –2.830, 3.082, 0.989)

upper limit at 95% confidence: $(a_3, a_2, a_1, a_0)$ = (–5.002, .4.462, –3.496, 0.9968)

# Normalization Approaches: Ratio Statistics (2)

Assume that the population mean $\mu_0 = 1$ and let the first approximation of the normalization parameter $m_1$ be equal to the calculated sample mean. A first approximation of $c$, $\hat{c}_1$, is calculated using

$$\hat{c}_i = \left[ \frac{1}{n} \sum_{j=1}^{n} \frac{(t_j - 1)^2}{(1 + t_j^2)} \right]^{1/2}$$

where the sum is over the $n$ elements taken initially between the one-half and twice the sample mean.

Upper and lower limits at the 95% confidence level, $\theta_1$ and $\theta_2$, are then calculated using $\hat{c}_1$ and the previous approximation.

A normalization factor $\hat{m}_1$ is calculated using

$$\hat{m}_i = \frac{1}{\hat{\mu}_{i-1}} \left( \frac{1}{n} \sum_{j=1}^{n} t_j \right),$$

where, again, we take $\hat{\mu}_0 = 1$, the sum is over the $n$ array elements used to estimate $\hat{c}_1$, and $i$ is an index used to count the number of iterations

The individual ratios are then rescaled using

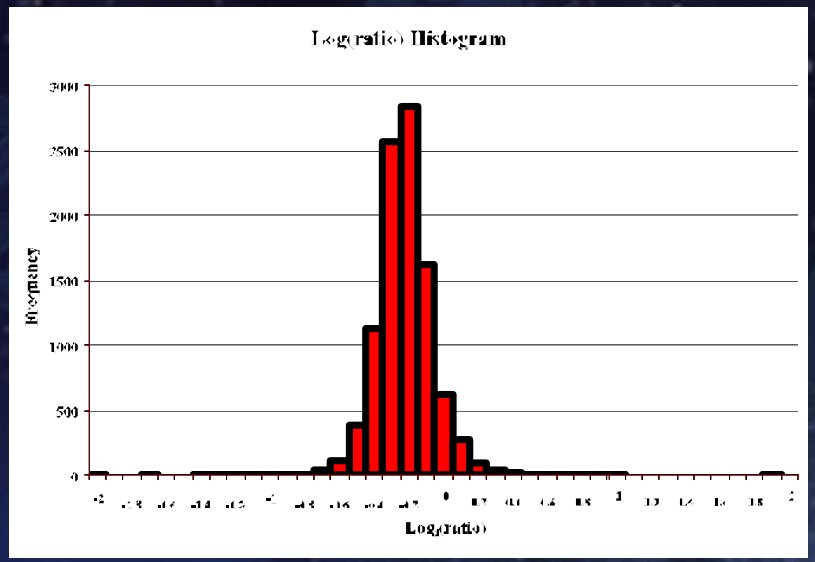
$$t'_k = \frac{t_k}{\hat{m}_i} = \frac{R_k}{(\hat{m}_i G_k)} = \frac{R'_k}{G'_k}.$$

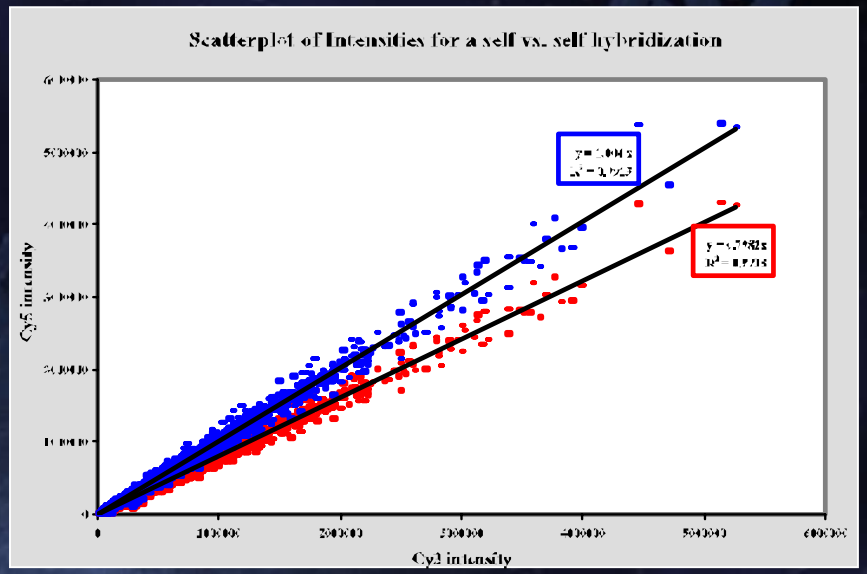
This process is then iterated until the calculated value of the mean estimator converges to a fixed value.

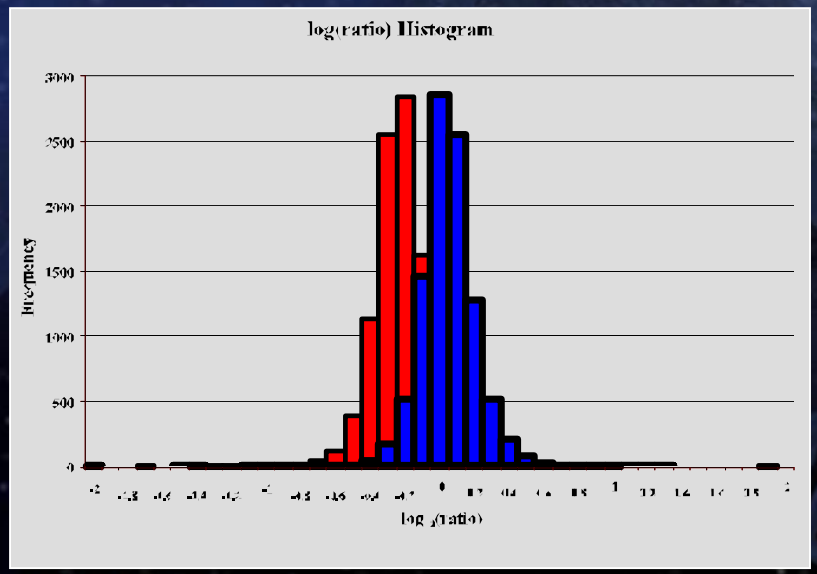
The upper and lower confidence limits for the normalized experimental data are then calculated as

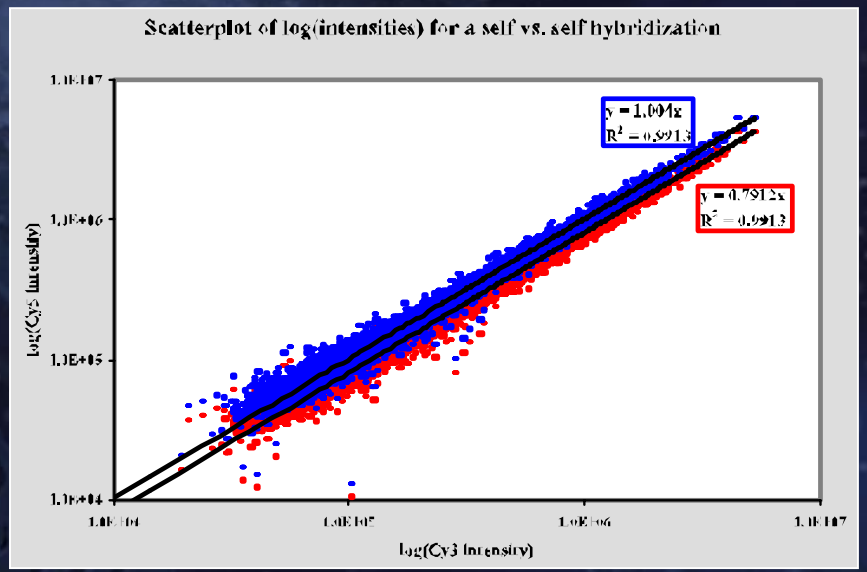
$$\theta'_1 = \hat{m}\theta_1 \text{ and } \theta'_2 = \hat{m}\theta_2$$

and $(\theta'_1, \theta'_2)$ are used to define the limits for identification of differentially expressed genes



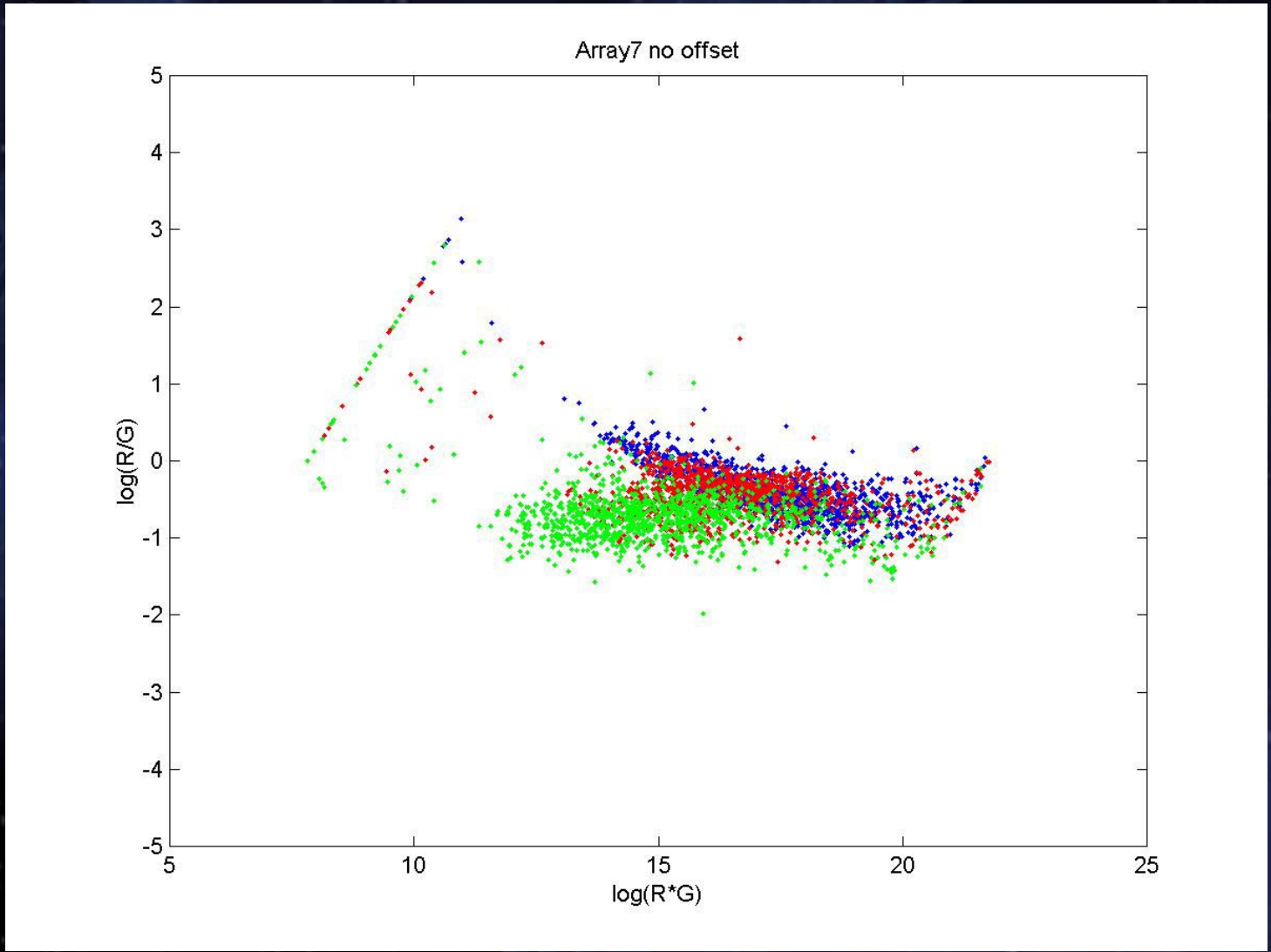




TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

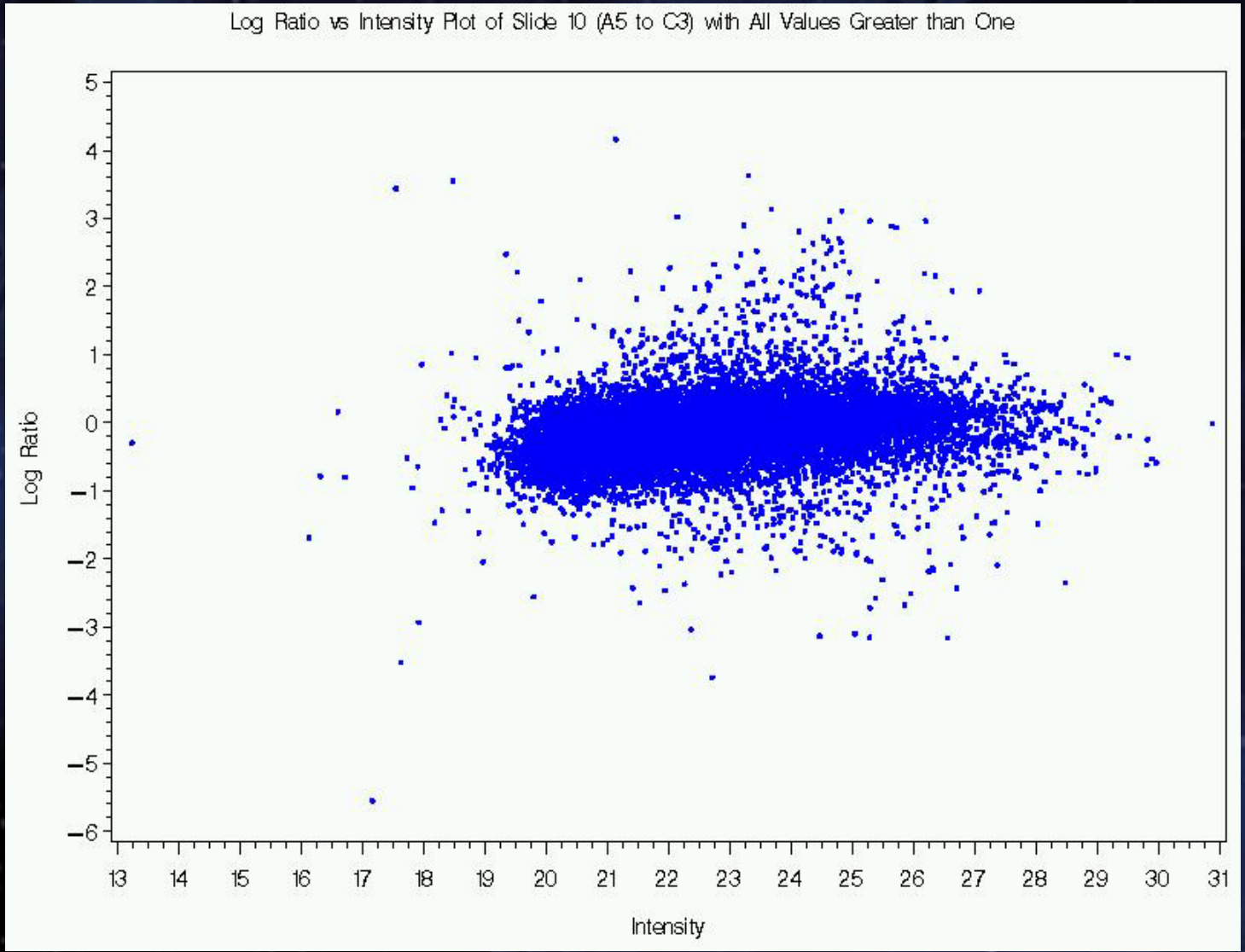
# Bad Data from Parts Unknown




TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

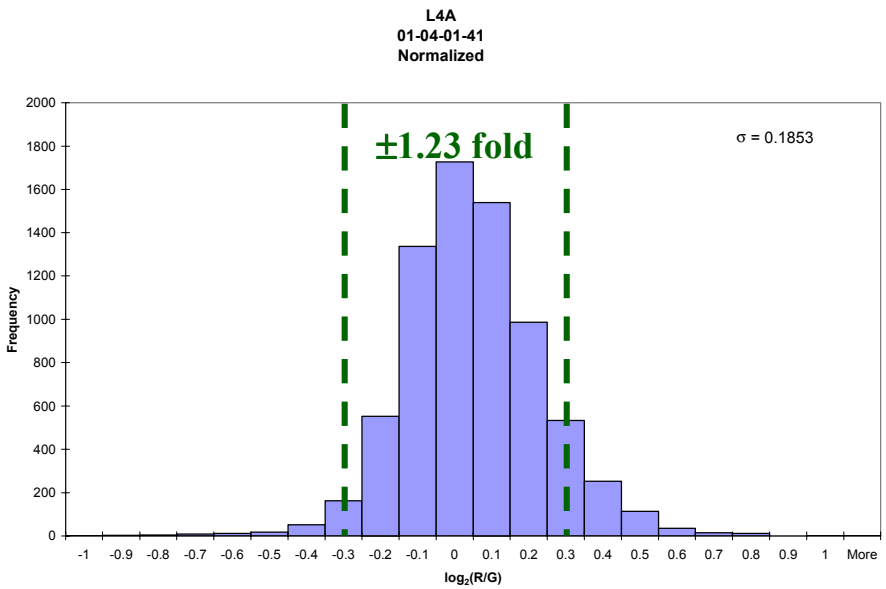
Gary Churchill

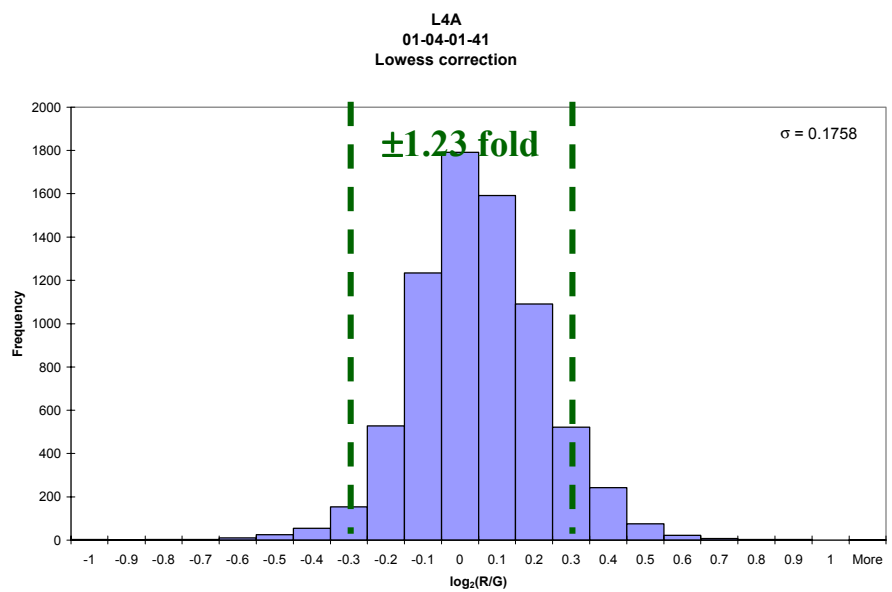
# Good Data from TREX

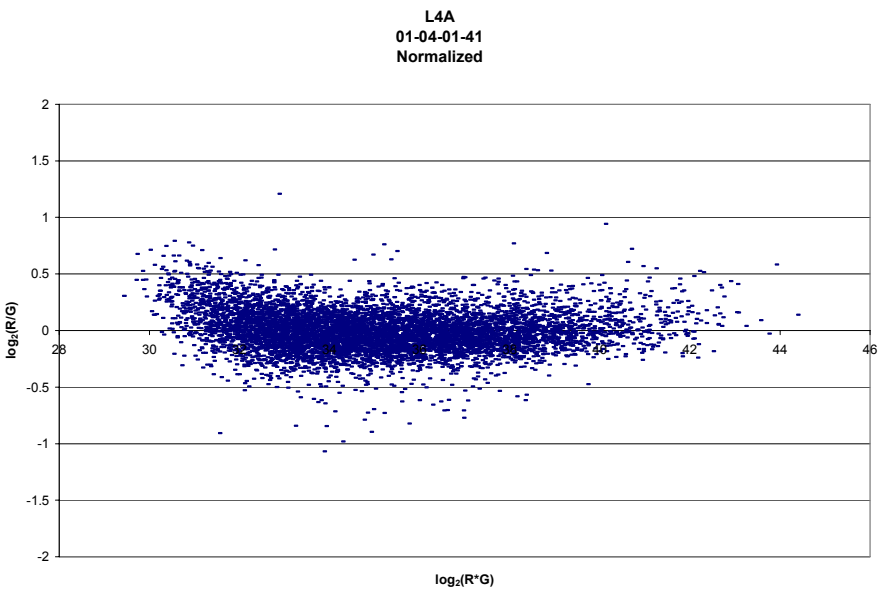


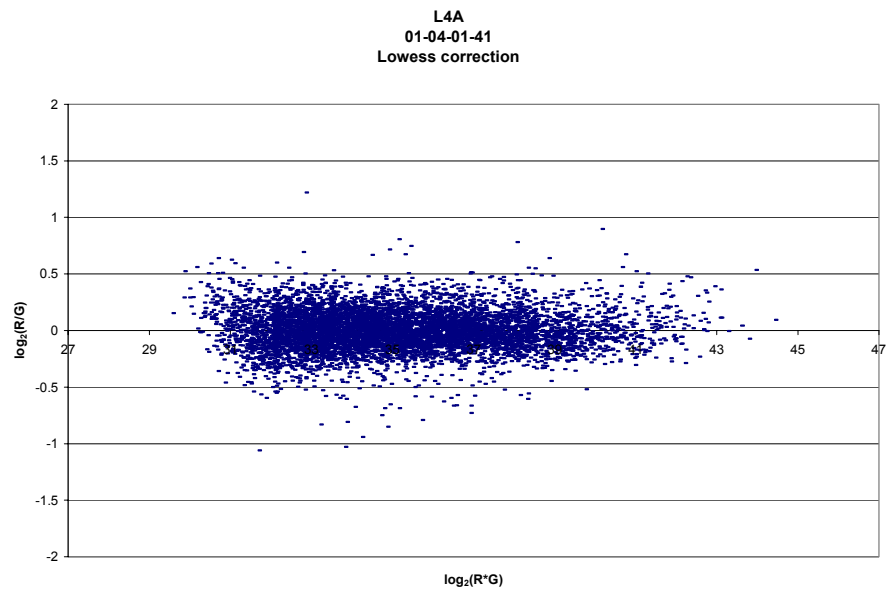

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

# Normalization using local linear regression



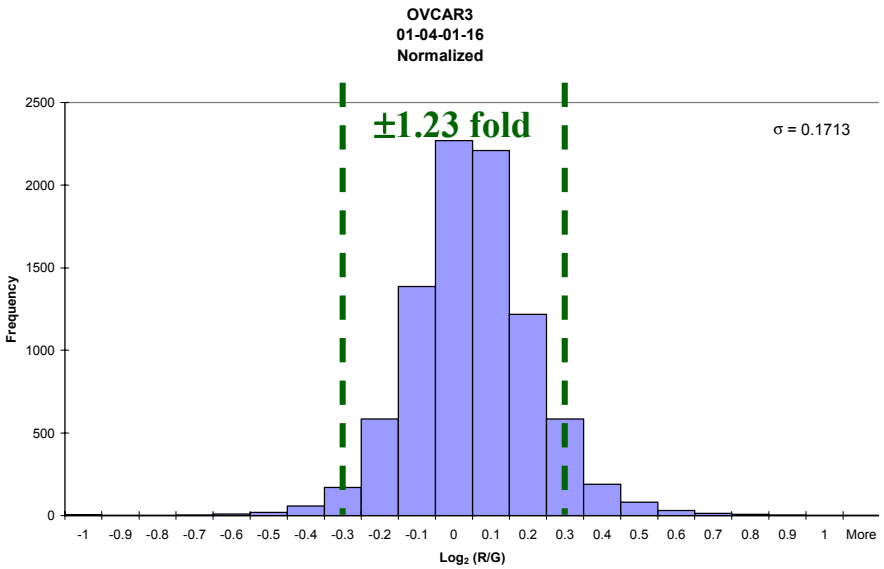



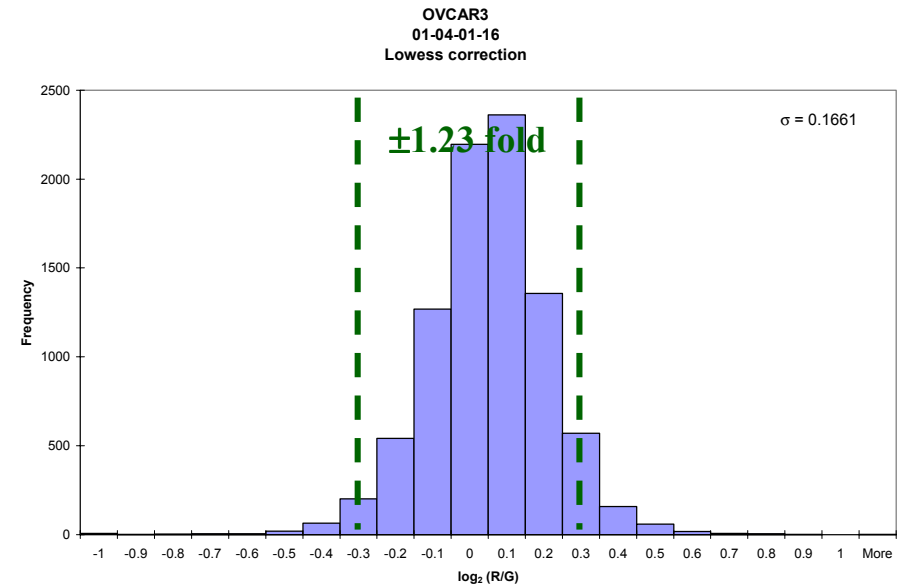



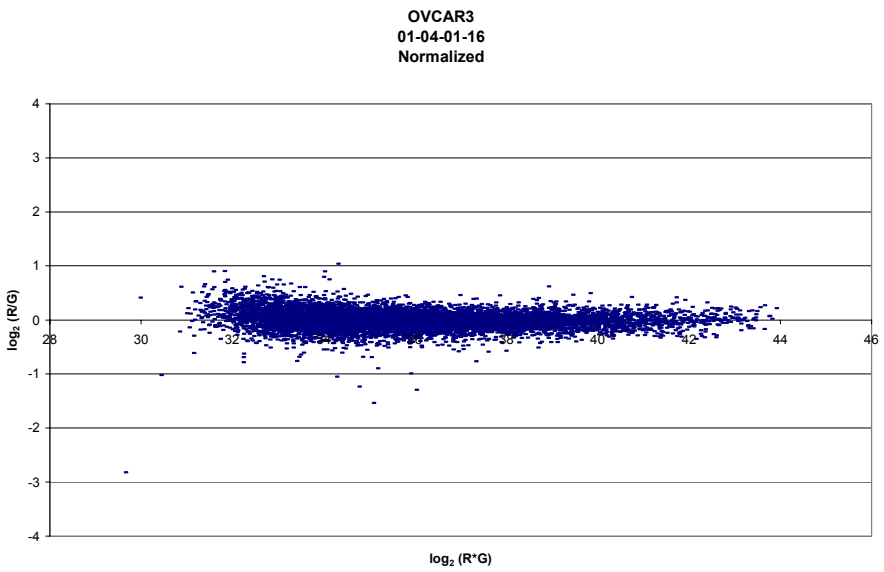


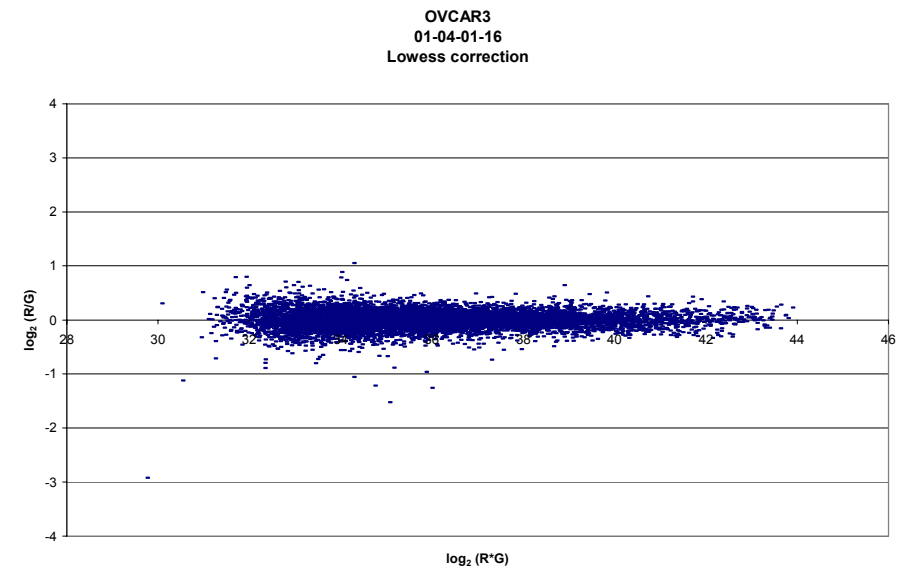

Ivana Yang , John Quackenbush

# Normalization using local linear regression



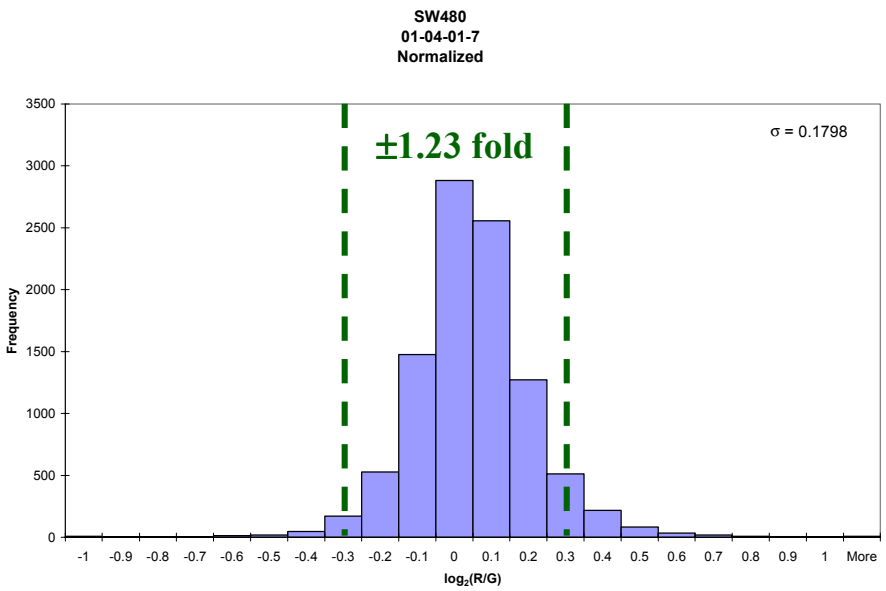



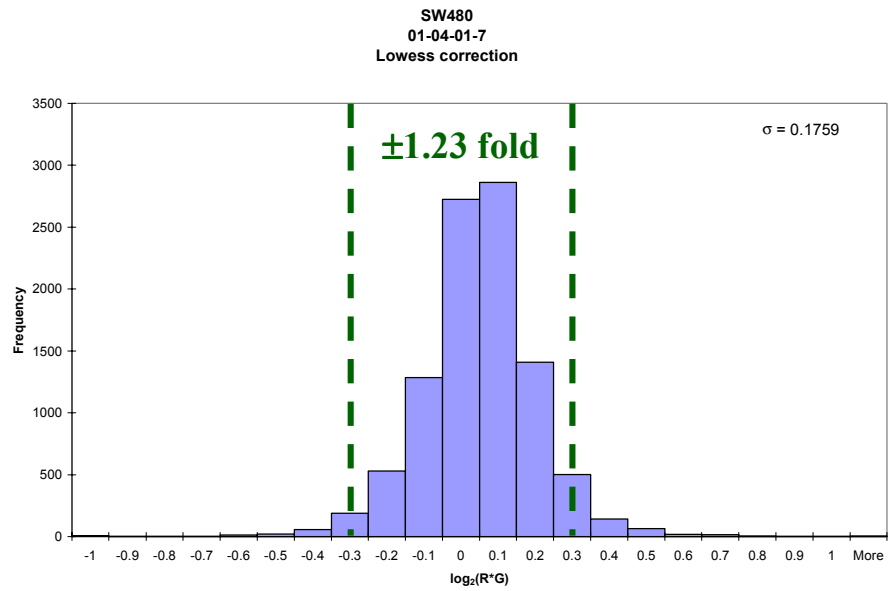



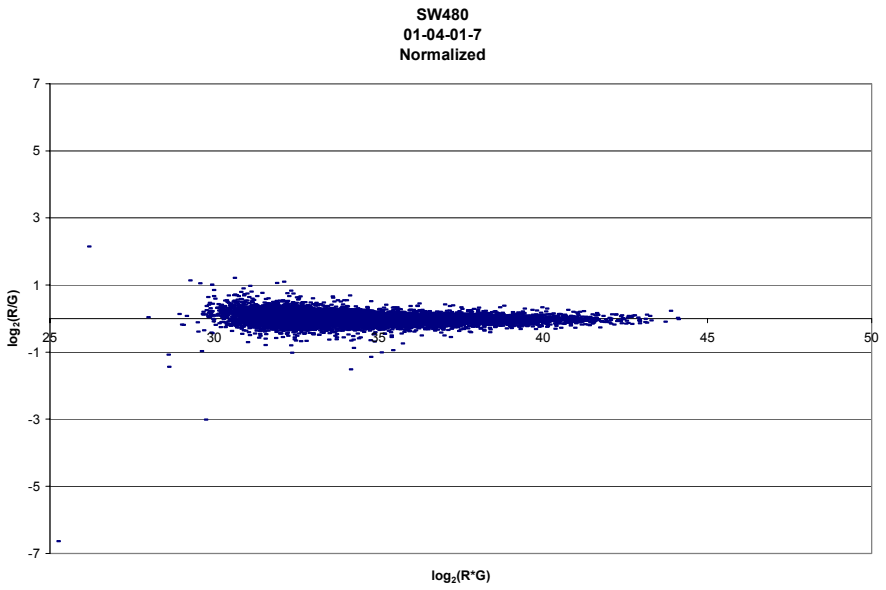


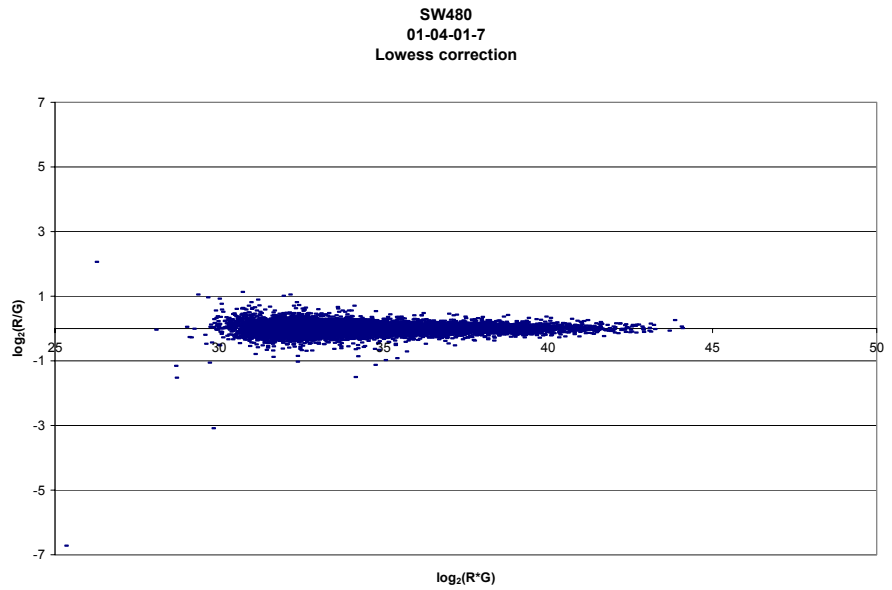

Ivana Yang , John Quackenbush

# Normalization using local linear regression






Ivana Yang , John Quackenbush

# Multiple Experiments?

- Goal is identify genes (or experiments) which have "similar" patterns of expression
- This is a problem in data mining
- "Clustering Algorithms" are most widely used
- Types
  - Agglomerative: Hierarchical
  - Divisive: *k*-means, SOMs
  - Others: Principal Component Analysis (PCA)
- All depend on how one measures distance

# Expression Vectors

- Crucial concept for understanding clustering

- Each gene is represented by a vector where coordinates are its values log(ratio) in each experiment
  - $x = \log(\text{ratio})_{\text{expt1}}$
  - $y = \log(\text{ratio})_{\text{expt2}}$
  - $z = \log(\text{ratio})_{\text{expt3}}$
  - etc.

- For example, if we do six experiments,
  - $\text{Gene}_1 = (-1.2, -0.5, 0, 0.25, 0.75, 1.4)$
  - $\text{Gene}_2 = (0.2, -0.5, 1.2, -0.25, -1.0, 1.5)$
  - $\text{Gene}_3 = (1.2, 0.5, 0, -0.25, -0.75, -1.4)$
  - etc.

# Expression Matrix

- These gene expression vectors of log(ratio) values can be used to construct an expression matrix

| | Expt 1 | Expt 2 | Expt 3 | Expt 4 | Expt 5 | Expt 6 |
|---|---|---|---|---|---|---|
| $Gene_1$ | -1.2 | -0.5 | 0 | 0.25 | 0.75 | 1.4 |
| $Gene_2$ | 0.2 | -0.5 | 1.2 | -0.25 | -1.0 | 1.5 |
| $Gene_3$ | 1.2 | 0.5 | 0 | -0.25 | -0.75 | -1.4 |
| etc. | | | | | | |

- This is often represented as a red/green colored matrix

# Distance metrics

- Distances are measured "between" expression vectors
- Distance metrics define the way we measure distances
- Many different ways to measure distance:
  - Euclidean distance
  - Pearson correlation coefficient(s)
  - Manhattan distance
  - Mutual information
  - Kendall's Tau
  - etc.
- Each has different properties and can reveal different features of the data

# Distance Matrix

- Once a distance metric has been selected, the starting point for all clustering methods is a "distance matrix"

| | $Gene_1$ | $Gene_2$ | $Gene_3$ | $Gene_4$ | $Gene_5$ | $Gene_6$ |
|---|---|---|---|---|---|---|
| $Gene_1$ | 0 | 1.5 | 1.2 | 0.25 | 0.75 | 1.4 |
| $Gene_2$ | 1.5 | 0 | 1.3 | 0.55 | 2.0 | 1.5 |
| $Gene_3$ | 1.2 | 1.3 | 0 | 1.3 | 0.75 | 0.3 |
| $Gene_4$ | 0.25 | 0.55 | 1.3 | 0 | 0.25 | 0.4 |
| $Gene_5$ | 0.75 | 2.0 | 0.75 | 0.25 | 0 | 1.2 |
| $Gene_6$ | 1.4 | 1.5 | 0.3 | 0.4 | 1.2 | 0 |

- The elements of this matrix are the pair-wise distances. Note that the matrix is symmetric about the diagonal.

# Hierarchical clustering

- Select the data you want to cluster
- "Filter" (normalize) the data appropriately and select distance
- Apply method:
  1. Search through the distance matrix and find the two most similar clusters. This is the first true stage in the "clustering" process. If several pairs share the same similarity, use a predetermined rule to decide between alternatives.
  2. Fuse the two selected clusters to produce a new cluster that now contains at least two objects.
  3. Calculate the distances between this new cluster and all other clusters. There is no need to calculate *all* distances since only those involving the new cluster have changed.
- Repeat steps 1-3 until all objects are in one cluster.

# *k*-means clustering

- Select the data you want to cluster and filter; select distance

- Apply method:

1. All initial objects are randomly assigned to one of *k* clusters (where *k* is an input parameter to the algorithm).
2. An average expression vector is then calculated for each cluster and this is used to compute the distances between clusters.
3. Objects are moved between clusters and intra- and inter-cluster distances are measured with each move. Objects are allowed to remain in the new cluster only if they are closer to it than to their previous cluster.
4. Following each move, the expression vectors for each cluster are recalculated.
5. The shuffling proceeds until moving any more objects would make the clusters more variable.

# Self Organizing Maps (SOMs)

- Select the data you want to cluster and filter; select distance

- Apply method:

1. Random vectors are constructed and assigned to each partition. (where the number and geometry are input parameters).
2. A gene is picked at random and using a selected distance metric, the reference vector that it is closest to the gene's is identified .
3. The reference vector is then adjusted so that it is more similar to the randomly picked gene's. The reference vectors that are nearby on the two dimensional grid are also adjusted so that they too are more similar to the randomly selected gene .
4. Steps 2 and 3 are iterated several thousand times, decreasing the amount by which the reference vectors are adjusted and increasing the stringency used to define closeness in each step. As the process continues, the reference vectors are converge to fixed values .
5. Finally, the genes are mapped to the relevant partitions depending on the reference vector to which they are most similar.

# Principal Component Analysis (PCA)

- Select the data you want to cluster and filter

- Apply method:
  OK, this gets a bit complicated. . . .

  Basically:
  1. We find the eigenvectors of the expression matrix
  2. We select those with the greatest eigenvalues
  3. We project our data on the eigenvectors with the three greatest eigenvalues
  4. And make pretty pictures

# Support Vector Machines (SVM)

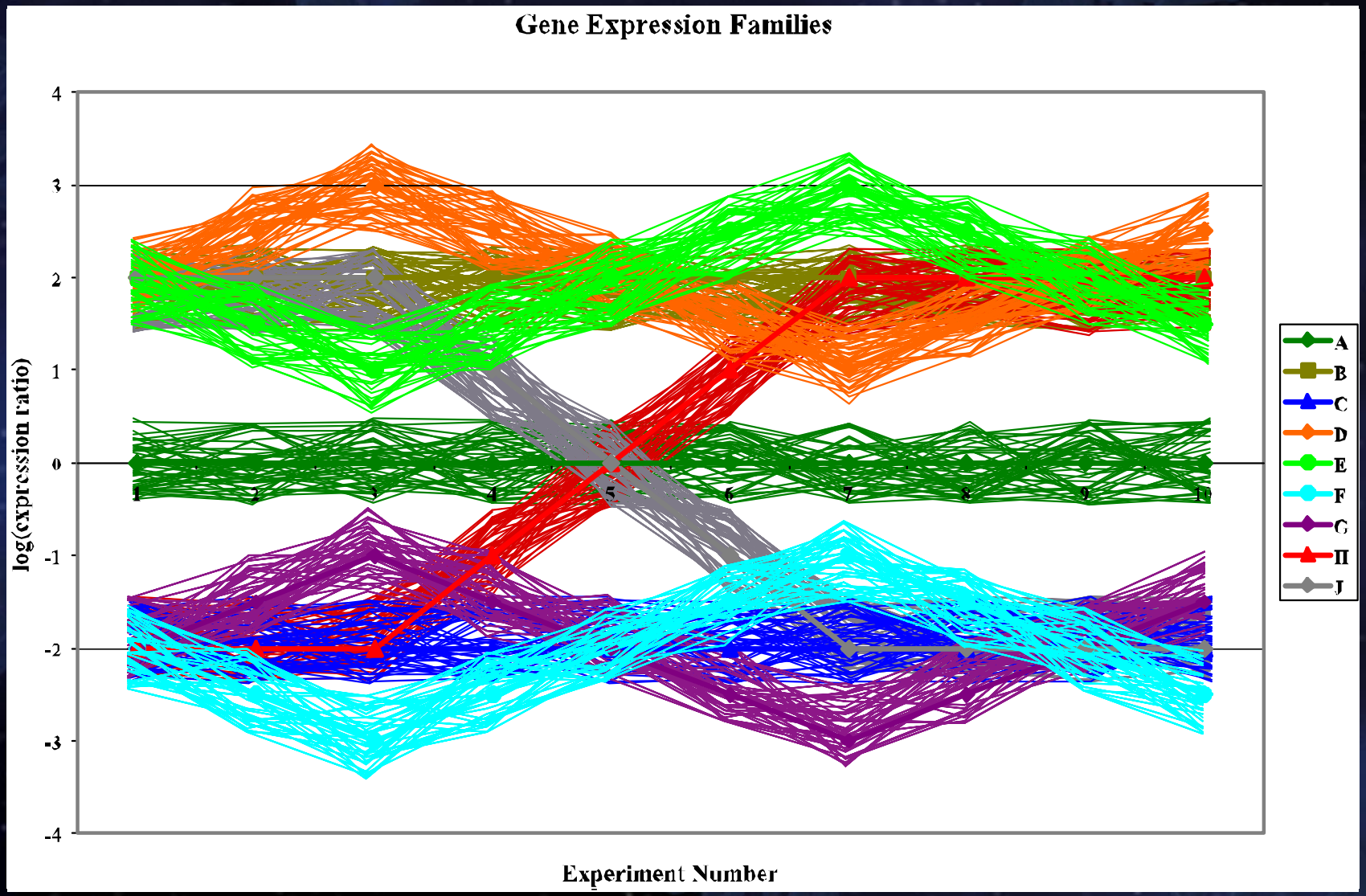
- Select the data you want to cluster and filter

- Apply method:
  OK, this gets even more complicated. . . .

  Basically this is a neural network approach to finding dividing your data into genes "like" and "unlike" a training set. . . .

  1. Pick a set of genes you are know about (your training set)
  2. Train the SVM. This produces a pattern that can be recognized
  3. Screen the data using the SVM model

# TIGR MeV: Test Data Set




TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

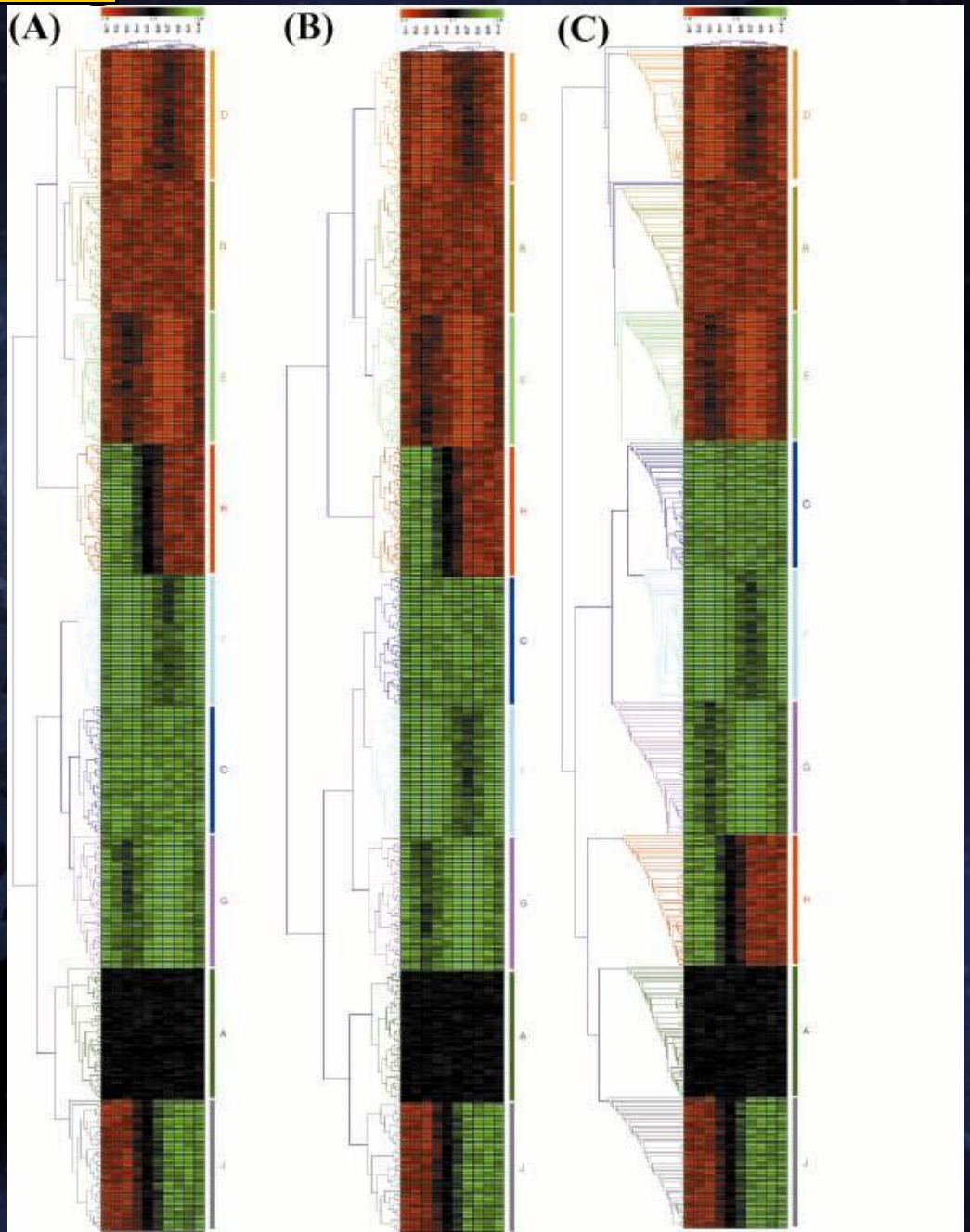
# Hierarchical Clustering

(A) Average Linkage

(B) Complete Linkage

(C) Single Linkage

Even related algorithms produce slightly different views of the data.



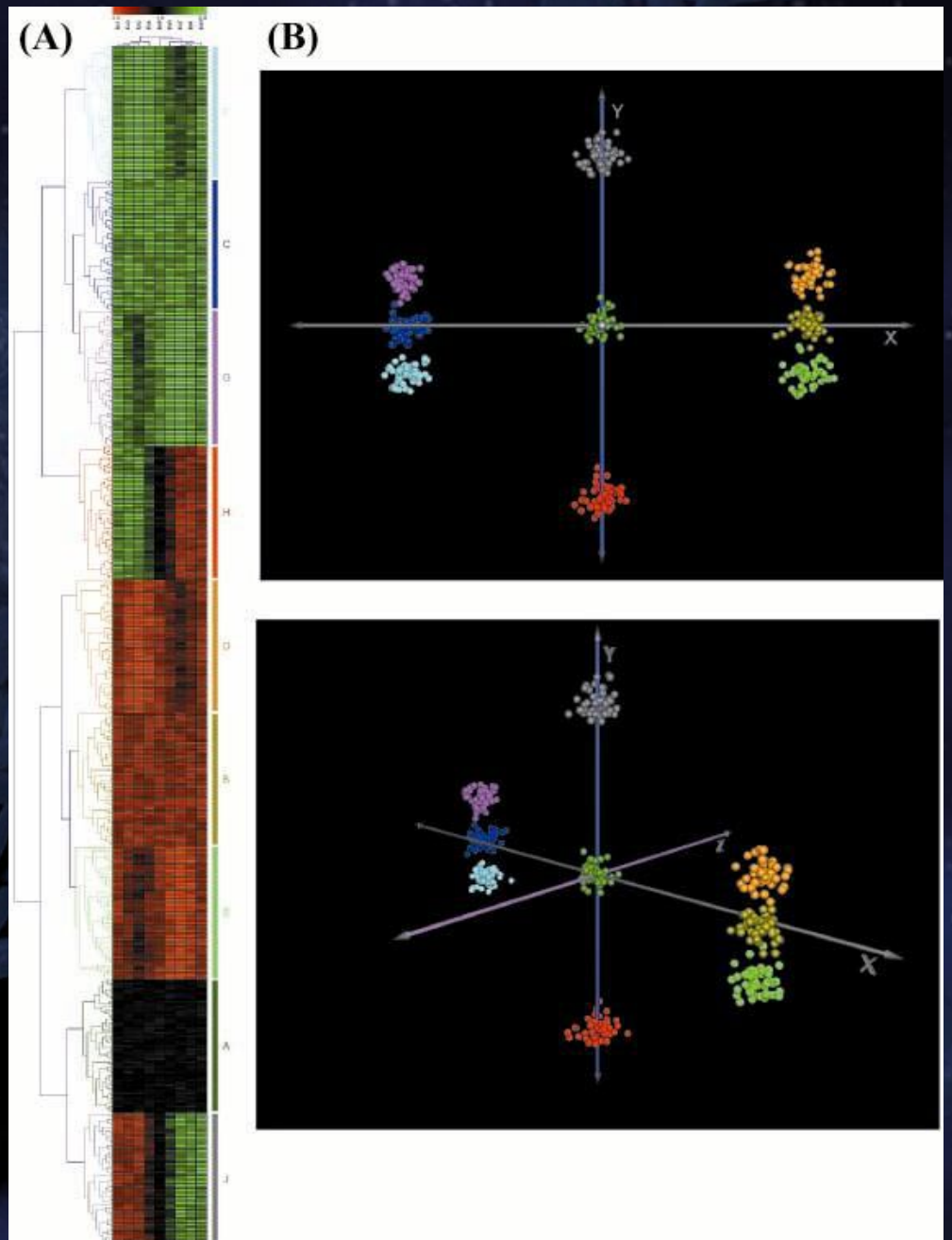

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

# Hierarchical Clustering and PCA

(A) Average Linkage

(B) PCA

Separate clusters may have more or less support when using different algorithms.

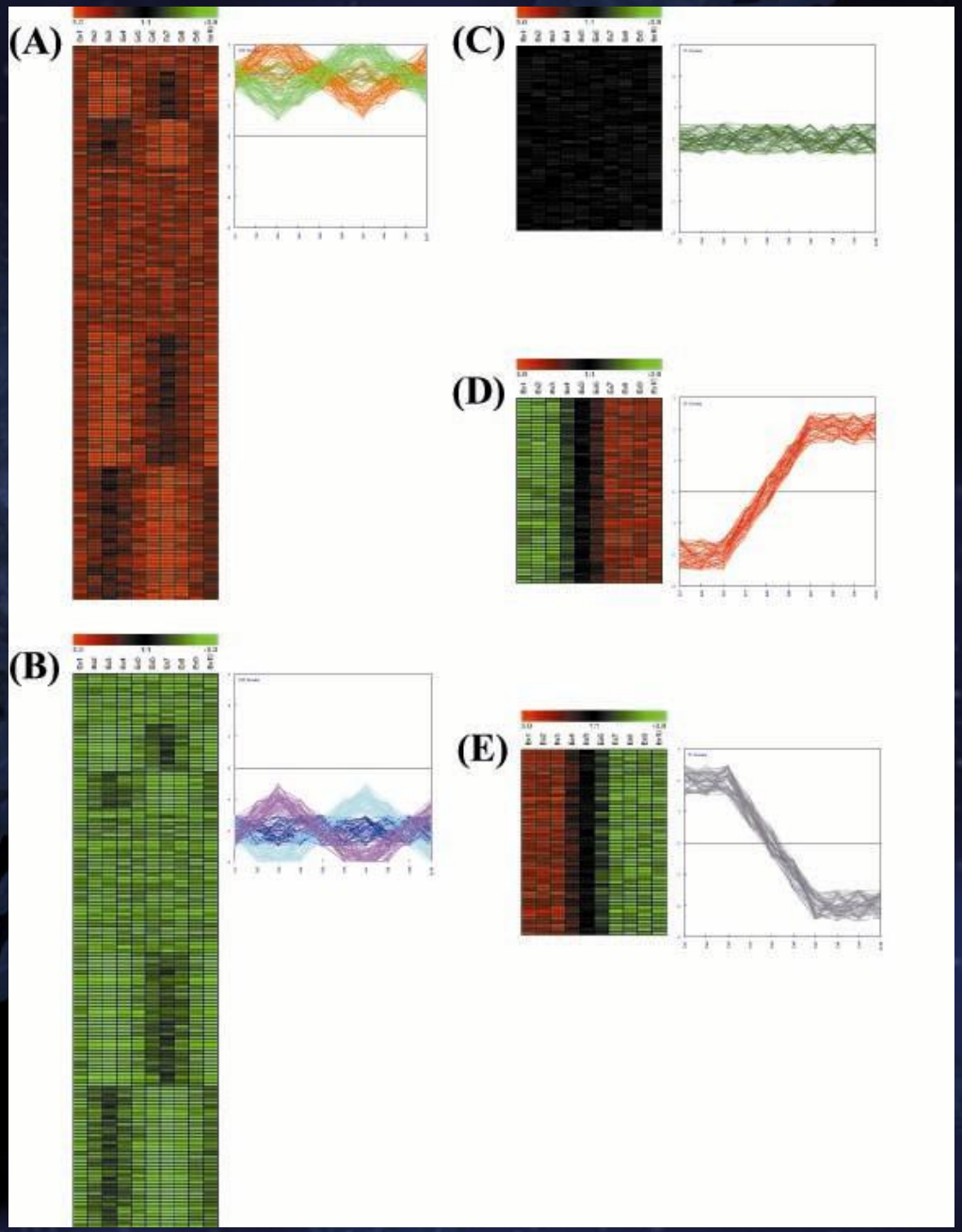



TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

# *k*-means Clustering

Separate clusters may have more or less support when using different algorithms.

Note colors are based on hierarchical clustering Results.

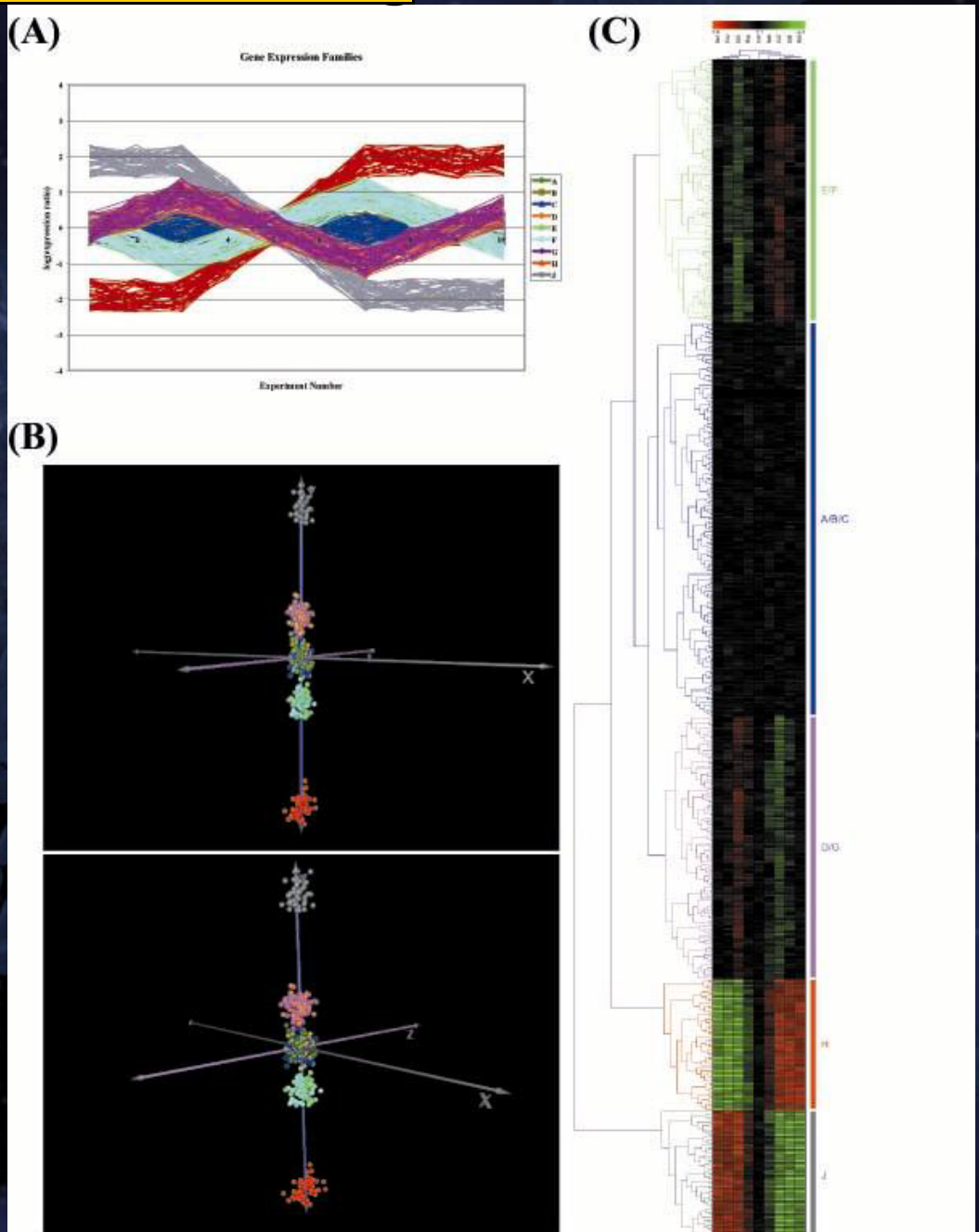



TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

# The effects on Mean Centering

(A) Expression Profiles
(B) PCA
(C) Hierarchical Clustering

Adjusting the data can have profound effects, but allow different patterns to be seen.




TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

# Very Useful Microarray URLs

| | |
|---|---|
| Leming Shi | http://www.gene-chips.com |
| TIGR | http://pga.tigr.org/tools |
| MGED | http://www.mged.org |
| Wentian Li | http://linkage.rockefeller.edu/wli/microarray |
| EBI | http://industry.ebi.ac.uk/~alan/MicroArray |
| Terry Speed | http://stat-www.berkeley.edu/users/terry/zarray/Html |
| Joe Derisi | http://www.microarrays.org/index.html |
| Pat Brown | http://cmgm.stanford.edu/pbrown/mguide/ |
| NCGR | http://www.ncgr.org/research/genex/other_tools.html |
| Stanford | http://www.dnachip.org |
| HAPI | http://array.ucsd.edu |

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

# Acknowledgments


**The TIGR Gene Index Team**

Jennifer Cho
Svetlana Karamycheva
Yudan Lee
Babak Parvizi
Geo Pertea
Razvan Sultana
Jennifer Tsai
John Quackenbush
Joseph White

**TIGR Collaborators**

Norman Lee
Renae Malek
Hong-Ying Wang
Truong Luu
Nnenna U. Nwokekeh

**TIGR Human/Mouse/Arabidopsis Expression Team**

Emily Chen
Renee Gaspard
Jeremy Hasseman
Heenam Kim
John Quackenbush
Erik Snesrud
Shiubang Wang
Ivana Yang
Yan Yu
Baoping Zhao

Funding provided by the Department of Energy and the National Science Foundation

**Array Software Hit Team**

Jerry Li
John Quackenbush
Alex Saeed
Vasily Sharov
Alexander Sturn
Joseph White

**Assistant**

Mary Mulholland

Funding provided by the National Cancer Institute, the National Heart, Lung, Blood Institute, and the National Science Foundation

TIGR Faculty, IT Group, and Staff


**Resources: <http://pga.tigr.org/tools.shtml>**


<johnq@tigr.org>


TIGR
THE INSTITUTE FOR GENOMIC RESEARCH