DETERMINING SIGNIFICANT FOLD DIFFERENCES IN GENE EXPRESSION ANALYSIS

A. J. BUTTE[1], J. YE[2], G. NIEDERFELLNER[3], K. RETT[3],
H. U. HÄRING[3], M. F. WHITE[2], I. S. KOHANE[1]

[1] *Children's Hospital Informatics Program,*
Boston, MA 02115, USA

[2] *Howard Hughes Medical Institute, Joslin Diabetes Center,*
Boston, MA 02115,USA

[3] *Department of Medicine, Universität Tübingen,*
Otfried-Müller-Straße 10
D-72076 Tübingen, Germany

A typical use for RNA expression microarrays is comparing the measurement of gene expression of two groups. There has not been a study reproducing an entire experiment and modeling the distribution of reproducibility of fold differences. Our goal was to create a model of significance for fold differences, then maximize the number of ESTs above that threshold. Multiple strategies were tested to filter out those ESTs contributing to noise, thus decreasing the requirements of what was needed for significance. We found that even though RNA expression levels appear consistent in duplicate measurements, when entire experiments are duplicated, the calculated fold differences are not as consistent. Thus, it is critically important to repeat as many data points as possible, to ensure that genes and ESTs labeled as significant are truly so. We were successfully able to use duplicated expression measurements to model the duplicated fold differences, and to calculate the levels of fold difference needed to reach significance. This approach can be applied to many other experiments to ascertain significance without *a priori* assumptions.

1 Background

1.1 Noise in expression measurements

Oligonucleotide microarrays currently allow the quantitation of expression of over 60,000 expressed sequence tags (EST) in a sample of RNA. A typical use for microarrays is the measurement of gene expression before and after an intervention, or the comparison of two groups. A fold difference for each gene is calculated by dividing its measurement in one group by its measurement in the other group. Expression can be measured using a two-dye microarray approach, where RNA from each of the two groups is labeled with a different color, then hybridized to a single microarray. (1, 2) Expression can also be measured in single-color microarrays, such as those available from Affymetrix.

Measurement noise can come from many theoretical and practical sources including, for example: varying microarray technology, nonspecific probes, intraprobe noise (from nonspecificity or differing concentrations of A/T), or biological noise (time of day for measurements). (3)

When RNA expression is measured using the same sample on two chips, correlation coefficients are commonly quoted as being high or near 1.0. Few studies have analyzed the reproducibility of these measurements. In a publicly available document, Incyte demonstrated high concordance between RNA expression measurements using Cy3 and Cy5 dye signals. Based on this, Incyte estimates that the limit of detection of fold differences is at 1.8, meaning 95% of fold differences between samples of 2.0 or higher are significant. (4)

There has been little other published data on reproducibility. Bertucci, et al., measured the expression of 120 genes in various cancer cell lines, using cDNA spotted filters. Close to 98% of the measurements showed less than a twofold difference when repeated. (5) Richmond, et al., studied differentially expressed genes in *E. coli* and filtered out genes under a minimum expression threshold as well as genes with less than a 5 fold difference. (6) Geiss, et al., used a Cy3/Cy5 system to measure genes differentially expressed during HIV infection. In their analysis, they determined that fold differences as little as 1.5 fold were statistically significant. However, this was determined to exclude 95% of the expression measurements seen, and not using an information-theoretic method. (7) Other publications citing differences between control and experimental groups as low as 1.7 fold continue to be published. (8)

To our knowledge, there has not been a study reproducing an entire experiment and modeling the distribution of reproducibility of fold differences.

2 Methods

2.1 Measurements of RNA expression

Steps needed to measure RNA expression levels using Affymetrix microarrays have been described previously. (9) Data was collected measuring RNA expression in muscle biopsies of four individuals. The overall goal here was to find the genes most significantly *different* between patients.

RNA was hybridized onto Affymetrix Hu35K microarrays. Expression levels for 35,714 ESTs across four microarrays were measured from each of the four persons. Duplicated measurements from the same samples were also made.

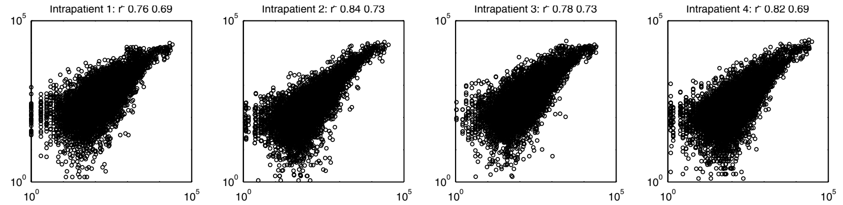

Figure 1: RNA samples from four humans were placed on duplicate GeneChips and the expression of 35,714 ESTs was measured. Each point represents an EST. Correlation coefficients were 0.69, 0.73, 0.73 and 0.69 when expression measures were transformed using the base-10 logarithm shown here.

2.2 Normalizing microarray scans and reproducibility of expression measurements

Measurements of the 35,714 ESTs using the four microarrays on the first patient were considered standard. The four microarrays measuring the three other patients were normalized to the standard by calculating a linear regression model and then multiplying the expression levels by the inverse slope of the linear model. The four duplicated microarray measurements for all four patients were also normalized to the same standard. Intrapatient fold differences (FD) were then calculated between the duplicated measurements for each of the four patients. Interpatient FD were calculated between all six possible pairs of patients, and duplicates of the interpatient fold differences were also calculated. The logarithm (base-10) fold differences (LFD) were used throughout this analysis, so that up and down regulation were represented equally.

2.3 Reproducibility of fold differences within and between patients

The correlation coefficients between all 35,714 repeated expression measures for ESTs measured in the four patients were 0.76, 0.84, 0.78 and 0.82. These correlation coefficients dropped to 0.69, 0.73, 0.73, and 0.69, respectively, when expression measures were transformed using base-10 logarithm (figure 1). This suggests that the wider splay of points seen at lower expression values worsens the correlation coefficient, and this splay is different between patients. However, we feel that the high correlation coefficients for duplicated measures were also due to bias by genes expressed at high levels; with such a large dynamic range of measurements, the fewer high values can overwhelm the pattern in the low measures.

When the interpatient LFD were calculated from these same expression measures *between* the six possible pairs of patients, the correlation coefficient for

LFD in the replicated measurements was very poor (figure 2), when almost all 35,714 ESTs were considered (those with any negative or zero expression value were already excluded, since these fold differences could not be calculated mathematically). Further analysis showed that the poor correlation coefficient in the replicated LFD was due to small expression values; when two small numbers (i.e. expression measures) were divided, it led to a high fold difference. This was particularly troublesome due to the more pronounced effects of noise on measurements at low expression levels.

Thus, we needed a strategy to filter out those ESTs with measurements that were contributing to the poor correlation between replicated LFD. However, we needed to determine the specific strategy without *a priori* knowledge or assumptions about this specific study. An overview of the data sources and types of calculated fold differences is shown in figure 3. The four intrapatient LFD were termed *goal-negative* because all of these should have equaled zero (i.e. there should have been no fold difference in the ESTs in the same patient). The six interpatient LFD were termed *goal-positive* because although few of the goal-positive LFD were non-zero, all of these should have equaled the repeated interpatient LFD.

Various strategies were used in a comprehensive manner to filter out those genes contributing to the noise. Models were created using the goal-negative and goal-positive LFD. Using these models, we determined the range of fold differences that could still be zero when replicated (the ranges of insignificance). Finally, the list of genes exceeding the range of both the goal-negative and goal-positive models was determined. An overview of this entire approach is shown in figure 4.

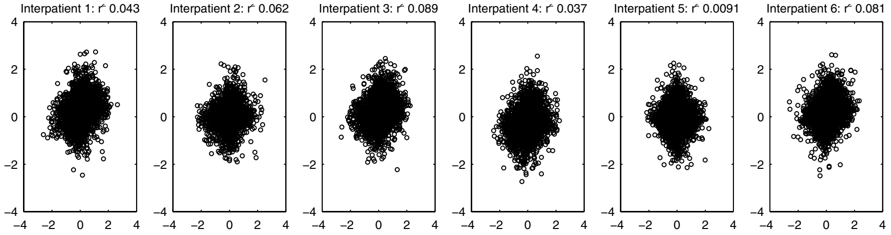

Figure 2: Fold differences of 35,714 ESTs were calculated between the six possible pairings of the four patients. Fold differences are expressed in logarithm base-10, so that ESTs that did not change between models are plotted in the center of each graph. The calculated fold differences from the duplicated measures are shown on the x- and y-axes. Even though the correlation coefficients were high between original and repeated expression values, the correlation coefficients were very low between original and repeated calculated fold differences.

2.4 Determining threshold fold differences where intrapatient fold-differences should have been zero (goal-negative)

Every gene had four calculated intrapatient LFD. Since the same samples were used in the duplicated measurements, all the intrapatient LFD should have been zero (i.e. fold difference of one). Instead, we found a bell-shaped distribution of LFD around zero. We calculated *threshold log fold differences* (TLFD), or the smallest and largest LFD that should have been zero. Together, these high and low thresholds define a range of LFD, called a *range of insignificance*. Specifically, the range of insignificance encompasses the LFD of 95% of the ESTs. Operationally, this means that if a new EST's log fold difference is inside this range, it is too close to zero, and could actually have been zero. When a range of insignificance is calculated, each EST can then be evaluated individually to determine whether its fold difference is significantly different than zero. The TLFD were empirically found at the 2.5th and 97.5th percentile of the bell-shaped distribution.

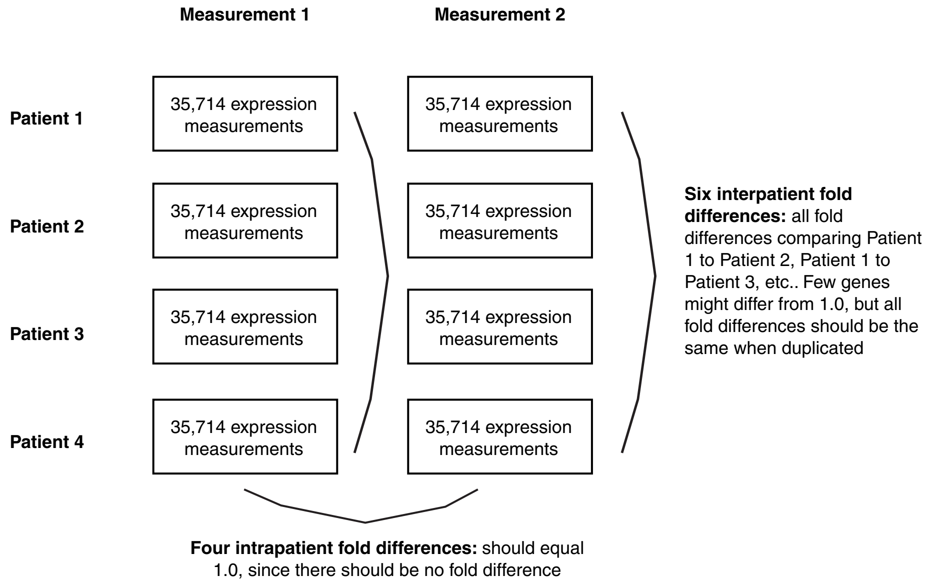
Measurement 1
Measurement 2
Patient 1
Patient 2
Patient 3
Patient 4
35,714 expression measurements
35,714 expression measurements
35,714 expression measurements
35,714 expression measurements
35,714 expression measurements
35,714 expression measurements
35,714 expression measurements
35,714 expression measurements
Six interpatient fold differences: all fold differences comparing Patient 1 to Patient 2, Patient 1 to Patient 3, etc.. Few genes might differ from 1.0, but all fold differences should be the same when duplicated
Four intrapatient fold differences: should equal 1.0, since there should be no fold difference

Figure 3: Data source. 35,714 gene expression measurements were made in duplicate in muscle samples from 4 patients. The four intrapatient fold differences should equal 1, since the measurements should have been equal (this assertion was used to make the *goal-negative* model, referred to in the text). The four intrapatient fold differences should equal the duplicated fold differences (used to make the *goal-positive* model).

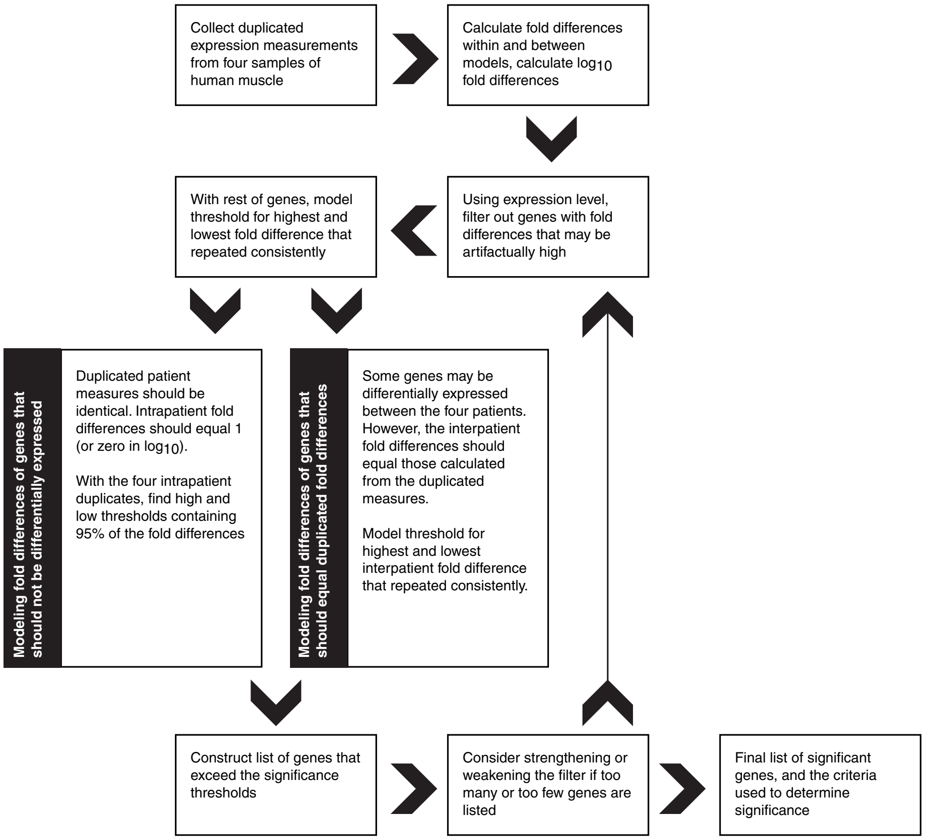
Collect duplicated expression measurements from four samples of human muscle
Calculate fold differences within and between models, calculate log10 fold differences
Using expression level, filter out genes with fold differences that may be artifactually high
With rest of genes, model threshold for highest and lowest fold difference that repeated consistently
Modeling fold differences of genes that should not be differentially expressed
Duplicated patient measures should be identical. Intrapatient fold differences should equal 1 (or zero in log10).
With the four intrapatient duplicates, find high and low thresholds containing 95% of the fold differences
Modeling fold differences of genes that should equal duplicated fold differences
Some genes may be differentially expressed between the four patients. However, the interpatient fold differences should equal those calculated from the duplicated measures.
Model threshold for highest and lowest interpatient fold difference that repeated consistently.
Construct list of genes that exceed the significance thresholds
Consider strengthening or weakening the filter if too many or too few genes are listed
Final list of significant genes, and the criteria used to determine significance

Figure 4: Strategy to filter out those ESTs with measurements that were contributing to the poor correlation between replicated LFD. Gene expression measurements were performed in duplicate, then fold differences were calculated for the original and repeated data set. Various strategies were used in a comprehensive manner to filter out those genes contributing to the noise. Using the goal-negative and goal-positive LFD, models were created to determine ranges of insignificance. Finally, the list of genes exceeding this range was determined, and the process repeated for all strategies.

2.5 Determining threshold fold differences where interpatient fold-differences should have been equaled the duplicated fold-differences (goal-positive)

Every gene had six interpatient LFD (i.e. the six possible pairings of four patients). Since the same samples were used in the duplicated measurements, the expectation was that each LFD would be equal to the LFD from the duplicated measurements. In reality, each LFD had a confidence interval, such that the duplicated LFD could have been larger, smaller or even zero when replicated. However, we found that the greater the LFD of a gene was from zero, the less likely that the replicated LFD was zero. For the interpatient LFD, we developed a statistic to choose *threshold log fold differences* (TLFD), which are the smallest LFD that are significantly likely to actually be differentially expressed. Together, the high and low thresholds define another *range of insignificance*. Similar to the previously defined range, an EST with an LFD inside this range is too close to zero and could actually be zero when replicated.

The method of determining the goal-positive TLFD is shown in figure 5. We created a linear regression model fitting the original and duplicated LFD with the equation $y = mx + b$, where x represents the original LFD, y is the duplicated LFD, m is the slope of the regression line, and b is the y-axis intercept. We then calculated the standard deviation of the differences of actual y from predicted y, using

$$LMSD = \sqrt{\sum \frac{(y - y')^2}{n}}$$

where y is an actual replicated LFD, y' is the predicted LFD for that same x using the regression model, and n is the number of duplicated points.

Based on this model, we were able to calculate the high and low significance thresholds. The high threshold was defined as

$$TLFD_{high} = \frac{(2 \cdot LMSD - b)}{m}$$

and the low threshold was defined as

$$TLFD_{low} = \frac{(-2 \cdot LMSD - b)}{m}$$

In other words, if a gene showed a fold increase greater than the high TLFD, it was significantly likely to still have a fold increase when the experiment was repeated.

Once both the goal-negative and goal-positive ranges of insignificance were known for a particular strategy, we counted the number of genes with at least one interpatient LFD outside both insignificance ranges and viewed these as significant. The goal was to maximize this count using the various combinations of strategies.

2.6 Strategies to improve the significant threshold fold differences

Once the ranges of insignificance were known, our goal was to maximize the number of genes with interpatient fold differences outside this range. There were two ways to do this (1) either eliminate the ESTs contributing to the noise, thus

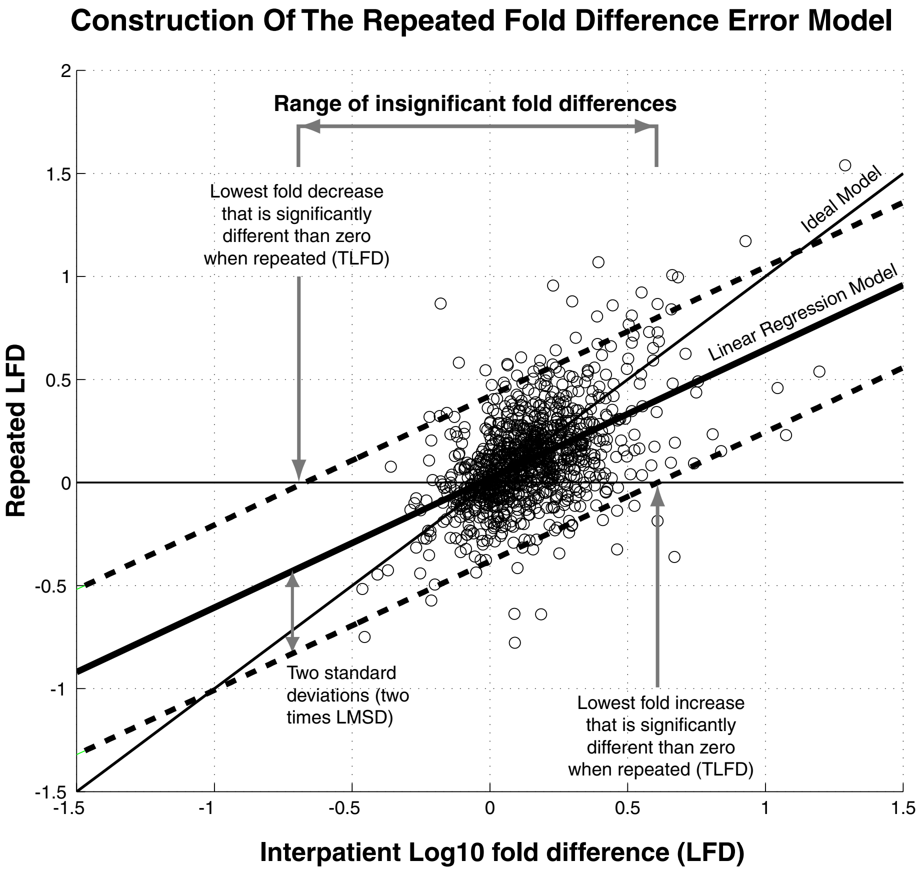

Figure 5: Finding the range of insignificance using the goal-positive LFD. Each point is an EST. Fold difference between two patients is on the *x*-axis, and fold difference using duplicated measurements is on the *y*-axis. The linear regression model is shown in the bold diagonal line, and the thin diagonal line is the ideal noise-free model ($x = y$). The distance of each point from the regression model was calculated, and the standard deviation was calculated (LMSD). The dashed lines are the regression line shifted by two times LMSD. The *y*-intercept of the dashed lines are the high and low threshold fold differences for significance (the high and low TLFD). The range between the high and low TLFD is the range of insignificance. An LFD in this range is too likely to be zero when repeated.

reducing the TLFD and allowing more ESTs to fall outside the range, or (2) include as many ESTs as possible, including those that may fall outside the range. In other words, both adding and decreasing the number of ESTs could improve the number of ESTs falling outside the insignificance range.

There were two strategies to limit the ESTs outside the range of insignificance:

- Limit ESTs to those with Affymetrix "present" calls
- Limit ESTs to those exceeding a defined minimum expression

Our methodology was to try all possible combinations of strategies to determine the best way to find the largest number of ESTs outside the range of insignificance. This way, there were no *a priori* assumptions as to the list of ESTs being chosen.

For the first strategy, we considered choosing ESTs based on the Affymetrix "A" and "P" calls from the scan. Affymetrix assigns "absent" and "present" calls during its quantitation algorithm. ESTs assigned a "P" demonstrated improved average contrast between perfectly-matching and mismatching probes, and thus have a tighter confidence interval around the quantitation. ESTs assigned an "A" have a weaker confidence interval. There were three possible sub-strategies:

- Choosing ESTs with a "P" call in all patient chips
- Choosing ESTs with a "P" call in at least one patient chip, and
- Ignoring "A" and "P" calls.

For the second strategy, we had found many of high LFD were artifacts resulting from the division of two small expression measures. We considered using a minimum expression level threshold (MELT) and eliminating those ESTs not meeting this minimum threshold. Again, there were three possible sub-strategies:

- Choosing ESTs meeting the MELT in all patient chips
- Choosing ESTs meeting the MELT in at least one patient chip, and
- Ignoring the MELT.

In addition, there were also many possible expression level thresholds to try.

All nine possible combinations of sub-strategies were tried. For those strategies involving the MELT, each MELT from 0 to 3000 was tried, in increments of 100.

3 Results and Discussion

3.1 Findings

Our goal was to create a model of significance for fold differences, then maximize the number of ESTs above that threshold. This was done by creating strict criteria to filter out those ESTs contributing to noise, thus decreasing the requirements of what was considered an insignificant fold difference. All nine combinations of criteria strategies were tested including using minimum expression level thresholds and the Affymetrix absent and present calls. The results of the testing are shown in the table. In each of the nine criteria strategies, models of insignificance were created. In seven of the nine strategies, ESTs were present that exceeded the range of insignificance. Two strategies, that of choosing all ESTs, and that of choosing ESTs on based on an Affymetrix "present" call on any one chip, failed to produce a list of significant ESTs because the measurement noise was so high, the range of insignificance was so wide that none of the ESTs exceeded the range.

For this data set, the largest number of ESTs meeting or exceeding significance occurred when either ESTs were removed if (1) expression levels were under 100 in all patients and duplicates, or (2) there was no "present" call on any chip. With this strategy, the correlation coefficient between interpatient LFD and duplicates was 0.61. Because of the high correlation coefficients, the interpatient range of insignificance was only 0.5 to 1.7 fold (in logarithm base-10: -0.33 to 0.22). Stated another way, and taking the asymmetry into account, genes with half the expression level in one patient compared to another were not significantly different than zero. Genes beyond this range were, however, significantly different than zero.

With this strategy, 389 ESTs had at least one interpatient LFD that exceeded both the goal-negative and goal-positive ranges of insignificance. This strategy could be viewed as having the highest sensitivity, in that using this strategy produced the largest number of significant ESTs., and highest specificity, in that the ESTs with no fold difference were maximally contained in a range of insignificance.

3.2 Advantages of Approach

There were two critical findings in this experiment. First, even though RNA expression levels appear consistent in duplicate measurements, when entire experiments are duplicated, the calculated fold differences are not necessary as consistent. Thus, it is critically important to repeat as many data points as possible, to ensure that genes and ESTs labeled as significant are truly so.

Second, we were successfully able to use duplicated expression measurements to model the duplicated fold differences and to calculate the levels needed to reach significance. Without *a priori* knowledge, we were able to comprehensively try multiple strategies to limit ESTs so that the duplicated fold differences would be more consistent. Dropping these noisier ESTs decreased the required levels of significance, permitting more ESTs to exceed the significance thresholds.

Table (next page): Nine combinations of strategies were used to filter out ESTs with measurement noise, to reduce the requirements of statistical significance and allow more ESTs to meet or exceed these requirements. Each strategy is listed in a column. Six of the nine strategies involved varying a minimum expression level threshold, which are listed in the rows. For each strategy, ranges of insignificance were calculated, such that an EST that demonstrated an interpatient fold difference within this range was still likely to be zero when measurements were repeated. The largest number of ESTs exceeding insignificance is highlighted.

Strategies to limit ESTs contributing to poorly correlating replicated fold differences									
No min expression level threshold	X			X			X		
ESTs must reach min expression level on at least one patient chip		X			X			X	
ESTs must reach min expression level on all eight patient chips			X			X			X
ESTs may have "A" or "P" calls	X	X	X						
ESTs must have "P" calls on at least one patient chip				X	X	X			
ESTs must have "P" calls on all eight patient chips							X	X	X
Minimum Expression Level Threshold	**Number of ESTs with significant fold difference using this strategy and threshold**								
0	0	0	2	0	0	3	363	363	363
100	0	0	337	0	3	389	363	363	367
200	0	4	354	0	8	359	363	363	284
300	0	12	243	0	12	254	363	367	200
400	0	16	156	0	29	157	363	366	137
500	0	31	93	0	56	93	363	369	78
600	0	55	50	0	87	50	363	368	50
700	0	84	48	0	136	49	363	384	55
800	0	103	25	0	161	25	363	386	41
900	0	112	19	0	161	19	363	380	29
1000	0	136	14	0	176	14	363	373	27
1100	0	158	10	0	186	10	363	373	18
1200	0	182	6	0	232	6	363	365	7
1300	0	226	5	0	252	5	363	355	4
1400	0	269	4	0	286	6	363	344	4
1500	0	283	2	0	299	3	363	331	1
1600	0	285	1	0	291	1	363	327	0
1700	0	272	2	0	298	2	363	319	1
1800	0	249	1	0	272	1	363	309	1
1900	0	287	0	0	308	0	363	308	0
2000	0	278	0	0	299	0	363	295	0
2100	0	261	0	0	275	0	363	286	0
2200	0	251	0	0	272	0	363	282	0
2300	0	261	0	0	280	0	363	274	0
2400	0	274	0	0	285	0	363	269	0
2500	0	257	0	0	270	0	363	266	0
2600	0	267	0	0	270	0	363	258	0
2700	0	258	0	0	256	0	363	250	0
2800	0	235	0	0	235	0	363	240	0
2900	0	250	0	0	248	0	363	229	0
3000	0	237	0	0	235	0	363	227	0

Unlike other methods of significance determination, this analysis makes no assumptions as to the distribution of expression measurements in these data sets. In other words, significance was only a function of the reproducibility of measures and calculated fold differences. It was not a function of the measurements themselves (e.g. significance was *not* set at the top 5%ile of genes by expression level).

3.3 Future Directions

This approach will be applied to many other data sets, to ascertain whether the specific strategy parameters found to be optimal in this experiment are also optimal in others. Further work will also be done on applying this technique to data sets collected under alternate microarray techniques. Other non-linear methods could be also be used to model duplicate data and set significance thresholds.

Acknowledgments

This research was supported in part by the grant "Research Training in Health Informatics" funded by the National Library of Medicine, 5T15 LM07092-07 and R01 LM06587-01, and in part by the Endocrine Fellows Foundation.

References

1. M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray" *Science*. 270:467-70 (1995).
2. G. Ramsay, "DNA chips: state-of-the art" *Nat Biotechnol*. 16:40-4 (1998).
3. J.G. Hacia, "Resequencing and mutational analysis using oligonucleotide microarrays" *Nat Genet*. 21:42-7 (1999).
4. Anonymous. GEM Microarray Reproducibility Study. Incyte Pharmaceuticals, Inc. (1999).
5. F. Bertucci, et al., "Expression scanning of an array of growth control genes in human tumor cell lines" *Oncogene*. 18:3905-12 (1999).
6. C.S. Richmond, J.D. Glasner, R. Mau, H. Jin, and F.R. Blattner, "Genome-wide expression profiling in Escherichia coli K-12" *Nucleic Acids Res*. 27:3821-35 (1999).
7. G.K. Geiss, et al., "Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays" *Virology*. 266:8-16 (2000).
8. C.K. Lee, R. Weindruch, and T.A. Prolla, "Gene-expression profile of the ageing brain in mice" *Nat Genet*. 25:294-7 (2000).
9. P. Tamayo, et al., "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation" *Proc Natl Acad Sci U S A*. 96:2907-2912 (1999).