Deposited research article

ResurfP: a response surface aided parametric test for identifying differentials in GeneChip based oligonucleotide array experiments

Suresh Gopalan

Addresses: 3207 Stearns Hill Road, Waltham, MA 02451, USA.

Correspondence: Suresh Gopalan. E-mail: gopalans2@hotmail.com

Posted: 28 September 2004

Genome ***Biology*** 2004, **5:**P14

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/11/P14

Received: 17 September 2004

This is the first version of this article to be made available publicly.

ResurfP: A response surface aided parametric test for identifying differentials in GeneChip based oligonucleotide array experiments

Suresh Gopalan*

Independent Investigator, Waltham, MA 02451, USA.

*Corresponding author: Phone/Fax: (781) 893-9065;

email: gopalans2@hotmail.com.

Running Title: Response surface aided Parametric test

Submitted to Genome Biology

09/17/2004

Keywords: ResurfP: Response surface assisted parametric test; ROC: Receiver operating characteristics; microarray; probe-level analysis; differential gene expression.

ABSTRACT

Background

Transcripts in a GeneChip type microarray is represented by multiple independent short oligonucleotide probes. One widely used approach is to compute a model based unified expression index for the transcript which is subsequently used for comparative data analysis. Alternative approach is to analyze the data at the probe-level. A good understanding of the effect of the number of probe-pairs included at different statistical threshold used for selection should aid optimal selection of differentials. A test dataset with known differentials was used to study this property in comparisons involving two datasets.

Results

A response surface was plotted by formulating an equation that captures the effect of varying threshold of probe-pairs and t-statistic on true positives and false positives identified. The resulting response surface indicate that a wide range of probe-pair and t-statistic combinations yield comparative results. The toplology of the surface was used to define one form of additive cost-based approach - involving t and number of probe-pairs used - to determine the optimum threshold to achieve a good balance of true positives and false positives when comparing two datasets at the probe-level. In addition a data scaling approach was used to study the impact of a selected threshold on the number of false negatives of differing magnitude of differentials in a given dataset.

Conclusions

The results indicate that this response surface assisted approach (termed ResurfP) would be effective in determining optimal data-specific threshold for number of probe-

pairs used and of the t-statistic when analyzing differentials between two datasets using probe-level data.

BACKGROUND

The recent availability of complete genome sequences of a number of organisms and the development of powerful microarray technologies [1-3] allow the determination of the comparative expression levels of all the genes in a cell, tissue or organism. A common paradigm is to compare the transcriptome patterns of two or more experimental treatments or biological backgrounds (e.g., mutant versus wild type, or infected versus control tissue) using two to several replicates in each data set at one or more time points with one or more replicates.

Two fundamental variations of microarray technologies are in common use. The first being represented by reasonably long large PCR fragments of the transcript of interest and the other using oligonucleotides representing regions of the transcript. One version of the latter technology that is in widespread use is GeneChip (Affymetrix, CA). In this version of the technology each transcript is represented by eleven or more oligonucleotides 25 nucleotides long and each of these also have another corresponding oligonucleotide with a mismatch in the exact middle nucleotide to account for non-specific hybridization. The chips are hybridized with labeled cRNA representative of all the transcripts at a given point of time in the organism/tissue/cells. In principle, having perfect match and mismatch probes together with multiple probes representing each transcript should aid selective and sensitive identification of differential expression between two conditions being compared. These same features also add to significant degree of technical complexity. For example, physico-chemical features of sequences under a given hybridization condition, differing kinetics of hybridization, lead to differing signals for sequences representing the same transcript. In addition, cross-hybridization e.g., due to regions that are not sequenced in an organism, and lack of hybridization of certain probes make certain probe-pairs unusable thus reducing the effective usable

number of probe-pairs. A common practice is to reduce the complexity of the multiple probe-pairs representing a probeset or transcript by extracting a single expression index after using an appropriate normalization technique [e.g., 4-6]. Statistical methods are applied to the expression index to identify differentials and reduce false discovery rate. The quality of most downstream numerical analyses (clustering, etc.) and biological interpretations depend on the sensitive and selective identification of differentially regulated genes. Many new measures and adaptations of statistical tests are constantly being proposed with varying degrees of success and none being accepted as a most effective approach yet.

Some of these limitations can be overcome by directly dealing with probe level data rather than a summary expression index. This is complicated due to the same reasons highlighted above and sometimes due to computational cost involved with such an approach. Here a response surface approach together with a cost factor - comprising the number of valid probe-pairs and the t statistic from Student's t-test [7] - is proposed to identify dataset dependent threshold, to apply statistics to probe level data that would aid sensitive and selective identification of differentials.

METHODS

The GeneChip expression data set used in these analyses is from the Affymetrix dataset released for purposes of algorithm development, and based on HG-U133A-Tag arrays Experiments 2 through 5, replicates R1 through R3 (http://www.affymetrix.com/support/technical/sample_data/datasets.affx). This data set was generated using a hybridization cocktail consisting of specific RNA spike-ins of known concentration mixed with total cRNA from HeLa cell line, by Affymetrix. All probe sets starting with AFFX not part of the spike-ins of known concentration were removed

for calculation of true and false positives involving spike-ins, since some of them had obviously discernible differences. Three probesets were reported to have perfect homology of 5 or more probe-pairs. Thus leaving 45 true positives and 22,185 false positives for each comparison in the dataset. Unless mentioned otherwise, values represented are based on average of three comparisons between experiments differing in spike-ins with two fold difference in concentration viz., experiments 2 with 3, 3 with 4 and 4 with 5. Probe level data were extracted from Cell files (using tiling coordinates defined by probesequence information supplied for the chip type – U133A-Tag by Affymetrix) and the mean of all signal values (of perfect matches and mismatches that were between the value 28 (the lowest background in the chips used) and a saturation value of 46,000) were scaled to target value of 500, i.e.,

$$x_i = [(\sum_{i=1}^{n} (x_{pi} - b \mid 0 <= (x_{pi} - b) <= 46{,}000) + \sum_{i=1}^{n} (x_{mi} - b \mid 0 <= (x_{mi} - b) <= 46{,}000))]/(N_p+N_m) * (1/500) * (x_{pi}-x_{mi}) \quad \textbf{[1]}$$

with i representing each probe-pair, n the total number of probe-pairs in that arrray, x_p and x_m are the intensity values for perfect matches and mismatches, respectively, N_p and N_m are the number of perfect and mismatches satisfying the conditions in the first term of the equation, and b is the background of that chip (as determined by Microarray Siute 5.0). When more than 11 probe-pairs represented a probeset only the first 11 (in their order of listing in Affymetrix probesequence file) were extracted and used. The difference between perfect match and mismatch value for each probe-pair was used for all further evaluations. Zero or negative differences were set to background.

The signal values were extracted using Microarray Suite 5.0 (Affymetrix, CA) with the trimmed mean (top and bottom 2% signal values are trimmed) for each array scaled to a

target intensity of 500, for representation in Figure 3. Standard definitions for sensitivity and positive prediction value (PPV) were used. Sensitivity was calculated as sn = TP / (TP + FN); PPV was calculated as: PPV = TP /(TP + FP), where TP is true positives, FP is false positives, and FN is false negatives. Weighted average for t was calculated as:

$$\bar{t} = \sum_{j=1}^{m} [\lambda_j t_j \mid \lambda_j = (1/\sigma_j^2) / \sum_{j=1}^{m} (1/\sigma_j^2)] \qquad \textbf{[2]}$$

where j runs over probe-pairs, m is the number of probe-pairs used, and σ is the standard deviation of t over selected probe-pairs.

For the preliminary evaluation on biological replicates, the data from human patients with aortic stenosis (samples JB-as_0806, JB-as_1504 and JB-as_1805 were compared against JB-as_2111, JB-as_2604 and JB-as_2708, hybridized to U75-Av2 chips), from Genomics of Cardiovascular Development, Adaptation, and Remodeling site. NHLBI Program for Genomic Applications, Harvard Medical School. URL: http://www.cardiogenomics.org [*accessed 28 May, 2004*]. This chip consisted of 16 probe-pairs for most transcripts and the average background was used as 60. Calculations were performed using C++ on MS-Developer environment in Windows XP background.

RESULTS

Typical analysis of GeneChip data for identification of differentials between datasets involve extraction of the probe level data using an unified expression index signifying the estimated level of expression of that transcript summarizing the information in the eleven or more probe-pairs, following normalization or scaling. Some common methods used for this purpose are dCHIP [4], RMA [5] and MAS (Microarray Suite, currently version 5,0, Affymetrix, CA). The use of unified expression index is advantageous in terms of computational simplicity and easy adaptation of statistical methods to high dimensional datasets. But, due the extremely variable behavior inherent to each probe representing the transcript the unified expression index do not always perform satisfactorily. Consequently, statistical approach to reduction of false positives based on ordered statistics or other Bayesian approaches does not satisfactorily address the issue of false positives. This aspect has recently been evaluated for a few test datasets such as the one used in this article [8]. While improvements in the aforementioned aspects are constantly being proposed, statistics applied directly to probe-level data is an attractive alternative. As discussed earlier, several biological and sequence related issues complicate simple selection of a statistical threshold such as a p-value when using the Student's t-test. The following approach is motivated by the fact that the multiple independent features measured signifying the expression level of a transcript should in principle allow selection of a threshold that is appropriate to the noise in a particular dataset. In many well behaved dataset this threshold should be lower than a commonly acceptable threshold, e.g., t signifying $p <= 0.05$.

A Response surface model involving probe-pair number and statistical threshold

In order to study the performance of differential expression measured at probe level the response surface of sensitivity, positive prediction value, number of true positives and

number of false positives were evaluated as a function of number of valid probe-pairs and a range of values for t (the Student's t statistic). This was done with triplicate datasets that had spike-ins of two fold difference with different probesets in concentration ranges (0 – 512 pM) between the two datasets. A valid probe-pair was defined as one that has a minimum difference of average signal value (difference between signal for perfect match and mismatch) above background, and the ratio of averages is at least 1.1 (selected intuitively, but can be determined empirically for different datasets) and above threshold t, to avoid values in very close range. In addition, a condition that there are no more than one-fifth the probesets that had change in opposite direction was enforced. In general this latter condition was never a determining factor in selection of differentials in this dataset. This selection criteria for can be expressed as:

$$[\sum_{i=1}^{m}(n \mid t >= t', x_{ie}/x_{ib} >= 1.1, (x_{ie} - x_{ib}) >= b)\ ] >= np \qquad \textbf{[3]}$$

where n is the number of probe-pairs satisfying the conditions, t' is the threshold for t statistic, np is the threshold for number of valid probe-pairs, x_{ie} and x_{ib} is the signal value for probe-pair i, in experimental and baseline chips, respectively. The above equation represents selection of probesets where the chip designated the experimental chip has higher value than the chip designated the baseline chip, the equation for probesets with value for baseline chip higher can be obtained by interchanging x_{ie} and x_{ib}. For example for a probeset that satisfies the threshold of 6 valid probe-pairs and t value of 7.0, at least 6 probe-pairs representing that probeset will individually have a t-statistic of 7.0 or above - all having the same direction of change. As can be seen from Figure 1A, and as expected, with increasing threshold of t and probe-pair threshold the positive prediction

value (PPV) increases i.e., a decreasing number of false positives are identified and sensitivity decreases i.e., lesser number of true positives are identified as differentials. Figure 1B, shows the decrease of true and false positives with increasing threshold of t and np.

Identification of optimal threshold

The above problem can in principle be viewed as area under the Receiver operating characteristic (ROC) curve problem [9] with two dimensions t threshold as one dimension and number of valid probe-pair number as another dimension. In this kind of situation, one would expect multiple thresholds involving the two dimensions that would have optimal area under the ROC curve. Alternatively, this can be viewed as an optimization problem with the goal of detecting as many true positives with optimal combination sensitivity and positive prediction value. In other words this can be written mathematically as, termed effective number of positives identified (N_{eff}):

$$N_{eff} = TP * TP/(TP + FP) * (1 – FP/TP) \quad \textbf{[4]}$$

Figure 2A shows the response surface of this effective number of positives as a function of t and number of valid probe-pairs (np). It can be seen from the figure that a range of t and np can result in comparable N_{eff}, with top two N_{eff} at (t',np) of (7,5) and (6,6) with (true positives, false positives) of (91,1),(89,1) and (87,0), respectively. The total possible number of true positives and false positives were 135, and 66,555, respectively. It should be noted that the lowest differential (two fold) was used from the dataset, higher differentials would lead to identification of higher number of true positives. The presence of a large portion of the surface across a range of t and np having similar N_{eff}

in Figure 2A suggests that it would be possible to achieve good sensitivity and selectivity for many np and t values thus potentially increasing the sensitivity of detection of small differentials and differentials in transcripts expressed at low levels. This can be achieved in principle by defining a cost factor consisting of the two parameters being tested. One form of defining such a cost adjusted effective number of positives picked (CAN_{eff}) would be:

$$CAN_{eff} = N_{eff} /(t' + np) \quad \textbf{[5]}$$

The response surface for CAN_{eff} as a function of t' and np is shown in Figure 2B. It can be seen from the surface of CAN_{eff} (Figure 2B) that the largely flat area near the peak of N_{eff} (in Figure 2A) can now be reduced to a few distinct and narrow peaks. The (t',np) values yielding the top three CAN_{eff} are (3,7), (4,6) and (4,7) with (true positives, false positives) (86, 2), (91,5) and (85,0), respectively. It should be highlighted that these values of true and false positives selected at this threshold are comparable to that of the maximum Neff mentioned before. For comparison, at t signifying p<= 0.05 and a threshold of six valid probesets the (true positives,false positives) was (85,0). The number of true and false positives identified and the concentration range of the spike-in positives for a selected set of t' and np values are summarized in Table 1.

The possibility of selecting a lower threshold and still being able to maintain high selectivity would especially be of interest (i) with certain datasets where there is a large increase in positives with a small reduction in threshold, whereas the training dataset indicative of variability in the experiment suggest that this would result in a very small number increase in selection of false positives, and (ii) for sensitive identification of small

differentials without significant loss of selectivity (illustrated in the next section with some test cases).

Evaluation of the threshold determined by ResurfP

The methodology outlined above is termed ResurfP, for Response surface assisted Parametic test. It can immediately be reckoned that lower the threshold that can give good selectivity, the better it is to select small differentials and differentials in transcripts with low expression levels. Thus, the advantage of the lowered threshold were evaluated by scaling one of the two datasets (i.e., the probe level data extracted as outlined in Methods section) used in above comparison to varying extents (1.5, 2, 3 and 4 fold) and comparing to the other dataset. This should allow comparison of data classes with wider variety of variances as opposed to a few signified by the spike-ins. Further, this should also reveal the sensitivity of the methodology in the context of technical replicates, thus revealing the maximum achievable sensitivity. The results for this evaluation at the thresholds yielding the top two CAN_{eff}, t signifying $p <= 0.05$, and the threshold specifying the top N_{eff} are represented in Table 2. As expected, the lower thresholds lead to higher sensitivity of detection at any given level. It should be noted that even at the lower threshold (t', np) of (3,6), the differentials (average of three comparisons compared to maximum identifiable differentials defined below) identified were only 42%, 61%, 81% and 86% of 1.5, 2, 3 and 4 fold respectively, which further emphasizes the need for and importance of the proposed approach. At a threshold of (7.71, 6) these values were significantly lower viz., 30%, 47%, 63% and 70%, respectively. For the purpose of calculating percentage of differentials identified the maximum identifiable differentials was set at 21,485, which is the differentials (average of three comparisons) identified at the threshold of (t' = 4, np = 5) with a scaling factor of 10. A steep decline

face on the surface of Figure 2B (right hand side) with increasing probe-pair threshold together with results indicated in Table 2 also indicate a higher penalty for increasing the probe-pair threshold than for increasing t statistic threshold. Additionally, these data indicate that an appropriate choice of a lower probe-pair threshold can lead to significantly higher number of true differentials without concomitant increase in false positives.

In order to have a preliminary idea of the nature of probesets/transcripts that are selected and are missed in this study, the distribution of the expression indices (to simplify the representation) of these probesets for one of the thresholds (t',np) of (3,7) is shown in Figure 3A. As can be seen from this figure and as expected the distribution of the expression indices of probesets, low expressors are detected better at higher differential ratios. Conversely, almost all the probesets missed at higher differential ratios were low expressors, which is consistent with observations that there is high variability in the low detection ranges.

The optimal application of ResurfP on biological samples with different properties need additional testing with an independent confirmation using another technology. Nevertheless, the results of a preliminary evaluation to test if the lower threshold identified by ResurfP would lead high false positives when tested on biological replicates are very encouraging. For this purpose (t',np) thresholds of (3,6) and (3,8) were tested on one set of biological replicates from cardiogenomics website (see methods). For this purpose, data from six human patients with aortic stenosis were split into two groups (of triplicates) and the method was evaluated. This lead to identification of only 52 and 21 of 12,624 probesets at (3,6) and (3,8), respectively, even though this chip type consisted of 16 probe-pairs for most probesets/transcripts.

Ranking differential genes identified

Another useful parameter will be to rank the genes in order of significance. For this purpose, the product of the number probe-pairs contributing to selection and the weighted average of the statistic t of those probe-pairs was evaluated. The resultant ranking index was used to order the probesets with ranks decreasing with ranking index. The results shown in Figure 3B indicated a tendency of probesets with higher expression levels to have higher rank, indicating higher reliability at higher signal intensities.

DISCUSSION

Studies on genome-wide analysis of transcripts is becoming increasingly popular, primarily due to availability of genome sequence of large number of organisms and technologies that permit arraying and probing sequences representing transcripts/other genomic regions. One primary problem with this increasing trend of large scale data generation and analysis is careful control of data quality at each stage of workflow (viz., data generation, first stage analysis to select genes/transcripts of interest), which has direct impact on the quality of all downstream analysis, hypothesis generation and testing.

Two common microarray platforms are widely in use, one representing long PCR fragments representing transcripts, or more recently long oligonucleotides to improve specificity and cross-hybridization and the other, a set of eleven or more probe-pairs each having a perfect match and a mismatch representing each transcript. Influenced by a variety of factors including intrinsic nature of the probe sequence, kinetics and efficiency of labeling and hybridization, labeled transcripts hybridize with varying degrees

of specificity and efficiency, thus yielding varying signal levels even for probe-pairs representing the same transcript. A commonly used approach is the use of an estimated signal measure that represents the summary of the signals from multiple probe-pairs, called the expression index. Different normalization schemes and models are used to achieve this index. Statistical and other data selection rules are applied to this index. An attractive and powerful alternative is to apply statistics directly to probe level data. In this article, an algorithm for the determination of threshold for identification of differentials between two datasets using analysis of probe level data from GeneChip type microarrays is proposed and evaluated on a test dataset. An earlier approach to identify differentials by application of statistics to probe-level data utilized a pairwise comparison (comparing one chip from each treatment) using non-parametric Wilcoxon signed rank test with a perturbation factor to account for technical variability [10]. Subsequently, standard p value cut-off (from Student's t-test) at median probe level analysis after applying logit transformation of the probe level data has been reported [8]. The former has the limitation of having to compare pairwise (i.e., one chip to one chip) and not directly applicable to datasets with replicates, the latter while being powerful does not exploit the full potential of the technological design. Another application involves more involved algorithm using combination of t-test p values, all pairwise ratios and Wilcoxon signed rank test [11]. More recently, application of two-factor ANOVA to probe-level data has been shown to be very sensitive and powerful than the method proposed here or the other methods discussed above on trial datasets [12]. However, this approach could potentially suffer from certain limitations including (i) between group deviations on either directions for different probe-pairs representing the same probesets and (ii) a large deviation of one or two probe-pairs among mostly invariant probe-pairs.

The approach proposed here determines data specific statistical threshold and a probe-pair threshold required for optimal selection of differentials using a response surface assisted model. In addition to the use of the response surface two simple equations are formulated: one to determine the optimal selection of true positives with maximal combination of sensitivity and selectivity, and other to achieve this goal considering the cost for this selection. The latter aids selecting the optimal threshold with the least cost. In this case, the cost factor is a simple additive value between the statistic t the number of probe-pairs (np). Application of additional safeguards to further control false positives are feasible. Further evaluation of the methodology on different biological datasets of varying properties with independent confirmation of the results using another technology should be valuable.

While the utility of this approach has been demonstrated with GeneChip type data it should have applicability in sensitive identification of differentials in time course data and in study of other data types where a response/phenotype is measured using multiple independent measurements.

ACKNOWLEDGEMENTS

The release of spike-in data sets by Affymetrix, CA for the purpose of algorithm development is gratefully acknowledged.

REFERENCES

1. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270**: 467-470.
2. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: **High density synthetic oligonucleotide arrays.** *Nat. Genet* 1999, **21**, 20-24.
3. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR et. al.: **Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.** *Nat. Biotechnol* 2001, **19**: 342-347.
4. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.** *Proc. Nat. acad. Sci. USA* 2001, **98**: 31-36.
5. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4**: 249-264.
6. Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003, **19**: 185-193.
7. Student: **The probable error of a mean.** *Biometrika* 1908, **6**: 1-25.
8. Lemon WJ, Liyanarachchi S, You M: **A high performance test of differential gene expression for oligonucleotide arrays.** *Genome Biology* 2003, **4**: R67.
9. Green DM, Swets J: A. *Signal Detection Theory and Psychophysics* Wiley, New York; 1966.

10. Liu WM, Mei R, Di X, Ryder TB, Hibbell E, Dee S, Webster T A, Harrington CA, Ho MH, Baid J, Smeekins SP: **Analysis of high density expression microarrays with signed-rank call algorithms.** *Bioinformatics* 2002, **18**; 1593-1599.
11. Aimone JB, Gage FH: **Unbiased characterization of high-density oligonucleotide microarrays using probe-level statistics.** *J Neurosci Methods* 2004, **135**: 27-332
12. Barrera L, Benner C, Tao Y-C, Winzeler E, Zhou Y: **Leveraging two-way probe-level block design for identifying differential gene expression with high-density oligonucleotide arrays.** *BMC Bioinformatics*. 2004, **5**: 42-55.

Table 1. Effect of different threshold of t-statistic and number of probe-pairs on the selection of spike-ins of varying concentrations

	3,6	3,7	4,6	4,7	6,6	7,5	12,6	7.71,6
0*	1	0	0	0	0	0	0	0
0.125	4	1	3	1	3	3	1	2
0.25	3	1	3	1	1	1	1	1
0.5	1	1	1	1	1	1	0	1
1	5	5	5	5	5	5	4	4
2	7	6	6	6	6	6	5	6
4	8	8	8	8	8	8	6	8
8	8	8	8	7	7	8	7	7
16	8	8	8	8	8	8	8	8
32	9	9	9	9	9	9	9	9
64	9	9	9	9	9	9	9	9
128	9	9	9	9	9	9	9	9
256	9	8	8	8	8	8	6	8
512*	9	9	9	9	9	9	9	9
CR†	5	4	5	4	4	5	3	4
Total identified	95	86	91	85	87	89	77	85
Total present	135	135	135	135	135	135	135	135
FP	16	2	5	0	0	1	0	0
PPV	0.86	0.98	0.95	1.00	1.00	0.99	1.00	1.00
Sensitivity	0.70	0.64	0.67	0.63	0.64	0.66	0.57	0.63

Indicated are the number of spike-ins of two fold difference identified at each threshold (out of 9, three in each comparison for three individual comparisons). The concentration of the spike-in (in pM) are indicated in the leftmost column in each case the concentration of the spike-in in the other dataset is twice this amount (except as indicated below). The threshold of t-statistic (t') and number of valid probe-pairs (np) is indicated in the first row as (t',np). FP is number of false positives, PPV is positive prediction value [TP/(TP+FP)], sensitivity is [TP/(TP + FN)].

* 0 pM spike-in was compared to 0.125 pM spike-in, and 512 pM spike-in is compared to 0 pM spike-in.

† CR indicates cross-reactive transcripts/probesets with homology to spike-ins (out of 9, three in each comparison for three individual comparisons).

Table 2. Effect of different thresholds of t statistic and minimum number of probe-pairs on the identification of 1.5, 2, 3 and 4 fold differentials

	1.5	2	3	4
3,5	9287	13595	18001	19251
3,6	7553	11031	15101	16657
4,6	8548	12753	16947	18287
4,7	6927	10333	13965	15431
6,6	7418	11444	15235	16677
7,5	8588	13500	17426	18690
12,6	5164	8993	12204	13596
7.71,6	6634	10584	14111	15600

Indicated are the number of probesets (average of three independent comparisons) detected (out of possible 22,301) at the given thresholds of t statistic cut-off (t') and minimum number of probe-pairs (np) satisfying this t', indicated as (t',np) in column 1. For the purpose of this evaluation three replicates were compared to three other independent replicates essentially representing the same samples scaled to the given differential (indicated in first row), and the values indicated are averages of three such independent evaluations.

Figure Legends

Figure 1: Effect of different combinations of the statistic t and threshold probe-pair number on sensitivity, positive prediction value, number of true and false positives selected.

(A) Positive prediction value - true positives/(true positives + false positives) identified – are indicated by black lines with values indicated in the primary y-axis, and sensitivity – true positives/(true positives + false negatives) – are indicated by grey lines with values indicated in the secondary y-axis (to the right) as a function of increasing t-statistic. (B) True positives identified are indicated by black lines with values indicated on primary y-axis and false positives by grey lines indicated on secondary y-axis as a function of increasing t-statistic. Each line on the graph represents the response for a specified number of probe-pair threshold increasing from 3 to 10. Arrows indicates direction of increasing probe-pair threshold.

Figure 2: The response surface of effective number of true positives picked (Neff) and cost adjusted N_{eff} (CAN_{eff}) as functions of changing t statistic probe-pair thresholds.

(A) The response surface of N_{eff}, defined as the product of true positives picked, positive prediction value and residual proportion of false positives to true positives on the two-dimensional plane of varying t (3 through 12) and minimum number of probe-pair satisfying the t threshold (3 through 10). (B) Response surface of cost adjusted $N_{eff,}$ CAN_{eff} – defined as the product of the sum of threshold t and probe-pair and N_{eff}. Values are colored from green at lower values and red at higher values.

Figure 3: Distribution of signal values identified (+) or missed (-) at different ratios of differential (A) and distribution of signal values as a function of rank for differential ratio of two (B).

Signal values were extracted with Microarray Suite 5.0 (Affymetrix, CA). Panel A represents signal values at 5^{th} percentile through 95^{th} percentile (at increments of 5%) - bottom to top - for the probesets identified (+) or missed (-) for each differential ratio are indicated. The number of probesets of that category is indicated in the top. The median value is indicated by "–" symbol, and the remaining values by diamonds. In each case the average of signals of three replicates from the dataset representing the lower value is indicated. Panel B represents the signal values as a function of rank (product of weighted average of t statistic and number of probe-pairs used for selection of differential) with higher ranks indicating higher significance. For the purpose of this evaluation three replicates were compared to three other independent replicates essentially representing the same samples scaled up by 2 fold. The data are with (t, probe-pair) threshold of (3,7).

Figure 1

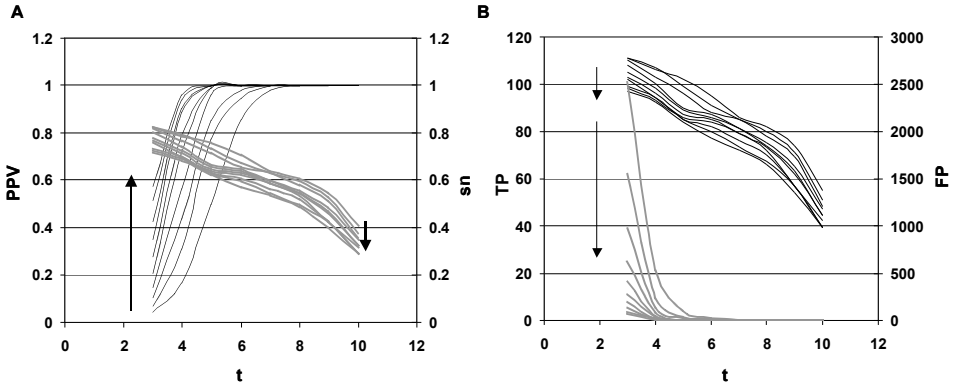


Figure 2

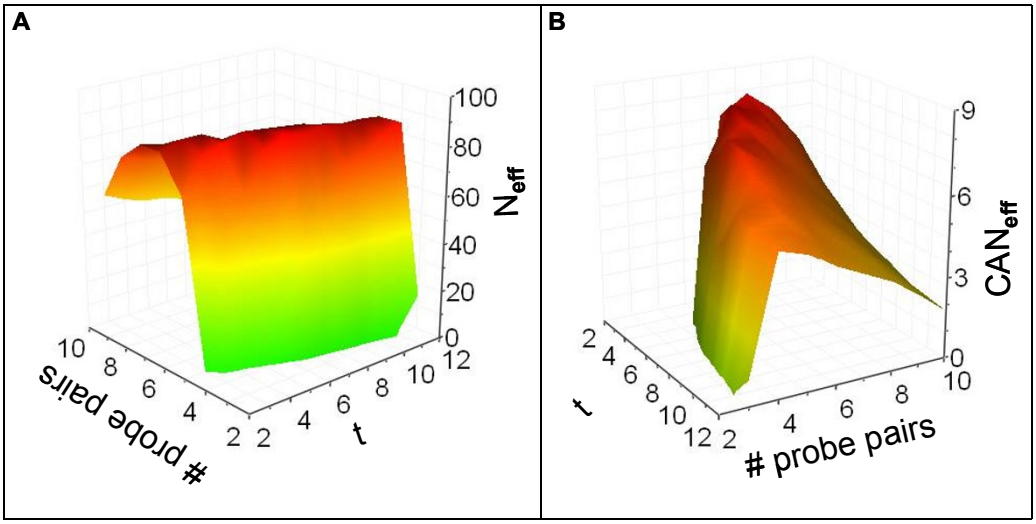
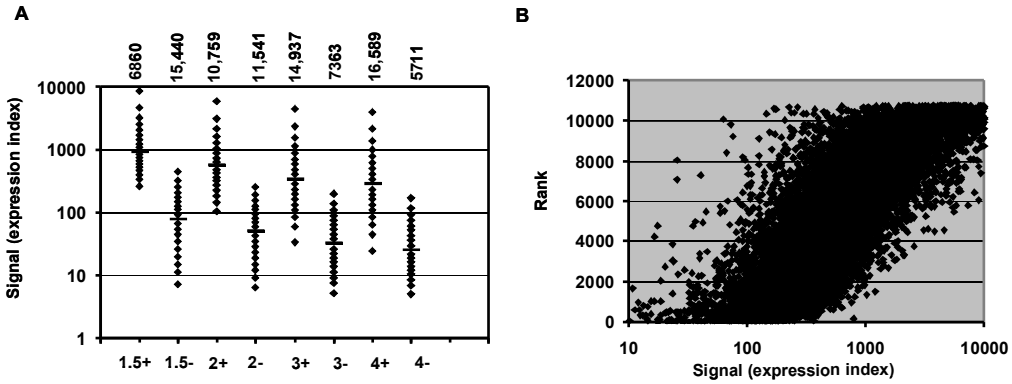

A
100
80
60
40
20
0
N_{eff}
10
8
6
4
2
probe pairs
2
4
6
8
10
12
t
B
9
6
3
0
CAN_{eff}
2
4
6
8
10
12
t
2
4
6
8
10
probe pairs

Figure 3


A
6860
15,440
10,759
11,541
14,937
7363
16,589
5711
10000
1000
100
10
1
Signal (expression index)
1.5+
1.5-
2+
2-
3+
3-
4+
4-
B
12000
10000
8000
6000
4000
2000
0
Rank
10
100
1000
10000
Signal (expression index)