This information has not been peer-reviewed. Responsibility for the findings rests solely with the author(s).

Deposited research article

Microarray data analysis: a practical approach for selecting differentially expressed genes

David M Mutch[1], Alvin Berger[1], Robert Mansourian[1], Andreas Rytz[2], Matthew-Alan Roberts[1]*

Addresses: [1]Metabolic and Genomic Regulation, Nestlé Research Center, Vers-chez-les-Blanc, CH-1000 Lausanne 26, Switzerland. [2]Applied Mathematics, Nestlé Research Center, Vers-chez-les-Blanc, CH-1000 Lausanne 26, Switzerland.

Correspondence: Matthew-Alan Roberts. E-mail: matthew-alan.roberts@rdls.nestle.com

Posted: 16 November 2001

Genome **Biology** 2001, **2(12)**:preprint0009.1-0009.29

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/12/preprint/0009

© BioMed Central Ltd (Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 15 November 2001

This was the first version of this article to be made available publicly. A peer-reviewed and modified version is now available in full at http://www.biomedcentral.com/1471-2105/3/17

.deposited research

AS A SERVICE TO THE RESEARCH COMMUNITY, GENOME **BIOLOGY** PROVIDES A 'PREPRINT' DEPOSITORY TO WHICH ANY ORIGINAL RESEARCH CAN BE SUBMITTED AND WHICH ALL INDIVIDUALS CAN ACCESS FREE OF CHARGE. ANY ARTICLE CAN BE SUBMITTED BY AUTHORS, WHO HAVE SOLE RESPONSIBILITY FOR THE ARTICLE'S CONTENT. THE ONLY SCREENING IS TO ENSURE RELEVANCE OF THE PREPRINT TO GENOME **BIOLOGY**'S SCOPE AND TO AVOID ABUSIVE, LIBELLOUS OR INDECENT ARTICLES. ARTICLES IN THIS SECTION OF THE JOURNAL HAVE **NOT** BEEN PEER-REVIEWED. EACH PREPRINT HAS A PERMANENT URL, BY WHICH IT CAN BE CITED. RESEARCH SUBMITTED TO THE PREPRINT DEPOSITORY MAY BE SIMULTANEOUSLY OR SUBSEQUENTLY SUBMITTED TO GENOME **BIOLOGY** OR ANY OTHER PUBLICATION FOR PEER REVIEW; THE ONLY REQUIREMENT IS AN EXPLICIT CITATION OF, AND LINK TO, THE PREPRINT IN ANY VERSION OF THE ARTICLE THAT IS EVENTUALLY PUBLISHED. IF POSSIBLE, GENOME **BIOLOGY** WILL PROVIDE A RECIPROCAL LINK FROM THE PREPRINT TO THE PUBLISHED ARTICLE.

Abstract

Background

The biomedical community is rapidly developing new methods of data analysis for microarray experiments, with the goal of establishing new standards to objectively process the massive datasets produced from functional genomic experiments. Each microarray experiment measures thousands of genes simultaneously producing an unprecedented amount of biological information across increasingly numerous experiments; however, in general, only a very small percentage of the genes present on any given array are identified as differentially regulated. The challenge then is to process this information objectively and efficiently in order to obtain knowledge of the biological system under study and by which to compare information gained across multiple experiments. In this context, systematic and objective mathematical approaches, which are simple to apply across a large number of experimental designs, become fundamental to correctly handle the mass of data and to understand the true complexity of the biological systems under study.

Results

The present report develops a method of extracting differentially expressed genes across any number of experimental samples by first evaluating the maximum fold change (FC) across all experimental parameters and across the entire range of absolute expression levels. The model developed works by first evaluating the FC across the entire range of absolute expression levels in any number of experimental conditions. The selection of those genes within the top X% of highest FCs observed within absolute expression bins was evaluated both with and without the use of replicates. Lastly, the FC model was validated by both real time polymerase chain reaction (RT-PCR) and variance data. Semi-quantitative RT-PCR analysis demonstrated 73% concordance with the microarray data from Mu11K Affymetrix GeneChips. Furthermore, 94.1%

of those genes selected by the 5% FC model were found to lie above measurement variability using a SD_{within} confidence level of 99.9%.

Conclusion

As evidenced by the high rate of validation, the FC model has the potential to minimize the number of required replicates in expensive microarray experiments by extracting information on gene expression patterns (*e.g.* characterizing biological and/or measurement variance) within an experiment. The simplicity of the overall process allows the analyst to easily select model limits which best describe the data. The genes selected by this process can be compared between experiments and are shown to objectively extract information which is biologically & statistically significant.

Abbreviations

CV: coefficient of variation
FC: fold change
HFC: highest fold change
LFC: limit fold change
MAE: mean absolute expression
RN: rank number
RT-PCR: real time polymerase chain reaction
RV: rank value
SD: standard deviation

Background

The complete sequencing of several genomes, including that of the human, has signaled the beginning of a post-genomic era in which scientists are becoming increasingly interested in functional genomics; that is uncovering the functional roles of different genes, and how these genes may interact with and/or influence one another. However, this question no longer need be answered by examining individual genes/proteins, but rather by simultaneously studying hundreds to thousands of unique genetic elements at a time. Already, the post-genomic era is beginning to subdivide into distinct 'omic' domains, such as genomics, proteomics, and metabolomics. This enables researchers to examine not only genetic elements, but also the corresponding proteins and metabolites derived from these genes. All such 'omic' technologies require fresh looks at data analysis issues, and many techniques are being shown to be applicable to them all. To date, the most widely studied of these 'omic domains' is that of transcriptomics, which is able to reveal subtle differences in thousands of mRNA levels between experimental samples and medical biopsies. Although mRNA is not the end product of a gene, the transcription of a gene is both critical and highly regulated, thereby providing an ideal point of investigation[1]. The development of DNA microarrays has enabled the global measurement of gene expression at the transcript level, and therefore a glimpse into the coordinated control and interactions between various genes.

At present, two technologies dominate the field of high-density microarrays: the cDNA printed array and the oligonucleotide array. The cDNA array has a long history of development [2] stemming from immunodiagnostic work done in the 1980s; however, it has been most widely developed in recent years by Stanford researchers, using the technique of depositing cDNA tags onto a glass slide with precise robotic printers [3]. Labeled cDNA fragments are then hybridized to the cDNA probes on a chip and differences in the mRNA between samples can be identified and visualized using an arbitrary coloring scheme, where typically red indicates a down regulation and green indicates an up regulation. The oligonucleotide array, largely developed by Affymetrix, Inc (Santa Clara, CA, USA) [4], involves synthesizing short (25-mers) probes directly onto a glass slide using photolithographic masks [5,6]. Sample processing includes the production of labeled cRNA, which is then hybridized to the chip, and a corresponding signal obtained after laser scanning. Regardless of the array used, the output can be readily transferred into various commercially available data analysis programs, where the selection and clustering of significantly modified genes can be examined.

Differentially expressed genes will be defined herein as gene data lying outside the normal distribution of differences with a control state, and which can not be ascribed to chance or natural variabilty. Various creative techniques have been proposed and implemented for the selection of differentially expressed genes; however, none has yet gained widespread acceptance in the analysis of microarray data. Despite this, there remains a great impulse to develop new data analysis techniques, in part driven by the obvious need to move beyond setting arbitrary fold-change cut-off which are out of context with the rest of the experimental and biological data at hand [7-9]. This is still the case for many studies, where selection of differential gene expression is performed through a simple fold-change cut-off between 1.8 to 3.0. There is an inherent problem with these simple selection criteria: it is far easier to see a 2-fold change in genes that are lowly expressed than a 2-fold change in highly expressed genes. Selecting significant genes based only on a single fold change across the entire range of experimental data preferentially selects genes that are lowly expressed [8]. Furthermore, this commonly used approach does not accommodate for background noise, variability, non-specific binding, or low copy numbers which are characteristics of gene data generated by a microarray. Other approaches entail the use of the simple statistical measures such as a *t-test* for every individual gene, however due to the cost of repeating microarray experiments, the *n* usually remains low and thereby can lead to inaccurate estimates of variance [8].

The present article describes a fold change model that takes into account both expression levels and fold changes for the selection of significantly differentiated genes. This simple model aids the experimenter to estimate the relationship between these two parameters and thereby extract information which becomes relevant to the estimation of variation. Subsequently those gene transcripts which can be determined to be outliers can be termed differentially expressed genes. An added strength to the model outlined within this document lies in is its ease of application to any dataset. This model can then be considered a progressive and cyclical process, where the data analyst can quickly and objectively identify a list of differentially regulated genes with a high level of confidence.

The development of this model was performed on data stemming from a nutritional experiment in a mouse model using Affymetrix Mu11K chips, where the effects of four diets were compared in a number of organs (pool of five mice for each sample in each organ): (1) control diet A in duplicate from the same pool, (2) diet B, (3) diet C, and (4) diet D. For the purposes of the current report, details of the dietary input are not required and will be reported elsewhere. The present article will take only the data from the liver as an example for the development of a gene selection model. The model is further validated by RT- PCR and indicates a high concordance between microarray data and RT- PCR data.

Selection of Differentially Regulated Genes & Data Analysis

A method of objective gene selection was sought to avoid the reliance simply on a single arbitrary fold-change cut-off, which is known to be overly influenced by both small and large absolute expression levels. The chosen method includes (A) the determination of the upper X% of highest fold changes within narrow bins of absolute expression levels, (B) the rejection of very small absolute values, and (C) the subsequent ranking of genes by a combined fold change/absolute difference calculation.

(A) Selection of the upper X% of highest fold changes within binned absolute expression levels

The data from a typical Affymetrix experiment contains an average difference *(Avg.Diff)* value, which can be described as the difference in intensity between a perfect match oligonucleotide and a mismatch oligonucleotide. In order to clarify this parameter in terms of the present model, the term "absolute expression" will be used in place of "average difference". As usually indicated in literature, both minimal and negative absolute expression values are set to a common number in order to eliminate genes with negative expression levels and to reject essentially uninterpretable information. Therefore, as a first-pass filter, genes with absolute expression values of less than 20 were set to 20 and all genes which had a value of 20 across all four diets were immediately rejected. This process left 9391 genes in the liver out of the original 13179 genes represented on the Mu11K GeneChip. An additional parameter, highest fold change, was then applied to these remaining genes. HFC can be defined as:

$$\mathrm{HFC} = \frac{\mathrm{Max(A,B,C,D...)}}{\mathrm{Min(A,B,C,D...)}} \qquad \textit{eqn. 1}$$

where A,B,C,D, etc… represent the individual microarray results for each gene

The proposed determination of HFC is highly influenced by absolute expression, and trends can readily be observed in our data set where HFC is negatively correlated with absolute expression. For example, it can be seen that with absolute expression values higher than 5000, it is unlikely to have HFC greater than 1.5, but with absolute expression values near 50, it is very easy to observe an HFC of ≥ 2. It should be noted that the present experiment is comprised of four diets or treatments; however, the HFC can be easily calculated for any number of experimental conditions. Furthermore, similar trends can be observed in numerous Affymetrix datasets we have examined (data not shown).

An ultimate goal was to develop a model that would account for absolute values when filtering genes on fold change. The selection of differentially expressed genes is essentially a search for outliers, *i.e.* gene data lying outside the normal distribution of differences relative to a control state, and which can not be ascribed to chance or natural variabilty. In order to determine those genes which are outliers, it is necessary to either measure the variability of the system or to make valid assumptions regarding the normal distribution of variability. In the present model we assume that: (1) variability in gene expression measurements are related to the absolute expression level; and (2) that if a broad sampling of the transcriptome is measured then only a small number of genes will actually be outliers even in the harshest of experimental treatments. Assumption (1) is a fairly general analytical concept, *i.e.* that the closer data is to the measurement threshold the higher the variability is in that measurement. Assumption (2) appears to be empiricaly valid when surveying the literature for high-density microarray experiments which evaluate severe biological events, from caloric restriction [10,11] to apoptosis [12,13]. In these experiments, through various selection techniques, it was found that less than 5% of the total number of genes probed were differentially regulated. Therefore, in order to develop the present model of gene selection, the validity of selecting outliers was evaluated for a range of highly variable genes, from 5% of the population on up.

The present model was developed by binning gene expression data into tight classes across the range of absolute expression values, *i.e.* 20-50, 50-100, 100-150, *etc.* and then selecting the upper 5% of HFC values for further consideration. Binning was carried out in such a manner as to ensure that there was never a bin containing zero genes or fewer genes than the proceeding bin, therefore bin sizes were not always equal. It is possible to search separately for the 5% of genes with the greatest HFCs in each class; however, in order to simplify the overall selection, we modeled the relationship between absolute expression, defined as MIN(diets A,B,C,D) value and HFC (*eqn 1*) in order to set a limit fold change (LFC). The relationship can be modeled using a simple equation of the form $LFC = a + b/x$ (with a and b depending on the number of genes to be selected). Figure 1a demonstrates that as the selection criteria becomes more strict (top 5% → 3% → 1% of genes), the LFC curves change, yet converge at expression levels above 1000. The simple equation contains two parameters that have various repercussions on gene selection. Firstly, a sets the asymptote, which corresponds to the minimum highest fold change value that can be observed at any given absolute value. Secondly, b affects the LFC at a given absolute value, and is therefore highly influenced by this latter value. For example, the lower the absolute values the greater the LFC, and vice versa.

Using the equations in Figure 1a, the selection of genes for further consideration is then objective, simple, and global. A gene is selected with the HFC approach if MAX(A,B,C,D)/Min(A,B,C,D) > a+b/Min(A,B,C,D). After applying the 5% LFC gene filter, 489 genes remained in the list out of the 9391 genes potentially differentially expressed, selected from the original 13179 genes represented on the GeneChip. When interested in only the top 3% or 1% of significant genes, the total number of genes that meet the LFC requirements, and correspondingly the number of genes per bin, drops off rapidly (245 and 102 genes, respectively).

(B) The rejection of very small absolute values

Lastly, in an effort to objectively determine a minimum expression level cut-off we examined the final distribution of absent & present calls (Absence Call) across gene bins in the remaining set of genes. It was determined that Affymetrix absence/presence calls would not be used *a priori* as criteria critical to the selection of significantly regulated genes, but that it would rather be used as a post-selection criteria. The absence call has been previously noted to be problematic, and has two potential drawbacks: 1) the assignment of an absence call is based on the *ad hoc* characterization of oligonucleotide matches & mismatches for which the validity has been previously challenged, and 2) is not empirically reliable for individual genes, *i.e.* the confidence in the call is not high [14]. However, it was expected that the distribution of absent calls across many genes at a range of absolute expression levels would not be random, and that the trend would be an important crosscheck for the confidence placed in changed genes at low expression levels.

As expected the distribution of absent calls demonstrated that it was predominantly the very lowly expressed genes (95% of genes called absent, absolute expression ≤ 207), which were called absent across all four diets by the Affymetrix analysis software. This analysis also supports the idea that a threshold for an absolute minimum expression level could be developed empirically for each data set examined. In the present case, this would imply that any gene, which didn't have at least a value of 207 in one experimental condition needs to be rejected independent of the fold change measured. In practice, more than 95% of genes meeting these criteria would also be rejected on the basis that they were consistently marked absent across all experimental conditions. Therefore, such genes were eliminated in the last method of gene filtration. After removing these lowly expressed genes, based on these objective criteria, 329 genes remained in the list out of the original 13179 gene probe sets. The selected genes were considered to be potentially differentially regulated by our dietary treatments in the sense that these are the most highly differentially regulated genes within the context of the present experiment.

(C) Assignment of Gene Rank

Following overall gene selection, a rank of "importance" or "interest level", defined as Rank Number (RN), based on both the magnitude of fold change and absolute expression values was assigned to each selected gene. The RN for each gene was determined by calculating a Rank Value (RV), which can be defined as: RV = HFC * (Max – Min). The RV is an abstract value that simply gives great importance to those genes that have a high fold change and simultaneously high differences in absolute expression values. After calculation of RV, gene lists were sorted and then assigned a simple rank of 1,2,3,4…329 in order of RV importance, where a gene with a RN of 1 corresponds to the gene with the highest RV. Both RV and RN are simply aids for the discussion of differential gene effects, which add the concept of relative weight or "importance" amongst selected genes. This concept then provides a further basis for the selection of genes for validation studies as is detailed below.

(D) Model validation

Real-time polymerase chain reaction

The results obtained from a microarray experiment are influenced by each step in the experimental procedure, from array manufacturing to sample preparation and application to image analysis [15]. The preparation of the cDNA sample is highly correlated to the efficiency of the reverse transcription step, where reagents and enzymes alike can influence the reaction outcome. All of these factors correspondingly affect the representation of transcripts in the final cDNA probe, which necessitate the need for validations by complementary techniques. Analysis by northern blot and RNAse protection assays are commonly reported in the literature; however, the emerging "gold-standard" validation technique is RT- PCR [16]. As microarrays tend to have low dynamic range, which leads to small but significant under-representations of fold changes in gene expression, RT-PCR with a higher dynamic range is used more to validate the observed trends rather than duplicate the absolute values obtained by chip experiments [17,16,18].

Having chosen genes that lie across the ranking system, RT- PCR was performed in triplicate for each experimental condition (Diet A, B, C, D) using the same pooled stocks of liver RNA (5 mice/experiment). Genes were compared to the endogenous controls β-actin and GAPDH, which were determined not to have significantly changed across the dietary treatments by both the LFC (microarray data) and a student's *t*-test (RT-PCR). Subsequently, significant changes by RT-PCR were calculated by the student's t-test with a predefined nominal α level of 0.05; where Diet B, C, and D were independently compared to the control diet A. The overall concordance of trends between the two techniques was 73% (*e.g.* an increase/decrease in gene expression seen by microarray was also seen by RT-PCR). For those genes whose results agreed between the two experiments, 68% of these results indicated larger fold changes by RT-PCR than those identified by array analysis. This concordance includes both genes determined as significantly changed as well as those genes determined not to have been significantly changed. When only those genes that were considered to be significantly changed by RT-PCR were examined, the concordance increased slightly to 80%.

What is immediately noticeable through the color scheme (Table 1) is that genes with high RN (low RV) have little to no concordance between the two techniques; where red indicates no concordance and blue indicates either one or two (out of three) of the results did not agree. When specifically examining fatty acid synthase (FAS), a highly expressed gene, one can quickly see that microarray fold changes of less than 2 can be corroborated between the two experimental techniques, reinforcing the strength of this fold change model.

As the selection criteria with the microarray data was that the HFC must be greater than the LFC model, the expectation is that the LFC trend line can be validated by RT-PCR. This is predominantly the case across the full dynamic range of data selected by the model; except for very lowly expressed genes such as the RAS oncogene. For genes with slightly lower RN (higher RV), such as ABCA1, and HSP5 some concordance is seen, indicating that confidence in gaining with these genes, and that as a group they can still be taken into account when looking for trends in gene expression. For genes with a RN lower than 176 (RV > 1156; e.g. USF-2) concordance quickly approaches 100%, indicating high confidence when discussing gene trends or individual gene results. These results in total reinforce the concept that RN is correlated with confidence / validity within the selected gene set resulting from the LFC model.

The genes discussed and validated in this report were identified using the 5% fold change model; however the fold change percentage can be varied to meet both the researcher's and experiment's needs. It must be stressed that the 5% fold change model was chosen under the assumption that a relatively small percentage of genes will have their expression altered under any given condition. Therefore, selecting a fold change model of 5% may be either too permissive, where false positives are selected as differentially changed, or too restrictive, where true positives are not selected. Within the context of the present study, validation of the microarray results indicates that genes with low rank values are often more difficult to confirm by complementary techniques. Using the data obtained from RT-PCR, if one assumes that all genes with a RN below 176 (corresponding to RV > 1156) can be validated, then one would expect that these genes would be concentrated at higher expression levels. However, when the spread of those genes with a rank of 1 to 176 is examined, it was observed that these genes comprise a wide range of expression levels, indicating that the fold change model is objectively selecting differentially regulated genes across a wide range of absolute expression levels (data not shown), and that confidence in that selection increases with RV.

Variance Analysis with Real-time PCR

Variability is introduced into microarray data from two sources: biological variation (whether *in vitro* or *in vivo*) and measurement variation (hybridization, processing, scanning, *etc*.). In a brief effort to examine variability between individual mice, *i.e.* biological variability, RT-PCR measurements across control mice were examined using a subset of the genes examined by RT- PCR. Each gene was examined in triplicate in each of the five mice, and the variation in ΔCt (detection threshold) was determined. The Ct indicates the relative abundance of any particular gene, and when normalized to an endogenous control (β-actin and GAPDH) allows the relative amounts of a gene to be calculated. RT- PCR indicated as did the microarray variance data, that lowly expressed genes have a higher variation; thereby hinting that biological and measurement variance are both influenced by absolute expression levels. The equation of the line was deemed significant (with a p-value of 0.014 and 0.013 when normalized against β-actin and GAPDH, respectively). This again confirms the concept that highly expressed genes have little variance, and that small fold changes do represent a meaningful biological event.

Validation of the LFC model via characterization of measurement variability

The concept that variability and absolute expression are related has recently been examined by Coombes and colleagues; however, they examined only the variability of replicate spots on a single slide [19]. This concept has now been further extended here to the examination of variability between genes on different microarrays. Measurement variance was examined following the development of the LFC model, and was therefore treated as a separate method for the confirmation of this model. To further understand the nature of measurement variability within the current study, duplicate Mu11K Affymetrix microarrays for the controls were examined. A pooled RNA sample from mice (*n*=5) fed the control diet was hybridized to two different chips, and the data was analyzed in order to characterize measurement variability (data not shown). It was apparent from the trend that as absolute expression levels increase, the coefficient of variation (CV= SD/MAE) decreases. By overlaying the trendline of the variability data on those genes determined to be significantly regulated by the LFC model, the CV upper confidence level for these selected genes could be elucidated.

In order to estimate the CV without taking into account extreme values of the duplicate we used a robust estimator, represented by the following equation:

$$Median.CV_{duplicate \cdot sample} * \sqrt{\frac{n-1}{\chi_{inverse}(p,n-1)}} = CV_{population} \qquad eqn.\ 2$$

Where n = 2 and p = 0.5 (as the median CV of duplicate gene sample was used), the above equation enabled the CV to be determined by narrow bins of mean expression level, where extreme values are not accounted for.

The mean absolute expression of 13057 data points (genes) across the four diets were plotted against CV, and indicated a similar trend for the variability data; where a high mean absolute expression results in a low CV (Figure 1b). Applying the CV derived from the duplicate sample data (*eqn. 2*) to the quadruplicate diet data enables the calculation of the CV upper confidence level (by bins of absolute expression level) using the following equation:

$$CV_{population} * \sqrt{\frac{\chi_{inverse}(p,n-1)}{n-1}} = CV.upper.confidence.level \qquad eqn.\ 3$$

Where n= 4 and p= 0.001, 0.00001, 0.0000003, depending on the level of confidence desired (1-p).

Equation 3 allows us to identify those genes with a variance above the measurement variability . This greater variability arose due to combined pool (biological) and treatment variabilities.

This confidence level, by altering *p*, could then be raised or lowered according to the level of confidence desired; therefore, modeling the variance data provides an objective method for examining the variation of genes across the complete range of absolute expression values. The spread of the data indicates that most of the 13000 genes are both lowly expressed and highly variable across the four chips. A further examination of the data indicated that 95% of the genes determined to be 'absent' across all four diets by Affymetrix software had a mean absolute expression less than 207.

With the LFC model, genes were initially selected if they were in the top X% of the bin highest fold changes; however the starting point (X%) was solely chosen based on the percentage of genes shown to be differentially regulated across a wide-range of published biological studies. However, the genes selected by the X% fold change model were then verified, with concordance results, by both RT- PCR and the variance data. Genes identified by the 5% fold change model were overlayed on the variance data corresponding to the four diets, and the confidence level for the X%-data selection was determined (Figure 1b). Concordance of 94.1%, 96.6% and 98.4% for the 5%, 3% and 1% fold change models, respectively, was observed with an upper confidence level selection of 99.9% (Figure 1b, inset table). In addition, overall concordance between microarray data and RT- PCR was examined in the different fold change models; and indicated 73.3%, 81.5%, and 94.4% concordance for the 5%, 3%, and 1% fold change models, respectively (Figure 1a). The degree of concordance with RT- PCR results and the high confidence level (99.9%) obtained with the variance data reinforces that the X% fold change model is a simple, efficient, objective and statistically valid method for the identification of significantly differentiated genes.

Conclusion

The analysis of microarray data is a new scientific field that has enabled researchers to establish novel and innovative methods for analyzing the results. Already, an evolution can be observed with regards to the methods of selecting significantly changed genes. Scientists are moving away from the arbitrary fold change cut-off, and incorporating robust statistical concepts into their line of thinking. The conclusion that highly expressed genes will rarely have a 2-fold change in mRNA levels, and that lowly expressed genes will commonly have a 2-fold or greater change, led to the development of models that would accommodate this real characteristic of gene expression measurements. The fold-change model presented in this paper takes into account expression level in addition to fold change, and allows for the selection of genes across the complete range of expression levels. Following gene selection using an initial criteria of X%, gene rank is introduced as a basis for choosing genes to validate the model. Therefore, a limited but judicious choice of model selected genes across a broad range of gene rank can then be used to reset X% in order to correspond with the data at hand (Figure 2). Further validation of this model in the current data set by RT- PCR confirmed these relationships, reinforcing that genes with fold changes even less than 1.8 can be consistently measured assuming adequate absolute expression levels. This demonstrates real changes in sample concentration of mRNA even at low fold-change levels. Additionally, the variance data characterizing measurement variability further supports the LFC model, indicating that selected genes lie outside measurement variability at very high confidence limits (> 99.9% CL). Although measurement variability was used here for model development, this concept can be extended to measurements of either experimental or biological variability. In summary, the X% LFC model enables one to define experiment specific selection stringency while maintaining simplicity and objectivity for the detection of differential gene selection. The LFC model can be used consistently across any number of experiments with widely varying numbers of experimental conditions, and can therefore be generalized to most types of microarray data.

Materials & Methods

Mice and feeding conditions:
Mice were male Rj:NMRI mice from Elevage Janvier, Le Genest-Saint-Isle France, weighing 10-11 g at delivery and 33-51 grams on day 42, housed 10 per cage in wire cages with bedding and normal light cycle. Mice received *ad libitum* quantities of bottled distilled water and purified powdered diets (7.5 g/mouse) in ceramic cups (10/group) for 42 d. Food was maintained at -80 °C in daily aliquots under nitrogen, thawed each afternoon before administration to mice, and uneaten food was discarded daily. Experimental diets are not relevant to the current manuscript on differential gene selection, but will be described in more detail in a following publication covering the biological significance of the experiment.

Dissection of Mice:
After administration of the aforementioned diets to 10 mice per group; 5 mice were randomly selected for inclusion in the gene expression analysis experiment. Organs were dissected according to standard protocols (Phoenix Laboratories), then cut into 100-150 mg subsections, flash frozen in liquid nitrogen, and finally stored at -80° C until gene expression analysis.

Nucleic Acid Preparation:
Tissue from each organ was extracted from 5 individual mice and extracted separately using Qiagen RNeasy mini-kits (Basel, Switzerland) according to the manufacters instructions with one exception: During extractions, all RNeasy columns were impregnated with DNase I (Roche, Basel, Switzerland) in order to remove possible genomic DNA contamination. After extraction, equal amounts of material were pooled to achieve a total of 10 µg total RNA per dietary group. All RNA samples were first quantified by the RiboGreen RNA Quantification Kit according to the manufactuers instructions (Molecular Probes, Eugene Oregon) and then analyzed via agarose gel electrophoresis for intact 18 and 28s rRNA. All samples included in the study were judged to

contain high-quality RNA in sufficient amounts for hybridization. Furthermore, Affymetrix "test chips" were run for each pool for all organs prior to hybridization with the GeneChip. All test arrays gave strong signals across an array of pre-selected genes.

Gene Expression Analysis using the Murine 11k GeneChip:

cRNA preparation.

15 µg total RNA was used as starting material for all samples. In all cases, a “test chip” provided by the manufacturer (Affymetrix, Santa Clara, CA) was run prior to using the Murine 11k GeneChip. In each case this confirmed that sufficient high quality RNA was present to detect gene expression in the various tissue samples. The first and second strand cDNA synthesis was performed using the SuperScript Choice System (Life Technologies) according to the manufacturers´instructions, but using oligo-dT primer containing a T7 RNA polymerase binding site. Labeled cRNA was prepared using the MEGAscript, In Vitro Transcription kit (Ambion). Biotin labeled CTP and UTP (Enzo) was used together with unlabeled NTP´s in the reaction. Following the IVT reaction, the unincorporated nucleotides were removed using RNeasy columns (Qiagen).

Array hybridization and scanning.

Ten µg of cRNA was fragmented at 94° C for 35 min. in buffer containing 40 mmol/L Tris-acetate pH 8.1, 100 mmol/L KOAc, 30 mmol/L MgOAc. Prior to hybridization, the fragmented cRNA in a 6xSSPE-T hybridization buffer (1mol/L NaCl, 10mM Tris pH 7.6, 0.005% Triton), was heated to 95° C for 5 min, subsequently cooled to 40° C and loaded onto the Affymetrix probe array cartridge. The probe array was then incubated for 16h at 40° C at constant rotation (60 rpm). The probe array was exposed to 10 washes in 6xSSPE-T at 25° C followed by 4 washes in 0.5xSSPE-T at 50° C. The biotinylated cRNA was stained with a streptavidin-phycoerythrin conjugate, 10 g/ml (Molecular Probes) in 6xSSPE-T for 30 min at 25° C followed by 10 washes in 6xSSPE-T at 25° C. The probe arrays were scanned at 560nm using a confocal laser scanning microscope (made for Affymetrix by Hewlett-Packard). The readings from the quantitative scanning were analysed by the Affymetrix Gene Expression Analysis Software.

References

1. Brazma A,Vilo J: **Gene expression data analysis**. *FEBS Lett* 2000, 480: 17-24.
2. Ekins RP: **Ligand assays: from electrophoresis to miniaturized microarrays**. *Clin Chem* 1998, 44: 2015-2030.
3. DeRisi JL, Iyer VR,Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale**. *Science* 1997, 278: 680-686.
4. Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS,Fodor SP: **Accessing genetic information with high-density DNA arrays**. *Science* 1996, 274: 610-614.
5. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP,Fodor SP: **Light-generated oligonucleotide arrays for rapid DNA sequence analysis**. *Proc Natl Acad Sci U S A* 1994, 91: 5022-5026.
6. Barone AD, Beecher JE, Bury PA, Chen C, Doede T, Fidanza JA,McGall GH: **Photolithographic synthesis of high-density oligonucleotide probe arrays**. *Nucleosides Nucleotides Nucleic Acids* 2001, 20: 525-531.
7. Woolf PJ,Wang Y: **A fuzzy logic approach to analyzing gene expression data**. *Physiol Genomics* 2000, 3: 9-15.
8. Baldi P,Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes**. *Bioinformatics* 2001, 17: 509-519.
9. Quackenbush J: **Computational analysis of microarray data**. *Nat Rev Genet* 2001, 2: 418-427.
10. Lee CK, Klopp RG, Weindruch R,Prolla TA: **Gene expression profile of aging and its retardation by caloric restriction**. *Science* 1999, 285: 1390-1393.
11. Kayo T, Allison DB, Weindruch R,Prolla TA: **Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys**. *Proc Natl Acad Sci U S A* 2001, 98: 5093-5098.
12. Voehringer DW, Hirschberg DL, Xiao J, Lu Q, Roederer M, Lock CB, Herzenberg LA,Steinman L: **Gene microarray identification of redox and mitochondrial elements that control resistance or sensitivity to apoptosis**. *Proc Natl Acad Sci U S A* 2000, 97: 2680-2685.
13. Cardozo AK, Kruhoffer M, Leeman R, Orntoft T,Eizirik DL: **Identification of novel cytokine-induced genes in pancreatic beta-cells by high-density oligonucleotide arrays**. *Diabetes* 2001, 50: 909-920.
14. Pavlidis P,Noble WS: **Analysis of strain and regional variation in gene expression in mouse brain**. *Genome Biol* 2001, 2: RESEARCH0042.
15. Rajeevan MS, Vernon SD, Taysavang N,Unger ER: **Validation of array-based gene expression profiles by real-time (kinetic) RT-PCR**. *J Mol Diagn* 2001, 3: 26-31.
16. Snider JV, Wechser MA,Lossos IS: **Human disease characterization: real-time quantitative PCR analysis of gene expression**. *Drug Discov Today* 2001, 6: 1062-1067.
17. Mayanil CS, George D, Freilich L, Miljan EJ, Mania-Farnell B, McLone DG,Bremer EG: **Microarray analysis detects novel Pax3 downstream target genes**. *J Biol Chem* 2001, 5: 5.

18. Wurmbach E, Yuen T, Ebersole BJ,Sealfon SC: **Gonadotropin releasing hormone receptor-coupled gene gene network organization**. *J Biol Chem* 2001, 1: 1.

19. Hess KR, Zhang W, Baggerly KA, Stivers DN,Coombes KF: **Microarrays: handling the deluge of data and extracting reliable data**. *TRENDS in Biotechnology* 2001, 19: 463-468.

Figure Legends

Figure 1. The relationship between absolute value, limit fold change (LFC), and variance across the absolute expression range. A) The various curves indicate the LFC required at different absolute values in order to be considered a significantly changed gene. As the selection criteria increases, the LFC increases, indicating that the 5% fold change model (green line) is more permissive than the 1% fold change model (red line). The various fold change models produced the curves with the following equations: A) in the liver: 5% LFC model = 1.52 + (100/absolute value); 3% LFC model = 1.55 + (140/absolute value); 1% LFC model = 1.70 + (185/absolute value). **B)** Examining the variance of each gene across the four dietary treatments enables the identification of those genes determined significantly changed. (•) represents genes below the 99.9% confidence level, (■) represents those genes selected by the 5% fold change model, and (+) represents those genes above the 99.9% confidence level. The various lines represent different confidence levels (**i.** 99.9%, **ii.** 99.999%, and **iii.** 99.99997%). As the fold change model increased (5%→1%), concordance between the fold change model and the variance data (at a confidence level of 99.9%) increased (embedded table: x(y%), where x represents the number of genes with concordance (and y the percentage of genes with concordance)).

Figure 2. Schemmatic representation of the cyclical nature of the LFC model. Selecting an initial X% limit fold change (1) provides a starting point for the identification of those genes differentially regulated. Genes can then be ranked (2) by a calculation combining fold change and absolute value in order to assign a degree of importance. Validation of the chosen LFC model with RT-PCR (3) and/or the characterization of variance (4)

enables the analyst to reexamine the initial LFC model and assign a confidence level to the results. Depending on the dataset, one could redefine the LFC model and repeat the cycle.

Table 1. Concordance data between an Affymetrix 11MuK microarray and RT- **PCR**. Through the coloring scheme, one can see that validation (confirmation by RT- PCR of the direction of fold change determined by microarray) of low rank value genes is not achieved; however as the rank value increases, concordance increases (red = genes with no concordance across the 3 diets; blue = genes with either one or two measurements in agreement; green = genes that have 100% concordance). Overall concordance with the 5% fold change model was 73%, which includes measurements found both significant and non-significant by microarray analysis. Numbers underlined indicate the HFC that resulted in this gene being selected as significantly different (70% concordance with RT-PCR results). Starred-numbers indicate significant fold changes, determined by a student's T-test, in RT- PCR (80% concordance).

Figure 1a

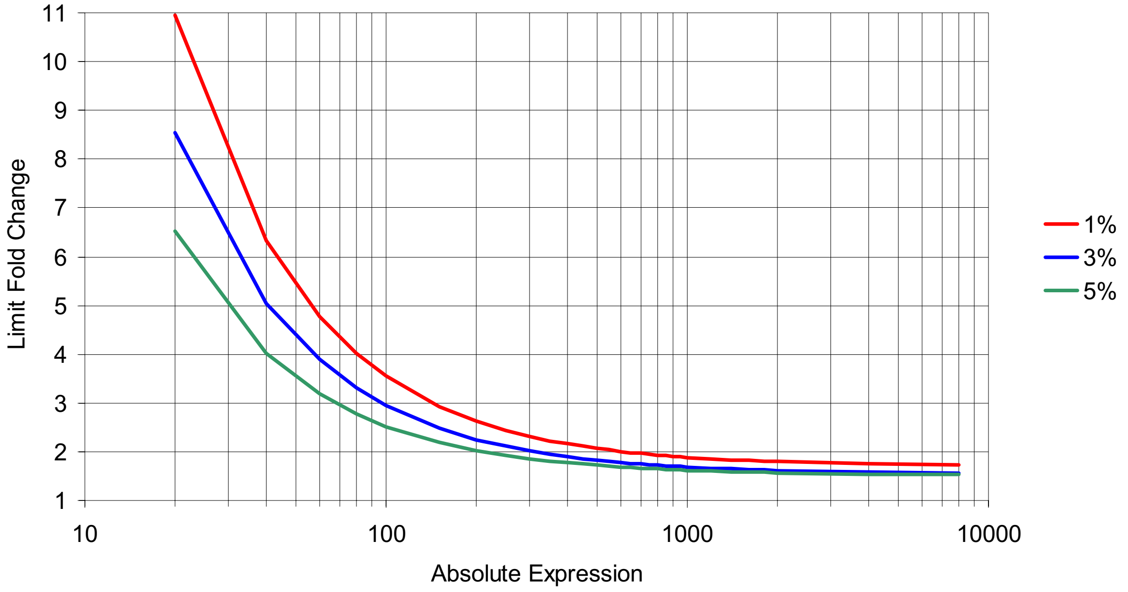
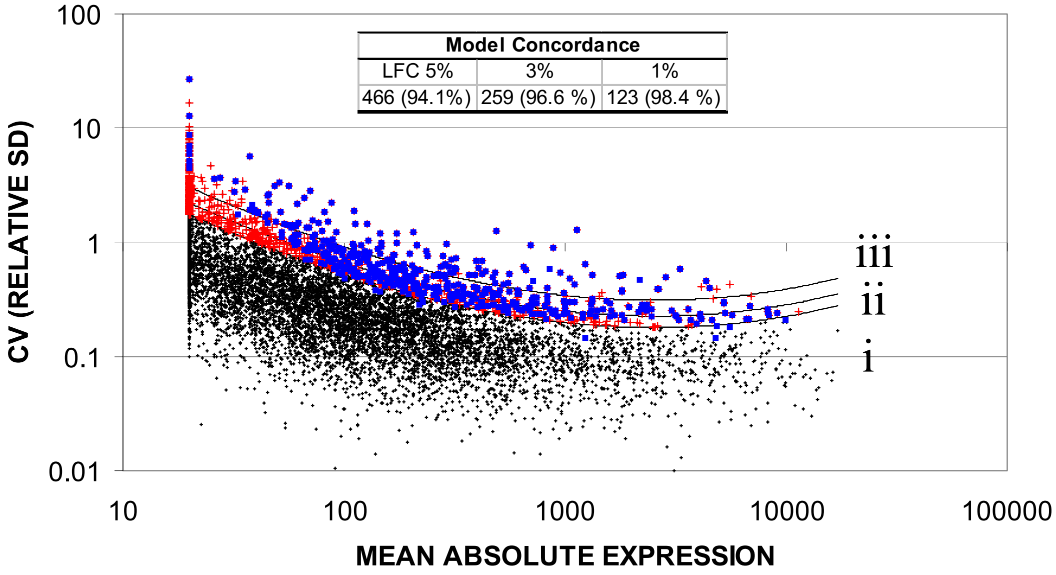


Figure 1b



Model Concordance		
LFC 5%	3%	1%
466 (94.1%)	259 (96.6 %)	123 (98.4 %)

Figure 2

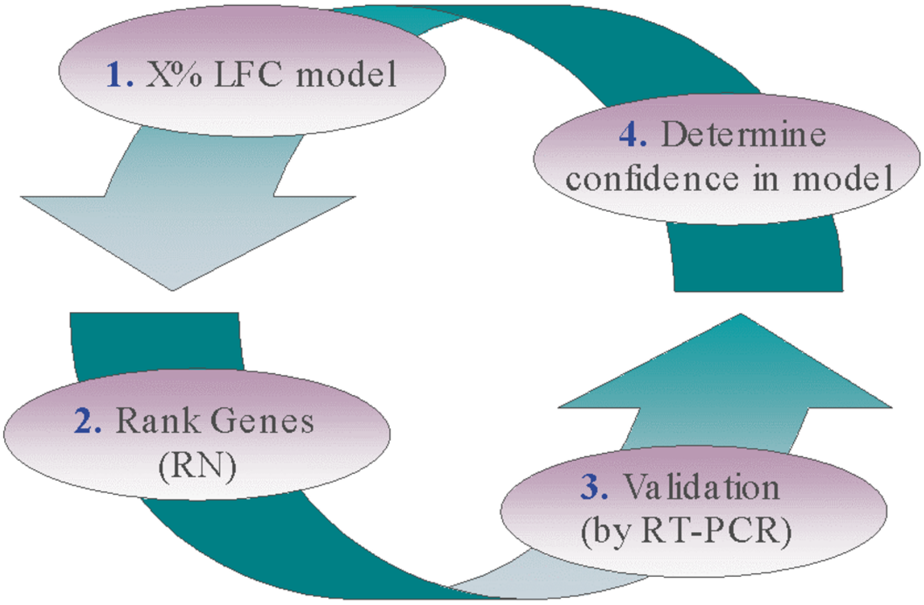

1. X% LFC model
2. Rank Genes (RN)
3. Validation (by RT-PCR)
4. Determine confidence in model

Table 1

Gene Name	Rank Number	Rank Value	Microarray			Real Time PCR		
			Diet D	Diet E	Diet F	Diet D	Diet E	Diet F
RAS	315	774	-2.49	-1.78	-1.67	1.81*	1.04	1.16
ABCA1	251	892	1.00	1.00	7.20	1.45	-1.12	-1.20
USF2	176	1156	1.16	-5.63	1.06	2.54	-1.08	1.98
HSP5	127	1559	-1.56	1.10	-1.60	2.11*	1.30	1.05
Cyp4a10	27	5754	2.67	4.67	3.01	15.00*	18.81*	6.78*
SCOD1	17	7488	-1.12	-1.03	-1.77	2.50*	-1.35	-2.65*
ALAS	7	12319	3.69	1.83	2.71	21.60*	8.51*	8.03*
FAS	4	22928	-1.92	-1.27	-5.40	-1.78	-1.40	-17.11*
ApoA4	2	32537	-2.57	-3.20	-17.18	-1.32	-3.01	-4.90*
FABP5	1	40749	-5.46	-8.43	-13.59	-2.94*	-8.49*	-16.37*