Defining the Human Endothelial Transcriptome

by

Sripriya Natarajan
M.Eng. Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 2001

S.B. Computer Science and Engineering
Massachusetts Institute of Technology, 2000

SUBMITTED TO
THE HARVARD-MIT DIVISION OF HEALTH SCIENCES AND TECHNOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTERS OF SCIENCE IN HEALTH SCIENCES AND TECHNOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2005

Signature of Author: .....
Harvard-MIT Division of Health Sciences and Technology
March 18, 2005

Certified by: .....
Guillermo García-Cardeña, Ph.D.
Assistant Professor of Pathology, Harvard Medical School
Thesis Supervisor

Accepted by: .....
Martha L. Gray, Ph.D.
Edward Hood Taplin Professor of Medical Engineering and Electrical Engineering, MIT
Co-director, Harvard-MIT Division of Health Sciences and Technology

Defining the Human Endothelial Transcriptome

by

Sripriya Natarajan

Submitted to the Harvard-MIT Division of
Health Sciences and Technology
on March 18, 2005 in Partial Fulfillment of the
Requirements for the Degree of Master of Science in
Health Sciences and Technology

ABSTRACT

Advances in microarray technology facilitate the study of biological systems at a genome-wide level. Meaningful analysis of these transcriptional profiling studies, however, demands the concomitant development of novel computational techniques that take into account the size and complexity of the data. We have devised statistical algorithms that use replicate microarrays to define a genome-wide expression profile of a given cell type and to determine a list of genes that are significantly differentially expressed between experimental conditions. Applying these algorithms to the study of cultured human umbilical vein endothelial cells (HUVEC), we have found approximately 54% of all genes to be expressed at a detectable level in HUVEC under basal conditions. The set of highest expressed genes is enriched in nucleic acid binding proteins, cytoskeletal proteins and isomerases as well as certain known markers of endothelium, and the complete list of genes can be found at http://vessels.bwh.harvard.edu/software/endo_xcriptome . We have also studied the effect of a 4-hour exposure of HUVEC to 10 U/mL of IL-1, and detected 491 upregulated and 259 downregulated statistically significant genes, including several chemokines and cytokines, as well as members of the TNFAIP3 family, the KLF family and the Notch pathway. Applying these rigorous statistical techniques to genome-wide expression datasets underscores known patterns of endothelial inflammatory gene regulation and unveils new pathways as well. Finally, we performed a direct comparison of direct-labeled microarrays with amplified RNA microarrays for an initial assessment of the effect of the additional noise of amplification on the outputs of the statistical algorithms. These techniques can be applied to additional genome-wide profiling studies of endothelium and other cell types to refine our understanding of transcriptomes and the gene regulatory network governing cellular function and pathophysiology.

Thesis Supervisor: Guillermo García-Cardeña, Ph.D.
Title: Assistant Professor of Pathology, Harvard Medical School

Table of Contents

1. Introduction

1.1. Gene Microarray Technology

Gene microarray technology now allows biologists to assay the transcriptional activity of tens of thousands of genes simultaneously [1-3]. The combination of this technology with the sequencing of the human genome has led to the development of total genome microarrays that allow systems biologists to view the comprehensive transcriptional activity of a cell or tissue type [4-6]. These total genome datasets possess a degree of richness that allows complex cellular regulatory mechanisms to be studied at a new level of detail.

These large datasets, however, also pose a myriad of analytical challenges. Data for over 30,000 genes, comprised of hundreds of thousands of individual data points, must be organized in a meaningful manner. In addition, for any given gene and condition, there are usually three replicates, and given the signal-to-noise characteristics of most microarray platforms, creative approaches must be used to analyze the data in a useful, statistically rigorous manner. Yet, when these aspects are taken into account, there is a vast potential for mining these datasets to uncover new biology.

1.2. Genome-wide Transcriptional Analysis of Endothelium

Vascular endothelium comprises a dynamic interface between blood and the vascular wall. Their intact function is essential for the regulation of many vital responses, including inflammation, haemostasis and vasodilation/constriction. Although this cell type is ubiquitous, it is far from being homogenous; for example, endothelial cells in the pulmonary, cardiac and brain microvessels [7] and the high endothelial venules of lymphoid tissue [8, 9], are specialized to cater to the special needs of their organs. Arterial and venous endothelial cells throughout the body have different morphologies, protein synthesis and levels of permeability [10, 11].

Discovering the similarities in the expression profiles between different endothelial cells will help us to define the set of genes required for basic endothelial identity. Determining the differences between these profiles will help elucidate which pathways confer the unique properties of specialized endothelial cells. In addition to developmental differences, a wide array of environmental factors can also affect endothelial cell phenotype; cytokines, hormones, metabolic products, hydrostatic pressures and flow-induced shear stress all modulate endothelial function [12, 13]. Studying the genome-wide transcriptional changes caused by such external stimuli can shed light on the regulatory mechanisms governing endothelial cell structure and function.

Our laboratory has recently embarked on an effort to extend our transcriptional profiling studies of endothelium [14, 15] by applying total genome microarray technology. Our recent foray into analyzing the transcriptional activity of endothelial cells at a global level has been enabled by a single-channel microarray platform containing 33,096 probes representing the entire genome of 29,791 genes, based on 60,808 transcripts. The ability to collect data simultaneously in such a comprehensive manner allows us to explore the transcriptional biology of endothelium in novel ways. Most immediately, the ability to assess the expression level for every gene allows us to define a putative endothelial transcriptome—the set of genes that are required for endothelial identity and are expressed at some detectable level under standard culture conditions. We can then assess how this global expression profile changes under different relevant stimuli. As data is collected for additional experimental conditions, we will be able to begin to decipher the complex gene regulatory networks governing endothelial function.

This study has developed computational techniques to define global expression profiles and to detect differential expression between profiles. We demonstrate the efficacy of these

techniques by defining the genome-wide expression profile for cultured human umbilical venous endothelial cells (HUVEC) and studying the changes caused by a potent inflammatory stimulus, IL-1. IL-1 was chosen as the first stimulus to examine because it has been previously well characterized and yet has been known to involve several pathways [16], thus increasing the potential for the discovery of new genes regulated in endothelium.

2. Genome-wide Expression Profile of Cultured HUVEC

2.1. Two-Population Hypothesis of Gene Intensity Distribution

Determining which genes are expressed in a given cell type under a set of standard control conditions sheds light on the genes that are required for the basic function of that cell type. Biologically, the question posed is: under defined baseline conditions, which subset of genes from the human genome are "turned on" or expressed, and which are the remaining genes that are "turned off" or not expressed. When microarray data is to be used to determine gene expression levels, the corresponding computational question becomes: what signal intensity corresponds to a gene transcript being expressed above noise levels?

Consider the distribution of intensity signal values (Fig 2.1a); i.e., the frequency at which signal values are observed among all the spots on a single microarray for different small ranges of signal values. This frequency corresponds to p(x), the probability of a random spot's intensity having a value of x. Given the biological premise that any gene falls into one of two categories, expressed or non-expressed, this distribution is actually comprised of two separate distributions—$p_E(x)$, the probability distribution of intensity values for a randomly selected spot detecting an expressed gene and $p_N(x)$, the probability distribution of intensity values for a randomly selected spot detecting a non-expressed gene. The net probability distribution, p(x), is a weighted sum of the two individual distributions, $f^* p_N(x) + (1-f)^* p_E(x)$, where f is the fraction of spots detecting non-expressed genes (Fig 2.1b).

This proposed mixed distribution is supported by the skewed shape of the net frequency distributions shown in Fig. 2.1a; the left tails resemble Gaussian distributions, which are commonly occurring distributions for noise (i.e., "signals" of spots probing non-expressed genes, that are caused by non-specific hybridization, instrument measurement error, etc.), but the right

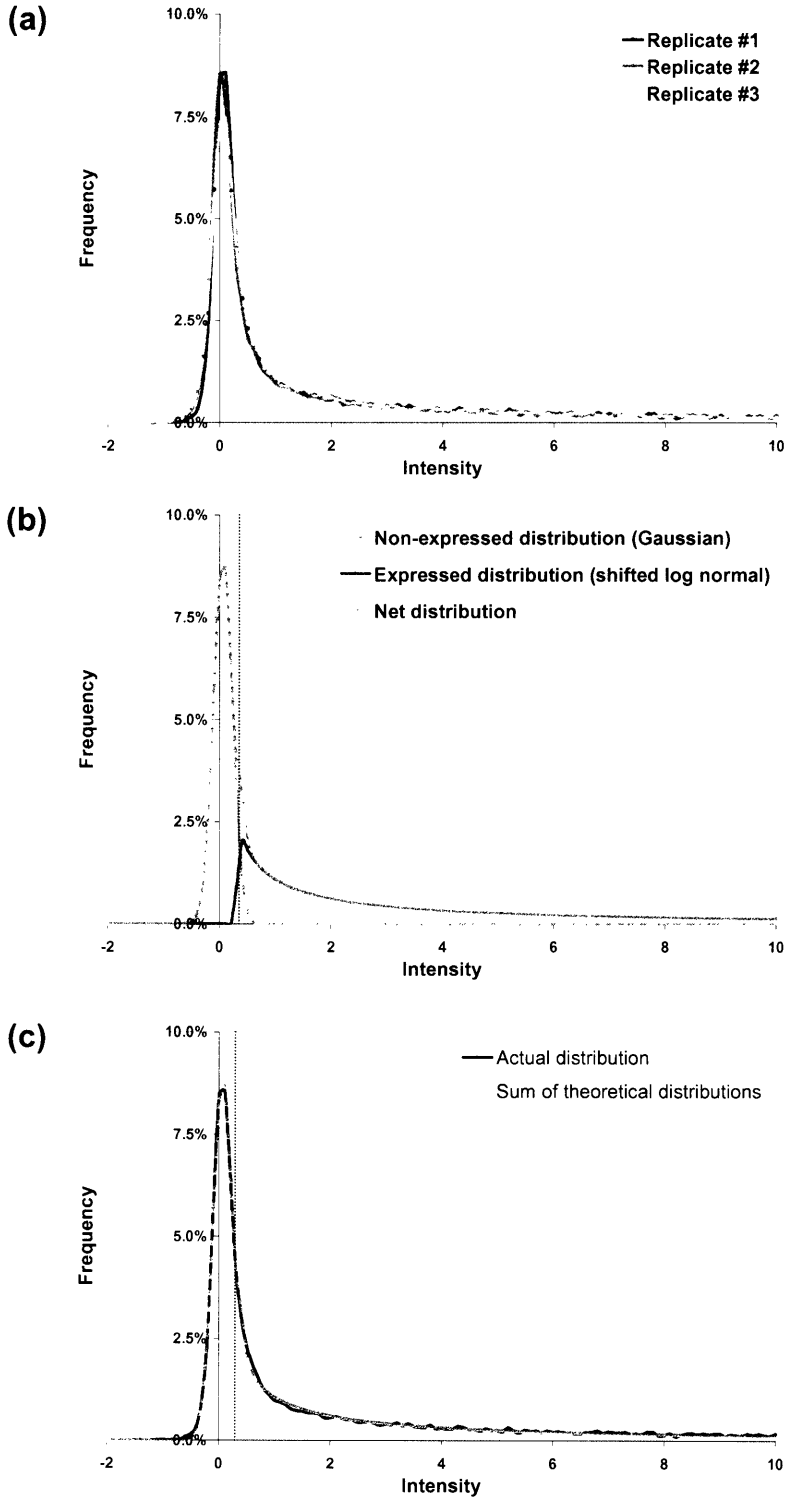

Fig. 2.1. *(a) Histogram of signal intensities for three independent replicate microarrays of cultured HUVEC. (b) Theoretical signal distributions of expressed and non-expressed spots and the net distribution generated by their sum. (c) Theoretical net signal distribution and actual signal distribution for replicate #1. The dashed line indicates the cutoff signal value for expressed genes that maximizes the theoretical true classification rate.*

tails fall off much more gradually. We hypothesized that this net distribution could be decomposed into a non-expressed Gaussian distribution (gray line, Fig. 2.1b) and an expressed distribution, which we modeled as a shifted log-normal distribution (red line, Fig. 2.1b). (A log-normal distribution implies that the logged values of the signal intensities follow a Gaussian distribution.). These two separate distributions could then be used to determine appropriate cutoffs for classifying a spot as representing an expressed on non-expressed gene.

2.2. Experimental Methods

2.2.1. Cell culture. Consistent culture conditions are important to generate a reliable gene expression profile of HUVEC since several factors such as cell cycle point, age, passage number, confluency or media can affect transcription. HUVEC were isolated from normal term cords and pooled from 5 to 7 donors were cultured in complete media supplemented with 20% fetal calf serum, 2mM L-glutamine, 50 mg/ml endothelial cell growth supplement, 100 mg/ml heparin and 100 unit/ml penicillin-G1100 mg/ml streptomycin, and incubated at 37°C in 5% CO2 in humidified air. Cells from the first subculture were plated at an initial density of 70,000 cells/cm^2 and grown for 24 hours, a time point at which we have documented that only 4-6% of the cells are in G2M phase [14]. Fresh media was added at this time point and cells were collected 4 hours later.

2.2.2. RNA isolation and purification. Cells were rinsed twice with PBS before collection, then scraped into Trizol. Total RNA was isolated by using TRIzol reagent (Invitrogen, Carlsbad, CA) and the RNeasy kit (Qiagen, Valencia, CA). Total RNA was DNase-treated with the RNase-free DNase kit (Qiagen) according to the manufacturer's protocol and purified on RNeasy mini columns (Qiagen). RNA quality was verified by Agilent's 2100

Bioanalyzer with RNA 6000 Nano LabChip Kit. The concentration of RNA was measured by spectrophotometric analysis at 260 nm.

2.2.3. Microarray preparation and scanning. Labeling, hybridization and scanning were performed according to the manufacturer's protocols for the AB1700 microarray scanner using total human genome microarrays (Applied Biosystems, Foster City, CA), with 30,096 spots representing 28,790 different genes. Each spot uses a 60 base pair probe that represents a region within the first 1500 base pairs of the 3' end of the target mRNA. Briefly, 40 μg of purified total RNA for each sample was used in an RT reaction that incorporated digoxigenin label into the cDNA products. The cDNA was then purified using a DNA purification column, DNA wash buffer and DNA elution buffer supplied by the manufacturer. Purified cDNA was then hybridized at 55°C under agitation at 100 rpm for 16 hours, to a glass microarray slide that was pre-hybridized with AB1700 Blocking Reagent for 1 hour. Slides were then washed as per manufacturer's protocol, then incubated under agitation with anti-digoxigenin-AP antibody for 20 minutes. Microarrays were then treated with a Chemoluminescence Enhancing Solution. Finally, the chemoluminescence substrate was added and the array scanned within one hour.

2.3. Methods for Decomposing Spot Intensity Distribution

2.3.1. Recovery of negative intensity values. The Applied Biosystems 1700 microarray processing software quantifies the spot intensities from two images, one capturing the chemoluminescent signals from the top half of the microarray and the other capturing the chemoluminescent signals from the bottom half of the microarray. The software then performs a number of normalization steps on the image quantification data. First, the chemoluminescent signal from each spot is subtractively corrected with a control fluorescent signal. This correction procedure can result in negative values for very dim (i.e., non-expressed) spots, which are

necessary to observe the true Gaussian nature of the noise distribution. The software, however, maps negative and other low-valued spots to a positively valued surrogate, the standard deviation of the different individual pixel values for the spot. The fluorescent-corrected and surrogated values are then normalized to map into a standard dynamic range of values, so that different arrays can be compared to one another meaningfully. This normalization step produces a different normalization factor for each spot, but the correction factors for the spots from the same region (top image or bottom image) are very similar. Thus, in order to generate normalized but unsurrogated intensity values, first the normalization factors from all spots that were not surrogated were averaged to generate a single uniform normalization factor for each region. Then the pre-normalized signal-to-noise ratio for each spot was multiplied by the inter-pixel standard deviation to recover the unnormalized, unsurrogated signal value, which was finally multiplied by the appropriate average normalization factor. This technique maps intensities to an appropriate dynamic range while preserving negative values. Spots flagged as poor quality (flag > 10,000) were excluded from analysis.

2.3.2. Net frequency distribution of signal intensity values. Three microarrays, each representing an independent biological replicate of HUVEC cultured under the conditions described above, were used to develop a baseline endothelial expression profile. The frequency of occurrence of normalized, unsurrogated signal values was counted using bin sizes of 0.1 intensity units, centered from –5.0 to 4500.0. The frequency distributions were calculated for each microarray separately. Figure 2.1a illustrates that the three distributions are highly similar to each other, with the squared Pearson's linear correlation coefficient (R^2) between any two replicates being 0.99.

2.3.3. Decomposition of net frequency distribution. The net frequency distribution was modeled as $p(x) = f^* p_N(x) + (1-f)^* p_E(x)$, i.e., the weighted sum of a Gaussian distribution, $p_N(x) = \frac{1}{\sigma_N \sqrt{2\pi}} e^{-(x-\mu_N)^2/(2\sigma_N)^2}$, where the mean, μ_N, and standard deviation, σ_N are adjustable parameters, and of a shifted log-normal, $p_E(x) = \begin{cases} \frac{1}{\sigma_E \sqrt{2\pi}} e^{-(\ln(x-x_0)-\mu_E)^2/(2\sigma_E)^2} & , x > x_0 \\ 0 & , x \le x_0 \end{cases}$, where the mean, μ_N, and standard deviation, σ_N and shift, x_0, are adjustable parameters. In addition, the relative fraction of the two distributions, f, is also an adjustable parameter. The associated cumulative distribution function for p(x) was used to determine the probability $p(x_1 < x < x_2)$ for each bin ranging from x_1 to x_2. The error between the actual and theoretical distributions was calculated as the sum of the squared differences between the actual frequency of occurence and theoretical cumulative probability for each bin. Starting with initial values of f= 0.5, μ_N=0, σ_N=0.5, μ_E=1, σ_N=1 and x_0=0.5, and applying the constraint that $x_0 > \mu_N+\sigma_N$, Microsoft Excel Solver, which applies the Generalized Reduced Gradient (GRG2) Algorithm for optimizing nonlinear problems [17], was used to determine a set of parameter values that minimized the error. This analysis was performed separately for each of the three replicate arrays. The final parameter values and root mean square (RMS) error (square root of the mean squared error across bins) are given in Table 2.1. The theoretical (dashed blue line) and actual (solid black line) distributions for the first replicate are shown in Fig. 2.1c, illustrating that the two curves are extremely similar.

2.3.4. Selecting a cutoff for classification of spots as expressed or not expressed. The two distributions, as seen in Fig. 2.1b, overlap with one another; thus any intensity cutoff used to classify spots as expressed or non-expressed will generate a certain number of false positives and false negatives. We generated ROC curves, shown in Fig. 2.2a, that graph the theoretical

Replicate	#1	#2	#3
Minimum signal value	-2.70	-1.66	-3.68
Maximum signal value	3245.93	3535.44	4429.74
f	0.38	0.37	0.40
μ_N	0.06	0.02	0.06
σ_N	0.17	0.17	0.21
μ_E	1.62	1.61	1.73
σ_E	2.09	2.02	2.04
x_0	0.30	0.20	0.38
RMS Error	6.9E-03	1.0E-02	6.8E-03
ML Cutoff	0.35	0.31	0.44
True Classification Rate	98.5%	96.6%	97.7%

Table 2.1. *Statistics and parameter values for two-population fit of signal intensity distributions from 3 replicate microarrays of cultured HUVEC RNA.*

sensitivity (true positive rate) vs. 1-specificity (false positive rate) for a range of cutoff values, to demonstrate the effect of choosing different cutoff values. That the ROC curves lie close to the left and upper borders of the graph indicate that the two distributions are well separated despite their overlap, and that a cutoff with low false classification rates can be selected.

One must therefore select a cutoff that meets some desired criteria for the false classification rates. For example, to reduce the false negative rate to 0, one should choose a cutoff $c < x_0$, guaranteeing that every spot belonging to the log normal (expressed) distribution will be classified as expressed; the tradeoff is, of course, a very high false positive rate. The other extreme choice to reduce the false positive rate to 0 by choosing a cutoff $c >> \mu_N + 4\sigma_N$, which lies far into the right-hand tail of the Gaussian (non-expressed) distribution; the tradeoff in this case would be a very high false negative rate. A more balanced option is to choose a cutoff c that maximizes $f*p_N(x<c)+(1-f)*p_E(x>c)$, the theoretical net true classification rate. A fourth option is to choose the cutoff to be the p^{th} percentile of all values, so that the highest (1-p)% of spots are classified as expressed. Fig. 2.2b demonstrates the different theoretical true positive,

Fig. 2.2

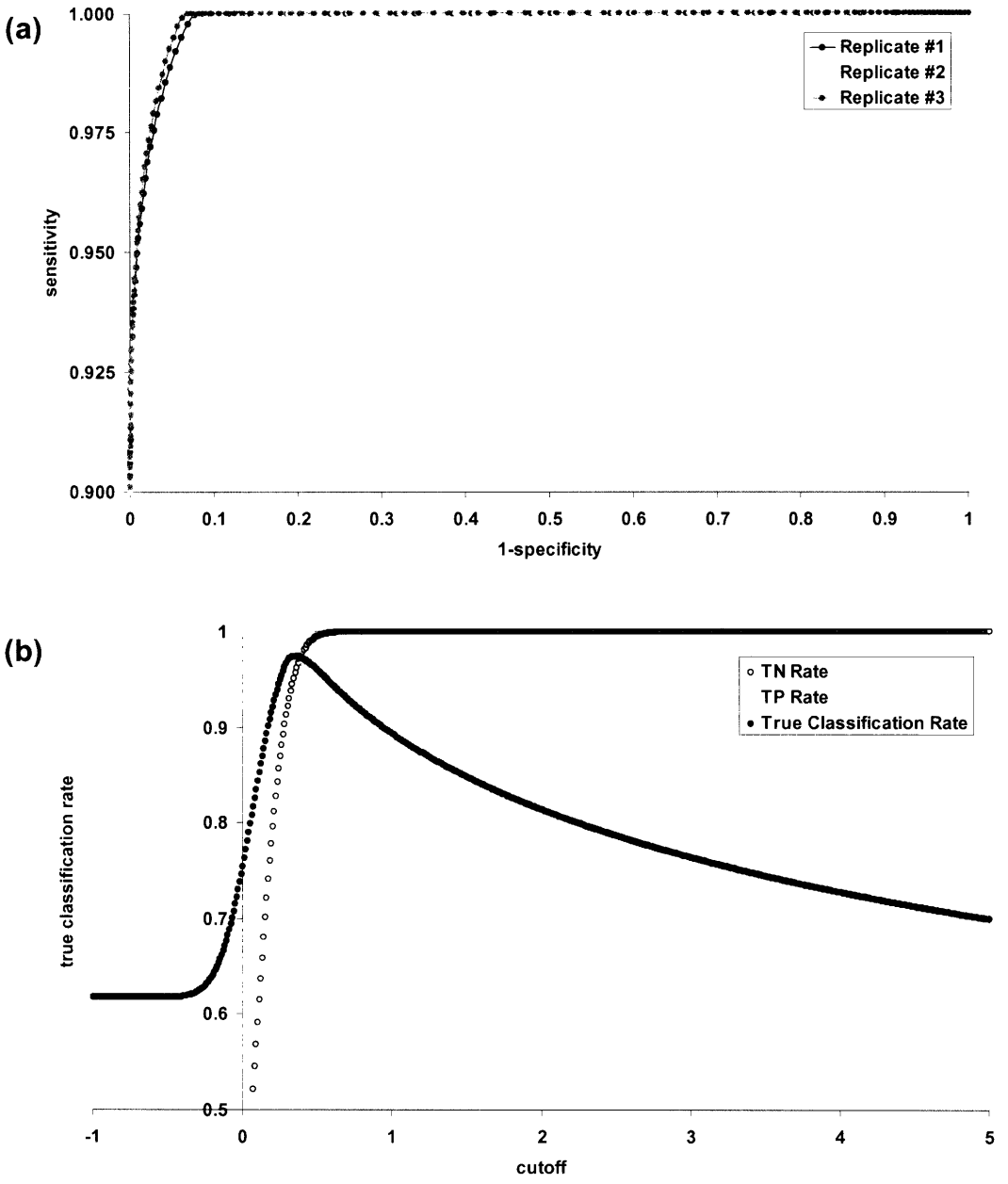

Fig. 2.2. *(a) Receiver operating curve (ROC) plotting the theoretical sensitivity vs. 1-specificity for different cutoffs. Data shown for all three replicates. (b) Changes in theoretical true positive, negative and net classification rates as cutoffs are varied for replicate #1.*

true negative and net true classification rates for a range of cutoffs, using the theoretical distribution fits for replicate #1.

We chose to use a cutoff for each array that maximized the theoretical net true classification rate. The cutoff and corresponding theoretical true positive and true negative rates for each array are given in Table 2.1. Finally, we selected spots that were classified as expressed on all replicates of good quality (for most spots there were 3 good quality replicates) to generate a list of expressed genes in quiescent cultured HUVEC, which included 18,472 (56%) spots representing 16,026 (54%) genes.

2.4. Patterns of gene expression in cultured HUVEC

2.4.1. Chromosome distribution. Fig. 2.3a shows the physical chromosome location of all annotated genes on the array in red and Fig. 2.3b shows the location of all expressed genes on the array in blue. HUVEC appear to express genes found on every single chromosome, and no regions appear to be enriched in expressed genes compared to the chromosomal distribution of all annotated genes.

2.4.2. Functional categorization of expressed genes. One of the most tractable ways of analyzing this genome-wide expression profile for cultured HUVEC is to group expressed genes by their function. One of the most widely used functional categorizations of genes is the Panther system [18]. The detailed Panther molecular function categories were partially merged to generate broader categories and the "molecular function unclassified" category was removed; these simplified categories are given in Table 2.2. Of the 33,096 spots on the array, 14,993 (45%) spots—representing 12,550 out of 29,791 (42%) genes—were associated with one or more simplified Panther categories. Of the expressed genes, 8,060 out of 16,207 genes

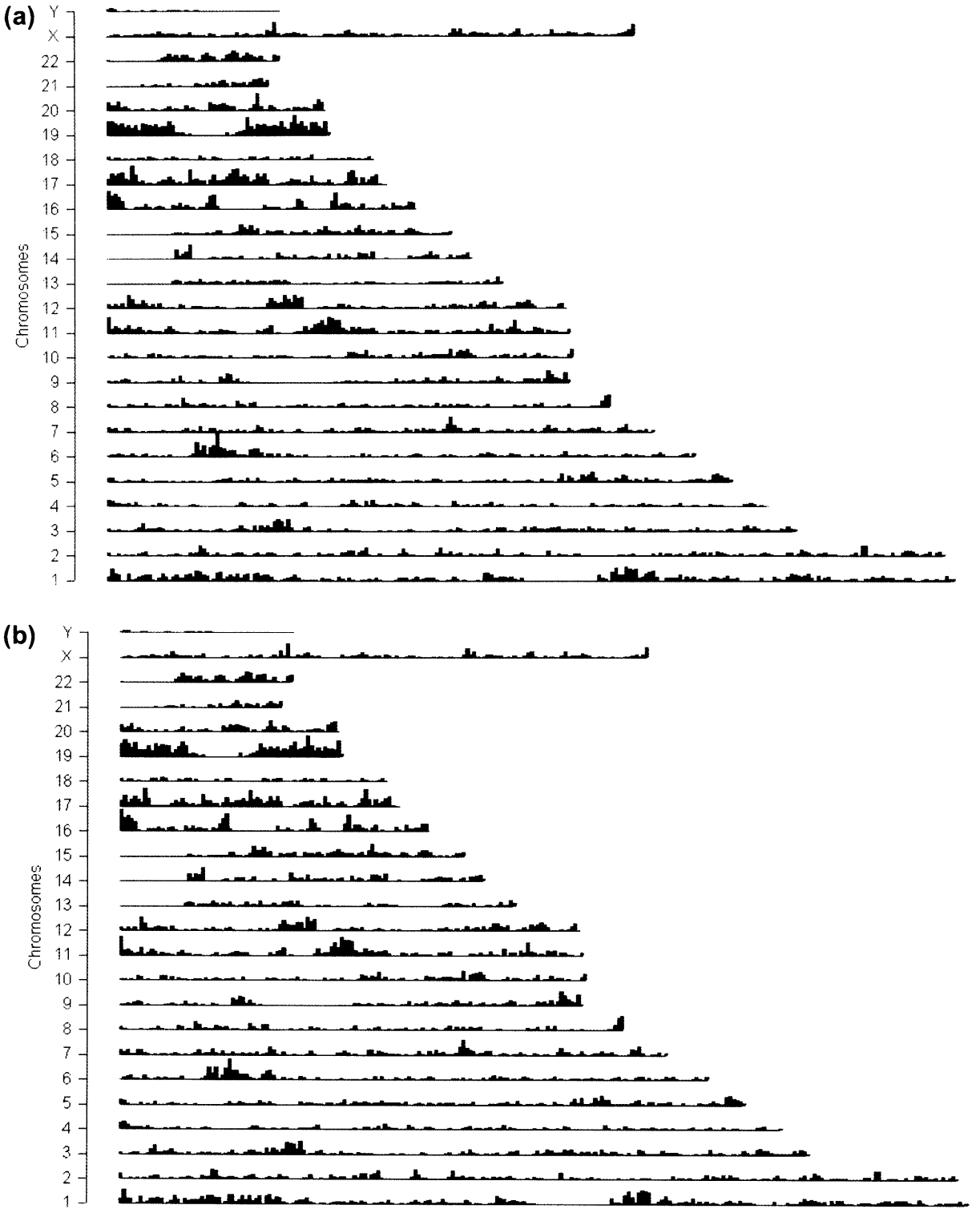

Fig. 2.3. *(a) Plot showing chromosomal location of annotated genes on microarray. Bar height corresponds to relative number of genes in given location. (b) Plot showing chromosomal location of annotated genes on microarray that were also classified as expressed.*

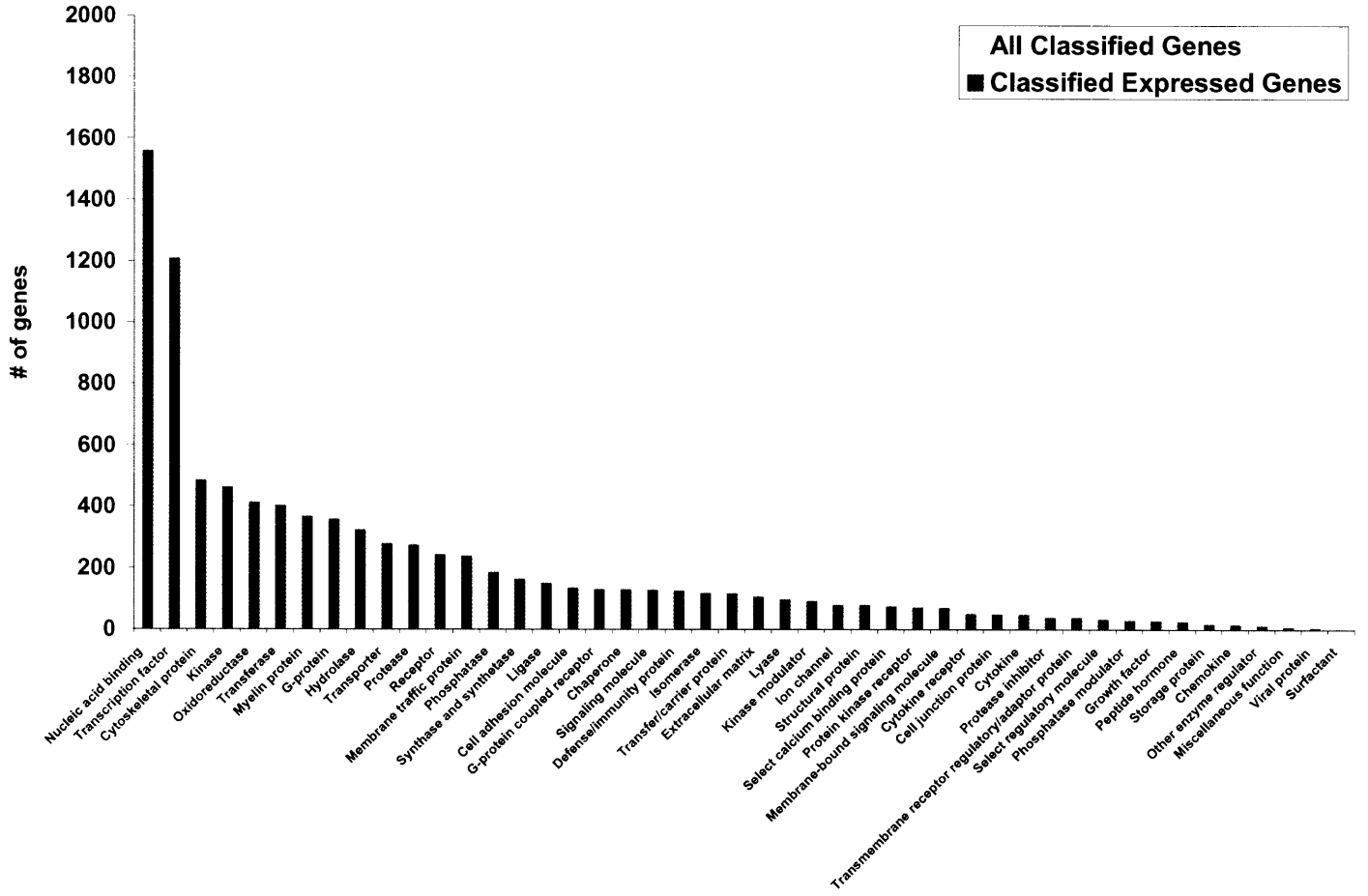

Fig. 2.4. *Histogram illustrating the total number of genes on the array (gray) and number of expressed genes (red) associated with each simplified Panther category.*

Simplified Panther Function	No. of Genes on Array	No. of Genes Expressed	Percentage of Category Enrichment	Percentage of Expressed Genes
Nucleic acid binding	1769	1559	88.1%	9.7%
Transcription factor	1674	1208	72.2%	7.5%
Cytoskeletal protein	730	484	66.3%	3.0%
Kinase	621	461	74.2%	2.9%
Oxidoreductase	614	412	67.1%	2.6%
Transferase	575	401	69.7%	2.5%
Myelin protein	476	366	76.9%	2.3%
G-protein	484	357	73.8%	2.2%
Hydrolase	491	322	65.6%	2.0%
Transporter	485	277	57.1%	1.7%
Protease	467	273	58.5%	1.7%
Receptor	633	243	38.4%	1.5%
Membrane traffic protein	309	237	76.7%	1.5%
Phosphatase	251	184	73.3%	1.1%
Synthase and synthetase	196	163	83.2%	1.0%
Ligase	180	148	82.2%	0.9%
Cell adhesion molecule	299	132	44.1%	0.8%
G-protein coupled receptor	642	129	20.1%	0.8%
Chaperone	147	128	87.1%	0.8%
Signaling molecule	229	127	55.5%	0.8%
Defense/immunity protein	458	124	27.1%	0.8%
Isomerase	124	117	94.4%	0.7%
Transfer/carrier protein	194	116	59.8%	0.7%
Extracellular matrix	229	105	45.9%	0.7%
Lyase	142	97	68.3%	0.6%
Kinase modulator	117	91	77.8%	0.6%
Ion channel	280	78	27.9%	0.5%
Structural protein	173	78	45.1%	0.5%
Select calcium binding protein	146	74	50.7%	0.5%
Protein kinase receptor	105	69	65.7%	0.4%
Membrane-bound signaling molecule	116	68	58.6%	0.4%
Cytokine receptor	81	49	60.5%	0.3%
Cell junction protein	84	47	56.0%	0.3%
Cytokine	132	46	34.8%	0.3%
Protease inhibitor	101	37	36.6%	0.2%
Transmembrane receptor regulatory/adaptor protein	58	37	63.8%	0.2%
Select regulatory molecule	39	32	82.1%	0.2%
Phosphatase modulator	38	28	73.7%	0.2%
Growth factor	71	26	36.6%	0.2%
Peptide hormone	82	24	29.3%	0.1%
Storage protein	23	16	69.6%	0.1%
Chemokine	40	15	37.5%	0.1%
Other enzyme regulator	18	11	61.1%	0.1%
Miscellaneous function	6	6	100.0%	0.04%
Viral protein	8	3	37.5%	0.02%
Surfactant	12	1	8.3%	0.01%

Table 2.2. *Summary of simplified Panther classifications for annotated genes determined to be expressed in cultured HUVEC.*

associated with one or more simplified functional categories. The number of expressed genes compared to the total number of genes on the array for each category is shown in Fig. 2.4.

Table 2.2, which gives the number of expressed genes as a percentage of the number of genes on the entire array for each category, indicates that genes from every simplified category are classified as expressed in cultured HUVEC. Of especial note are the nucleic acid binding, chaperone and isomerase categories, which are enriched by over 85%.

We examined the top 5% of expressed genes, ranked by average intensity, to study which genes were the most highly expressed. (A list of all expressed genes, their average intensities and percentiles can be found at http://vessels.bwh.harvard.edu/software/endo_xcriptome .) A remarkable 57% of these genes were classified as related to nucleic acid binding. 422 of these genes were ribosomal, ribonuclear or other RNA-binding proteins, 23 were translation initiation or elongation factors and 6 were histones, DNA helicases or chromatin-binding proteins. In addition, 4% of the highest expressed genes coded for cytoskeletal proteins, mostly tubulin and actin-related, and another 4% of these genes coded for isomerases. Thus, rather than being endothelial-specific, the highest expressed genes are mostly related to the function of any transcriptionally active cell, and we would expect to see most of these genes expressed at high levels for most other cultured cell types.

However, certain genes known to be crucial for endothelial function and identity were also found within the top 5% of expressed genes, including PECAM/CD31 and Hsp-90 protein 1 alpha and beta. PECAM is considered to be an endothelial cell-surface marker, also plays an important role in regulating endothelial permeability, cell signaling and cell survival [19]. Hsp90 associates with and activates eNOS, the most important vasodilatory molecule produced by endothelial cells [20]. Other highly expressed genes—within the top 15% of expressed

genes—include vonWillebrand factor, a molecule important in hemostasis that is found in endothelial-specific Weibel-Palade bodies and may be a shear-stress sensitive molecule [21] and reticulon 4/NOGO, which is a regulator of vascular remodeling [22].

A number of vasoactive proteins in addition to Hsp90 are expressed at high levels. These genes include caveolin-1, which has been shown to be prominent in vascular endothelium [23] as well as caveolin-2, both of which binds and disables eNOS [24], and endoglin, which upregulates eNOS protein expression [25]. Suprisingly, eNOS itself is expressed, but at approximately the 45th percentile of expressed genes, a lower level than some of its regulators. Several genes from the endothelin pathway, the most important vasorestrictive system in endothelial cells[26], are also expressed, including endothelin-1, as well as endothelin converting enzyme 1 [27] at high levels and endothelin converting enzyme 2, shown to be expressed previously in HUVEC [28], at lower levels (intensity of 1.3, 15th percentile of expressed genes). Surprisingly, the endothelin receptor B, the chief endothelin receptor for endothelial cells [29], is expressed in cultured HUVEC, but at a very low level (average intensity ~0.50, 0.7th percentile of expressed genes). If this result is true, it suggests that perhaps the receptor's turnover from the cell surface membrane is low. Other vasoactive factors that are expressed in HUVEC include ACE-1 and 2, and adrenomedullin.

The Notch pathway, initially known for its role in neuronal development, has been shown to play an important role in vascular development [30, 31] and injury response [32]. Several genes involved in this pathway were found to be expressed in cultured HUVEC, including notch-1, 2 and 3 homologs, delta-like 1, 3 and 4, deltex 2 and 3, jagged 1 and 2 (jagged 1 among the top 15% of genes), presenilin-1 and 2 and suppressor of hairless. Two transcription factors downstream of suppressor of hairless [33] were also expressed in cultured HUVEC at

appreciable levels, HEY-1 and HEY-2. HEY-2 is considered to be important in embryonic vascular development [34] and may also be an arterial-specific marker [35].

A variety of other transcription factors known to play an important role in endothelial function were shown to be expressed in cultured HUVEC through our analysis. Several members of the Kruppel-like factor family, KLF-2, 3, 5, 7, 14, 15 and 16 are expressed, the highest of which is KLF-2, which appears to be a anti-inflammatory and pro-vasodilatory transcription factor in endothelial cells [36]. Three myocyte enhancing factors, MEF2A, C and D, were all shown to be expressed. MEF2A has recently been implicated as an endothelial gene whose mutation causes inherited cardiovascular disease [37].

Those transcription factors whose roles are not well characterized in endothelium include several members of the foxhead box family, which may play a role in regulation of cell proliferation of HUVEC [38], the dachshund homolog, which has been shown to play an important role in optic development [39], as well as several jumonji family genes that are thought to have a role in neural development [40] but are not further characterized in mammalian systems. The expression of these transcription factors in endothelium may provide further insight into their transcription factors.

The pattern of expression of arterial and venous markers in cultured HUVEC is extremely interesting. A number of Ephrins (A1-5, B1 and 2) and their Eph receptors (EphA2 and 4, B1, 2, 4 and 6). The highest expressed of these, Ephrin B2, a putative arterial marker, and EphB4, a venous marker [41], are both expressed at similar levels (over the 75^{th} percentile among expressed genes). Expression of both these molecules together is generally not seen after development. Interestingly, cultured HUVEC appears to express other markers of both veins and arteries, such as HEY-2 and neuropilin 1, which are arterial, and Tie-2 and neuropilin 2, which

are venous [30]. This dichotomy may be explained by the unique physiological role of HUVEC since the umbilical vein receives low flows but oxygenated blood, and is meant to atrophy after birth. It may also be a result of phenotypic drift after being cultured.

Several endothelial development and angiogenesis-related genes are also expressed in cultured HUVEC. Among these are tissue inhibitor of metalloproteases 2, which is one of the top 5% of expressed genes, thought to mediate inhibition of angiogenesis [42], and angiopoietin-2, among the top 15% of expressed genes, known to control the ratio of arterial/venous vessels during angiogenesis [43]. Among the 20 collagens that are expressed in cultured HUVEC, the two highest expressed are collagen IVα2, the primary component of basement membranes, which would be important for newly plated HUVEC to lay down, and collagen XVIIIα1, whose C-terminal fragment is endostatin, a potent anti-angiogenic factor [44]. IL-8, which is both a chemokine and an angiogenic stimulus [45], is also expressed appreciably in cultured HUVEC. The role of these genes under these circumstances could be an influence of culture conditions, or these genes could possibly play other important roles in endothelial function outside of their angiogenic roles. For example, VEGF-A (expressed at low levels) and VEGF-B and C (expressed at higher levels) are considered to be pro-angiogenic factors, but they also play a role in mediating endothelial permeability.

Endothelial cells play an important role in regulating inflammatory responses. The genes regulating this function would most likely not be expressed until the cells were exposed to an inflammatory stimulus, but surprisingly, two adhesion molecules known to be induced by inflammatory conditions, E-selectin and VCAM, both have intensities above the 30th percentile of expressed genes. These genes may be false positives, or alternatively, may indeed be expressed at extremely low basal levels in quiescent conditions. Over 35 MAP kinases and MAP

kinase-related proteins, which are known to be important in the signaling pathways triggered by inflammatory responses [46], were found to be expressed basally, presumably primed for being activated by upstream factors in the case of an inflammatory response.

Several genes involved in hemostasis, which is closely linked to the inflammatory response, appear to be expressed in cultured HUVEC, including calmodulin, which is among the top 5% of expressed genes and is involved in von Willebrand factor-dependent shear-induced platelet aggregation [47], and thrombomodulin, which is an anti-thrombotic factor. Culture conditions that do not mimic the blood flow endothelial cells see in a physiological setting could affect the expression of hemostatis-related genes.

Interestingly, the expression data supports recent hypotheses regarding endothelial function. For example, tight junctions, thought to be primarily an epithelial feature, have recently been shown to exist in dermal microvascular endothelium [48]. Our data show that several genes involved in tight junction formation are among the highest expressed genes in HUVEC, including connexin 43 (in the top 6% of expressed genes) as well as ECAM, zona occludens 1 and 2 (in the top 20%). Another hypothesis sparking much discussion is the possible role of endothelial cells as professional antigen presenting cells. They have been shown to play such a role in the liver [49] and small intenstine [50], but whether endothelial cells play this role in a generalized fashion is still under debate. A few MHC class II genes, coding for the molecules used by professional antigen-presenting cells, are expressed in cultured HUVEC, the highest of which was HLA-DPA1 (average intensity 10.2, approximately 60^{th} percentile of expressed genes). A similar number of MHC class I genes were also expressed, but at higher levels (65^{th} to 90^{th} percentiles). Studies have shown that the development of the vascular and neural system are linked. Interestingly, several of the genes expressed in cultured HUVEC are

well known for their neural role, such as neuropilin-1 and 2 [51], neurexin-2, a neural cell surface protein [52] and adrenomedullin and MEF2A, both of which influence neuronal differentiation [53, 54].

2.5. Caveats of Expression Profile Analysis

A total-genome microarray system allows one to examine the entire pattern of gene expression across the genome, and provides new data to estimate how many genes are expressed in a typical mammalian cultured cell type. Our estimate of approximately 16,000 genes is about 50% greater than previous estimates of approximately 10,000 genes in an endothelial-derived cell line [55]. Several factors could contribute to this difference. For one, our analysis does not necessarily exclude genes that may have only a few transcript copies; we seek only to distinguish between signals due to measurement noise and signals due to specific hybridization. We also make the assumption that the noise level is identical for every gene, when in fact the binding properties of each probe may be slightly different; a signal level of 1 may indicate specific hybridization for one probe while it may represent non-specific hybridization for another probe. Finally, the results are highly dependent on the choice of distributions. We have made the assumption that the noise distribution itself is a symmetric Gaussian; if this noise distribution itself is actually skewed towards the right, then our assumption would be generating additional false positives. However, the strong matching between the theoretical and actual distribution well, as seen in Fig. 2.1c, supports the use of the current choice of distributions.

3. Statistical Detection of Regulated Genes Using Intensity-Based Variance Estimation*

3.1. Statistical Methods for Detecting Differential Expression from Gene Microarray Data

3.1.1. The Necessity for Statistical Detection of Differential Expression. Biologists can now use microarray technology to determine the expression levels of tens of thousands of genes simultaneously, in less time than it previously took to measure the expression level of a single gene. However, there remains the challenge of processing the microarray data from array images into a format that best facilitates the discovery of new biological insights. The potential of gene microarray technology is limited without an estimate of the statistical significance of the observed changes in gene expression. Algorithms beyond standard statistical methods, such as the Student's t-test, are necessary to produce reliable results. We strongly believe, as do others, that the quality of the data processing steps is critical to the overall success of a microarray experiment [56].

A typical data processing pipeline consists of several steps. (See [57] for a review, and see [58, 59] for a review of microarray processing software.) First, image analysis software locates the arrayed spots in the scanned image, quantifies the foreground and background brightness of each spot, and notes any irregularities in spot morphology. The background intensity value is then subtracted from the foreground intensity value. The background-subtracted intensity data from each array must then be normalized, or rescaled, to remove artifactual differences in signal brightness due, for example, to different labeling efficiencies that produced arrays of different overall intensity. Normalization techniques are often based on the assumption that a large number of spots will have similar expression levels between conditions.

* This chapter is modified from Comander, J.*, Natarajan, S.*, et al. *BMC Genomics*. 2004 Feb 27;5(1):17 (* equal contributors).

Curve-fitting techniques, such as a locally weighted regression, are used to equalize expression values between arrays, or between array channels for two-color arrays [60, 61]. After this normalization, the intensity values can be used by a variety of algorithms for detecting differences in expression between the measured biological conditions. This processing is applied whether two samples are compared directly or a "reference sample" experimental design is used. In a reference sample design, the same reference RNA sample is hybridized to one channel of all arrays, and the other channel is hybridized with each individual experimental sample. This design is often used when multiple biological conditions are being investigated and it becomes impractical to perform every pairwise combination of conditions directly [57, 62].

Given a list of normalized intensity values across various biological conditions, the next step is to determine which genes are differentially regulated among the conditions being studied. In the early days of microarray experimentation, an emphasis was placed on analyzing the data using exploratory data mining techniques, such as hierarchical clustering [63] and self-organizing maps [64]. Clustering algorithms measure the similarity between observed gene regulation patterns across the various conditions, and assemble clusters such that similarly regulated genes are grouped together. The resulting clusters produce an effective overview of the data, showing which of the many possible patterns of regulation are actually present in the data. Since these patterns are somewhat robust, a few erroneous spots are unlikely to change them dramatically. For a researcher who is simply interested in the overall pattern of the data, performing replicate arrays to reduce the number of errors is not particularly efficient. Many researchers choose instead to explore a greater number of experimental conditions.

Increasingly, microarrays are being used in a different context; researchers want to know with high confidence which *specific* genes are regulated across a small number of experimental

conditions (e.g., treatment vs. control, or mutant vs. wildtype). To answer this question, it becomes extremely important to use an accurate method to rank individual genes by their probability of truly being regulated, especially since this information may be used to plan more labor-intensive experiments around biological questions raised by a small number of such putatively regulated genes. In the absence of replicate arrays, the reliability of the data can be estimated (e.g. [65], [66]), but such "single slide" methods require a model of the expected noise characteristics of the system, a property that can potentially change between datasets. Performing replicate arrays can significantly improve predictions of differentially regulated genes, thereby decreasing the false positive (false detection) rate and false negative rate [62, 67, 68]. Using replicate arrays allows the calculation of more accurate significance estimates (p-values) that will aid in the interpretation of a list of "top regulated genes," which are commonly ranked by ratio alone.

Here we address the problem of accurately detecting genes that are significantly differentially regulated between a pair of biological conditions, given microarray datasets with a small number of replicates (e.g. N=3 arrays). If the number of replicates were very large (e.g., hundreds), the task would be relatively easy; since the ratio of expression levels between the two conditions would be well estimated by the average ratio or median ratio, the genes could simply be ranked by one of these estimates. In practice, however, the number of replicate arrays is rarely greater than 3, and estimates of average expression ratios are not always sufficiently accurate to predict which genes are truly regulated. The *variation* of a measured expression ratio is critical in determining whether the observed ratio is due to random measurement fluctuations or to a true difference between the quantities being measured. Genes with larger measured expression ratios between conditions are more likely to be truly regulated, while genes whose

ratios have a high measured variance are less likely to be truly regulated. This idea can be expressed mathematically as a test statistic where the numerator contains an estimate of the size of the effect, i.e. the ratio of gene expression intensities between conditions, and the denominator includes an estimate of the variance, i.e. the standard deviation of the ratio. A variety of such statistical tests have been applied to microarray data (reviewed in [57, 69]); the challenge is to choose the numerator and denominator of the test statistic such that it makes the best use of all available data in order to get the most accurate determination of which genes are most likely to be regulated.

3.1.2. Comparing Statistical Tests Used to Find Differentially Regulated Genes. The familiar Student's t-test (hereafter, "standard t-test") is the most straightforward method of calculating whether there is a significant difference in expression levels between conditions for each gene. Suppose that mRNAs from two biological conditions, "X" and "Y", are hybridized to a small number of replicate arrays (N two-color arrays or 2N one-color arrays). M_{avg}, the average logged ratio of expression levels between conditions X and Y, and its sample standard deviation, σ_M, are given by the standard formulas (see Methods). A standard t-statistic is calculated as $t = \frac{M_{avg}}{\sigma_M / \sqrt{N}}$. From this formula, it is clear that a large t-statistic (and the corresponding highly significant p-value) can occur because of either a large M_{avg} (high ratio) or a small σ_M (low noise). Although the standard t-statistic (or derivates thereof based on permutation [70] or Bayesian analysis [71]) can produce acceptable results for larger numbers of replicates (e.g., N=8), the results are less than satisfactory when applied to a small number of microarray replicates (e.g., N=3, Fig. 3.1). Fig. 3.1a shows data from an experiment that was repeated 6 times on two-color arrays. The six arrays were split into two random groups of three arrays, and the t-statistic described above was calculated for each gene in each group of three.

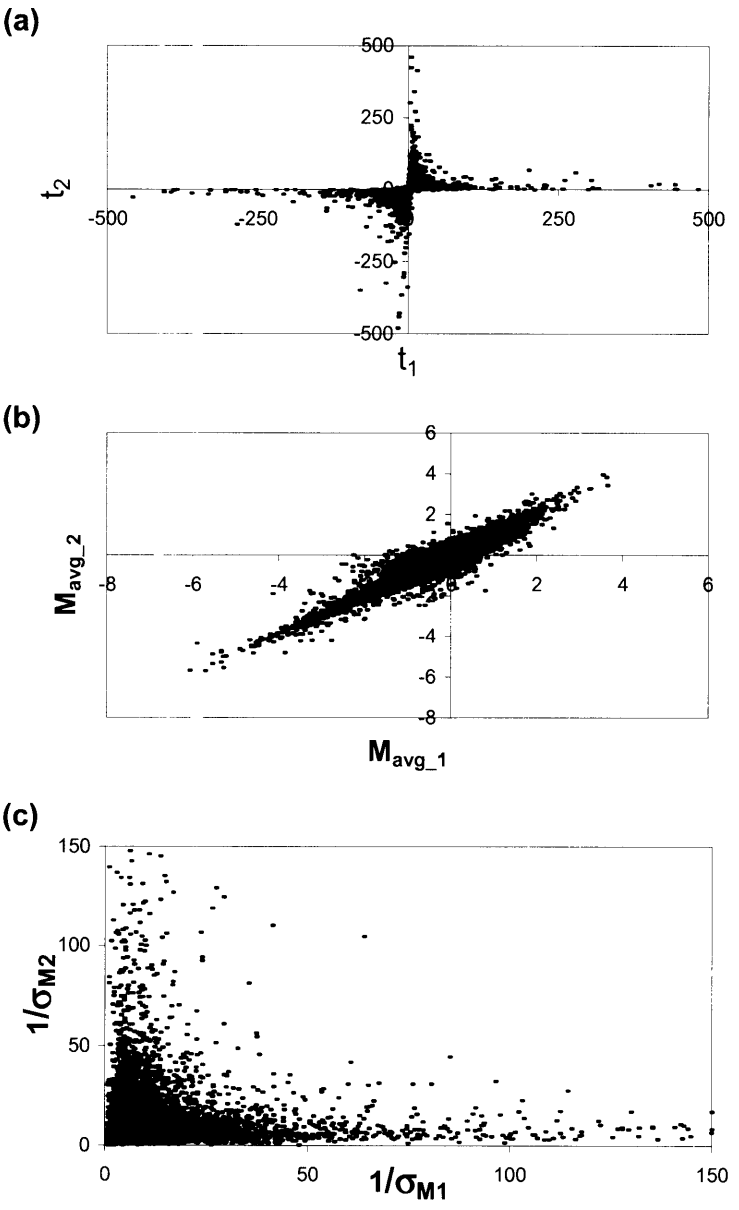

Fig. 3.1. *Evaluating the reproducibility of t-statistics between spots using a standard t-test. Two subsets of Dataset 4 each contain three replicate arrays derived from identical biological experiments. (a) Comparison of t-statistics for each subset. Values greater than ±500 are not shown. (b) Comparison of average logged ratios M_{avg}, which is the numerator of the t-statistic. (c) Comparison of the inverse of the standard deviation σ_M, which is in the denominator of the t-statistic. Values greater than 150 are not shown.*

The t-statistics from the two groups are graphed against each other in Fig. 3.1a. Although the two groups contain replicate arrays from the same experimental conditions, the t-statistic is clearly not reproducible between the groups. Fig. 3.1b and c demonstrate that M_{avg}, the numerator of the t-statistic, is more reproducible between the two groups, while $1/\sigma_M$, representing the denominator of the t-statistic, is not reproducible. This example highlights the major shortcoming of the t-statistic: due to random chance, the replicate ratios can occasionally be extremely similar, producing an artificially low σ_M and high t values. False positives stemming from this effect prevent the standard t-statistic from serving as a reliable or useful test of which genes are truly regulated.

To overcome this limitation, various modifications to the t-statistic have been proposed. First, a "penalized" t-statistic (also called a "moderated" or "regulated" t-statistic) can be used, where a constant value is added to the denominator. Tusher et al. use a penalized t-statistic of the form $\frac{M_{avg}}{\sqrt{(\sigma_M + s_0)^2 / N}}$ [72]. The addition of the constant s_0 prevents the denominator from becoming small for low σ_M, reducing the false positive rate of genes with unusually low σ_M. Choosing too large an s_0, however, effectively makes the denominator a constant, removing useful information about the variability of genes. Estimating the optimal s_0 for a particular dataset can be based on minimizing the coefficient of variation of the absolute t-statistic values ("SAM") [72], minimizing false positive and false negative estimates obtained through permutation ("SAMroc") [69], or simply choosing s_0 as the 90^{th} percentile of the σ_M values [73]. These studies have demonstrated that when ranking genes from a microarray dataset, a penalized t-statistic can perform better than a standard t-statistic in terms of decreasing the false positive and false negative rate [57, 69, 71-74], but it also has the potential disadvantage of showing bias against genes of high intensity [69].

An alternative to using a penalized t-statistic is obtaining a more precise estimate of the standard deviation σ_M. Such an estimate should be less susceptible to a chance concordance of measurements of M that occasionally produces an extremely low σ_M and a high t-statistic. For this purpose, knowledge of the relationships between the data points can be used to improve the estimate. Namely, the variance values, or σ_M^2, for one spot can be pooled, or smoothed, with the σ_M^2 values of spots that are likely to have similar variances. The variance of microarray data has often been observed to be a function of the spot intensity [65, 68, 74-83], raising the possibility that the variances of individual spots can be pooled with those of spots of similar intensity to produce a more precise estimate of the standard deviation. Several studies have taken into account this intensity-dependent heteroscedasticity. For example, Rocke et al. [80] and Newton et al. [66] have presented models of measurement error in microarrays that can explicitly take into account higher variance at lower expression levels. More general approaches to variance pooling have been implemented in a variety of ways, using loess-based curve fits [68], robust nonparametric spline fits [81] and sliding windows for calculating either local averages [79, 82, 83] or interquartile ranges [77]. These more reliable estimates of the standard deviation can be used directly to calculate Z-statistics, which are calculated according to the same formula as the standard t-statistic, but correspond to lower p-values [79, 84].

3.1.3. Strategies for pooling standard deviations. The studies cited above use methods that pool spots together based on their average intensity or logged intensity. For example, consider one set of replicate spots with an average intensity of 128 (2^7) in one channel and 16384 (2^{14}) in the other channel compared to a set of replicate spots with an average intensity of 1024 (2^{10}) in one channel and 2048 (2^{11}) in the other channel (Fig. 3.2). Since both of these sets of spots have the same average $\log_2$ intensity of 10.5, the standard deviations of their ratios would

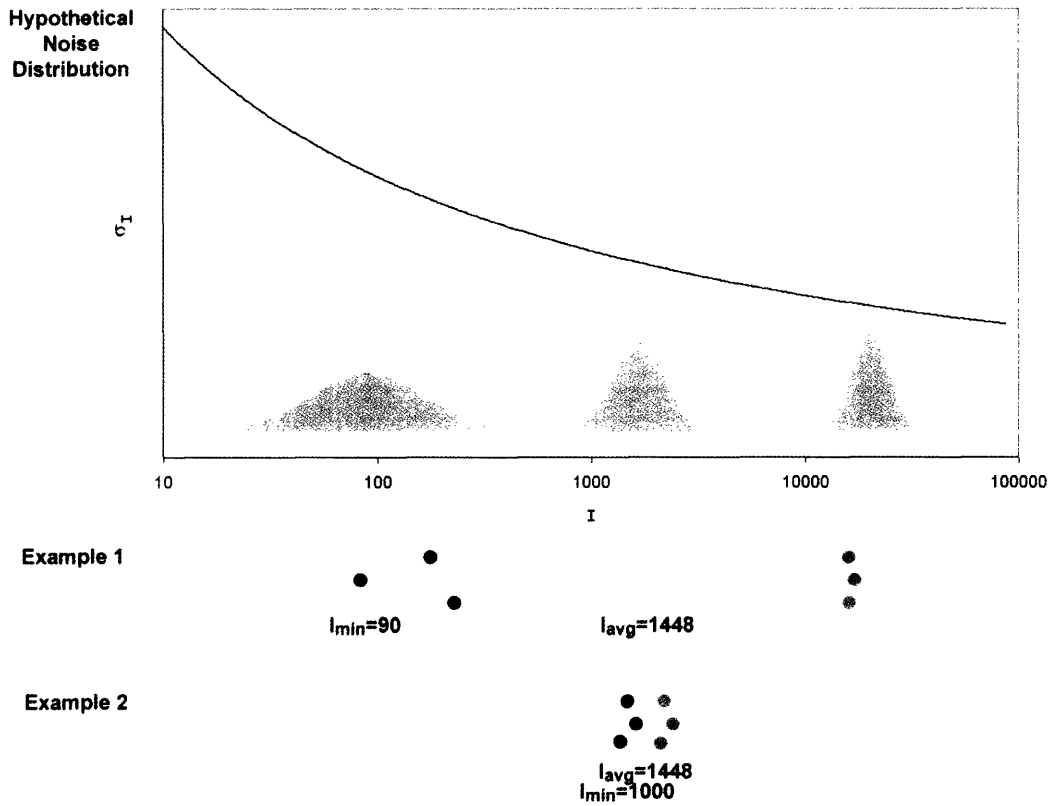

Fig. 3.2. *Motivation for pooling standard deviations by minimum intensity. A hypothetical noise distribution is given with higher noise at low intensities. Two sets of replicate spots (N=3 arrays) that have the same average intensity are shown. However, example 1 produces a higher standard deviation of the logged ratio compared to example 2, because example 1 contains very low intensity measurements that fall into the noisiest range of the intensity scale. In this case, the minimum intensity would differentiate between these two examples while the average intensity would not.*

be presumed to be similar and would be pooled together using the pooling methods described above. However, these spots may actually be expected to have quite different standard deviations; we have noted that many ratios with high variances result from spots that have a medium or high intensity in one channel and a very low intensity in the other (data not shown). Thus, the ratios for the first spot are expected to be more variable because of the very low intensities (~100) in one channel. In this study, we test the hypothesis that if spots are pooled together with other spots of similar *minimum* intensity over both channels (I_{min}), rather than *average* intensity over both channels (I_{avg}), then a larger proportion of the high-variance spots will be grouped together, resulting in a tighter fit of the pooled standard deviation curve to the actual variance and generating more accurate estimates of the standard deviation.

This study expands upon previous work on intensity-dependent variance estimation for microarray data by introducing a new metric, I_{min}, for pooling standard deviations. We evaluate the performance of the I_{avg} and I_{min} metrics by explicitly comparing the reproducibility and accuracy of the Z-statistics calculated using these two metrics. We also compare the performance of the Z-statistics to the performance of other statistical techniques in current use, the standard and penalized t-tests. Finally, we extend our technique for pooling standard deviations to two-color microarray data from a reference sample experimental design.

3.2. Methods

3.2.1. Data Acquisition. The analyses in this study were performed on five different datasets. Datasets 1-4 use the direct comparison experimental design, i.e. labeled cDNA from two biological conditions, "X" and "Y," were co-hybridized onto a single array. Each dataset was generated from a different biological experiment using two-color Agilent cDNA arrays. For Datasets 1-3, microarrays were prepared essentially according to the manufacturer's instructions

[85]. Briefly, 20 μg of total RNA were direct-labeled with Cy-3 and Cy-5, and labeled cDNAs were hybridized overnight to Agilent Human 1 cDNA arrays (G4100a, Agilent Technologies, Palo Alto, CA) containing 16,142 features representing approximately 10,500 unique genes. After washing, the microarrays were scanned in an Agilent model G2505A microarray scanner.

Dataset 3 contains 3 replicate two-color arrays with condition X in the Cy-5 channel and condition Y in the Cy-3 channel. Dataset 1 contains 3 replicates from another experiment, including one dye-swapped array; i.e. condition X in the Cy-3 channel and condition Y in the Cy-5 channel. Dataset 2 contains 3 replicate arrays without dye-swap, but each array was hybridized with a different amount of RNA, 5, 10 or 20 μg.

Dataset 4 consists of 23 replicate Agilent cDNA arrays from the Alliance for Cellular Signaling. The files MAE030201N00.txt to MAE030223N00.txt were downloaded from http://www.signaling-gateway.org/data/micro/cgi-bin/microcond.cgi. These arrays correspond to the conditions "B-cell + SIMDM exposure=0 minutes" vs. "Spleen". Four additional arrays are available for this condition (numbered MAE02070xN00.txt), but these arrays appeared to be slightly different from the other 23 arrays (using hierarchical clustering, data not shown) and were excluded from further analysis. The B-cell RNA was derived from 23 preparations, each from a different set of mice, while the spleen RNA was drawn from a single large pool (Rebecca Hart, Alliance for Cellular Signaling at the California Institute of Technology, Pasadena, CA, USA, personal communication).

Dataset 5 uses a reference sample design, where RNA from each experimental condition is co-hybridized on an array with a standardized reference RNA sample. Dataset 5 contains three replicates arrays for each experimental condition, for a total of 6 microarrays, generated in our laboratory. Each of the arrays contains a reference RNA sample in the Cy-3 channel. Three

have condition "X" samples in the Cy-5 channel and the other three have condition "Y" samples in the Cy-5 channel. Since corresponding biological specimens for conditions X and Y were prepared together for each replicate, a natural pairing exists for the condition X and Y arrays.

3.2.2. Computer Techniques. Statistical modules were programmed in Perl v5.8. Microsoft Visual Basic 6.0 was used to integrate the image processing and statistical modules.

3.2.3. Image Processing. For Datasets 1-3 array images were processed using Agilent Feature Extraction software version A.6.1.1. The Feature Extraction Software provides normalized Cy-3 and Cy-5 channel intensity values for each spot on an array (in the gProcessedSignal and rProcessedSignal fields of the output files). The default settings were used for all options. Quality control algorithms in the software detect unusual (poor quality) spots; spots were excluded from analysis that contained a nonzero value any of the following fields: IsSaturated, IsFeatNonUnifOL, IsBGNonUnifOL, IsFeatPopnOL, IsBGPopnOL, IsManualFlag. For a detailed description of the Agilent Feature Extraction software and the algorithms it uses, see the Agilent Feature Extraction Version 6.1 Users' Manual. Briefly, Agilent Feature Extraction determines the foreground value for each channel based on the pixel values in a fixed-size circle centered on each spot. The median of pixel values in a concentric ring around the circle, with an excluded region between the outer boundary of the circle and the inner boundary of the ring, gives the spot background value. The raw spot value is calculated as its foreground value less its background value. A surrogate raw value is assigned when the foreground value does not exceed the background value by two standard deviations of the spot's background pixel values. Intensity-based normalization between channels using a linear regression and a lowess curve-fit technique is then applied to remove any systematic dye incorporation biases.

Images were also processed using SPOT (CSIRO, New South Wales, Australia)[86], an R-based implementation which uses seeded region growing to determine the foreground pixels for each spot and morphological opening to determine the background value for each spot. The raw spot values, foreground less the background values, are normalized between channels using an intensity-based Loess implementation in R available in the maNorm function of the marrayNorm package of the open-source Bioconductor software (www.bioconductor.org). We considered three image processing techniques: Agilent Feature Extraction output alone, SPOT output alone with maNorm-based normalization and Agilent foreground (gMedianSignal and rMedianSignal columns) less SPOT background (morphG and morphR columns) with maNorm-based normalization.

3.2.4. Pooled Standard Deviations—Direct Comparison Design. Three replicate arrays were processed for each direct comparison experiment. To map intensities from different replicates onto similar scales without altering the absolute ratio values, we multiplied the intensity values on each array by a constant such that mean square error between the intensities of that array and the intensities of the first replicate array was minimized. The multiplicative factor for array j is given by $\frac{\sum_{g=1}^{G}(x_{1g}x_{jg}+y_{1g}y_{jg})}{\sum_{g=1}^{G}(x_{jg}^{2}+y_{jg}^{2})}$, where G is the total number of spots and x and y are intensities for condition X and condition Y. Then, for each spot, the mean and sample (measured) standard deviation (σ) across array replicates were calculated for the logged ratio M = X-Y, where X and Y are $\log_2(x)$ and $\log_2(y)$. The sample standard deviation of M, σ_M, is calculated as $\sigma_M = \sqrt{\frac{\sum_{i=1}^{N}(M_i - M_{avg})^2}{N-1}}$. A replicate spot for which either channel was flagged as poor quality was excluded from these calculations. Spots for which there were less than two replicates of good quality were discarded from analysis.

The pooled logged ratio standard deviation, σ'_M, was calculated by sorting all the spots by the average logged intensity $I_{avg} = \frac{X_{avg}+Y_{avg}}{2}$ or the minimum logged intensity I_{min} across both channels of all replicates and then taking the square root of the moving average of the variance σ_M^2 with a window of 501 spots. We averaged the variance instead of the standard deviation, since averaging the standard deviation directly will produce a negatively biased (~13%) estimate for N=3 [87]. The Z-statistic was then calculated as $\frac{M_{avg}}{\sqrt{\sigma'^2_M / N}}$. Note that I_{avg} and M as defined above are equivalent to the symbols $\bar{A}$ and M, respectively, as used in other studies [70]. The common "M-A plot" would be called an "M-I plot" using the notation of this study.

3.2.5. Pooled Standard Deviations—Reference Sample Design. Three pairs of arrays were processed for each reference sample experiment. For the unpaired analysis, the arrays within a given condition were linearly normalized to each other, in order to map intensities from different replicates onto similar scales without altering the absolute ratio values (as described above). For each condition, the mean M_{avg} and sample standard deviation σ_M of the logged ratio were calculated for each feature. The pooled standard deviation of the logged ratio, σ'_M, was calculated by sorting all the spots by the average intensity, I_{avg}, or the minimum intensity, I_{min}, across both channels of all replicates for the condition and then taking the square root of the moving average of the variance σ_M^2, with a window of 501 spots, centered on the given spot. The Z-statistic was calculated as $\frac{M_{X_{avg}} - M_{Y_{avg}}}{\sqrt{\sigma'^2_{M_X}/N_X + \sigma'^2_{M_Y}/N_Y}}$ where N_X and N_Y are the number of replicates for the given spot for condition X and condition Y, respectively.

For the paired reference sample analysis, the intensity vectors were all linearly normalized to the vector for the first replicate array of condition X to put all intensity values from both conditions on the same scale without changing the value of the ratios. Then the paired

difference of logged ratios $\mu = M_X - M_Y$ for each pair of replicates was computed. The mean and sample standard deviation of μ was then calculated across replicates. The pooled standard deviation of μ, σ'_μ, was calculated by sorting all the spots by the average intensity I_{avg} or the minimum intensity I_{min} across both channels of all replicates for both conditions, and then taking the square root of the moving average of the variance σ_μ^2, with a window of 501 spots. The Z-statistic was calculated as $\frac{\mu_{avg}}{\sqrt{\sigma'^2_\mu / N}}$ where N is the number of paired replicates for the spot.

To compare Z-statistic values between the paired and unpaired methods, the linear regression slope coefficient with intercept set to 0 was calculated between corresponding Z-statistics from the two methods.

3.2.6. Calculation of p-values. For a Z-statistic Z, the two-tailed p-value is given by $1-2\Phi(|Z|)$, where Φ is the cumulative distribution function for the zero-mean, unit-variance Gaussian. The p-value is corrected for multiple tests using Sidak's formula, $p' = 1-(1-p)^L$, where L is the total number of spots being examined. Note that we did not find it necessary to use more sophisticated means of controlling the error rate [70, 88], as we are primarily concerned with ranking regulated genes and not in establishing firm statistical cutoffs.

3.2.7. Calculation of standard t-statistics and penalized t-statistics. Standard t-statistics for direct comparison arrays were calculated with the formula $t = \frac{M_{avg}}{\sqrt{\sigma_M^2 / N}}$. The two-tailed p-value was calculated using a t distribution with N-1 degrees of freedom. In a penalty-based technique, a constant penalty s_0 is included in the denominator of the t-statistic. The new statistic, d, is given by $\frac{M_{avg}}{\sqrt{(\sigma_M + s_0)^2 / N}}$. Two different methods of choosing s_0 were used: setting s_0 to equal the 90th percentile of the actual standard deviations and the significance analysis of microarrays (SAM) technique, which chooses s such that the coefficient of variation of d is

minimized. The SAM technique was implemented using software developed at Stanford University Labs [72, 82]. This software imputes missing logged ratio values before calculating s_0, and this feature cannot be disabled. The K-nearest-neighbor technique was selected for imputation.

3.2.8. Outlier Detection. When outlier detection was enabled, Z-statistics were calculated using the measured standard deviation instead of the pooled standard deviation for outlier spots. Outliers were determined by calculating σ_ε, the standard deviation of the residual error $\varepsilon = \sigma - \sigma'$ for spots with $\sigma > \sigma'$. Spots for which $\varepsilon > 2\sigma_\varepsilon$ were treated as outliers, similar to [79]. The measured standard deviations for the outlier points were considered to be valid sample measurements of the variance process and were not excluded from the calculation of the pooled standard deviations for spots with similar intensities.

3.2.9. Comparison of Z-statistic and penalty-based statistics. In order to test the reproducibility of different test statistics (cf. Fig. 3.6), two sets of three arrays were randomly selected from the 23 replicate arrays in Dataset 4. For both of these subsets, we calculated the several different test statistics described above. For each gene, the value of each of the test statistics from one 3-array subset was compared to the corresponding value from the other subset, using the squared Pearson's linear correlation coefficient, R^2, and two non-parametric, rank-based correlation coefficients, Spearman Rho and Kendall Tau, which were calculated using JMP (SAS Inc., Cary, NC). This entire process was repeated twice with the remaining arrays in Dataset 4, yielding a total of three independent comparisons. In total, six non-overlapping sets of three arrays—18 arrays in all—were drawn from the original pool of 23 arrays, leaving 5 arrays that were not used in this analysis. As the sets are non-overlapping, each comparison is based on independent data.

In order to evaluate the accuracy of the different test statistics, we compared these statistics to an approximate "gold standard" measure (cf. Fig. 3.5). 3 arrays were randomly selected from the 23 arrays in Dataset 4; the other 20 were used to calculate "gold standard" t-statistics to which the results from the n=3 dataset could be compared. The R^2 value and the linear regression slope coefficient with intercept set to 0 were calculated between the corresponding experimental statistic and "gold standard" t-statistic for each gene. Only spots for which there were at least 15 replicates in the "gold standard" set of arrays were used. This process was repeated on a total of 6 random subsets.

3.3. Results

3.3.1. Average Logged Intensity (I_{avg}) vs. Minimum Logged Intensity (I_{min}) Pooling Metric. We demonstrate our technique of pooling standard deviations using the three arrays in Dataset 1 as a representative example of a "direct comparison" dataset. For each spot, we calculate the average logged ratio M_{avg} and the standard deviation of the logged ratio σ_M, across the three replicates. The spots are then sorted by either average intensity (I_{avg}) or minimum logged intensity (I_{min}) before pooling. Fig. 3.3a and b show the results of pooling standard deviations for Dataset 1, using either the I_{avg} or I_{min} metrics; the measured standard deviation σ_M and the pooled standard deviation σ_M' are plotted together against either I_{avg} or I_{min}. For better comparison, the pooled standard deviation curves for $\sigma_M'(I_{avg})$ and $\sigma_M'(I_{min})$ are both plotted together on Fig. 3.3c against their respective intensity metric, I_{avg} or I_{min}. Fig. 3.3 is based on data produced using the Agilent Feature Extraction software Version A.6.1.1 to quantify spot intensities in the original microarray image. This entire analysis was repeated on Datasets 2 and 3, as well as using two additional image processing techniques: SPOT Processing [86] and a combination of Agilent foreground and SPOT background values (see Methods).

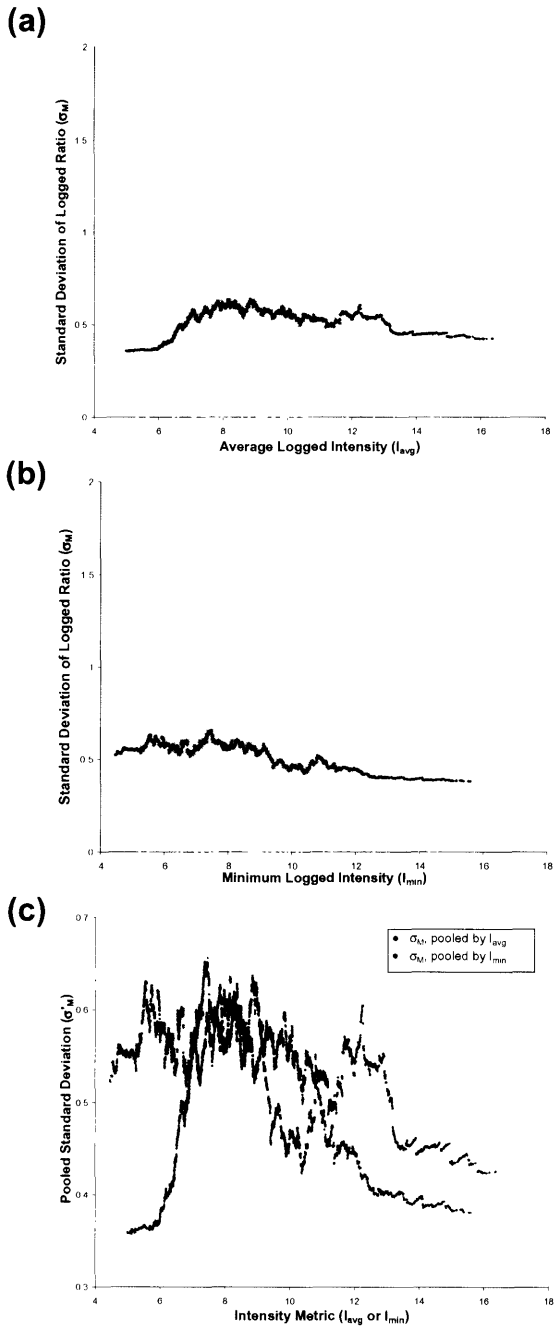

Fig. 3.3. *Two methods of pooling standard deviations of M: sorting by I_{avg} or by I_{min}. The standard deviation (σ_M) is pooled by taking the moving average of the variance (σ_M^2). (a) Measured (σ_M, gray) and pooled ($\sigma_M'(I_{avg})$, black) standard deviation of the logged ratio M, plotted against I_{avg}. For spots with $\sigma_M' > \sigma_M$, the average residual error is 0.28; for spots with $\sigma_M' < \sigma_M$, the average residual error is 0.31. (b) Measured (σ_M, gray) and pooled ($\sigma_M'(I_{min})$, black) standard deviation of M, plotted against I_{min}. For spots with $\sigma_M' > \sigma_M$, the average residual error is 0.28; for spots with $\sigma_X' < \sigma_X$, the average residual error is 0.31. (c) Pooled standard deviation of M (σ_M') plotted against the intensity metric used for pooling, I_{avg} or I_{min}. Data are from Dataset 3.*

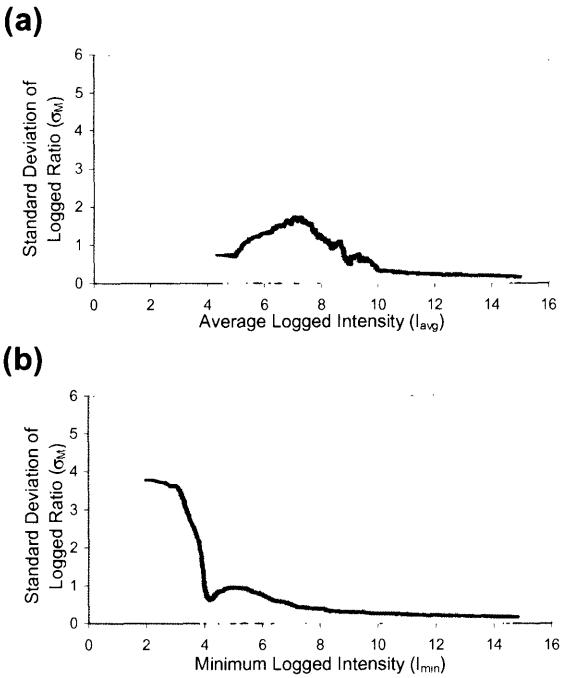

Fig. 3.4. *Comparison of pooled standard deviation curves using I_{avg} or I_{min} pooling metrics. The pooling algorithms are applied to a noisy three-array subset of Dataset 4. (a) Measured (σ_M, gray) and pooled ($\sigma_M'(I_{avg})$, black) standard deviation of M, plotted against I_{avg}. For spots with $\sigma_M' > \sigma_M$, the average residual error is 0.45; for spots with $\sigma_M' < \sigma_M$, the average residual error is 0.49. (b) Measured (σ_M, gray) and pooled ($\sigma_M'(I_{min})$, black) standard deviation of M, plotted against I_{min}. For spots with $\sigma_M' > \sigma_M$, the average residual error is 0.24; for spots with $\sigma_M' < \sigma_M$, the average residual error is 0.23.*

	Agilent Feature Extraction				*SPOT Processing*				*Agilent FG + SPOT BG*			
	$\sigma_M'^2 > \sigma_M^2$		$\sigma_M'^2 < \sigma_M^2$		$\sigma_M'^2 > \sigma_M^2$		$\sigma_M'^2 < \sigma_M^2$		$\sigma_M'^2 > \sigma_M^2$		$\sigma_M'^2 < \sigma_M^2$	
	I_{avg}	I_{min}	I_{avg}	I_{min}	I_{avg}	I_{min}	I_{avg}	I_{min}	I_{avg}	I_{min}	I_{avg}	I_{min}
Dataset 1	0.28	0.28	0.31	0.31	0.19	0.20	0.22	0.25	0.19	0.20	0.22	0.25
Dataset 2	0.24	0.23	0.25	0.25	0.15	0.15	0.16	0.17	0.15	0.15	0.16	0.17
Dataset 3	0.20	0.18	0.21	0.18	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09
Dataset 4 #1	0.15	0.14	0.17	0.15	NA	NA	NA	NA	NA	NA	NA	NA
Dataset 4 #2	0.20	0.17	0.23	0.19	NA	NA	NA	NA	NA	NA	NA	NA
Dataset 4 #3	0.45	0.24	0.49	0.23	NA	NA	NA	NA	NA	NA	NA	NA
Dataset 4 #4	0.21	0.20	0.23	0.21	NA	NA	NA	NA	NA	NA	NA	NA
Dataset 4 #5	0.25	0.19	0.28	0.20	NA	NA	NA	NA	NA	NA	NA	NA
Dataset 4 #6	0.18	0.17	0.21	0.18	NA	NA	NA	NA	NA	NA	NA	NA

Table 3.1. *Mean residual errors for spots with $\sigma_M' > \sigma_M$ and $\sigma_M' < \sigma_M$, using I_{avg} or I_{min} pooling metric. Data is given for three datasets using different image processing techniques (Agilent Feature Extraction, SPOT Image Processing and Agilent foreground combined with SPOT background), and for 6 independent three-array subsets of Dataset 4.*

We evaluated the tightness of the I_{avg}-pooled vs. I_{min}-pooled standard deviation curve fits to the measured standard deviations. Figs. 3.4a and b plot both measured (σ_M) and pooled (σ_M') standard deviations against either the I_{avg} or I_{min} pooling metric, analogous to Fig. 3.3a and b but using an especially noisy three-array subset of Dataset 4 that includes a population of extremely high variance spots. Instead of pooling together spots with similar variance, the I_{avg} metric combines the high-variance spots with the lower-variance spots. In contrast, the I_{min} metric pushes the high-variance spots to the left end of the curve, apart from the less noisy spots. This effect is reflected in the lower mean residual errors between σ_M and σ_M' for the I_{min} metric, calculated for Datasets 1-3 and six independent three-array subsets of Dataset 4 (see Table 3.1). For all of the datasets processed with Agilent Feature Extraction software only, the mean residual errors from using the I_{min} pooling metric are always less than or equal to the corresponding mean residual errors from using the I_{avg} pooling metric. This observation is most striking for Dataset 4 subset #3, which corresponds to the data in Fig. 3.4. The tighter fit that is obtained using the I_{min} metric is also reflected in the improved accuracy of the final Z statistic

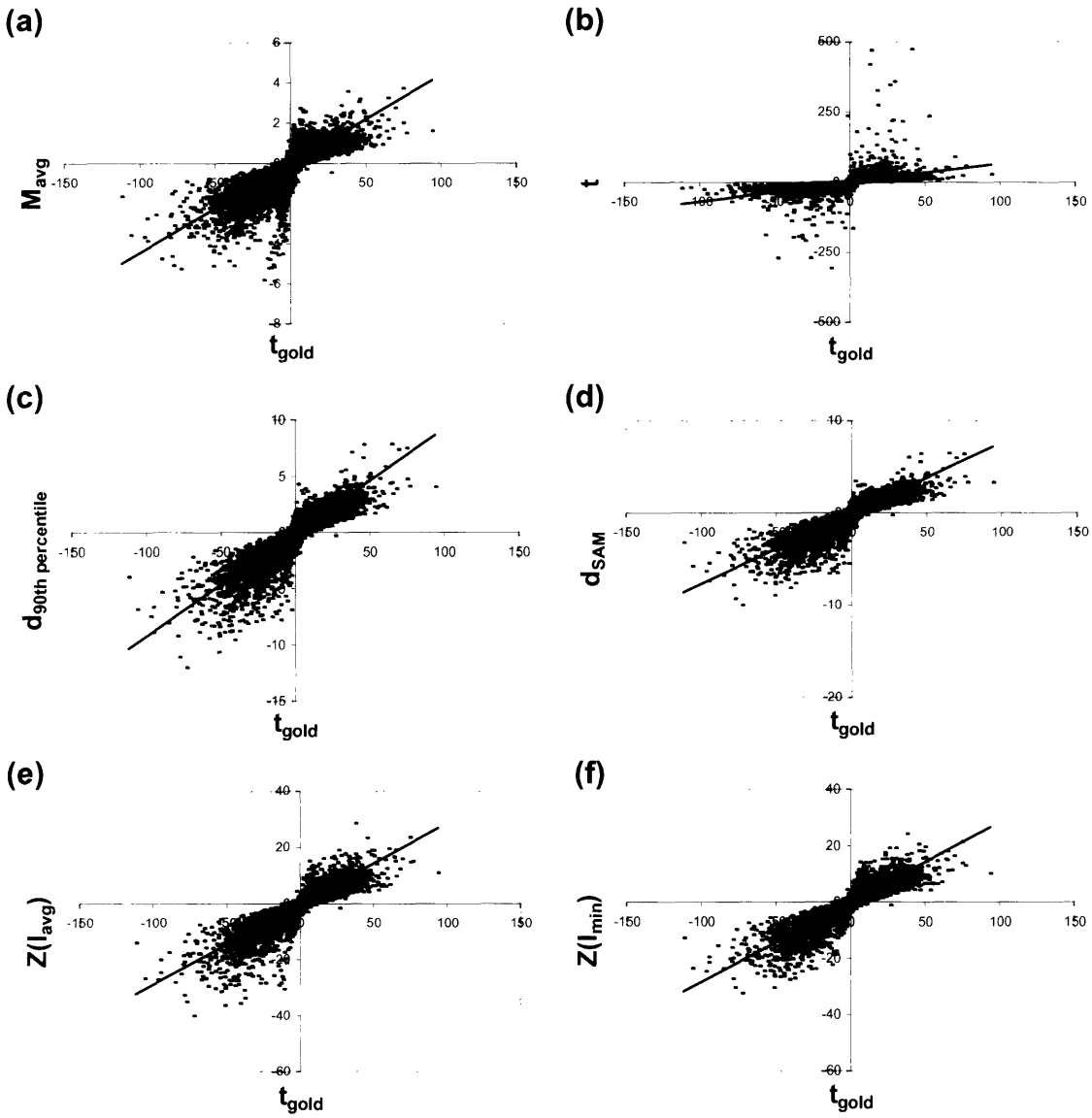

Fig. 3.5. *Comparing the accuracy of different test statistics.* Statistics were calculated for 3 replicate arrays from Dataset 4 and compared to the "gold standard" t-statistic for the remaining 20 arrays. The x-axis for all plots is the "gold standard" t-statistic. The y-axis shows: (a) average logged ratio M_{avg}, (b) standard t-statistic, (c) 90^{th} percentile penalized t-statistic, (d) SAM penalized t-statistic, (e) Z-statistic using the I_{avg} pooling metric, or (f) Z-statistic using the I_{min} pooling metric.

calculated using $\sigma_M'(I_{min})$, which is demonstrated in Fig. 3.5 and discussed below. The trend in residual values is not present when datasets are processed with the SPOT technique or with Agilent foreground and SPOT background.

3.3.2. Comparing the Accuracy of Different Ranking Statistics. In order to test the accuracy of the different test statistics—M_{avg}, the standard t-statistic, the 90th percentile penalized t-statistic, the SAM penalized t-statistic, $Z(I_{avg})$ and $Z(I_{min})$—a subset of three arrays was randomly selected from the total set of 23 replicate arrays in Dataset 4 (see Methods). Each statistic was calculated for each gene in this set. The large number of remaining replicate arrays allowed us to calculate an approximate "gold standard" statistic, t_{gold}, by computing the standard t-statistic over the set of 20 remaining replicates. The value of each test statistic from the three-array subset was compared to the value of the "gold standard" t-statistic, t_{gold}, as shown in Fig. 3.5. The squared Pearson's linear correlation coefficient value (R^2), representing the degree of concordance between the test statistic and t_{gold}, was calculated. This analysis was repeated five additional times, selecting different subsets of experimental and "gold standard" arrays from Dataset 4 each time, and the R^2 values from all six repetitions are given in Table 3.2. The Z-statistics and penalized t-statistics both have appreciably higher R^2 values than either M_{avg} or the standard t-statistic. The R^2 values for $Z(I_{min})$ are greater than the R^2 value for any other technique across all six datasets. Note that there is less scatter for high-magnitude values when using $Z(I_{min})$ instead of $Z(I_{avg})$ (Fig. 3.5e and f respectively). Accordingly, the R^2 value is higher for the $Z(I_{min})$ than the $Z(I_{avg})$ ranking metric for all three datasets, confirming that the tighter curve fits seen in Fig. 3.4a and b and Table 3.1 (see above) translate into improved accuracy of using the I_{min} pooling metric over I_{avg}.

	M_{avg}	t	$d_{90th\ percentile}$	d_{SAM}	$Z(I_{avg})$	$Z(I_{min})$
Dataset 4 #1	0.69	0.09	0.80	0.80	0.79	0.83
Dataset 4 #2	0.66	0.08	0.78	0.78	0.79	0.82
Dataset 4 #3	0.54	0.04	0.79	0.70	0.77	0.84
Dataset 4 #4	0.68	0.02	0.80	0.79	0.80	0.83
Dataset 4 #5	0.64	0.06	0.80	0.74	0.78	0.84
Dataset 4 #6	0.67	0.01	0.80	0.80	0.79	0.83

Table 3.2. *Accuracy of each test statistic when compared to a "gold standard" t-statistic. Each column contains the R^2 value calculated between each experimental test statistic and the "gold standard" t-statistic for (left to right): the average logged ratio M_{avg}, the standard t-statistic, the 90^{th} percentile penalized t-statistic, the SAM penalized t-statistic, the Z-statistic using the I_{avg} pooling metric and the Z-statistic using the I_{min} pooling metric. Data are from six independent three-array subsets of Dataset 4. Although M_{avg} is not a statistical test, it is included in this Table 3.for comparison.*

3.3.3. Comparing the Reproducibility of Different Ranking Statistics. We also evaluated the reproducibility of these different test statistics, by constructing test datasets that split six replicate arrays from Dataset 4 into two subsets of 3 arrays (see Methods). Each test statistic— M_{avg}, the standard t-statistic, the 90th percentile penalized t-statistic, the SAM penalized t-statistic, $Z(I_{avg})$ and $Z(I_{min})$— was calculated for both three-array subsets. A precise, i.e., reproducible, test statistic should produce similar values for both subsets since all of the arrays in both subsets were drawn from a pool of replicates prepared from identical biological experiments. Fig. 3.1a-b and Fig. 3.6a-d show the correlation for each test statistic between the two subsets, including a linear regression line in Fig. 3.6. The slope coefficient of the linear regression indicates whether overall magnitudes of the test statistics are different between the two subsets, while R^2 indicates the degree of correlation on a gene-by-gene basis (Table 3.3). This analysis was repeated for an additional two pairs of independent three-array subsets of Dataset 4 (graphs not shown), with the slope coefficients and R^2 values given in Table 3.3.

The R^2 values for the two Z-statistics were similar to each other and consistently higher than those of the other techniques. Nonparametric measures of correlation, the Spearman Rho

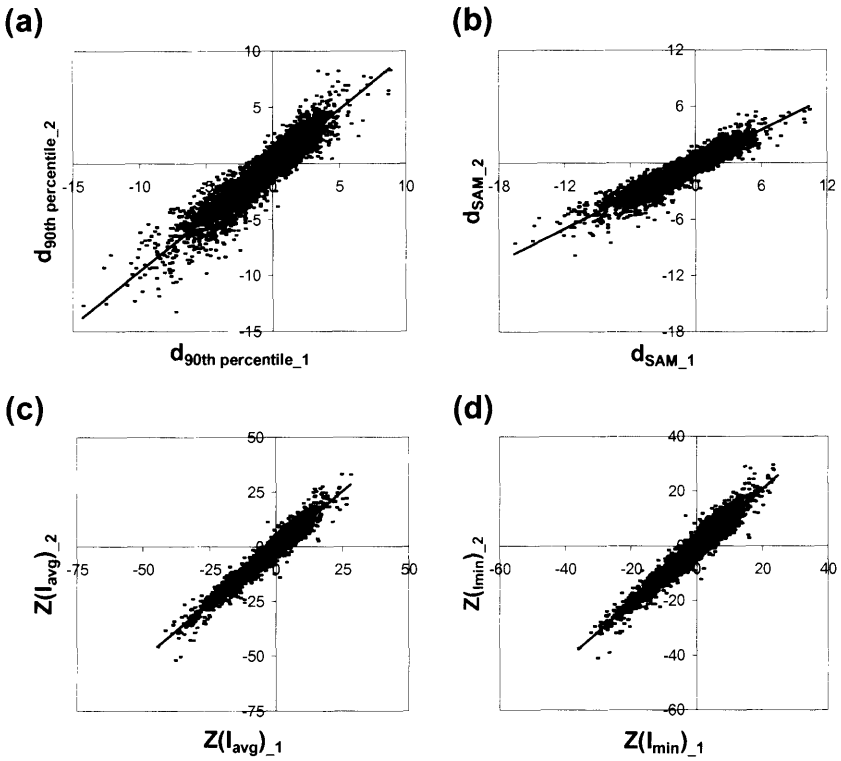

Fig. 3.6. *Comparing the reproducibility of different test statistics. Two subsets of Dataset 4 each contain three replicate arrays derived from identical biological experiments. Each test statistic is calculated twice, once for each subset, and the two statistics are plotted against each other. (a) Comparison of 90th percentile penalized statistics. (b) Comparison of SAM penalized statistics. (c) Comparison of Z-statistics using I_{avg} pooling metric. (d) Comparison of Z-statistics using I_{min} pooling metric. Also see Fig. 1a for comparison of the standard t-test.*

and Kendall Tau rank correlation coefficients, were also higher for both Z-statistics than any of the other statistics for all three pairs of subsets (data not shown). All three calculations of the slope coefficients for both Z-statistics, as well as M_{avg} and the 90th percentile penalized t-statistic, are close to 1, indicating that the overall magnitudes of the Z-statistics are consistent across datasets, whereas the standard t-statistic and the SAM penalized t-statistic produced test statistics whose overall magnitudes vary across the subsets.

		M_{avg}	t	$d_{90th\ percentile}$	d_{SAM}	$Z(I_{avg})$	$Z(I_{min})$
R^2	Dataset 4 #1	0.89	0.00	0.86	0.87	0.93	0.94
	Dataset 4 #2	0.90	0.00	0.87	0.87	0.93	0.93
	Dataset 4 #3	0.89	0.00	0.88	0.88	0.95	0.95
Slope	Dataset 4 #1	0.90	0.00	0.97	1.63	1.00	1.00
	Dataset 4 #2	0.93	0.01	0.96	0.58	1.02	1.04
	Dataset 4 #3	1.00	0.07	0.96	0.74	1.09	1.08

Table 3.3. *Reproducibility of each test statistic when used on replicate datasets. Linear regression slope coefficients and R^2 coefficients are calculated between corresponding statistics from two replicate three-array subsets of Dataset 4. Columns represent (left to right): the average logged ratio M_{avg}, the t-statistic, the 90th percentile penalized statistic, the SAM penalized statistic, the Z-statistic using I_{avg} pooling metric and the Z-statistic using I_{min} for three different pairs of subsets. Although M_{avg} is not a statistical test, it is included in this Table 3.for comparison.*

3.3.4. Outlier Detection. When calculating the Z-statistic, using a much smaller pooled σ_M' in place of a large σ_M has the potential to overestimate the significance of gene regulation in the case where one of the replicates is an outlier measurement and the large measured standard deviation provides a better estimate of the variability. As seen in Fig. 3.3a-c, there are several spots that lie far above the pooled standard deviation curve. Datasets 1 and 4 were reprocessed using an outlier detection technique (see Methods). Fig. 3.7a shows σ_M and σ_M' from Dataset 1 plotted against I_{min}, as in Fig. 3.3c, except that the y-axis has been rescaled to show all spots detected as outliers, which are now highlighted in black.

The accuracy of this outlier detection technique was also evaluated by comparing the Z-statistic to t_{gold} using Dataset 4. Fig. 3.7c plots $Z(I_{min})$ vs. t_{gold} for Dataset 4 set #2. The outliers,

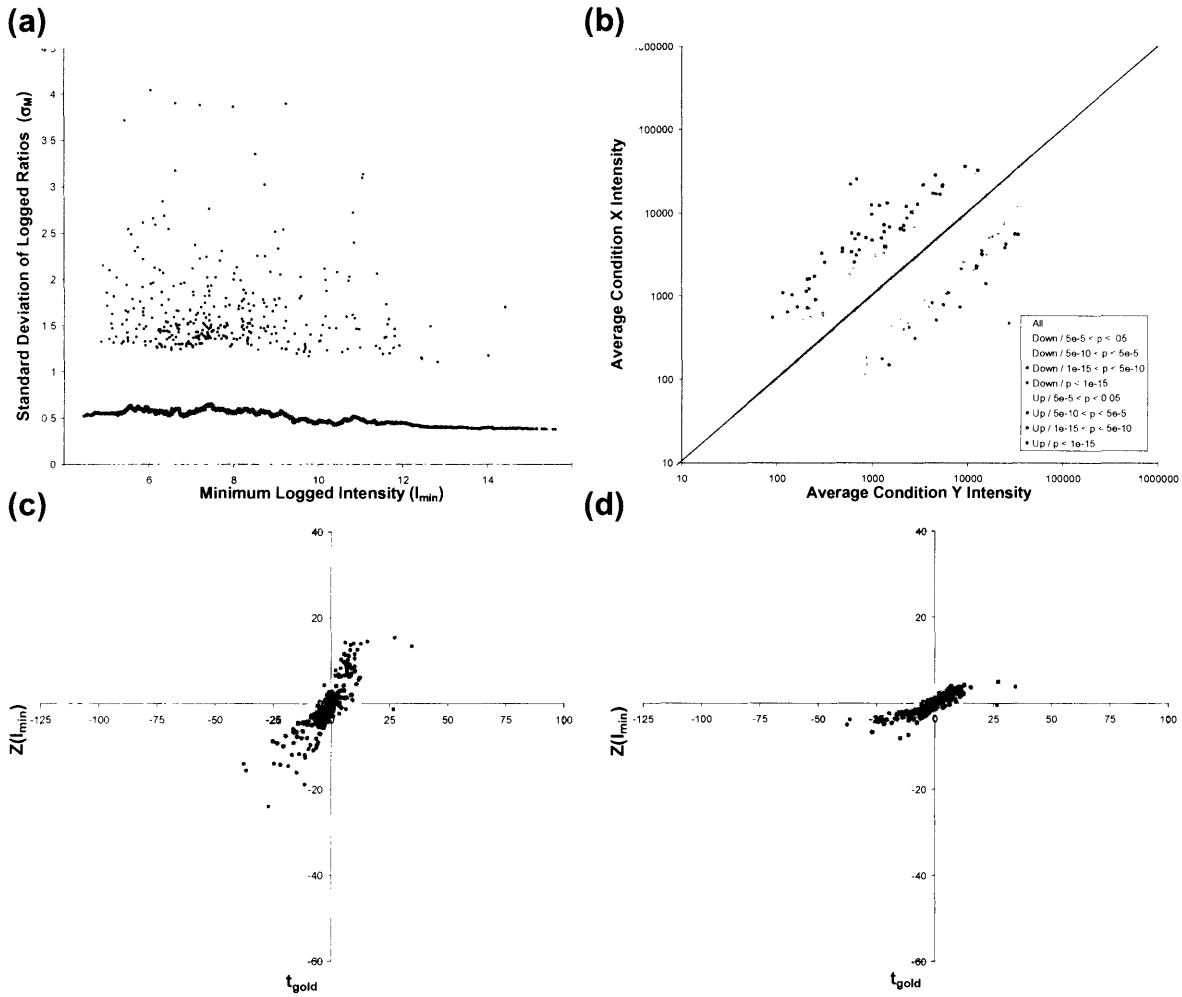

Fig. 3.7. *Implementation of outlier (high experimental standard deviation) detection. (a) Measured (σ_M, gray) and pooled ($\sigma_M'(I_{min})$, black curve) standard deviation of M, plotted against I_{min}, with the outlier spots highlighted (black points), for Dataset 3. (Compare to Fig. 3.3b.) (b) Scatterplot of average condition X intensity vs. average condition Y intensity for Dataset 3, with p-values indicated in color. (c) Z-statistic using I_{min} pooling metric vs. "gold standard" t-statistic with outliers highlighted in black, for a 3-array subset and 20-array "gold-standard" subset of Dataset 4. Outlier Z-statistics calculated using the pooled standard deviation. (d) Z-statistic using I_{min} pooling metric vs. "gold standard" t-statistic, with outliers highlighted in black, for the same data in (c). Outlier Z-statistics calculated using the measured standard deviation, for Dataset 4 subset #2.*

which are highlighted, include false positive spots for which t_{gold} is low and $Z(I_{min})$ is high, although not all such points are detected as outliers. Fig. 3.7d is an identical plot to Fig. 3.7c except that the Z-statistics for the outlier spots are calculated using the higher-valued measured standard deviation σ_M instead of the pooled value σ_M'. The outliers are now mostly clustered around the origin with the other non-significant spots. A few spots with moderately high t_{gold} values are detected as outliers and have low corrected Z-statistics, and some potential false positives with high Z-statistic values and low t_{gold} values are not detected as outliers.

At the end of the analysis, the outlier-corrected Z statistics are converted to p-values. To demonstrate the additional information that the p-values provide, Fig. 3.7b shows a scatterplot of X vs. Y for Dataset 1, with statistically significant spots colored according to their multiple-test-corrected p-values (see Methods). Spots with similar ratios may have different p-values due to their different standard deviations. In addition, after outlier detection, some spots with high ratios are not found to be significant.

3.3.5. Analysis of Reference Sample Arrays. The techniques used above for a direct comparison experimental design were extended to a reference sample design (see Methods). Under a reference sample design, one can estimate either the standard deviation of the individual logged ratios comparing experimental samples to the reference sample, M_x and M_y, or the standard deviation of the paired differences of these logged ratios, $\mu = M_x - M_y$. Under the first, or unpaired method, the Z-statistic is calculated as $\frac{M_{X_{avg}} - M_{Y_{avg}}}{\sqrt{\sigma'^{2}_{M_X}/N_X + \sigma'^{2}_{M_Y}/N_Y}}$ where N_X and N_Y are the number of replicates for the given spot for condition X and condition Y, respectively. Under the second, or paired method, the Z-statistic is calculated as $\frac{\mu_{avg}}{\sqrt{\sigma'^{2}_{\mu}/N}}$ where N is the number of paired

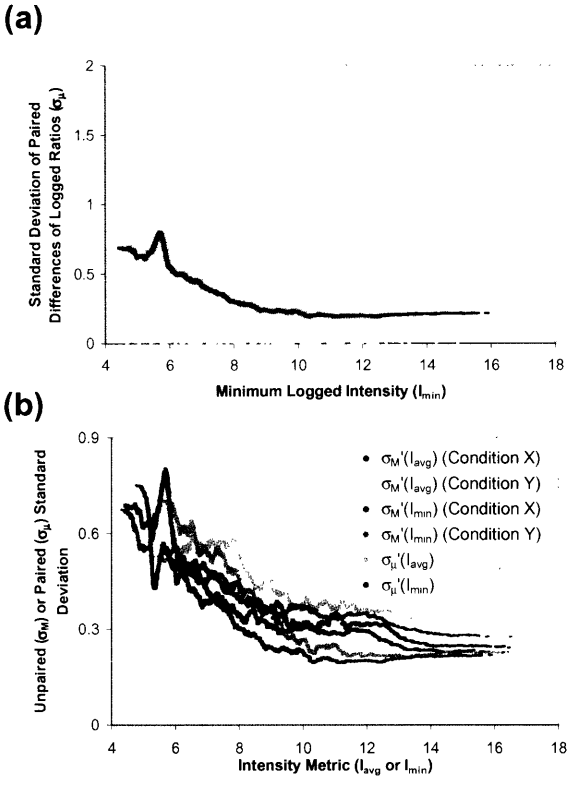

Fig. 3.8. *Methods of pooling the standard deviation for a reference sample design. The standard deviation (σ) is pooled by taking the moving average of the variance(σ^2). (a) Measured (σ_m, gray) and pooled (σ_m' (I_{min}), black) standard deviation of the difference of logged ratios μ, plotted against I_{min}. (b) Pooled standard deviation of M_X, M_Y and μ plotted against the intensity metric used for pooling, I_{avg} or I_{min}. Data are from Dataset 5.*

replicates for the spot. The samples used in a reference sample design may not always have been collected or processed in pairs, so we evaluated both of these methods.

For each replicate in reference sample Dataset 5, the biological specimens for conditions X and Y were prepared on the same day, so a natural pairing exists for the condition X and Y arrays. These data were processed using all three image processing techniques and then analyzed using both paired and unpaired methods, and using either the I_{avg} or I_{min} pooling metric for each approach. Fig. 3.8a shows the measured and pooled standard deviation of the paired differences of logged ratios (σ_μ and σ_μ') plotted together against the pooling metric, I_{min}. Curve fits were analogously constructed using the I_{avg} pooling metric with the paired method (with results similar to using the I_{min} metric, data not shown), and using both I_{avg} and I_{min} with the unpaired method (with results similar to using the ratio method with direct comparison arrays, data not shown). The unpaired σ_M' and paired σ_μ' curves are plotted together against their the I_{avg} or I_{min} pooling metric in Fig. 3.8b. The paired standard deviations are lower than the unpaired standard deviations except at low intensity metric values.

Linear regression was performed between Z-statistics calculated using the paired and unpaired methods for all spots. Table 3.4 gives the linear regression slope coefficients when either the I_{avg} or I_{min} pooling metric was used, for Dataset 5 processed with the three different image processing techniques. For most spots, both the difference of logged ratios (μ) and number of replicates (N) are the same, except for the occasional difference between the two conditions in the number of low quality spots that are excluded from the analysis. Thus, differences in the Z-statistic primarily reflect differences in the standard deviations. The slope coefficients are all greater than 1, indicating that the paired technique produced higher Z-statistic values, due to the lower standard deviations that are produced with paired analysis.

	Agilent Feature Extraction	SPOT	Agilent FG + SPOT BG
I_{avg}	1.70	1.68	1.70
I_{min}	1.59	1.61	1.63

Table 3.4. *Linear regression slope coefficients calculated between the corresponding Z-statistics using independent or pairwise analysis. Coefficients given for reference sample design Dataset 5. Values greater than 1 indicate higher Z-statistics with the pairwise technique. Data is shown for both pooling metrics I_{avg} and I_{min} and for three different image processing techniques. Every linear regression analysis produced an R^2 value greater than 0.89 (data not shown).*

The mean residual errors for spots with $\sigma' < \sigma$ and $\sigma' > \sigma$ were calculated when using the I_{avg} or I_{min} pooling metric in unpaired or paired analyses of Dataset 5, and are given in Table 3.5. For the unpaired analysis, as in the direct comparison experiments, mean residual values produced by using the I_{min} pooling metric are less than or equal to those produced by the I_{avg} pooling metric. The same trend is seen between the two pooling metrics for the paired analysis. These results are consistent regardless of the image processing technique used.

	Agilent Feature Extraction				*SPOT Processing*				*Agilent FG + SPOT BG*			
	$\sigma^{2'} > \sigma^2$		$\sigma^{2'} < \sigma^2$		$\sigma^{2'} > \sigma^2$		$\sigma^{2'} < \sigma^2$		$\sigma^{2'} > \sigma^2$		$\sigma^{2'} < \sigma^2$	
	I_{avg}	I_{min}	I_{avg}	I_{min}	I_{avg}	I_{min}	I_{avg}	I_{min}	I_{avg}	I_{min}	I_{avg}	I_{min}
σ_M^2, Cond. X	0.19	0.18	0.21	0.20	0.14	0.14	0.15	0.15	0.13	0.13	0.14	0.14
σ_M^2, Cond. Y	0.22	0.18	0.25	0.23	0.16	0.14	0.17	0.17	0.15	0.13	0.16	0.16
σ_μ^2	0.19	0.18	0.22	0.19	0.13	0.13	0.14	0.14	0.12	0.12	0.12	0.12

Table 3.5. *Mean residual errors for spots with $\sigma^{2'} > \sigma^2$ and $\sigma^{2'} < \sigma^2$, using I_{avg} or I_{min} pooling metric. Analysis was performed on Dataset 5 (reference sample design). Data is given for both unpaired and paired analyses, using three different image processing techniques: Agilent Feature Extraction, SPOT Image Processing, and Agilent foreground combined with SPOT background.*

3.4. Evaluation of Algorithm Performance

Building up new knowledge about biological systems is the ultimate purpose of microarray experiments, but all such insights have to be built on a solid analytical foundation to be accurate and useful. Proper normalization of data and accurate detection of which genes are

regulated are vital to the success of downstream exploration of microarray data. Even for exploratory cluster analyses, the genes that are significantly regulated must be selected beforehand. This task of detecting these genes is a difficult statistical problem; a statistical hypothesis is made for each of tens of thousands of genes tested, but only a small number of replicate arrays are available to test those hypotheses. The statistical methods presented in this study attempt to draw as much information as possible out of a small number of array replicates to determine which genes are likely to be regulated.

It is clear that looking at the measurements of each gene in isolation can produce a test with low statistical power (e.g. using the standard t-test, Fig. 3.1). To improve statistical power, we can use knowledge about the relationships among the many thousands of points in the arrays. Specifically, we group together spots that have similar standard deviations and then pool together many less accurate estimates of standard deviation into a single, more accurate estimate. Our data also show that the Z-statistics are more precise than either standard or penalized t-statistics for detecting differential gene expression in microarray data. We further demonstrate that pooling standard deviations using the minimum intensity metric produces Z-statistics that are more accurate than the standard t-test, the penalized t-tests, and the average intensity-based Z-statistic.

3.4.1. Average Combined Logged Intensity (I_{avg}) vs. Minimum Logged Intensity (I_{min}) Pooling Metric. We evaluated two different intensity-based metrics for pooling standard deviations. There are many reports that the variance is a function of intensity, but the exact shape of this relationship could depend on many factors extrinsic to the biological experiment, such as the array technology being used, the signal-to-noise ratio of the data, the similarity between the two conditions[83], the normalization technique or the background subtraction

technique. For this reason, we favor an estimation of the standard deviation using a curve-fitting technique rather than a fixed model based on previous data. Furthermore, when dealing with two-channel arrays, there are two different intensity values associated with each replicated spot. It is possible that the variation is best described as a function of the average intensities of both channels. However, our own experience and many other reports [74, 78, 89] suggest that the highest variances are often seen for low intensity spots. If so, the variance may be better described as a function of the minimum intensity over all the spots.

The data presented here show that the mean residual errors are either equal or lower when using the I_{min} compared to the I_{avg} pooling metric, for every dataset using the Agilent Feature Extraction image processing technique. The subset of Dataset 4 for which this difference is most striking, #3 in Table 3.1, also has a population of spots with particularly high variance (see Fig. 3.4). The I_{avg} metric pools these spots together with other spots that have a much lower variance. In contrast, the I_{min} metric moves these spots to the low end of the x-axis, and the curve fit tracks the standard deviation of the spots much better. The noisiest spots on microarrays are often those where at least one channel is "blank", i.e. a noisy, low level of signal that presumably represents no expression. The I_{min} metric is better at grouping such spots together. For datasets with low background levels, there is a smaller difference in the performance of the two pooling metrics.

The trends in the mean residual errors from the unpaired reference sample analysis agree with the results from the direct comparison analyses. This similarity is to be expected, since processing each reference sample condition separately is equivalent to doing a direct comparison between each condition and reference RNA samples. Both pooling metrics generate similar mean residual error values when pooling σ_μ, but one dataset is not enough to make any

generalizations about which pooling metric will perform best for all paired reference sample datasets. The improved performance of the I_{min} pooling metric is lost when using SPOT processing or combined Agilent foreground and SPOT background image processing, suggesting that these image processing techniques may be more effective at removing noise at low intensities.

The I_{avg} and I_{min} pooling techniques are reproducible to the same degree, since their R^2 coefficients between Z-statistics from paired datasets (see Table 3.1) are similar to each other. The I_{min} pooling technique generates slightly more accurate results, as indicated by the greater R^2 coefficients between $Z(I_{min})$ and t_{gold} compared to those between $Z(I_{avg})$ and t_{gold} (see Table 3.2). This trend holds for all six subsets of Dataset 4.

3.4.2. The Higher Accuracy of $Z(I_{min})$. The Z-statistic calculated using the I_{min} pooling metric provides an improvement in accuracy over the other techniques. The t-statistic derived from datasets with 20 replicates was used as a surrogate "gold standard" since 8 or more replicates can be considered sufficient to give power to the t-statistic [70]. The t-statistic was chosen as the "gold standard" instead of the average logged ratio since the latter does not take variability into account. For each of the six permuted subsets of Datasets 4, the 90^{th} percentile penalized t-statistic, SAM penalized t-statistic, and $Z(I_{avg})$ had similar R^2 values when correlated with the "gold standard" t-statistic, although the SAM statistic did perform poorly for the noisiest subset of Dataset 4 (#3 in Table 3.2) with an R^2 value of only 0.70. $Z(I_{min})$, however, consistently produced the highest R^2 value for each of the six datasets. Since the ratios used in each of these statistics is identical, this result indicates that the standard error generated with the I_{min} technique produces the best correlation with the gold standard t-statistic based on 20 replicates. Although excluding spots with very low intensity could eliminate the difference in

performance between the I_{min} and I_{avg} pooling metrics, this approach would make it impossible to detect low-expressed regulated genes, which may be biologically significant.

The Z-statistics from the I_{min} technique do not correlate perfectly with the "gold standard" t-statistic, however. Some disagreement can be expected because the $Z(I_{min})$ data was based on only three replicate arrays, which contain much less information than the 20 replicates used to calculate the "gold standard" t-statistic. Also the significance estimates calculated using the "gold standard" t-statistic may still contain some inaccuracies, even with 20 replicates. Kerr et al. found this to be true with 12 replicates, where accuracy is reduced if the error distribution for each gene is modeled separately instead of using a pooled estimate [68]. Analyzing the large (N=20) replicate dataset using robust estimators of ratio and standard deviation may be able to create a more accurate "gold standard" to use for further testing of the Z-statistic or other statistics. Note that we do not employ an explicit permutation-based approach to estimate the false detection rates of the statistics investigated in this study, as in Ref. [69]. Rather than permute gene labels from a small set of arrays to estimate the distribution of expected test statistics, with the availability of the large (N=23) replicate dataset described herein, we preferred to use this rich source of actual test statistics directly.

3.4.3. The Higher Reproducibility of Z-statistics. The Z-statistic—calculated with either pooling the I_{min} or I_{avg} pooling metric—provides an appreciable improvement in reproducibility over the average logged ratio alone, the standard t-test and the 90^{th} percentile and SAM penalized t-statistics. Both linear (R^2) and non-parametric rank correlation coefficients were highest for the Z-statistic when comparing corresponding spots between three independent pairs of replicate datasets. Also, the standard t-statistic and SAM penalized t-statistic generate linear regression slope coefficients that vary greatly from pair to pair, indicating that their absolute

magnitude is not as reproducible as the Z-statistics, whose linear regression slope coefficients are much closer to 1.

The high correlation values and near-unity slope coefficients for the Z-statistic support the hypothesis that pooling the standard deviations of spots with similar intensities provides a stable, precise estimate of the standard deviation. This assumption of a well-estimated standard deviation supports the use of the Gaussian distribution to map the Z-statistic to a p-value. Using only the measured standard deviation, one is forced to use a t-distribution with only 2 degrees of freedom to generate a p-value. This test does not have sufficient power to generate any significantly regulated points; because of the very small number of degrees of freedom, not a single spot seen in Fig. 3.1a is found to be significant after multiple test correction. In contrast, even after a conservative multiple test correction that makes the cutoff for statistical significance much more stringent, many spots are found significant using the Z-statistic. The penalized t-statistics do not produce a stable estimate of the standard deviation with these data, perhaps because the constant added to the denominator of the test statistic showed a large variation between replicate datasets. Therefore they cannot be mapped to a p-value in a reproducible manner.

3.4.4. Outlier Detection. One limitation of using a pooled standard deviation is that for a spot with replicate ratios that include one or more outliers, the appropriately high measured standard deviation will be replaced by an inappropriately low pooled standard deviation. This substitution could produce a false positive result. We have sought to minimize this limitation by implementing an overlying outlier detection algorithm. (For other implementations of outlier detection, see Ref. [79, 83].) The algorithm in this study uses the measured standard deviation instead of the pooled standard deviation for spots for which the pooling model may not hold.

These spots are identified as ones for which residual error $\sigma-\sigma'$ is positive and greater than twice the standard deviation of the positive residual errors.

The measured standard deviations for these outlier points are valid sample measurements of the variance process and should be used to calculate the pooled standard deviations for spots with similar intensities. These ratio measurements, however, are too widely varying for one to have the same confidence in the average ratio as one would have for other spots; thus, it is appropriate to substitute the measured standard deviation for the pooled standard deviation in these cases. Fig. 3.7c-d, which highlight outlier spots on a plot of the $Z(I_{min})$ vs. the "gold standard" t-statistic for Dataset 4b, show that this outlier detection technique correctly detects many of the presumably false positive spots that have a high Z-statistic and low t_{gold} value. The plots also show some false positive spots that are not detected through this algorithm, as well as a few spots that become false negatives after outlier detection. Other, more complex outlier detection algorithms may perform better, and should be explored. A simple modification to the current algorithm, using local instead of global estimates of the standard deviation of the residual error, may improve outlier detection. Alternative implementations include modifying the pooling window shape to give more weight to a spot's measured standard deviation or that of its nearest neighbors by intensity. Strictly speaking, the p-values for outlier spots should be calculated using a t-distribution instead of a Gaussian distribution since the measured standard deviation is being used. We have shown, however, that with 3 replicates, no spots in our datasets can be found statistically significant using the t-test and strict multiple test correction. In order to preserve detection of spots, we continue to use the Gaussian distribution to convert outlier Z-statistics to p-values, which may slightly increase the false positive rate for spots detected as outliers. In practice, however, such spots are rarely found to be significantly regulated.

3.4.5. Unpaired vs. Paired Analysis for Reference Sample Experiments. Finally, we have extended our algorithms to apply to data from a reference sample experimental design. This design gives one the flexibility to compare many different conditions to one another, but the trade-off is a loss in precision. In theory, using a reference sample design instead of a direct comparison design should increase the variance by a factor of 2. This increase has in fact been observed in practice [90].

The paired analysis method can reduce the measured variation in a reference sample design. The linear regression slope coefficients in Table 3.1 indicate that the Z-statistic values using the paired analysis are higher than the unpaired Z-statistic values. Thus, the paired difference of logged ratios, μ, is less variable than the independent logged ratios, M_X and M_Y. This observation suggests that the effects of biological or analytical variation from replicate to replicate can be reduced if comparisons are made between paired samples. Whether this reduction is due to using paired biological samples or paired array processing dates [91] is still an open question, and probably will be context-dependent. Although it may not always be practical, it would be beneficial for investigators to design reference sample experiments to be performed in parallel whenever possible to take advantage of the lower standard deviations produced by paired analysis.

3.4.6. Finding the optimal statistical test. Several areas remain for further refinement of our implementation of pooling-based statistical analysis of microarray data. Currently, the standard deviation is pooled using a simple moving rectangular window of 501 spots, but other window sizes and shapes may improve performance slightly. More generally, we have not explicitly compared the moving average estimator with the spline-fit or loess-based techniques to

estimate the standard deviation used in other studies (see Background). While we expect performance to be similar, further testing may reveal an advantage.

Following Ref. [92], we do not try to estimate the dye-specific bias of individual spots or genes (i.e., dye- gene interaction) in order to preserve degrees of freedom needed to estimate the variance. Informally we noted that dye bias in some spots produced high measured variances that caused those spots to be considered non-significant outliers. A post-hoc test to warn of potential dye bias of individual spots may be appropriate for small numbers of array replicates (e.g. N=3), especially if the experimental design is unbalanced (i.e., the number of dye-swapped and unswapped arrays is not equal).

Note that this study only considered statistics of the general form (ratio) / (standard deviation). ANOVA models that consider the variance as intensity-dependent, as seen in Ref. [68, 78], can be seen as an extension of this concept. An ANOVA framework, however, also allows for a more complicated experimental model that can incorporate normalization and multiple biological conditions. Pooling standard deviations as a function of minimum intensity instead of average intensity may benefit such models. Permutation tests can also be used to detect regulated genes, and are known to be robust to outliers but can have low power for small N. Xu et al. found a permutation test to be equally or less accurate than parametric methods in ranking genes [93]. Bayesian analysis can also be applied to microarray data [66, 73, 74], and may be useful in this context to draw more information out of the distribution of intensities and ratios in the data.

In this study, data is first normalized, and then detection of regulated genes is performed in a separate step. In contrast, other approaches incorporate normalization and statistical inference into a unified model [82, 92]. Furthermore, the options for normalizing the data are

numerous, including algorithms based on local regression (loess) [60], splines [94], a constant shift [68], or more exotic transforms that tend to remove the intensity dependence of the variance [95]. Increased attention to the low-level details of scanning and image processing may also improve accuracy [75, 90, 96], while at the same time potentially changing the intensity dependence of the variance. It remains to be seen how the techniques used for normalization or variance-stabilizing transforms will impact the accuracy and precision of regulated gene detection. In addition, we are concerned that some of these transforms may create a systematic bias for or against genes of low intensity (e.g., [97]).

3.4.7. Test performance can depend on data characteristics. Although many datasets have a variance that is intensity-dependent [65, 68, 74-79], some studies have analyzed datasets whose variance characteristics are not strongly intensity-dependent (e.g., [92]). In general, we have experienced that microarray datasets with a low background relative to signal, loess-based normalization, and conservative background subtraction (e.g. SPOT Image Processing) produce standard deviations that are not strongly intensity-dependent. In this context, the differences between the I_{min} and I_{avg} metrics disappear. In fact, for data with unusually low noise, the standard deviations is nearly constant across all spots and all of the statistical tests considered in this paper, even simply the average logged ratio, tend to converge. This observation is not unexpected; as the standard deviations converge to the same value, the denominator of the test statistics will become constant, leaving the test statistics simply proportional to the ratio. We would recommend finding a normalization [60, 82, 90, 94] and background subtraction technique [75, 86, 96] that produces low, intensity-independent standard deviations. Applying variance stabilizing transforms may eliminate the intensity dependence of the standard deviation [95], but might also reduce statistical power or bias the test toward spots of certain intensities. It

cannot be predicted in advance whether all intensity dependence of the variation will be removed, so we continue to use the more robust statistic $Z(I_{min})$ for all of our datasets. Furthermore, in situations where changing the background subtraction or normalization technique is not possible because the original data is not available, using a more robust statistic like $Z(I_{min})$ will be advantageous.

While the pooling techniques described herein can compensate for intensity-dependent variation, this intensity dependence can be minimized or exaggerated by different normalization techniques and background subtraction techniques. These techniques may have subtle effects on the power to detect regulated genes at different intensities, perhaps creating bias for or against detection of low-expressed genes. For this reason, until the most sensitive and unbiased normalization and background subtraction methods are optimized for each microarray system, we would encourage creators of microarray data archives to preserve unnormalized intensity and background data, and the original image data when possible.

Of the many useful tests used to detect regulated genes from a small number of microarray replicates, we see the intensity-based variance estimation and Z-statistic described here to be a good combination of simplicity, robustness, precision, and accuracy. This technique allows meaningful p-values to be added to a list of regulated genes. With this assessment of statistical significance, an investigator can proceed to focus on genes that are most likely to be regulated. Implementations of the Z-test algorithms are available at http://vessels.bwh.harvard.edu/software/papers/bmcg2004.

4. The Transcriptional Response of Cultured HUVEC to IL-1β

4.1. The Response of HUVEC to an Inflammatory Stimulus

The changes caused by an external stimulus to the genome-wide basal expression profile of cultured HUVEC (cf. Chapter 2) can be studied using the statistical algorithms detailed in the previous chapter. The stimulus we have chosen to study here is interleukin-1 beta (IL-1β), a potent inflammatory stimulus for endothelial cells [98]. Responding to inflammatory stimuli is one of the most critical roles of the endothelium, as it is the capillary and venule barrier that regulates extravasation of leukocytes into tissues. By overseeing leukocyte tethering and migration, and amplifying or abrogating signaling cascades, endothelial cells modulate the effect and severity of an inflammatory response. In addition, a number of other roles of endothelium—e.g., regulation of haemostasis and control of permeability —are modulated during inflammation. Thus, a genome-wide endothelial snapshot of endothelial cells exposed to an inflammatory stimulus will begin to provide insight into the complex regulatory networks that comprise the genetic control of several endothelial cell functions. Understanding the nature of this response also provides insights into targets for clinical manipulations in the context of sepsis, atherosclerosis, autoimmune disorders and other inflammatory pathologies.

4.2. Methods

4.2.1. Experimental conditions. We chose to stimulate cells with IL-1β (10 U/mL) in our *in vitro* model of inflammation as it has been well-characterized in our laboratory as a potent inflammatory stimulus [98]. HUVEC isolated from normal term cords and pooled from 5 to 7 donors were cultured in complete media supplemented with 20% fetal calf serum, 2mM L-glutamine, 50 mg/ml endothelial cell growth supplement, 100 mg/ml heparin and 100 unit/ml penicillin-G1100 mg/ml streptomycin, and incubated at 37°C in 5% CO2 in humidified air.

Cells from the first subculture were plated at an initial density of 70,000 cells/cm^2 and grown for 24 hours. At this time point, the cells were exposed for 4 hours either to IL-1 at a concentration of 10 U/mL ("IL-1"), or to media ("control"). These paired experiments were repeated with three independent batches of HUVEC. RNA from these cells was collected and microarrays for each sample were prepared as described in Chapter 2.

4.2.2. Image Quantification and Statistical Processing. The images produced by the AB1700 scanner were quantified by the scanner software, which subtractively normalizes each spot's chemoluminescent signal against a corresponding fluorescent control signal, provides surrogate values for negative or especially faint spots and normalizes all the values in an attempt to map intensities from each array to a uniform dynamic range. In contrast with the expression profile analysis, the normalized and *surrogated* signal values were used to enable the calculation of meaningful, non-negative ratios (although these ratios may be underestimates for surrogated values). We observed that when comparing any two microarrays, there was usually a bias in signal values towards one array over the other at high intensities. Thus, before comparing any two arrays to each other, we applied an intensity-based Lowess correction to normalize the array values [60]. Spots with flag values over 10,000 were excluded from analysis, and all remaining genes, whether classified as expressed or not expressed in cultured HUVEC (cf. Chapter 2), were retained. Genes differentially expressed between the two conditions, IL-1 and control, were statistically identified by applying the minimum-intensity-based variance estimation technique with outlier detection described in Chapter 3 to the three Lowess-normalized pair of replicates.

4.2.3. QRT-PCR Validation of Select Genes. Ninety-five genes known to be involved in inflammatory processes were validated using real-time quantitative Taqman PCR with the same three pairs of RNA samples used to generate the microarray data. Briefly, the purified, DNase-

treated RNA (1.5 μg) was reversed-transcribed by using a MultiScribe based reverse transcription reaction (Applied Biosystems). The cDNA were then subjected to a real-time TaqMan PCR in a GeneAmp 5700 sequence detection system (Applied Biosystems). The relative gene expression was normalized to 18s RNA.

4.3. Results

4.3.1. Validation of Statistical Techniques. To validate the results presented in Chapter 3 for this single-channel array platform, the microarray data were processed using both the minimum and average-intensity-based variance estimation algorithm without outlier detection. The average residual error for spots whose actual standard deviation fell below the pooled standard deviation was 0.23 using the minimum intensity metric and 0.25 using the average intensity metric; the average residual error for spots whose actual standard deviation fell above the pooled standard deviation was 0.17 using the minimum intensity metric and 0.19 using the average intensity metric. Thus, the minimum intensity metric produced a slightly better fit than the average intensity metric.

4.3.2. Genes Transcriptionally Regulated by IL-1 in HUVEC. Fig. 4.1 plots the average intensity of spots from the IL-1 arrays against the average intensity of the corresponding spots from the control arrays, with statistically significantly regulated genes highlighted in color. Overall, 491 genes (523 spots) were upregulated by the IL-1 stimulus and 259 genes (275 spots) were downregulated by the IL-1 stimulus. These numbers indicate that approximately 4.7% of the human genome is transcriptionally regulated in HUVEC by IL-1. The list of all these genes is found in Appendix A.

Fig. 4.2 shows the physical chromosome location of annotated up- and down-regulated genes. These genes are distributed throughout the genome, with notable clusters at on

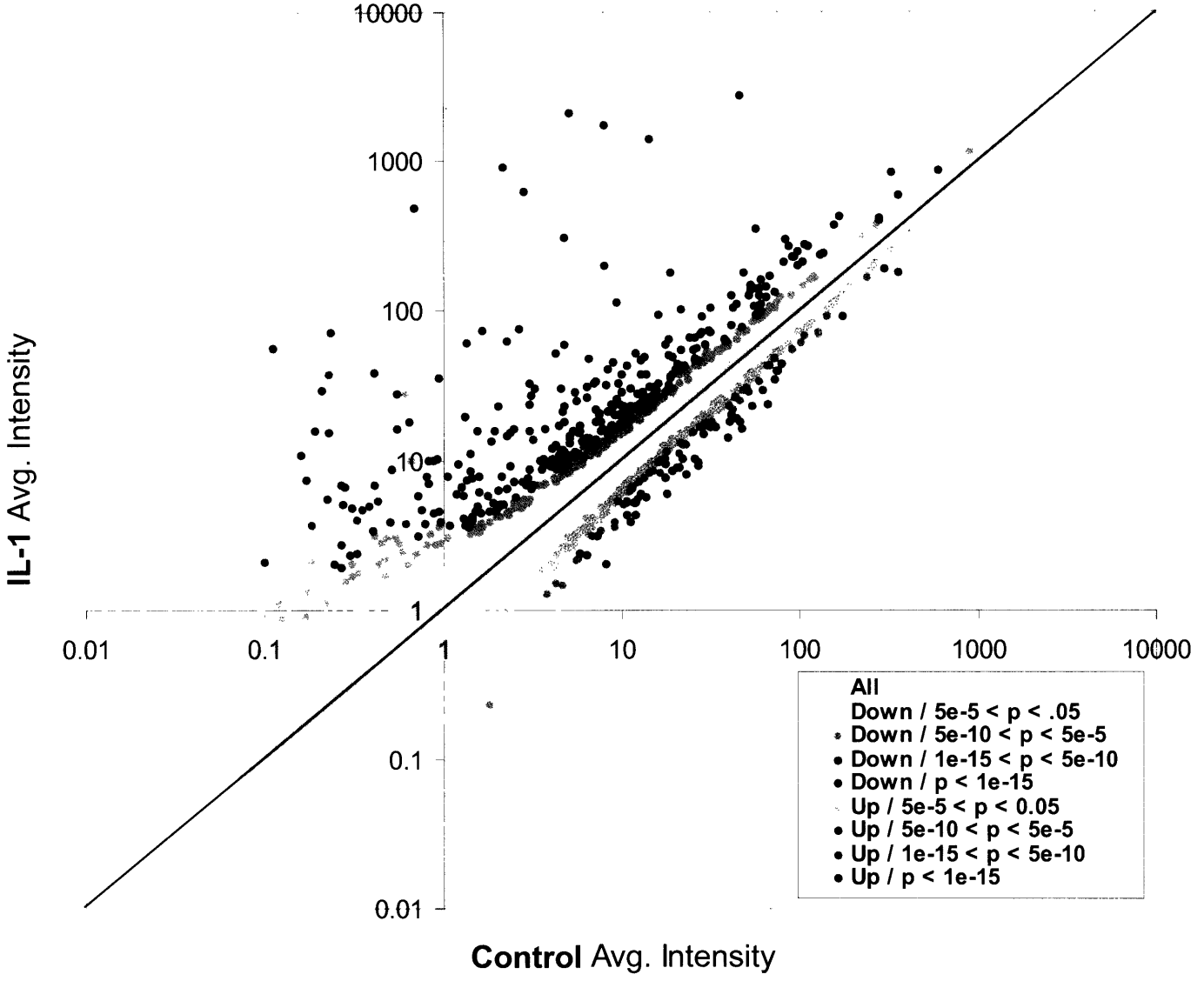

Fig. 4.1. *Scatterplot of average spot intensitiies for control vs. IL-1 treated samples. Statistically significant differentially expressed genes highlighted in color.*

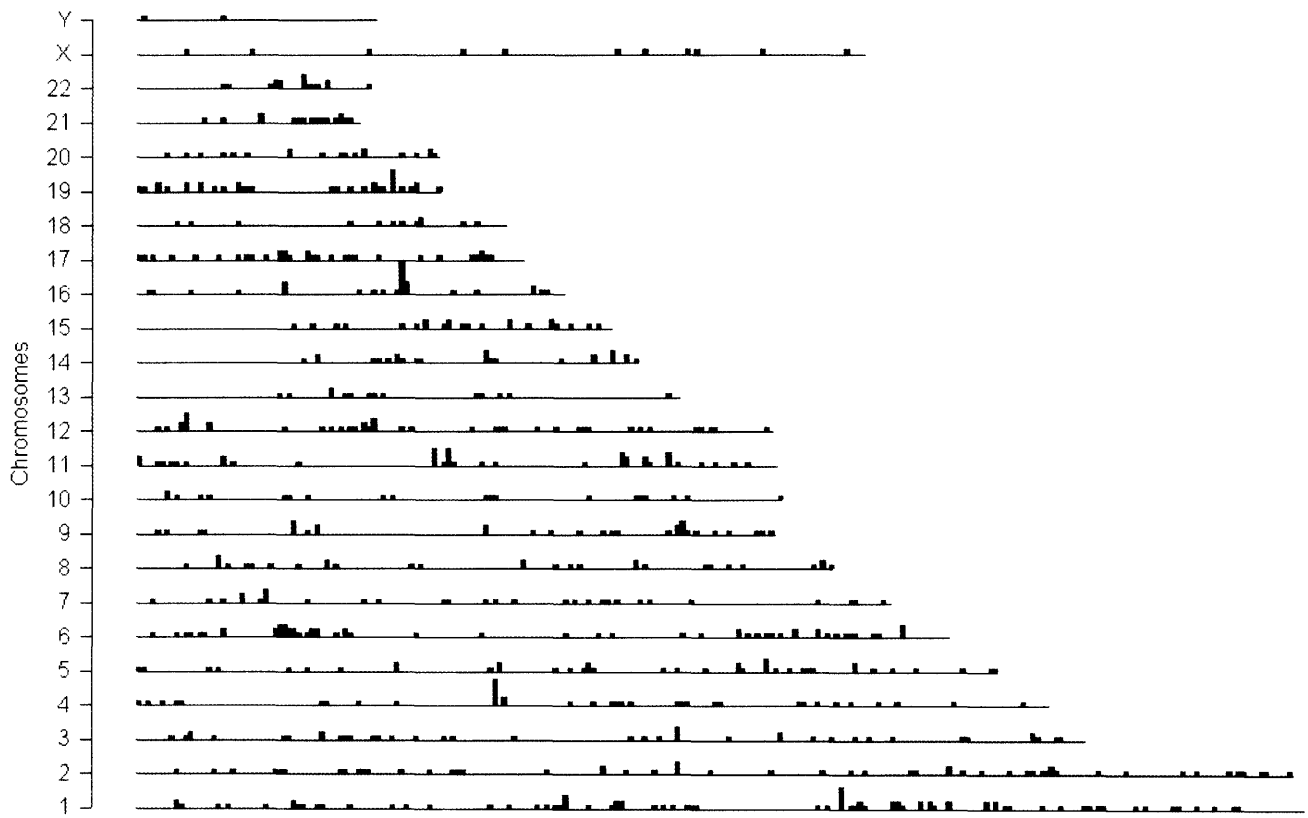

Fig. 4.2. *Plot showing chromosomal location of annotated genes on microarray that were statistically upregulated (red) or downregulated (blue) by IL-1 treatment. Bar height corresponds to relative number of genes in given location.*

Simplified Panther Function	No. of Up-regulated Genes	No. of Down-regulated Genes	Percentage of Up-regulated Genes	Percentage of Down-regulated Genes
Transcription factor	51	49	10.4%	18.9%
Nucleic acid binding	12	13	2.4%	5.0%
Cytoskeletal protein	12	11	2.4%	4.2%
Cytokine	18	0	3.7%	0.0%
Myelin protein	17	7	3.5%	2.7%
G-protein	17	6	3.5%	2.3%
Defense/immunity protein	17	2	3.5%	0.8%
Kinase	15	8	3.1%	3.1%
Transporter	12	7	2.4%	2.7%
Receptor	12	7	2.4%	2.7%
Phosphatase	12	1	2.4%	0.4%
Signaling molecule	11	4	2.2%	1.5%
Cell adhesion molecule	11	3	2.2%	1.2%
Protease	11	2	2.2%	0.8%
G-protein coupled receptor	6	5	1.2%	1.9%
Transferase	9	4	1.8%	1.5%
Chaperone	8	2	1.6%	0.8%
Oxidoreductase	8	2	1.6%	0.8%
Chemokine	8	0	1.6%	0.0%
Cytokine receptor	8	0	1.6%	0.0%
Membrane traffic protein	7	3	1.4%	1.2%
Extracellular matrix	6	0	1.2%	0.0%
Membrane-bound signaling molecule	5	3	1.0%	1.2%
Transfer/carrier protein	5	2	1.0%	0.8%
Kinase modulator	5	1	1.0%	0.4%
Protease inhibitor	4	0	0.8%	0.0%
Growth factor	3	2	0.6%	0.8%
Cell junction protein	3	2	0.6%	0.8%
Select calcium binding protein	2	2	0.4%	0.8%
Protein kinase receptor	1	2	0.2%	0.8%
Isomerase	0	2	0.0%	0.8%
Synthase and synthetase	3	1	0.6%	0.4%
Hydrolase	3	1	0.6%	0.4%
Ligase	2	1	0.4%	0.4%
Ion channel	2	1	0.4%	0.4%
Transmembrane receptor regulatory/adaptor protein	0	1	0.0%	0.4%
Peptide hormone	0	1	0.0%	0.4%
Other enzyme regulator	1	0	0.2%	0.0%
Storage protein	1	0	0.2%	0.0%
Phosphatase modulator	1	0	0.2%	0.0%
Select regulatory molecule	1	0	0.2%	0.0%
Lyase	0	0	0.0%	0.0%
Structural protein	0	0	0.0%	0.0%
Surfactant	0	0	0.0%	0.0%
Viral protein	0	0	0.0%	0.0%

Table 4.1. *Summary of simplified Panther classifications for annotated genes determined to be statistically regulated by IL-1 in cultured HUVEC.*

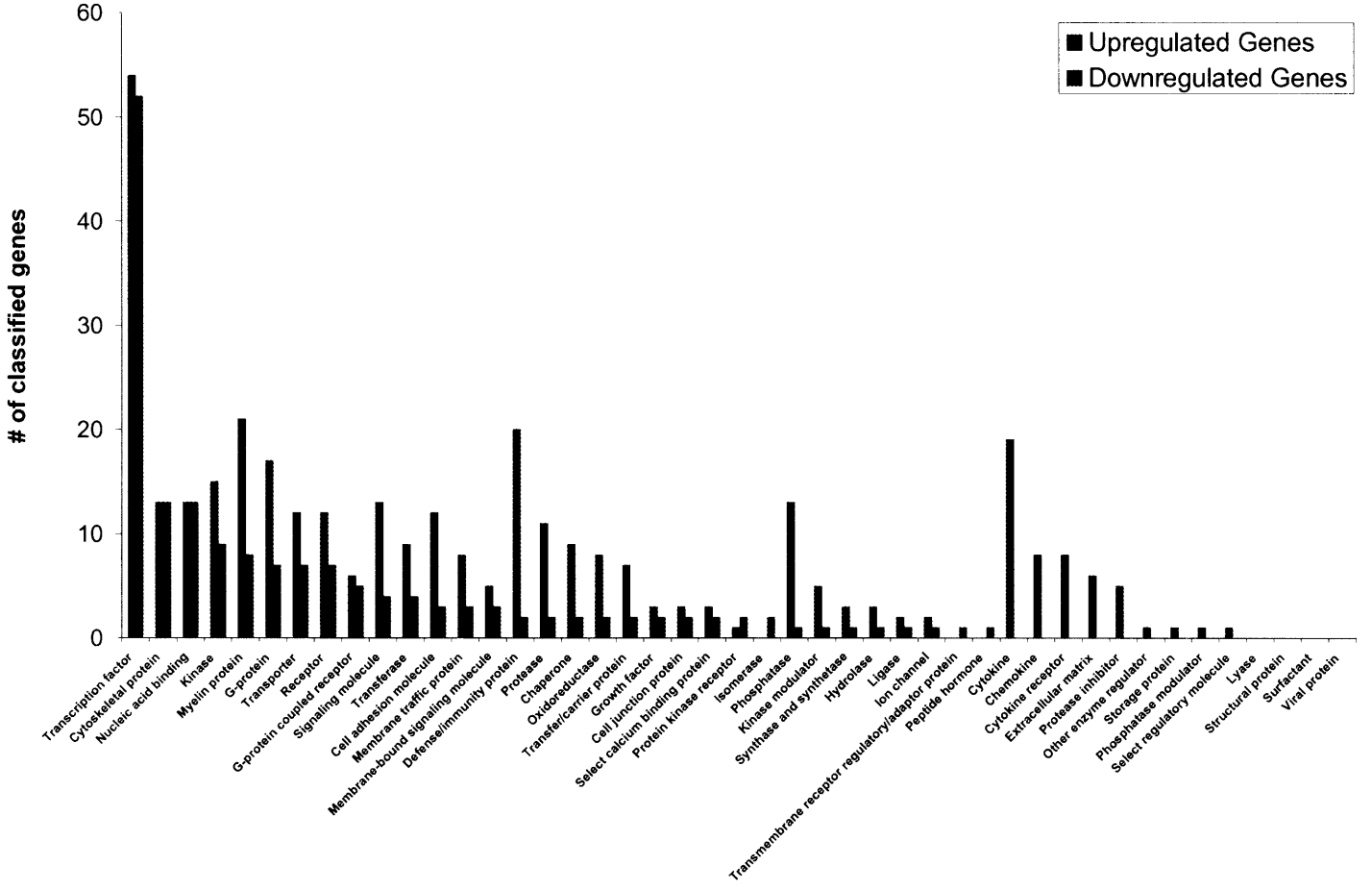

Fig. 4.3. *Histogram illustrating the number of IL-1-upregulated (red) and downregulated (blue) genes on the array associated with each simplified Panther category.*

chromosomes 4, 6, and 11. The chromosome 4 cluster includes upregulated chemokines IL-8, and CXCL1, 2, 3, 5 and 6. The chromosome 16 cluster includes upregulated genes ubiquitin D, MHCI-HLA-A and E, ABCF1 and IER3. The chromosome 11 cluster includes 3 upregulated genes—BIRC2 and 3, and MMP10, as well as the downregulated metalloprotease MMP13.

Given such a large number of genes, it is useful to group them into functional categories based on the simplified Panther categories described in Chapter 2. Of the 750 regulated genes, 439 were associated with one or more simplified categories, as given in Table 4.1; Fig. 4.3 presents a bar graph showing the number of regulated genes found in each category. Most categories contain both up- and down-regulated genes, while a few contain only upregulated genes.

4.3.3. Select QRT-PCR Validation of Microarray Results. Of the 95 genes validated by QRT-PCR (see Table 4.2), 20 were found to be significantly regulated using a t-test on the 3 replicate logged PCR ratios. These 95 PCR probes corresponded to 108 microarray spots (due to duplicate spots for a given gene). Of the 21 spots that corresponded to a statistically regulated gene by PCR, 14 (sensitivity of 67%) were detected as statistically significantly regulated on the microarrays. Of the 7 genes that were not detected on the microarrays, 3 had average ratios greater than 2, but were not statistically detected due to noise and another 3 had ratios less than 2 fold by PCR. Interferon gamma was the only validated gene regulated greater than 2x (22.9 fold) by PCR and not detected by microarray. Of the remaining 87 spots that corresponded to a non-regulated gene by PCR, 3 (false positive rate of 3.4%) were detected as statistically significantly regulated on the microarrays. Of note, all 3 of these genes were regulated by 1.9-fold or higher on PCR, but results were too variable to be found statistically significant. Since

the PCR was performed specifically on genes associated with the inflammatory process, the specificity may be artificially high.

Name	Avg PCR Ratio	Direct Labeled Ratio
CSF2	**630.6**	**518.1**
CSF3	**595.1**	**51.4**
SELE	**471.1**	**432.2**
CXCL10	**176.6**	*9.2*
MCP1	**122.4**	**107.8**
ICAM1	**71.0**	**64.0**
IL8	**69.2**	**58.9**
TNF	**49.3**	*2.0*
IFNg	**22.9**	*-1.2*
IL6	**21.8**	**23.8**
IL1b	**11.1**	**6.6**
IL1a	**10.5**	**16.9**
NFKB2	**8.8**	**11.1**
IL15	**7.9**	**7.1**
CSF1	**7.2**	**10.1**
CSF1	**7.2**	*6.0*
MADH3	**2.1**	**2.4**
LTA	**1.8**	*1.3*
MADH7	**-2.2**	**-1.7**
CD68	**-1.3**	*1.2*
ACE	**-1.2**	*-1.1*
CCL5	17.4	2.2
CXCL11	2.8	2.3
IL12A	2.4	-1.1
TNFRSF18	2.2	1.2
VEGF	2.2	3.3
CYP7A1	2.0	-1.5
PTGS2	1.9	6.2
IL7	1.6	1.2
IL7	1.6	-1.2
IL7	1.6	-1.2
CCR4	1.4	-1.1
CD28	1.2	1.1
IL18	1.1	-1.3
EC1	1.1	1.4
CD38	1.1	-1.1

Name	Avg PCR Ratio	Direct Labeled Ratio
Stat3	1.1	1.2
Stat3	1.1	1.3
Stat3	1.1	1.1
CCL3	-3.6	-1.2
CCL3	-3.6	-1.5
CCL3	-3.6	2.1
PRF1	-3.1	1.1
CXCR3	-3.0	-1.4
PTPRC	-2.7	1.1
PTPRC	-2.7	-1.0
CD3	-1.9	-1.5
BCL2	-1.8	-1.1
BCL2	-1.8	-1.0
COL4A5	-1.7	-1.0
IL5	-1.6	-1.1
IkB2	-1.5	1.0
SKI	-1.5	-1.1
AGTR1	-1.5	-1.1
GZMB	-1.5	-1.2
HMOX1	-1.4	-1.2
GNLY	-1.4	-1.3
TNFRSF6	-1.4	-1.1
ACTB	-1.3	-1.1
HLADR	-1.3	1.4
TGFB1	-1.3	-1.1
Nos2A	-1.3	-1.8
BAX	-1.3	1.1
BAX	-1.3	1.1
CD8	-1.3	1.6
CD8	-1.3	1.6
SELP	-1.3	1.0
TFRC	-1.2	-1.1
CD34	-1.2	1.0
GUSB	-1.2	-1.2
EDN1	-1.1	-1.1
FN	-1.1	1.1

Name	Avg PCR Ratio	Direct Labeled Ratio
TNFRSF5	-1.1	1.2
GAPDH	-1.1	-1.0
GAPDH	-1.1	-1.1
TBX21	-1.0	-1.0
BCL2L1	-1.0	1.1
C3	-1.0	1.2
AGTR2	n/a	-2.2
CCL19	n/a	1.2
CCR2	n/a	1.2
CCR5	n/a	1.2
CCR7	n/a	-1.0
CD19	n/a	-1.1
CD4	n/a	-1.2
CD80	n/a	1.8
CD86	n/a	-1.1
CTLA4	n/a	1.6
CYP1A2	n/a	1.3
HLADRA	n/a	1.0
Hs00411908	n/a	-1.0
ICOS	n/a	1.5
ICOS	n/a	2.0
IL10	n/a	1.8
IL12B	n/a	-1.0
IL13	n/a	-1.4
IL17	n/a	1.2
IL2	n/a	-1.2
IL2RA	n/a	1.4
IL3	n/a	2.2
IL4	n/a	-1.0
IL4	n/a	1.2
IL9	n/a	2.7
LRP2	n/a	1.4
REN	n/a	1.2
RPL3L	n/a	-1.2
TNFSF5	n/a	1.3
TNFSF6	n/a	-1.4

Table 4.2. *Average IL-1 to control expression ratios in cultured HUVEC as determined by RT-PCR and microarray for a select set of inflammation-related genes. Ratios in bold were determined to be statistically significant. Microarray results that agree with PCR data are shown in red (regulated) and blue (not regulated) and results that disagree with PCR data are shown in pink (not regulated on microarray) and cyan (regulated on microarray). A value of "n/a" indicates RNA levels in both samples were undetected by PCR for more than 1 replicate.*

4.4. Discussion

4.4.1. Comparison to Basal Expression Profile. Of the 750 genes regulated by IL-1, 701 genes were found to be expressed by the statistical analysis detailed in Chapter 2. Thus, only 49 (10%) of the upregulated genes, are "turned on" from a non-expressed state; the others are already "on" and their expression is only modulated. Although this number may be an underestimate due to false positives from the expression analysis, it does reflect that the majority of genes involved in mediating the endothelial inflammatory response are transcribed basally. Thus, the response appears not so much to switch on the activation of completely silent pathways, but rather to amplify the processes of already enabled pathways.

Interestingly, there is a bias towards upregulation of genes over downregulation in response to an IL-1 stimulus. The gene expression profile of cultured HUVEC (cf. Chapter 2) indicates that the majority of highly expressed genes under basal conditions are those required for basic cellular function (e.g., transcription) or key endothelial function (e.g., hemostasis). Thus, the scope for modulating phenotype by downregulating these genes without impairing necessary functions is limited. In fact, in Panther functional categories that contain a number of highly expressed genes that may not be critical for cell survival or identity—e.g., transcription factors or cytotskeletal proteins—there are comparable numbers of IL-1 upregulated and downregulated genes. For the most part, however, the endothelial cell under basal conditions appears to activate at high levels only critical pathways, and responds to stimuli by amplifying additional pathways.

4.6.2. Biological roles of regulated genes. A number of the genes found to be regulated have previously been shown to be IL-1 responsive in HUVEC, such as the upregulation of cell adhesion molecules including ICAM-1, VCAM-1 and E-selectin [99],[100], cytokines such as

IL-6, IL-8, MCP-1, and GM-CSF [99], pro-thrombotic factors such as tissue factor, and the downregulation of anti-thrombotic factors such as thrombomodulin [101] and the glycoprotein thrombospondin [102]. The Panther classifications underscore their patterns; the largest categories containing only upregulated genes are cytokine, chemokine and cytokine receptor and categories highly enriched in upregulated genes include defense/immunity protein.

A number of genes not previously recognized to be regulated by IL-1 were also discovered in this dataset. These genes include upregulated transcripts, but downregulated genes are of especial interest in this genome-wide study since previous studies on the effect of IL-1 on endothelial cells have focused primarily on which gene products are upregulated by the inflammatory stimulus, and report far more upregulated than downregulated genes [5, 103]. As noted earlier, the largest number of downregulated genes fall in the transcription factor, nucleic acid binding and cytoskeletal protein Panther categories (see Fig. 4.3).

Two Kruppel-like factors, KLF3 and KLF7, are upregulated by IL-1. KLF7 has been shown to be upregulated in HUVEC by IL-1 at other time points [5], but this study is the first to note upregulation of KLF3. Interestingly, KLF2, which shares a close homology with these other factors, is downregulated by IL-1. While KLF2 has been characterized as an anti-inflammatory transcription factor [36], the others have not been implicated functionally in any inflammatory process to date.

Several TNF-alpha-induced proteins (TNFAIP's) and TNFAIP-interacting proteins, which appear to have a net anti-inflammatory and cytoprotective effect, are upregulated in our dataset. TNFAIP2, 3, and 6 have been shown to be regulated by IL-1 in previous transcriptional profiling studies as well [5, 103]. Our dataset also includes the upregulation of TNFAIP1 and 8, as well as TNFAIP3 interacting proteins 1, 2 and 3. The roles of TNFAIP1, 2 and 8 are not well

characterized; TNFAIP3 has been shown to inhibit NF-Kappa B activation and apoptosis [104] and TNFAIP6 activates inter-alpha-inhibitor, an anti-inflammatory agent [105]. TNFAIP3 interacting protein 2 (TNIP2) inhibits endothelial apoptosis [106]. In contrast, TNIP1 appears to attenuate ERK2 signaling, which may have a pro-apoptotic effect [107]. TNIP3, which was among the top 50 upregulated genes (11.8x), is a cytoskeletal protein whose only known role is its induction by Listeria in macrophages [108]. Thus, IL-1 appears to activate in concert the transcription of several related and possibly redundant genes preventing apoptosis, perhaps as a part of a negative feedback loop.

The Notch pathway has previously been shown to inhibit the NF-Kappa B pathway [109], which plays an important role in mediating inflammatory responses. It has also been shown to be activated under TNF stimulation in the context of rheumatoid arthritis [110]. Notch ligand jagged 1 has been seen as upregulated in HUVEC by IL-1 in previous transcriptional profiling studies, but our dataset indicates that several other genes related to the Notch pathway are regulated in HUVEC by IL-1 at the 4-hour timepoint. The downstream Gridlock homolog Hey-1 is upregulated (1.5 fold), along with Notch activator presenilin 1 (1.6 fold) and Notch ligand jagged 1 (2.6 fold). The ligand jagged 2, however, is downregulated (1.7 fold). The increased expression of Notch pathway members under IL-1 stimulation may indicate a negative feedback loop to moderate the NF-Kappa B response.

The Notch pathway, which is involved in vascular development, may also be regulated as part of an angiogenic response to an inflammatory stimulus. Concurrent regulation of other developmental or angiogenesis-related genes include both Ephrin A1 and B1, which were upregulated (4.3 and 2.5 fold, respectively), and Ephrin ligand EphA4, which was downregulated (1.7 fold). The Ephrin-Eph genes have been implicated mostly in development and angiogenesis

[111], and their regulation under IL-1 suggests either that even at an early time point, this stimulus causes endothelial cells to modulate their angiogenic phenotype. Several other known or putative pro-angiogenic factors were found to be upregulated, including the cytokine IL-8, VEGF, and adrenomedullin [112]. In contrast, another angiogenic factor, placental growth factor [113], was downregulated.

Atherosclerosis, considered to be an inflammatory process, also includes the accumulation of many lipids in its lesions. Interestingly, two species of apolipoprotein L are upregulated by IL-1 in cultured HUVEC. Apolipoprotein L has previously been demonstrated to be upregulated in a TNF-alpha-induced endothelial inflammatory response and to be present in atherosclerotic lesions [114]. Also upregulated (1.6 fold) is seipin, the gene implicated in Bernardinelli-Seip congenital lipodystrophy; loss of function of this gene leads to loss of fat accumulation [115]. Upregulation of this gene in endothelial cells may be involved in the lipid dysregulation that is often seen in the context of vascular inflammation.

Suprisingly, also upregulated is EDG-1, a molecule implicated in angiogenesis and formation of adherens junctions in endothelial cells [116], although increased permeability would be an expected endothelial response to an inflammatory stimulus. Perhaps this gene is upregulated as part of a negative feedback loop in response to the increased permeability one would expect to find in an endothelial inflammatory response.

4.6.3. Advantages of Genome-wide Screening of Differential Gene Expression. The ability to interrogate the entire genome for changes in HUVEC gene expression due to IL-1 stimulation has revealed a number of new genes regulated, such as KLF2, Ephrin B1, HEY-1 and jagged 2, in addition to several genes known to be IL-1 responsive such as E-selectin, VCAM and IL-8. This study is the first to look at the effect of IL-1 stimulation of HUVEC with total

genome array technology and with replicate measurements. Comparisons with other studies are confounded by the use of different time points and experimental conditions such as IL-1 dosage and source of endothelial cells [5, 103, 117]. Compared to the two closest studies that profiled HUVEC treated with IL-1 10 U/mL across a range of time points [5, 103], this study has found far more regulated genes, and a greater percentage of downregulated genes. Zhao, et al. looked only at a subset of the genome, 4,000 genes, and found 33 genes to be regulated in at least 1 of 5 different time points, of which 10% were downregulated at 4 hours [103]. Mayer, et al. used a more comprehensive genome array examining approximately 30,000 genes, and found 137 regulated genes in at least 1 of 3 different time points, of which approximately 10% were downregulated. Our previous transcriptional profiling experiments with non-total-genome arrays (using identical experimental conditions and comparable statistical processing) have also produced a smaller number of regulated genes and smaller fraction of downregulated genes [118]. The bias towards upregulated genes may indicate that previous arrays that spotted a limited number of genes were enriched in genes characterized as having increased expression under activating stimuli.

One factor affecting the overall high numbers of regulated genes that we found at just one time point compared to other studies is our use of replicate microarrays and statistical analysis. Instead of setting arbitrary fold cutoffs, we have been able to select genes according to a statistical cutoff of the likelihood of their being truly regulated. Thus, we are not limited by the size of ratios, e.g., 4x in Mayer, et al.'s study, and can detect important genes that are regulated at lower fold differences but consistently so. Such genes include both upregulated species such as apolipoprotein L2 (3.3x), TNFAIP3 interacting protein 2 (1.8x), the apoptosis-related caspase 7 (1.7x) and downregulated species such as connexin 37 (2.8x) and thrombospondin (-1.9x). The

use of replicates also reduces the number of false positives in our lists of differentially regulated genes, since reproducibility is taken into account. The reliability of our statistical approach is reflected in the PCR data, especially in the high sensitivity (96.6%).

In addition to unveiling additional genes regulated under an inflammatory stimulus, a genome-wide exploration of the HUVEC response to IL-1 has enriched the existing knowledge we have of which pathways are affected by this stimulus. For example, jagged 1 expression has been previously shown to be regulated by IL-1, but this study is the first to illustrate the concomitant regulation of several other elements of the Notch pathway. Similarly, a number of additional molecules in the TNFAIP family and the Ephrin/Eph family have been newly shown to be regulated by IL-1. These results underscore the complexity of the endothelial response to a simple chemical stimulus; both receptors and ligands, as well as genes with overlapping function, are regulated at a transcriptional level. Such results can be used to discover novel DNA binding sites for transcription factors that affect multiple genes regulated by IL-1 [5].

Perhaps most revealing from a genome-wide study of the endothelial response to a simple stimulus are the conflicting directions in which genes involved in the same pathway or function are regulated. The contrasting regulation of jagged 1 and 2, or of the pro-angiogenic VEGF and anti-angiogenic placental growth factor, highlights the complexity of the feedback mechanisms that are activated in response to IL-1. By further defining the genome-wide expression profiles of HUVEC responding to the same stimulus at different time points, as well as to different stimuli that modulate some of the same pathways, will provide the data required to reverse engineer these rich regulatory transcriptional networks.

5. Effect of Linear Amplification of RNA on Microarray Analysis

5.1. Introduction

The genome-wide expression profiling techniques we have described can be applied to the study of an endless spectrum of tissues and conditions, limited only by sample availability. It may be difficult to obtain sufficient RNA material (~40 μg of total RNA) to perform a microarray hybridization using patient biopsy specimens or under certain experimental conditions. RNA amplification techniques have been developed to allow enough labeled material to be generated starting from only 1-10 μg of RNA, a fraction of the material required for a direct-labeling experiment [119].

RNA amplification techniques must be mostly linear to be of use in a differential expression study. Although most RNA amplification techniques have been demonstrated to act linearly across a broad range of concentrations [120, 121], there is undoubtedly an introduction of additional noise via the amplification processes [122]. For example, the amplified RNA sample may be enriched in shorter RNA sequences due to preferential transcription or degradation [123]. RNA amplification has higher fidelity on the 3' end of an mRNA compared to the 5' end; thus genes queried with probes that represent the 5' end may not be accurately detected on microarrays [124, 125]. Some of these systematic biases affect both control and treated samples and may have only a minor effect on the ratios of gene expression between conditions. Concentration biases may have a larger impact on differential expression studies. If the amplification factor is greater for species at lower concentrations than for those at higher concentrations, the differential expression ratios will be damped by the amplification process. If the reverse is true, genes whose transcript levels differ only slightly may appear to be strongly regulated. To characterize the effect of amplification on our expression profiling results, we

compared results from our original data and from microarrays using material amplified from the same original RNA samples.

5.2. Methods

5.2.1. RNA isolation. RNA from IL-1 treated and control HUVEC was isolated and quantified in triplicate as detailed in Chapter 4. The same RNA was used for reverse transcriptase – *in vitro* transcription (RT-IVT) amplification so that the effects of amplification could be compared to the data from the original biological specimen.

5.2.2. Amplification and Hybridization. RT-IVT amplification was performed for each sample according to manufacturer's protocols (all reagents supplied by Applied Biosystems, Foster City, CA, unless otherwise specified). Briefly, first a reverse transcription amplification process generated cDNA from 10 μg of original total RNA. The total RNA was mixed with 2.0 μL T7-oligo (dT) primer and 4.0 μL control RNA, then heated to 70°C for 5 min. and cooled to 4°C. 2.0 μL 10X First Strand Buffer Mix and 3.0 μL RT Enzyme Mix were added and the reaction held at 25°C for 10 min., 42°C for 2 hours, 70°C for 15 minutes and then cooled to 4°C. Second strand DNA was generated by adding 95.0 μL nuclease-free water, 30.0 μL 5x Second Strand Buffer and 5.0 μL Second Strand Enzyme Mix, and this reaction was held at 16°C for 2 hrs., 70°C for 15 min. and cooled to 4°C. Double-stranded DNA was then purified by the addition of 150 μL of DNA Binding Buffer, followed by transfer to a DNA purification column, two washes with 700 μL each of DNA Wash Buffer and finally three elutions with 30 μL each of DNA Elution Buffer.

Labeled cRNA for microarray hybridization was generated from 15 μL of the purified cDNA. 11 μL nuclease-free water, 8.0 μL 5x IVT buffer, 4.0 μL digoxigen-11-UTP (Roche) and 2.0 μL IVT Enzyme Mix were added to the cDNA and reacted at 37°C for 9 hrs. then cooled to

4°C. Labeled cRNA was then purified by the addition of 20 μL nuclease-free water, 200 μL RNA Binding Buffer and 140 μL 100% ethanol, transfer to an RNA purification column, two washes with 500 μL each of RNA Wash Buffer and finally two elutions with 50 μL each of RNA Elution Buffer. Quality of the cRNA was verified with the Agilent Bioanalyzer 2100, and absorbance of a 1:30 dilution of the purified cRNA was measured at 260 nm and 320 nm to calculate cRNA concentration as $(A_{260} - A_{320})*1.2$. 10 μg of cRNA in 80 μL of nuclease-free water was combined 10 μL cRNA Fragmentation Buffer, heated to 60°C for 30 minutes then stopped by addition of 50 μL of cRNA Fragmentation Stop Buffer. 150 μL of this fragmented labeled cRNA was used in the microarray hybridization procedure as detailed in Chapter 2. Normalized, surrogated values were used for comparison of individual replicate data and of differential expression analyses; recovered unsurrogated values (see Chapter 2 methods) were used for expression profile analysis.

5.3. Correlation of RT-IVT Signal Values with Non-Amplified Signal Values

For each replicate in each condition, the set of signal values from the RT-IVT microarray was compared to its corresponding set of non-amplified signal values, resulting in squared Pearson's linear correlation coefficients (R^2) of 0.65 to 0.67. If signal values are biased similarly for samples from both conditions, then the ratios may still be valid. For each replicate, the set of ratios after Lowess-normalizing between conditions were compared between the direct and RT-IVT datasets, resulting in fitted slopes of 1.0, 0.93 and 1.1 and R^2 of 0.05, 0.65 and 0.16 for replicates 1, 2 and 3 respectively. This poor correlation, however, appears to be due to a small number of false classifications and many small differences in the ratios of non-regulated genes. Fig. 5.1, which plots the corresponding RT-IVT and direct ratio values against each other for each replicate, illustrates that there are a similar number of points for all three replicates where

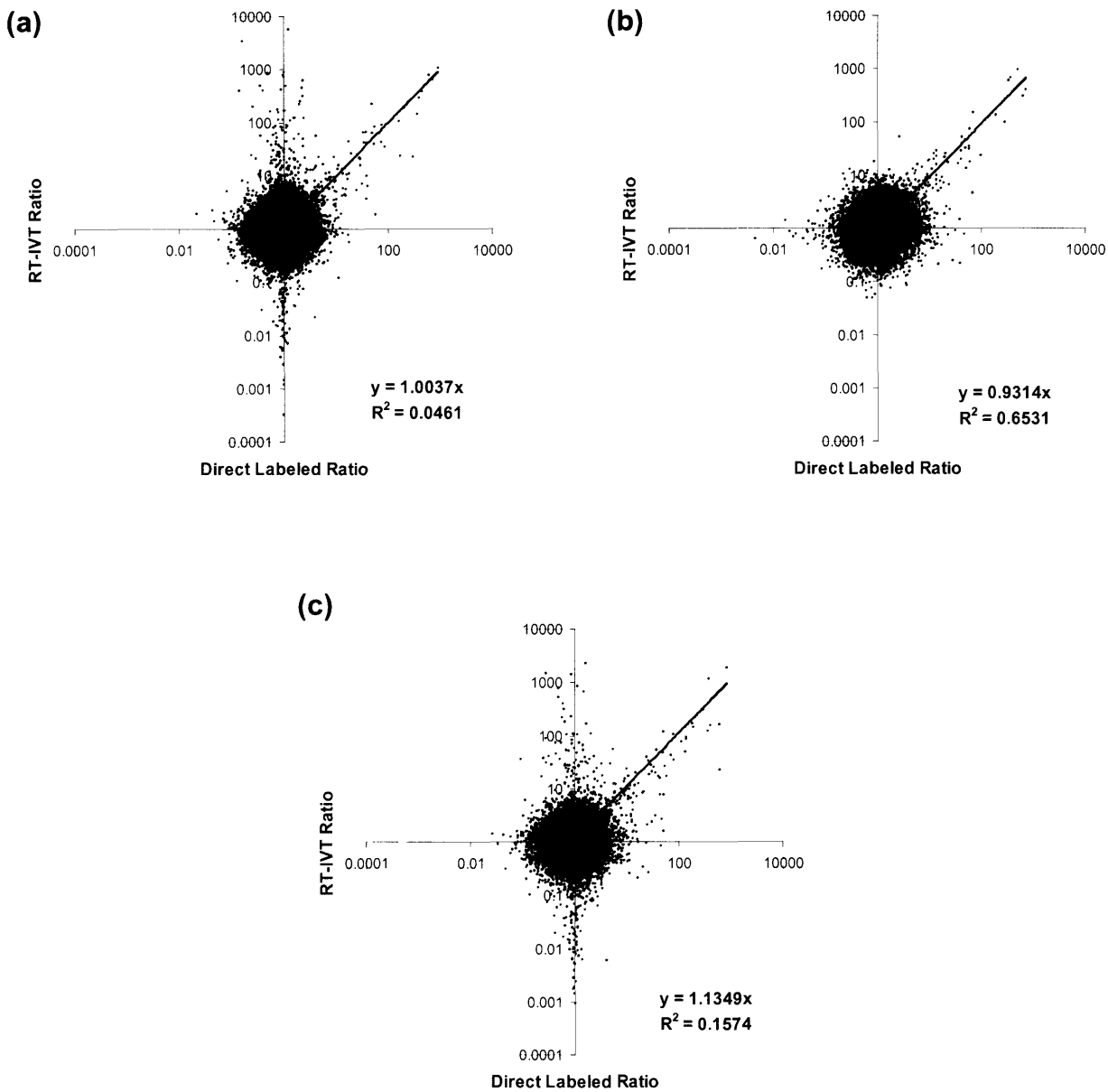

Fig. 5.1. *Plots of corresponding IL-1 to control signal ratios from the RT-IVT dataset vs. the direct labeled dataset. (a) Replicate #1. (b) Replicate #2. (c) Replicate #3.*

the log ratio magnitude is large in the direct dataset, but small for the RT-IVT dataset (scatter parallel to the x-axis), representing false negatives. Examining false positives, however, there are a number of points in replicates #1 and #3, compared to replicate #2, that have small log ratio magnitudes in the direct microarray dataset but large log ratios magnitudes in the RT-IVT dataset. The largest ratios lie close to the y=x line for all replicates; thus the behavior of the most regulated genes appear to be accurately captured when using RNA amplification.

5.4. Expression Analysis with RT-IVT Data

The gene expression profile analysis detailed in Chapter 2 was repeated using the RT-IVT microarray data from control samples. Parameter values for the theoretical distributions and the RMS errors between the theoretical and actual distributions are given in Table 5.1. Most noticeably, the Gaussian standard deviation values are larger, suggesting that the RT-IVT signal values for non-expressed genes are noisier and fall in a broader range of values that direct signal values. Thus, the expressed and non-expressed distributions have a greater overlap and are harder to separate absolutely. The increased noise reduces the fraction of spots predicted to be expressed for each replicate, from ~60% to ~50% ($f = 0.42 - 0.50$). Selecting spots for which all good quality replicates were classified as expressed (using cutoffs that maximized the theoretical true classification rate), a total of 17,848 spots (53.9% of all spots on array) representing 15,914 genes (53.4% of all genes on array) were considered to be expressed. 83% of the spots originally classified as expressed were also classified as expressed based on the RT-IVT data. 3,141 spots classified as expressed based on the unamplified microarray data were not classified as expressed using the RT-IVT data (e.g., taxilin, with an average signal value of 141.08 vs. 0.26), and 2,517 spots were newly classified as expressed based on the RT-IVT data (e.g., the chemokine CCL3, with an average signal value of 31.40 vs. 0.87). Thus, the majority of genes are classified

consistently as expressed or non-expressed under both direct labeling and amplification conditions.

Replicate	#1	#2	#3
Minimum signal value	-3.49	-4.36	-2.44
Maximum signal value	1457.66	2634.80	2647.40
f	0.42	0.47	0.50
μ_N	0.06	0.12	0.21
σ_N	0.33	0.44	0.45
μ_E	1.47	1.67	1.58
σ_E	1.76	1.72	1.60
x_0	0.39	0.56	0.69
RMS Error	8.3E-03	9.2E-03	7.7E-03
ML Cutoff	0.59	0.86	1
True Classification Rate	95.4%	95.3%	96.0%

Table 5.1. *Statistics and parameter values for two-population fit of signal intensity distributions from 3 replicate microarrays of cultured HUVEC RT-IVT-amplified RNA.*

5.5. Differential Expression of Genes

5.5.1. Differentially Expressed Genes Detected on RT-IVT Microarrays. Ultimately, it is the net results from combining replicate data that will be used to interpret microarray data. Thus, we compared the statistical detection of differentially expressed genes after RT-IVT amplification to the genes detected using direct labeling. Using the same statistical algorithms with outlier detection as applied to the unamplified microarray data, 219 genes (234 spots) were found to be statistically significantly upregulated and 82 genes (87 spots) were found to be statistically significantly downregulated on the RT-IVT microarrays. Fig. 5.2a highlights these points on a scatterplot showing the average RT-IVT IL-1 intensity vs. the average RT-IVT control intensity. Overall, a smaller number of genes (301 vs. 706) were statistically detected as regulated. 498 genes originally detected as regulated were not found and 60 spots were newly classified as regulated in the RT-IVT dataset; these points are circled both in the RT-IVT

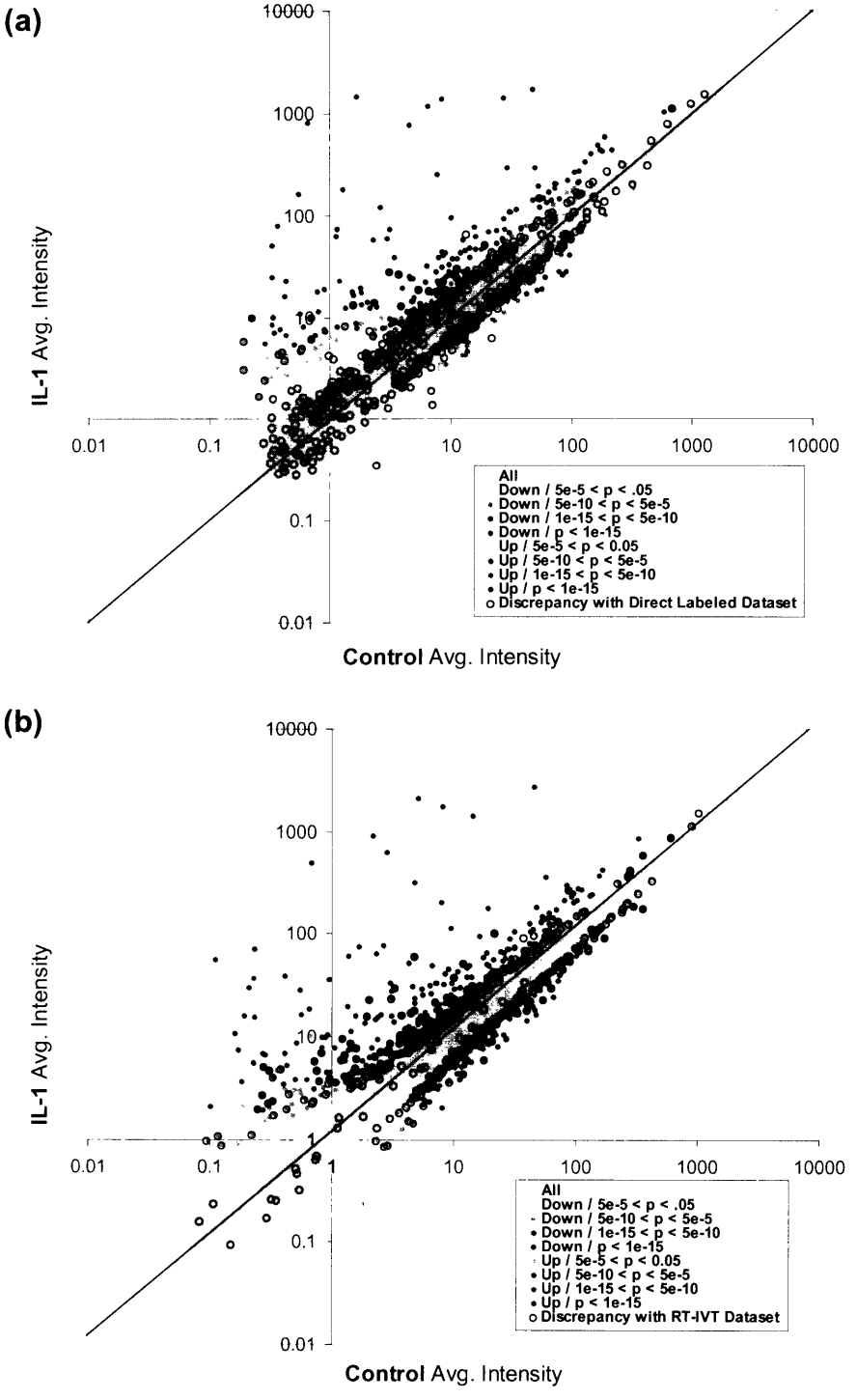

Fig. 5.2. *(a) Scatterplot of average spot intensitiies for control vs. IL-1 treated samples in the RT-IVT dataset. Statistically significant differentially expressed genes highlighted in color and spots whose classification as statistically regulated differed from the direct labeled dataset are circled in black. (b) Scatterplot of average spot intensitiies for control vs. IL-1 treated samples in the direct labeled dataset. Statistically significant differentially expressed genes highlighted in color and spots whose classification as statistically regulated differed from the RT-IVT dataset are circled in black.*

scatterplot shown in Fig. 5.2a, as well as in a scatterplot of the original unamplified data shown in Fig. 5.2b.

Results can vary between the RT-IVT and direct labeled methods for several major reasons. First, non-linear amplification can cause an mRNA species present at low concentration under one condition to be amplified by a greater factor than the same species present at a higher concentration under the other condition. Thus, the expression ratio between conditions would be blunted, leading to "false negative" results. Possible examples of such genes would be CSF1 (10.1x reduced to 1.3), VEGF (3.0x reduced to 1.1x), connexin 40 (-3.8x to –1.8 x) and connexin 37 (-2.8x to –1.5). A bias in amplification may even lead to a reverse ratio; Ephrin A1, for example, appears to be 4.3-fold upregulated in the direct comparison dataset, but 10.7-fold downregulated (not significant) in the RT-IVT dataset.

If the amplification bias is instead towards the higher expressed condition, however, the expression ratio will be enhanced, possibly generating false negative values. For example, CCL3, which was not statistically upregulated according to PCR validation (see Table 5.2), appears to be statistically significantly upregulated by 42-fold in the RT-IVT dataset but only 2-fold upregulated (and not statistically significant) in the direct labeled dataset. A few of the 60 "false positive" genes, however, may actually have genuinely regulated at a low level, and the RT-IVT process, by artificially increasing the ratio, may have enhanced statistical detection of these genes. Possible genes in this category include inflammation-related factors such as CCL7 (8.4x upregulated in the RT-IVT dataset vs. 2.4x upregulated and not significant in the direct labeled dataset), cytokine receptor-like factor 1 (10.4x vs. –1.4x) and CD79A antigen (-.6.7x vs. 1.1x).

The additional biochemical manipulations in the RT-IVT process change the noise characteristics of the dataset. The overall trend is an increase in noise; thus, fewer genes are found statistically significant. This effect can be due solely to the increased standard deviation, seen with genes such as CXCL11 (2.8x regulated in the direct labeled dataset vs. 2.9x regulated and not statistically significant in the RT-IVT dataset), HEY-1 (1.5x vs. 1.6x), presenilin 1 (1.6x vs. 1.7x) and jagged 2 (-1.7x vs. –1.7x). In contrast, the amplification process may proportionally increase the concentration and therefore the intensity of a few mRNA species under both conditions. Since standard deviations of ratios tend to be lower at higher intensities, the RNA amplification process could actually enhance statistical detection of such genes, e.g., the intermediate filament binding protein plectin 1 (ratio of 18.5 to 8.9 (2x) in the RT-IVT dataset vs. 0.2 to 0.1 (also 2x) in the direct labeled dataset). Thus, effects on ratio magnitudes and ratio variance, as well as the combination of the two, cause the discrepancies seen between the RT-IVT and direct labeled results.

5.5.2. Validation with PCR. Comparing the data to the PCR results given in Chapter 4, 12 out of 21 (57.1% sensitivity or true positive rate) spots representing genes shown to be regulated by QRT-PCR and 2 out 87 (2.3% false positive rate) of the spots representing genes not shown to be regulated by QRT-PCR are found to be statistically significantly regulated on the RT-IVT microarrays (see Table 5.2). The sensitivity value of 57.1% is slightly worse than the value of 67% seen with unamplified microarray data. Of the 9 "false negative" spots not found to be regulated in the RT-IVT microarray dataset, 5 have ratios over 2-fold on RT-PCR but under 2-fold on the microarray. This large difference in computed ratios is seen in only 1 out 7 "false negative" spots for the direct labeled dataset. The false positive rate of 2.3% compared to 3.4% for unamplified data is not appreciably different. Of the 2 false positives, however, the

RT-PCR dataset shows CCL3 to be upregulated by 42-fold, whereas the PCR results indicate a non-significant 3.6-fold downregulation. None of the false positive discrepancies were of this magnitude in the direct labeled dataset.

Name	Avg PCR Ratio	RT-IVT Ratio	Direct Labeled Ratio
CSF2	**630.6**	**406.9**	**518.1**
CSF3	**595.1**	**6.1**	**51.4**
SELE	**471.1**	**177.6**	**432.2**
CXCL10	**176.6**	**11.3**	*9.2*
MCP1	**122.4**	**57.2**	**107.8**
ICAM1	**71.0**	**59.5**	**64.0**
IL8	**69.2**	**37.2**	**58.9**
TNF	**49.3**	*-1.3*	*2.0*
IFNg	**22.9**	*-1.7*	*-1.2*
IL6	**21.8**	**31.4**	**23.8**
IL1b	**11.1**	**8.2**	**6.6**
IL1a	**10.5**	**11.9**	**16.9**
NFKB2	**8.8**	*2.6*	**11.1**
IL15	**7.9**	**8.3**	**7.1**
CSF1	**7.2**	*1.3*	**10.1**
CSF1	**7.2**	*-1.1*	*6.0*
MADH3	**2.1**	**2.5**	**2.4**
LTA	**1.8**	*1.4*	*1.3*
MADH7	**-2.2**	*-1.6*	**-1.7**
CD68	**-1.3**	*-1.3*	*1.2*
ACE	**-1.2**	*-1.6*	*-1.1*
CCL5	17.4	1.8	2.2
CXCL11	2.8	2.9	*2.8*
IL12A	2.4	1.3	-1.1
TNFRSF18	2.2	-1.4	1.2
VEGF	2.2	1.1	*3.0*
CYP7A1	2.0	1.3	-1.5
PTGS2	1.9	*6.0*	*6.2*
IL7	1.6	-1.3	1.2
IL7	1.6	-1.0	-1.2
IL7	1.6	-1.0	-1.2
CCR4	1.4	-1.7	-1.1
CD28	1.2	-1.0	1.1
IL18	1.1	1.0	-1.3
EC1	1.1	1.5	1.4
CD38	1.1	-1.1	-1.1
Stat3	1.1	1.2	1.2
Stat3	1.1	1.1	1.3
Stat3	1.1	1.3	1.1
CCL3	-3.6	1.2	-1.2
CCL3	-3.6	-1.1	-1.5
CCL3	-3.6	*42.3*	2.1
PRF1	-3.1	1.2	1.1
CXCR3	-3.0	1.8	-1.4
PTPRC	-2.7	1.1	1.1
PTPRC	-2.7	1.4	-1.0
CD3	-1.9	-2.6	-1.5
BCL2	-1.8	-1.2	-1.1
BCL2	-1.8	1.1	-1.0
COL4A5	-1.7	-1.0	-1.0
IL5	-1.6	1.2	-1.1
IkB2	-1.5	-1.1	1.0
SKI	-1.5	-1.1	-1.1
AGTR1	-1.5	1.1	-1.1
GZMB	-1.5	-1.3	-1.2
HMOX1	-1.4	-1.5	-1.2
GNLY	-1.4	1.5	-1.3
TNFRSF6	-1.4	-1.1	-1.1
ACTB	-1.3	-1.2	-1.1
HLADR	-1.3	-1.3	1.4
TGFB1	-1.3	-1.2	-1.1
Nos2A	-1.3	-1.0	-1.8
BAX	-1.3	-1.0	1.1
BAX	-1.3	1.2	1.1
CD8	-1.3	1.8	1.6
CD8	-1.3	-1.1	1.6
SELP	-1.3	-1.0	1.0
TFRC	-1.2	-1.2	-1.1
CD34	-1.2	-1.1	1.0
GUSB	-1.2	-1.1	-1.2
EDN1	-1.1	1.0	-1.1
FN	-1.1	1.2	1.1
TNFRSF5	-1.1	1.4	1.2
GAPDH	-1.1	-1.1	-1.0
GAPDH	-1.1	1.1	-1.1
TBX21	-1.0	1.5	-1.0
BCL2L1	-1.0	-1.7	1.1
C3	-1.0	-1.6	1.2
AGTR2	n/a	-1.1	-2.2
CCL19	n/a	-1.3	1.2
CCR2	n/a	1.0	1.2
CCR5	n/a	1.1	1.2
CCR7	n/a	-1.6	-1.0
CD19	n/a	2.4	-1.1
CD4	n/a	1.7	-1.2
CD80	n/a	-1.2	1.8
CD86	n/a	1.1	-1.1
CTLA4	n/a	1.8	1.6
CYP1A2	n/a	-2.3	1.3
HLADRA	n/a	1.5	1.0
Hs00411908	n/a	1.3	-1.0
ICOS	n/a	-1.3	1.5
ICOS	n/a	1.8	2.0
IL10	n/a	-1.9	1.8
IL12B	n/a	1.5	-1.0
IL13	n/a	2.0	-1.4
IL17	n/a	1.0	1.2
IL2	n/a	-1.6	-1.2
IL2RA	n/a	-1.2	1.4
IL3	n/a	1.2	2.2
IL4	n/a	1.5	-1.0
IL4	n/a	1.4	1.2
IL9	n/a	1.4	2.7
LRP2	n/a	1.1	1.4
REN	n/a	-1.2	1.2
RPL3L	n/a	-1.1	-1.2
TNFSF5	n/a	1.1	1.3
TNFSF6	n/a	1.3	-1.4

Table 5.2. *Average IL-1 to control expression ratios in cultured HUVEC as determined by RT-PCR, RT-IVT microarray and direct labeled microarray for a select set of inflammation-related genes. Ratios in bold were determined to be statistically significant. Microarray results that agree with PCR data are shown in red (regulated) and blue (not regulated) and results that disagree with PCR data are shown in pink (not regulated on microarray) and cyan (regulated on microarray). A value of "n/a" indicates RNA levels in both samples were undetected by PCR for more than 1 replicate.*

These qualitative differences highlight the limitations of RT-IVT amplification. With this dataset alone, it is impossible to determine if the errors introduced by RT-IVT are systematic or random. For example, genes certain sequences may be more prone to non-linear amplification. Also, non-linearities may be pronounced in specific ranges of concentration. Other datasets that directly compare direct labeling and RT-IVT data for other conditions under which different genes are regulated may provide further insight into the nature of the incurred noise due to RT-IVT amplification. Nonetheless, our data illustrates that this technique reliably detects the majority of top statistically regulated genes when the amount of starting material is limited.

6. Conclusions

We have studied gene expression in cultured human endothelium at a genome-wide level under both basal and inflammatory conditions. We have developed generalized analytical techniques, both to define a comprehensive profile of expressed genes under a specific condition and to determine statistically differentially regulated genes between two conditions. These methods have been applied to define an endothelial transcriptome for cultured HUVEC under basal conditions and to study the genome-wide changes to this transcriptome that are caused by an IL-1 stimulus. This analysis represents the largest snapshot of the transcriptional activity of cultured and IL-1-stimulated HUVEC to date. Finally, we have applied these techniques to data collected using amplified RNA samples, allowing us to characterize the effect that the additional noise may have on the output of statistical analyses.

The tools described here can be used for the rigorous analysis of microarray data studying any biological question. Our laboratory is currently generating genome-wide expression profiles for different types of endothelial cells to continue developing our definition of the endothelial transcriptome—the set of genes required for endothelial identity. By applying the differential expression algorithms to data from more and more experimental conditions, we should enhance our understanding of how this transcriptome is modulated by different stimuli and begin to determine the gene regulatory networks that control endothelial structure and function. As our new bioinformatics tools are used to explore endothelium along these two orthogonal paths—diversity of endothelial cell origin and diversity of environmental stimuli—we will develop a more detailed picture of the physiological and pathophysiological behavior of this vascular interface.

References

1. Schena M, Shalon D, Davis RW, Brown PO. *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. 270(5235): p. 467-70.

2. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J. *Gene-expression profiles in hereditary breast cancer.* N Engl J Med, 2001. 344(8): p. 539-48.

3. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. *Genetic analysis of genome-wide variation in human gene expression.* Nature, 2004. 430(7001): p. 743-7.

4. Khanna S, Cheng G, Gong B, Mustari MJ, Porter JD. *Genome-wide transcriptional profiles are consistent with functional specialization of the extraocular muscle layers.* Invest Ophthalmol Vis Sci, 2004. 45(9): p. 3055-66.

5. Mayer H, Bilban M, Kurtev V, Gruber F, Wagner O, Binder BR, de Martin R. *Deciphering regulatory patterns of inflammatory gene expression from interleukin-1-stimulated human endothelial cells.* Arterioscler Thromb Vasc Biol, 2004. 24(7): p. 1192-8.

6. Rhodes DR, Chinnaiyan AM. *Bioinformatics strategies for translating genome-wide expression analyses into clinically useful cancer markers.* Ann N Y Acad Sci, 2004. 1020: p. 32-40.

7. Ghitescu L, Robert M. *Diversity in unity: the biochemical composition of the endothelial cell surface varies between the vascular beds.* Microsc Res Tech, 2002. 57(5): p. 381-9.

8. Girard JP, Springer TA. *High endothelial venules (HEVs): specialized endothelium for lymphocyte migration.* Immunol Today, 1995. 16(9): p. 449-57.

9. Pepper MS, Skobe M. *Lymphatic endothelium: morphological, molecular and functional properties.* J Cell Biol, 2003. 163(2): p. 209-13.

10. Chang YS, Munn LL, Hillsley MV, Dull RO, Yuan J, Lakshminarayanan S, Gardner TW, Jain RK, Tarbell JM. *Effect of vascular endothelial growth factor on cultured endothelial cell monolayer transport properties.* Microvasc Res, 2000. 59(2): p. 265-77.

11. Wagner WH, Henderson RM, Hicks HE, Banes AJ, Johnson G, Jr. *Differences in morphology, growth rate, and protein synthesis between cultured arterial and venous endothelial cells.* J Vasc Surg, 1988. 8(4): p. 509-19.

12. Topper JN, Gimbrone MA, Jr. *Blood flow and vascular gene expression: fluid shear stress as a modulator of endothelial phenotype.* Mol Med Today, 1999. 5(1): p. 40-6.

13. Bevilacqua MP, Gimbrone MA, Jr. *Inducible endothelial functions in inflammation and coagulation.* Semin Thromb Hemost, 1987. 13(4): p. 425-33.

14. Garcia-Cardena G, Comander J, Anderson KR, Blackman BR, Gimbrone MA, Jr. *Biomechanical activation of vascular endothelium as a determinant of its functional phenotype.* Proc Natl Acad Sci U S A, 2001. 98(8): p. 4478-85.

15. Dai G, Kaazempur-Mofrad MR, Natarajan S, Zhang Y, Vaughn S, Blackman BR, Kamm RD, Garcia-Cardena G, Gimbrone MA, Jr. *Distinct endothelial phenotypes evoked by arterial waveforms derived from atherosclerosis-susceptible and -resistant regions of human vasculature.* Proc Natl Acad Sci U S A, 2004. 101(41): p. 14871-6.

16. Gimbrone MA, Jr., Topper JN, Nagel T, Anderson KR, Garcia-Cardena G. *Endothelial dysfunction, hemodynamic forces, and atherogenesis.* Ann N Y Acad Sci, 2000. 902: p. 230-9; discussion 239-40.

17. *XL2000: Solver Uses Generalized Reduced Gradient Algorithm.* 2003, Microsoft.

18. Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, Ladunga I, Ulitsky-Lazareva B, Muruganujan A, Rabkin S, Vandergriff JA, Doremieux O. *PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification.* Nucleic Acids Res, 2003. 31(1): p. 334-41.

19. Ilan N, Madri JA. *PECAM-1: old friend, new partners.* Curr Opin Cell Biol, 2003. 15(5): p. 515-24.

20. Garcia-Cardena G, Fan R, Shah V, Sorrentino R, Cirino G, Papapetropoulos A, Sessa WC. *Dynamic activation of endothelial nitric oxide synthase by Hsp90.* Nature, 1998. 392(6678): p. 821-4.

21. Tsai HM. *Shear stress and von Willebrand factor in health and disease.* Semin Thromb Hemost, 2003. 29(5): p. 479-88.

22. Acevedo L, Yu J, Erdjument-Bromage H, Miao RQ, Kim JE, Fulton D, Tempst P, Strittmatter SM, Sessa WC. *A new role for Nogo as a regulator of vascular remodeling.* Nat Med, 2004. 10(4): p. 382-8.

23. Feron O, Belhassen L, Kobzik L, Smith TW, Kelly RA, Michel T. *Endothelial nitric oxide synthase targeting to caveolae. Specific interactions with caveolin isoforms in cardiac myocytes and endothelial cells.* J Biol Chem, 1996. 271(37): p. 22810-4.

24. Garcia-Cardena G, Martasek P, Masters BS, Skidd PM, Couet J, Li S, Lisanti MP, Sessa WC. *Dissecting the interaction between nitric oxide synthase (NOS) and caveolin. Functional significance of the nos caveolin binding domain in vivo.* J Biol Chem, 1997. 272(41): p. 25437-40.

25. Jerkic M, Rivas-Elena JV, Prieto M, Carron R, Sanz-Rodriguez F, Perez-Barriocanal F, Rodriguez-Barbero A, Bernabeu C, Lopez-Novoa JM. *Endoglin regulates nitric oxide-dependent vasodilatation.* Faseb J, 2004. 18(3): p. 609-11.

26. Yanagisawa M, Kurihara H, Kimura S, Tomobe Y, Kobayashi M, Mitsui Y, Yazaki Y, Goto K, Masaki T. *A novel potent vasoconstrictor peptide produced by vascular endothelial cells.* Nature, 1988. 332(6163): p. 411-5.

27. Sawamura T, Kimura S, Shinmi O, Sugita Y, Kobayashi M, Mitsui Y, Yanagisawa M, Goto K, Masaki T. *Characterization of endothelin converting enzyme activities in soluble fraction of bovine cultured endothelial cells.* Biochem Biophys Res Commun, 1990. 169(3): p. 1138-44.

28. Russell FD, Davenport AP. *Evidence for intracellular endothelin-converting enzyme-2 expression in cultured human vascular endothelial cells.* Circ Res, 1999. 84(8): p. 891-6.

29. Murakoshi N, Miyauchi T, Kakinuma Y, Ohuchi T, Goto K, Yanagisawa M, Yamaguchi I. *Vascular endothelin-B receptor system in vivo plays a favorable inhibitory role in vascular remodeling after injury revealed by endothelin-B receptor-knockout mice.* Circulation, 2002. 106(15): p. 1991-8.

30. Lawson ND, Weinstein BM. *Arteries and veins: making a difference with zebrafish.* Nat Rev Genet, 2002. 3(9): p. 674-82.

31. Shawber CJ, Kitajewski J. *Notch function in the vasculature: insights from zebrafish, mouse and man.* Bioessays, 2004. 26(3): p. 225-34.

32. Lindner V, Booth C, Prudovsky I, Small D, Maciag T, Liaw L. *Members of the Jagged/Notch gene families are expressed in injured arteries and regulate cell phenotype via alterations in cell matrix and cell-cell interaction.* Am J Pathol, 2001. 159(3): p. 875-83.

33. Rones MS, Woda J, Mercola M, McLaughlin KA. *Isolation and characterization of Xenopus Hey-1: a downstream mediator of Notch signaling.* Dev Dyn, 2002. 225(4): p. 554-60.

34. Fischer A, Schumacher N, Maier M, Sendtner M, Gessler M. *The Notch target genes Hey1 and Hey2 are required for embryonic vascular development.* Genes Dev, 2004. 18(8): p. 901-11.

35. Chi JT, Chang HY, Haraldsen G, Jahnsen FL, Troyanskaya OG, Chang DS, Wang Z, Rockson SG, van de Rijn M, Botstein D, Brown PO. *Endothelial cell diversity revealed by global expression profiling.* Proc Natl Acad Sci U S A, 2003. 100(19): p. 10623-8.

36. SenBanerjee S, Lin Z, Atkins GB, Greif DM, Rao RM, Kumar A, Feinberg MW, Chen Z, Simon DI, Luscinskas FW, Michel TM, Gimbrone MA, Jr., Garcia-Cardena G, Jain MK. *KLF2 Is a Novel Transcriptional Regulator of Endothelial Proinflammatory Activation.* J Exp Med, 2004. 199(10): p. 1305-15.

37. Wang L, Fan C, Topol SE, Topol EJ, Wang Q. *Mutation of MEF2A in an inherited disorder with features of coronary artery disease.* Science, 2003. 302(5650): p. 1578-81.

38. Potente M, Fisslthaler B, Busse R, Fleming I. *11,12-Epoxyeicosatrienoic acid-induced inhibition of FOXO factors promotes endothelial proliferation by down-regulating p27Kip1.* J Biol Chem, 2003. 278(32): p. 29619-25.

39. Caubit X, Thangarajah R, Theil T, Wirth J, Nothwang HG, Ruther U, Krauss S. *Mouse Dac, a novel nuclear factor with homology to Drosophila dachshund shows a dynamic expression in the neural crest, the eye, the neocortex, and the limb bud.* Dev Dyn, 1999. 214(1): p. 66-80.

40. Berge-Lefranc JL, Jay P, Massacrier A, Cau P, Mattei MG, Bauer S, Marsollier C, Berta P, Fontes M. *Characterization of the human jumonji gene.* Hum Mol Genet, 1996. 5(10): p. 1637-41.

41. Shin D, Garcia-Cardena G, Hayashi S, Gerety S, Asahara T, Stavrakis G, Isner J, Folkman J, Gimbrone MA, Jr., Anderson DJ. *Expression of ephrinB2 identifies a stable genetic difference between arterial and venous vascular smooth muscle as well as endothelial cells, and marks subsets of microvessels at sites of adult neovascularization.* Dev Biol, 2001. 230(2): p. 139-50.

42. Seo DW, Li H, Guedez L, Wingfield PT, Diaz T, Salloum R, Wei BY, Stetler-Stevenson WG. *TIMP-2 mediated inhibition of angiogenesis: an MMP-independent mechanism.* Cell, 2003. 114(2): p. 171-80.

43. Visconti RP, Richardson CD, Sato TN. *Orchestration of angiogenesis and arteriovenous contribution by angiopoietins and vascular endothelial growth factor (VEGF).* Proc Natl Acad Sci U S A, 2002. 99(12): p. 8219-24.

44. O'Reilly MS, Boehm T, Shing Y, Fukai N, Vasios G, Lane WS, Flynn E, Birkhead JR, Olsen BR, Folkman J. *Endostatin: an endogenous inhibitor of angiogenesis and tumor growth.* Cell, 1997. 88(2): p. 277-85.

45. Heidemann J, Ogawa H, Dwinell MB, Rafiee P, Maaser C, Gockel HR, Otterson MF, Ota DM, Lugering N, Domschke W, Binion DG. *Angiogenic effects of interleukin 8 (CXCL8) in human intestinal microvascular endothelial cells are mediated by CXCR2.* J Biol Chem, 2003. 278(10): p. 8508-15.

46. Berk BC, Abe JI, Min W, Surapisitchat J, Yan C. *Endothelial atheroprotective and anti-inflammatory mechanisms.* Ann N Y Acad Sci, 2001. 947: p. 93-109; discussion 109-11.

47. Oury C, Sticker E, Cornelissen H, De Vos R, Vermylen J, Hoylaerts MF. *ATP augments von Willebrand factor-dependent shear-induced platelet aggregation through Ca2+-calmodulin and myosin light chain kinase activation.* J Biol Chem, 2004. 279(25): p. 26266-73.

48. Ruffer C, Strey A, Janning A, Kim KS, Gerke V. *Cell-cell junctions of dermal microvascular endothelial cells contain tight and adherens junction proteins in spatial proximity.* Biochemistry, 2004. 43(18): p. 5360-9.

49. Limmer A, Knolle PA. *Liver sinusoidal endothelial cells: a new type of organ-resident antigen-presenting cell.* Arch Immunol Ther Exp (Warsz), 2001. 49 Suppl 1: p. S7-11.

50. Haverson K, Singha S, Stokes CR, Bailey M. *Professional and non-professional antigen-presenting cells in the porcine small intestine.* Immunology, 2000. 101(4): p. 492-500.

51. Chilton JK, Guthrie S. *Development of oculomotor axon projections in the chick embryo.* J Comp Neurol, 2004. 472(3): p. 308-17.

52. Ushkaryov YA, Petrenko AG, Geppert M, Sudhof TC. *Neurexins: synaptic cell surface proteins related to the alpha-latrotoxin receptor and laminin.* Science, 1992. 257(5066): p. 50-6.

53. Heidenreich KA, Linseman DA. *Myocyte enhancer factor-2 transcription factors in neuronal differentiation and survival.* Mol Neurobiol, 2004. 29(2): p. 155-66.

54. Kobayashi H, Itoh S, Yanagita T, Yokoo H, Sugano T, Wada A. *Expression of adrenomedullin and proadrenomedullin N-terminal 20 peptide in PC12 cells after exposure to nerve growth factor.* Neuroscience, 2004. 125(4): p. 973-80.

55. Rieber AJ, Marr HS, Comer MB, Edgell CJ. *Extent of differentiated gene expression in the human endothelium-derived EA.hy926 cell line.* Thromb Haemost, 1993. 69(5): p. 476-80.

56. Tilstone C, *Vital statistics*, in *Nature*. 2003. p. 610-612.

57. Smyth GK, Yang YH, Speed T. *Statistical issues in cDNA microarray data analysis.* Methods Mol Biol, 2003. 224: p. 111-36.

58. Comander J, Weber GM, Gimbrone MA, Jr., Garcia-Cardena G. *Argus--a new database system for Web-based analysis of multiple microarray data sets.* Genome Res, 2001. 11(9): p. 1603-10.

59. Dudoit S, Gentleman RC, Quackenbush J. *Open source software for the analysis of microarray data.* Biotechniques, 2003. Suppl: p. 45-51.

60. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.* Nucleic Acids Res, 2002. 30(4): p. e15.

61. Kepler TB, Crosby L, Morgan KT. *Normalization and analysis of DNA microarray data by self-consistency and local regression.* Genome Biol, 2002. 3(7): p. RESEARCH0037.

62. Churchill GA. *Fundamentals of experimental design for cDNA microarrays.* Nat Genet, 2002. 32 Suppl: p. 490-5.

63. Eisen MB, Spellman PT, Brown PO, Botstein D. *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. 95(25): p. 14863-8.

64. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.* Proc Natl Acad Sci U S A, 1999. 96(6): p. 2907-12.

65. Chen Y, Kamat V, Dougherty ER, Bittner ML, Meltzer PS, Trent JM. *Ratio statistics of gene expression levels and applications to microarray data analysis.* Bioinformatics, 2002. 18(9): p. 1207-15.

66. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. *On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data.* J Comput Biol, 2001. 8(1): p. 37-52.

67. Pan W, Lin J, Le CT. *How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach.* Genome Biol, 2002. 3(5): p. research0022.

68. Kerr MK, Afshari CA, Bennett L, Bushel P, Martinez J, Walker NJ, Churchill GA, *Statistical Analysis of a Gene Expression Microarray Experiment with Replication.* 2001, The Jackson Laboratory: Bar Harbor.

69. Broberg P. *Statistical methods for ranking differentially expressed genes.* Genome Biol, 2003. 4(6): p. R41.

70. Dudoit S, Shaffer JP, Boldrick JC, *Multiple Hypothesis Testing in Microarray Experiments. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 110.* 2002.

71. Lonnstedt I, Speed T, *Replicated microarray data.* 2001, Uppsala University: Uppsala.

72. Tusher VG, Tibshirani R, Chu G. *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. 98(9): p. 5116-21.

73. Efron B, Tibshirani R, Storey JD, Tusher V. *Empirical Bayes analysis of a microarray experiment.* J Am Stat Assoc, 2001. 96: p. 1151-1160.

74. 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(6): p. 509-19.

75. Colantuoni C, Henry G, Zeger S, Pevsner J. *SNOMAD (Standardization and NOrmalization of MicroArray Data): web-accessible gene expression data analysis.* Bioinformatics, 2002. 18(11): p. 1540-1.

76. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH. *Functional discovery via a compendium of expression profiles.* Cell, 2000. 102(1): p. 109-26.

77. Baggerly KA, Coombes KR, Hess KR, Stivers DN, Abruzzo LV, Zhang W. *Identifying differentially expressed genes in cDNA microarray experiments.* J Comput Biol, 2001. 8(6): p. 639-59.

78. Coombes KR, Highsmith WE, Krogmann TA, Baggerly KA, Stivers DN, Abruzzo LV. *Identifying and quantifying sources of variation in microarray data using high-density cDNA membrane arrays.* J Comput Biol, 2002. 9(4): p. 655-69.

79. Quackenbush J. *Microarray data normalization and transformation.* Nat Genet, 2002. 32 Suppl: p. 496-501.

80. Rocke DM, Durbin B. *A model for measurement error for gene expression arrays.* J Comput Biol, 2001. 8(6): p. 557-69.

81. Nadon R, Shi P, Skandalis A, Woody E, Hubschle H, Susko E, Rghei N, Ramm P, *Statistical inference methods for gene expression arrays.* 2001, Imaging Research Inc.: St. Catharines.

82. Tsodikov A, Szabo A, Jones D. *Adjustments and measures of differential expression for microarray data.* Bioinformatics, 2002. 18(2): p. 251-60.

83. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, Quackenbush J. *Within the fold: assessing differential expression measures and reproducibility in microarray assays.* Genome Biol, 2002. 3(11): p. research0062.

84. Cheadle C, Vawter MP, Freed WJ, Becker KG. *Analysis of microarray data using Z score transformation.* J Mol Diagn, 2003. 5(2): p. 73-81.

85. *Agilent Fluorescent Direct Label Kit Protocol Rev. 2.1.* 2003, Agilent Technologies.

86. Yang YH, Buckley MJ, Dudoit S, Speed T, *Comparison of Methods for Image Analysis on cDNA Microarray Data, Technical report #584.* 2000.

87. Sokal R, Rohlf F, *Biometry: the principles and practice of statistics in biological research.* 3rd ed. ed. 2000: W.H. Freeman and Co. 53:53.

88. Ge Y, Dudoit S, Speed T. *Resampling-based multiple testing for microarray data hypothesis.* Test, 2003. 12(1): p. 1-77.

89. Dudoit S, Yang YH, Callow MJ, Speed T. *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.* Statistica Sinica, 2002. 12(1): p. 111-39.

90. Dudley AM, Aach J, Steffen MA, Church GM. *Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range.* Proc Natl Acad Sci U S A, 2002. 99(11): p. 7554-9.

91. Livesey FJ, Furukawa T, Steffen MA, Church GM, Cepko CL. *Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene Crx.* Curr Biol, 2000. 10(6): p. 301-10.

92. Kerr MK, Martin M, Churchill GA. *Analysis of variance for gene expression microarray data.* J Comput Biol, 2000. 7(6): p. 819-37.

93. Xu R, Li X. *A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data.* Bioinformatics, 2003. 19(10): p. 1284-9.

94. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HB, Saxild HH, Nielsen C, Brunak S, Knudsen S. *A new non-linear normalization method for reducing variability in DNA microarray experiments.* Genome Biol, 2002. 3(9): p. research0048.

95. Durbin B, Rocke DM. *Estimation of transformation parameters for microarray data.* Bioinformatics, 2003. 19(11): p. 1360-7.

96. Qian J, Kluger Y, Yu H, Gerstein M. *Identification and correction of spurious spatial correlations in microarray data.* Biotechniques, 2003. 35(1): p. 42-4, 46, 48.

97. Manduchi E, Grant GR, McKenzie SE, Overton GC, Surrey S, Stoeckert CJ, Jr. *Generation of patterns from gene expression data by assigning confidence to differentially expressed genes.* Bioinformatics, 2000. 16(8): p. 685-98.

98. Bevilacqua MP, Pober JS, Wheeler ME, Cotran RS, Gimbrone MA, Jr. *Interleukin-1 activation of vascular endothelium. Effects on procoagulant activity and leukocyte adhesion.* Am J Pathol, 1985. 121(3): p. 394-403.

99. Unger RE, Krump-Konvalinkova V, Peters K, Kirkpatrick CJ. *In vitro expression of the endothelial phenotype: comparative study of primary isolated cells and cell lines, including the novel cell line HPMEC-ST1.6R.* Microvasc Res, 2002. 64(3): p. 384-97.

100. Yoshida M, Westlin WF, Wang N, Ingber DE, Rosenzweig A, Resnick N, Gimbrone MA, Jr. *Leukocyte adhesion to vascular endothelium induces E-selectin linkage to the actin cytoskeleton.* J Cell Biol, 1996. 133(2): p. 445-55.

101. Maruyama I, Soejima Y, Osame M, Ito T, Ogawa K, Yamamoto S, Dittman WA, Saito H. *Increased expression of thrombomodulin on the cultured human umbilical vein endothelial cells and mouse hemangioma cells by cyclic AMP.* Thromb Res, 1991. 61(3): p. 301-10.

102. Morandi V, Cherradi SE, Lambert S, Fauvel-Lafeve F, Legrand YJ, Legrand C. *Proinflammatory cytokines (interleukin-1 beta and tumor necrosis factor-alpha) down regulate synthesis and secretion of thrombospondin by human endothelial cells.* J Cell Physiol, 1994. 160(2): p. 367-77.

103. Zhao B, Stavchansky SA, Bowden RA, Bowman PD. *Effect of interleukin-1beta and tumor necrosis factor-alpha on gene expression in human endothelial cells.* Am J Physiol Cell Physiol, 2003. 284(6): p. C1577-83.

104. Huang J, Teng L, Li L, Liu T, Chen D, Xu LG, Zhai Z, Shu HB. *ZNF216 Is an A20-like and IkappaB kinase gamma-interacting inhibitor of NFkappaB activation.* J Biol Chem, 2004. 279(16): p. 16847-53.

105. Fries E, Kaczmarczyk A. *Inter-alpha-inhibitor, hyaluronan and inflammation.* Acta Biochim Pol, 2003. 50(3): p. 735-42.

106. Tadros A, Hughes DP, Dunmore BJ, Brindle NP. *ABIN-2 protects endothelial cells from death and has a role in the antiapoptotic effect of angiopoietin-1.* Blood, 2003. 102(13): p. 4407-9.

107. Zhang S, Fukushi M, Hashimoto S, Gao C, Huang L, Fukuyo Y, Nakajima T, Amagasa T, Enomoto S, Koike K, Miura O, Yamamoto N, Tsuchida N. *A new ERK2 binding protein, Nafl, attenuates the EGF/ERK2 nuclear signaling.* Biochem Biophys Res Commun, 2002. 297(1): p. 17-23.

108. Staege H, Brauchlin A, Schoedon G, Schaffner A. *Two novel genes FIND and LIND differentially expressed in deactivated and Listeria-infected human macrophages.* Immunogenetics, 2001. 53(2): p. 105-13.

109. Wang J, Shelly L, Miele L, Boykins R, Norcross MA, Guan E. *Human Notch-1 inhibits NF-kappa B activity in the nucleus through a direct interaction involving a novel domain.* J Immunol, 2001. 167(1): p. 289-95.

110. Ando K, Kanazawa S, Tetsuka T, Ohta S, Jiang X, Tada T, Kobayashi M, Matsui N, Okamoto T. *Induction of Notch signaling by tumor necrosis factor in rheumatoid synovial fibroblasts.* Oncogene, 2003. 22(49): p. 7796-803.

111. Pandey A, Shao H, Marks RM, Polverini PJ, Dixit VM. *Role of B61, the ligand for the Eck receptor tyrosine kinase, in TNF-alpha-induced angiogenesis.* Science, 1995. 268(5210): p. 567-9.

112. Witlin AG, Li ZY, Wimalawansa SJ, Grady JJ, Grafe MR, Yallampalli C. *Placental and fetal growth and development in late rat gestation is dependent on adrenomedullin.* Biol Reprod, 2002. 67(3): p. 1025-31.

113. Luttun A, Tjwa M, Carmeliet P. *Placental growth factor (PlGF) and its receptor Flt-1 (VEGFR-1): novel therapeutic targets for angiogenic disorders.* Ann N Y Acad Sci, 2002. 979: p. 80-93.

114. Horrevoets AJ, Fontijn RD, van Zonneveld AJ, de Vries CJ, ten Cate JW, Pannekoek H. *Vascular endothelial genes that are responsive to tumor necrosis factor-alpha in vitro are expressed in atherosclerotic lesions, including inhibitor of apoptosis protein-1, stannin, and two novel genes.* Blood, 1999. 93(10): p. 3418-31.

115. Simha V, Garg A. *Phenotypic heterogeneity in body fat distribution in patients with congenital generalized lipodystrophy caused by mutations in the AGPAT2 or seipin genes.* J Clin Endocrinol Metab, 2003. 88(11): p. 5433-7.

116. Lee MJ, Thangada S, Claffey KP, Ancellin N, Liu CH, Kluk M, Volpi M, Sha'afi RI, Hla T. *Vascular endothelial cell adherens junction assembly and morphogenesis induced by sphingosine-1-phosphate.* Cell, 1999. 99(3): p. 301-12.

117. Bandman O, Coleman RT, Loring JF, Seilhamer JJ, Cocks BG. *Complexity of inflammatory responses in endothelial cells and vascular smooth muscle cells determined by microarray analysis.* Ann N Y Acad Sci, 2002. 975: p. 77-90.

118. Comander J, *Transcriptional and functional modulation of the endothelial cell inflammatory response by a biomechanical stimulus*, in *Program in Biological and Biomedical Sciences*. 2004, Harvard University: Cambridge, MA.

119. Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH. *Amplified RNA synthesized from limited quantities of heterogeneous cDNA.* Proc Natl Acad Sci U S A, 1990. 87(5): p. 1663-7.

120. Schneider J, Buness A, Huber W, Volz J, Kioschis P, Hafner M, Poustka A, Sultmann H. *Systematic analysis of T7 RNA polymerase based in vitro linear RNA amplification for use in microarray experiments.* BMC Genomics, 2004. 5(1): p. 29.

121. Polacek DC, Passerini AG, Shi C, Francesco NM, Manduchi E, Grant GR, Powell S, Bischof H, Winkler H, Stoeckert CJ, Jr., Davies PF. *Fidelity and enhanced sensitivity of differential transcription profiles following linear amplification of nanogram amounts of endothelial mRNA.* Physiol Genomics, 2003. 13(2): p. 147-56.

122. Ji W, Zhou W, Gregg K, Lindpaintner K, Davis S. *A method for gene expression analysis by oligonucleotide arrays from minute biological materials.* Anal Biochem, 2004. 331(2): p. 329-39.

123. Spiess AN, Mueller N, Ivell R. *Amplified RNA degradation in T7-amplification methods results in biased microarray hybridizations.* BMC Genomics, 2003. 4(1): p. 44.

124. Dumur CI, Garrett CT, Archer KJ, Nasim S, Wilkinson DS, Ferreira-Gonzalez A. *Evaluation of a linear amplification method for small samples used on high-density oligonucleotide microarray analysis.* Anal Biochem, 2004. 331(2): p. 314-21.

125. Baugh LR, Hill AA, Brown EL, Hunter CP. *Quantitative analysis of mRNA amplification by in vitro transcription.* Nucleic Acids Res, 2001. 29(5): p. E29.

Acknowledgements

I am deeply indebted to my research advisor, Dr. Guillermo Garcìa-Cardeña, for his unwavering support, guiding insights and constant encouragement. His mentorship has been invaluable and his example inspiring. I also thank Dr. Michael A. Gimbrone for his insightful feedback. I thank Jason Comander for his creativity and enthusiasm during our collaboration on variance estimation algorithms, and for his patience in introducing me to many of the aspects of molecular biology. I am also grateful to lab members Guohao Dai and Jeanne-Marie Kiely for their insight and advice, as well as Yuzhi Zhang, Kush Parmar, Saran Vaughn, Christina Pham, Eric Wang and Keith Anderson for their support. Summer students Nathan Becharzyzk, Christopher Doucette and Justin vanKlein were invaluable in helping to develop software implementing these algorithms, being especially accomodating with my constant requests to tweak the algorithms here and there. I thank Yuzhi Zhang for preparing all the RNA samples and the RT-IVT microarrays, and Kay Case, Vanessa Davis and Deanna LaMont for harvesting human umbilical venous endothelial cells. Finally, we thank Applied Biosystems for their collaborative efforts on gene expression analysis.

Appendix A: Genes Significantly Regulated by 4-hr. IL-1β Exposure in Cultured HUVEC

Color corresponds to p-value indicating statistical significance:

- All
- Down / 5e-5 < p < .05
- Down / 5e-10 < p < 5e-5
- Down / 1e-15 < p < 5e-10
- Down / p < 1e-15
- Up / 5e-5 < p < 0.05
- Up / 5e-10 < p < 5e-5
- Up / 1e-15 < p < 5e-10
- Up / p < 1e-15

Ratio	Gene Name
719.7	chemokine (C-X-C motif) ligand 3
518.1	colony stimulating factor 2 (granulocyte-macrophage)
432.2	selectin E (endothelial adhesion molecule 1)
431.9	chemokine (C-X-C motif) ligand 2
352.4	vascular cell adhesion molecule 1
291.5	chemokine (C-X3-C motif) ligand 1
240.3	chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)
172.3	TNF receptor-associated factor 1
138.7	ubiquitin D
107.8	chemokine (C-C motif) ligand 2
100.6	chemokine (C-C motif) ligand 20
79.4	baculoviral IAP repeat-containing 3
70.8	CD69 antigen (p60, early T-cell activation antigen)
67.1	thymic stromal lymphopoietin
64.0	intercellular adhesion molecule 1 (CD54), human rhinovirus receptor
58.9	interleukin 8
51.4	colony stimulating factor 3 (granulocyte)
47.5	tumor necrosis factor, alpha-induced protein 6
45.5	tumor necrosis factor, alpha-induced protein 2
43.2	CCAAT/enhancer binding protein (C/EBP), delta
31.0	<no annotation>
29.4	chemokine (C-X-C motif) ligand 6 (granulocyte chemotactic protein 2)
28.3	tumor necrosis factor, alpha-induced protein 3
27.5	Rho family GTPase 1
25.7	leukocyte receptor cluster (LRC) member 9
24.9	likely ortholog of rat SNF1/AMP-activated protein kinase
23.8	interleukin 6 (interferon, beta 2)
23.2	tumor necrosis factor receptor superfamily, member 9
20.4	nuclear receptor subfamily 4, group A, member 3
20.3	hypothetical protein FLJ23231
17.7	coagulation factor III (thromboplastin, tissue factor)

Ratio	Gene Name
16.4	solute carrier family 7 (cationic amino acid transporter, y+ system), member 2
15.9	inducible T-cell co-stimulator ligand
13.9	<no annotation>
13.7	S100 calcium binding protein A3
13.6	CD83 antigen (activated B lymphocytes, immunoglobulin superfamily)
13.5	<no annotation>
12.5	molecule possessing ankyrin repeats induced by lipopolysaccharide (MAIL), homolog of mouse
12.3	chemokine orphan receptor 1
12.2	tumor necrosis factor receptor superfamily, member 11b (osteoprotegerin)
11.8	TNFAIP3 interacting protein 3
11.7	superoxide dismutase 2, mitochondrial
11.2	chromosome 6 open reading frame 128
11.2	<no annotation>
11.1	nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (p49/p100)
11.0	TRAF2 binding protein
10.1	colony stimulating factor 1 (macrophage)
9.5	<no annotation>
9.4	v-rel reticuloendotheliosis viral oncogene homolog B, nuclear factor of kappa light polypeptide gene enhancer in B-cells 3 (a
9.3	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha
9.2	activating transcription factor 3
9.1	<no annotation>
8.8	apolipoprotein L, 3
8.4	<no annotation>
8.2	human immunodeficiency virus type I enhancer binding protein 2
8.2	jun B proto-oncogene
8.1	tenascin C (hexabrachion)
8.0	hypothetical protein MGC52057
7.9	undifferentiated embryonic cell transcription factor 1
7.8	mitogen-activated protein kinase kinase kinase 8
7.7	leukemia inhibitory factor (cholinergic differentiation factor)
7.7	interleukin-1 receptor-associated kinase 2
7.2	interleukin 18 receptor 1
7.2	nuclear receptor coactivator 7
7.1	interleukin 15
6.8	histone deacetylase 9
6.8	<no annotation>
6.7	potassium intermediate/small conductance calcium-activated channel, subfamily N, member 2
6.6	receptor-interacting serine-threonine kinase 2
6.4	interferon regulatory factor 1
6.3	tumor necrosis factor, alpha-induced protein 8

Ratio	Gene Name
6.2	NK3 transcription factor related, locus 1 (Drosophila)
6.2	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)
5.9	<no annotation>
5.7	<no annotation>
5.6	TNFAIP3 interacting protein 1
5.3	<no annotation>
5.3	msh homeo box homolog 1 (Drosophila)
5.2	ATP-binding cassette, sub-family G (WHITE), member 1
5.2	antigen identified by monoclonal antibody MRC OX-2
5.2	<no annotation>
5.1	serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 2
4.9	tumor necrosis factor (ligand) superfamily, member 18
4.8	BCL2-related protein A1
4.8	talin 2
4.7	Down syndrome critical region gene 1
4.7	follistatin-like 3 (secreted glycoprotein)
4.7	syndecan 4 (amphiglycan, ryudocan)
4.6	carbonyl reductase 3
4.6	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (p105)
4.5	interleukin 7 receptor
4.5	hepatocellular carcinoma-associated antigen 66
4.5	<no annotation>
4.4	solute carrier family 2 (facilitated glucose transporter), member 6
4.4	hypothetical protein DKFZp434K0427
4.3	v-ets erythroblastosis virus E26 oncogene homolog 1 (avian)
4.3	ephrin-A1
4.3	<no annotation>
4.3	sterile alpha motif domain containing 4
4.1	chromosome 8 open reading frame 4
4.1	<no annotation>
3.9	transporter 1, ATP-binding cassette, sub-family B (MDR/TAP)
3.9	syntaxin 11
3.8	solute carrier family 12 (potassium/chloride transporters), member 7
3.8	<no annotation>
3.8	<no annotation>

Ratio	Gene Name
3.8	<no annotation>
3.8	sequestosome 1
3.7	v-maf musculoaponeurotic fibrosarcoma oncogene homolog F (avian)
3.7	plasminogen activator, urokinase
3.6	solute carrier family 12 (sodium/potassium/chloride transporters), member 2
3.6	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, epsilon
3.5	immediate early response 3
3.4	tumor necrosis factor (ligand) superfamily, member 15
3.3	<no annotation>
3.3	SRY (sex determining region Y)-box 7
3.3	DNA-damage-inducible transcript 4
3.3	apolipoprotein L, 2
3.2	glutamine-fructose-6-phosphate transaminase 2
3.2	leucine-rich repeats and immunoglobulin-like domains 1
3.1	interleukin 4 induced 1
3.1	a disintegrin-like and metalloprotease (reprolysin type) with thrombospondin type 1 motif, 9
3.1	tumor necrosis factor, alpha-induced protein 1 (endothelial)
3.1	<no annotation>
3.1	agouti signaling protein, nonagouti homolog (mouse)
3.0	neuronal pentraxin I
3.0	metallothionein 1F (functional)
3.0	<no annotation>
3.0	sialyltransferase 1 (beta-galactoside alpha-2,6-sialyltransferase)
2.9	interferon stimulated gene 20kDa
2.9	B-cell CLL/lymphoma 6, member B (zinc finger protein)
2.9	peroxisomal proliferator-activated receptor A interacting complex 285
2.9	cytoplasmic linker 2
2.8	FOS-like antigen 2
2.8	interferon gamma receptor 1
2.8	neuron navigator 2
2.8	<no annotation>
2.8	membrane associated guanylate kinase interacting protein-like 1
2.8	UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 1
2.8	<no annotation>
2.8	laminin, gamma 2
2.7	collagen triple helix repeat containing 1
2.7	colony stimulating factor 2 receptor, beta, low-affinity (granulocyte-macrophage)
2.7	signal transducer and activator of transcription 5A
2.7	phosphoprotein regulated by mitogenic pathways
2.7	coagulation factor II (thrombin) receptor-like 1

Ratio	Gene Name
2.7	<no annotation>
2.7	UDP-glucose ceramide glucosyltransferase
2.7	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide
2.7	slingshot 1
2.7	solute carrier family 31 (copper transporters), member 2
2.7	hypothetical protein FLJ23375
2.7	immediate early response 5
2.7	tripartite motif-containing 47
2.6	cathepsin S
2.6	dual specificity phosphatase 16
2.6	Ras and Rab interactor 2
2.6	natural killer cell transcript 4
2.6	metallothionein 1B (functional)
2.6	elongation factor, RNA polymerase II, 2
2.6	jagged 1 (Alagille syndrome)
2.6	calcium-binding transporter
2.6	<no annotation>
2.6	cathepsin K (pycnodysostosis)
2.6	<no annotation>
2.6	<no annotation>
2.6	phorbol-12-myristate-13-acetate-induced protein 1
2.5	<no annotation>
2.5	metallothionein IV
2.5	metallothionein 1A (functional)|metallothionein 1K|metallothionein 1E (functional)|metallothionein 2A
2.5	suppression of tumorigenicity 5
2.5	TRAF family member-associated NFKB activator
2.5	ninjurin 1
2.5	zinc finger protein 36, C3H type, homolog (mouse)
2.5	interferon gamma receptor 2 (interferon gamma transducer 1)
2.5	SEC14-like 2 (S. cerevisiae)
2.5	<no annotation>
2.5	<no annotation>
2.5	metallothionein 1X
2.5	CDC14 cell division cycle 14 homolog A (S. cerevisiae)
2.5	uridine phosphorylase 1
2.5	ephrin-B1
2.5	DnaJ (Hsp40) homolog, subfamily B, member 9
2.4	<no annotation>
2.4	nicotinamide N-methyltransferase
2.4	pannexin 1
2.4	v-rel reticuloendotheliosis viral oncogene homolog (avian)
2.4	C-type lectin-like receptor-1
2.4	tumor necrosis factor (ligand) superfamily, member 10
2.4	heat shock 27kDa protein 8
2.4	optineurin
2.4	TNF receptor-associated factor 3
2.4	MAD, mothers against decapentaplegic homolog 3 (Drosophila)

Ratio	Gene Name
2.4	zinc fingers and homeoboxes 2
2.4	interleukin 15 receptor, alpha
2.3	junctional adhesion molecule 2
2.3	hypothetical protein FLJ10276
2.3	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3
2.3	chromosome 6 open reading frame 197
2.3	opioid growth factor receptor-like 1
2.3	bone morphogenetic protein 2
2.3	protein tyrosine phosphatase, receptor type, K
2.2	guanylate binding protein 1, interferon-inducible, 67kDa
2.2	salvador homolog 1 (Drosophila)
2.2	hypothetical protein FLJ90440
2.2	PDZ and LIM domain 4
2.2	pentaxin-related gene, rapidly induced by IL-1 beta
2.2	TIR domain containing adaptor inducing interferon-beta
2.2	<no annotation>
2.2	hypothetical protein FLJ90005
2.2	B-cell CLL/lymphoma 6 (zinc finger protein 51)
2.2	<no annotation>
2.1	MO25 protein
2.1	transducin-like enhancer of split 3 (E(sp1) homolog, Drosophila)
2.1	START domain containing 10
2.1	CDC42 effector protein (Rho GTPase binding) 2
2.1	pleckstrin homology-like domain, family A, member 1
2.1	Kruppel-like factor 3 (basic)
2.1	phospholipase A2, group IVC (cytosolic, calcium-independent)
2.1	cylindromatosis (turban tumor syndrome)
2.1	solute carrier family 7, (cationic amino acid transporter, y+ system) member 11
2.1	<no annotation>
2.1	endothelial cell-specific molecule 1
2.1	matrix metalloproteinase 10 (stromelysin 2)
2.1	endothelial differentiation, sphingolipid G-protein-coupled receptor, 1
2.1	tubulin, beta polypeptide
2.0	selenoprotein SelM
2.0	tripartite motif-containing 56
2.0	KIAA1404 protein
2.0	<no annotation>

Ratio	Gene Name
2.0	serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 8
2.0	a disintegrin-like and metalloprotease (reprolysin type) with thrombospondin type 1 motif, 4
2.0	tyrosylprotein sulfotransferase 1
2.0	SRY (sex determining region Y)-box 1
2.0	regulator of G-protein signalling 3
2.0	zinc finger protein, multitype 2
2.0	WD repeat endosomal protein
2.0	ubiquitous tetratricopeptide containing protein RoXaN
1.9	formin binding protein 1
1.9	LIM domain kinase 2
1.9	DKFZP586N0721 protein
1.9	<no annotation>
1.9	CASP8 and FADD-like apoptosis regulator
1.9	eukaryotic translation initiation factor 2C, 2
1.9	mitogen-activated protein kinase kinase 3
1.9	ras homolog gene family, member B
1.9	X-box binding protein 1
1.9	G protein-coupled receptor 56
1.9	<no annotation>
1.9	hypothetical protein FLJ22344
1.9	DnaJ (Hsp40) homolog, subfamily A, member 1
1.9	<no annotation>
1.9	mitogen-activated protein kinase kinase kinase 7 interacting protein 2
1.9	mitogen-activated protein kinase kinase 3
1.9	nucleolar protein 1, 120kDa
1.9	guanylate binding protein 4
1.9	v-ets erythroblastosis virus E26 oncogene homolog 2 (avian)
1.9	Ras association (RalGDS/AF-6) domain family 1
1.9	monocyte to macrophage differentiation-associated
1.9	poliovirus receptor
1.8	<no annotation>
1.8	roundabout, axon guidance receptor, homolog 1 (Drosophila)
1.8	<no annotation>
1.8	stannin
1.8	tumor necrosis factor receptor superfamily, member 10b

Ratio	Gene Name
1.8	guanine nucleotide exchange factor for Rap1
1.8	TNFAIP3 interacting protein 2
1.8	phosphatidylinositol transfer protein, cytoplasmic 1
1.8	<no annotation>
1.8	<no annotation>
1.8	hypothetical protein MGC17791
1.8	GPP34-related protein
1.8	<no annotation>
1.8	cyclin-dependent kinase inhibitor 1A (p21, Cip1)
1.8	<no annotation>
1.8	melanoma differentiation associated protein-5
1.8	hypothetical protein FLJ12484
1.8	activated leukocyte cell adhesion molecule
1.8	transforming growth factor, beta receptor II (70/80kDa)
1.7	regulator of G-protein signalling 2, 24kDa
1.7	component of oligomeric golgi complex 3
1.7	myelin protein zero-like 1|hypothetical protein FLJ21047
1.7	signal transducer and activator of transcription 6, interleukin-4 induced
1.7	BCL2-related ovarian killer
1.7	tryptophanyl-tRNA synthetase
1.7	cysteine and glycine-rich protein 2
1.7	hypothetical protein MGC10986
1.7	pleckstrin homology domain containing, family C (with FERM domain) member 1
1.7	EH-domain containing 1
1.7	hepatocellular carcinoma related protein 1

Ratio	Gene Name
1.7	H2.0-like homeo box 1 (Drosophila)
1.7	promyelocytic leukemia
1.7	CD47 antigen (Rh-related antigen, integrin-associated signal transducer)
1.7	mitochondrial folate transporter/carrier
1.7	phosphatidic acid phosphatase type 2B
1.7	armadillo repeat protein ALEX2
1.6	baculoviral IAP repeat-containing 2
1.6	UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 5
1.6	cysteine-rich hydrophobic domain 2
1.6	dishevelled associated activator of morphogenesis 1
1.6	pleckstrin homology, Sec7 and coiled-coil domains 1(cytohesin 1)
1.6	Wilms tumor 1 associated protein
1.6	transcription factor binding to IGHM enhancer 3
1.6	<no annotation>
1.6	mitogen-activated protein kinase kinase 1

Ratio	Gene Name
1.6	**spermidine/spermine N1-acetyltransferase**

Ratio	Gene Name
1.5	<no annotation>
1.4	tubulin, alpha 3
1.4	<no annotation>
-4.0	thioredoxin interacting protein
-3.0	DnaJ (Hsp40) homolog, subfamily B, member 4
-3.0	thrombomodulin
-2.9	Kruppel-like factor 2 (lung)
-2.9	papillomavirus regulatory factor PRF-1|hypothetical protein DKFZp434K1210
-2.8	core-binding factor, runt domain, alpha subunit 2; translocated to, 1; cyclin D-related
-2.8	adrenomedullin
-2.8	gap junction protein, alpha 4, 37kDa (connexin 37)
-2.8	SRY (sex determining region Y)-box 18
-2.6	prickle-like 1 (Drosophila)

Ratio	Gene Name
-2.6	G protein-coupled receptor 126
-2.5	protein kinase, cAMP-dependent, catalytic, beta
-2.5	C1q domain containing 1
-2.5	hypothetical protein LOC51063
-2.4	<no annotation>
-2.4	Meis1, myeloid ecotropic viral integration site 1 homolog (mouse)
	[illegible] protein CG30021
-2.4	Rho GTPase activating protein 18
-2.4	EH-domain containing 3
-2.4	lymphoblastic leukemia derived sequence 1
-2.3	palmdelphin
-2.3	mesenchymal stem cell protein DSCD75
	[illegible] nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 1
-2.3	hypothetical protein FLJ20674
-2.3	chemokine-like factor super family 8
	[illegible]
	v-maf [illegible] musculoaponeurotic fibrosarcoma oncogene homolog B (avian)
-2.2	<no annotation>
-2.1	dachshund homolog (Drosophila)
	[illegible]
-2.1	calcitonin receptor-like
-2.1	<no annotation>
	[illegible]
-2.0	v-yes-1 Yamaguchi sarcoma viral related oncogene homolog
	[illegible] CUB and LCCL domain containing 1
	[illegible] protein MGC13024
-2.0	scavenger receptor class F, member 2
-2.0	homeo box A10
	[illegible] homolog 8 (Drosophila)
	LIM [illegible] domain protein (SRF accessory protein 2)
-1.9	LIM domain only 4
-1.9	early hematopoietic zinc finger
	[illegible] receptor B

Ratio	Gene Name
	[illegible]
-1.9	**interferon regulatory factor 2 binding protein 2**
	[illegible]
	[illegible]
	[illegible]
	[illegible]
	[illegible]
	[illegible]
	[illegible]
	[illegible]
	[illegible]
-1.8	**<no annotation>**
	[illegible]
-1.8	**immune associated nucleotide**
	[illegible]
	[illegible]
	[illegible]
-1.8	**<no annotation>**
	[illegible]
	[illegible]
	[illegible]
	[illegible]
	[illegible]
	[illegible]
	[illegible]
	[illegible]
	[illegible]

Ratio	Gene Name
	[illegible] (Opitz/BBB syndrome)
	[illegible] 4 open reading frame 132
	[illegible] solute carrier family 25 (phosphate transporter), member [illegible]
	[illegible] ant gen 2
	[illegible] neocortex (with BTB-like domain)
	[illegible] transmembrane fibronectin domain containing 12
	[illegible] synthetic protein LOC152687;hypothetical protein FLJ90036
	[illegible] mutation
	[illegible] mediator and activator of transcription 1, 91kD
-1.6	**connective tissue growth factor**
	[illegible] kinase receptor

Ratio	Gene Name
	[illegible] chromosome 14 open reading frame 37
	[illegible] homolog (Xenopus laevis)
	[illegible]

Ratio	Gene Name
	[illegible] signalling 5