BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 21 2007, pages 2873–2880
doi:10.1093/bioinformatics/btm450

Gene expression

Systematic order-dependent effect in expression values, variance, detection calls and differential expression in Affymetrix GeneChips®

Kathe E. Bjork[1,2,]* and Karen Kafadar[1]

[1]Department of Mathematical Sciences, University of Colorado at Denver and Health Sciences Center, Denver, CO 80217 and [2]PriMetrics Inc., Arvada, CO 80007, USA

Received on June 17, 2007; revised on August 24, 2007; accepted on August 25, 2007

Advance Access publication September 25, 2007

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Affymetrix GeneChips® are common 3′ profiling platforms for quantifying gene expression. Using publicly available datasets of expression profiles from human and mouse experiments, we sought to characterize features of GeneChip® data to better compare and evaluate analyses for differential expression, regulation and clustering. We uncovered an unexpected order dependence in expression data that holds across a variety of chips in both human and mouse data.

Results: Order dependence among GeneChips® affected relative expression measures pre-processed and normalized with the Affymetrix MAS5.0 algorithm and the robust multi-array average summarization method. The effect strongly influenced detection calls and tests for differential expression and can potentially significantly bias experimental results based on GeneChip® profiling.

Contact: Kathe.bjork@cudenver.edu

Supplementary information: Supplementary Material, including links to files of ordered transcripts and supporting analyses, is available at the authors' websites, at http://www-math.cudenver.edu/~kbjork/research/, and http://www-math.cudenver.edu/~kk/research/.

1 INTRODUCTION

Affymetrix GeneChips® are common 3′ profiling platforms for quantifying gene expression. These chips are dense microarrays of single-stranded 25-base oligonucleotides synthesized *in situ* for hybridization to single-stranded mRNA-derived complementary RNA (cRNA) from target tissues, capable of quantifying relative expression of tens of thousands of genes simultaneously. The current version of the human GeneChip®, the HG-U133 Plus 2.0 array, contains 54 675 probe sets for querying the full complement of known human mRNA transcripts and variants and a set of known Affymetrix controls via 1.3 million distinct oligonucleotide features (Affymetrix, 2004). Each gene or control is represented on the HG-U133 Plus 2.0 array by 11–20 oligonucleotide segments selected for uniqueness, with each segment tiled into probe pairs of perfect matched (PM) and mismatched (MM) oligonucleotides. Each PM strand hybridizes to its complement in target cRNA, while MM strands, with a central base switch (at number 13) to destabilize and repulse binding, serve as a vague control. To reduce potential spatial effects, Affymetrix has distributed probe pairs throughout the chip. Some investigators have suggested that MM values are not exact controls, and have developed probe summarization methods that disregard them.

Systematic perturbations arising from chip design or manufacturing, sample processing or instrumentation can be non-trivial in such massive datasets. As we demonstrate herein, data from two current and three historical GeneChip® platforms contain a systematic order-dependent pattern that manifests as a major determinant of PM and MM values, with the resulting effect carried forward into estimates of relative expression and variance, detection calls and significance tests for differential expression. Neither of two of the most commonly employed algorithms for summarizing PM and MM values detects, reports or adjusts for this effect. Experiments conducted without adjustment for the effect are likely to have inflated type I and II error rates.

2 APPROACH

Using publicly available datasets of Affymetrix GeneChip® 3′ expression profiles from human and mouse experiments, we began this study with the intention of trying to characterize the features of gene expression data, to better model such data and thereby compare and evaluate the performance of various detection algorithms for differential expression, regulation and clustering. During this investigation we uncovered an unexpected order dependence in GeneChip® expression data that appears to hold, sometimes to a significant extent, across a variety of chips. We describe order-dependent patterns in multiple chip types, demonstrate their significant influence in expression analysis, inference and estimation, and introduce a potential method for correction.

*To whom correspondence should be addressed.

Table 1. Data sources, number of files, GeneChip® type and example file identifiers[1]

Data source	Number of files	Type of chip	Example filename
Affymetrix Demonstration Data	4	HG-U133A	HG-U133A-1-121502
	8	HG-U95Av2	1521l99hpp-av06r
	1	Mouse430A	Mouse430A_031903
	4	MG-U74Av2	MG-U74Av2-1-121502
Harvard CardioGenomics	46	HG-U133 Plus 2.0	PA-N_300
NCBI-GEO GSE3325	8	HG-U133 Plus 2.0	GSM74489

[1]See Supplementary Material for a complete list of data sources, filenames and links to original data.

3 METHODS

GeneChip® data from hybridization of cRNA targets to probes are measured as relative levels of emission in 64 (8 × 8) pixels in each PM or MM tile. The outer perimeter of pixels is discarded and the 75th percentile of the central 36 (6 × 6) pixels is recorded as the probe pair PM or MM value. These values are collected into .CEL files with accompanying probe location information.

Various statistical methods have been designed and implemented to summarize probe pair data in .CEL files into relative quantitative expression measures. We used pre-processing software from the Bioconductor project (http://www.bioconductor.org, version 1.8), an open source repository for routines for probe- and higher-level genomic analyses. The 'affy' package (Gautier *et al*., 2003) implements routines for several summarization algorithms, including two commonly used, Affymetrix' Microarray Suite Statistical Algorithm version 5.0 (MAS5.0) signal value (SV) (Affymetrix, 2002) and the robust multiarray average (RMA) (Irizarry *et al*., 2003). SV summarization proceeds chipwise with a zone-dependent background correction and normalization, winsorization, computation of an 'ideal' match for probe pairs with MM > PM, and a robust procedure, Tukey's one-step biweight, to reduce sensitivity to outlier probe pairs (Affymetrix, 2002). In our analyses, SVs were $\log_2$-transformed after SV expression computation. Developers of the 'affy' package for computing SVs caution that their MAS5.0 routine ('mas5()') was compiled from descriptions in published corporate literature and not from code obtained directly from the company, and hence results may not exactly replicate those obtained with the company's proprietary software (Gautier *et al*., 2003). Some pre-processed SVs and 'detection calls' (of present versus absent transcripts, see below) were available online with corresponding .CEL files. If SVs were not available with the .CEL files, we used 'affy' package software to compute them. In all analyses with SVs, we note the package used to compute them.

RMA values are multi-chip model-based relative expressions based solely on PM values. RMA summarizes a collection of arrays via background correction, log-transformation and normalization, and gene-by-gene processing with robust median polish. Developers recommend the default quantile normalization for RMAs, and, unless noted otherwise, we accepted this default in our analyses (Bolstad *et al*., 2003). All .CEL files were processed with the 'rma()' routine from the 'affy' package to obtain RMA values.

Publicly available human and mouse GeneChip® data were obtained from Affymetrix corporate, Harvard CardioGenomics Project (CardioGenomics PGA, 2007) and National Center for Biotechnology Information Gene Expression Omnibus (NCBI-GEO) website repositories (Table 1). Affymetrix Inc. shares .CEL file data with researchers and developers for algorithm development. Their 'Demonstration Data' website (http://www.affymetrix.com/support/technical/sample_data/demo_data.affx) contains data files of GeneChip® experiments from several organisms. We used experimental data from four human HG-U133A, eight HG-U95Av2, one Mouse430A and four mouse MG-U74Av2 chips in this analysis.

The NCBI-GEO (http://www.ncbi.nlm.nih.gov) is a gene expression and molecular abundance repository supporting MIAME compliant data submissions and is a curated, online resource for gene expression data browsing, query and retrieval. Eight files were accessed from an investigation of prostate cancer progression, GEO Accession GSE3325 (www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc = GSE3325), four each of benign and metastatic tumor profiles. We used .CEL files, SVs and detection calls contributed by the primary investigators (Varambally *et al*., 2005) in our analyses.

The CardioGenomics Project (http://cardiogenomics.med.harvard.edu/) shares data on genomic investigations of human and animal models of cardiovascular disease. The project goal is to explore gene linkages with functional, dysfunctional and structural abnormalities of the cardiovascular system caused by clinically relevant genetic and environmental stimuli. Forty-six HG-U133 Plus 2.0 cardiac tissue .CEL files were selected from collections (14 control 'clinically normal' and 32 ischemic profiles). Both .CEL files and SVs computed by the CardioGenomics project were used in these analyses. To control for potential confounding by array lot, we used 11 normal and 12 ischemic tissue profiles with the same probe array lot number reported in quality control data, and confirmed our results with the complete set of 46 arrays.

Typically, PM and MM values are recorded in .CEL files in an ordered sequence of Affymetrix gene identifiers, a series of 3–8 numbers, 1–2 underscores, and 1–3 letters. Transcript order reporting varies according to pre-processing routine. Bioconductor software for SVs and RMA expressions records transcripts in alphanumerically ascending sequence beginning with the leftmost character and proceeding through the rightmost character. Among mRNA transcripts the leftmost characters are usually digits; in Affymetrix controls it is an alpha prefix 'AFFX'. In the chips we examined Bioconductor software placed AFFX controls in the terminal 62–67 rows of data (Table 2). We accepted Bioconductor ordering in all of our analyses. For ease of display, transcripts and controls in .CEL, SV and RMA files were assigned a numeric index according to sequence. In the HG-U133 Plus 2.0 chip, transcripts were indexed 1:54 675 starting with Affymetrix probe identifier 1007_at and ending with AFFX-TrpnX-M_at. Patterns of order dependence were demonstrated by plotting gene expression values according to index. We demonstrate order-dependence in the current human HG-U133 Plus 2.0 and mouse MOE430A chips as well as earlier generation chips HG-U133A, HG-U95Av2 and MG-U74Av2.

To investigate potential biological explanations for the order dependence, 40 genes were selected from regions of HG-U133 Plus 2.0 chips, 20 each on either side of the estimated point where clear pattern shifts occur. Chromosomal location, gene name, tissue type and common biochemical pathways were recorded for these genes and evaluated for potential commonalities.

The MAS5.0 statistical algorithm computes a 'present', 'marginal' or 'absent' detection call and an associated detection *P*-value for each transcript via statistical analysis of probe pairs (Affymetrix, 2002). The detection call algorithm defines as absent any expression level below a threshold of detection, i.e. 'not provably different from 0'. Detection calls are determined for each probe set by removing saturated probe pairs or those with minimal differences, calculating a discrimination score for each probe pair; and testing for significance of the score. Genes with detection $p < 0.04$ are called 'present'. 'Marginal' detections are those genes with $0.04 < p < 0.06$, and genes with $p > 0.06$ are reported 'absent'. Investigators may use these detection calls and

Table 2. Fifteen initial and 15 terminal sequences of Affymetrix identifiers in HG-U133 Plus 2.0 and HG-U133A GeneChips®[1]

HG-U133 Plus 2.0		HG-U133A	
1:15	54 661:54 675[2]	1:15	22 269:22 283[2]
1007_at	-r2-Ec-bioB-3_at	1007_at	-r2-Ec-bioD-5_at
1053_at	-r2-Ec-bioB-5_at	1053_at	-r2-Hs18SrRNA-3_s_at
117_at	-r2-Ec-bioB-M_at	117_at	-r2-Hs18SrRNA-5_at
121_at	-r2-Ec-bioC-3_at	121_at	-r2-Hs18SrRNA-M_x_at
1255_g_at	-r2-Ec-bioC-5_at	1255_g_at	-r2-Hs28SrRNA-3_at
1294_at	-r2-Ec-bioD-3_at	1294_at	-r2-Hs28rRNA-5_at
1316_at	-r2-Ec-bioD-5_at	1316_at	-r2-Hs28rRNA-M_at
1320_at	-r2-P1-cre-3_at	1320_at	-r2-P1-cre-3_at
1405_i_at	-r2-P1-cre-5_at	1405_i_at	-r2-P1-cre-5_at
1431_at	-ThrX-3_at	1431_at	-ThrX-3_at
1438_at	-ThrX-5_at	1438_at	-ThrX-5_at
1487_at	-ThrX-M_at	1487_at	-ThrX-M_at
1494_f_at	-TrpnX-3_at	1494_f_at	-TrpnX-3_at
1552256_a_at	-TrpnX-5_at	1598_g_at	-TrpnX-5_at
1552257_at	-TrpnX-M_at	160020_at	-TrpnX-M_at

[1] See Supplementary Material for links to files with complete lists of ordered probeset identifiers (Affymetrix IDs) for HG-U133 Plus 2.0, HG-U133A, HG-U95Av2, Mouse430A and MG-U74Av2 GeneChips®.
[2] Affymetrix control identifiers in terminal sequences all begin with 'AFFX-'.

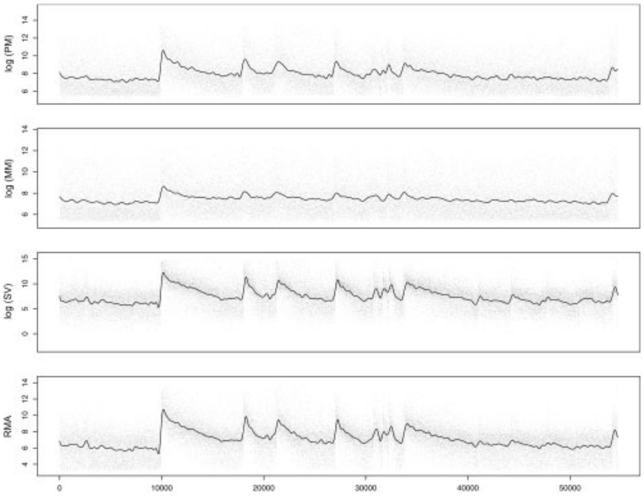

Fig. 1. Log(PM) (top), log(MM), log(SV) and RMA (bottom) expression values by index, HG-U133Plus2.0 GeneChip®, sample PA-N_300, CardioGenomics project data. All probe level summarizations were computed with 'affy' package, BioConductor version 1.8 software.

P-values as selection criteria for consideration of genes for further analysis. We used detection calls and SVs from prostate cancer study profiles (NCBI-GEO Accession GSE3325) in assessing order dependence effects on detection calls.

All statistical analyses, data processing and plots of pre-processed and normalized data were performed using 'R' project software (www.r-project.org, version 2.4.1). Determinations of block breakpoints were estimated by visually assessing regions with obvious shifts in ordered plots of log(SV)s and RMAs in multiple tissue and sample profiles. Estimated breakpoints were checked with plots of running means and medians.

Order dependence in expression value variance according to pre-processing method was evaluated using cardiac tissue profiling with HG-U133 Plus 2.0 chips from the CardioGenomics Project. MAS5.0 SVs were downloaded from the project website and compiled, normalized with median-based methods, and log-transformed with National Cancer Institute's Biometrics Research Branch Array Tools software version 3.5 (Simon and Lam, 2006). RMA expressions were computed using Bioconductor routines. Expression value variance and SD were evaluated according to index and relationship with mean expression values. Results were replicated via an identical analysis using prostate cancer profiles with HG-U133 Plus 2.0 chips obtained from the NCBI-GEO. (See Supplementary Material for further results of analyses of prostate cancer expression.)

To explore the effect of order dependence on differences between group means and on basic measures of differential expression, we computed the between-group mean difference for each transcript, i.e. *mean*(*normal*)−*mean*(*ischemic*) and two-sample significance tests with log(SV)s and RMA expressions from CardioGenomics data. Conventional (normal-based) *t*-tests (applying both a Welch *t*-test allowing for unequal variance and a pooled *t*-test assuming equal variances) and Wilcoxon's rank sum non-parametric test were calculated for each transcript to assess the significance of the effect. For data reduction and pattern detection by index, we applied a linear smoother to mean differences, test statistics and *P*-values, first by computing running means of length 7 to remove fine noise and then computing a smoothing spline for gross patterns. Mean differences, test statistics and associated *P*-values were plotted by index.

Once order dependence is detected or suspected, chip(s) may require correction at the probe pair (PM and MM) or probe-summary level for valid inference. Using CardioGenomics data, we adjusted each individual array's RMA and median-based normalized log(SV) expressions with a smoothing spline by subtracting the smoothed fitted value from the observed value and replacing the observed value with the residual. Plots of ordered residuals, variance and SD of the residuals were evaluated for ability to remove the systematic effect. Residuals were also used in Welch *t*-tests and Wilcoxon rank sum tests to evaluate diminution of order dependence via smoothing.

4 RESULTS

A systematic order-dependent effect was evident and pronounced in indexed plots of log(PM), log(MM), log(SV) and RMA values from publicly available datasets of five human and mouse GeneChip® types, and within those chip types transcript order was a major determinant of relative expression value and variance. Figure 1 shows the complex order-dependent patterns in HG-U133 Plus 2.0 expressions in one sample, PA-N_300, from the CardioGenomics project data. This sample is typical of all HG-U133 Plus 2.0 chip data that we examined. Log(SV)s ranged approximately −2 to 15 units with some truncation in the higher values. RMAs ranged approximately 3–14 units, with distributions truncated in the lower levels in the first block, and in the higher-indexed values in the other blocks. Chips were characterized by 14 grossly visible variably sized blocks, with block intersections at indices approximately 9937, 18 028, 21 230, 27 000, 30 744, 31 663, 32 255, 33 814, 40 933, 44 210, 50 592, 54 148, and 54 618. Some blocks were very large, such as the first block, and some blocks were quite small, such as those around index 30 000. Blocks with higher indices, such as those greater than 40 000, tended to have less discernable intersections.

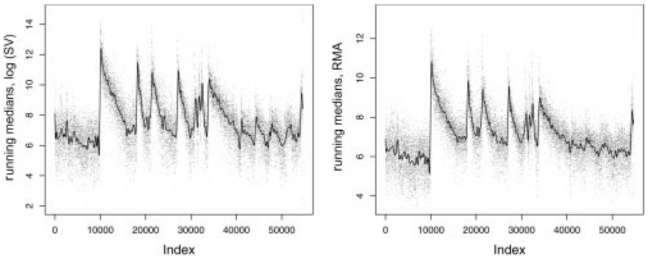

Fig. 2. Running medians of length 7, log(SV)s (left) and RMA expressions (right) by index, HG-U133Plus2.0 GeneChip® sample PA-N_300, CardioGenomics project data.

A subterminal block of approximately 465 genes had uniformly greater expression values among both RMAs and SVs, and the 62 Affymetrix controls in the terminal location extended the full range of expression value. These patterns can be more easily discerned in plots of running medians of length 7 (Fig. 2), wherein fine noise has been reduced. These plots indicate that the 14 blocks may be comprised of smaller sub-blocks.

For both summarization methods, genes with the highest expression levels within blocks tended to be those with the lowest indices, and high-index genes tended to have expressions roughly 2–4 units smaller. This disparity is most obvious between the first two blocks; the mean expression of high-index block 1 genes is similar to the low-expressed (perhaps non-expressed) block 2 low-index genes. This relationship at block junctures does not hold across all blocks, and there are some blocks, especially in the highest level indices (>40 000), that appear U-shaped or uniformly elevated.

No clear biological or tissue-related explanation for pattern shifts could be found in comparisons of the groups of genes surrounding the first four and the final block intersections when examined by chromosome number or location, gene name or function. However, two intersecting regions, blocks 1:2 and blocks 13:14 (Table 3), showed major changes in Affymetrix identifier numberings.

Profiles from the earlier versions of the human GeneChip® HG-U133A obtained from Affymetrix Demonstration Data exhibited similar systematic pattern shifts (Fig. 3). The patterns were less complex, with four major peaked blocks, a fifth U-shaped block, and a terminal collection of Affymetrix controls spanning the full breadth of expression. Patterns were evident in log(PM), log(MM), log(SV) and RMA expression values. Log(SV)s and RMA expressions in three other chip types all exhibited a high degree of order dependence in highly variable and complex patterns (Fig. 4). Indexed HG-U95Av2 data exhibited a systematic oscillatory pattern more prominent in log(SV)s, and mouse chips, Mouse430A and MG-U74Av2, contained mixtures of narrow and wide oscillations. Overall, within-chip block patterns were identical from disparate data and tissue sources using the same chip type. Block patterns, lengths, discernible breaks and shifts differed between different chip types and species.

The impact of order dependence is evident in detection calls and their associated P-values. Plots of HG-U133 Plus 2.0 log(SV)s, detection calls and P-values from one sample of the prostate cancer data (Fig. 5) show that the highest expression levels are directly correlated with the lowest P-values. Detection P-values associated with expression peaks are lower than those of the surrounding genes, and are therefore more likely to be called ‘present’. Closer inspection of this relationship among genes indexed 5000:15 000 (Fig. 6), at the intersection of blocks 1:2, shows that at index approximately 9937, the distribution of P-values abruptly changes from one more uniformly discrete, to one with a sparsity of genes with large P-values. Given their lower P-values, genes indexed 9937 to approximately 14 000 would preferentially be selected for further analysis.

A cursory examination of the relationship between transcript means and log(SD)s of log(SV)s and RMA expressions, respectively, (Fig. 7) of the eight arrays of the prostate cancer dataset shows the relationships between probe summary measures and variance. To better illustrate the trends within

Table 3. Block, index, Affymetrix identifier and transcript information for genes at three block intersections, HG-U133 Plus 2.0 GeneChip®

Block intersection	index	Affymetrix identifier	Gene symbol	Chromosomal location
1:2	9935	1570639_at	CASC4	15q15.3
	9936	1570644_at	NA[1]	NA
	9937	1570645_at	NA	NA
	9938	1570650_at	CCBL1	9q34.11
	9939	1570651_at	CCBL1	9q34.11
	9940	1570653_at	NA	NA
	9941	1598_g_at	GAS6	13q34
	9942	160020_at	MMP14	14q11-q12
	9943	1729_at	TRADD	16q22
	9944	1773_at	FNTB	14q23-q24
	9945	177_at	PLD1	3q26
	9946	179_at	PMS2L11	7q11.23
	9947	1861_at	BAD	11q13.1
	9948	200000_s_at	PRPF8	17p13.3
2:3	18 024	208605_s_at	NTRK1	1q21-q22
	18 025	208606_s_at	WNT4	1p36.23-p35.1
	18 026	208607_s_at	SAA1, SAA2	11p15.1
	18 027	208608_s_at	SNTB1	8q23-q24
	18 028	208609_s_at	TNXA,TNXB	6p21.3
	18 029	208610_s_at	SRRM2	16p13.3
	18 030	208611_s_at	SPTAN1	9q33-q34
	18 031	208612_at	PDIA3	15q15
	18 032	208613_s_at	FLNB	3p14.3
13:14	54 141	244890_at	SLC22A6	11q13.1-q13.2
	54 142	244891_s_at	PVT1	8q24
	54 143	266_s_at	CD24	6q21
	54 144	31637_s_at	THRA, NR1D1	17q11.2
	54 145	31799_at	COPB2	3q23
	54 146	31807_at	DDX49	19p12
	54 147	31826_at	FKBP15	9q32
	54 148	31835_at	HRG	3q27
	54 149	31837_at	TMEM153	22q13.33
	54 150	318 45_at	ELF4	Xq26

[1]NA = not available.

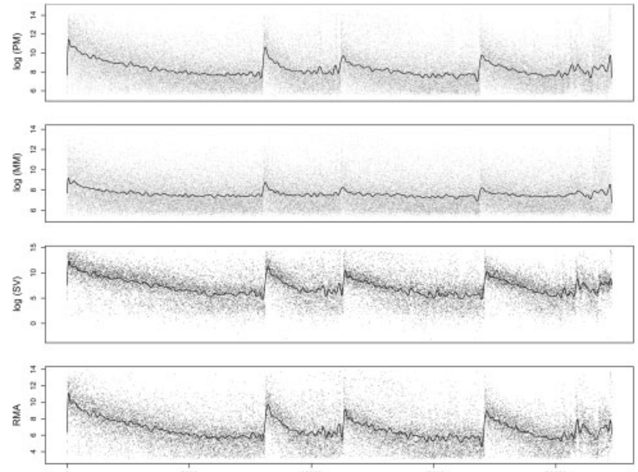

Fig. 3. Log(PM) (top), log(MM), log(SV) and RMA (bottom) expression values by index, HG-U133A-1-121502, Affymetrix Demonstration Data.

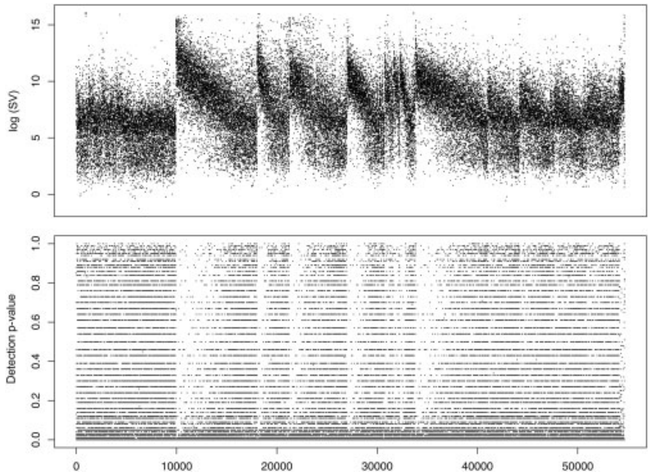

Fig. 5. Log(SV) (top) and detection call *P*-values (bottom) by index, HG-U133Plus2.0 GeneChip®, sample GSM74489, NCBI-GEO Accession GSE3325 prostate cancer dataset.

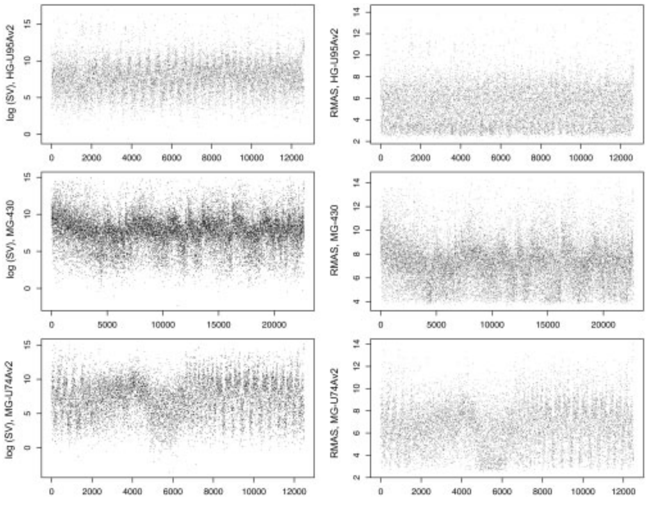

Fig. 4. Log(SV) (left) and RMA (right) expression values by index, human HG-U95Av2 (top), and mouse Mouse430A (middle) and MG-U74Av2 (bottom) GeneChips®, Affymetrix Demonstration Data.

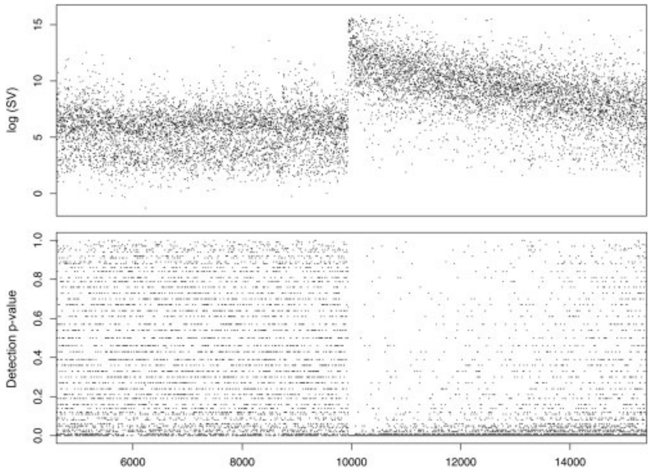

Fig. 6. Log(SV) (top) and detection call *P*-values (bottom), indices 5000-15 000, HG-U133Plus2.0 GeneChip® sample GSM74489, NCBI-GEO prostate cancer dataset.

the blocks, highlighted by vertical segments, the values have been smoothed by a simple running means smoother (length 7) followed by application of a smoothing spline, indicated by the solid line. For both RMA and log(SV), the averaged gene expression values tend to *decline* with order number within each block. However, the trends in log(SD)s are reversed for the two types of expression summarizations: the SDs *decrease* with *increasing* average log(SV)s, but SDs of the RMA values *increase* with *increasing* average RMA value (see Supplementary Material for an expanded analysis of prostate cancer data).

The order dependence effect carried forward into between-group mean differences and significance tests for differential expression (Figs 8 and 9). Smoothed plots of mean differences by comparison group, and Welch *t*- and Wilcoxon rank sum statistics and associated *P*-values show characteristic peaks and troughs by index and block, corresponding with block breakpoints. These figures show that the 'block effect' is dampened compared with that in Figure 1, but is not eliminated by the calculated *t*- or Wilcoxon statistic, because the effect appears to remain in the mean difference (numerator of the *t*-statistic) as well as in the SDs. No difference was detected by type of *t*-test variance, as both pooled-variance and Welch *t*-tests produced similar results. A review and exploration of published quality control data relative to group status of the CardioGenomics data produced no explanation for the effects (see Supplementary Material for abstracted quality control data related to the CardioGenomics project).

Application of a smoothing spline to expression values and replacing observed values with residuals from the fit was

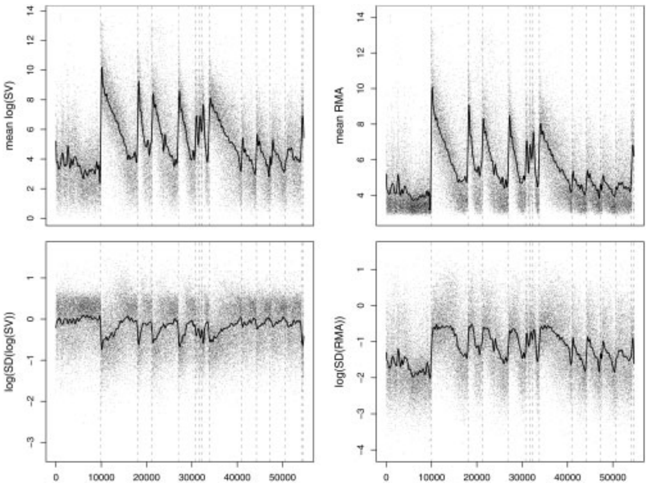

Fig. 7. Means of log(SV)s and RMA expressions (top panels), and log(SD)s of log(SV)s and RMAs (lower panels), by index, HG-U133Plus2.0 GeneChips®, NCBI-GEO prostate cancer data.

successful in zeroing the mean expression level baseline (Fig. 10, upper left). However, transcript variance was not stabilized among residuals (Fig. 10, upper right), as shown by the plots of the smoothed log(SD)s of log(SV)s and RMA expressions. Consequently, significance tests using the residuals were not substantially improved (Fig. 10, bottom panels), as smoothed plots of *P*-values from both *t*-tests and Wilcoxon tests still exhibited within- and between-block effects.

5 DISCUSSION

We observed an order dependence in relative gene expression values calculated from publicly available data collected using five types of Affymetrix GeneChips®, including two types currently in use in human and mouse expression profiling. Order dependence patterns were very similar across disparate tissue types and laboratories. There was no obvious biological basis for the observed patterns, suggesting that the dependence did not arise from specific chromosomal, tissue or other target sources.

Order dependence is characterized by baseline shifts and trends both within and between blocks and can be visualized by plotting transcripts in the order reported from the Affymetrix GeneChip® analysis system (Affymetrix, 2002) and Bioconductor (Gautier *et al*., 2003) routines. In the five types of chips we examined (HG-U133 Plus 2.0, HG-U133A, HG-U95Av2, Mouse430A and MG-U74Av2), the number and morphology of blocks varied widely by chip type. Absolute block boundaries were often difficult to distinguish. The majority of blocks began with high relative expression levels, underwent linear or quadratic downward drifts and were often terminated at a juncture with the next block. At such junctures, mean relative expressions could shift by up to four units on the $\log_2$ scale. Some blocks' expression values had U-shaped distributions or contained oscillating baselines (as in the HG-U95Av2 chip), and some chips contained a mixture of blocks with abrupt shifts, drifts and oscillations. These patterns were very similar in expression levels computed and normalized with two of the most common expression value summarization methods, the MAS5.0 Statistical Algorithm (SV) (Affymetrix, 2002) and the robust multi-chip average RMA (Irizarry *et al*., 2003).

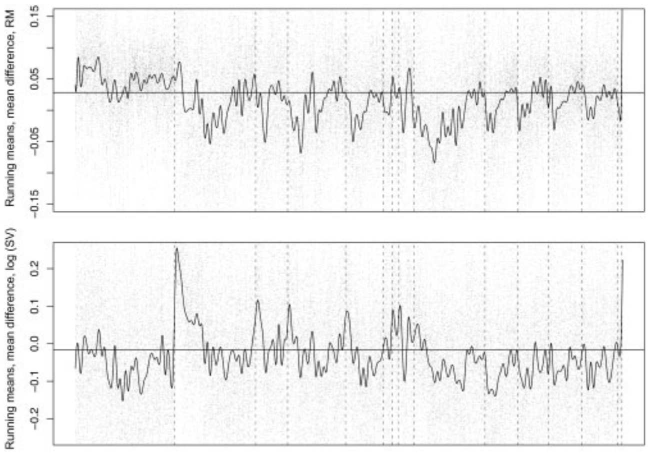

Fig. 8. Mean differences of RMA values (top) and log(SV)s (bottom), by index, HG-U133Plus2.0 GeneChips®, CardioGenomics data. Dots are running means (length 7) of the between-group mean difference (*mean*(*normal*)–*mean*(*ischemic*)); solid horizontal line is the overall median of the running means; wavy solid line is a smoothing spline applied to the running means and dashed vertical lines indicate estimated blocks.

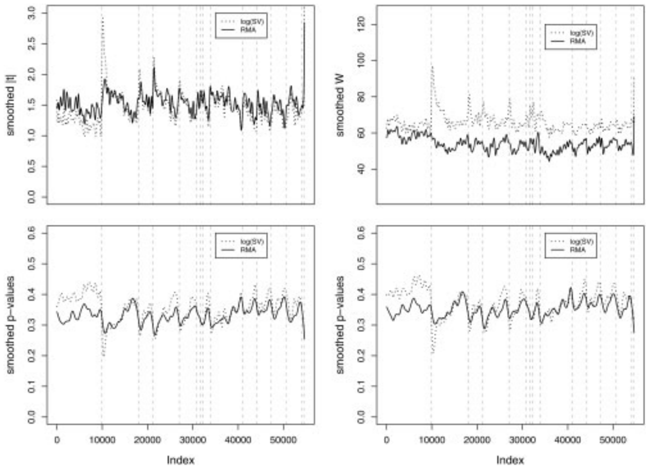

Fig. 9. Test statistics (upper panel) and *P*-values (bottom panel) from significance testing via *t*-test (left panel) and Wilcoxon rank sum test (right panel), smoothed by index, control versus ischemic cardiac tissues, HG-U133Plus2.0 GeneChips®, CardioGenomics project data.

Variance of GeneChip® expression profiles also exhibited order dependence, as shown in plots of mean expression values and logarithms of SDs (log(SD)s) relative to index. Smoothed log(SD)s had baseline shifts and within-and between-block variability similar to that noted in mean expression levels. However, whereas log(SD)s and means of RMA expressions were positively associated, log(SD)s and means of log(SV)s had an inverse relationship. These results differed from those reported by the developers of the RMA expression (Irizarry *et al*., 2003), who based their model on the HG-U95Av2 chip.

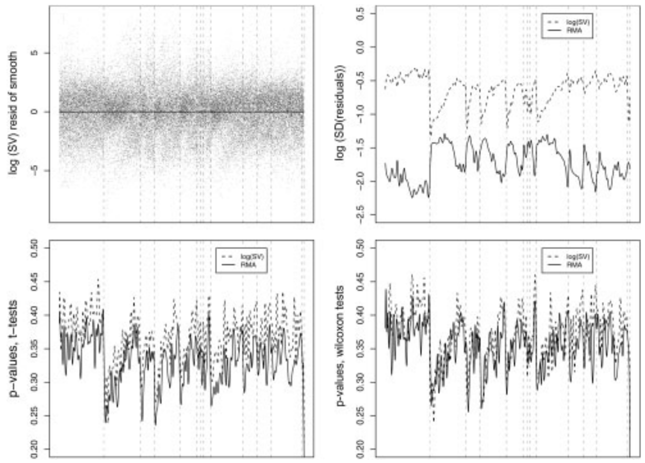

Fig. 10. Residuals of smoothed log(SV)s of one control tissue profile, by index (upper left); log(SD) of residuals from smoothed log(SV)s and RMA values, smoothed by index, (upper right); *P*-values from *t*-tests (lower left) and Wilcoxon rank sum tests (lower right), by index, normal versus ischemic cardiac tissues, HG-U133Plus2.0 GeneChips®, CardioGenomics project data.

In the RMA development report, SDs increased with average expression level for all summarization methods examined. Because our analyses were based on these same summarization methods, it is not clear why the relationship between the mean and log(SD) among transcripts processed with MAS5.0 SV algorithm did not hold in the newer generation GeneChip®. Studies are underway to evaluate this relationship in other GeneChip® varieties and summarization methods, and to explore appropriate models and potential transformations for expression value and variance adjustment (see also Supplementary Material for results using two different summarization methods on NCBI-GEO Accession GSE3325 prostate cancer study data).

When detection call *P*-values are used to select transcripts for further study, order dependence may bias the selection process. We observed in multiple tissue types that transcripts with the highest expression values tended to occur among lower indexed genes within blocks and that these genes had the lowest detection *P*-values. This was especially apparent at the intersection of the first two blocks in the HG-U133 Plus 2.0 chip.

A second potential bias due to order dependence was demonstrated in our examination of the differences between group means and in differential expression studies. In plots of between-group mean differences and two-sample tests comparing control and ischemic cardiac tissues, we observed that, in most blocks,the largest (absolute) magnitude mean differences and test statistics and the smallest *P*-values were strongly associated with spikes in expression values occurring among the lowest block indices. These effects occurred with both log(SV)s and RMAs and were duplicated with 46 arrays from the CardioGenomics project. The same analysis with eight prostate cancer tissue profiles from the NCBI-GEO database (see Supplementary Material) produced similar results. This suggests that the effect is not localized to specific datasets or laboratories. The application of median-based normalization methods using the BRB ArrayTools software did not appear to mitigate the effect. Although it seems plausible that the order-dependence effect should cancel out in significance tests between 'target' and 'control' groups, it did not. We found no explanation for this in available quality control data. We also explored potential misalignment of blocks as a possible explanation; however, block shifts appeared to occur at the same breakpoints in the data. These results may be a spurious finding in these particular datasets and may not be replicable in other laboratories or under different experimental conditions. More detailed investigations are warranted to clarify these findings.

The full implications of the order dependence effect on differential expression remains to be seen. Because the expression levels are elevated for genes that are indexed at the beginning of a block, for all arrays in an experiment, *P*-values likewise tend to indicate 'significance' even in the absence of any real effect. Consequently, multiple testing procedures such as the method of false discovery rate (Benjamini and Hochberg, 1995) that focus on genes with the smallest *P*-values will likewise preferentially select those genes that are indexed at the beginning of blocks.

The cause of the order-dependence effect was not readily discernable from examinations of publicly available data. Because details of GeneChip® design and manufacturing specifications, processes and instrumentation are proprietary, we were unable to relate our findings to particular characteristics of the system, or to undertake any meaningful directed diagnostics such as those that can be performed with cDNA microarrays generated in-house (Balazsi *et al.*, 2003). Lacking this proprietary production-specific information, these findings remain descriptive of symptoms and not etiology. This severely limits clarification of the source of the bias and potential avenues for remediation. Furthermore, because these effects varied across GeneChip® versions and generations, comprehensive diagnostics, etiologies and chip-specific remediation may require access to and review of historical system-wide manufacturing information and specifications. The determination of causes for these trends will require collaboration among biologists, R&D and processing engineers, data monitors and statisticians.

One proposed method for bias correction is the *post hoc* adjustment of computed expression measures with localized non-parametric regression, the application of a smoothing spline and removal of fitted, or expected, expression level and replacement with the residual from the regression. Although our smoothing spline method was effective in stabilizing the baseline of expression values and removing the majority of order-dependent effects from expression values, it was not satisfactory in stabilizing the variance and the influence of the effect on differential expression. Further work is required to understand and remediate systematic chip effects on the mean difference and variance.

Other methods of adjustment are currently under investigation, including multiple-array *z*-score methods for stabilizing variance and individual array 'change point detection' methods such as those used in industrial process control monitoring (Vardeman and Jobe, 1999) to discern between and within-block trends, for detecting both abrupt and gradual up- and downward shifts in series of data. If subsets can be identified and transcripts correctly classified into blocks, the subsets can be better characterized with improved parameterizations and analytical methods selected to produce more reliable and valid summary, detection and differential expression measures. In addition, studies of the effects of order dependence on sensitivity, specificity and synchronization of GeneChip® data with other platforms are underway.

While numerous corrective methods may be applied, and likely successfully, a more fruitful approach to solving order-dependence related problems will be to identify and remediate the source(s) for the varied within- and between-block effects. Current pre-processing and normalization procedures for GeneChip® analyses may be adequate for assuring common levels and ranges of gene expression values *across* arrays, but may be inadequate for normalizing out the order effects noted *within* arrays.

The impact of order dependence on Affymetrix expression profiling experiments is likely to extend beyond transcript measurements and investigations of differential expression. Results from the prostate cancer progression study (Varambally *et al.*, 2005) suggest that controlling or adjusting for order dependence in transcript profiling holds promise for improving concordance between measurements of mRNA transcripts and translated peptides and proteins. In the prostate cancer study, investigators estimated 61% concordance in gene:protein expression for clinically localized prostate cancer relative to benign prostate tissue, and 48% concordance for metastatic prostate cancer relative to clinically localized disease. Once the impact of order dependence in 3′ profiling experiments with Affymetrix GeneChips® is more fully characterized and explained, it is possible that improvements in concordance of gene:protein expression may follow.

We found no mention of order dependence or its consequences in any corporate literature or in references of commonly used summarizations for probe-pair data computed from .CEL files, including Affymetrix' MAS5.0 algorithm, their newer probe summarization method 'probe logarithmic intensity error estimation' (PLIER) (Affymetrix, 2007), RMA, or the Li–Wong model-based expression index (MBEI, (Li and Wong, 2001)), although it is likely some of these summary measures were developed using older generations of chips, (such as the HuGeneFl chip noted in the MBEI method), when greater levels of noise perhaps obscured the order dependence. Summarization methods for probe level data continue to be an active area of research for GeneChips®, and up to 50 different probe summary methods have been proposed (Chen *et al.*, 2007). One difficulty we encountered in comparing different summary measures was differing output transcript order by summarization designer. We suggest that, if order dependence is proven to be an important factor in Affymetrix GeneChip® analyses, all summarization methods be output in a standardized order for ease of comparison (such as in Supplementary Material). We also recommend that all current and previous generations of GeneChips®, across all organisms, be evaluated for order dependence and its potential biases, and that these investigations be conducted in a variety of laboratories, under different experimental conditions, and using the full spectrum of 3′ expression profiling GeneChips® to fully assess the occurrence and importance of the effect.

6 CONCLUSION

This report of a significant within-array order-dependent effect in multiple types of Affymetrix 3′ expression profiling GeneChips® demonstrates that GeneChip®-based experiments likely require both within- and between-chip transformation and normalization to produce valid inference, and that further investigations are warranted to better understand the etiology and impacts of the effect. If order dependence is proven as a significant source of bias in the Affymetrix 3′ expression profiling system, appropriate technical and statistical methods for remediation should be developed and applied. Lacking a full understanding of the source of the effect, investigators will be restricted to addressing symptoms with chip- and species-specific statistical patches. Once order dependence is better understood and adjustments applied, the benefits may be far-reaching. Historical datasets can be re-analyzed and may reveal previously obscured information, including improved inference from differential expression and concordance with protein expression studies.

ACKNOWLEDGEMENTS

We thank Tzulip Phang and Sanjiv Bhave, UCDHSC, for their contributions and clarifications of GeneChip® system processes, and Richard Jones, UCDHSC, for helpful discussions. Research was supported in part by the Army Research Office, Grant Number 48391-MA, awarded to the Department of Mathematical Sciences, UCDHSC (K.K).

Conflict of Interest: none declared.

REFERENCES

Affymetrix Inc. (2002) Statistical Algorithms Dscription Document.

Affymetrix Inc. (2004) GeneChip®Human Genome Arrays Data Sheet.

Affymetrix Inc. (2007) Guide to probe logarithmic intensity error (PLIER) estimation.

Balazsi,G. *et al.* (2003) Spurious spatial periodicity of co-expression in microarray data due to printing design. *Nucleic Acids Res.*, **31**, 4425–4433.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Bolstad,B.M. *et al.* (2003) a comparison of normalization methods for high-density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Chen,Z. *et al.* (2007) A distribution free summarization method for Affymetrix GeneChip®arrays. *Bioinformatics*, **23**, 321–327.

Gautier,L. *et al.* (2003) affy–analysis of Affymetrix GeneChip®data at the probe level. *Bioinformatics*, **20**, 307–315.

Genomics of cardiovascular development, adaptation, and remodeling. (2007) NHLBI Program for Genomic Applications. Harvard Medical, School URL: http://www.cardiogenomics.org.

Irizarry,R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Li,C. and Wong,W. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Simon,R. and Lam,A.P. (2006) BRB-ArrayTools Version 3.5 User's Manual National Cancer Institute Biometrics Research Branch.

Varambally,S. *et al.* (2005) Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell*, **8**, 393–406.

Vardeman,S.B. and Jobe,J.M. (1999) *Statistical Quality Assurance Methods for Engineers*. Wiley, New York.