Genomics 100 (2012) 337–344

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

A single-sample microarray normalization method to facilitate personalized-medicine workflows

Stephen R. Piccolo [a,b], Ying Sun [c], Joshua D. Campbell [b], Marc E. Lenburg [b], Andrea H. Bild [a,c,*,1], W. Evan Johnson [b,c,**,1]

[a] Department of Pharmacology and Toxicology, University of Utah, 201 Presidents Circle, Salt Lake City, UT 84112, USA
[b] Division of Computational Biomedicine, Boston University School of Medicine, 72 East Concord Street, Boston, MA 02118, USA
[c] Huntsman Cancer Institute, 2000 Circle of Hope Drive, Salt Lake City, UT 84112-5550, USA

ARTICLE INFO

Article history:
Received 19 April 2012
Accepted 14 August 2012
Available online 19 August 2012

Keywords:
Method
Normalization
Microarray
Linear model
Mixture model
Single-sample technique

ABSTRACT

Gene-expression microarrays allow researchers to characterize biological phenomena in a high-throughput fashion but are subject to technological biases and inevitable variabilities that arise during sample collection and processing. Normalization techniques aim to correct such biases. Most existing methods require multiple samples to be processed in aggregate; consequently, each sample's output is influenced by other samples processed jointly. However, in personalized-medicine workflows, samples may arrive serially, so renormalizing all samples upon each new arrival would be impractical. We have developed Single Channel Array Normalization (SCAN), a single-sample technique that models the effects of probe-nucleotide composition on fluorescence intensity and corrects for such effects, dramatically increasing the signal-to-noise ratio within individual samples while decreasing variation across samples. In various benchmark comparisons, we show that SCAN performs as well as or better than competing methods yet has no dependence on external reference samples and can be applied to any single-channel microarray platform.

1. Introduction

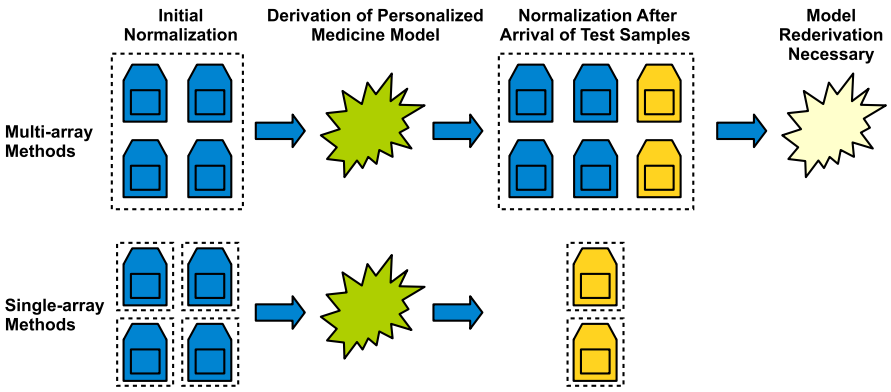
Gene-expression microarrays enable high-throughput characterizations of biological phenomena. They are used widely in the investigation of biological and clinical phenotypes and have begun to be disseminated in clinical settings [1–3] via "personalized-medicine" workflows. For example, expression profiles representing prognostic subclasses of breast cancer are being used to guide treatment decisions and have been shown to influence medical-oncologist treatment recommendations [4]. Developing such a workflow involves two distinct steps: 1) identification of expression profiles that reliably characterize subclasses and 2) comparison of new breast-cancer samples against those profiles to predict each sample's subclass. However, a critical preliminary step is to normalize each microarray sample to adjust for technological biases and non-biological variabilities that arise during sample collection and processing and that obfuscate true signal. Traditional normalization methods require groups of samples to be processed jointly—thus each time a new sample arrived, all samples would need to be renormalized. Clearly, this type of approach presents serious logistical and reproducibility challenges and obviates the need for methods that process samples individually. To address this need, we developed Single Channel Array Normalization (SCAN), a normalization technique that uses data observed within a given microarray sample to correct for probe-nucleotide biases and to produce standardized output values that remain constant irrespective of any other sample. Furthermore, unlike competing methods, SCAN requires no ancillary reference samples to perform these corrections and can be applied to any one-color microarray.

Many normalization methods–including the popular Robust Multi-chip Average (RMA) [5] and its probe-sequence-adjusting counterpart, GeneChip RMA (GC-RMA) [6]–borrow information from other samples to estimate probe-level effects and to standardize variances across arrays. However, with such approaches, output values for any given array depend on the quantity and quality of other arrays that are processed jointly (see Fig. 1A). In an attempt to address such limitations, two normalization techniques capable of processing individual samples have gained popularity. The MAS5 algorithm [7], developed by Affymetrix, uses mismatch probes with single-base differences to correct for background variation. As such, this method attempts to minimize probe-level effects by observing such effects in probes with nearly identical sequences. Despite the relative simplicity of its algorithmic approach, MAS5 has performed well against multi-array methods [8]; however, this algorithm can only be applied to platforms containing mismatch probes, which excludes the newer Affymetrix

* Correspondence to: A.H. Bild, Department of Pharmacology and Toxicology, University of Utah, 201 Presidents Circle, Salt Lake City, UT 84112, USA. Fax: +1 801 581 5111.
** Correnspondence to: W.E. Johnson, Division of Computational Biomedicine, Boston University School of Medicine, 72 East Concord Street, Boston, MA 02118, USA. Fax: +1 617 414 6947.
E-mail addresses: andreab@genetics.utah.edu (A.H. Bild), wej@bu.edu (W.E. Johnson).
[1] These authors contributed equally to the work.

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:http://dx.doi.org/10.1016/j.ygeno.2012.08.003

A



B

	Multi-array	Single-array		
Affymetrix Array Version	RMA	fRMA	MAS5	SCAN
HG-U95Av2	X		X	X
HG-U133A	X	X	X	X
HG-U133B	X		X	X
HG-U133A_2	X		X	X
HG-U133 Plus 2	X	X	X	X
HT_HG-U133A	X		X	X
HG Gene 1.0 ST	X			X
HG Exon 1.0 ST	X			X

Fig. 1. Multi-array versus single-array normalization. A) With multi-array methods, such as RMA, samples are processed in groups. Thus when new samples have been hybridized–for example, in personalized-medicine settings–all samples, old and new, may need to be renormalized as a group, which may require reanalysis of the data or recalibration of biomarkers. Contrarily, with single-array methods, including SCAN, each sample is normalized individually. Thus newly arrived samples remain separate during normalization, and data values for existing samples do not change. B) Affymetrix offers many different array versions to quantify human gene expression. SCAN can normalize any version. However, fRMA does not currently support most array versions because an inadequate number and diversity of previously hybridized samples have been made available publicly. And because MAS5 relies on mismatch probes, it is unable to normalize samples from newer array versions.

Gene ST, Exon ST, and GG-H arrays (see Fig. 1B). Recently, McCall et al. introduced another single-sample approach, frozen RMA (fRMA) [9], which borrows information from publicly available arrays of the same platform. A key prerequisite to applying fRMA is that a reference database be derived from a large set of biologically diverse samples from a wide variety of studies. For the array platforms currently supported by fRMA, approximately 6000 "well-annotated" samples (from which 850 were randomly selected) were used to derive each reference database. However, obtaining such a large and diverse sample set is currently impossible for many platforms, and this approach requires arbitrary decisions of which samples to consider and a deliberate effort to construct a reference set for each unique platform.

To overcome these limitations, SCAN estimates the effect of probe-nucleotide composition on fluorescence intensity through a linear statistical model and distinguishes between background noise and biological signal with a mixture-modeling approach. Probe-level values are standardized against the variance observed for other probes on the same array that have similar binding-affinity-adjusted intensities. In this light, our approach is more appropriately deemed a standardization–rather than the more commonly termed normalization–approach; however, to be consistent with the current terminology in the field, we use both terms interchangeably here. Because SCAN processes a given array only using information intrinsic to that array, normalized values are independent of any other sample that may have been hybridized in the same batch. Also importantly, to process a large number of arrays, computer memory must only be adequate to process a single array at a time, making SCAN a versatile approach that can be used on low-memory desktop computers or compute nodes, even when existing methods fail to process the set of arrays simultaneously. However, in the case where a large number of processors and memory are available, SCAN can be used to process multiple arrays in parallel.

In this study, we demonstrate that SCAN corrects for probe-level effects associated with GC bias and drastically increases the signal-to-noise ratio within each array. We show that SCAN compares favorably against RMA, fRMA, and MAS5 in its ability to detect differences in expression across a wide spectrum of RNA spike-in concentrations. Using data from the Connectivity Map (CMAP) [10], we demonstrate SCAN's methodological advantages in dealing with batch effects [11] and differences between array versions. Strikingly, we show that SCAN enables effective concurrent use of independent gene-expression datasets. In particular, in an analysis of two independent microarray compendia that profile identical cell lines–Broad-Novartis Cancer Cell Line Encyclopedia (CCLE) and GSK Cancer Cell Line Genomic Profiling Data–we show that SCAN produces highly consistent values for samples grown, processed, and hybridized by independent groups at separate times. Finally, in a pathway-based analysis, we show that SCAN-based predictions are highly concordant across the data sets which are known to contain batch effects.

2. Results

2.1. General description of SCAN standardization approach

The most common approaches to normalizing and summarizing microarray data simultaneously process a series of samples, causing measurements for each sample to be dependent on other samples in the group. This approach can lead to significant batch effects and can create discrepant results between datasets that are normalized separately, thus creating problems for applications in personalized medicine. Furthermore, processing large datasets may require excessive amounts of physical memory or splitting large datasets in order to successfully complete the normalization process. SCAN has been

developed to resolve these limitations, and is highly effective at standardizing array data to control for background and individual array signal.

SCAN utilizes a modification of the MAT standardization approach developed for Affymetrix tiling arrays [12]. The MAT method models expected probe behavior based on the individual probe sequence and other probes on the array with similar nucleotide composition and removes effects due to base-pair content and array bias. Kapur et al. successfully applied the MAT model to Affymetrix Exon ST arrays [13], but in their approach parameters were estimated using all data from each array. This assumption is reasonable for tiling arrays where the number of probes measuring signal is small with respect to the number of probes measuring background signal, but is not appropriate for RNA-based expression array experiments due to the high percentage of probes that measure true signal. Therefore, our novel modeling procedure is based on a mixture of two Gaussian distributions with MAT-like models for the means of both distributions—one mixture component measuring background noise and the other measuring biological signal (plus background). The component of interest is the background distribution, which is used by SCAN to standardize the probe-level data by subtracting out the background mean and then standardizing the variance based on the estimated variance of other probes with similar expected background. As illustrated below, this approach is highly effective at removing GC effects and other probe and array-specific variation from each sample individually while leaving the biological signal intact, thus increasing the signal-to-noise ratio within the array while reducing technical variation between arrays.

2.2. SCAN corrects for GC bias inherent to one-color arrays

SCAN accounts for GC bias using a refined linear statistical model that includes effects for the counts of each nucleotide type separately, the nucleotide count squared, and also the location of each nucleotide type on the probe. Using publicly available CEL files downloaded from Gene Expression Omnibus (GEO) [14], we have tested SCAN on a wide variety of Affymetrix platforms, including those that are widely used–including Human Genome U133A, Human Genome U133 Plus 2.0, and Human Exon 1.0 ST–as well as those that have been released recently or that have been less utilized. Upon analyzing these samples, we observed a strong GC bias prior to normalization—probes associated with higher GC content tended to have higher raw-intensity values (see Fig. 2A). We assessed for GC bias after normalization and observed that SCAN abrogated this bias (see Fig. 2B). This result illustrates that SCAN can correct for binding-affinity biases using only data from within a given sample.

2.3. Overall microarray signal-to-noise explained by SCAN model

A unique advantage of SCAN is that it directly identifies systematic within-array variation that can be explained as probe effects and removes this variation from the data. For probes measuring both signal and noise, this method only removes noise and leaves the signal intact. When applied to a set of 100 Affymetrix Human Exon ST 1.0 arrays (GSE25219), SCAN estimated that an average of 79% of probes are only measuring background noise and exhibit no signal above background. In these arrays, we estimate that an average of 35% of the total squared variation in the array data is attributable to background signal. In all, the probe-mixture model accounts for an average of 79% of the variation in the data, and 44% of this explained variation is attributable to background noise (probe effects) and thus removed. Therefore, we estimate the average overall signal-to-noise ratio to be 1.4 in the raw data; this is improved to 6.5 after removing probe-specific noise, resulting in a relative increase of 482% in the overall signal-to-noise ratio. Table 1 summarizes these results and gives ranges for these values across all arrays in the dataset we examined. Therefore these results show that SCAN is highly effective at removing background variation and magnifying signal within each array.

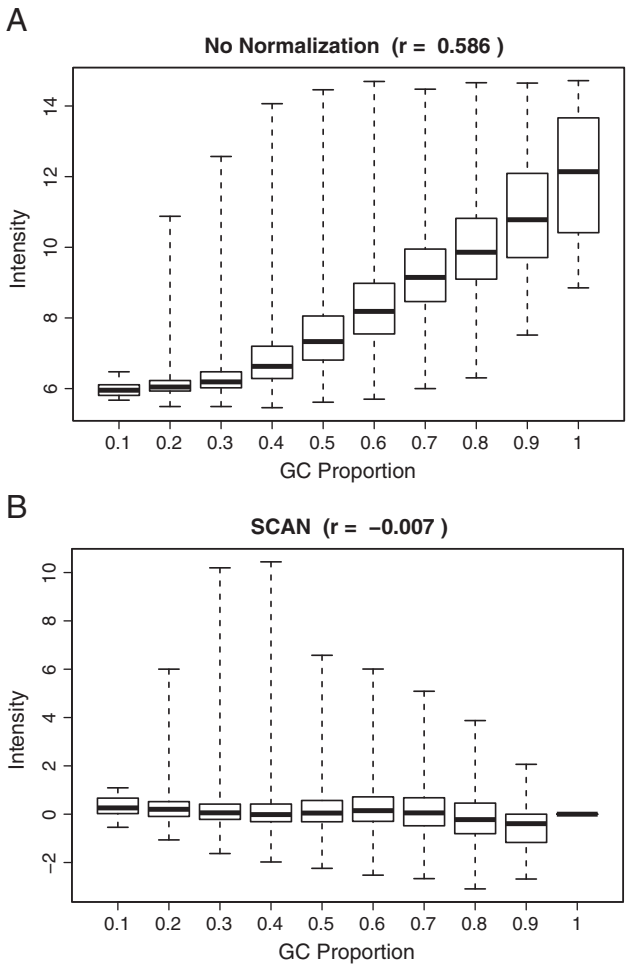


Fig. 2. Proportion GC content versus probe intensity. A) For an Affymetrix Human Exon ST 1.0 array (GSE25219), this figure illustrates the relationship between probe-level GC content and expression intensities. As GC content increases, raw intensity also tends to increase. B) After SCAN processing, this bias is removed.

2.4. SCAN detects transcription levels with high sensitivity and specificity

Next we assessed SCAN's ability to detect known transcription levels in the Affymetrix Human Genome U133 Latin Square data (http://www.affymetrix.com/support/technical/sample_data/datasets.affx), which is designed to support comparisons of analysis methods. This data set

Table 1
Summary of probe-level metrics for SCAN normalization. SCAN was applied to 100 Affymetrix Human Exon 1.0 ST arrays from publicly available data representing spatio-temporal transcription patterns in the human brain (GSE25219) [26]. This table lists various probe-level metrics that characterize SCAN's ability to account for variation and to magnify signal in the data.

Metric	Average	Minimum	Maximum
Proportion of active probes in sample	0.208	0.154	0.263
Proportion of total variation explained by model	0.792	0.757	0.814
Proportion of variation attributed to background noise	0.439	0.280	0.634
Estimated signal-to-noise ratio before application of SCAN	1.357	0.578	2.574
Estimated signal-to-noise ratio after SCAN	6.554	2.632	13.13
Percent increase in signal-to-noise ratio after SCAN	482%	412%	538%

contains 14 distinct concentrations of spiked transcripts, ranging from 0.125 picomolar (pM) to 512 pM, which were hybridized to Affymetrix U133A microarrays. After normalization and summarization, we sorted expression levels for the human cell-line probeset groups by pM concentration and then averaged the values across all samples. For each concentration, we generated receiver-operating-characteristic (ROC) curves to evaluate how well "expressed" transcripts could be differentiated from "non-expressed" transcripts, using that concentration as a threshold. We then calculated the mean area under the ROC (AUC) across all thresholds and used this value as a comparison metric. Identical processing steps were followed for the SCAN, fRMA, RMA, and MAS5 normalization methods. As shown in Supplementary Figs. 1 and 2, MAS5 and SCAN performed best, attaining mean AUCs of 0.977 and 0.976, respectively. Results for fRMA and RMA were slightly lower (0.960 and 0.959, respectively). Note that although these differences are only small in these spike-in data, the differences will be magnified in more complex analysis scenarios, as illustrated in the examples below. These results suggest that SCAN is highly effective at detecting relatively minor differences in abundance, a crucial capability for "real-world" studies where differences in transcription between conditions may be subtle.

2.5. SCAN adjusts for batch and platform effects

In processing large data sets, differences in personnel, equipment, and laboratory conditions may lead to non-biological variability in expression, known as batch effects [11]. When multiple array versions are used for profiling, such effects may be amplified further. CMAP contains thousands of microarrays that were used to profile cancer cell line responses to dosage levels of many drugs. We examined the subset that was used to profile MCF7 cell lines treated with valproic acid and observed strong batch effects in the raw data (see Fig. 3A). However, after SCAN normalization, values fell within a similar range for each array, irrespective of batch or array type (see Fig. 3B). RMA also standardized the value ranges within each array type; however, because RMA can only be applied to samples from a single platform at a time, substantial differences persisted across

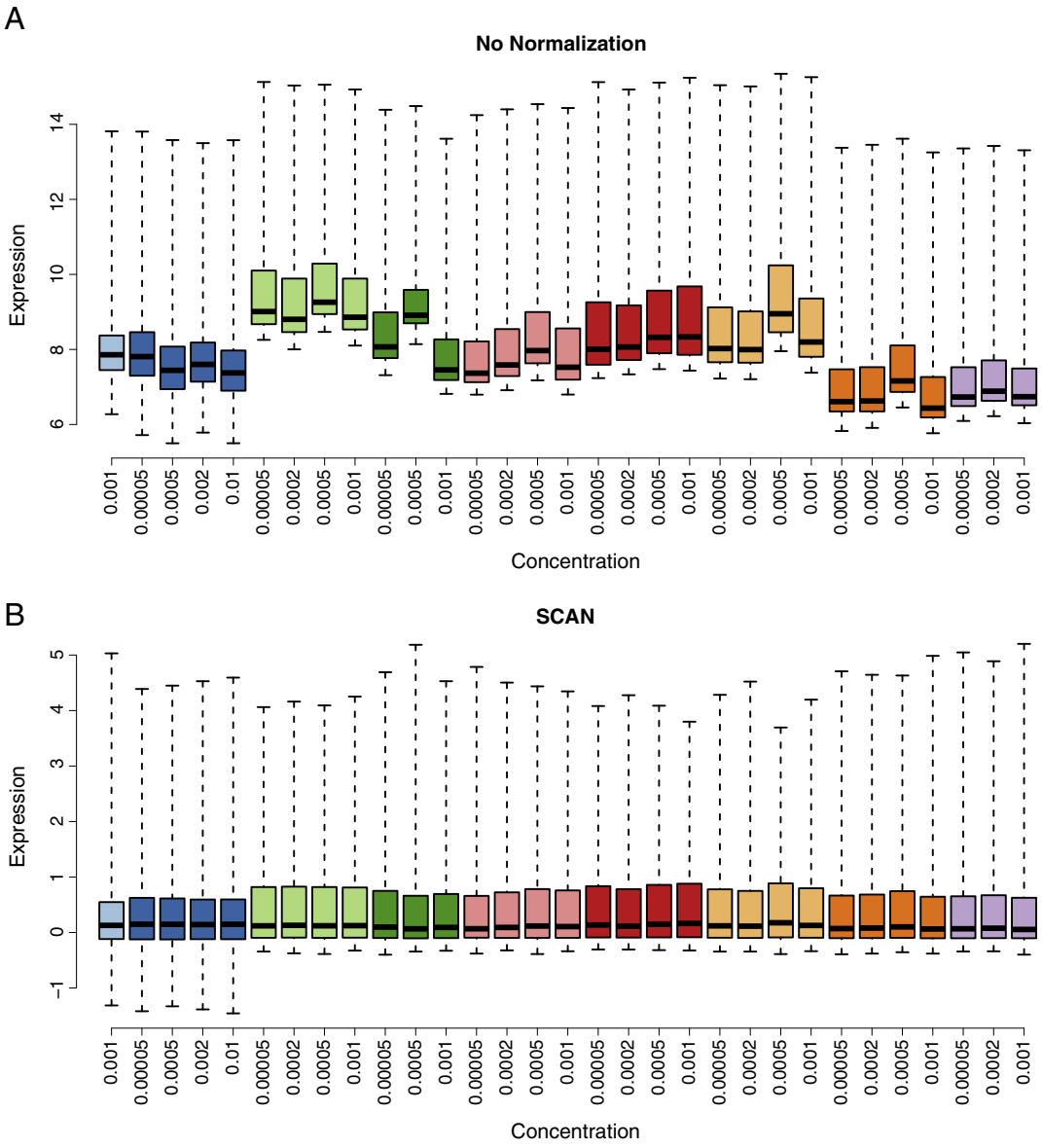


Fig. 3. SCAN adjusts for sample-level variations in expression intensity arising from platform and batch effects. In A), log2 intensity values are shown for CMAP samples treated with valproic acid. In B), SCAN normalized values are shown for the same samples. Before normalization, value ranges varied largely by batch and/or platform (each color represents a distinct batch of samples profiled on a specific Affymetrix platform). After normalization, values fell within a similar range, irrespective of batch or platform.

samples measured on different platforms; a similar pattern was observed for MAS5 (see Supplementary Fig. 4).

For an additional comparison, we tested how well valproic acid response could be estimated across the various concentration levels. For each normalization method, we generated a gene signature using cell-line data from Cohen et al. [15] and projected this signature onto the CMAP samples using BinReg 2.0 [16] with no quantile or shift-scale normalization that might bias the comparison of methods. As illustrated in Supplementary Fig. 5, each method produced response predictions that increase by concentration. However, SCAN produced predictions that ranged nearly from zero to one, whereas RMA and MAS5 predictions ranged between 0.4 and 0.8. Accordingly, SCAN appeared to be more effective at characterizing cellular activity associated with valproic acid treatment, thus enabling a clearer distinction between samples that had been treated at lower versus higher concentrations.

It is important to note that the fRMA software currently provides external reference vectors for only one (HG-U133A) of the three array versions used in CMAP. The other two versions (HT_HG-U133A and HT_HG-U133_EA), which constitute 89.0% of perturbed samples in CMAP, are unsupported because an inadequate number and diversity of reference samples have been deposited in GEO. Thus we did not include fRMA in the CMAP comparisons. This scenario illustrates an important methodological advantage of SCAN over fRMA; SCAN uses only data from a given array for normalization and thus can be applied to any one-color array version, whereas fRMA requires platform-specific ancillary samples that may or may not be available. Additionally, because fRMA reference vectors are derived from different sets of samples for each platform, platform-specific noise may be introduced.

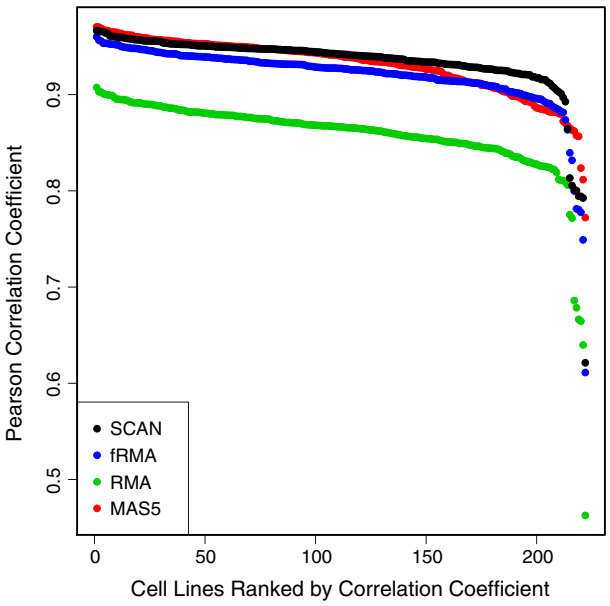


Fig. 4. Correlation of sample-wise expression levels between CCLE and GSK cell lines. CCLE and GSK data were processed using each normalization method, and sample-wise Pearson correlation coefficients were calculated. Across the samples, SCAN values were more highly correlated than for the other normalization methods, suggesting its ability to produce consistent output values for analogous samples processed in independent facilities by different personnel at different times. (Correlation coefficients were sorted for each normalization method independently before plotting.)

2.6. Concordance between two cancer cell-line compendia

Next we assessed whether the various normalization methods would produce consistent values for identical cell lines that were prepared and hybridized by different personnel, in different facilities, and at different times. Such differences may cause non-biological noise that could confound downstream conclusions. We posited that an effective normalization procedure would reduce the effects of such noise. To perform this comparison, we obtained raw files from the Cancer Cell Line Encyclopedia (CCLE) (http://www.broadinstitute.org/ccle) and a corresponding data set generated by GlaxoSmithKline (GSK) [17]. For both data sets, the Affymetrix Human Genome U133 Plus 2.0 platform was used for profiling. We identified samples that had been profiled in both CCLE and GSK ($n = 222$). For the GSK data set, which contained triplicate hybridizations of each cell line, we selected the first sample that had a GAPDH 3′/5′ ratio greater than 3.0. Next we calculated the mean Pearson correlation coefficient between corresponding samples in CCLE and GSK. SCAN produced highly consistent values across the data sets and had overall higher correlation coefficients than the other methods (see Fig. 4 and Supplementary Fig. 3).

In a follow-on analysis, we estimated RAS pathway activation within the treated cell lines using a methodology developed by Bild et al. [18] In this approach, a pathway signature is constructed by comparing expression levels of human primary mammary epithelial cell cultures infected with recombinant adenoviruses and control cells infected with green fluorescent protein. After renormalizing RAS samples from the Bild et al. study, we derived expression signatures representing pathway activation and projected those signatures onto the CCLE and GSK samples using the BinReg 2.0 algorithm. Under the expectation that normalization methods better at filtering non-biological artifacts would result in greater consistency between the two data sets, we calculated Pearson correlation coefficients based on the predicted probabilities of pathway activation. As illustrated in Fig. 5, SCAN and fRMA performed substantially better ($r = 0.870$, 0.869) than MAS5 and RMA ($r = 0.757$, 0.323).

Taken together, these findings suggest that SCAN produces standardized data that account for unavoidable variations that occur in microarray studies.

3. Discussion

Prior capital investments in microarray equipment and enormous troves of institutional and publicly available data will surely lead to microarrays' continuance as a prominent research tool for years to come. In the Gene Expression Omnibus alone, hundreds of thousands of raw data files are available for secondary analyses [14]. Optimization of microarray normalization procedures remains imperative to genomic research by enabling more accurate characterizations of biomedical phenomena and providing insights into the design of RNA-sequencing projects. In this study, we have described SCAN, a novel normalization technique capable of processing one microarray sample at a time. Through analyses of spike-in data and cell-line experiments, we have demonstrated that SCAN performs similarly or better than several existing methods, including other single-sample techniques and the most popular multi-sample method.

Various normalization techniques have been proposed over the years, but most depend on groups of samples for processing. Although in some simple cases such an approach may be adequate for standardizing values within a given data set, sample aggregation introduces logistical challenges that may limit their usefulness in settings where samples accrue over time—for example, in personalized-medicine workflows or meta-analyses that span multiple studies. SCAN takes advantage of the large amount of raw data produced by a given array and corrects for binding-affinity biases on a single-sample basis. This approach provides various methodological advantages, including that no information is borrowed from external samples. Additionally, SCAN is robust to differences among microarray platforms because the same linear model is applied for every platform. Indeed, we observed that

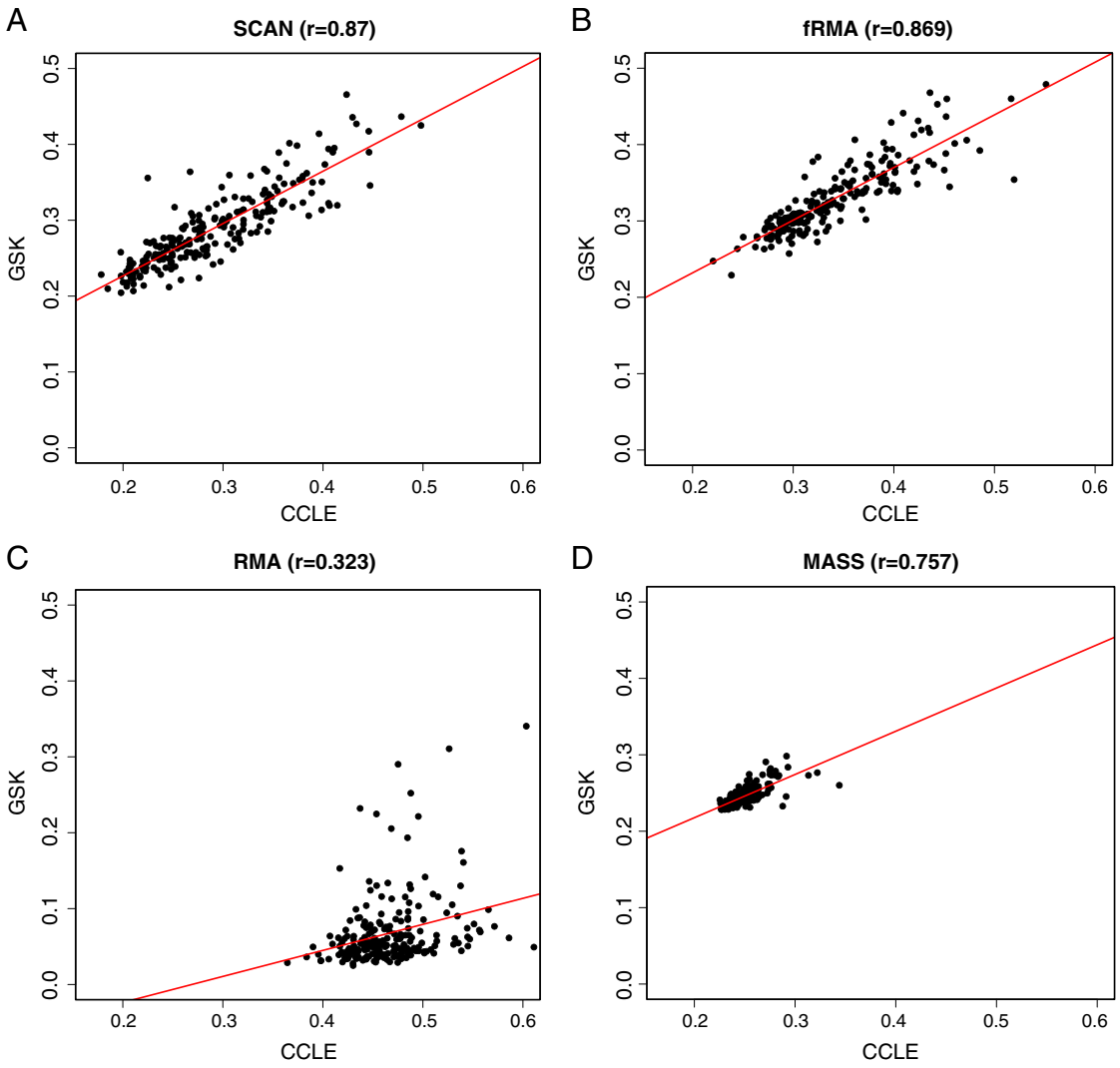


Fig. 5. Comparison of RAS pathway activation probabilities for CCLE and GSK cell line data. The probability of RAS pathway activation was estimated for individual samples in CCLE and GSK cell lines. For each normalization method, a pathway signature was derived through a comparison of expression levels in RAS-activated cell cultures and controls. The signatures were then projected onto the cell line data, and a probability of pathway activation was calculated. For SCAN and fRMA (A and B), probabilities were highly concordant across the data sets, whereas for RAS and MAS5 (C) and D) the probabilities were less concordant and fell within narrower ranges. Lack of replication between data sets may lead to markedly different biological conclusions, depending on which data set is examined.

SCAN produced highly concordant (see Supplementary Fig. 6) gene-summarized values for Human Exon ST 1.0 and HG-U133A arrays from the same ovarian-cancer sample (TCGA-04-1335) [19], even though the array platforms differ considerably in design and in the number of probes that they contain.

Across our analyses, SCAN performed as well as or better than all other techniques, which suggests its superior utility and consistency as a gene-expression normalization method. In our analyses, the most popular [9] multi-array method, RMA, underperformed in relation to the single-array methods. In particular, RMA faltered in its attempts to produce congruous expression values for the cancer cell-line data that had been processed and hybridized at different facilities. Conceivably, RMA's attempt to standardize across arrays introduces subtle biases that currently hamper cross-study replicability. In fact, Giorgi et al. demonstrated that RMA's median-polish summarization step can introduce between-array correlation artifacts [20]. fRMA attempts to overcome the limitations of multi-array processing by constructing an external reference database of probe-level effects. Indeed, across the various analyses that we performed, fRMA performed better than its multi-array counterpart. However, fRMA comes with practical limitations that hamper its broad utility. Researchers who desire to apply this method to samples from array versions that are currently unsupported by fRMA must go through an extensive process of constructing and validating a new reference database for each platform. Importantly, variation in the reference data can itself affect data quality. The SCAN algorithm, on the other hand, can be applied to any single-channel array (such as those manufactured by Affymetrix and Nimblegen) that provides a support file matching probe sequences with probe coordinates. In some of our analyses, MAS5 performed similarly to SCAN; however, it also lacked consistency when applied to the cell-line data. And perhaps most importantly, MAS5 can only be applied to older generations of Affymetrix arrays that contain mismatch probes. Affymetrix has also provided the probe logarithmic intensity error (PLIER) method (http://media.affymetrix.com/support/technical/technotes/plier_technote.pdf), which corrects for GC content and can be applied to older- and newer-generation arrays. However, PLIER relies upon external samples for its summarization approach and thus was not evaluated in this study.

By accurately distinguishing between true biological signal and background noise, SCAN helps elucidate the cellular mechanisms driven by complex variations in transcriptional activity. Effectively correcting for technological biases on a single-sample basis is particularly important for microarrays to gain acceptance in the clinic.

4. Materials and methods

4.1. Model-based adjustment for probe-nucleotide composition

To correct for binding-affinity biases, we developed a linear model that is a variation on the Model-based Analysis of Tiling arrays (MAT) method [12], which was designed for Affymetrix tiling arrays. MAT has been shown to account for as much as 50% of the variation that results from binding-affinity and cross-hybridization effects in tiling-array samples [12]. As with MAT, our model contains a nucleotide indicator for each position on a given 25-mer probe sequence, along with quadratic terms representing nucleotide counts. However, the novel contribution of SCAN, which makes it appropriate for expression-array experiments, is the use of a mixture of two Gaussian distributions, one mixture component measuring background noise and the other measuring signal (plus background). The main component of interest is the background distribution, as the other component will likely result in a confounded distribution of true biological signal of interest, cross-hybridization, etc. More specifically, letting P_i represent the probe-intensity value for probe $i = 1,\ldots, N$, we assume that $\log(P_i) = (1-\Delta_i)\, Y_{1i} + \Delta_i\, Y_{2i}$ where $Y_{1i} \sim N(X_i\theta_1, \sigma_1^2)$ and $Y_{2i} \sim N(X_i\theta_2, \sigma_2^2)$, where $X_i\theta_m$ is the m^{th} mixture-component MAT model for probe i presented in the following equation for a given m:

$$Y_{mi} = \alpha_m n_{iT} + \sum_{j=1}^{25} \sum_{k \in \{A,C,G\}} \beta_{mjk} I_{ijk} + \sum_{l \in \{A,C,G,T\}} \gamma_{ml} n_{il}^2 + \varepsilon_{mi} \quad (1)$$

where α_m is the baseline value based on the number of T's in the probe sequence, n_{il} is the nucleotide l count, β_{mjk} is the effect of each nucleotide k (except T which is already modeled in α_m) at each position j, I_{ijk} is an indicator function such that $I_{ijk} = 1$ if the nucleotide at position j is k in probe i, γ_{ml} is the effect of nucleotide count squared, and $\boldsymbol{\varepsilon}_{mi}$ is the probe-specific error term, assumed to follow a normal distribution. The probe mixture component indicators, Δ_i, take on the values either 0 or 1, indicating from which mixture component each probe was drawn. However, these indicators are not observable and will be estimated in the mixture-modeling procedure. We formulate the above model based on a 'missing data' approach (where the Δ_is are the missing data), and apply the Expectation-Maximization (EM) algorithm to estimate the model parameters [21]. Specifically, assuming the previously described Gaussian mixture for log (P_i) and assuming $\Delta_i \sim Bernoulli\ (\pi)$ where π is the proportion of non-background probes on the array, we obtain the complete data likelihood of

$$L(\mathbf{y}, \Delta) \propto \prod_{i=1}^{N} [(1-\pi) f_1(y_i|\theta_1)]^{1-\Delta_i} [\pi f_2(y_i|\theta_2)]^{\Delta_i}, \quad (2)$$

where $f_1(y_i|\theta_1)$ and $f_2(y_i|\theta_2)$ are Gaussian densities representing the distributions of the 'background' and 'background + signal' distributions of probe respectively. Because the complete data log-likelihood is linear in the missing data (Δ_is), applying the EM algorithm is straight-forward in this case. The Expectation Step (E-Step) consists of imputing the missing data in the complete data log-likelihood Eq. (2) with their expected values using the most recent estimates of the model parameters from the previous Maximization Step (M-Step). The M-Step then consists of maximizing the model parameters given using Eq. (2) and the most recent imputed missing data from the E-step. The E-Step and M-Step are then iterated until the model parameters converge to the Maximum Likelihood Estimates (MLEs). These MLEs are then used to standardize the data as described below.

The model (for each mixture component) contains 80 parameters, 1 for α, 3×25 for β, and 4 for γ. SCAN estimates these parameters by ordinary least squares using a subset of probes ($n = 50{,}000$) on the array. After the parameters have been estimated, a baseline intensity is predicted for each probe. After fitting the model, we use the background mixture component to obtain an estimate for the background expected intensity for probe i, m_i, to remove the probe-specific variation. We bin probes into groups of 5000 with similar m_i and normalize the data as follows:

$$t_i = \frac{\log(P_i) - m_i}{s_{i\ bin}} \quad (3)$$

where m_i is the baseline intensity predicted by the model based on the sequence of probe i, and $s_{i\ bin}$ is the standard deviation of the bin to which probe i belongs.

The MAT statistical model was previously applied to Affymetrix exon arrays [13]. However, we have incorporated modifications that were not included in their approach. Most importantly, their approach did not take into account that most of the variation in an expression experiment is likely due to the true biological signal of interest. The original MAT approach estimates parameters based on all the variation in the data (background and signal), which will remove some biological signal from the data and result in reduction of the signal-to-noise ratio. SCAN's novel introduction of a mixture modeling approach results in the removal of mostly background while preserving the biological signal in the data.

4.2. Software

SCAN has been implemented as an open-source software package that can be downloaded freely from http://jlab.bu.edu/software/scan-upc. It can be applied to any single-channel platform for any species, so long as metadata describing the coordinates and sequence of each probe is available. SCAN outputs probe-level expression values, which by default are summarized to gene-level values using a trimmed mean.

In this study, the R statistical software [22] and its associated ROCR package [23] were used for analyzing results and producing graphics. The implementations of fRMA, RMA, and MAS included in the affy [24] and frma [9] R/Bioconductor packages were applied with default parameters.

4.3. Data summarization

Affymetrix offers a variety of microarray platforms that target various applications. For each platform, Affymetrix provides a tab-delimited file that indicates the position and sequence of each probe. We downloaded these files from http://www.affymetrix.com and parsed the relevant information into consistently formatted "probe-tab" files. (The probe-sequence information for the recently released Affymetrix GG-H arrays was downloaded from http://gluegrant1.stanford.edu/~DIC/GGHarray.) By design, Affymetrix places a variety of background and control probes on the Affymetrix ST and GG-H arrays. To avoid overfitting the linear model, SCAN ignores these probes when estimating model parameters but does output normalized expression values for these probes. Unlike previous generations of Affymetrix microarrays, the ST and GG-H arrays also include a small proportion of probes that have sequences shorter than 25 base pairs; because the linear model requires all probes to have the same length, SCAN ignores these probes.

Upon normalizing a CEL file, SCAN outputs a tab-delimited text file containing three columns. The first column lists a unique identifier for each probe (a concatenation of the probe's X and Y position on the array). The second column lists the normalized expression value for each probe. And the third column contains values representing the probability that each probe is expressed above background noise. (A companion publication describing the latter is pending.)

To support gene-level summaries, probe values are mapped to genes using either the mapping files provided by Affymetrix or custom mapping files. By default, we use custom mapping files provided by the BrainArray resource [25]; if no such mapping file is available, we default to mappings provided by Affymetrix. However, users also have the

option to specify alternative mapping files that adhere to a simple tab-delimited format. Currently, each human GeneChip platform listed on http://www.affymetrix.com is supported.

After probes have been mapped to genes, gene-level values are calculated using a 10% trimmed mean. If fewer than 3 probes are associated with a gene, no gene-level value is calculated.

Author contributions

SRP, AHB, and WEJ conceived and designed the study. SRP performed all data analyses. SRP and YS developed the software. MEL and JDC provided feedback on the manuscript. WEJ developed the statistical approach. AHB designed experiments testing SCAN on external datasets. SRP, AHB, and WEJ wrote the manuscript. All authors read and approved the manuscript.

Acknowledgments

We thank Adam Cohen for his help with microarray quality control. We gratefully acknowledge an allocation of computer time from the Fulton Supercomputing Lab at Brigham Young University. We thank Affymetrix, the Broad Institute, the Novartis Institutes for Biomedical Research, and GlaxoSmithKline for making raw data available publicly. This work was partially supported by funds from the United States National Institutes of Health (1R01HG00569 and 5T32CA093247).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2012.08.003.

References

[1] L.J. van 't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A.M. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, S.H. Friend, Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (2002) 530–536.
[2] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F.L. Baehner, M.G. Walker, D. Watson, T. Park, W. Hiller, E.R. Fisher, D.L. Wickerham, J. Bryant, N. Wolmark, A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer, N. Engl. J. Med. 351 (2004) 2817–2826.
[3] U. McDermott, J.R. Downing, M.R. Stratton, Genomics and the continuum of cancer care, N. Engl. J. Med. 364 (2011) 340–350.
[4] S.S. Lo, P.B. Mumby, J. Norton, K. Rychlik, J. Smerage, J. Kash, H.K. Chew, E.R. Gaynor, D.F. Hayes, A. Epstein, K.S. Albain, Prospective multicenter study of the impact of the 21-gene recurrence score assay on medical oncologist and patient adjuvant breast cancer treatment selection, J. Clin. Oncol. 28 (2010) 1671–1676.
[5] R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, T.P. Speed, Summaries of Affymetrix GeneChip probe level data, Nucleic Acids Res. 31 (2003) e15.
[6] Z. Wu, R.A. Irizarry, R. Gentleman, F. Martinez-Murillo, F. Spencer, A model-based background adjustment for oligonucleotide expression arrays, J. Am. Stat. Assoc. 99 (2004) 909–917.
[7] E. Hubbell, W. Liu, R. Mei, Robust estimators for expression analysis, Bioinformatics 18 (2002) 1585–1592.
[8] W.K. Lim, K. Wang, C. Lefebvre, A. Califano, Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks, Bioinformatics 23 (2007) i282–i288.
[9] M.N. McCall, B.M. Bolstad, R.A. Irizarry, Frozen robust multiarray analysis (fRMA), Biostatistics 11 (2010) 242–253.
[10] J. Lamb, E.D. Crawford, D. Peck, J.W. Modell, I.C. Blat, M.J. Wrobel, J. Lerner, J.P. Brunet, A. Subramanian, K.N. Ross, M. Reich, H. Hieronymus, G. Wei, S.A. Armstrong, S.J. Haggarty, P.A. Clemons, R. Wei, S.A. Carr, E.S. Lander, T.R. Golub, The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease, Science 313 (2006) 1929–1935.
[11] J.T. Leek, R.B. Scharpf, H.C. Bravo, D. Simcha, B. Langmead, W.E. Johnson, D. Geman, K. Baggerly, R.A. Irizarry, Tackling the widespread and critical impact of batch effects in high-throughput data, Nat. Rev. Genet. 11 (2010) 733–739.
[12] W.E. Johnson, W. Li, C.A. Meyer, R. Gottardo, J.S. Carroll, M. Brown, X.S. Liu, Model-based analysis of tiling-arrays for ChIP-chip, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 12457–12462.
[13] K. Kapur, Y. Xing, Z. Ouyang, W.H. Wong, Exon arrays provide accurate assessments of gene expression, Genome Biol. 8 (2007) R82.
[14] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, R.N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, A. Soboleva, NCBI GEO: archive for functional genomics data sets–10 years on, Nucleic Acids Res. 39 (2011) D1005–D1010.
[15] A.L. Cohen, R. Soldi, H. Zhang, A.M. Gustafson, R. Wilcox, B.E. Welm, J.T. Chang, E. Johnson, A. Spira, S.S. Jeffrey, A.H. Bild, A pharmacogenomic method for individualized prediction of drug sensitivity, Mol. Syst. Biol. 7 (2011) 513.
[16] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson, J.R. Marks, J.R. Nevins, Predicting the clinical status of human breast cancer by using gene expression profiles, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 11462–11467.
[17] J. Greshock, K.E. Bachman, Y.Y. Degenhardt, J. Jing, Y.H. Wen, S. Eastman, E. McNeil, C. Moy, R. Wegrzyn, K. Auger, M.A. Hardwicke, R. Wooster, Molecular target class is predictive of in vitro response profile, Cancer Res. 70 (2010) 3677–3686.
[18] A.H. Bild, A. Potti, J.R. Nevins, Linking oncogenic pathways with therapeutic opportunities, Nat. Rev. Cancer 6 (2006) 735–741.
[19] The Cancer Genome Atlas Research Network, Integrated genomic analyses of ovarian carcinoma, Nature 474 (2011) 609–615.
[20] F.M. Giorgi, A.M. Bolger, M. Lohse, B. Usadel, Algorithm-driven artifacts in median polish summarization of microarray data, BMC Bioinform. 11 (2010) 553.
[21] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. B Methodol. 39 (1977) 1–38.
[22] R Development Core Team, R: A language and environment for statistical computing [computer program]. Available from http://www.r-project.org/2011.
[23] T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, ROCR: visualizing the performance of scoring classifiers [computer program]. Available from http://cran.r-project.org/package=ROCR2009.
[24] L. Gautier, L. Cope, B.M. Bolstad, R.A. Irizarry, Affy–analysis of Affymetrix GeneChip data at the probe level, Bioinformatics 20 (2004) 307–315.
[25] M. Dai, P. Wang, A.D. Boyd, G. Kostov, B. Athey, E.G. Jones, W.E. Bunney, R.M. Myers, T.P. Speed, H. Akil, S.J. Watson, F. Meng, Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data, Nucleic Acids Res. 33 (2005) e175.
[26] H.J. Kang, Y.I. Kawasawa, F. Cheng, Y. Zhu, X. Xu, M. Li, A.M.M. Sousa, M. Pletikos, K.A. Meyer, G. Sedmak, T. Guennel, Y. Shin, M.B. Johnson, Z. Krsnik, S. Mayer, S. Fertuzinhos, S. Umlauf, S.N. Lisgo, A. Vortmeyer, D.R. Weinberger, S. Mane, T.M. Hyde, A. Huttner, M. Reimers, J.E. Kleinman, N. Sestan, Spatio-temporal transcriptome of the human brain, Nature 478 (2011) 483–489.