JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 17, Number 10, 2010
© Mary Ann Liebert, Inc.
Pp. 1385–1395
DOI: 10.1089/cmb.2010.0049

Subset Quantile Normalization Using Negative Control Features

ZHIJIN WU[1] and MARTIN J. ARYEE[2]

ABSTRACT

Normalization has been recognized as a necessary preprocessing step in a variety of high-throughput biotechnologies. A number of normalization methods have been developed specifically for microarrays, some general and others tailored for certain experimental designs. All methods rely on assumptions about data characteristics that are expected to stay constant across samples, although some make it more explicit than others. Most methods make assumptions that certain quantities related to the biological signal of interest stay the same; this is reasonable for many experiments but usually not verifiable. Recently, several platforms have begun to include a large number of negative control probes that nonetheless cover nearly the entire range of the measured signal intensity. Using these probes as a normalization basis makes it possible to normalize without making assumptions about the behavior of the biological signal. We present a subset quantile normalization (SQN) procedure that normalizes based on the distribution of non-specific control features, without restriction on the behavior of specific signals. We illustrate the performance of this method using three different platforms and experimental settings. Compared to two other leading nonlinear normalization procedures, the SQN method preserves more biological variation after normalization while reducing the noise observed on control features. Although the illustration datasets are from microarray experiments, this method is general for all high throughput technologies that include a large set of control features that have constant expectations across samples. It does not require an equal number of features in all samples and tolerates missing data. Supplementary Material is available online at www.liebertonline.com.

Key words: DNA arrays, functional genomics, genes chips, gene expression.

1. INTRODUCTION

High-throughput biotechnologies have become increasingly important in biomedical research. Among these, DNA microarrays are probably the most widely used in applications studying variation of the transcriptome, genome and epigenome. Microarray technology simultaneously quantifies a large number of DNA or RNA species with various sequences by labeling the target sample with

[1]Center for Statistical Sciences and Department of Community Health, Brown University, Providence, Rhode Island.
[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, and Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, Maryland.

fluorescent dyes and hybridizing it to features (probes) with sequences complementary to the target molecules. The target concentration is reflected in the fluorescent intensities observed on the complementary feature. Because the hybridization efficiency of each feature is different and unknown, microarrays only provide a measure of relative abundance of the target molecules. Thus, the quantity of interest on each feature is the biological variation of its target molecule in different samples. In addition to the quantity of interest, a number of other factors affect the feature intensities on an array in a systematic way. These factors include sample preparation, hybridization, and array processing. The systematic variations caused by these factors are of no biological interest and are sometimes referred to as "*obscuring variations*" (Bolstad et al., 2003). It is not uncommon for the magnitude of obscuring variation to exceed that of true biological variation. It has been well recognized that normalization is a necessary step to remove or reduce such variation in order to make data from different arrays comparable before further analysis can be carried out.

A number of normalization methods have been developed since the introduction of microarray technology. Some methods are general and some are tailored towards specific experimental designs. Regardless of the biological application, all normalization methods reflect the observation that many factors causing obscuring variation affect the entire array in some systematic fashion. For example, one sample may have higher labeling efficiency and the feature intensities in this sample would tend to be higher in general than those from other samples, or one scanner may give higher readings than another. If the overall hybridization in different samples is not expected to differ, we would like to remove the global "array effect" in normalization. When the labeling or scanner effect is the same for all features, a scaling normalization such as aligning the medians or means of each sample would suffice.

This example illustrates a key issue in normalization: What is expected to be constant across samples? All normalization methods make such assumptions, explicitly or implicitly. Normalization is then achieved by equalizing certain summary statistics based on the assumption(s). The scaling normalization works well only when the obscuring variation is a linear effect, which is rare in reality. A number of non-linear methods have been proposed to allow more flexible normalization. Using gene expression as an example, if one assumes that only a small set of genes have differential expression, or that up- and down-regulation are approximately symmetrical, the distribution of gene expression measures on an array should be similar across all samples. *Quantile normalization* (QN) (Amaratunga and Cabrera, 2001; Bolstad et al., 2003) can be applied in such cases and has been demonstrated to have favorable properties (Bolstad et al., 2003). Loess normalization removes intensity-dependent biases in differential expression and works well under similar assumptions (Yang et al., 2002). Sometimes a group of house keeping genes are assumed to have constant expression across samples and are used as internal controls for normalization. However, there have been numerous reports that housekeeping genes are found to be quite variable in given situations (Thellin et al., 1999). Another choice is to use a rank-invariant set of genes (Li and Wong, 2001; Tseng et al., 2001), whose expression levels remain a similar rank on an array across samples, and normalize so that expression measures of these genes are constant. The size of the rank-invariant set depends on the sample data and measurements from this set may not span the entire intensity range (Yang et al., 2002).

All of the normalization methods mentioned above have one aspect in common: assumptions are made on the stability of expression levels of certain genes, that is, assumptions on the behavior of specific biological signal. These assumptions are rarely verifiable and in some situations known to be violated. Fortunately, a series of recent oligonucleotide arrays have started to include a large number (over ten thousand) of negative control probes. These new platforms include the Affymetrix Exon arrays, Gene St arrays, tiling arrays, and Illumina arrays. The negative control probes in these arrays are designed to monitor the extent of non-specific binding and to assist in background estimation and correction. They, however, also provide valuable information about systematic obscuring variation that prove to be vital to normalization.

One crucial observation about these control probes is that their intensity range spans almost the entire range of the array. This may be counter-intuitive since these probes are designed to have no complementary match to the genome/transcriptome of the targeted species. Since their intensities are a result of non-specific binding only, one may expect them to be much lower than that on the probes with specific binding. However, due to great heterogeneity in the probe's affinity for non-specific binding, the intensities observed on control probes actually vary greatly. Fig. 1 shows the empirical distribution of log intensities on all features with specific binding from an Affymetrix Human Gene St array. The 99*th* percentile and maximum of log intensity on the control probes are marked, showing the nearly complete coverage of the dynamic range of specific binding. This is consistent with observations on other microarray platforms (Wu et al., 2004).

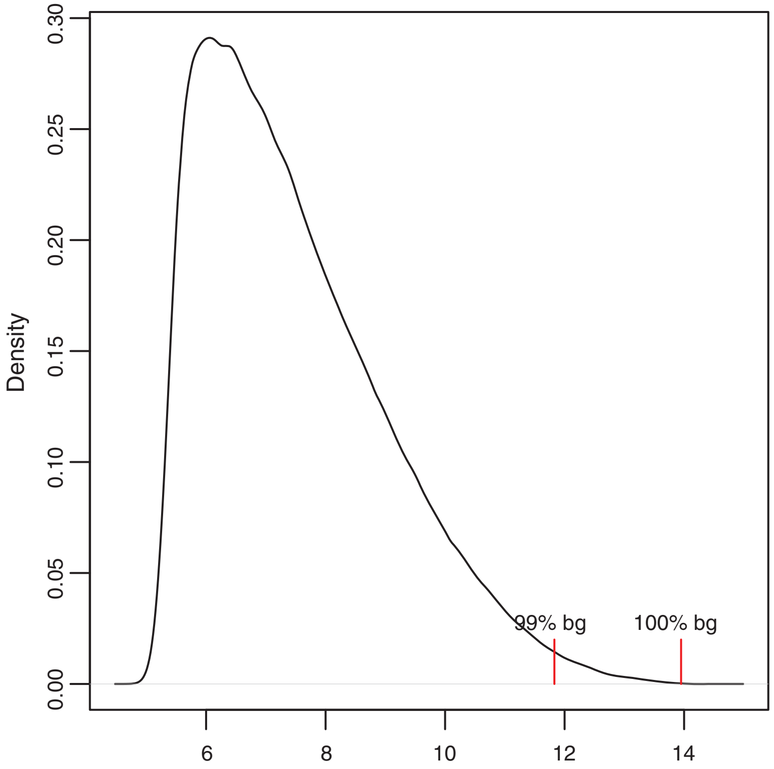

FIG. 1. Probability density function of feature intensities (log_2 scale) from an Affymetrix Human Gene 1.0 St array. The 99*th* percentile and the maximum of negative control (bg) probes are marked to show that the intensity of control probes span almost the entire range of probe intensity.

The availability of a large number of negative control probes makes it possible to observe the impact of systematic obscuring variation that normalization procedures hope to remove. The intensities on the control probes are affected by the factors that have overall impact for the entire array, such as labeling efficiency, scanner setting, time of experiment, and hybridization conditions. This makes them ideal controls for normalization, since we do expect the intensities on these features to stay constant regardless of how the biological signal varies from sample to sample.

In this article, we propose a normalization method based on a group of negative control features that are expected to produce constant measurements across samples. We present results from three different platforms, two gene expression datasets and a DNA methylation dataset. We demonstrate improvements in preserving biological variation while reducing systematic noise in all three examples. Although the examples are all from microarrays, the methodology is applicable to other platforms that include such controls, regardless of the specific technology or biomedical application.

2. DATA

- *Tissue experiment:* A collection of 11 tissues (brain, breast, heart, kidney, liver, pancreas, prostate, skeletal muscle, spleen, testes, and thyroid), each with three biological replicates, are included in the experiment, giving a dataset of 33 samples (arrays). The samples are hybridized to Affymetrix Human Gene 1.0 ST Arrays. On this platform, there are 16,943 negative control probes and 764,885 perfectMatch probes. This data set is provided to the public by Affymetrix at www.affymetrix.com/support/technical/sample_data/gene_1_0_array_data.affx.
- *HG-U95 platform data:* This dataset includes two groups of samples. The first set includes 10 samples randomly selected from a database of microarray data on HG-U95 platform. These samples can be obtained from GEO with accession numbers GSM43986, GSM134229, GSM44814, GSM44199, GSM2821, GSM15807, GSM44906, GSM134105, GSM44044, and GSM2894. The second set

includes three background experiments, in which either yeast genomic DNA, polyC RNA or polyG RNA are used as target sample. Only non-specific binding are expected in these three samples. This platform includes perfectMatch (PM) and mismatch (MM) probes.

- *DNA methylation data:* DNA methylation is an epigenetic mark related to the control of gene expression (Bird, 2002). In this experiment, adult cells (fibroblasts) were reprogrammed into induced pluripotent stem cells (iPS) (Doi et al., 2009). Differences in transcriptional programs between the differentiated adult state and undifferentiated stem-cell like state are mediated in part by significant changes in DNA methylation profiles. Six samples of each type were processed and hybridized to two-color Nimblegen CHARM DNA methylation microarrays (GEO accession number GSE18227) (Irizarry et al., 2008).

3. METHODS

The proposed normalization method does not rely on assumptions on the property of biological signals in different samples. Instead, we use quantiles of the negative control probes as "anchors" and require that these statistics are equalized after normalization. Intensities of all probes on an array are adjusted according to their relationship to the control quantiles on the same array. We term this method the "subset quantile normalization (SQN)," in order to differentiate with the complete QN that makes the distributions of the entire array equal.

Specifically, for each array a, we estimate the cumulative distribution function (CDF) F_a of the subset of probes that serve as controls. We use the estimates of F_a to define a reference control distribution F as the target distribution for the control probes. Now consider any probe on an array, if its raw intensity equals the q^{th} quantile of the control probes on the same array, its normalized intensity is defined as the q^{th} quantile from the reference distribution F. That is,

$$\tilde{y}_{a,j} = F^{-1}\{F_a(y_{a,j})\}.$$

We can estimate F_a by the empirical CDF $\hat{F}_a$ and use the median of each quantile to define a reference control distribution $\hat{F}$. This works well for most of the data except the tails of the distributions because $\hat{F}^{-1}$ is bounded by the observed intensities from the control probes. The probes whose intensity is beyond the maximum of the control set would have normalized values truncated at $\hat{F}^{-1}(1)$. To avoid the problem in tail areas, we use a semi-parametric approach. We first estimate F_a parametrically as a mixture of k normal distributions,

$$\Phi_a^k(x) = \sum_{i=1}^{k} \pi_{ai}\Phi(\frac{x-\mu_{ai}}{\sigma_{ai}}),$$

where Φ is the standard normal CDF.

The final estimate of F_a is defined as a weighted average of the empirical CDF and the normal mixture CDF,

$$\tilde{F}_a(x) = w\Phi_a^k(x) + (1-w)\hat{F}_a(x).$$

F is defined with a similar approach. From each array a, let $y_{a,(j)}$ be the j^{th} order statistic of the control probes. The medians of the order statistics, $\bar{y}_{(j)} = \text{Median}_a(y_{a,(j)})$, define the reference control intensity vector. We use the values $\bar{y}_{(j)}$ to estimate a normal mixture distribution $\Phi^k = \sum_{i=1}^{k} \pi_i \Phi(\frac{x-\mu_i}{\sigma_i})$ and compute a weighted average of Φ^k and the empirical CDF,

$$\tilde{F}(x) = w\Phi^k(x) + (1-w)\hat{F}(x).$$

With the CDF estimates $\tilde{F}(x)$ and $\tilde{F}_a(x)$ available, we define the normalized intensities as

$$\tilde{y}_{a,j} = \tilde{F}^{-1}\{\tilde{F}_a(y_{a,j})\}.$$

We leave k and w as tuning parameters for smoothness. Given k, the parameters for the mixture distributions πs,μs and σs can be estimated by maximum likelihood using EM algorithm. In practice, $k = 5$

gives very good flexibility to a wide variety of distributions. We have used $k = 5$ and $w = .9$ in our example. Our implementation of the method uses the R package *Mclust* (Fraley and Raftery, 2009) for the EM estimation of mixture parameters and the R package *nor1mix* (Mchler, 2007) for the computation of mixture distribution quantiles.

RESULTS

Tissue experiment. We apply SQN to the *Tissue* experiment data and compare the results with two other most widely used nonlinear normalization methods, the complete QN, and the loess normalization—both of which are found to perform well and are incorporated in numerous preprocessing modules (Irizarry et al., 2003a; Smyth, 2005; Yang et al., 2007; Gautier et al., 2004). There are several variants of the loess normalization, and we have used the cyclic loess normalization (Bolstad et al., 2003) in this comparison.

First, we examine the distributions of raw intensities on the negative control probes. Since the sequences of these probes are not present in the human genome, we expect the their intensities to reflect the systematic technical array variation instead of biological variation. The empirical distributions of the raw intensities on these antigenomic probes appear to have different location and scale, as demonstrated by the various medians and inter-quartile-ranges in Figure 2A, indicating the need for normalization.

The loess and complete QN normalization do not make a lot progress in making these control probes more comparable across samples, as seen in Figure 2B,C. This shows that after loess or complete QN, what is expected to be constant may still have considerable variation. In contrast, the subset quantile normalization forces the control probes to have the identical distribution by design (Fig. 2D). The loess normalization is very similar to the complete QN in all comparisons to follow. The results from loess normalization are

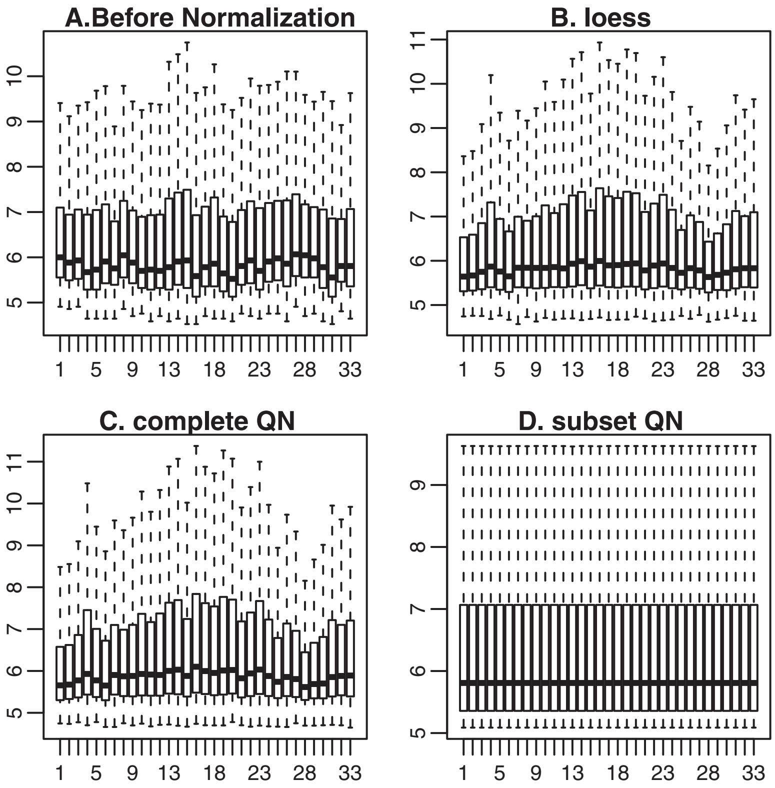

FIG. 2. The boxplots of the feature intensities (log $_2$ transformed) from the negative control probes in the 33 samples from 11 tissue types. **(A)** Raw intensities without normalization. **(B)** Cyclic loess normalization is applied to the complete set of probes including the perfectMatch probes and the negative controls. Shown are boxplots of the negative controls after normalization. **(C)** Like B but with complete quantile normalization. **(D)** Like B but with subset quantile normalization.

therefore omitted in the main text and are provided in Supplementary Material (for online Supplementary Material, see www.liebertonline.com).

The goal of normalization is certainly not just to make the control data stable. More importantly, we want to reduce or remove the obscuring variability on the signal-bearing probes. Since the variation observed in the raw data is a combination of biological variation and obscuring processing variation, we would expect a good normalization procedure to reduce the variation among replicates and make them more similar. We thus compute the within-tissue variances for all 11 tissue types and average the 11 variances for each probe as a pooled within-tissue variance. Since the probe intensity variance is commonly observed to vary across log intensity levels, we stratify the probes into 20 groups based on the average raw intensity level.

Figure 3 compares the pooled within-tissue variances before and after normalization, across the range of observed intensities. All normalization methods reduce the variability among replicates, compared to the unnormalized data. Interestingly, the complete QN appears to do a more aggressive adjustment (a greater reduction of variance), although Figure 2C shows that it does not fully normalize the control probes.

Reducing certain variability alone is only one side of the story and never enough to show the benefit of one normalization method over another. We also want to make sure that variation of interest, i.e, the actual biological variation, are preserved in the normalized data. We proceed to compare the cross tissue type variances. A good normalization would reduce technical variation among replicates, but retain real biological variation. Figure 4 shows that the complete QN reduces a lot of the between tissue variance, while the subset QN preserves the variation at the same level as the unnormalized data. This suggests that the complete QN may have paid the price of reducing signal in order to reduce noise, and may have over adjusted in this example.

In order to evaluate the variance and bias trade-off of these normalization methods, we compare the between- and within-tissue variances of each probe. Since a greater extent of differential expression is

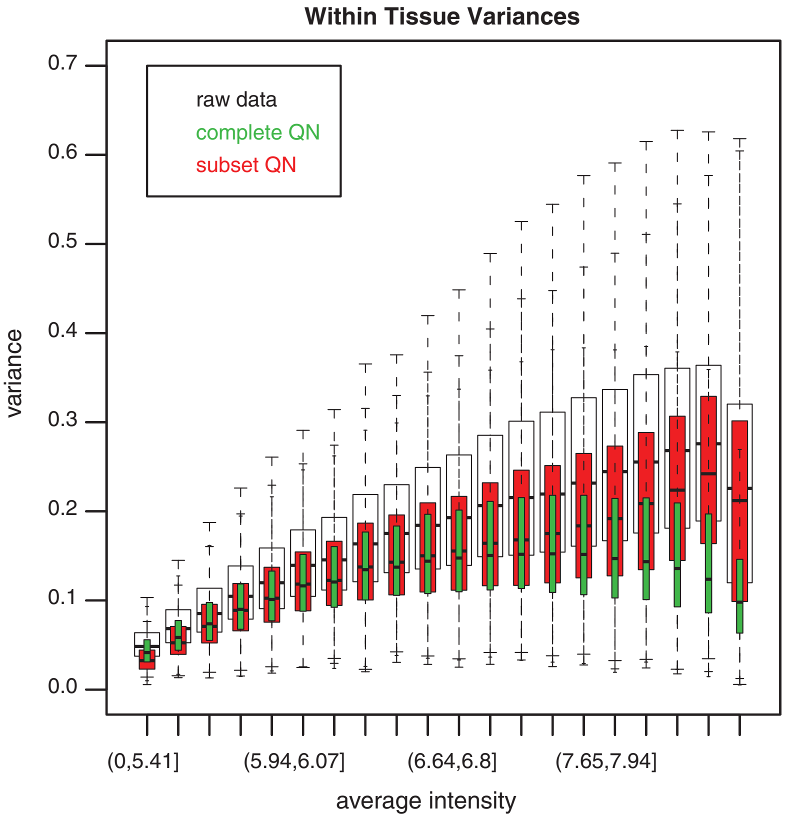

FIG. 3. The within-tissue variance comparison. For each probe, the biological replicate variances for each of the 11 tissue types are averaged to give a pooled within-tissue variance. The probes are stratified by average raw intensities into 20 equal-sized groups. Boxplots of the within-tissue variances from before and after normalization are overlaid for easy comparison.

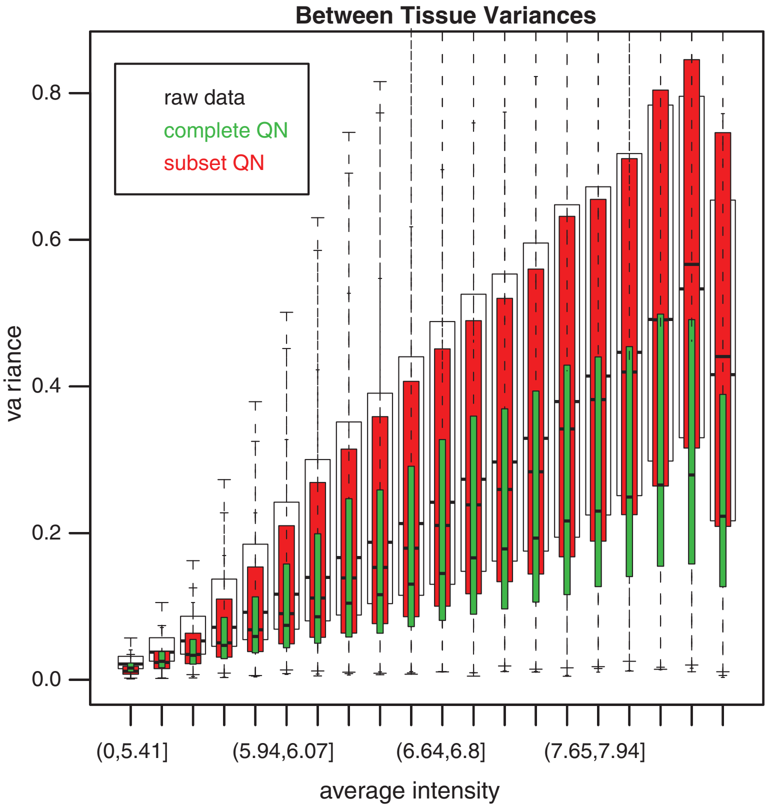

FIG. 4. The between-tissue variance comparison. For each probe, the average intensity from the triplicates of each tissue type is computed and the variance of the tissue average intensities is computed. The probes are stratified by average raw intensities into 20 equal-sized groups. Boxplots of the between-tissue variances from before and after normalization are overlaid for easy comparison.

expected between different tissue types than between biological replicates of the same tissue, we compute the ratio of between and within tissue variances. Because there is real biological variation even between replicates, a good normalization method may not have the greatest reduction of variance, but will increase the ratio of between and within tissue variances. In Figure 5, we compare the variance ratio over the range of average intensity. The ratio in the subset QN group is greater than that from the complete QN over the entire range of log intensities. The loess normalization result is again very similar to that from complete QN.

All the above results are done at the feature level. Since this is an experiment on the transcriptome, we also summarized the data at feature set level, using the RMA (Irizarry et al., 2003b) method. We compared the within-, between-tissue variances and the ratio of variances again. The results are very similar to those shown in Figures 3–5, and are provided in the supplementary file.

HG-U95 platform database and control experiment. The HG-U95 platform has different probe designs from the Human Gene 1.0 ST platform. Instead of non-specific background probes, this platform includes a mismatch (MM) probe for each perfect match (PM) probe that is complementary to the target transcript. The MMs are designed to measure non-specific binding background. Since the MMs differ from the pairing PMs by only one nucleotide in the center, they respond to the target transcript to some extent (Irizarry et al., 2003a), and are not ideal background probes. However, we do expect them to reflect much less biological signal compared to the PMs.

In order to evaluate whether a normalization procedure removes the obscuring variation and retains biological variation, we identify three sets of probes. We randomly select 10 thousand PM/MM probe pairs but use only half of MMs as negative controls in SQN. SQN will force the distributions of these MMs to be identical, but not impose any particular outcome on the other 5000 MMs or the 10,000 PMs, our *testing probes*. We expect little biological variation across samples on the testing MMs, regardless of the hybridization target. The 10,000 testing PMs form the set from which we expect biological variation in real

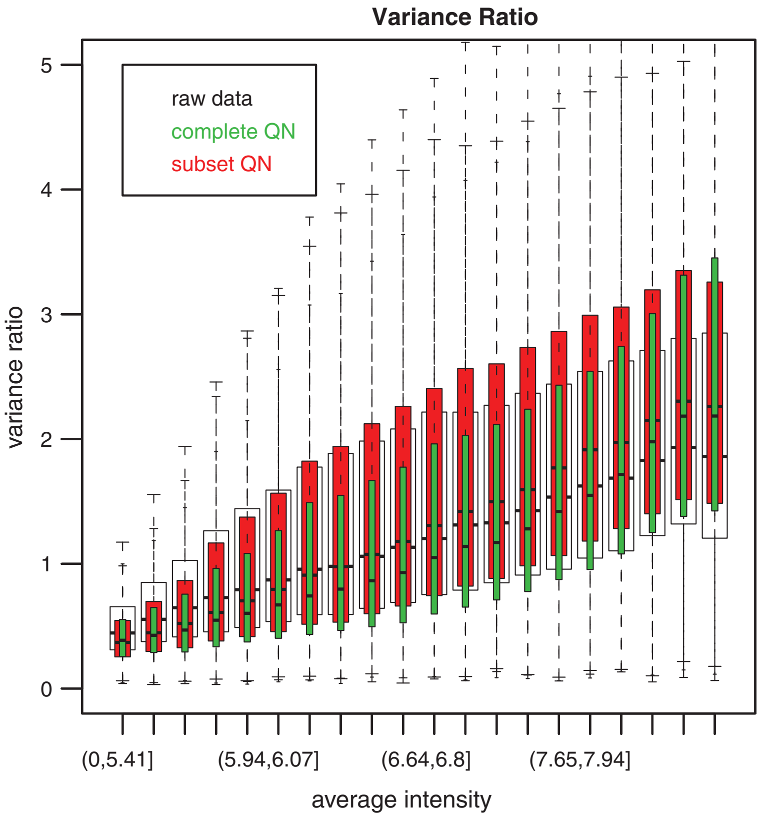

FIG. 5. The ratio of between and within-tissue variances stratified by average raw intensities. Boxplots of the variance ratio from before and after normalization are overlaid for easy comparison.

hybridization samples. However, in the three background experiments in which poly-C, poly-G or yeast DNA are used as target, we expect the PMs to be much lower and also indistinguishable from the MMs.

SQN delivers exactly what we expect biologically. Figure 6a shows a great deal of systematic variation across samples in the raw data before normalization. The open boxes are from the testing PMs and solid green boxes are from testing MMs. After SQN (Fig. 6c), the testing MMs in all samples have almost identical distribution across samples. Notice that this is not imposed by SQN since these MMs were not used as controls in SQN. The testing PMs in samples *a* to *j* show biological variation, but their distribution is identical to the MMs in the background experiments, and much lower than samples *a* to *j*. As a contrast, the distribution of the testing MMs are much more variable after complete QN (Fig. 6b). Complete QN also inflates the distribution of the MM probes in the background experiments and make them even higher than the MMs in real transcriptomes.

DNA methylation assay. DNA methylation patterns in differentiated (fibroblasts) and undifferentiated (iPS) cells are compared in 6 samples of each type in this experiment. For each sample, the genomic DNA is digested with the McrBC enzyme, which selectively cuts methylated DNA. Fragment size selection is used to enrich for unmethylated DNA. The enriched fraction is co-hybridized with a total genomic DNA fraction on a two-color CHARM DNA methylation microarray. The ratio between the enriched and total genomic fractions at each probe reflects the percentage of methylated cells at the targeted locus. Since differentiated cells and stem cells are known to exhibit global differences in methylation patterns (Lister et al., 2009) the assumptions of complete quantile normalization are violated. We identify a set of negative control probes in order to assess the obscuring variation across samples. Since the restriction enzyme McrBC binds only at CpG dinucleotides these control probes are chosen from CpG-free regions of the genome. The signal ratios at these probes are thus expected to be constant across all samples, regardless of methylation status of other probes. Using these probes as negative controls in SQN allows us to normalize the samples without forcing the global pattern of methylation to be equal. Figure 7 compares the distributions of

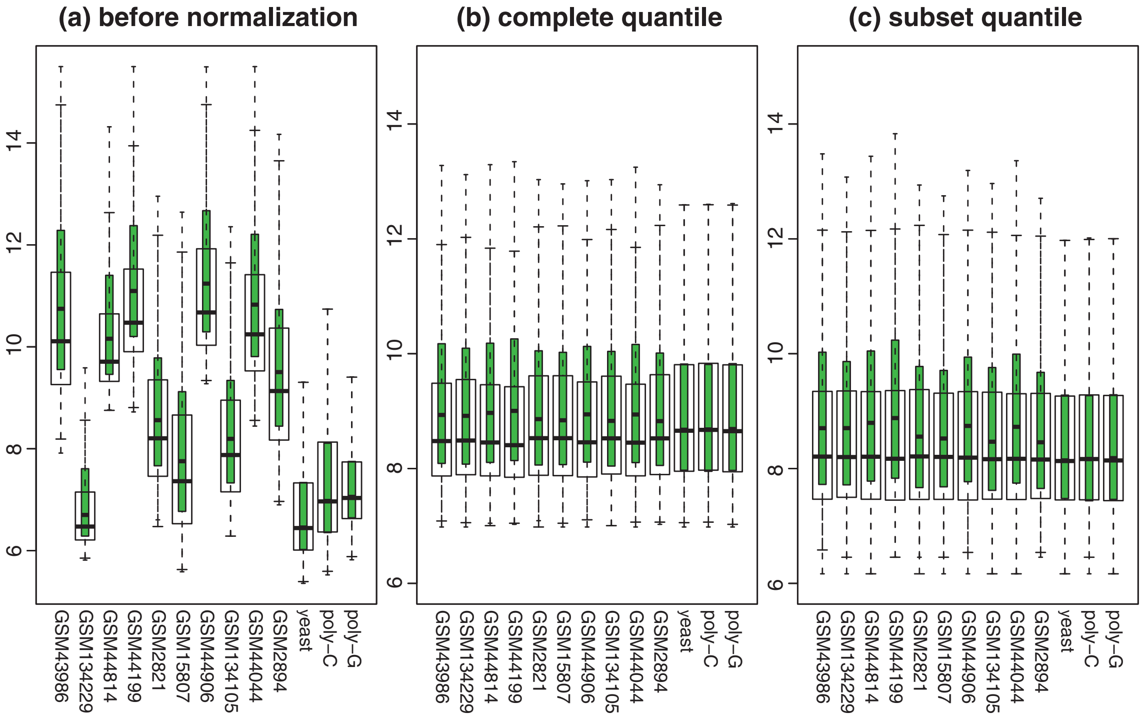

FIG. 6. Boxplots of the log intensities on testing PMs (solid green boxes) and MMs (open boxes) on a human gene expression array, HG-U95a. MMs are expected to measure non-specific binding, and thus are ideally equal in all samples. PMs are expected to measure specific binding and vary across real biological samples, but not in control samples without human RNA (samples *yeast, poly-C*, and *poly-G*).

probe log ratios before and after normalization, using complete QN or SQN. The biological variation in the methylation extent is apparent in the unnormalized log ratios (Fig. 7a), and there is clear separation between fibroblast and iPS samples. This variation between classes is washed away when a complete QN is applied (Fig. 7b). However, when we apply SQN, the variation between samples is reduced but not completely removed. Most importantly, the two classes, fibroblast and iPS, are now well separated (Fig. 7c), demonstrating a between class difference beyond the variation among biological replicates within class. To provide a more quantitave comparison, we also compute the F-statistics comparing the between-group to within-group variability for the 10,000 most variable probes. The average F statistic for the unnormalized data is 50.5. QN shows a slight increase to 52, while SQN normalization results in a dramatic increase to an average F of 147. Since class membership is blinded for the SQN procedure, this result clearly demonstrates that SQN is able to reduce the obscuring, non-biological variation due to technical issues, while retaining the meaningful biological variation.

5. DISCUSSION

In this article, we present a normalization method using a subset of negative control features. Examples of suitable controls include antigenomic features designed to measure the extent of non-specific binding in a number of microarray platforms. Although the original purpose of these features was to estimate and adjust for background, they also serve as great resource for normalization for two reasons. First, they are not expected to hybridize to the target genome or transcriptome, thus the observed intensities on these features reflect the systematic variation that we hope to remove, and not the biological signal that we hope to preserve. Second, empirical observation reveals that intensities from these features span almost the entire range of array intensities, so that we do not have to extrapolate much in the normalization. Using these probes as a basis for normalization allows us to avoid making assumptions about the biological signal behavior in various samples. This is especially useful in situations when we are not comfortable with the usual assumptions such as that the majority of genes do not have significant differential expression or symmetric

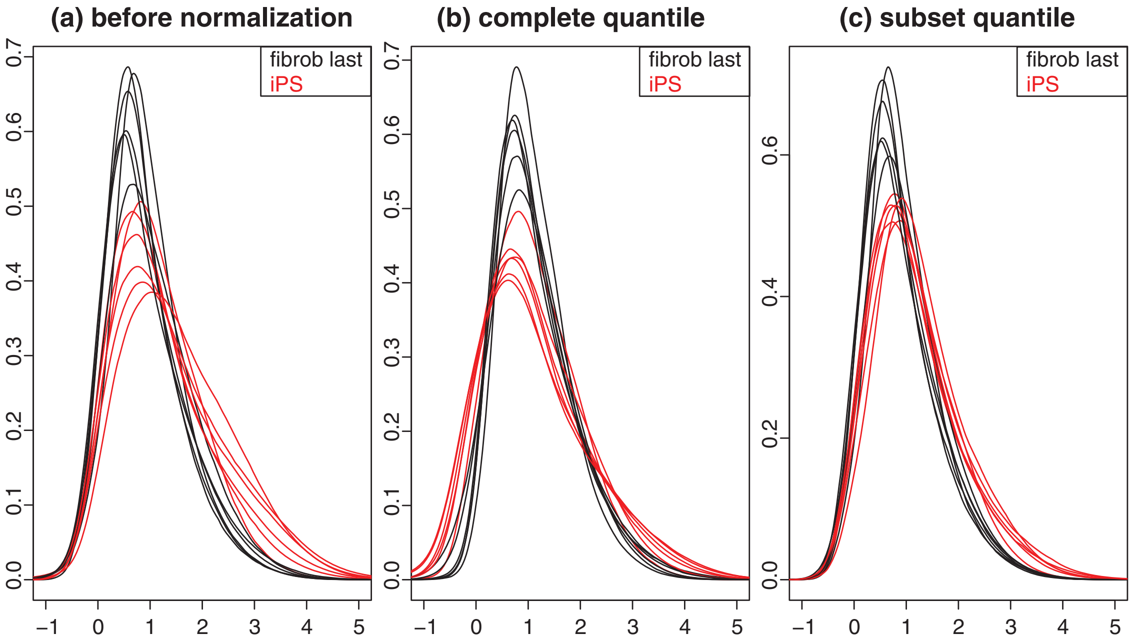

FIG. 7. Probability density functions of M-values in 6 fibroblast and 6 iPS samples. **(a)** Before normalization. There is considerable within-group variation in the iPS samples. Average F-statistic comparing the two groups among the 10,000 most variable probes is 50.5. **(b)** After complete quantile normalization. Both within group and between group variation are reduced. Average F-statistic is 52. (**c**) After SQN. Within group variation is reduced while the fibroblasts and iPS samples are well seperated. Average F-stastitics increased to 147.

up- and down-regulation. Examples include radiation experiments that disrupt the transcription of a large fraction of genes (Rea et al., 2003) and asymmetric expression regulation under stress (Weber et al., 2006).

We have also presented an example of using specially designed control features in a DNA methylation assay. The negative controls in this case are probes where we expect a constant ratio across all samples, regardless of methylation status. The flexibility of SQN allows us to normalize without making assumptions about overall levels of methylation.

Microarray technology has become a routine technique in biomedical research. Scientists are constantly designing novel experiments and as a result many computational methods originally developed for gene expression assays may no longer be appropriate for these new applications. Normalization procedures are one such example. Our proposed method, SQN, allows flexible normalization based on the distribution of a group of negative control probes and can thus be applied to a wide variety of experiments without making assumptions on the biological signal. Another benefit of SQN is that we no longer require the number of signal probes to be the same across samples. This makes it easier to handle missing values, or even combine data from different platforms, as long as the same set of control features are used. This flexibility of subset QN also allows it to be applied to other high throughput technologies beyond microarrays, as long as a subset of control features can be identified.

ACKNOWLEDGMENTS

We thank Thomas A. Louis for helpful discussion regarding the estimation of reference distribution. We thank Rafael Irizarry for helpful comments and for sharing the microarray data from background experiments. This publication is supported by NIH grant 5R01GM083084.03.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Amaratunga, D., and Cabrera, J. 2001. Analysis of data from viral DNA microchips. *J. Am. Statiss. Assoc.* 96, 1161–1170.

Bird, A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev.* 16, 6–21.

Bolstad, B., Irizarry, R., Åstrand, M., et al. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinfromatics* 19, 185–193.

Doi, A., Park, I.-H., Wen, B., et al. 2009. Differential methylation of tissue- and cancer-specific CPG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat. Genet.* 41, 1350–1353.

Fraley, C., and Raftery, A. 2009. *mclust: Model-Based Clustering/Normal Mixture Modeling*. R package, version 3.2.

Gautier, L., Cope, L., Bolstad, B.M., et al. 2004. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics* 20, 307–315.

Irizarry, R.A.B., Hobbs, F.C., Beaxer-Barclay, Y., et al. 2003a. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264.

Irizarry, R.A.B., Hobbs, F.C., Beaxer-Barclay, Y., et al. 2003b. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264.

Irizarry, R.A., Ladd-Acosta, C., Carvalho, B., et al. 2008. Comprehensive high-throughput arrays for relative methylation (charm). *Genome Res.* 18, 780–90.

Li, C., and Wong, W. 2001. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* 2, 0032–0031.

Lister, R., Pelizzola, M., Dowen, R.H., et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 416, 315–322.

Mchler, M. 2007. *nor1mix: Normal (1-d) Mixture Models (S3 Classes and Methods)*. R package, version 1.0-7.

Rea, M., Gregg, J., Qin, Q., et al. 2003. Global alteration of gene expression in human keratinocytes by inorganic arsenic. *Carcinogenesis* 24, 747.

Smyth, G.K. 2005. Limma: linear models for microarray data, 397–420. In Gentleman, R., Carey, V., Dudoit, S., et al., eds. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York.

Thellin, O., Zorzi, W., Lakaye, B., et al. 1999. Housekeeping genes as internal standards: use and limits. *J. Biotechnol.* 75, 291–295.

Tseng, G., Oh, M., Rohlin, L., et al. 2001. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* 29, 2549–2557.

Weber, C., Guigon, G., Bouchier, C., et al. 2006. Stress by heat shock induces massive down regulation of genes and allows differential allelic expression of the Gal/GalNAc lectin in *Entamoeba histolytica*. *Eukaryotic Cell* 5, 871–875.

Wu, Z., Irizarry, R., Gentlemen, R., et al. 2004. A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Statist. Assoc.* 99, 909–917.

Yang, Y., Dudoit, S., Luu, P., et al. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 30, e15.

Yang, Y.H.J., Paquet, A., and Dudoit, S. 2007. *marray: Exploratory Analysis for Two-Color Spotted Microarray Data*. R package, version 1.22.0.

Address correspondence to:
Dr. Zhijin Wu
Center for Statistical Sciences and Department of Community Health
Brown University
121 South Main Street
Providence, RI 02912

Email: zhijin_wu@brown.edu