An Integrated Approach for RNA-seq Data Normalization

Shengping Yang[1,2], Donald E. Mercante[2], Kun Zhang[3] and Zhide Fang[2]

[1]Department of Pathology, School of Medicine, Texas Tech University Health Sciences Center, Lubbock, TX, USA. [2]Biostatistics Program, School of Public Health, LSU Health Sciences Center, New Orleans, LA, USA. [3]Department of Computer Science, Xavier University of Louisiana, New Orleans, LA, USA.

ABSTRACT

BACKGROUND: DNA copy number alteration is common in many cancers. Studies have shown that insertion or deletion of DNA sequences can directly alter gene expression, and significant correlation exists between DNA copy number and gene expression. Data normalization is a critical step in the analysis of gene expression generated by RNA-seq technology. Successful normalization reduces/removes unwanted nonbiological variations in the data, while keeping meaningful information intact. However, as far as we know, no attempt has been made to adjust for the variation due to DNA copy number changes in RNA-seq data normalization.

RESULTS: In this article, we propose an integrated approach for RNA-seq data normalization. Comparisons show that the proposed normalization can improve power for downstream differentially expressed gene detection and generate more biologically meaningful results in gene profiling. In addition, our findings show that due to the effects of copy number changes, some housekeeping genes are not always suitable internal controls for studying gene expression.

CONCLUSIONS: Using information from DNA copy number, integrated approach is successful in reducing noises due to both biological and nonbiological causes in RNA-seq data, thus increasing the accuracy of gene profiling.

KEYWORDS: DNA copy number alterations, RNA-seq, normalization

CITATION: Yang et al. An Integrated Approach for RNA-seq Data Normalization. *Cancer Informatics* 2016:15 129–141 doi: 10.4137/CIN.S39781.

TYPE: Original Research

RECEIVED: March 29, 2016. **RESUBMITTED:** May 12, 2016. **ACCEPTED FOR PUBLICATION:** May 30, 2016.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Four peer reviewers contributed to the peer review report. Reviewers' reports totaled 794 words, excluding any confidential comments to the academic editor.

FUNDING: ZF's research was supported by 1 U54 GM104940 from the National Institute of General Medical Sciences of the National Institutes of Health that funds the Louisiana Clinical and Translational Science Center of Pennington Biomedical Research Center. KZ was supported by NIH NIMHD-RCMI grant # 2G12MD007595 and the DOD ARO grant # W911NF-15-1-0510 and the Louisiana Cancer Research Consortium (LCRC). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: zfang@lsuhsc.edu; kzhang@xula.edu

Background

Next-generation sequencing (NGS) technologies have been widely used in various areas of genetic studies, such as the use of RNA-seq technology in detecting differentially expressed genes and measuring genome-wide gene expression profiles[1] and the use of DNA-seq technology in detecting single-nucleotide polymorphisms (SNP) and copy number alterations (CNAs).[2]

DNA stores genetic information that can be transcribed into RNA, and subsequently RNA may be translated into proteins that are directly involved in biological activity. Thus, DNA and gene expression data are often used for investigating whether normal biological processes are altered and how these alterations are associated with various disease conditions such as tumorigenesis. In order to obtain a holistic view of the interrelationship between DNA, RNA, and protein, many studies have incorporated DNA-, RNA-, and protein-level experiments to better understand the complex nature of genetic diseases such as cancers. The results thus far are promising.[3]

DNA point/structural aberrations and RNA differential expression are very often found to be associated with diseases such as cancers. Specifically, DNA CNAs can alter the expression levels of genes that are associated with genetic disorders. For example, the deletion of tumor suppression genes or the amplification of oncogenes can lead to expression-level change in these genes and subsequent diseases.[4–6] In a study of B-progenitor acute lymphoblastic leukemia, Mullighan et al.[7] found that about 40% of patients had DNA sequence deletions and/or amplifications, indicating a possible association between CNAs and acute lymphoblastic leukemia. Therefore, it is natural to assume that many of the genes associated with tumorigenesis are harbored within the CNA regions. On the other hand, studies have also revealed that, while epigenetic factors and other modifications of the genome attribute to a minor proportion of the changes in gene expression,[8] about 15% of the variations in gene expression can be explained by CNAs.[9]

A common approach used in the analysis of gene expression data is to interrogate whether individual genes behave differently for subjects receiving treatment or control or between subtypes of a disease. Although both expression array and RNA-seq platforms are widely used for generating expression data, the latter is more powerful in that it can

also pick up novel microRNAs that have not yet been incorporated into arrays in microRNA sequencing.[10,11] However, analytically, it remains a challenge to accurately identify differentially expressed genes from RNA-seq data. For example, the lists of top differentially expressed genes identified using various software packages may not agree in many cases, and a lower type I error rate is usually accompanied with a decreased statistical power. While controlling the false-positive/discovery rate[12–14] is one important component for proper expression data analyses, appropriate normalization is indispensable to ensure accurate and reliable results.[15]

Many microarray-based gene expression studies have demonstrated that normalization is essential for proper analysis of expression data. Normalization ensures that signal intensities are properly centralized and that differences in signal intensities are due to biological differences between samples, rather than nonbiological noise.[16,17]

RNA-seq data, generated by the *sequencing-by-synthesis* procedure, are fundamentally different from microarray data. Read counts generated from NGS have a discrete distribution and are commonly considered to be *digital*. Thus, directly adopting the normalization methods developed for log-normally (continuously) distributed microarray data may not be the best strategy.

Many normalization methods have been developed for RNA-seq data to mitigate the between/within-sample differences, such as those in library size (sequencing depth) and GC content.[18,19] Some authors propose scaling the data so that the total read counts are the same across samples.[20–22] Others directly adopt the microarray data normalization methods, such as quantile normalization (QN) on the log-transformed, length-standardized read counts.[23] Alternatively, Mortazavi et al.[18] proposed using the per kilobase per million mapped method to adjust for read count bias due to gene-coding region length. Robinson and Oshlack[15] employed an empirical strategy to equate the overall expression levels of genes among samples based on the commonly made assumption that the majority of genes are not differentially expressed. Specifically, a weighted trimmed mean of the log expression ratios is used to estimate the ratio of RNA productions so that a large number of genes that are highly expressed in one condition can be properly adjusted. In addition, Hansen et al.[24] proposed using conditional QN to remove technical variability. We comment that these methods do not take into account the effect of DNA copy number on the gene expression, and thus, almost every method has obvious limitations when dealing with samples with DNA copy number changes.

CNA is a hallmark of cancer. Studies have shown that gene expressions are correlated with DNA copy number.[25,26] DNA copy number change alters the number of DNA templates from which genes can be transcribed and, thus, directly affects the corresponding expression level. To properly normalize expression data, ideally, the global average of the expression levels of genes with the same copy number should be the same across samples, given that differentially expressed genes account for only a small percentage of all genes interrogated. In fact, the majority of current RNA-seq normalization methods assume that the whole genome has a copy number of two (except for the sex chromosomes) by default. While this assumption is acceptable for samples with no structural genetic alterations, it would not work well for data generated from cancer samples, which usually have CNAs.

In fact, since gene expression is downstream of DNA alteration and if gene expression levels can be accurately measured, then the underlying mechanisms (for example, CNAs and point mutations) are not relevant. However, we cannot assume by default that the expression values of individual genes obtained by sequencing/microarray technology are accurate, before performing appropriate normalization, for example, adjusting for CNAs.

As far as we know, gene expression measurements are all relative. In general, some reference values are used to ensure that the actual expression level of a gene (per cell) is proportional to the measured value, for example, read count is used for gene expression in RNA-seq. However, if such reference values are not correctly chosen, then the proportion between the actual expression level and the measured value will not be consistent across samples, and thus, gene expression cannot be considered to be accurately measured. Consequently, all the downstream analyses will become problematic. Note that accurately controlling the number of cells used in the laboratory experiments might be an alternative for improving data normalization; however, this is not the focus of this study.

For the samples with CNAs, choosing such reference values is not straightforward in DNA data normalization.[17,27] In fact, it is equally challenging in RNA data normalization – we will demonstrate this by showing the existence of CNA-oriented correlation between DNA copy number and gene expression in the "Methods" section.

With the rapid development of biotechnology, an increasingly larger number of contemporary genomic studies collect data generated by multiple platforms simultaneously. Today, many experiments interrogating DNA, RNA, and protein activities are being carried out on samples from the same group of subjects, which thus provide opportunities for statisticians to apply integrated data analysis strategies to improve the accuracy and reliability of the statistical tests. Specifically, in cancer genetic studies, statistical methods have been proposed to integrate analysis of copy number and gene expression data, with the belief that more genetic information could be revealed by incorporating complementary information from DNA and RNA data together into a statistical model.[28,29]

Using either a two-step approach or an integrated approach, many investigations have demonstrated improved power for detecting DNA and RNA variations that would not otherwise be detected using either DNA data or RNA data alone. Commonly used approaches very often include the following steps: (1) identifying the regions with CNAs,

(2) calling the copy number status of each CNA region, and (3) comparing the expression levels of genes in the CNA regions to obtain a list of differentially expressed genes for further investigations.[28,29] Many test statistics have been proposed, and generally speaking, they are often based on either correlation or externally centered correlation between CNA and gene expression.[25,26,30,31] Although these approaches are mainly designed to partition regions with strong, equally directed abnormalities from other regions, none of them addresses the possibility of using the available laboratory obtained DNA data, that is, addition information, to improve RNA data normalization.

In this article, we propose an integrated RNA-seq normalization method, referred to as integrated normalization, which takes advantage of the availability of DNA-seq data to appropriately normalize RNA-seq data. In the next section, we describe the proposed method. Then, we utilize a publicly available dataset to demonstrate how the normalization method affects DE gene detection as well as RNA expression profiling results. At the end of this article, we discuss the advantages and limitations of the proposed method and possible improvements.

Methods

Distribution of the ratios of sample read counts. Suppose that we have both the DNA-seq and RNA-seq data generated from each patient who has one of two subtypes of a cancer. Specifically, for DNA, preferably, we have both the tumor and the matched normal tissue samples for each patient; otherwise, we can generate a generic control by averaging the data of several unrelated samples. Note that having paired tumor and germline samples would be ideal but is not a requirement. However, for RNA, we have the expression data for each tumor sample and a generic control, which is generated by averaging the expressions of a few matched normal tissues.

Denote x_{ij}^{C1} (x_{ij}^{C2}) as the RNA-seq read count for the ith gene of the jth tumor sample in condition C1(C2) and $N_{ij}^{C1}(N_{ij}^{C2})$ as the sum of read counts from the tumor sample and an appropriately matched normal tissue sample. Similarly, denote $y_{ij}^{C1}(y_{ij}^{C2})$ and $M_{ij}^{C1}(M_{ij}^{C2})$ as the corresponding DNA-seq read count and the sum of read counts for condition C1(C2). We define the following ratios:

$$r_{ij}^{C1}=\frac{x_{ij}^{C1}+u_1}{N_{ij}^{C1}+u_1+u_2},\quad r_{ij}^{C2}=\frac{x_{ij}^{C2}+u_3}{N_{ij}^{C2}+u_3+u_4},$$
$$d_{ij}^{C1}=\frac{y_{ij}^{C1}+u_5}{M_{ij}^{C1}+u_5+u_6},\quad d_{ij}^{C2}=\frac{y_{ij}^{C2}+u_7}{M_{ij}^{C2}+u_7+u_8},$$

where u_is ($i=1,2,3,4,5,6,7,8$) are random variables with uniform distribution (0, 1). These ratios will be used in the proposed normalization procedure. Note that in the rest of this article, we drop the subscription j for notational simplification.

Note that the above constructed variables are the ratios of modified read counts and take values between 0 and 1. Thus, when x_{ij}s and y_{ij}s are large, under the assumption that, for the ith gene, the read counts of the jth tumor and the control sample(s) are independent, it can be shown that each ratio approximately follows a beta distribution. Specifically, we assume that r_i^{C1}, r_i^{C2}, d_i^{C1}, and d_i^{C2} have the following beta distributions, $B(r_i^{C1};\alpha_i^{C1},\beta_i^{C1})$, $B(r_i^{C2};\alpha_i^{C2},\beta_i^{C2})$, $B(d_i^{C1};u_i^{C1},v_i^{C1})$, and $B(d_i^{C2};u_i^{C2},v_i^{C2})$, respectively, with parameters α^g, β^g, u^g, v^g, where $g\in(C1,C2)$. In addition, with the assumption that observations in nearby genome locations approximately follow the same distribution, we can, due to the central limit theory (CLT), approximate the distribution of the average ratio by a normal distribution. That is,

$$\frac{1}{m}\sum_i r_i^g \overset{d}{\approx} N\left(\pi^g,\frac{\sigma_g^2}{m}\right),$$
$$\frac{1}{m}\sum_i d_i^g \overset{d}{\approx} N\left(\rho^g,\frac{\tau_g^2}{m}\right),$$

where $g\in\{C1,C2\}$, m is the number of observations to be averaged, and π^g (ρ^g) and σ_g^2 (τ_g^2) are the mean and variance of the ratios r_i^g (d_i^g).

It is evident that a beta distribution is approximately symmetric if the two parameters are close. Also, if the two parameters of a beta distribution are both greater than 1, then the beta distribution is unimodal (see Supplementary File 1 for details). In this case, even for a small window size m, the approximation based on the CLT works considerably well (see Supplementary Fig. 1 for details). Furthermore, we suggest adding an independent random number with a uniform distribution to both tumor and control read counts to ensure the read count ratios approximately have a beta distribution when both read counts are small.[32] In fact, in many situations, the paired read counts are both large, and thus, adding a random uniform variable to each of them has very little effect on the estimation of the ratio. In addition, for locations where both tumor and control samples have zero count, we claim that these locations carry little/no information and will not be included in the analysis.

Integrated RNA-seq data normalization. By assuming that differentially expressed genes only account for a small proportion of the genes interrogated, QN has been successful for microarray-based expression data normalization. However, QN might not be the best choice for normalizing DNA samples with substantial (whole chromosome/arm amplifications/deletions) CNAs.[17] In fact, it has been shown that for samples with CNAs, using QN for microarray DNA data can generate unwanted results, that is, the average log ratio of the regions without CNAs does not align to 0. Several methods have been proposed to address this problem such that alignment can be appropriately performed.[17,27,33,34] Motivated by these methods, we propose an integrative approach for RNA-seq data normalization in this article.

Given the availability of both DNA-seq and RNA-seq data in many genomic studies, our proposed method includes

two steps. First, identify genomic regions that do not have CNAs using DNA-seq data. Second, perform RNA-seq data normalization using the regions without CNAs as references. We comment that, compared to DNA-seq data, RNA-seq data exhibit far greater variability due to switching between the *on* and *off* expression of certain genes. Thus, genomic regions that do not have CNAs can be more precisely identified using DNA-seq data. Note that many studies have shown that CNAs substantially affect regional gene expression, and the key to successfully implement our proposed method is to effectively remove the variation in gene expression associated with DNA CNAs to ensure that reference values used for RNA-seq data normalization are consistent across samples.

An operational sequence of integrated RNA-seq normalization. In Figure 1, we present a flow chart that outlines the operational sequence in the proposed procedure.

Identification of regions without CNAs using DNA-seq data. We propose using the following steps to identify genomic regions without CNAs.

A. Calculate the averages of d_{ij}^{C1} and d_{ij}^{C2} (the subscript j and the superscript C1/C2 will be dropped for notational simplification purpose), separately, in a fixed window of size s (consecutive observations) to reduce the variance of the raw data (see Supplementary Fig. 2 for details). We name this variable mean$d$. That is,

$$\text{mean}d_p = \frac{1}{s}\sum_{k=(p-1)s+1}^{\min(ps,T)} d_k, \quad p = 1,2,\ldots,\lceil T/s \rceil, \qquad (1)$$

where $\lceil\ \rceil$ is the ceiling of a number, T is the total number of genes, and p is the index of each window. We will use s = 10 as default unless otherwise stated. The variable mean$r$, the average of r, is defined in the same fashion. Note that since our goal is to identify the reference genomic regions, the resolution of CNAs detection is not a main concern.

B. Apply a modality test[35] on mean$d$. Note that if there are no CNAs in the whole genome, then mean$d$ is supposed to have a unimodal distribution; otherwise, it would have a multimodal distribution. Unimodal and a series of multimodal model fits will be compared based on AIC.[36] If the multimodal model has a better fit, we will estimate the mean and variance of every component distribution (see Supplementary Fig. 3 for details).

C. Initially assign the two-copy state to the component distribution harboring regions without CNAs. This step aims at roughly identifying the regions that do not have CNAs. Although a modality test would not provide the exact locations of these regions, it provides initial estimates of the mean of these regions, so that the subsequent analysis can use them as the references. In fact, there are two ways to obtain the initial two-copy state information: (1) If laboratory data such as cytogenetic data are available, regions that do not have CNAs can be identified straightforwardly, and the mean of the component distribution harboring these regions will be set to 0.5. This will ensure that in the following CNA detection, regions without CNAs can be identified and used as the references. (2) If laboratory data are not available, we can assume that the component distribution that contains the majority of genes reflects the regions without CNAs. Note that the second option is simple but less optimal because sometimes, CNAs occur in the majority of the genome, for example, triploid or tetraploid samples. However, should reference regions be incorrectly identified for some samples, they are very likely to be detected in the

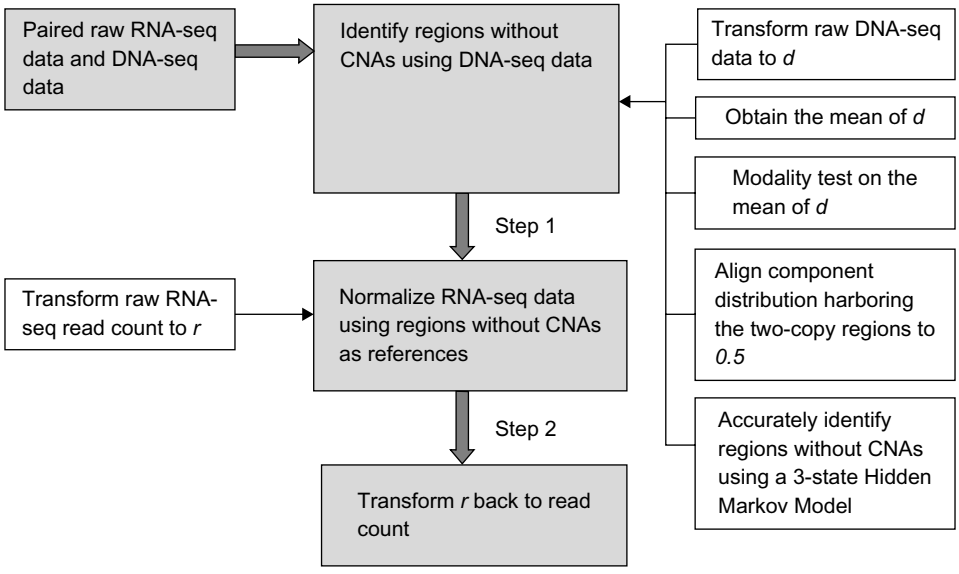

Figure 1. A flow chart of the proposed integrated normalization method.

laboratory validation step, and appropriate correction can be made. In fact, there is a third option, in which an extra step of genotype calling using DNA-seq data is needed. A component distribution with heterozygous SNP proportion less than 15% is referred to as having loss of heterozygosity. The component distribution, which is next to it and has a higher mean, represents the regions that do not have CNAs.[17]

D. Identify regions without CNAs using a hidden Markov model (HMM).

We have previously developed a method, with accompanying software called PAIR[33,37] for SNP array data normalization using a two-state HMM.[33] The key component of PAIR is to identify regions without CNAs. Note that in many HMM-based CNA detection algorithms, parameters for each hidden state need to be estimated iteratively. However, the accuracy of parameter estimation largely depends on whether every possible hidden state has been defined. Since CNAs are sample specific and cannot be predicted in advance, any method for CNA detection in a large number of samples will require a large number of hidden states, resulting in a heavy computation burden. As an alternative, in PAIR, we proposed an HMM with only the following two states: regions that harbor SNPs without CNAs (two-copy regions) and regions that do not. By setting up upper and lower bounds for the two-copy regions, we can accurately partition the regions with and without CNAs, while the parameters for the regions that harbor various forms of CNAs do not have to be estimated individually.

In this article, we adopt the same principle as that in PAIR in identifying regions without CNAs, while making the following modifications.

i. Use mean$d$ instead of d in the HMM. The reasons are twofold: (1) by doing so, the normality assumption is approximately satisfied by the CLT and (2) though using the average of a fixed number of observations may reduce the resolution for detecting a change point, it has minimal effect on the normalization process.
ii. Instead of partitioning mean$d$ into a unimodal or bimodal state (part I of PAIR), we propose to use a HMM with the following three states: copy number gain, two copy, and copy number loss. There are two reasons for this modification: (1) SNP array data have log intensity readings for both A and B alleles, and taking the difference between them can partition all SNPs into unimodal and bimodal states. DNA-seq data consist of read counts for every genomic locations, not just the SNP locations, thus, adding read counts within fixed size windows utilizes information on all these locations. (2) DNA-seq data have much larger dynamic range; thus, regions with different copy numbers can be more easily separated.
iii. Use an iterative approach to estimate the mean of mean$d$ for regions without CNAs. Specifically, we set the iteration,

$$\mu_0^{(i+1)} = \frac{\sum_{w=1}^{n} d_w I_{\mathrm{CN}_w^{(i)}=2}}{\sum_{w=1}^{n} I_{\mathrm{CN}_w^{(i)}=2}},$$

where $\mathrm{CN}_w^{(i)}$ is the copy number state of the wth observation at iteration i. The calculation is repeated until the difference in means between two consecutive iterations is less than 0.001. This iterative approach is necessary for accurately assigning a copy number state to an observation because the initial mean value estimated from the modality test is often slightly different from the estimated mean in the HMM.

RNA-seq data normalization. Using the regions without CNAs detected by the procedure in the proceeding subsection as references, the normalization process for RNA-seq data can be carried out through the following steps.

A. Align the mean of the ratios, r_is, of RNA-seq counts that correspond to the regions without CNAs to 0.5 by iteratively updating r_i,

$$r_i^{\mathrm{update}} = r_i - \left(\bar{r}_{\mathrm{CN}=2} - 0.5\right)\left(1 - 2^{|2r_i-1|-1}\right),$$

until the average r_i^{update} of the regions without CNAs equals to 0.5. Note that the magnitude of the adjustment is close to the numerical difference when r_i is close to 0.5 and close to 0 when r_i is close to either 0 or 1 (see Supplementary Fig. 4 for details). Therefore, all updated values would not go beyond the support of a beta distribution (0, 1).

B. Adjust tumor read counts based on the updated r_i:

$$x_i^{\mathrm{update}} = \left\| \frac{(N_i - x_i) r_i^{\mathrm{update}}}{1 - r_i^{\mathrm{update}}} \right\|,$$

where $||\cdot||$ represents the closest integer of a number.

Compare the DE gene detection power before and after normalization. Bioconductor package DESeq was used to detect DE genes, as well as perform expression profiling. Default parameters were used for analyzing the raw data, the DESeq normalized data, and the integrated normalized data, except that we forced the size factor (the coverage of a specific library)[38] for all samples to be 1 when the raw and integrated normalized data were analyzed.

The housekeeping genes. In order to understand the effect of CNA on GE, as well as how copy number change affects the expression of housekeeping genes, we downloaded the human housekeeping genes from a public database[39] and utilized the online Clone/Gene ID converter[40] to obtain gene names. In the end, we obtained approximately

460 housekeeping genes from this database. Spline smoothing was performed on the housekeeping gene expression data, and the correlation between the expressed levels of these genes and the corresponding DNA read count ratios was calculated.

Results and Discussion

In this section, we demonstrate the performance of the proposed integrated normalization method by applying it to a public dataset. Both RNA-seq and DNA-seq data for eight breast cell lines, including seven from breast cancer and one from nontumor breast epithelial cells, are available for download from the public domain Gene Expression Omnibus with access ID, GSE27003.

Four of the seven cancer cell lines were established from estrogen receptor positive (ER+) and the other three from ER– breast tumors. Both DNA and RNA of these cell lines were extracted from mid-log phase populations of low passage number cultures, as described in Ref. 41.

RNA-seq and DNA-seq data preparation. *RNA-seq data preparation.* Short reads were aligned and annotated using the pipeline described in Ref. [41]. Specifically, Illumina's alignment tool, Eland_RNA, was used to align the short reads to genome and exon junctions. The aligned sequence tags were summarized and annotated using Illumina's CASAVA tool. As a result, a total of 18,517 genes were annotated, and the total read counts for these genes were used for the downstream analysis.

DNA-seq data preparation. The 50 base pair paired-end short reads from Illumina Genome Analyzer were aligned by Bowtie.[42] A maximum of two mismatches were allowed in each alignment, and reads mapped to multiple genomic locations were discarded. SAMtools was used to generate the BAM format output files,[43] and subsequently read count for each location was calculated using HTseq-count.[44] As a result, a total read count within the start and end locations of each of 18,517 annotated genes was calculated.

The correlation between GE on DNA copy number. We calculated the correlation coefficients (CCs) between the RNA-seq and DNA-seq data using the variables introduced in the "Methods" section. The range of the CC is from 0.100 to 0.186, indicating a weak relationship. However, considering that the random variation in signal intensity (especially from RNA-seq data) might have disguised the true association, we performed circular binary segmentation (CBS)[45] on both DNA-seq and RNA-seq data using the average of four observations to reduce variation, as well as to meet the normal assumption. Then, we replaced the original mean$r$/mean$d$ values in the same segment with the corresponding segment mean values. The range of CC calculated from these segment mean values increases substantially to (0.247, 0.761), indicating that DNA copy number does play an important role in gene expression. This is the motivation of the proposed method.

Figure 2 shows that the segmentation patterns are very similar between RNA-seq and DNA-seq data (Fig. 2a and b). For a comparison, we performed the same analysis on other variables constructed by a commonly used transformation, the log ratio between tumor and control read counts, and observed much lower correlation (Fig. 2c and d). The correlation between DNA copy number and gene expression has been evaluated in other studies. This correlation is comparatively weak, at round 10%–20%, in most of these studies. However, by applying the proposed variable transformation/construction, that is, constructing r and d, as well as a fixed window average to reduce variation, the correlation increases dramatically. Supplementary Figure 5 demonstrates that comparing r/d with log ratio, the former magnifies the difference in the middle range of the data, and this increases the power to detect differential expression with lower magnitude.

Identification of two-copy DNA regions using DNA-seq data. The high correlation described in the previous section demonstrates the necessity and advantage of incorporating the information on CNA in the RNA-seq data normalization.

More specifically, a three-state HMM was applied to identify regions without CNAs using the DNA-seq data. Here, we did not use CBS for segmentation. While CBS can accurately identify change points in a sequence, it is not designed for automatically assigning segments with the same copy number state to a specific copy number. In the HMM, we defined the three hidden states as copy number loss, no CNAs, and copy number gain. Based on the first-difference[46] estimated variance, we set upper and lower bounds for the two-copy regions by simulation (details in Ref. 33).

For the initial values of the means of the regions without CNAs in the HMM, we used the copy number status confirmed in Ref.[41] Note that this type of copy status information can be obtained if cytogenetic data are available. Specifically, we obtained the copy number status of two regions (regions with or without CNAs) in chromosomes 8 and 17 separately and identified the component distributions that harbor these two regions. We then set the initial mean of the component distribution harboring the regions without CNAs to 0.5. After applying the HMM, the copy number states of these regions were compared with those in Ref. 41 to ensure consistency. As an example, we presented the identified regions without CNAs in an ER– cell line sample in Figure 3. It was confirmed in Ref.[41] that this sample has a copy number loss in the region from 125,504,248 to 126,521,417 on chromosome 8, and no CNAs on chromosome 17 from 44,246,133 to 63,413,540. This is consistent with Figure 3 (see two yellow vertical bars in Fig. 3b), indicating that our segmentation results are consistent with the lab results.

Note that, if the lab results are not available initially, we will assume that the largest component distribution harbors the two-copy regions. If this assumption is found to be incorrect during the lab validation process, we can reset the initial value and repeat the process.

Normalization of RNA-seq signals. Based on the fact that RNA-seq and DNA-seq segment means are correlated, and the assumption that DE genes only account for a small

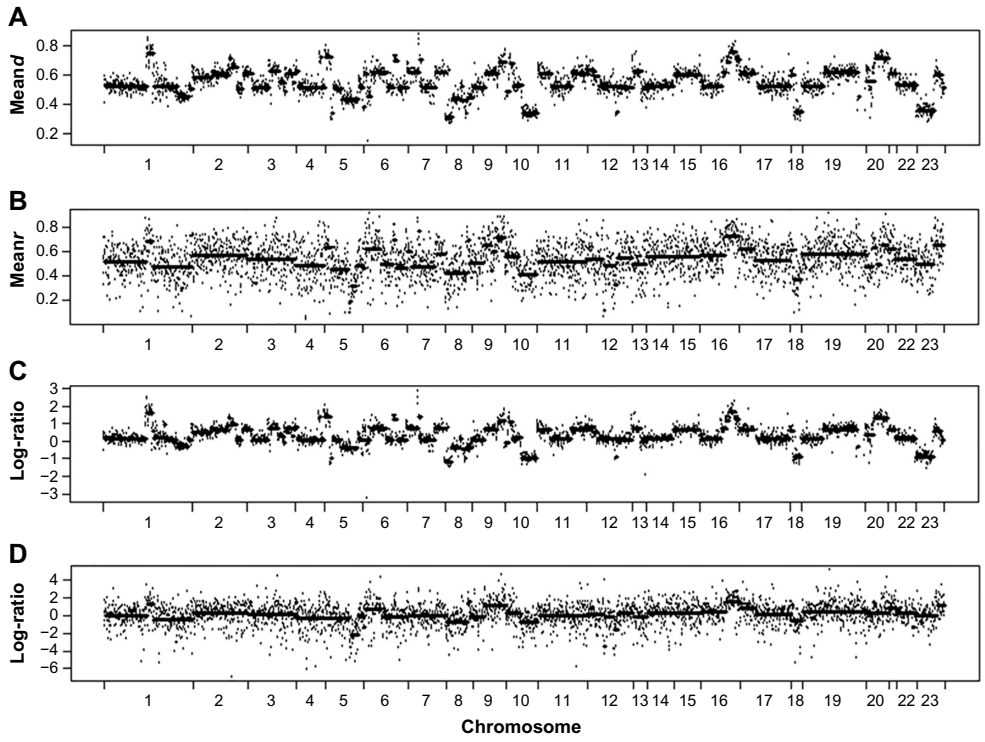

Figure 2. Comparison of the correlation between DNA-seq and RNA-seq data, using proposed variables *d* and *r* and log ratio separately. Average data, over a fixed window of size 4 consecutive genes, were used. CBS segmentation results were superimposed in bold lines. (**A**) For DNA-seq data, the ratio *d* was used. (**B**) For RNA-seq data, the ratio *r* was used. (**C**) For DNA-seq data, log ratio of tumor and control signal intensities was used. (**D**) For RNA-seq data, log ratio of tumor and control signal intensities was used.

proportion of all the genes interrogated,[15,47] we normalized the RNA-seq data such that the *r* values of genes in the regions without CNAs have a mean close to 0.5. In other words, given that the majority of genes in the regions without CNAs are not differentially expressed, they should have similar expression levels in both tumor and normal cells. The normalized *r* values were then transformed back to read counts as described in the "Methods" section for the downstream DE gene detection. Note that the generic control sample was only used as the baseline for constructing the variables we have proposed.

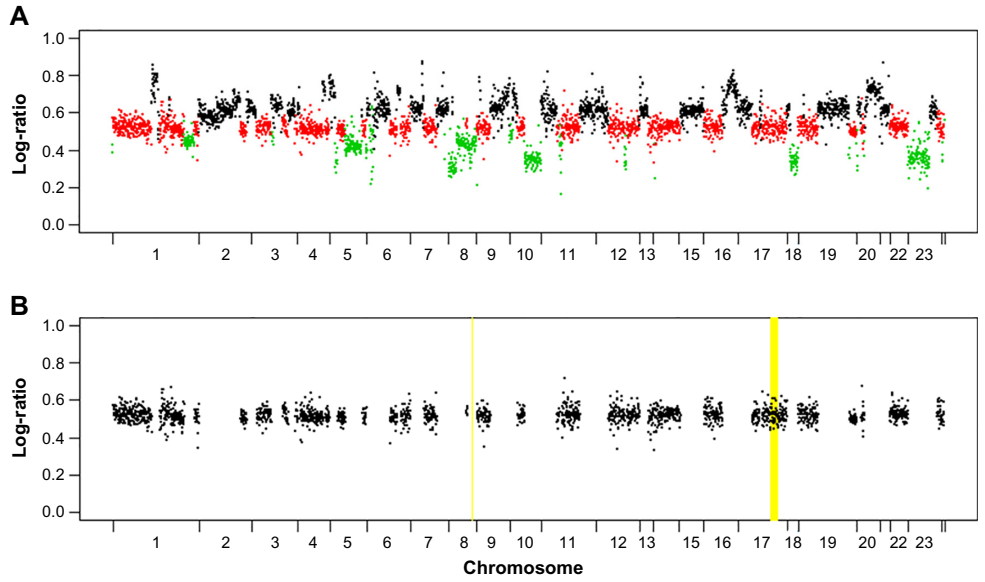

Figure 3. Identify two-copy regions using a three-state HMM. (**A**) Regions in red are those genes in two-copy state, regions in black are those in copy number gain state, and regions in green are those in copy number loss state. (**B**) Regions in two-copy state were highlighted in black. Note that we did not *normalize* DNA-seq data in this step, but instead, we identified two-copy regions. The two yellow vertical bars are the two regions mentioned in the text.

After the normalization process, DE gene detection will be performed using the normalized read count data of the four ER+ and three ER– cell lines only.

DE gene detection comparison. To evaluate the effect of normalization method on the downstream analysis, DE gene detection was performed by using Bioconductor package DESeq.[38] The distributions of the *P* values obtained after applying DESeq to the raw, DESeq default normalized, and integrated normalized data are presented in Figure 4. Though many of the DE genes are the same when applying DESeq to the three datasets, discrepancies do exist. For example, the smallest *P* values obtained are 6.58377×10^{-10} (raw), 6.460862×10^{-14} (DESeq), and 7.390189×10^{-18} (integrated), respectively. This, together with Table 1, indicates that integrated normalization increases the power of detecting DE genes. Furthermore, the histograms of the *P* values from the raw and DESeq default normalized data are very similar (Fig. 4a and b), but there are more small *P* values when using the integrated normalized data (see the most left bars in Fig. 4). In addition, except for the very small *P* values, the majority of the *P* values from the integrated normalized data are more uniformly distributed. Our interpretation is that if correctly normalized, the true DE genes can be detected with increased power, while *P* values of the *true* non-DE genes should be uniformly distributed.[48] Note that since read counts take discrete values and the expressions of some genes are quite low, a large fraction of *P* values being very close to 1 is expected.

We further generated and presented the volcano plots in Figure 5. We can see that a substantially larger number of smaller *P* values [equivalently, larger –log 10(*P* value)], which are associated with large absolute fold changes, are found for the integrated normalized data.

Next, we compared the total numbers of DE genes at different Benjamini and Hochberg (BH)[12] cutoff values (Table 1). It is clear that for all the cutoff *P* values, the integrated normalization method detects most DE genes compared to other methods.

The numbers of DE genes (with BH-adjusted *P* value <0.001) found in common from three datasets are listed in Table 2. Among the 33 DE genes identified by using the raw data, 30 genes were identified by using the DESeq normalized data, out of which 25 genes were detected by using all three normalization methods. This indicates that the majority of DE genes identified by using the raw or DESeq normalization data can also be identified by using the integrated normalized data. On the other hand, using the integrated normalization data detects 21 (46 – 25) DE genes that were not detected by using the raw or DESeq normalized data.

We also used the *concordance at the top* plots to display the concordance among the genes discovered by the three methods. It is clear that the concordance rate of top differentially expressed genes detected by using the raw data and the DESeq normalized data were very high. In addition, the top genes obtained from integrated normalized data slightly differed (Fig. 6).

We preformed the same analysis by using the edgeR software, and similar results were observed. We further performed GC content and gene length adjustments using the conditional QN (CQN).[24] However, adding CQN step resulted in smaller numbers of DE genes for small cutoff values (Supplementary Table 1).

Last, we constructed the heat maps in Figure 7 using DE genes with BH-adjusted *P* value less than 0.001. We observed that, when applying DESeq to the raw data and the DESeq default normalized data, unsupervised clustering did not provide the expected classification: all three ER– cell lines belonging to one cluster and all four ER+ cell lines belonging to another cluster. More specifically, T47D, which is an ER+ cell line, was assigned to a separate cluster (Fig. 7a and b) other than the ER+ cluster. On the other hand, when applying DESeq to the integrated normalized data, the clustering results are consistent with the biological classification of the seven samples (Fig. 7c). Note that ER+ and ER– breast cancers are clinically different diseases with distinct responses

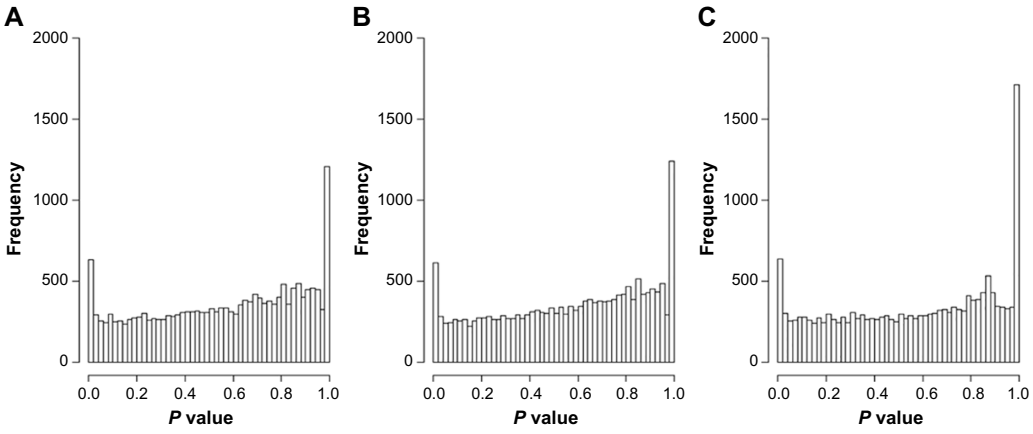

Figure 4. Histograms of *P* values: (**A**) apply DESeq to raw data; (**B**) apply DESeq to DESeq default normalized data; and (**C**) apply DESeq to integrated normalized data.

Table 1. Total numbers of DE genes identified at different cutoff values.

BH ADJUSTED CUTOFF VALUE	RAW DATA	DESeq NORMALIZED	INTEGRATED NORMALIZED
<0.01	75(2)	69(2)	77(2)
<0.001	33(2)	30(2)	46(2)
<0.0001	17(1)	17(1)	30(2)
<0.00001	7(0)	7(0)	22(2)

Note: Numbers in parentheses are for housekeeping genes.

Table 2. The numbers of DE genes (BH-adjusted *P* values <0.001) found in common for different normalization methods.

	RAW DATA	DESeq NORMALIZED	INTEGRATED NORMALIZED
Raw data	33	30	25
DESeq normalized		30	25
Integrated normalized			46

to hormonal therapy or other treatments. It has been shown by microarray and serial analysis of gene expression studies that ER+ and ER– subtype breast cancers have distinct gene expression profiles that can be used for both diagnosis and outcome prediction.[49]

To check the effect of using different BH-adjusted *P* value cutoffs on sample classification, we listed a series of classification results using different cutoff values in Supplementary Figures 10–12. Our comment is that the classification results are consistent regardless of which cutoff value is used for the integrated normalized data. However, the results are less consistent for the raw data and the DESeq default normalized data.

The housekeeping genes. Housekeeping genes are those whose expressions, under ideal situations, should not be regulated or influenced by the experimental conditions/tissue types.[50] However, based on our results, the expression levels of these housekeeping genes are correlated with DNA copy number. As a comparison, the correlations between DNA copy numbers and expression levels of the housekeeping genes are somewhat higher than those between DNA copy numbers and RNA expressions for all genes. For example, we observed that the highest correlation is 0.280 for housekeeping genes, while the correlations for all genes range from 0.10 to 0.186. In addition, the spline smoothing fitted lines for DNA data and RNA data for the housekeeping genes look similar, indicating that the expressions of housekeeping genes are, against our intuition, affected by DNA copy number changes (Fig. 8a and b). Our interpretation is that changes in expression level caused by CNAs are less compensated by gene regulation mechanisms, and thus, CNAs play a more important role in affecting housekeeping gene expression. These results provide strong evidence to suggest that, when CNAs exist, it might not be appropriate to use housekeeping genes as the reference genes for quantitative real-time reverse transcription polymerase chain reaction normalization, and these genes also should not be used as references in NGS data normalization.

Conclusions

RNA-seq and expression array technologies have been widely used in cancer genetic studies to identify DE genes and reveal the biologic spectrum of cancers, as well as provide diagnostic tools and identify new therapeutic targets. RNA data normalization, especially for microarrays, has been well discussed. However, many normalization methods do not take into consideration how DNA copy number change alters gene expression. Consequently, though these methods work well for normal samples without CNAs, they may not be sufficient for cancer samples with substantial CNAs.

It can be very difficult to identify CNAs using RNA expression data directly, due to the on/off nature of gene expression, and the substantial differences in expression levels

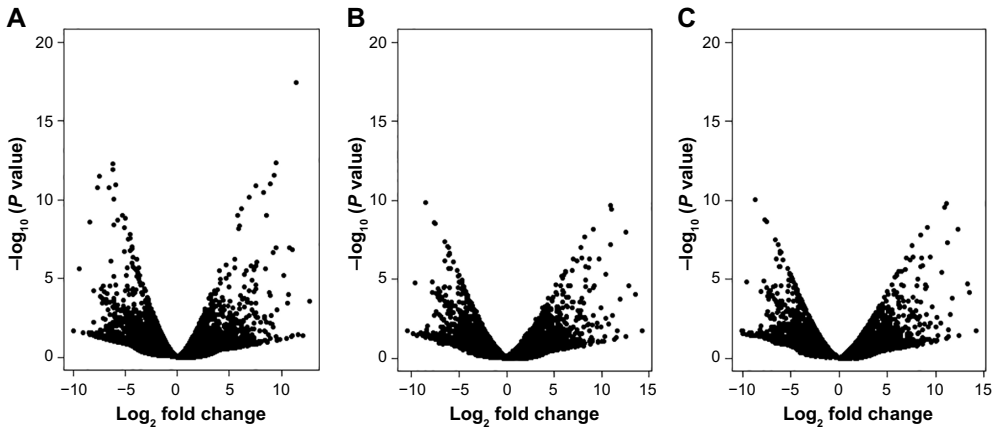

Figure 5. Volcano plots: (**A**) integrated normalized data; (**B**) DESeq normalized data; and (**C**) raw data. A few genes expressed in only one condition and, thus, did not show in the volcano plots.

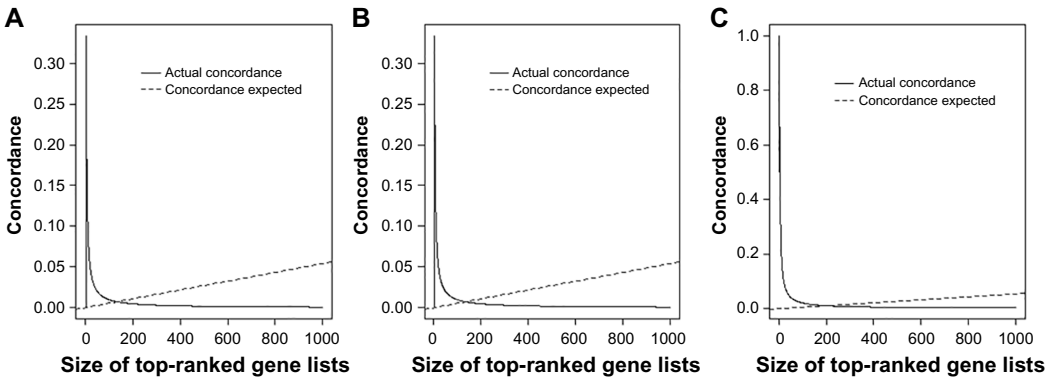

Figure 6. Concordance at the top plots. (**A**) Concordance between top genes obtained by analyzing the raw data and the integrated normalized data. (**B**) Concordance between top genes obtained by analyzing the DESeq and integrated normalized data. (**C**) Concordance between top genes obtained by analyzing the raw data and DESeq normalized data.

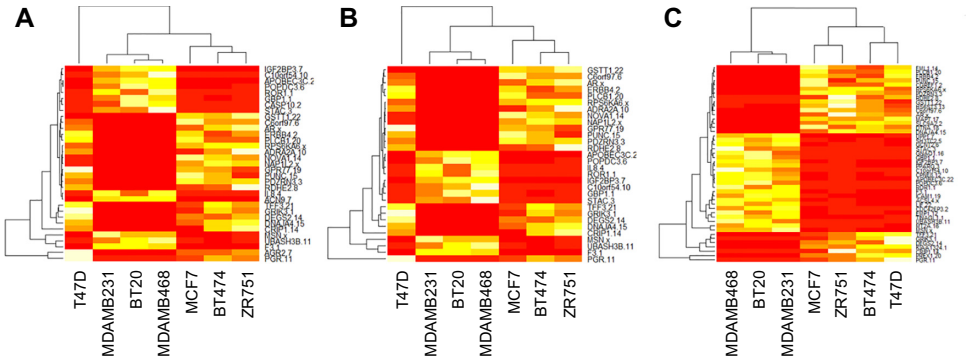

Figure 7. Heatmaps of the genes with BH-adjusted *P* values less than 0.001. (**A**) Heat map was generated by applying DESeq to the raw data. (**B**) Heat map was generated by applying DESeq to the DESeq normalized data. (**C**) Heat map was generated by applying DESeq to the integrated normalized data.

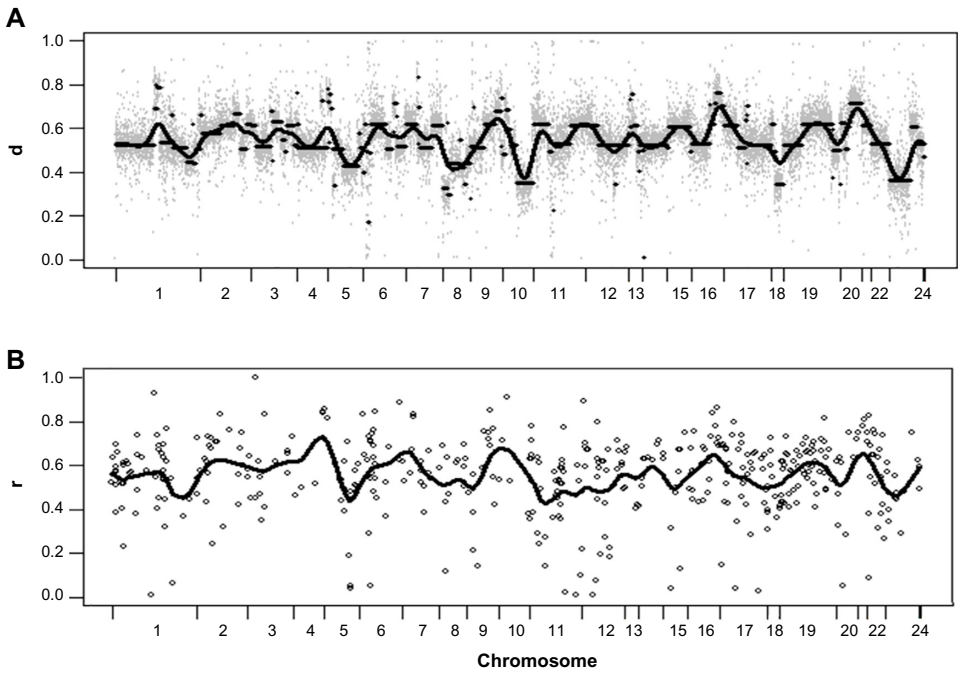

Figure 8. The plots of CNA and RNA-seq housekeeping gene expression. Spline smoothing curve was superimposed in bold. (**A**) For DNA-seq data, *r* is used. CBS segmentation results were superimposed in bold lines. (**B**) For RNA-seq data for housekeeping genes.

of different genes. On the other hand, DNA level data, along with RNA level data, have been collected in many cancer genetic studies to better understand the causes of tumorigenesis. This provides an excellent opportunity for biostatisticians/bioinformaticians to develop integrated methodologies to extract more meaningful information from the data.

We demonstrated in this article that moderate-to-strong correlation between DNA and RNA data does exist, especially for cancer samples with substantial CNAs. This was achieved by taking the average of signal intensities within a window of fixed size for both RNA and DNA data to reduce signal intensity variation and using segment mean values in correlation calculation.

The magnitude of the correlation between DNA copy number and gene expression level is a good indication of whether DNA copy number change affects gene expression. The integrated normalization method proposed in this article benefits more with stronger correlation, in other words, when CNAs occur in many genomic regions.

Many current RNA-seq normalization methods do not have the capacity to make adjustments for CNA-caused variations. The integrated method in this article aims to address this issue by appropriately aligning the global signal intensity. The integrated normalization resulted in more differentially expressed genes being detected than using either the raw data or the DESeq default normalized data. Parenthetically, nuisance variations due to differences in CNAs and sequencing depths would not be removed without proper normalization, thus, diluting the power for DE gene detection. After appropriate normalization, genes that are *truly* differentially expressed have a greater chance of being detected.

We demonstrated the utility of the proposed method using publicly available cell line data due to the facts that there is no restriction on the access of these data and that the cell lines are well studied, with many of their genomic characteristics well known. Meanwhile, we do not foresee any theoretical challenges in applying the method on patient data. Specifically, there should be no difference in terms of aligning the reference two-copy regions across samples regardless whether the data came from cell lines or patient samples. Gene expression profiling is commonly used to provide accurate diagnosis/classification results. In this study, we expect that expression profiling for all the cell lines agrees with the known biological classification. Meanwhile, we anticipate that such classification results should not be affected by how many DE genes are used in the classification algorithm. However, to our surprise, other than using the integrated normalized data, using either the raw data or the DESeq default normalized data, resulted in inconsistent classification results (for different cutoff *P* values). Note that when including all genes in the profiling, ER+ and ER– breast cancers can be correctly classified, indicating that the difference between the ER+ and ER– cells are substantial.

We did not perform GC content and gene length adjustments in the integrated normalization for two reasons:

1. GC content is supposed to affect the numerator and denominator of the constructed read count ratios similarly; thus, it is not necessary to adjust the GC content for the ratios. We plotted the GC content versus the constructed ratio and did not observe obvious relationship between them (Supplementary Figs. 6 and 7);
2. In the DE gene detection stage, since GC content is the same for both the comparison and reference groups, GC content adjustment is also not necessary. Note that the relationship between GC content and DNA/RNA read counts are quite consistent across samples (Supplementary Figs. 8 and 9). In fact, many of the DE gene detection methods, for example, the protocol proposed by Trapnell et al.[51], the DESeq software we used, and edgeR, do not suggest GC content adjustment because of the nature of such comparisons.

Nevertheless, to evaluate the performance of GC content and gene length adjustments, we applied CQN on those samples. Not surprisingly, CQN did not perform well when we used smaller cutoff values. Our interpretation is as follows:

1. CQN makes data adjustments based on GC content and gene length; however, since the comparison and the reference groups have the same GC content and gene length, such adjustments have little effect on the comparisons;
2. Since CQN made certain adjustments based on GC content and gene length, if such adjustments are not necessary, then the CQN process turns out to introduce noises and, thus, may have negative effect on the comparisons. Note that, when the effect of GC content is consistent across the samples, then as mentioned in (1), GC content adjustment is not necessary; otherwise, if the effects are not consistent, then we cannot rule out the possibility that we are actually adjusting the spurious correlation between expression and GC content, which is not appropriate. As a comparison, the integrated normalization utilizes information obtained from real experiments, and thus, the adjustments bear biological meanings.

Performing RNA-seq data normalization by using CNAs detected from DNA-seq data can be very useful when CNAs do exist, and such normalization ensures that downstream analyses are more meaningful. For samples that have only focal CNAs or without any CNAs (normal samples), this approach is still applicable except that all the genome will be used as the reference in RNA-seq normalization.

The limitation of our method is that we need a control sample(s), preferably paired normal samples, for CNA detection (it is not needed for RNA-seq data). In fact, accurately

identifying CNAs without using any control sample(s) is not straightforward and may need some manual trial and error. In addition, our method improves the power of detecting truly DE genes, but it is not capable of discriminate the driver and passenger genes. Although DE genes located within CNA regions are more likely to be the driver genes than the passenger genes, the nonuniqueness nature of the cause of gene expression level change, along with the very often near-random distribution of CNAs, determines that seeking such an association is quite challenging.

Authors Contributions

Designed the experiments: SY, ZF. Did the data analysis: SY, ZF. Wrote the article: SY, DM, KZ, ZF. All authors read and approved the final article.

Supplementary Material

Supplementary Figure 1. A QQ plot for the average of the ratio *d*. The plot indicates that there is no serious violation on the normality assumption.

Supplementary Figure 2. An illustration of reduced variance of meand comparing with *d*.

Supplementary Figure 3. The result of modality test.

Supplementary Figure 4. The magnitude of the adjustment factor for the variable r.

Supplementary Figure 5. An illustration of the property of different transformations.

Supplementary Figure 6. Scatter plots of GC content versus DNA read count ratio.

Supplementary Figure 7. Scatter plots of GC content versus RNA read count ratio.

Supplementary Figure 8. Scatter plots of GC content versus log-transformed DNA read count ratio.

Supplementary Figure 9. Scatter plots of GC content versus log-transformed RNA read count ratio.

Supplementary Figure 10. Heatmaps with different BH cut-off values (Integrated normalized).

Supplementary Figure 11. Heatmaps with different BH cut-off values (Raw data).

Supplementary Figure 12. Heatmaps with different BH cut-off values (DESeq normalized).

Supplementary Table 1. Total numbers of DE genes identified by edgeR at different cutoff values, with different normalization methods.

REFERENCES

1. Costa V, Aprile M, Esposito R, Ciccodicola A. RNA-seq and human complex diseases: recent accomplishments and future perspectives. *Eur J Hum Genet*. 2013;21:134–42.
2. Altmann A, Weber P, Bader D, Preuss M, Binder EB, Müller-Myhsok B. A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Hum Genet*. 2012;131:1541–54.
3. Diouf B, Cheng Q, Krynetskaia NF, et al. Somatic deletions of genes regulating MSH2 protein stability cause DNA mismatch repair deficiency and drug resistance in human leukemia cells. *Nat Med*. 2011;17(10):1298–303.
4. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. *Nat Genet*. 2007;39:S37–42.
5. Fanciulli M, Norsworthy PJ, Petretto E, et al. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat Genet*. 2007;39(6):721–3.
6. Yang Y, Chung EK, Wu YL, et al. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am J Hum Genet*. 2007;80(6):1037–54.
7. Mullighan CG, Goorha S, Radtke I, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature*. 2007;446(7137):758–64.
8. Ha G, Roth A, Lai D, et al. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res*. 2012;22(10):1995–2007.
9. Stranger BE, Forrest MS, Dunning M, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*. 2007;315(5813):848–53.
10. Illumina Inc. *RNA-Seq Data Comparison with Gene Expression Microarrays. White Paper: Sequencing*. 2011. Available at: http://www.europeanpharmaceuticalreview.com/wp-content/uploads/Illumina_whitepaper.pdf. Accessed August 1, 2015.
11. Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57–63.
12. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Methodol*. 1995;57(1):289–300.
13. Storey JD. A direct approach to false discovery rates. *J R Stat Soc Series B Stat Methodol*. 2002;64(3):479–98.
14. Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW. Significance analysis of time course microarray experiments. *Proc Natl Acad Sci U S A*. 2005;102(36):12837–42.
15. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11:R25.
16. Smyth GK, Speed TP. Normalization of cDNA microarray data. *Methods*. 2003;31(4):265–73.
17. Pounds S, Cheng C, Mullighan C, Raimondi SC, Shurtleff S, Downing JR. Reference alignment of SNP microarray signals for copy number analysis of tumors. *Bioinformatics*. 2009;25(3):315–21.
18. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods*. 2008;5(7):621–8.
19. Pickrell JK, Marioni JC, Pai AA, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*. 2010;464(7289):768–72.
20. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008;18(9):1509–17.
21. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*. 2008;9:321–32.
22. Robinson MD, Strbenac D, Stirzaker C, et al. Copy-number-aware differential analysis of quantitative DNA sequencing data. *Genome Res*. 2012;22(12):2489–96.
23. Cloonan N, Forrest ARR, Kolle G, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods*. 2008;5:613–9.
24. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*. 2012;13(2):204–16.
25. Gu W, Choi H, Ghosh D. Global associations between copy number and transcript mRNA microarray data: an empirical study. *Cancer Inform*. 2008;6:17–23.
26. Kotliarov Y, Kotliarova S, Charong N, et al. Correlation analysis between single-nucleotide polymorphism and expression arrays in gliomas identifies potentially relevant target genes. *Cancer Res*. 2009;69(4):1596–603.
27. Chen HI, Hsu FH, Jiang Y, et al. A probe-density-based analysis method for array CGH data: simulation, normalization and centralization. *Bioinformatics*. 2008;24(16):1749–56.
28. Bicciato S, Spinelli R, Zampieri M, et al. A computational procedure to identify significant overlap of differentially expressed and genomic imbalanced regions in cancer datasets. *Nucleic Acids Res*. 2009;37(15):5057–70.
29. Orozco LD, Cokus SJ, Ghazalpour A, et al. Copy number variation influences gene expression and metabolic traits in mice. *Hum Mol Genet*. 2009;18(21):4118–29.
30. Haverty PM, Hon LS, Kaminger JS, Chant J, Zhang Z. High-resolution analysis of copy number alterations and associated expression changes in ovarian tumors. *BMC Med Genomics*. 2009;2:21.
31. Menezes RX, Boetzer M, Sieswerda M, van Ommen GJ, Boer JM. Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC Bioinformatics*. 2009;10:203.
32. Yang S, Fang Z. Beta approximation of ratio distribution and its application to next generation sequencing read counts. *Journal of Applied Statistics*, in press, DOI: 10.1080/02664763.2016.1158798.

33. Yang S, Pounds S, Zhang K, Fang Z. PAIR: paired allelic log-intensity-ratio-based normalization method for SNP-CGH arrays. *Bioinformatics*. 2013;29(3):299–307.
34. Staaf J, Jönsson G, Ringnér M, Vallon-Christersson J. Normalization of array-CGH data: influence of copy number imbalances. *BMC Genomics*. 2007;8:382.
35. Hasselblad V. Estimation of parameters for a mixture of normal distributions. *Technometrics*. 1966;8(3):431–44.
36. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr*. 1974;19(6):716–23.
37. *PAIR Software*. 2012. Available at: http://publichealth.lsuhsc.edu/pair.html. Accessed August 1, 2015.
38. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11(10):R106.
39. *Human Housekeeping Genes*. 2003. Available at: http://www.tau.ac.il/~elieis/Housekeeping_genes.html. Accessed August 1, 2015.
40. *BioMart Clone/Gene ID Converter*. 2015. Available at: http://central.biomart.org. Assessed August 1, 2015.
41. Sun Z, Asmann YW, Kalari KR, et al. Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing. *PLoS One*. 2011;6(2):e17490.
42. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25.
43. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.
44. Anders S, Pyl PT, Huber W. HTSeq – a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015;31(2):166–9.
45. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 2004;5:557–72.
46. Rice J. Bandwidth choice for nonparametric regression. *Ann Stat*. 1984;12(4):1215–30.
47. Bolstad BM, Irizarry RA, Astrand H, Speed TP. A comparison of normalization methods for high density oligonucleotide array based on variance and bias. *Bioinformatics*. 2003;19(2):185–93.
48. Fodor AA, Tickle TL, Richardson C. Towards the uniform distribution of null *P* values on affymetrix microarrays. *Genome Biol*. 2007;8:R69.
49. Sotiriou C, Neo SY, McShane LM, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A*. 2003;100(18):10393–8.
50. Radonić A, Thulke S, Mackay IM, Landt O, Siegert W, Nitsche A. Guideline to reference gene selection for quantitative real-time PCR. *Biochem Biophys Res Commun*. 2004;313(4):856–62.
51. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012;7(3):562–78.