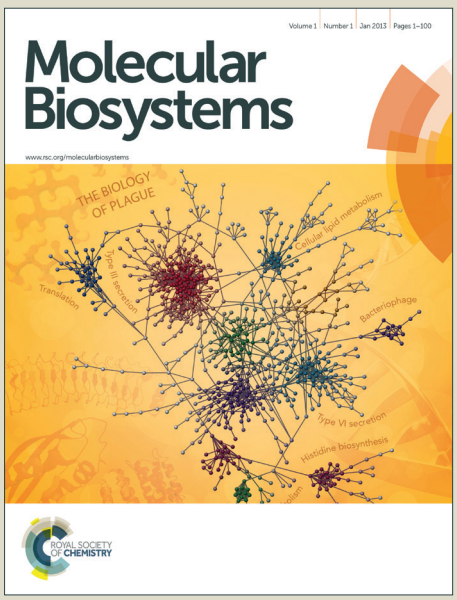
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Molecular Biosystems Molecular BioSystems Dynamic Article Links ►

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

PAPER

Deciphering global signal features of high-throughput array data from cancers

Deng Wu, [a†] Juanjuan Kang, [a†] Yan Huang, [a†] Xiang Li, [a†] Xiansong Wang,[a] Dan Huang,[a] Yuting Wang,[a] Bin Li,[a] Dapeng Hao,[a] Qi Gu, [a] Nelson Tang,[c] Kongning Li,[a] Xia Li,*[a] Zheng Guo,*[a] Jianzhen Xu,*[b] and Dong Wang*[a]

Normalization of array data relies on the assumption that most genes are not altered, which means that the signals for different samples should be scaled to have similar median or average values. However, accumulating evidence suggests that gene expressions could be widely up-regulated in cancers. Our previous results and subsequent finding have shown violation of the assumption led to erroneous interpretations of microarray data. To decipher the global signal features of microarray data from cancer samples, we empirically evaluated a large collection of gene and miRNA expression profiles and copy-number variation arrays. Our results showed that, at the transcriptomic level, genes and miRNAs are widely over-expressed in a large proportion of cancers. In contrast, at the genomic level, global raw signal intensities for methylation and copy number variation show negligible differences between cancer and normal samples. These results force us to re-evaluate the proper use of normalization procedures under different experimental conditions and for different array platforms.

Introduction

High-throughput array technology is a powerful tool for transcriptome and genome analysis[1,2,3,4,5]. Gene and miRNA expression microarrays provide quantitative information about the population of RNA species in a cell or tissue[1,6]. Using microarrays to monitor global transcriptome expression under various conditions has a tremendous influence on modern biological research[1,6]. Similarly, methylation arrays and SNP technology were developed for the investigation of methylation status and copy number variation on a genome-wide scale, and they have provided insights into cancer mechanisms, biomarker prediction and drug target identification[7,8,9,10].

Typically, array data are subject to multiple sources of variation, including variation in the preparation of the biological sample, scanning effects and the characteristics of the different arrays. Thus, normalization is a critical initial step in data analysis[1,3,4,5]. Usually, the same number or a similar number of DNA/RNA molecules from each sample should be applied to each array so that the total signal intensities will be similar for each sample[3,11,12]. Researchers normalize the signal intensities across all arrays to have the same distribution or a similar distribution regardless of the disease state, under the assumption that only a few genes are altered by disease and that similar numbers of genes are up-regulated or down-regulated. Hence, if the cellular sources produce equivalent amounts of DNA/RNA molecules, and if the yields of the molecules and their derivatives are equivalent throughout the experimental manipulation, then the normalized expression data should produce an accurate representation of the relative levels of each gene product (Figure 1)[3,11,12]. However, emerging evidence suggested that this commonly used assumption may not hold true in certain situations. Based on a previous analysis of 16 pair-matched cancer and normal gene expression datasets, we observed extensive increases in the microarray signals for the cancer samples[13,14]. Subsequently, Loven et al. also showed that cells with high levels of c-Myc can amplify their gene expression program, producing two to three times more total RNA and generating cells that are larger than their low-Myc counterparts[15]. Recently, we also found that gene expression may be widely up-regulated in several non-cancerous complex diseases[16]. In such cases, normalization would distort the global data distribution and lead to erroneous interpretations of gene expression profiles (Figure 2)[13,15].

c-Myc is just one of the many master transcription factors governing the transcriptional programs in cancers[17,18,19]. It is unclear how many cancers show increased transcription and how often it may lead to misinterpretation of genome-wide expression data. More importantly, in addition to mRNA, what are the global signal features in cancers compared to normal samples for miRNA and copy number variations array data. These issues for miRNA and copy number variation array data also represent fundamental questions that relate to all subsequent data analysis and interpretation, but surprisingly, they have not been systematically analyzed. In this study, we comprehensively analyzed the raw global signal intensities of multiple cancer datasets in an unbiased collection from the NCBI GEO database[20]. Transcriptomic (gene and miRNA expression) and genomic (copy number variation) datasets were used.

Results

Global over-expression in cancer gene expression data

In previous work, our results showed that genes are extensively up-regulated in cancers. We observed this phenomenon in 14 of the 16 datasets[13]. Recently, we assembled an unbiased collection of 23 pair-matched gene expression datasets for 12 cancer types. For each of the 23 datasets, we computed the median of the raw signal intensities, and we compared the medians between cancer samples and normal samples. The medians of the raw signal intensities in the cancer samples was increased in 21 of the 23 datasets; this was unlikely to happen by chance if the probability of observing a larger median in the cancer-state dataset is 0.5 in each independent dataset ($P<0.05$, binomial test). The increase in the median signal in the cancer samples was significant ($P<0.05$, Wilcoxon rank-sum test) in eight datasets and was marginally significant ($P<0.1$, Wilcoxon rank-sum test) in another four datasets (Supplementary Table 1).

These results show that the raw signal intensities of cancer samples tend to be significantly or marginally significantly higher in more than half of the datasets (12/23=52%). Because of the low statistical power of detecting significant differences in a small set of samples[21,22], we focused next on five larger datasets with at least 35 samples (cancer or normal), and we found that the percentages were increased further 4/5=80% (significantly and marginally significantly). Thus, the assumption that all arrays for a particular cancer would have the same probe intensity distribution regardless of the physiological state might be misleading. Common normalization methods would distort the global over-expression signal distribution and lead to erroneous interpretations of gene expression data (Figure 2 and 3)[13,15].

Global over-expression in cancer miRNA data

Several commonly used normalization methods for miRNA expression are similar to mRNA gene profiling normalization approaches[23,24], but the global features in cancers have not been investigated in detail. Using the same criteria as for gene expression datasets, we assembled an unbiased collection of 12 pair-matched single-channel miRNA expression datasets for 8 cancer types. The median of the raw signal intensities in the cancer samples was increased in 9 of 12 datasets. The increase in the median signal in the cancer samples was significant ($P<0.05$, Wilcoxon rank-sum test) in three datasets and marginally significant ($P<0.1$, Wilcoxon rank-sum test) in one dataset: Colon168 ($P=6.7\times10^{-7}$), Esophageal152 ($P=1.19\times10^{-3}$), Esophageal206 ($P=1.03\times10^{-3}$), Liver184 ($P=6.3\times10^{-2}$) (Table 1). If we only focus on the five largest miRNA sample datasets with at least 50 samples for each state (cancer or normal), the percentage is 80% (4/5=80%) (significant and marginally significant). Hence, the traditional normalization methods would again distort the global over-expression signal distribution of miRNA expression data (Figure 2 and 3).

Normalization might over-normalize signals in cancer transcriptome data

Based on the results described above, we can see that transcription (for both genes and miRNAs) differs greatly between cancer samples and normal samples. Thus, the underlying assumption for normalization is not satisfied. This means that normalization may over-normalize the global signal features in cancer transcriptome data (Figure 2 and 3). As illustrated in Figure 2-C, the raw signal intensities of genes in cancer samples were moderately significantly higher than that in the normal samples, but differentially expressed genes (DEGs) could not be identified after normalization. For example, as illustrated in Figure 3-B for the mRNA Colon34 dataset, the gene PABPC1L2B was moderately significantly higher ($P=6.69\times10^{-4}$, Wilcoxon rank-sum test) in terms of raw signal intensity in the cancer samples, but it was not identified as a DEG by RMA. Thus, this represents a false-negative result. As illustrated in Figure 2-D, a gene could be selected as a down-regulated differential gene after normalization, even though its raw signal intensities in the cancer samples were similar to those of the normal samples. Figure 3-C shows another example; the gene PRR12 had similar raw intensities in the cancer samples and the normal samples, but it was identified as significantly down-regulated ($P=2.06\times10^{-2}$, Wilcoxon rank-sum test) in the cancer samples by RMA. Thus, it represents a false-positive result. In addition, the expression directions of genes with moderate or low raw signal intensity differences between cancer and normal samples could be reversed after normalization. These results indicate that commonly used normalization methods might over-normalize the data. This can cause many up-regulated differentially expressed genes/miRNAs to be missed, and it can lead to a non-negligible fraction of down-regulated differentially expressed genes/miRNAs in cancer transcriptome data.

Effect of normalization on the expression directions of differentially expressed genes/miRNAs and the Pearson correlation coefficient distribution

Next, we focused on the mRNA and miRNA datasets with significant increases in the raw signal intensities in the cancer samples. As shown in Figure 4, we compared the expression directions of the DEGs detected before and after normalization (RMA, dChip and LVS) in the mRNA colon34 dataset. Our results showed that many genes (1204, 1324, 623) were identified as up-regulated DEGs in the raw signal data, but these genes were not identified as DEGs after RMA, dChip or LVS normalization. Furthermore, 98% (1204/1233), 84% (1324/1575) and 99% (623/627) of the DEGs selected based on the raw signal data were identified as up-regulated DEGs, respectively. Similarly, as shown in Figure 5, 1017 and 1113 miRNAs were detected as up-regulated differentially expressed miRNAs in the raw signal data, but these miRNA were not detected as differentially expressed miRNAs after quantile normalization or LVS in the miRNA Esophagus152 dataset. Furthermore, 100% (1017/1017) and 95% (1113/1166) of the differentially expressed miRNAs selected based on the raw signal data were identified as up-regulated differentially expressed miRNAs. Similar results were observed in the miRNA Colon168 and Esophagus206 datasets (Supplementary Figures 1 and 2). These results indicate that normalization may cause a large fraction of truly up-regulated differentially expressed genes/miRNAs to be overlooked in cancer samples. On the other hand, as shown in Figures 4 and 5, a large fraction of the genes/miRNAs in mRNA dataset Colon34 and miRNA dataset Esophagus152 were identified as down-regulated differentially expressed genes/miRNAs after normalization, but these genes/miRNAs were not identified as differentially expressed genes/miRNAs in the raw signal data.

This shows that RMA, quantile normalization, dChip and LVS can produce many false down-regulated differentially expressed genes/miRNAs in cancer samples.

Due to the complexity of biological regulation, most functional mechanisms might be explained not by individual differentially expressed genes but by the combined effects of many moderately changed genes[25,26]. A recent trend is to construct gene co-expression networks based on microarray data[25,27,28]. Thus, we have also compared the pair-wise Pearson correlation coefficient distribution before and after normalization for both the mRNA and miRNA arrays. As shown in Figure 6, in the mRNA Colon34 dataset, 68% of genes pairs tend to be positively correlated before normalization, but this percentage decreases to 50%, 52% or 58% after RMA, dChip or LVS, respectively. Similarly, as shown in Figure 7, in the miRNA Esophagus152 dataset, 98% of the miRNA pairs tend to be positively correlated before normalization, but this decreases to 54% or 52% after quantile normalization or LVS. Similar results were observed for the miRNA Colon168 and Esophagus206 datasets (Supplementary Figures 3 and 4). Our results show that, before normalization, the genes pairs tend to be positively correlated, but normalization severely affects the signal distribution and causes that tendency to disappear.

Minor differences in raw signal intensities for genomic data between cancer samples and normal samples

For methylation and copy number variation array data, although some researchers normalize these data after considering the statistical benefit, most do not perform normalization to avoid systemic bias[3,9]. Using an unbiased collection of nine pair-matched single-channel datasets based on SNP arrays, we found that, in eight of the nine datasets, the median raw signal intensities in the cancer samples were not significantly different from those of the normal samples. Further analysis has shown that, in six of nine datasets, the raw signal intensities are slightly increased at the 75th quantiles in the cancer samples. Only one dataset (colon188) showed a significant increase of the median and the 75th quantile raw signal intensities ($P=1.54x10^{-2}$ and $2.44x10^{-2}$, respectively) (Table 2). Similarly, our previous results also demonstrated that, in all of the eight analyzed methylation datasets, the median raw signal intensities in the cancer samples were not significantly different ($P>0.05$) from those of normal samples. Further analysis indicated that slightly more genes were hypomethylated in the cancer samples compared to the normal tissues around the 75th quantile[29]. These results suggest that, except for colorectal cancer, normalization algorithms may have more positive effects by reducing technical variations compared to the negative effects that remove biological signals of methylation or copy number variation in array data.

Discussion

Our results show that cancer tissue samples tend to show elevated intensities in most pair-matched cancer and normal expression datasets (21 of 23). Similar results were also observed for miRNA expression datasets (9 of 12). Notably, these samples were taken from a variety of cancer types and were produced by different laboratories around the world. Thus, this phenomenon is not likely to be limited to cancers overexpressing c-Myc. Rather, the up-regulation of mRNA and miRNA products is a general feature of cancer cells. Alterations of many essential cellular functions, which are referred to as cancer hallmarks, collectively dictate malignant growth for almost all human cancers[30,31], so gene expression could be globally changed in cancer, as evidenced by the fact that we often detect thousands of differentially expressed genes in comparisons of cancer samples with normal controls[13,32]. On the other hand, the global raw signal distributions of genomic array datasets, such as datasets for methylation or copy number variation, show little difference between cancer and normal samples. This is perhaps because both DNA hypomethylation and hypermethylation have been associated with carcinogenesis in numerous investigations[33,34]. Similarly, there are both gains and losses of copy number that occur in cancer genomes[35,36]. Thus, the global raw signals of genomic data may follow similar distributions in both cancer samples and normal samples.

These results have great implications for cancer biology. First, we suggest that, at least for cancer microarray data, it is better to compare the raw global signal distributions between cancer and normal samples, as was done in this study. This critical quality control step will enable proper use of normalization methods and more accurate interpretation of array data. Second, when there is the potential for global signal changes in transcriptome data, conventional normalization methods should be used with caution because such analyses are based on an unreliable assumption and may actually distort the biological signals. This may cause more harm than good during the biological analysis of array data. However, in genomics analyses (methylation and copy number variation data), the situation is very different. Our results show that the current practice of avoiding normalization for these types of arrays may represent an over-abundance of caution. In fact, normalization for these arrays can increase the power of detecting differentially methylated sites and SNP changes among many weak and complicated signals. Based on our analysis, we encourage the use of normalization when analyzing genomic array data, but we advise against using normalization uncritically for the analysis of transcriptome data. Finally, we think it is important to comprehensively evaluate the effects of normalization procedures on the subsequent bioinformatics analysis of array data, especially for cancer datasets. For example, if our goal is to find a small number of significant biomarkers for disease diagnosis[37], normalization might have less of an influence on the results because the most significant differentially expressed genes/miRNAs tend to remain significant after normalization (Figure 2-B, Figure 3-A). However, as our results demonstrated, global over-expression of genes/miRNAs in cancers may lead to erroneous conclusions about the underlying cancer mechanism if normalization is used (Figure 2-C and D, Figure 3-B and C, Figure 4-7)[13,15], especially for the expression directions of differentially expressed genes/miRNAs and the Pearson correlation coefficient distribution. Finally, we suggest that, at least for transcriptome data from cancer studies, a critical initial step is to evaluate the differences in raw global signal distribution between cancer and normal samples. Additionally, the use of spike-in controls, as suggested by Loven et al.[15] and others[38,39,40], may be a useful, robust, cross-platform quality control method to enable more accurate detection of cancer-

associated genes/miRNAs in transcriptome data. Furthmore, we should pay more attention towards optimizing experimental designs and stringently randomizing potential experimental artifacts across biological groups, use of sufficient sample sizes, more conducive to probe hybridization and developing the novel normalization strategy for transcriptome data may also help for the solution of this problem in cancer studies.

Methods

High-throughput omics datasets and signal intensities

We specifically selected pair-matched datasets in which the normal samples were taken from the same subjects as the cancer samples, so that the effects of certain variables would be eliminated, such as familial effects, individual effects and environmental differences[13]. Using the NCBI GEO database[20] for gene expression datasets, we obtained an unbiased collection of 23 pair-matched Affymetrix single-channel datasets for 12 cancer types. Each dataset had to include at least 10 samples for each state (cancer and normal). Similarly, we collected a total of 12 pair-matched single-channel miRNA datasets for 8 cancer types. High-density SNP arrays provide a robust, effective method for detecting and analyzing genomic copy number variation[41,42], so we also collected 9 pair-matched single-channel datasets for 6 cancer types. All of the datasets that were analyzed in this study are described in Table 3.

For the raw signal intensities of the high-throughput omics data in .cel format, we only used the PM intensities to represent "signal intensity" because it has been shown that ignoring the MM values is preferable for background correction[43,44]. For .txt and .gpr format data, the background intensities were subtracted from the raw signal intensities to obtain the final signal intensities.

Normalization algorithms and identification of differentially expressed genes

In this work, RMA[44], quantile[45] and dChip[46] normalization algorithms were used for the mRNA and miRNA data. These normalization methods are based on the traditional assumption that only a few genes are altered by disease status and that the numbers up- and down-regulated genes are similar. Furthermore, we analyzed another algorithm (LVS)[47,48] that relies less on this assumption. SAM (significance analysis of microarrays)[49] was used to identify DEGs in cancer samples with an FDR control level of 0.05.

Acknowledgments

This work was supported by the National High Technology Research and Development Program of China (863 Program, 2014AA021102), the Major State Basic Research Development Program of China (973 Program, 2014CB910504), the National Natural Science Foundation of China (31100901, U1304311, 91129710, 61170154), China Postdoctoral Science Foundation (2013M531064), Natural Science Foundation of Heilongjiang Province of China (QC2010012 and ZD201114), Heilongjiang Postdoctoral Foundation (LBH-Z12171), Plan of Nature Science Fundamental Research in Henan University of Technology(11JCYJ11) and China Undergraduate Training Programs for Innovation (201210226016).

Author contributions

Conceived and designed the experiments: WD, XJZ, LX, GZ. Performed the experiments: HY, KJJ, WD, LX, XSW, DH. Analyzed the data: LKN, HDP, YTW, BL. Wrote the paper: WD, XJZ, TN.

Conflict of interest

The authors declare that they have no conflict of interest.

Notes and references

[a] *College of Bioinformatics Science and Technology, Harbin Medical University, 157 Baojian Road, Harbin, China.*
Corresponding author E-mail: Dong Wang, wangdong@ems.hrbmu.edu.cn or Zheng Guo, guoz@ems.hrbmu.edu.cn or Xia Li, lixia@hrbmu.edu.cn; Tel: 86 045186699584

[b] *College of Bioengineering, Henan University of Technology, Zhengzhou, China*
Corresponding author E-mail: Jianzhen Xu, xujz0451@gmail.com

[c] *Department of Chemical Pathology, the Chinese University of Hong Kong, Hong Kong, China*

† *These authors contributed equally to this work.*

1. Quackenbush, J. *N Engl J Med* **354**, 2463-2472 (2006).
2. Attiyeh, E. F. *et al. Genome Res* **19**, 276-283 (2009).
3. van de Wiel, M. A., Picard, F., van Wieringen, W. N. & Ylstra, B. *Brief Bioinform* **12**, 10-21 (2011).
4. Bock, C. *Nat Rev Genet* **13**, 705-719 (2012).
5. Pritchard, C. C., Cheng, H. H. & Tewari, M. *Nat Rev Genet* **13**, 358-369 (2012).
6. Calin, G. A. *et al. Proc Natl Acad Sci U S A* **101**, 11755-11760 (2004).
7. Yu, Y. P. *et al. Carcinogenesis* **26**, 471-479 (2005).
8. Carter, N. P. *Nat Genet* **39**, S16-21 (2007).
9. Laird, P. W. *Nat Rev Genet* **11**, 191-203 (2010).
10. Network, C. G. A. *Nature* **490**, 61-70 (2012).
11. Quackenbush, J. *Nat Genet* **32 Suppl**, 496-501 (2002).
12. Smyth, G. K. & Speed, T. *Methods* **31**, 265-273 (2003).
13. Wang, D. *et al. Mol Biosyst* **8**, 818-827 (2012).
14. Wang, D. *et al. Comput Biol Chem* **35**, 126-130 (2011).
15. Loven, J. *et al. Cell* **151**, 476-482 (2012).
16. Wu, Y. *et al. Molecular BioSystems*, DOI:10.1039/C3MB70524B (2014).
17. Liu, H. *et al. Nat Cell Biol* **14**, 567-574 (2012).
18. Dang, C. V. *Cell* **149**, 22-35 (2012).
19. Cascon, A. & Robledo, M *Cancer Res* **72**, 3119-3124 (2012).
20. Barrett, T. *et al. Nucleic Acids Res* **33**, D562-566 (2005).
21. Ein-Dor, L., Zuk, O. & Domany, E. *Proc Natl Acad Sci U S A* **103**, 5923-5928 (2006).
22. Zhang, M. *et al. Bioinformatics* **24**, 2057-2063 (2008).
23. Nelson, P. T., Wang, W. X., Wilfred, B. R. & Tang, G. *Biochim Biophys Acta* **1779**, 758-765 (2008).

24.Meyer, S. U., Pfaffl, M. W. & Ulbrich, S. E. *Biotechnol Lett* **32**, 1777-1788 (2010).
25.Taylor, I. W. *et al. Nat Biotechnol* **27**, 199-204 (2009).
26.Subramanian, A. *et al. Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
27.Przytycka, T. M., Singh, M. & Slonim, D. K. *Brief Bioinform* **11**, 15-29 (2010).
28.Choi, J. K., Yu, U., Yoo, O. J. & Kim, S. *Bioinformatics* **21**, 4348-4355 (2005).
29.Wang, D. *et al. Gene* **506**, 36-42 (2012).
30.Hanahan, D. & Weinberg, R. A. *Cell* **144**, 646-674 (2011).
31.Segal, E., Friedman, N., Koller, D. & Regev, A. *Nat Genet* **36**, 1090-1098 (2004).
32.Yao, C. *et al. PLoS One* **7**, e29686 (2012).
33.Kulis, M. & Esteller, M. *Adv Genet* **70**, 27-56 (2010).
34.Sincic, N. & Herceg, Z. *Curr Opin Oncol* **23**, 69-76 (2011).
35.Beroukhim, R. *et al. Nature* **463**, 899-905 (2010).
36.Riddick, G. & Fine, H. A. *Nat Rev Neurol* **7**, 439-450 (2011).
37.Martin, C. M. *et al. Methods Mol Biol* **511**, 333-359 (2009).
38.Jiang, L. *et al. Genome Res* **21**, 1543-1551 (2011).
39.Benes, V. & Muckenthaler, M. *Trends Biochem Sci* **28**, 244-249 (2003).
40.Hill, A. A. *et al. Genome Biol* **2**, RESEARCH0055 (2001).
41.Wang, K. & Bucan, M. *CSH Protoc* **2008**, pdb top46 (2008).
42.Nowak, D., Hofmann, W. K. & Koeffler, H. P. *Transfus Med Hemother* **36**, 246-251 (2009).
43.Naef, F., Hacker, C. R., Patil, N. & Magnasco, M. *Genome Biol* **3**, RESEARCH0018 (2002).
44. Irizarry, R. A. *et al. Biostatistics* **4**, 249-264 (2003).
45.Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P.. *Bioinformatics* **19**, 185-193 (2003).
46.Li, C. & Wong, W. H. *Proc Natl Acad Sci U S A* **98**, 31-36 (2001).
47.Calza, S., Valentini, D. & Pawitan, Y. *BMC Bioinformatics* **9**, 140 (2008).
48.Suo, C., Salim, A., Chia, K. S., Pawitan, Y. & Calza, S. *RNA* **16**, 2293-2303.
49.Tusher, V. G., Tibshirani, R. & Chu, G. *Proc Natl Acad Sci U S A* **98**, 5116-5121 (2001).

Figure legends:

Fig. 1 An illustration of how normalization affects data interpretation when global signals are comparable between two states. The yellow and blue samples represent different conditions. The common assumptions for normalization are reasonable if similar global signal distributions are seen in the different conditions. In such cases, normalization has little influence on the interpretation of expression data.

Fig. 2 An illustration of how normalization affects data interpretation when global raw signals are significantly different between two states. (A) The yellow and blue samples represent cancer samples and normal samples with large differences in signal patterns. The signal intensities were normalized across all arrays to have the same distribution. (B) A gene shows strong up-regulation in cancer samples in the raw signals. Though normalization may reduce the size of the difference, this gene could be still selected as a differential up-regulated gene after normalization. (C) A gene shows moderate up-regulation in cancer samples in the raw signals. After normalization, it cannot be identified as a differentially expressed gene. (D) A gene shows little difference in expression between cancer samples and normal samples in the raw signals. After normalization, it may be identified as a differential down-regulated gene.

Fig. 3 Effects of RMA normalization on expression directions in the mRNA colon34 dataset.

Fig. 4 Normalization may change the expression directions of DEGs in the mRNA Colon34 dataset. The overlap-consistent area represents the intersection of the sets of DEGs selected before and after normalization that have the same expression direction. Overlap-inconsistent represents the intersection of DEGs selected before and after normalization that have inconsistent expression directions. Non-overlap-up represents the set of up-regulated genes among DEGs that were selected only before or only after normalization. Non-overlap-down represents the set of down-regulated genes among DEGs that were selected only before or only after normalization.

Fig. 5 Normalization may change the expression directions of differentially expressed miRNAs in the miRNA Esophagus152 dataset. Overlap-consistent represents the set of miRNAs that have the same expression direction within the intersection of the sets of differentially expressed miRNAs selected before and after normalization. Overlap-inconsistent represents the set of miRNAs that have inconsistent expression directions within the intersection of the differentially expressed miRNAs selected before and after normalization. Non-overlap-up represents the up-regulated miRNAs among the differentially expressed miRNAs that were selected only before or only after normalization. Non-overlap-down represents the down-regulated miRNAs among differentially expressed miRNAs that were selected only before or only after normalization.

Fig. 6 The density distributions of pair-wise Pearson correlation coefficients before and after normalization of the mRNA colon34 dataset.

Fig. 7 The density distributions of pair-wise Pearson correlation coefficients before and after normalization of the miRNA Esophagus152.

Tables:

Table 1 Comparison of the median raw signal intensities of miRNA expression in cancer samples and normal samples.

Dataset	GSE ID	Median of Cancer	Median of Normal	P value
Colon168	GSE7828	5.95	5.70	6.70E-7
Bladder20	GSE39093	6.01	5.98	0.92
Esophagus32	GSE16456	3.75	3.61	0.60
Esophagus152	GSE13937	4.09	3.91	1.19E-3
Esophagus206	GSE6188	7.63	7.32	1.03E-3
Gastric56	GSE23739	5.73	5.72	0.77
Liver146	GSE21362	5.78	5.78	0.38
Liver184	GSE22058	7.51	7.37	0.063
Lung54	GSE14936	4.67	4.52	0.29
Prostate40	GSE23022	4.97	5.01	0.18
Prostate56	GSE21036	6.01	6.01	0.34
Renal30	GSE41282	5.51	5.54	0.39

Note: Median of Cancer represents the median of the raw signal intensities in the cancer samples. Median of Normal represents the median of the raw signal intensities in the normal samples. P value represents the P value of the medians between cancer and normal samples according to the Wilcoxon rank-sum test.

Table 2 Comparison of the medians and 75th quantile values of raw signal intensities for copy number variation data in cancer samples and normal samples.

Dataset	GSE ID	Median			75th quantile		
		Cancer	Normal	P value	Cancer	Normal	P value
Colon188	GSE11417	11.17	10.89	0.015	11.56	11.54	0.024
Gastric166	GSE31168	8.30	8.48	0.22	8.92	8.73	0.29
MDS58	GSE31174	8.61	8.75	0.63	9.02	9.00	0.43
Prostate36	GSE29569	8.26	8.52	0.72	8.92	8.77	0.54
Prostate40Nsp	GSE12702	7.47	7.13	0.32	8.01	7.63	0.36
Prostate40Sty		7.86	7.67	0.35	8.19	7.92	0.38
Prostate76	GSE18333	8.75	8.77	0.95	9.24	9.38	0.68
Sarcoma410	GSE25540	6.67	6.83	0.20	7.54	7.65	0.37
Renal44	GSE21123	9.24	9.53	0.53	9.59	9.80	0.65

Note: Median of Cancer represents the median of the raw signal intensities in the cancer samples. Median of Normal represents the median of the raw signal intensities in the normal samples. The 75th quantile of cancer is the 75th quantile value of the raw signal intensities in the cancer samples. The 75th quantile of normal is the 75 quantile value of the raw signal intensities in the normal samples. P value: the P value of the medians/75th quantiles between cancer and normal samples according to the Wilcoxon rank-sum test.

Table 3 The high-throughput omics datasets analyzed in this study.

mRNA	GSE ID	miRNA	GSE ID
Breast22	GSE10780	Bladder20	GSE39093
Breast24	GSE16873	Colon168	GSE7828
Colon34	GSE18105	Esophagus32	GSE16456
Colon64	GSE8671	Esophagus152	GSE13937
Esophagus24	GSE29001	Esophagus206	GSE6188
Esophagus34	GSE20347	Gastric56	GSE23739
Esophagus106	GSE23400	Liver146	GSE21362
Gastric24	GSE19826	Liver184	GSE22058
Gastric62	GSE13911	Lung54	GSE14936
Hnc44	GSE6631	Prostate40	GSE23022
Liver20	GSE29721	Prostate56	GSE21036
Lung54	GSE7670	Renal30	GSE41282
Lung60	GSE31552	Copy number variation	
Lung66	GSE10072	Colon188	GSE11417
Lung88	GSE18842	Gastric166	GSE31168
Otscc24	GSE9844	Mds58	GSE31174
Otscc40	GSE13601	Prostate36	GSE29569
Pancreatic30	GSE16515	Prostate40Nsp Prostate40Sty	GSE12702
Pancreatic78	GSE15471		
Pancreatic90	GSE28735	Prostate76	GSE18333
Prostate116	GSE6919	Renal44	GSE25540
Ptc40	GSE29265	Sarcoma410	GSE21123
Renal20	GSE6344		

Note: Each dataset is denoted using the following nomenclature: cancer type is followed by the total number of samples. Profiles of copy number variation datasets (GSE12702) were generated on two Affymetrix platforms (Affymetrix Mapping 250K Nsp/Sty SNP Array).

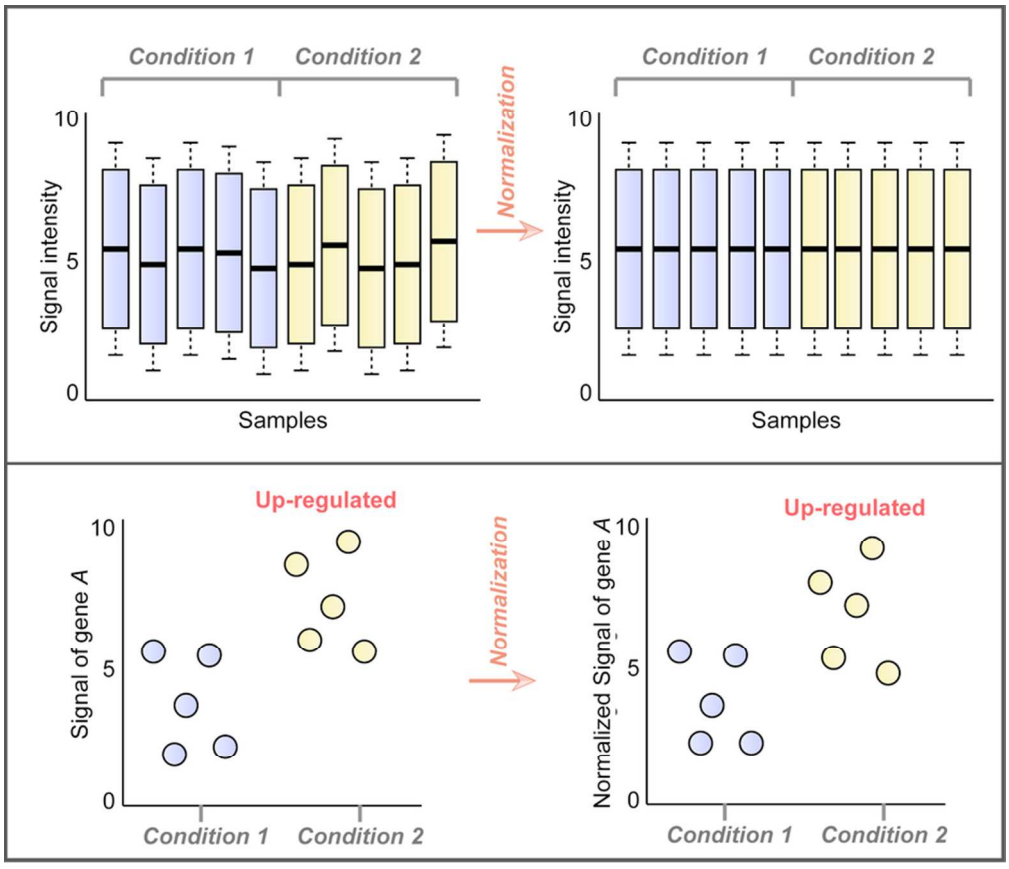


Figure 1
78x67mm (300 x 300 DPI)

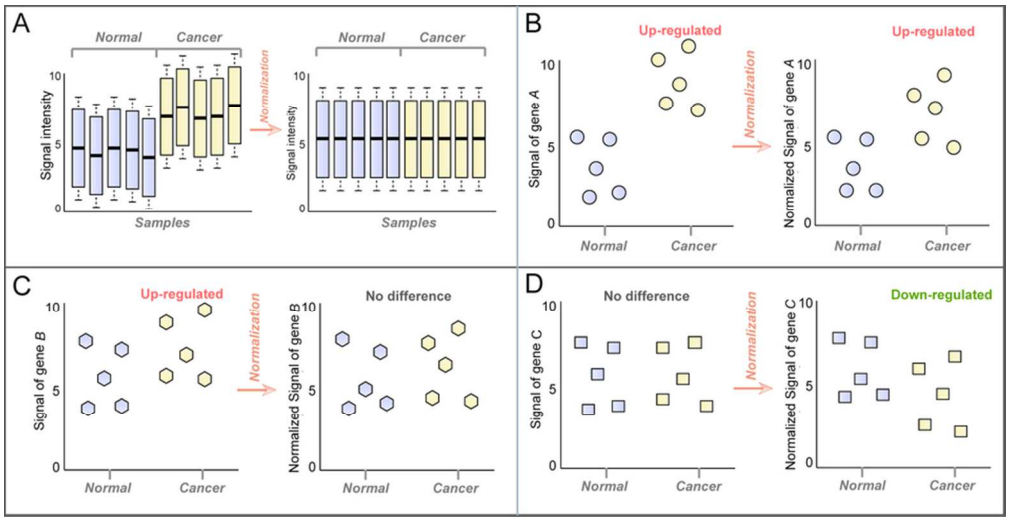


Figure 2
80x41mm (300 x 300 DPI)

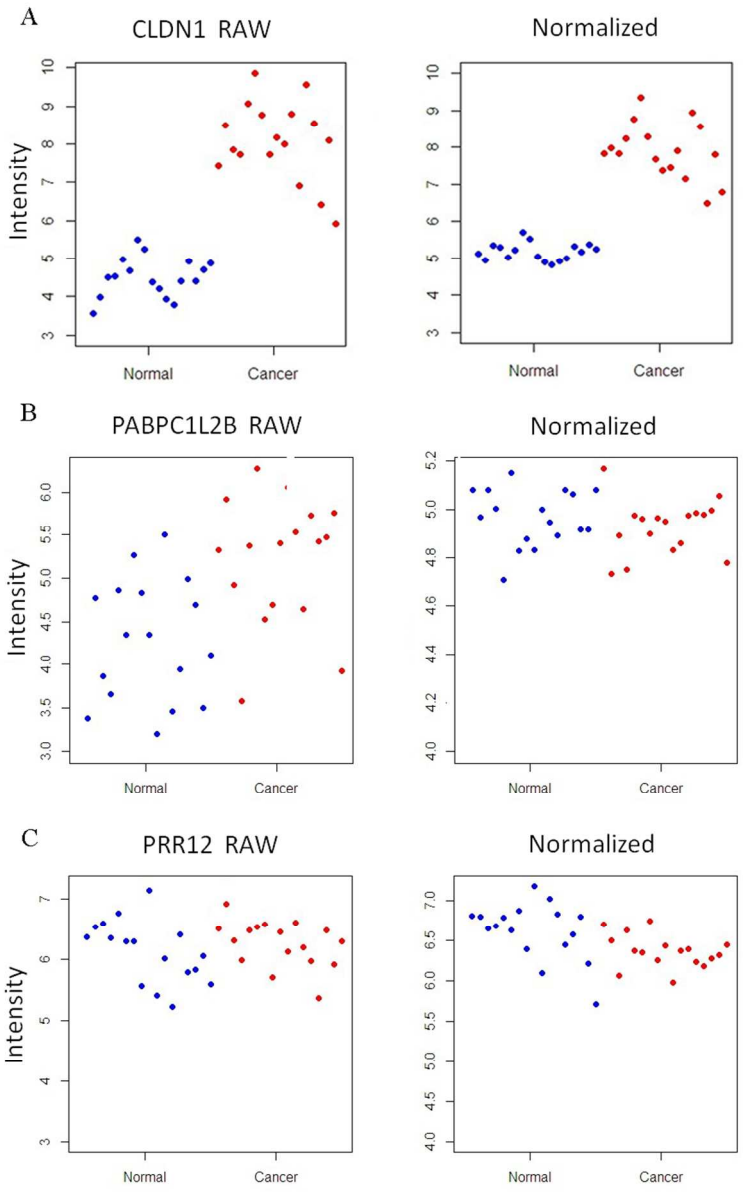


Figure 3
76x118mm (300 x 300 DPI)

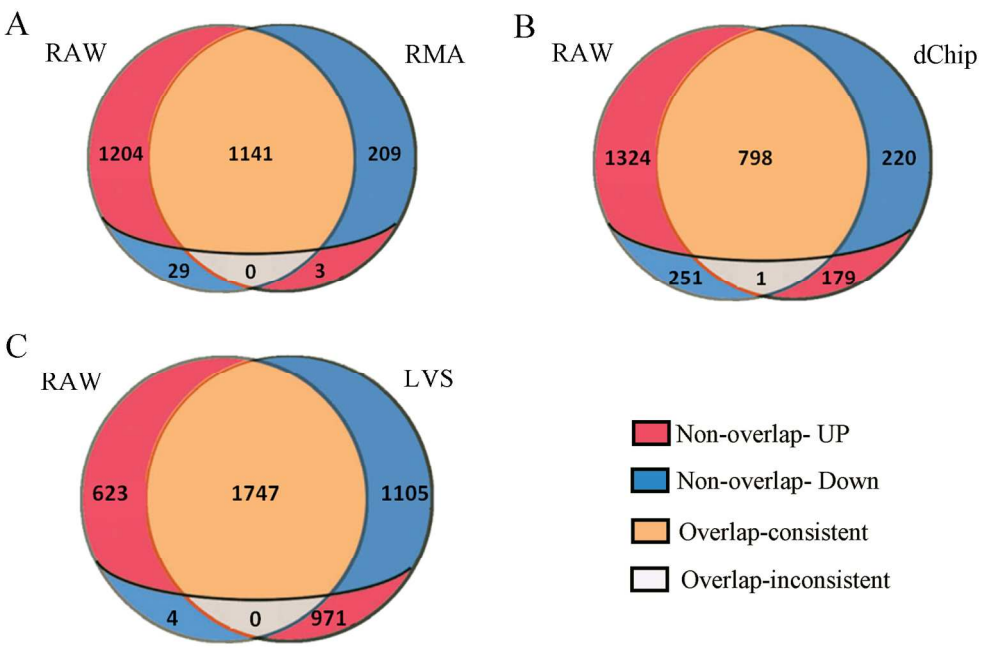


Figure 4
187x121mm (300 x 300 DPI)

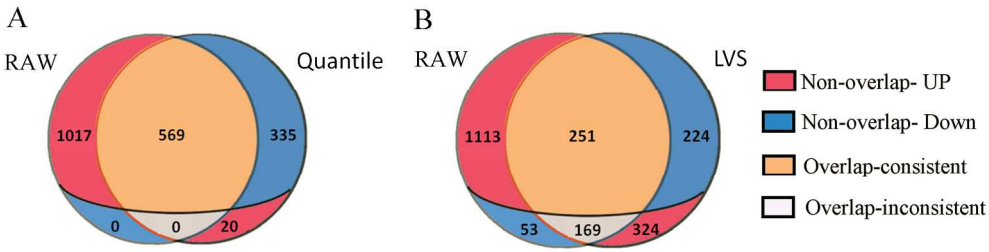


Figure 5
217x55mm (300 x 300 DPI)

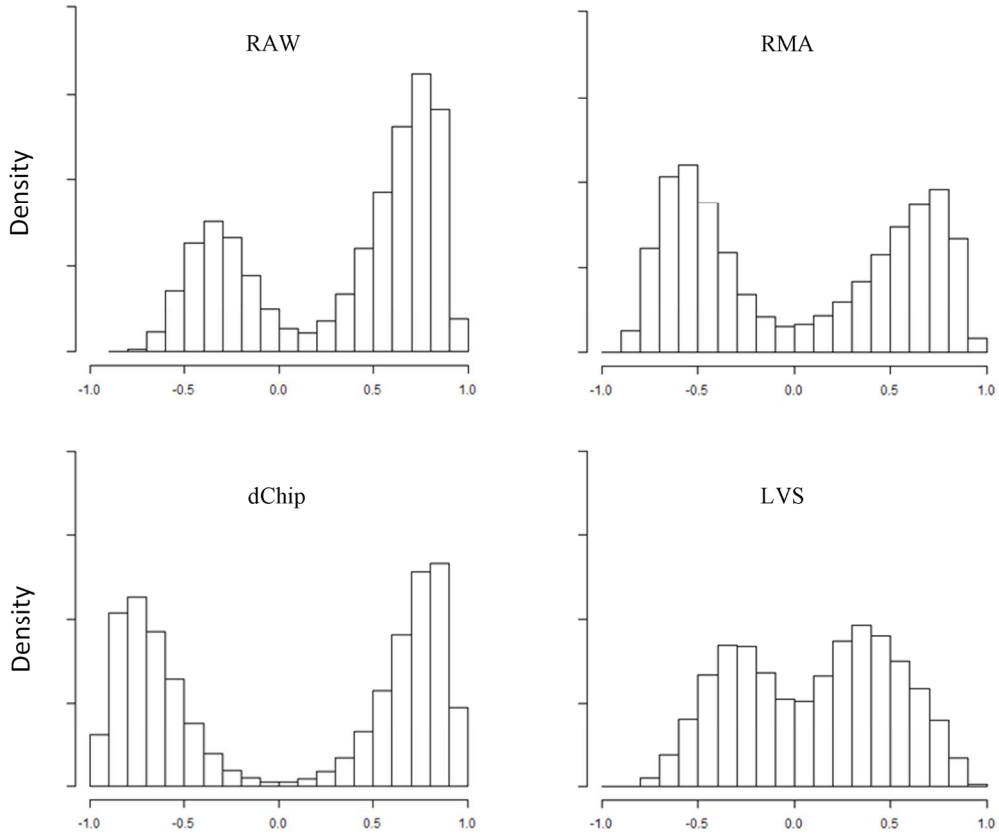


Figure 6
169x141mm (300 x 300 DPI)

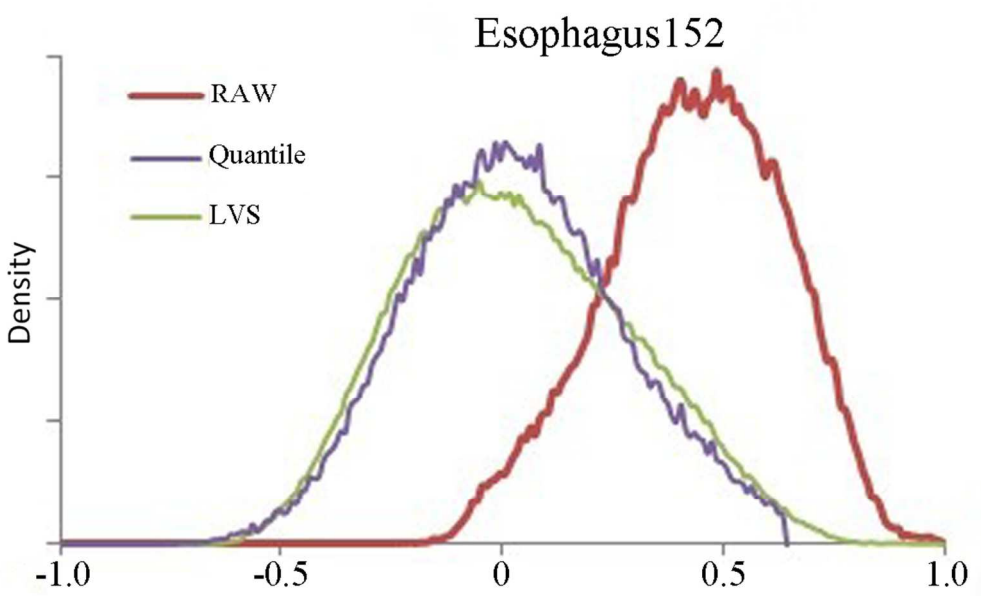


Figure 7
98x60mm (300 x 300 DPI)