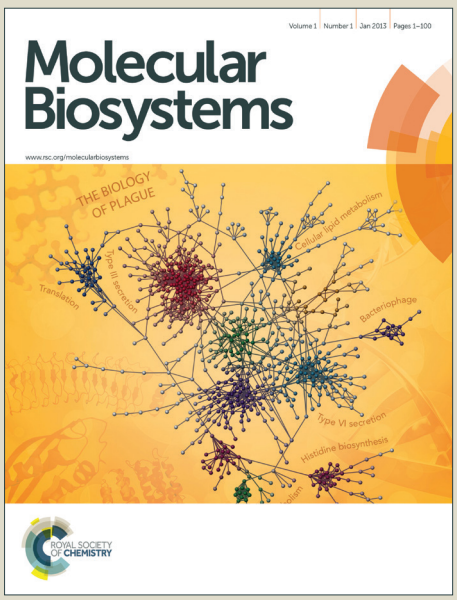
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Molecular BioSystems

RSCPublishing

COMMUNICATION

Global gene expression distribution in non-cancerous complex diseases

Cite this: DOI: 10.1039/x0xx00000x

Received ooth January 2012,
Accepted ooth January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

Yun Wu,[a†] Nana Jin,[a†] Haiyang Zhu,[b†] Chunmiao Li,[a] Nannan Liu,[a] Yan Huang,[a] Zhengqiang Miao,[a] Xiaoman Bi,[a] Deng Wu,[a] Xi Chen,[a] Yun Xiao,[a] Dapeng Hao,[a] Chuanxing Li,[a] Binsheng Gong,[a] Shaojun Zhang,[a] LiweiZhuang,*[a] Kongning Li*[a] and Dong Wang*[a]

For gene expression in non-cancerous complex diseases, we systemically evaluated the sensitivities of biological discoveries to violation of the common normalization assumption. Our results indicated gene expression may be widely up-regulated in digestive system and musculoskeletal diseases. However, global signal intensities showed little difference in other four disease types.

Gene expression analysis provides quantitative and systematic characterization about the population of transcriptome species in a tissue and cell.[1] Monitoring the global transcriptome expression by microarray has had a tremendous influence on modern biological research.[2-4] However, it is known that microarray experiments are subject to multiple sources of technical variations.[5] Data normalization is supposed to adjust the global properties of measurements of individual samples so that they can be more appropriately compared by removing large technical variations.[5] Thus, data normalization plays a critical role in minimizing the impact of technical variations.

The common normalization methods assume that most genes are not differentially changed and the numbers of up-and down-regulated genes are roughly equal.[6, 7] Thus the distributions of global signal intensities for each experiment should be similar and the signal intensities for different samples from different experiments should be scaled to have the same or similar median or average value.[5, 8, 9] However, emerging evidences suggested this commonly used assumption may not hold true under some situations. Such as, based on 16 pair-matched normal and cancer samples gene expression datasets, previously we observed extensive increase of microarray signals in cancers datasets.[10] Subsequently, Love´n, et al. also found that cells with high levels of c-Myc can amplify their gene expression program, producing two to three times more total RNAs and generating cells that were larger than their low-Myc counterparts.[1] Thus, under above circumstances, normalization would distort the global data distribution and lead to erroneous interpretations of gene expression profile.[1, 10]

With the widespread use of gene expression data, many researchers start to study non-cancerous complex disease by expression profiles, such as digestive system and musculoskeletal diseases.[11, 12] It is not certain how prevalent transcription increases in diverse non-cancerous complex diseases and misinterpretation of genome-wide expression data. More importantly, for gene expression array generated from high throughput platforms, what are the global features in non-cancerous complex diseases compared to normal samples? These issues are fundamental questions and basically related to all subsequent data analysis and interpretations, but surprisingly they have not been systematically analysed until now. In this study, using the NCBI GEO database,[13] we analysed the global gene expression distribution for unbiased collected 21 Affymetrix single channel datasets for six non-cancerous complex disease types that each dataset must include at least eight samples for each state (normal or disease), including seven datasets for digestive system disease, three for female urogenital diseases and pregnancy complication, neuropsychiatric disorder, respiratory tract disease, skin disease, respectively, two for musculoskeletal disease (Table 1). Especially, four pair-matched datasets from the same individuals for female urogenital diseases and pregnancy complication, musculoskeletal disease, neuropsychiatric disorder and skin disease were included. For the signal intensities of these gene expression data, we only used perfect match (PM) probe intensities as signal intensities because it has been shown that ignoring the mismatch (MM) values is preferable for background correction.[7, 14]

In our previous work, the results showed genes were extensive up-regulated in a high proportion of cancers. Especially in digestive system cancer (colon, esophagus and pancreas, etc.), gene expression profiles were significantly extensive up-regulated in five of eight datasets.[10] However, it is not certain how prevalent global signal intensities increased is as well as that in non-cancerous digestive system diseases. Thus, we conducted gene expression analysis on seven datasets from four types of digestive system diseases with normal and disease samples, collected from the Gene Expression Omnibus (GEO) database,[13] including hepatitis B virus-associated acute liver failure(HBV-ALF), inflammatory bowel disease (IBD), irritable bowel syndrome (IBS), and ulcerative colitis (UC). For each dataset, we computed the median of the raw signal intensities in each sample and compared the medians between the normal and disease samples. As shown in Fig. 1, the median of raw signal intensities in the disease samples increased in six of the seven datasets. The increase in the median of the raw signal intensities in the disease state was significant ($P<0.05$) in one datasets and marginally significant ($p<0.1$) in another two datasets according to the Wilcoxon rank-sum test: HBV-ALF27 (P=4.50E-06), IBD58 (P=0.069) and UC54 (P=0. 063) (Table 2).

Table 1 The microarray datasets that were analysed in this study.

Dataset	GEO Accession number	Dataset	GEO Accession number
HBV-ALF26	GSE38941	AD20	GSE4757
IBD23	GSE4183	Asthmatics86	GSE4302
IBD58	GSE10616	COPD38	GSE37768
IBS221	GSE36701	COPD54	GSE8545
JIA27	GSE15645	Endometriosis20	GSE7305
Tendinopathy46	GSE26051	PCOS23	GSE10946
UC20	GSE22619	PCOS29	GSE6798
UC26	GSE9452	PD20	GSE20146
UC54	GSE13367	Psoriasis48	GSE41662
		Psoriasis54	GSE14905
		Psoriasis122	GSE13355
		Schizophrenic51	GSE17612

Each dataset is denoted using the following nomenclature: disease type followed by the total number of samples.

Totally, the gene expression signals of disease samples tended to significantly or marginally significantly increase in nearly half of the seven datasets (3/7 =43%). Due to the low statistical power of detecting significant differences in small samples,[15, 16] if we focused on the datasets with larger sample size (sample size at least 25), the percentage was up to 60% (3/5). Therefore, similarly with the normal and cancer samples, it might be also misleading that all of the arrays in digestive system disease should have the same or similar gene expression distribution regardless of the physiological state. Normalizing the expression data by common normalization methods would lead to erroneous results in gene expression analysis, especially for the large sample size datasets.[1, 10]

Further, using the same criteria with digestive system datasets, we unbiased collected two single channel musculoskeletal disease datasets for human tendinopathy and polyarticular juvenile idiopathic arthritis (JIA). Similar results were observed that the median of the raw signal intensities in the disease samples increased in both of two datasets. Further, the increase in the medians of raw signal intensities between normal and disease samples was significant ($P<0.05$) in two musculoskeletal disease datasets by the Wilcoxon rank-sum test: Tendinopathy46 ($P=0.030$), JIA27 ($P=0.019$). Especially in one of these two datasets (Tendinopathy46), the significantly increase was identified in disease samples compared to pair-matched normal samples from 23 patients. Hence, the gene expression signals of disease samples tended to extensive increase in musculoskeletal diseases, indicating a non-negligible adverse impact on gene expression analysis that simply pre-processing the expression data by common normalization methods (Fig. 1, Table 2).

By focusing on the five datasets with significantly/marginally significantly increases in their probe intensities in diseases samples, we analyzed the effects of data normalization on the percentage of up- and down- regulated genes between raw signal intensities and normalized intensities. By using RMA normalization method, our results showed the percentage of up-regulated genes in raw signal intensities was 84%, 95%, 92%, 98% and 99% in HBV-ALF26, IBD58, JIA27, Tendinopathy46 and UC54, respectively, while it decreased around 50% (49%, 57%, 55%, 43% and 45%) in normalized samples, indicating expression data may be over-normalized by common normalization methods. Then, we further analyzed the number of up- and down- regulated differentially expressed genes (DEGs) selected by SAM (FDR=0.1) between raw signal intensities and normalized intensities. Our results showed that in the raw signal intensities, the number of up-regulated DEGs was 11068, 11089, 11657, 10844, and 11289 while the number of down-regulated DEGs was 2174, 319, 9, 0, and 15 in HBV-ALF26, IBD58, JIA27, Tendinopathy46 and UC54, respectively, suggesting that the number of up- and down- regulated DEGs was asymmetry and a large fraction (more than half) of the overall number of genes showed differentially expressed in the raw signal intensities. However, after normalization, the number of up- and down-regulated DEGs was 7900 and 8290, 7603 and 3940, 1839 and 1780, 830 and 1333, 1309 and 1003 in HBV-ALF26, IBD58, JIA27, Tendinopathy46 and UC54, respectively, suggesting that the number of up- and down- regulated DEGs shows no significant difference between samples and DEGs still were a big fraction of the overall number of genes in most of these five datasets.

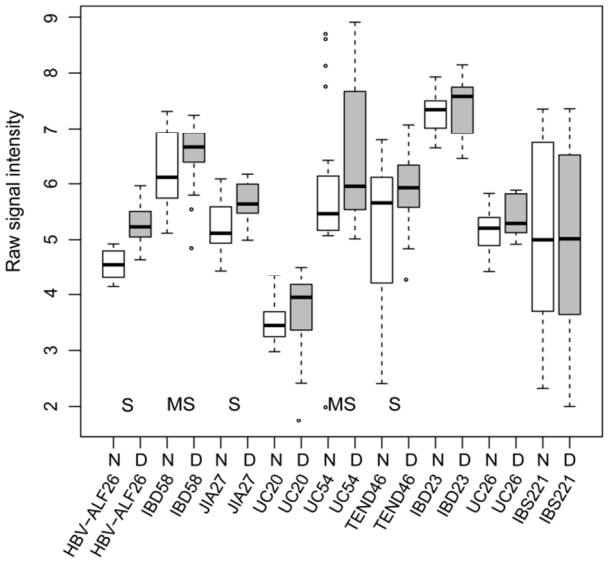


Fig. 1 The distributions of the raw signal intensities for the normal (white) and disease (grey) states in digestive system and musculoskeletal diseases. For each dataset, the raw signal intensities were averaged across all of the samples in each state. The datasets were ranked in descending order of the differences between the medians of the raw signal intensities of the normal and disease states. Three datasets with significant differences in the median are marked by (S), and two datasets with marginally significant differences are marked by (MS).

Table 2 Comparison of the medians of raw signal intensities of gene expression in digestive system and musculoskeletal disease datasets.

Disease Type	Datasets	Normal	Disease	P value
digestive system disease	HBV-ALF26	4.55	5.23	4.50E-06
	IBD23	7.35	7.58	0.64
	IBD58	6.12	6.66	0.069
	IBS221	5.06	5.04	0.51
	UC20	3.45	3.95	0.17
	UC26	5.21	5.29	0.16
	UC54	5.47	5.96	0.063
musculoskeletal disease	JIA27	5.12	5.64	0.019
	Tendinopathy46	5.66	5.93	0.030

HBV-ALF, IBD, IBS and UC represent hepatitis B virus (HBV)-associated acute liver failure (ALF), inflammatory bowel disease, irritable bowel syndrome and ulcerative colitis respectively of digestive system diseases. JIA represents polyarticular juvenile idiopathic arthritis of musculoskeletal disease.

Next, for other non-cancerous complex diseases, we collected a total of 12 datasets for female urogenital diseases and pregnancy complications, neuropsychiatric disorders, respiratory tract diseases and skin diseases, including Alzheimer's disease (AD), asthmatics, chronic obstructive pulmonary disease (COPD), endometriosis, Parkinson's disease (PD), polycystic ovary syndrome (PCOS),

psoriasis and schizophrenia. Especially, three pair-matched datasets from the same individuals were included (AD20, Psoriasis48 and Psoriasis20). On the contrary with digestive system and musculoskeletal diseases, we found non-significant differences in medians of the raw signal intensities between normal and disease samples for these 12 datasets (Table 3). Similar as the methylation datasets between normal and cancer samples, our previous results [10] also demonstrated that, in all of the eight analysed methylation datasets, the medians of the raw signal intensities in the cancer samples were not significantly different (P>0.05) from those in normal samples. These results suggested that the common assumption normalization may bring more positive effects in reducing technical variations than negative effects in removing biological signal in these four disease types.

Table 3 Comparison of the medians of raw signal intensities of gene expression in other non-cancerous disease datasets.

Disease Type	Datasets	Normal	Disease	P value
female urogenital disease and pregnancy complication	Endometriosis20	4.83	4.86	0.52
	PCOS23	5.41	3.96	0.41
	PCOS29	5.52	5.66	0.73
neuropsychiatric disorder	AD20	4.37	4.48	0.97
	PD20	3.75	4.20	0.14
	Schizophrenic51	7.33	7.36	0.87
respiratory tract disease	Asthmatics86	6.18	6.12	0.67
	COPD38	4.86	5.11	0.16
	COPD54	5.49	5.35	0.95
skin disease	Psoriasis48	7.70	7.73	0.57
	Psoriasis54	5.13	5.09	0.78
	Psoriasis122	5.21	5.51	0.48

PCOS represents polycystic ovary syndrome of female urogenital diseases and pregnancy complications. AD and PD represent Alzheimer's disease and Parkinson's disease of neuropsychiatric disorders. COPD represents chronic obstructive pulmonary disease of respiratory tract disease.

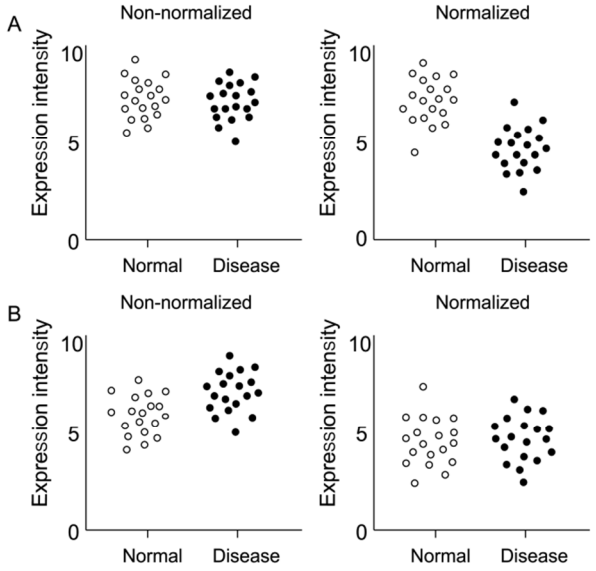


Fig. 2 An illustration of normalization causing over-normalizing signals. (A) A gene is detected as a down-regulated differential expressed gene, after normalization, the raw signal intensities in disease samples are similar with those in normal samples. (B) A gene is detected as an up-regulated differential expressed gene in non-normalized data, after normalization, it shows the similar distribution in disease and normal samples as non-differential expressed gene. The left parts of panels A and B show the raw signal intensities and the right parts show the normalized intensities.

Our previous and subsequent researches exhibited extensive increasing of microarray signals in a high proportion of cancer datasets.[1, 10] In this work, our results similarly showed that gene expression signals extensively up-regulated in five of nine digestive system and musculoskeletal disease datasets, indicating expression data may be over-normalized by common normalization methods. As shown in Fig. 2, normalizing all arrays to have the same distribution of signal intensities regardless of the disease state tends to result a non-negligible portion of falsely down-regulated differentially expressed genes while missing a number of truly up-regulated differentially expressed genes. In Fig. 2A, one gene has little difference as non-differential expressed gene between normal and disease samples in raw signal intensities. After normalization, it could be identified as a down-regulated differential expressed gene. In the other situation (Fig. 2B), another gene has moderate difference as an up-regulated differential expressed gene between normal and disease samples in raw signal intensities. After normalization, it could not be identified as a differential expressed. Accordingly, normalizing should be more precautious in gene expression analysis of these diseases and thus produce more accurate assessments of changes in gene expression programs. Furthermore, because it has been suggested that the variability in the microarray data by technical noise might be lower than the biological variation and its role in statistical data analysis might not be critical[17], we should pay more attention to optimizing experimental designs, stringently randomizing potential experimental artifacts across biological groups, using sufficient sample sizes or containing more RNA in diseased tissues and more conducive to probe hybridization. On the contrary, no significant differences were found in medians between normal and disease samples in other 12 non-cancerous complex diseases datasets, indicating that common normalization assumption may bring more positive effects on reducing technical variations than negative effects on removing biological signals. Also the spike-in controls, as suggested by Love´n, et al and others,[1, 18-20] may be an indispensable, robust, cross-platform, quality control method to enable more accurate detection of disease-associated gene in transcriptome data.

Along with the rapid spring-up of high-throughput technologies, high-throughput arrays as a potential biological tool have been increasingly used in the analysis of transcriptome and genome.[2, 8, 21-23] Gene and miRNA expression profiles offer quantitative information of RNA in a cell or tissue;[2, 24] similarly, methylation arrays and SNP technology were developed for investigating the methylation status and copy number variations (CNVs) on a genome-wide scale. Hence, besides gene expression, for miRNA expression,[25, 26] methylation[27] and copy-number variations[8] array data signals generated from high throughput platforms, how the raw signal intensities distribution in non-cancerous complex diseases also need to validate in a warrant future detailed studies.

Acknowledgements

This work was supported by National Natural Science Foundation of China (31100901), Oversea Scholars Project funded by Education Department of Heilongjiang Province (1155H012).

Notes and references

[a] *College of Bioinformatics Science and Technology, the Fourth Affiliated Hospital, Harbin Medical University, 157 Baojian Road, Harbin, China. Corresponding author E-mail: Dong Wang, wangdong@ems.hrbmu.edu.cn or Kongning Li, kongningli@hotmail.com or Liwei Zhuang, zhuangliweiyd@126.com; Tel: 86 045186699584*

[b] *The infection department, fifth affiliated hospital of Zhengzhou University, Zhengzhou, China*
[†] *These authors contributed equally to this work.*

1. J. Loven, D. A. Orlando, A. A. Sigova, C. Y. Lin, P. B. Rahl, C. B. Burge, D. L. Levens, T. I. Lee and R. A. Young, *Cell*, 2012, **151**, 476-482.
2. J. Quackenbush, *N Engl J Med*, 2006, **354**, 2463-2472.
3. S. Mohr, G. D. Leikauf, G. Keith and B. H. Rihn, *J Clin Oncol*, 2002, **20**, 3165-3175.
4. E. S. Lander, *Nat. Genet.* , 1999, **21**, 3-4.
5. J. Quackenbush, *Nat. Genet.* , 2002, **32 Suppl**, 496-501.
6. C. Li and W. H. Wong, *Proc Natl Acad Sci U S A*, 2001, **98**, 31-36.
7. R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs and T. P. Speed, *Nucleic Acids Res.* , 2003, **31**, e15.
8. M. A. van de Wiel, F. Picard, W. N. van Wieringen and B. Ylstra, *Brief Bioinform*, 2011, **12**, 10-21.
9. G. K. Smyth and T. Speed, *Methods*, 2003, **31**, 265-273.
10. D. Wang, L. Cheng, Y. Zhang, R. Wu, M. Wang, Y. Gu, W. Zhao, P. Li, B. Li, H. Wang, Y. Huang, C. Wang and Z. Guo, *Mol. Biosyst.* , 2012, **8**, 818-827.
11. R. Hasler, Z. Feng, L. Backdahl, M. E. Spehlmann, A. Franke, A. Teschendorff, V. K. Rakyan, T. A. Down, G. A. Wilson, A. Feber, S. Beck, S. Schreiber and P. Rosenstiel, *Genome Res.*, 2012, **22**, 2130-2137.
12. S. Kugathasan, R. N. Baldassano, J. P. Bradfield, P. M. Sleiman, M. Imielinski, S. L. Guthery, S. Cucchiara, C. E. Kim, E. C. Frackelton, K. Annaiah, J. T. Glessner, E. Santa, T. Willson, A. W. Eckert, E. Bonkowski, J. L. Shaner, R. M. Smith, F. G. Otieno, N. Peterson, D. J. Abrams, R. M. Chiavacci, R. Grundmeier, P. Mamula, G. Tomer, D. A. Piccoli, D. S. Monos, V. Annese, L. A. Denson, S. F. Grant and H. Hakonarson, *Nat. Genet.* , 2008, **40**, 1211-1215.
13. T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky and R. Edgar, *Nucleic Acids Res.* , 2007, **35**, D760-765.
14. F. Naef, C. R. Hacker, N. Patil and M. Magnasco, *Genome Biol.*, 2002, **3**, RESEARCH0018.
15. L. Ein-Dor, O. Zuk and E. Domany, *Proc Natl Acad Sci U S A*, 2006, **103**, 5923-5928.
16. M. Zhang, C. Yao, Z. Guo, J. Zou, L. Zhang, H. Xiao, D. Wang, D. Yang, X. Gong, J. Zhu, Y. Li and X. Li, *Bioinformatics*, 2008, **24**, 2057-2063.
17. L. Klebanov and A. Yakovlev, *Biol Direct*, 2007, **2**, 9.
18. L. Jiang, F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, T. R. Gingeras and B. Oliver, *Genome Res.*, 2011, **21**, 1543-1551.
19. A. A. Hill, E. L. Brown, M. Z. Whitley, G. Tucker-Kellogg, C. P. Hunter and D. K. Slonim, *Genome Biol.*, 2001, **2**, RESEARCH0055.
20. V. Benes and M. Muckenthaler, *Trends Biochem. Sci* 2003, **28**, 244-249.
21. C. Bock, *Nat. Rev. Genet.* , 2012, **13**, 705-719.
22. E. F. Attiyeh, S. J. Diskin, M. A. Attiyeh, Y. P. Mosse, C. Hou, E. M. Jackson, C. Kim, J. Glessner, H. Hakonarson, J. A. Biegel and J. M. Maris, *Genome Res.*, 2009, **19**, 276-283.
23. C. C. Pritchard, H. H. Cheng and M. Tewari, *Nat. Rev. Genet.* , 2012, **13**, 358-369.
24. G. A. Calin, C. G. Liu, C. Sevignani, M. Ferracin, N. Felli, C. D. Dumitru, M. Shimizu, A. Cimmino, S. Zupo, M. Dono, M. L. Dell'Aquila, H. Alder, L. Rassenti, T. J. Kipps, F. Bullrich, M. Negrini and C. M. Croce, *Proc Natl Acad Sci U S A*, 2004, **101**, 11755-11760.
25. D. Sarkar, R. Parkin, S. Wyman, A. Bendoraite, C. Sather, J. Delrow, A. K. Godwin, C. Drescher, W. Huber, R. Gentleman and M. Tewari, *Nucleic Acids Res.* , 2009, **37**, e17.
26. S. Pradervand, J. Weber, J. Thomas, M. Bueno, P. Wirapati, K. Lefort, G. P. Dotto and K. Harshman, *RNA*, 2009, **15**, 493-501.
27. P. W. Laird, *Nat. Rev. Genet.* , 2010, **11**, 191-203.

