Data in Brief 6 (2016) 121–123

Contents lists available at ScienceDirect

Data in Brief

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

Data article

Detection bias in microarray and sequencing transcriptomic analysis identified by housekeeping genes

Yijuan Zhang [a], Oluwafemi S. Akintola [a], Ken J.A. Liu [b], Bingyun Sun [a,b,*]

[a] Department of Chemistry, Simon Fraser University, Burnaby, British Columbia, Canada
[b] Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada

ARTICLE INFO

Article history:
Received 6 October 2015
Received in revised form
20 November 2015
Accepted 20 November 2015
Available online 27 November 2015

Keywords:
Transcriptome
Microarray
Sequencing
RNA-seq
Next-generation sequencing
Housekeeping genes

ABSTRACT

This work includes the original data used to discover the gene ontology bias in transcriptomic analysis conducted by microarray and high throughput sequencing (Zhang et al., 2015) [1]. In the analysis, housekeeping genes were used to examine the differential detection ability by microarray and sequencing because these genes are probably the most reliably detected. The genes included here were compiled from 15 human housekeeping gene studies. The provided tables here comprise of detailed chromosomal location, detection breadth, normalized expression level, exon count, total exon length, and total intron length of each concerned gene and their related transcripts. We hope this information can help researchers better understand the differences in gene ontology-bias we discussed (Zhang et al., 2015) [1] and can encourage further improvement on these two technology platforms.

Specifications Table

Subject area	*Biology*
More specific subject area	*Transcriptomics*

DOI of original article: http://dx.doi.org/10.1016/j.gene.2015.09.041
* Corresponding author at: Department of Chemistry, Simon Fraser University, Burnaby, British Columbia, Canada.
E-mail address: bingyun_sun@sfu.ca (B. Sun).

http://dx.doi.org/10.1016/j.dib.2015.11.045

Type of data	*Excel table*
How data was acquired	*Microarray and sequencing*
Data format	*Downloaded from public domain, compiled and analyzed*
Experimental factors	*Gene identifier was unified*
Experimental features	*Analysis of gene chromosomal location, gene structure, and gene expression*
Data source location	
Data accessibility	Data is with the article

Value of the data

- Housekeeping genes are the most reliably detected genes in high throughput fashion that have the least detection errors for examining differences in analysis.
- The detailed value of all concerned factors including the chromosomal location, the exon count, total exon length, total intron length, normalized expression value, detection breadth are provided here in a per gene or per transcript basis such that the data can be further queried or analyzed.
- The information included here should also help further improvement on these two popular technology platforms.

1. Data

Table S1, chromosomal location of housekeeping (HK) genes exclusively detected by MA alone, sequencing alone, as well as jointly. Table S2, exon count, total exon length, total intron length, and GC content of HK genes exclusively detected by MA alone, sequencing alone, as well as jointly. Table S3, detection breadth and the normalized maximum expression quantity of each HK gene exclusively detected by MA alone, sequencing alone, as well as jointly.

2. Experimental design, materials and methods

The data included here were downloaded from 15 published human housekeeping studies, i.e. Warrington [2], Hsiao [3], Eisenberg_03 [4], Tu [5], Dezso [6], She [7], Chang [8], Shyamsundar [9], Zhu_MA, Zhu_EST [10], Podder [11], Reverter [12], Ramskold [13], Eisenberg_13 [14] and Fagerberg [15], in which nine studies used microarray (MA) analysis, i.e. Warrington [2], Hsiao [3], Eisenberg_03 [4], Tu [5], Dezso [6], She [7], Chang [8], Shyamsundar [9], Zhu_MA, and the rest used sequencing analysis. The gene identifiers used in different studies were first converted to entrez gene ID using Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 (http://david.abcc.ncifcrf.gov/) [16,17] as detailed in [1,18]. The chromosomal location was queried against National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov). Genes with unknown genome locations were removed. The obtained entrez gene list was further converted to Refseq mRNA IDs using DAVID, and the Refgene information on exon count, exon starting and ending position as well as the coding sequences were obtained by querying the Refgene information from University of California, Santa Cruz (UCSC) genome browser (http://genome.ucsc.edu/index.html) against the latest human genome assembly (GRCh38) [19]. The total intron length was calculated by the total gene length minus total exon length. The GC content was deduced by the coding sequence only. Again transcripts could not be mapped to Refgene in UCSC database, and those without exon count or exon starting or ending information as well as sequencing information, were removed from the table. The expression quantity was collected from Chang [8], Eisenberg_03 [4], She [7], Warrington [2], Shyamsundar [9] and Fagerberg [15]. The raw expression quantity was first normalized against the maximum value in each individual list to make them comparable. For entrez genes having multiple quantification values in a single list (for example in cases where a single entrez gene ID was mapped to several IDs, each IDs in that particular study had an expression value), the maximum normalized

expression value was used. The detective breadth (DB) [1,18] described the number of studies, in which a HK gene had been identified. For example, if a gene was detected in 8 out of 9 MA studies, its DB value would be 8, and similarly if a gene was detected in 5 out of 6 sequencing studies, its DB value would be 5.

Acknowledgments

This work was financially supported by Simon Fraser University, Stem Cell Network of Canada, Compute Canada, and Westgrid. Y. Z. was supported in part by NNSFC (National Natural Science Foundation of China), Grant no. 21336009.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2015.11.045.

References

[1] Y. Zhang, O.S. Akintola, K.J.A. Liu, B. Sun, Membrane gene ontology bias in sequencing and microarray obtained by housekeeping-gene analysis, Gene 575 (2 Pt 2) (2016) 559–566. http://dx.doi.org/10.1016/j.gene.2015.09.041.
[2] J. Warrington, A. Nair, M. Mahadevappa, M. Tsyganskaya, Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes, Physiol. Genom. 2 (2000) 143–147.
[3] L.-L. Hsiao, F. Dangond, T. Yoshida, R. Hong, R.V. Jensen, J. Misra, W. Dillon, K.F. Lee, K.E. Clark, P. Haverty, A compendium of gene expression in normal human tissues, Physiol. Genom. 7 (2) (2001) 97–104.
[4] E. Eisenberg, E. Levanon, Human housekeeping genes are compact, Trends Genet. 19 (2003) 362–365.
[5] Z. Tu, L. Wang, M. Xu, X. Zhou, T. Chen, F. Sun, Further understanding human disease genes by comparing with housekeeping genes and other genes, BMC Genom. 7 (1) (2006) 31.
[6] Z. Dezső, Y. Nikolsky, E. Sviridov, W. Shi, T. Serebriyskaya, D. Dosymbekov, A. Bugrim, E. Rakhmatulin, R.J. Brennan, A. Guryanov, A comprehensive functional analysis of tissue specificity of human gene expression, BMC Biol. 6 (1) (2008) 49.
[7] X. She, C.A. Rohl, J.C. Castle, A.V. Kulkarni, J.M. Johnson, R. Chen, Definition, conservation and epigenetics of housekeeping and tissue-enriched genes, BMC Genom. 10 (1) (2009) 269.
[8] C.-W. Chang, W.-C. Cheng, C.-R. Chen, W.-Y. Shu, M.-L. Tsai, C.-L. Huang, I.C. Hsu, Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis, PLoS One 6 (7) (2011) e22859.
[9] R. Shyamsundar, Y.H. Kim, J.P. Higgins, K. Montgomery, M. Jorden, A. Sethuraman, M. van de Rijn, D. Botstein, P.O. Brown, J. R. Pollack, A DNA microarray survey of gene expression in normal human tissues, Genome Biol. 6 (3) (2005) R22.
[10] J. Zhu, F. He, S. Song, J. Wang, J. Yu, How many human genes can be defined as housekeeping with current expression data? BMC Genom. 9 (2008) 172.
[11] S. Podder, T.C. Ghosh, Exploring the differences in evolutionary rates between monogenic and polygenic disease genes in human, Mol. Biol. Evol. 27 (4) (2010) 934–941.
[12] A. Reverter, A. Ingham, B.P. Dalrymple, Mining tissue specificity, gene connectivity and disease association to reveal a set of genes that modify the action of disease causing genes, BioData Min. 1 (1) (2008) 8.
[13] D. Ramsköld, E.T. Wang, C.B. Burge, R. Sandberg, An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data, PLoS Comput. Biol. 5 (12) (2009) e1000598.
[14] E. Eisenberg, E.Y. Levanon, Human housekeeping genes, revisited, Trends Genet. 29 (10) (2013) 569–574.
[15] L. Fagerberg, B.M. Hallström, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, M. Habuka, S. Tahmasebpoor, A. Danielsson, K. Edlund, Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics, Mol. Cell Proteom. 13 (2) (2014) 397–406.
[16] D.W. Huang, B.T. Sherman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat. Protoc. 4 (1) (2008) 44–57.
[17] D.W. Huang, B.T. Sherman, R.A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, Nucleic Acids Res. 37 (1) (2009) 1–13.
[18] Y. Zhang, D. Li, B. Sun, Do housekeeping genes exist? PLoS One 10 (5) (2015) e0123691.
[19] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, The human genome browser at UCSC, Genome Res. 12 (6) (2002) 996–1006.