OPEN ACCESS Freely available online

Journal of Data Mining in Genomics & Proteomics

Research Article

Gene Expression Level and Gene Set Enrichment Analysis of Host Genes

Yosuke Kondo, Satoru Miyazaki*

Department of Medicinal and Life Science, Tokyo University of Science, Tokyo, Japan

ABSTRACT

Genome-wide analysis has shown that there are non-protein-coding RNAs (ncRNAs) that are localized on intronic regions of protein-coding genes. The intronic ncRNAs are hosted in introns of protein-coding genes that are referred to as host genes. Our previous study reported genomic features of intronic ncRNA genes and host genes. However, transcriptomic features of host genes have not been investigated. Here we report gene expression level analysis of host genes and investigate biological functions of host genes. Our results showed that gene expression levels of host genes tend to be higher than those of non-host genes. Host genes orthologous between human and mouse have more conserved expression levels than non-host orthologous genes. And host genes with high expression levels involve nervous system, gene expression, protein modification and cytoskeleton whereas there were mostly no enriched biological functions in host genes with low expression levels. These results suggest that host genes have characteristic transcript quantification and biological functions. The characteristics may be useful for further analysis of regulatory ways of host gene expression.

Keywords: Host gene; Intronic ncRNA; Enrichment analysis; Gene expression

INTRODUCTION

Non-protein-coding RNAs (ncRNAs) are widely present and genetically conserved in eukaryotic cells and broadly classified on the basis of their sequence length, structure, genomic localization and biological processes [1]. Intragenic ncRNAs are originated from exons or introns of protein-coding genes and have diverse biological functions [2]. MicroRNA (miRNA) in an intronic region called a mirtron can regulate gene expression of a protein-coding gene by affecting the messenger RNA (mRNA) [3,4]. Circular RNA (circRNA) arises from splicing of a transcript from a protein-coding gene [5]. Genome-wide analysis has proven that thousands of circRNAs in nervous system increase with aging in *Drosophila* species [6].

Host genes are protein-coding genes that contain ncRNAs in the intronic regions. When host genes are expressed, the intronic ncRNA genes are also expressed at the same time. A host gene produces precursor mRNA (pre-mRNA) and splicing of the pre-mRNA makes mature mRNA and also generates intronic ncRNA [7]. Mirtrons can regulate gene expression of a protein-coding gene that are not its host gene. But some mirtrons can regulate their host genes [8,9]. Therefore, gene expression of host genes is possible to be regulated by intronic ncRNAs in a post-transcriptional way. There are independent and dependent transcriptional regulatory ways of intronic ncRNA. Some intronic ncRNA genes share transcriptional regulatory regions with their host genes but some genes have independent expression regulatory elements [10].

Previously, we reported genomic features of intronic ncRNA genes and their host genes [11]. This analysis showed that more than 10% of protein-coding genes were host genes in human. 20% of human host genes were estimated as orthologs of mouse host genes. In addition, human host genes is more likely to be orthologous with mouse host genes if the intronic ncRNA sequences are conserved. We also analyzed biological functions of host genes. The results showed host genes relate with some specific biological functions. However, transcriptomic features of host genes have not been elucidated yet. These features are important to discuss the transcriptional regulation of host genes and predict biological functions of host genes. In this study, we analyze gene expression levels and biological functions of host genes as the first step for clarification of host gene expression.

Correspondence to: Satoru Miyazaki, Faculty of Pharmaceutical Sciences, Department of Medicinal and Life Science, Tokyo University of Science, Tokyo, Japan, E-mail: smiyazak@rs.noda.tus.ac.jp

Received: March 11, 2020; **Accepted:** March 24, 2020; **Published:** March 31, 2020

Citation: Kondo Y, Miyazaki S (2020) Gene Expression Level and Gene Set Enrichment Analysis of Host Genes. J Data Mining Genomics Proteomics. 11:221. DOI: 10.35248/2153-0602.20.11.221.

MATERIALS AND METHODS

Gene expression level analysis

We obtained RNA-seq data from Expression Atlas (https://www.ebi.ac.uk/gxa/home, E-MTAB-3716 and E-MTAB-3718) [12]. Transcript per million (TPM) was used to estimate transcript quantification [13]. Host genes in human or mouse were extracted from our database and protein-coding genes other than host genes were defined as non-host genes. We analyzed distributions of gene expression levels of host and non-host genes and also investigated those of host or non-host genes orthologous between human and mouse. Pearson's correlation coefficients were calculated for gene expression levels of orthologous host and non-host genes between human and mouse.

Gene set enrichment analysis

We used host genes in human for gene set enrichment analysis (GSEA) by gene ontology (GO) terms [14]. Statistical analysis was performed by the Benjamini-Hochberg method and GO terms with p<0.05 were obtained by the goatools python package [15,16]. We divided host genes by a most frequent value of the expression levels and performed GSEA for host genes with high or low expression levels.

RESULTS

Gene expression level of host genes

Our database contains 19,961 human and 22,050 mouse protein-coding genes, which include 2,691 and 1,633 host genes, respectively. Expression levels of the protein-coding genes in human and mouse were obtained from Expression Atlas. We used baseline expression levels that are healthy cell data in human or mouse tissues (frontal lobe, brain, cerebellum, heart, liver, kidney and testis) [17]. Expression levels in TPM were used and logarithm of the expression levels were calculated.

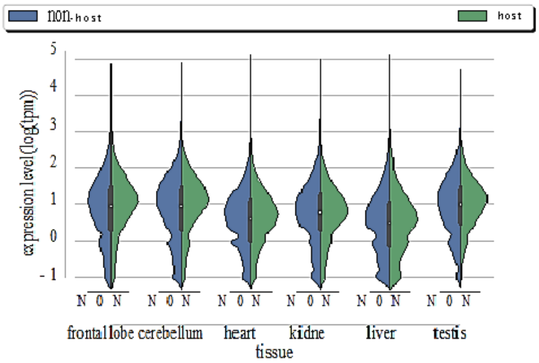

Figure 1: Expression levels of host and non-host genes in human. Horizontal and vertical lines indicate tissues and logarithm of expression levels (tpm), respectively. The horizontal line in each tissue shows the number of genes. Green and blue display host and non-host genes, respectively.

Figure 1 shows distributions of expression levels of host or non-host genes in each human tissue. The most frequent values of the numbers of genes in each tissue were from 0 to 2. The numbers of host genes were larger than those of non-host genes at the most frequent values. On the other hand, there were some smaller peaks around 0 and -1 in non-host genes. In the peaks, the most frequent values of non-host genes around these regions tend to be higher than those of host genes.

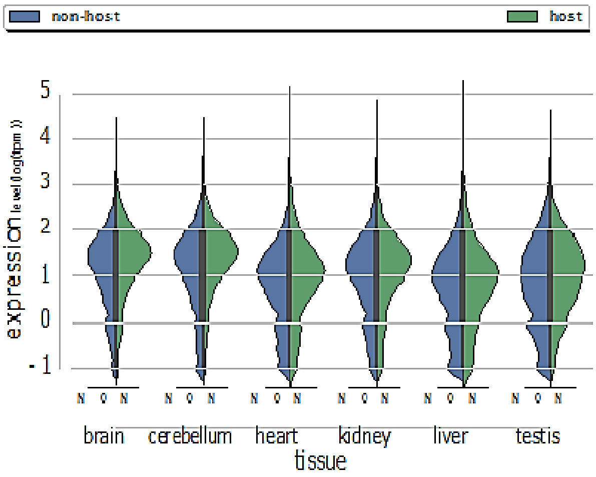

Figure 2: Expression levels of host and non-host genes in mouse. Horizontal and vertical lines indicate tissues and logarithm of expression levels (tpm), respectively. The horizontal line in each tissue shows the number of genes. Green and blue display host and non-host genes, respectively.

Figure 2 shows that host genes in mouse have similar tendency with those in human.

Gene expression level analysis of orthologous host genes

We extracted host genes orthologous between human and mouse.

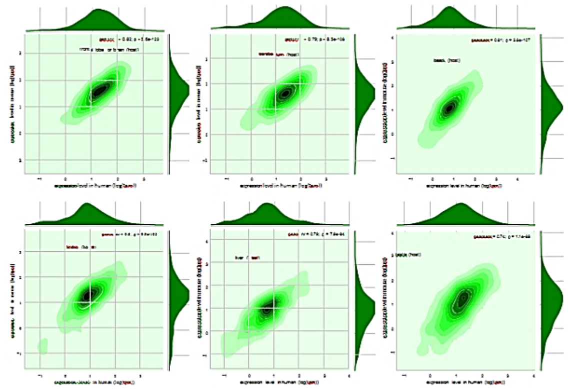

Figure 3: Expression levels of orthologous host genes. Horizontal and vertical lines indicate expression levels in human and mouse, respectively. Expression levels were displayed by logarithm of transcript per million. The number of orthologous host genes are greater if the green color is darker.

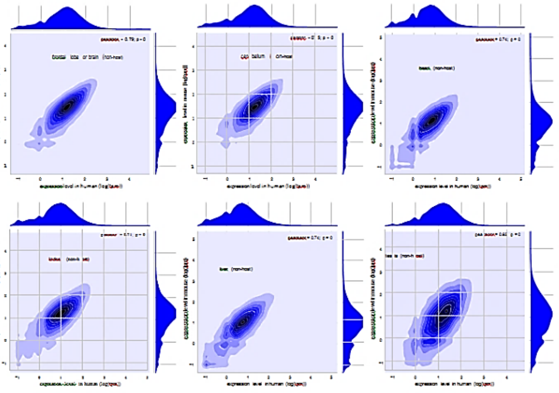

Figure 4: Expression levels of orthologous non-host genes. Horizontal and vertical lines indicate expression levels in human and mouse, respectively. Expression levels were displayed by logarithm of transcript per million. The number of orthologous non-host genes are greater if the blue color is darker.

Figures 3 and 4 show that expression levels of both host and non-host genes orthologous between human and mouse have linear relationships. The correlation coefficients of host genes are 0.82, 0.79, 0.81, 0.80, 0.79 and 0.74 in frontal lobe, cerebellum, heart, kidney, liver and testis, respectively. This shows that these values are larger than correlation coefficients of non-host genes because they were 0.79, 0.75, 0.74, 0.71, 0.74 and 0.63 in brain (frontal lobe), cerebellum, heart, kidney, liver and testis, respectively. Expression levels of host genes are more correlated than those of non-host genes.

Enrichment analysis of host genes

GSEA was performed for host genes in human. We used host genes expressed in cerebellum because our previous study showed that host genes involve neuron related terms. There were 98 enriched or purified GO terms in host genes (Table S1). The GO terms contain 22 biological process, 54 cellular component and 22 molecular function terms. As the enriched biological process terms, there were neuron related terms such as neuron projection development. There were gene expression related terms such as viral transcription, translational initiation, translation, positive and negative regulation of transcription by RNA polymerase II, and so on. There were protein modification related terms such as protein phos- phorylation and ubiquitination. As the purified GO terms, there were olfaction related terms such as detection of chemical stimulus involved in sensory perception of smell and olfactory receptor activity.

GSEA was performed for host genes by dividing high and low expression levels. We used the most frequent value of expression levels to divide host genes. In host genes with high expression levels, there were 129 enriched or purified GO terms (Table S2). This showed that the number of enriched or purified GO terms was increased compared with all host genes described above. The GO terms contain 28 biological process, 72 cellular component and 29 molecular function terms. The GO terms showed similar results with all host genes but cytoskeleton related terms such as regulation of filopodium assembly, regulation of microtubule cytoskeleton organization and stress fiber assembly were newly detected. In host genes with low expression levels, there were 7 enriched or purified GO terms (Table S3). The GO terms contain 1 biological process, 3 cellular component and 3 molecular function terms. There were detection of chemical stimulus involved in sensory perception of smell, cytosol, cytoskeleton, centrosome, ATP binding, protein binding and G protein-coupled receptor activity. This showed that the number of enriched or purified GO terms was remarkably decreased compared with all or high expression level host genes.

DISCUSSION

Intronic ncRNAs are in an intragenic region of a protein-coding gene. The intronic ncRNAs are expressed with the protein-coding genes. This inclusive relationship of genes is useful to estimate biological functions of intronic ncRNAs. Our previous study showed that host genes are conserved between species if the intronic ncRNA sequences are conserved and host genes involve specific biological functions. However, we have not analyzed transcriptomic features of host genes.

In order to analyze baseline gene expression, we used RNA-seq data from a variety of healthy cells in human and mouse. The results showed that distributions of expression levels of host genes are not much different among tissues. We then compared expression levels of host genes with those of non-host genes. The comparison showed that expression levels of host genes tend to be higher than those of non-host genes at the most frequent values. This tendency was also observed in expression levels of host and non-host genes in mouse. These results suggest that host genes tend to be high expression levels in human and mouse. We also compared host genes orthologous between human and mouse. The results showed that correlation coefficients of orthologous host genes tend to be larger than those of orthologous non-host genes. This result suggests that expression levels of host genes are more conserved than non-host genes in human and mouse.

In order to investigate biological functions of host genes, we performed GSEA. In our previous study, GSEA was performed by dividing host genes into ortholog and non-ortholog between human and mouse. The previous study showed that host genes are related with neuron related terms. In this study, we used gene expression in cerebellum and divided host genes by their gene expression levels. In host genes with high expression levels, there were many enriched GO terms related with neuron, gene expression, protein modification and cytoskeleton. On the other hand, in host genes with low expression levels, several GO terms were only detected. The results show that the host genes with high expression levels may involve specific biological functions. However, host genes with low expression levels may relate with a variety of biological functions.

CONCLUSION

In conclusion, this study found that gene expression levels are slightly different between host and non-host genes and biological functions of host genes are different between high and low expression levels. These results are useful for further investigation of transcriptomic features of host genes. In future works, we will analyze gene expression difference of host genes and also investigate transcriptional regulation of host genes by using regulatory regions of host genes to predict and clarify biological functions of host genes and intronic ncRNAs.

REFERENCES

1. Mattick JS. The Genetic Signatures of Noncoding RNAs. PLoS Genet. 2009;5(4): e1000459
2. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Morales RD, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc of the Nat Acad of Sci of the USA. 2009;106(28): 11667-11672.
3. Berezikov E, Chung WJ, Willis J, Cuppen E, Lai EC. Mammalian mirtron genes. Molecular Cell. 2007;28(2):328-336.
4. Curtis HJ, Sibley CR, Wood MJA. Mirtrons, an emerging class of atypical mirna. Wiley Interdiscip Rev RNA. 2012;3(5):617-632.
5. Shang Q, Yang Z, Jia R, Ge S. The novel roles of circRNAs in human cancer. Molecular Cancer. 2019;18:1-10.
6. Westholm JO, Miura P, Olson S, Shenker S, Joseph B, Sanfilippo P, et al. Genome-wide Analysis of *Drosophila* Circular RNAs Reveals Their Structural and Sequence Properties and Age-Dependent Neural Accumulation. Cell Reports. 2014;9(5): 1966-1980.
7. He C, Li Z, Chen P, Huang H, Hurst LD, Chen J, et al. Young intragenic miRNAs are less coexpressed with host genes than old ones: implications of miRNA-host gene coevolution. Nucleic Acids Res. 2012;40(9):4002-4012.
8. Qian J, Tu R, Yuan L, Xie W. Intronic mir-932 targets the coding region of its host gene, *drosophila neuroligin2*. Experimental Cell Res. 2012;344(2):183-193.
9. Shimony AS, Shtrikman O, Margalit H. Assessing the functional association of in- tronic mirnas with their host genes. RNA. 2018;24(8):991-1004.
10. Baskerville S, Bartel D. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA. 2005;11(3):241-247.
11. Kondo Y, Hayashi C, Miyazaki S. Comparative analysis of intronic noncoding rna genes among organisms. J Mol Genet Med. 2017;11(271):1747-0862.
12. Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, et al. Expression atlas update-a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. Nucleic Acids Res. 2014;42(1):D926-D932.
13. Wagner GP, Kin K, Lynch VJ. Measurement of mrna abundance using rna-seq data: Rpkm measure is inconsistent among samples. Theory in Biosciences. 2012;131: 281-285.
14. Carbon S, Douglass E, Dunn N, Good B, Harris NL, Lewis CJ, et al. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019;47(D1):D330-D338.
15. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful aprroach to multiple testing. Journal of the Royal Statistical Society Series B-Statistical Methodology. 1995;57(1):289-300.
16. Klopfenstein DV, Zhang L, Pedersen BS, Ramirez F, Vesztrocy AW, Naldi A, et al. GOATOOLS: A Python library for Gene Ontology analyses. Scientific Reports. 2018;18(1):1-7.
17. Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011;478(7369): 343-348.