Gene 576 (2016) 560–570

Contents lists available at ScienceDirect

Gene

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

Research paper

Identification and analysis of house-keeping and tissue-specific genes based on RNA-seq data sets across 15 mouse tissues

Jingyao Zeng [a,b,1], Shoucheng Liu [a,b,1], Yuhui Zhao [a,b,1], Xinyu Tan [a], Hasan Awad Aljohi [c], Wanfei Liu [a,c,*], Songnian Hu [a,**]

[a] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, No. 1 Beichen West Road, Chaoyang District, Beijing 100101, China
[b] University of Chinese Academy of Sciences, No. 19A Yuquan Road, Beijing 100049, China
[c] Joint Center for Genomics Research (JCGR), King Abdulaziz City for Science and Technology and Chinese Academy of Sciences, Prince Turki Road, Riyadh 11442, Saudi Arabia

ARTICLE INFO

Article history:
Received 17 June 2015
Received in revised form 27 October 2015
Accepted 3 November 2015
Available online 6 November 2015

Keywords:
RNA sequencing
Mouse
Transcriptome profiling
Housekeeping
Tissue specificity
Real-time PCR

ABSTRACT

Recently, RNA-seq has become widely used technology for transcriptome profiling due to its single-base accuracy and high-throughput speciality. In this study, we applied a computational approach on an integrated RNA-seq dataset across 15 normal mouse tissues, and consequently assigned 8408 house-keeping (HK) genes and 2581 tissue-specific (TS) genes among UCSC RefGene annotation. Apart from some basic genomic features, we also performed expression, function and pathway analysis with clustering, DAVID and Ingenuity Pathway Analysis, indicating the physiological connections (tissues) and diverse biological roles of HK genes (fundamental processes) and TS genes (tissue-corresponding processes). Moreover, we used RT-PCR method to test 18 candidate HK genes and finally identified a novel list of highly stable internal control genes: *Ywhae*, *Ddb 1*, *Eif4h*, etc. In summary, this study provides a new HK gene and TS gene resource for further genetic and evolution research and helps us better understand morphogenesis and biological diversity in mouse.

1. Introduction

In the past decades, biology researchers have long been concentrating on the important issues about the morphogenesis of diverse organisms and what makes them morphologically different from each other. Many results indicated that the gene differential expression in distinct tissues should account for this difference. In the process of research, two important concepts arise spontaneously. The notion of housekeeping (HK) genes was firstly used in paper about 40 years ago (Levy and Dixon, 1977), it describes genes ubiquitously expressed in almost all tissues/cell-types regardless of its developmental stage, physiological condition and external stimuli, which are considered essential for the maintenance of fundamental cellular functions (Chang et al., 2011; Butte et al., 2001; Tu et al., 2006), and HK genes are generally required to be expressed in a relatively constant level across tissues in some previous studies (Eisenberg and Levanon, 2013). Therefore, several traditional HK genes with highly constant expression such as *Gapdh* and *Ubc* are frequently used as internal controls in experimental testing (Qi et al., 2010; Heckmann et al., 2006). While tissue-specific (TS) genes represent those genes which are expressed in a single specific tissue, unlike tissue-selective genes, TS genes present a more extreme expression manner than the latter, and many previous studies have clarified their potential roles in the tissue-corresponding functions. Thus, TS genes can be used as drug targets and disease markers (Dezso et al., 2008; Liang et al., 2006). Furthermore, related studies have also analyzed their genomic and evolutional features and found that HK genes generally are shorter in genomic structure, lower conserved in their promoter regions (Farre et al., 2007) and contained more repeated sequences in 5′-UTRs than TS genes (Lawson and Zhang, 2008). These studies provided us a basic understanding of the two special groups of genes.

To address the fundamental questions mentioned above, researchers endeavored to extensively study the composition and complexity of mammalian tissue transcriptomes and used methods such as microarray, serial analysis of gene expression (SAGE) and sequencing of

Abbreviations: RNA-seq, RNA sequencing; HK, house-keeping; TS, tissue-specific; RefSeq, NCBI Reference Sequence Database; DAVID, Database for Annotation Visualization and Integrated Discovery; IPA, Ingenuity Pathway Analysis; Real-time RT-PCR, real-time quantitative reverse transcriptional-polymerase chain reaction; SAGE, serial analysis of gene expression; EST, expressed sequence tag; ENCODE, Encyclopedia of DNA Elements; FPKM, fragments per kilobase of exon per million fragments mapped; FPR, false positive rate; FNR, false negative rate; CV, coefficient variation; JSD, Jensen–Shannon distance; MGI, Mouse Genome Informatics; LINE, long interspersed nuclear elements; LTR, long terminal repeat; SINE, short interspersed nuclear elements; CDS, coding sequence; TSS, transcription start site; UTR, untranslated regions; UCSC, University of California Santa Cruz; CT, cycle threshold.

* Correspondence to: W. Liu, CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, No. 1 Beichen West Road, Chaoyang District, Beijing 100101, China.
** Corresponding author.
E-mail addresses: liuwf@big.ac.cn (W. Liu), husn@big.ac.cn (S. Hu).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.gene.2015.11.003
0378-1119/

expressed sequence tags (ESTs) to detect HK genes and TS genes (Zhu et al., 2008a; Velculescu et al., 1999; Su et al., 2002). Results showed that these methods are good at identifying TS genes but the lists of HK genes defined by different methods generally have low consensus (Ramskold et al., 2009). Fortunately, the rapid expansion of RNA-seq technology has largely improved the accuracy and reliability of transcriptome profiling (Haas and Zody, 2010; Cloonan et al., 2008; Yassour et al., 2009; Cui et al., 2010). Compared to previous technologies, RNA-seq is more sensitive which enables us to capture a vaster amount of transcripts even for those genes with low expression level or are only expressed in a few tissues (Eisenberg and Levanon, 2013). Most importantly, RNA-seq has been widely used to measure and quantify the expression level of genes and alternative isoforms. All these advantages make RNA-seq the most powerful technology for the global study of transcriptomes and the detection for HK genes and TS genes.

The number of tissues/cell-types and how to choose tissues are both key factors which could directly impact on the amount and quality of HK genes and TS genes detection. Chang et al. (2011) once illustrated that whatever expression indicator they used, more than 10 tissues were required to reconstruct the representative transcripts, and a high quality transcriptome library with long reads could certainly increase the follow-up statistical power (Eisenberg and Levanon, 2013). Here, we focused on an integrated RNA-seq data sets including 14 mouse tissues from Encyclopedia Of DNA Elements (ENCODE) (Consortium, 2004) representing 7 mouse physiology systems, and mouse cerebrum generated from our own lab., using the fragments per kilobase of exon per million fragments mapped (FPKM) as our normalized value to estimate the gene expression in a tissue. Referring to the method used by Ramskold et al. and Chen et al. (Ramskold et al., 2009; Chen et al., 2013), we adopted a definition that coordinated false positive rates (FPR) and false negative rates (FNR) to investigate whether a candidate gene is expressed or not. As a result, we generated a total number of 8408 HK genes and 2581 TS genes across 15 tissues, respectively, and further classified the HK genes into two categories: constantly expressed HK genes (2%) and variably expressed HK genes (98%). In addition, we applied the functional enrichment analysis for HK genes and TS genes and found out that these two sets of genes possess totally different functional roles in diverse biological processes.

2. Results

2.1. Reads alignment and expression estimation of 23,374 candidate genes

In this study, we used the 23,374 mouse annotation genes from UCSC RefGene as our candidate gene set (Karolchik et al., 2014). To investigate their expression patterns in different tissues, we downloaded the polyadenylated RNA-seq data sets across 14 mouse tissues including ovary, mammary gland, stomach, small intestine, adrenal gland, large intestine, thymus, testis, kidney, liver, lung, spleen, colon and heart from the ENCODE project (http://www.ncbi.nlm.nih.gov/sra), and additionally sequenced mouse cerebrum using the ribo-minus RNA-seq (rmRNA-seq) method in our own lab, ENCODE data are 76 bp in read length and cerebrum data are 101 bp, and all data are paired-end and strand-specific (Table S1). Then, we applied these 15 RNA-seq data on a computational pipeline to identify mouse HK genes and TS genes (Fig. S1).

Before the analysis, a purification procedure was required to increase the quality of these numerous raw data. Therefore, we used an in-house Perl script to eliminate those low-quality reads and obtained about 96.6% of the original reads on average for ENCODE data. However, a lower fraction (50.87%) of high-quality reads in cerebrum passed initial quality thresholds due to a large percent (42.70%) of ribosomal RNA reads (Table S1). For the task of HK gene and TS gene detection, GSNAP was used to align high-quality reads for each tissue against mouse genome (Genome Reference Consortium Mouse Build 38, mm10) individually, resulting in a high mapping rate (92.68% on average) for ENCODE data but a relatively lower rate for cerebrum due to the larger percent of half-mapping rate (Table S2). However, the mappable data are saturated and adequate for further analysis in cerebrum (Fig. S2). As we know, the expression level of a transcript or a gene is generally estimated according to the number of supporting reads, and on the basis of mapping result, we adopted Cuffdiff to calculate the FPKM value for each gene and estimate the gene differential expression across tissues at the same time. Thus, we have generated the expression data for each gene for the follow-up HK gene and TS gene identification.

2.2. HK genes and TS genes identification

After we got the FPKM values for each gene in distinct tissues, the most thorny problem was how to judge a gene is expressed or not and it is known to be challenging to determine the expression cutoff due to experiment contamination and background noises. Referring to the commonly used method in many HK gene studies (Ramskold et al., 2009; Chen et al., 2013), we adopted the same definition to find the background threshold value. By comparing the expression levels between exons (positive) and intergenic regions (negative), we used two mathematical formulas (see Materials and methods) to calculate the FPR and FNR values and took their equilibrium value as our primary criterion (Fig. 1A as for instance for lung tissue). For individual tissues, we obtained their equilibrium values ranging from 0.5 to 1.1, and we also compared the number of defined HK genes and TS genes respectively under different thresholds (Table S3). Eventually, we took the median value of all threshold values across tissues (0.7) as our ultimate threshold (Fig. 1B).

According to this expression cutoff, genes with FPKM larger than 0.7 were marked as expressed genes in the corresponding tissue, and genes with FPKM less than 0.7 were considered to be non-expressed. We thus obtained a total number of 8408 HK genes (Supplemental-Data1) which ubiquitously expressed in all 15 tissues and 2581 genes (Supplemental-Data2) expressed in an extreme tissue-specific manner (Fig. 1C). To assess the accuracy of our defined TS genes, we used the Jensen–Shannon (JS) distance model performed by CummeRbund to calculate the tissue-specificity score for each gene, and found that more than 75% of defined TS genes obtained a JS score larger than 0.75 (the larger the JS score, the more tissue-specific the gene is) (Fig. S3), which effectively proved the reliability of our detection. In the light of the results, the numbers of expressed genes in different tissues ranged from 11,569 to 14,661, among them, those genes expressed in testis accounted for the largest fraction (62.72%) of the total genes. In addition, testis and cerebrum remarkably presented a more specificity because of the larger fractions (10.07% for testis and 3.95% for cerebrum) of TS genes assigned to these two tissues, and the percentage of TS genes for other tissues were all less than 1% (Table 1; Fig. 1D).

In comparison with the human annotation genes from the MGI database, there are 17,171 mouse genes homologous to human, and as many as 95.21% (8005) of our defined HK genes are homologous genes in human. Actually, many previous HK genes studies have reported a significant amount of human HK genes, for example, Daniel et al. defined 7897 human HK genes from 16 tissues in 2009 (Ramskold et al., 2009), Chang et al. identified 2064 human HK genes across 46 tissues (Chang et al., 2011) and Eisenberg et al. also detected 3804 human HK genes from 16 tissues (Eisenberg and Levanon, 2013) etc. By comparing our defined 8005 homologous HK genes to the human HK genes annotated by these studies, we have surprisingly found that 6360 HK genes defined in our study have been annotated as HK genes as well in human, which indicated the high consensus between mouse and human. However, there are 1645 HK genes only defined in our study which has not ever reported by others, it may have resulted from the species difference and tissues difference (only large intestine, small intestine stomach and spleen are used in our study). Similarly, the particular 1684 HK genes only reported by Daniel et al. can be also attributed to their 4 unique tumor tissues (Fig. 2). All together, the list of our defined HK