Membrane Gene Ontology Bias
in Sequencing and Microarray Obtained by Housekeeping-Gene Analysis

Yijuan Zhang[1], Oluwafemi S. Akintola[1], Ken J. A. Liu[2], and Bingyun Sun[1,2,]*

1. Department of Chemistry; 2. Department of Molecular Biology and Biochemistry,

Simon Fraser University, Burnaby, British Columbia, Canada

* Correspondence: Bingyun Sun, Simon Fraser University, Burnaby, British Columbia,

Canada, Email: bingyun_sun@sfu.ca

Key words:

transcriptome, microarray, sequencing, RNA-seq, next-generation sequencing,

housekeeping genes, probe coverage, gene expression, gene ontology, gene structure

Highlights:

- Housekeeping genes can be used for reliable analysis of differences between microarray and sequencing technology.
- Microarray tends to identify gene ontology related to membrane, cell surface, and secreted proteins that sequencing technology often misses.
- Sequencing tends to identify gene ontology related to nuclear transcripts that microarray technology misses.
- Both the probe coverage and detection sensitivity has contributed to the missing gene identification by microarray
- Sequencing technology has greatly improved current microarray probe coverage on the nuclear transcripts

Abstract

Microarray (MA) and high-throughput sequencing are two commonly used detection systems for global gene expression profiling. Although these two systems are frequently used in parallel, the differences in their final results had not to be examined thoroughly. Transcriptomic analysis of housekeeping (HK) genes provides a unique opportunity to reliably examine the technical difference between these two systems. We investigated here the structure, genome location, expression quantity, microarray probe coverage, as well as biological functions of differentially identified human HK genes by 9 MA and 6 sequencing studies. These in-depth analyses allowed us to discover, for the first time, a subset of transcripts encoding membrane and cell surface proteins, as well as nuclear proteins regulating transcription prone to differential identification by the two platforms. We hope the discovery can aid the future development of these technologies for comprehensive transcriptomic studies.

INTRODUCTION

Sequencing and microarray (MA) are two major high-throughput technologies in gene expression profiling [1, 2]. The former is based on reading at individual nucleotide resolution of nucleic acids or their fragments; the latter is based on the hybridization principle through the use of nucleic acid template known to bind with the sample targets. The differences between these two technologies have been discussed frequently in recent publications[1-8]. Even though sequencing, especially the fast evolving next generation sequencing (NGS), is well regarded for its higher sensitivity than MA, the complementarity between these two systems has been well recognized. It is, however, not clear why sequencing is to miss some MA-identified genes (abbreviated as MA genes below).

The reasons for lacking clear understanding, in our opinion, are following. Firstly, in the existing comparison studies, sample types used were constrained to one or a couple and each with limited measurements [4, 9]. Stochastic errors resolved from the small number of analyses can be prominent, which renders missing identification by each technique easily justifiable, and any further discussion on the causality less reliable. Secondly, because different tissue types at various biological states have drastically different gene expression profiles, results derived from one study cannot be easily merged with the others for more robust analysis.

The studies of housekeeping (HK) genes are, however, different. These studies address gene expression at the entire organismal level. Not only that any gene in HK studies was derived from repeated analysis of a broad collection of different organs, tissues, and cell types [10, 11], such that these identifications (often derived from more than a hundred runs) are very reliable; but also that if the studying species is the same, these genes are comparable[12]. Therefore HK genes are ideal candidates for examining the differences in MA and sequencing analyses. To date, numerous human HK-gene studies have been carried out. Initially, these studies were mostly conducted by MA [13-22] because of its low cost. Recently, the significant drop of spend in sequencing also enabled its analysis of HK genes [23]. The results derived from these large-scale analyses provided us a unique opportunity to reliably and comprehensively reveal the identification difference between the two technical platforms.

To seek differences, we conducted here a series of in-depth analyses with consideration of the structural, localizational, and functional aspects of the detected genes besides commonly compared gene number and expression quantity. From our investigation of 15 human HK-gene datasets including 9 MA and 6 sequencing based studies, we discovered some interesting biases and identified gene ontology of membrane, cell surface, extracellular space, and nuclear related is prone to differential identification by the two techniques. The obtained information will help guide future selection and design of these techniques for comprehensive transcriptomic analyses.

METHODS

Data collection We obtained lists of HK genes from 15 published studies. MA was used in 9 studies, including Warrington[13], Hsiao [15], Eisenberg_03[14], Tu [16], Dezso [18], She [19], Chang [20], Shyamsundar [21], Zhu_MA; sequencing was employed in the rest 6 studies as summarized in Table 1. In particular, Zhu_EST [22] and Podder [24] used expressed sequence tag technique (EST), Reverter [25] used massively parallel signature sequencing (MPSS), while Ramskold [26], Eisenberg_13[27] and Fagerberg [23] used RNA-sequencing (RNA-seq).

To ease of comparison, we converted all the gene identifiers to Entrez gene ID using the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 (http://david.abcc.ncifcrf.gov/)[28, 29]. The resolved genes were divided into three categories based on their detection technique, i.e. MA unique, sequencing unique, and common to both techniques.

Probe coverage (PC) As the probes on MAs will determine whether a gene can be detected or not, we studied the probe coverage of genes exclusively identified by sequencing. We obtained the probe information of 12 chips used in the 9 MA studies (some studies used more than one type of chips) from three sources, i.e. Gene Expression Omnibus (GEO) Database [30] (http://www.ncbi.nlm.nih.gov/gds), NetAffx Analysis Center [31] (https://www.affymetrix.com/analysis/index.affx), and Applied Biosystems Human Genome Survey Microarray (https://www.lifetechnologies.com/).

To quantify the frequency that a gene is covered by MAs, we defined Probe Coverage (PC). If a gene is not included by any MA studies, the PC value will be zero; if a gene is included in all 9 MA studies, its PC value will be 9. If multiple gene symbols were mapped and each with different PC value, the highest PC value was considered.

Chromosomal location To localize the identified HK genes on chromosomes as well as mitochondria genome, we queried the database from National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov). Because gene annotation is not distributed evenly across chromosomes, we normalized our results based on annotated genes in each chromosome offered by the Ensembl genome (release 68, July 2012) (http://jul2012.archive.ensembl.org/index.html).

Gene structure To analyze the potential structural difference among the three groups of genes, we extracted the exon count, total exon and intron lengthes, as well as the coding sequences using RefGene information from University of California, Santa Cruz (UCSC) genome browser (http://genome.ucsc.edu/index.html). In detail, the gene lists were translated into Refseq gene ID in DAVID, and then queried against human genome assembly, GRCh38 [32] in UCSC genome browser. When multiple Refseq IDs were mapped, all IDs are considered. The total intron length was obtained by subtracting the total exon length that was the sum of all exons in a transcript. Only the coding sequence was considered in the analysis of GC content.

Gene abundance We analyzed the expression level of HK genes with quantitative information, which includes datasets of Chang [20], Eisenberg_03 [14], She [19], Warrington [13], Shyamsundar [21] and Fagerberg [23]. To compare, we normalized the expression quantity by the highest value in each list. If a gene had several expression values, the highest one was used.

Detection breadth (DB) To quantitatively characterize across studies how effectively a HK gene was detected by the two technologies, we defined Detection Breadth (DB)[12], i.e. the number of studies in which a particular gene was resolved. If a gene was only resolved in one study, its DB value would be 1; if a gene was resolved in all MA studies, its DB value would be 9; similarly genes identified in all sequencing analyses would have DB value of 6.

Functional analysis We conducted functional enrichment analysis on Gene Ontology (GO) Biological Process (BP_FAT) (i.e. the summarized version of BP in GO) using DAVID v6.7 [33] (http://david.abcc.ncifcrf.gov/) separately for four lists, i.e. MA genes, sequencing genes with PC=0 and PC > 0 values, as well as the shared genes. Top 10 enriched functions in each list after filtering through default Fisher Exact p value (≥0.1) offered by DAVID [34] were obtained.

Results

Data collection

We compiled 12,501 HK genes from 15 studies using either MA or sequencing based transcriptomic techniques[12] in which we removed 13 genes annotated to Homo sapiens neanderthalensis. Table 1 summarizes the 15 studies, in which 9 studies employed MA and 6 studies used sequencing. Sixty-five percent of these genes were discovered by MA, whereas almost all of them (90%) were identified by sequencing. Among them, sequencing technique identified 4,395 (35%) HK genes exclusively (i.e. sequencing genes); whereas MA alone identified only 1, 304 (10%) (i.e. MA genes); commonly identified genes were 6802. A Venn diagram of this comparison is shown in Fig. 1. The lists of MA and sequencing specific genes as well as the commonly identified genes using two techniques were listed in Supplementary Table S1.

Microarray probe coverage

The probe coverage of the sequencing-specific HK genes was examined, and the results are summarized in Fig. 2. The chips used in 9 MA studies are summarized in Supplementary Table S2. Among 4,394 sequencing unique genes, 1,389 of them (32%) did not have any probe coverage in all the 9 MA studies, the rest 3,005 genes spread between PC values of 1 to 7. These results implied that the coverage of arrays used for HK-gene studies was limited. The missing identification of 3,005 sequencing genes with PC>0 suggested that besides probe coverage other factors had also contributed. In analysis below, we separately considered sequencing genes with PC=0 and PC>0 coverage to seek potential explanation.

Chromosomal location

The genome location of genes identified by MA and sequencing alone, as well as by commonly identified genes, is summarized in Supplementary Table S3, in which we entailed the gene count and the percentage of HK gene distribution along 24 chromosomes (X and Y chromosomes were considered separately) and the mitochondria genome. In the table, the percentage was computed by HK genes obtained in each chromosome against both the total of number of HK genes detected and the total number of annotated genes in each chromosome on all three types of detections, i.e. MA alone, sequencing alone, and commonly identified. A detailed list of chromosomal location of each addressed gene can be found in Data In Brief, Table 1 [35]. For ease of visualization, we plotted distribution of MA, sequencing, and commonly identified genes against the total annotated chromosomal genes as shown in Fig. 3. The results resembled the overall comparison in Fig. 1, in which sequencing identified much more genes than MA, and the trend is similar across all the chromosomes.

An extreme case was appeared in mitochondria genome, in which there were no MA unique genes, and sequencing genes had particularly high (~85%) percentage of identification. In Ensembl database, mitochondria has only 13 confirmed genes, whereas sequencing analysis identified 11 of them, the other two genes were identified by both methods. A result suggesting mitochondria genome was quite conserved in all organs and tissues. Interestingly, there were also 11 Homo sapiens neanderthalensis mitochondria genes in the original dataset compiled only from sequencing analysis. As all the analyzed samples in the studied datasets were obtained from human samples, it was interesting to observe that this genome was also conservative across closely evolved species. The high efficiency of sequencing to identify mitochondria genes and the complete absence of MA contribution demonstrated the capacity of sequencing in sensitivity and de novo discovery of closely related non-modern human genes that were lacking in MA analysis.

Gene structure characteristics

It was reported that sequencing tends to identify genes with longer exons [36, 37] We here analyzed the structural features in differentially identified genes by the two techniques. For all gene lists we complied above, Fig 4A shows their average exon number per gene; Fig. 4B shows their total exon length per gene; Fig. 4C shows the total intron length per gene. Slight difference was observed in all three comparisons, and the trend was similar to previous reports [36, 37].

It was also known that GC content can affect both sequencing and hybridization efficiencies [38-41]. Miss-sequencing has been reported to genomes with extreme GC contents [42, 43], and increased GC content (such as those exceeding 55%) has been associated with increased cross hybridization in microarray[39]. We therefore analyzed the GC content of gene detected by MA alone, sequencing alone, and jointly. Fig. 4D shows the GC percentage of each compared category.

The average of exon count, total exon and intron length as well as the GC percentage of each analyzed category, i.e. MA alone, sequencing (PC=0), sequencing (PC>0), and commonly identified, is provided in Supplementary Table S4. The detailed value for each transcript concerned here is included in Data In Brief, Table 2 [35].

Gene abundance and detection breadth (DB)

To analyze whether the observed identification difference was due to the low expression, we studied the distribution of expression quantity among the exclusively identified genes as shown in Fig. 5A. In the figure, MA genes were evenly distributed across all normalized abundance levels; whereas almost all sequencing genes (98%) occupied the lowest expression level.

We also examined the DB distribution of MA and sequencing genes in Fig. 5B. Opposing to the profile in Fig. 5A, Fig. 5B showed that MA genes significantly skewed toward lower values, whereas sequencing genes spread much more evenly.

The detailed value of expression and DB status for each gene detected by MA alone, sequencing alone, as well as jointly was summarized in Data In Brief, Table 3 [35]. The detailed statistical value shown is Fig. 5 was provided as Supplementary Table S5.

Functional analysis

For effective comparison, we also obtained the top-10 most enriched Gene Ontology (GO) biological process (BP) terms from four gene lists, i.e. MA genes, sequencing genes with PC=0, sequencing genes with PC > 0 values, and the shared genes. The visualization was achieved by plotting the results in heatmap in Fig. 6 using MultipleExperiment Viewer (MeV, a software package for displays of high throughput analysis) [44]. Clear distinction was observed between MA genes and the rest, in which the MA unique genes were uniquely enriched in cell surface and secretion related biological processes including “cell-cell signaling”, “defense response”, “regulation of response to external stimulus”, and “regulation of hormone levels”. The details of the

MA genes contributed to these biological processes were summarized in Supplementary Table S6. The common and sequencing genes have broadly covered other processes such as those occurred in nucleus and cytosol, including “RNA processing”, “translation”, and “mRNA metabolic process”. In addition, sequencing genes also uniquely enriched GO terms of “transcription”, “regulation of transcription”, “chromatin modification”, “tRNA metabolic process”, “DNA repair” in which “transcription related terms” were solely from PC=0 group while the rest from PC>0 group.

Discussion

The advantages of sequencing and MA technologies as well as their complementarity have been discussed recently[1-8]. It is well known that sequencing based analysis comparing to MA is direct and probe-free; therefore it is more accurate and sensitive with unlimited dynamic range [7, 27, 45-47]. The challenges of sequencing based transcriptome analysis are however centered on the sequencing depth, genomic DNA contamination, and alignment errors raised by short reads among others [48, 49].

Even though sequencing is known for its high sensitivity and unlimited dynamic range, yet the counting based nature rendered the measurement sensitive to sample complexity and sequencing depth. To achieve high coverage to low-expressing transcripts, larger quantity of samples and more reads are necessary[50, 51]. It was reported that at extremely high mapped-reads of 50 million in the fly modENCODE samples, the detection of new transcripts was still not saturated [52]. Based on Malone and Oliver's report [7] at 6-8 million mapped reads, the coverage of fly transcriptome was roughly 80-90%; whereas Blencowe et al. had estimated that in mammalians, 700 million reads would be necessary for quantitatively access 95% of transcript expression [53]. However, based on studies of Tarazona et al.[54], excessive increase of sequencing depth will drastically increase false discovery rate, in some cases to 60%, with a significant increase on non-protein coding RNA detection and a minor increase on protein coding RNA. Therefore a balance between the depth of coverage and the false discovery rate was recommended. As a result, a compromised coverage of transcriptome by sequencing is anticipated.

Because of these reasons, the complementarity in data obtained by parallel analysis using NGS and MA is well recognized. The missing identification of sequencing genes by MA can be easily explained. Yet the missing identification of certain MA genes in sequencing results has been, however, less discussed. Given the fact that NGS cannot practically obtain complete transcriptome, and considering that MA functions on a different technological principle than sequencing, it is reasonable to expect that the miss-identified genes by the two methods are different. As the missed genes can possess important biological functions [55-57], the awareness of types of transcripts that are prone to miss identification therefore is critical, and will help design strategies to compensate such loss.

To obtain reliable information on genes that are sensitive to measurement biases, we chose to analyze HK genes. The reason is that in HK-gene analyses, each gene is derived from hundreds of measurements; therefore, random error from small number of experiments can be effectively eliminated. To our knowledge, HK-gene studies are the largest systematic analysis of transcriptomes of the single species using the same technical platform; thereof their results can be the most reliable measurement and are comparable when the examined species is the same.

After investigating a total of 15 human HK-gene datasets from 9 MA studies and 6 sequencing studies, we obtained 4,395 sequencing-specific HK genes, 1,304 MA-specific genes, and 6,802 common genes. The large number of common genes suggested that the two methods worked equally well, which was consistent with previous studies. The much smaller number of MA genes compared to those of sequencing also agreed with previous studies. The congruence suggested that results from HK genes can well represent results from others. Previously, the miss-identification of sequencing genes by MA had been largely believed due to the sensitivity of analysis[4, 58], and the missing-identification of MA genes by sequencing had no clear explanation.

We here focused on the in-depth investigation of the differentially identified genes with an aim to reveal the characteristics in the missing genes by each technique. Firstly, our quantitative analysis in Fig. 5A agreed with previous knowledge that sequencing was much more sensitive and identified genes with much less expression levels [12]. Similarly in Fig. 5B, the high sensitivity of sequencing allowed more common and reliable identification of transcripts among different studies supported by broad DB distribution, which is also consistent with existing knowledge [4, 6, 7, 40]. Conversely, MA genes in Fig. 5B had much smaller DB value.

Secondly, our probe coverage (PC) and genome location analyses in Fig. 2 and 3 respectively further suggested that for MA to miss-identify sequencing HK genes was also due to the limited probe coverage besides its sensitivity. The lack of probe coverage appeared evenly distributed across chromosomes. The particular high percentage of sequencing genes in mitochondria genome as shown in Fig. 3 was due to its much smaller genome.

To note, the arrays used for HK-gene studies were relatively old, and the probe design was based on past knowledge of human gene annotation. Current sequencing technology has largely advanced our knowledge of human genome and transcriptome, in such the UCSC human genome and NCBI Refseq human transcripts databases have been frequently updated. The more complete genome annotation and the improved array fabrication technology has made modern array probe collection significantly improved, and is able to provide the coverage of the entire Refseq database plus other sources such as long non-coding RNA database (www.Incrnadb.org). To examine current array coverage efficiency, we compared the sequencing detectable genes against the SurePrint G3 Human Gene Expression v3 8x60K microarray probe set from Agilent Technologies (Santa Clara, CA, USA). Only a marginal 152 sequence genes were not on this chip; therefore for modern array, probe coverage is no longer a major factor for missing identification. Yet the technology still suffers from the probe limitation. For example, this Agilent array had missed all 11 Homo sapiens neanderthalensis transcripts identified by sequencing. For de novo and discovery based transcriptome analysis, sequencing will be a favorable choice.

Thirdly, recent reports indicated a length bias in sequencing on gene identification and quantification[36, 59]. Our structure analysis in Fig. 4 (A-C) showed a slightly higher count and longer length of exons in sequencing genes compared with MA and common genes, which supported the existence of length bias, yet the influence seemed marginal than a major a factor contributing to differential HK-gene identification we observed here. In addition, GC content has also been discussed for their interference to sequencing identification efficiency and microarray cross hybridization and signal intensity [38-41]. Our results in Fig. 4D showed close ratios in MA alone, sequencing alone and commonly identified genes with similar standard deviation in all categories, suggesting GC content was also not the major contributor to the observed differential detection.

Lastly, we sought out how different detection systems can impact the functional understanding of the examined biological samples by analyzing enriched GO_BP terms among studied genes in Fig. 6. Interestingly, we observed a unique pattern of MA genes that were distinct from the rest of genes. A preference on membrane and surface related protein-coding transcripts was clearly observed in MA analysis specifically. Interestingly, sequencing genes showed unique enrichment on nuclear related GO terms besides their high similarity in enrichment pattern to commonly identified genes. Surprisingly, the enrichment of "transcription" related GO terms was solely contributed by PC=0 sequencing genes. A result explained a lack of annotation of these genes that causing early probe design to fail including them. In the meantime, this result also indicated the strength of sequencing technology to identify these important transcripts.

Our observation on the identification strength of MA on membrane GO terms coincided with the computational analysis carried out by Young et al.[37]. Young et al. noticed in their study that gene ontology in RNA-seq had biased towards gene length as well as read counts, in which read counts contributed more than length. In their statistical algorithm after adjusting these biases, GO terms of membrane and extracellular region were ranked highest in their analysis.

In addition, in previous proteomic studies conducted on cell surface and membrane proteins by us as well as by others when proteomics and sequencing based transcriptomics results were compared[60, 61], it was noticed that membrane and extracellular proteins were always under-represented in sequencing based transcriptomics. It is also well known that membrane proteins are lowly abundant, for proteomic analysis in which no amplification of protein quantity is possible, membrane proteins are extremely difficult to be detected and always require sophisticated enrichment and/or sensitive instruments for analysis [62]. Therefore, our current study as well as previous results supported the discovery of Young et al. that membrane and extracellular related GO terms are hampered in sequencing by abundant transcripts, and statistical corrections are needed for these GO terms to be identified.

It is quite valuable for MA to detect these rare transcripts. The reason for MA to be less sensitive to the abundant transcripts is likely due to the methodology design, in which each gene is represented by a set of probes with certain spotting density, and the hybridization process is relatively independent from one another. This design penalizes abundant transcripts when they saturate the probes, yet offers equal opportunity for rare transcripts to compete with abundant transcripts for binding based on the hybridization principle. The design also renders the technology insensitive to the total read counts and length biases suffered by sequencing. Nevertheless, it is inevitable that abundant transcripts can cause non-specific binding to array probes and, therefore, introduce high background. Currently there are no ideal methods for transcriptome analysis yet; however, the better understanding of their limitations will encourage future improvement for accurate interrogation of biological systems.

Conclusion

The technical differences of MA and sequencing have been frequently discussed in the past[9, 63, 64]. Most of these comparisons were conducted on a small scale with single sample type and limited number of measurements. Through our analysis on human HK-gene studies that were derived from hundreds of measurements covered the entire major organ and tissue types, we discovered that MA can more effectively identify surface and extracellular gene ontology that tends to be missed by sequencing analysis, while sequencing tends to identify more efficiently nuclear transcripts regulating transcription, DNA repair, and chromatin modifications. Our results coincided with the ontology bias predicted statistically by considering the length and count biases in sequencing technique[37]. From our current study, the two platforms show complementary biases. Since these two technologies also function on different principles, they are good orthogonal methods to each other, and can effectively validate in high throughput many potential measurement errors. In our opinion, both technologies should exist and need to be constantly improved. We hope the obtained information can help advance current transcriptome analysis for more accurate and comprehensive studies.

Acknowledgement:

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.

REFERENCES

1. Hurd, P. J., Nelson, C. J.: Advantages of next-generation sequencing versus the microarray in epigenetic research. *Briefings in Functional Genomics* 2009:elp013.

2. Mantione, K. J., Kream, R. M., Kuzelova, H., Ptacek, R., Raboch, J., Samuel, J. M., Stefano, G. B.: Comparing bioinformatic gene expression profiling methods: Microarray and RNA-Seq. *Medical science monitor basic research* 2014, 20:138.

3. Elingarami, S., Li, X., He, N.: Applications of nanotechnology, next generation sequencing and microarrays in biomedical research. *Journal of nanoscience and nanotechnology* 2013, 13(7):4539-4551.

4. 't Hoen, P. A. C., Ariyurek, Y., Thygesen, H. H., Vreugdenhil, E., Vossen, R. H., de Menezes, R. X., Boer, J. M., van Ommen, G.-J. B., den Dunnen, J. T.: Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic acids research* 2008, 36(21):e141-e141.

5. Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R.: Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* 2009, 10(1):161.

6. Nookaew, I., Papini, M., Pornputtpong, N., Scalcinati, G., Fagerberg, L., Uhlén, M., Nielsen, J.: A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic acids research* 2012:gks804.

7. Malone, J. H., Oliver, B.: Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology* 2011, 9(1):34.

8. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., Liu, X.: Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* 2014, 9(1):e78644.

9. Git, A., Dvinge, H., Salmon-Divon, M., Osborne, M., Kutter, C., Hadfield, J., Bertone, P., Caldas, C.: Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *Rna* 2010, 16(5):991-1006.

10. Butte, A. J., Dzav, V. J., Glueck, S. B.: Further defining housekeeping, or “maintenance,” genes Focus on “A compendium of gene expression in normal human tissues”. *Physiol Genomics*

2001, 7:95-96.

11. Jonge., H. J. M. d., Fehrmann, R. S. N., Bont, E. S. J. M. d., Hofstra, R. M. W., Gerbens, F., Kamps, W. A., Vries, E. G. E. d., Zee, A. G. J. v. d., Meerman, G. J. t., Elst, A. t.: Evidence Based Selection of Housekeeping Genes. *PLOS ONE* 2007, 2(9):e898.

12. Zhang, Y., Li, D., Sun, B.: Do housekeeping genes exist? *PLOS ONE* 2015, 10(5):e0123691.

13. Warrington, J., Nair, A., Mahadevappa, M., Tsyganskaya, M.: Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics* 2000, 2:143 - 147.

14. Eisenberg, E., Levanon, E.: Human housekeeping genes are compact. *Trends Genet* 2003, 19:362 - 365.

15. Hsiao, L.-L., Dangond, F., Yoshida, T., Hong, R., Jensen, R. V., Misra, J., Dillon, W., Lee, K. F., Clark, K. E., Haverty, P.: A compendium of gene expression in normal human tissues. *Physiological genomics* 2001, 7(2):97-104.

16. Tu, Z., Wang, L., Xu, M., Zhou, X., Chen, T., Sun, F.: Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics* 2006, 7(1):31.

17. Jiang Zhu, F. H., Songnian Hu1 and Jun Yu: On The Nature Of Human Housekeeping Genes. *Trends in Genetics* 2008, Vol.24 No.10.

18. Dezső, Z., Nikolsky, Y., Sviridov, E., Shi, W., Serebriyskaya, T., Dosymbekov, D., Bugrim, A., Rakhmatulin, E., Brennan, R. J., Guryanov, A.: A comprehensive functional analysis of tissue specificity of human gene expression. *BMC biology* 2008, 6(1):49.

19. She, X., Rohl, C. A., Castle, J. C., Kulkarni, A. V., Johnson, J. M., Chen, R.: Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC Genomics* 2009, 10(1):269.

20. Chang, C.-W., Cheng, W.-C., Chen, C.-R., Shu, W.-Y., Tsai, M.-L., Huang, C.-L., Hsu, I. C.: Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PloS one* 2011, 6(7):e22859.

21. Shyamsundar, R., Kim, Y. H., Higgins, J. P., Montgomery, K., Jorden, M., Sethuraman, A., van de

Rijn, M., Botstein, D., Brown, P. O., Pollack, J. R.: A DNA microarray survey of gene expression in normal human tissues. *Genome biology* 2005, 6(3):R22.

22. Zhu, J., He, F., Song, S., Wang, J., Yu, J.: How many human genes can be defined as housekeeping with current expression data? *BMC Genomics* 2008, 9:172.

23. Fagerberg, L., Hallström, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K.: Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular Proteomics* 2014, 13(2):397-406.

24. Podder, S., Ghosh, T. C.: Exploring the differences in evolutionary rates between monogenic and polygenic disease genes in human. *Molecular biology and evolution* 2010, 27(4):934-941.

25. Reverter, A., Ingham, A., Dalrymple, B. P.: Mining tissue specificity, gene connectivity and disease association to reveal a set of genes that modify the action of disease causing genes. *BioData Min* 2008, 1(1):8.

26. Ramsköld, D., Wang, E. T., Burge, C. B., Sandberg, R.: An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS computational biology* 2009, 5(12):e1000598.

27. Eisenberg, E., Levanon, E. Y.: Human housekeeping genes, revisited. *Trends in Genetics* 2013, 29(10):569-574.

28. Huang, D. W., Sherman, B. T., Lempicki, R. A.: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* 2008, 4(1):44-57.

29. Huang, D. W., Sherman, B. T., Lempicki, R. A.: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* 2009, 37(1):1-13.

30. Edgar, R., Domrachev, M., Lash, A. E.: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* 2002, 30(1):207-210.

31. Liu, G., Loraine, A. E., Shigeta, R., Cline, M., Cheng, J., Valmeekam, V., Sun, S., Kulp, D., Siani-Rose, M. A.: NetAffx: Affymetrix probesets and annotations. *Nucleic acids research* 2003, 31(1):82-86.

32. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., Haussler, D.: The human genome browser at UCSC. *Genome research* 2002, 12(6):996-1006.

33. Dennis Jr, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., Lempicki, R. A.: DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 2003, 4(5):P3.

34. Tonevitsky, A. G., Maltseva, D. V., Abbasi, A., Samatov, T. R., Sakharov, D. A., Shkurnikov, M. U., Lebedev, A. E., Galatenko, V. V., Grigoriev, A. I., Northoff, H.: Dynamically regulated miRNA-mRNA networks revealed by exercise. *BMC physiology* 2013, 13(1):9.

35. Zhang, Y., Akintola, O. S., Liu, K. J. A., Sun, B.: Detection Bias in Microarray and Sequencing Transcriptomic Analysis Identified by Housekeeping Genes *Data In Brief* 2015, submitted.

36. Oshlack, A., Wakefield, M. J.: Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* 2009, 4(1):14.

37. Young, M. D., Wakefield, M. J., Smyth, G. K., Oshlack, A.: Method Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 2010, 11:R14.

38. Benjamini, Y., Speed, T. P.: Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic acids research* 2012, 40(10):e72.

39. Kucho, K.-i., Yoneda, H., Harada, M., Ishiura, M.: Determinants of sensitivity and specificity in spotted DNA microarrays with unmodified oligonucleotides. *Genes Genet. Syst.* 2004, 79:189-197.

40. Swindell, W. R., Xing, X., Voorhees, J. J., Elder, J. T., Johnston, A., Gudjonsson, J. E.: Integrative RNA-seq and microarray data analysis reveals GC content and gene length biases in the psoriasis transcriptome. *Physiol. Genomics* 2014, 46:533-546.

41. Xia, X.: The effect of probe length and GC% on microarray signal intensity: characterizing the functional relatioship. *International J. Systems Synthetic Biology* 2010, 1:171-183.

42. Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S. *et al*: Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* 2002, 419(6906):498-511.

43. Perelygina, L., Zhu, L., Zurkuhlen, H., Mills, R., Borodovsky, M., Hilliard, J. K.: Complete Sequence and Comparative Analysis of the Genome of Herpes B Virus (Cercopithecine Herpesvirus 1) from a Rhesus Monkey. *Journal of Virology* 2003, 77(11):6167-6177.

44. Howe, E., Holton, K., Nair, S., Schlauch, D., Sinha, R., Quackenbush, J.: Mev: multiexperiment viewer. In: *Biomedical Informatics for Cancer Research.* Springer; 2010: 267-277.

45. Ansorge, W. J.: Next-generation DNA sequencing techniques. *New biotechnology* 2009, 25(4):195-203.

46. Trapnell, C., Pachter, L., Salzberg, S. L.: TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009, 25(9):1105-1111.

47. Maher, C. A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N., Chinnaiyan, A. M.: Transcriptome sequencing to detect gene fusions in cancer. *Nature* 2009, 458(7234):97-101.

48. Jun, G., Flickinger, M., Hetrick, K. N., Romm, J. M., Doheny, K. F., Abecasis, G. R., Boehnke, M., Kang, H. M.: Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *The American Journal of Human Genetics* 2012, 91(5):839-848.

49. Jeck, W. R., Reinhardt, J. A., Baltrus, D. A., Hickenbotham, M. T., Magrini, V., Mardis, E. R., Dangl, J. L., Jones, C. D.: Extending assembly of short DNA sequences to handle error. *Bioinformatics* 2007, 23(21):2942-2944.

50. Wilhelm, B. T., Landry, J.-R.: RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* 2009, 48(3):249-257.

51. Gawronski, J. D., Wong, S. M., Giannoukos, G., Ward, D. V., Akerley, B. J.: Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. *Proceedings of the National Academy of Sciences* 2009, 106(38):16422-16427.

52. Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W.: The developmental transcriptome of Drosophila melanogaster. *Nature* 2011, 471(7339):473-479.

53. Blencowe, B. J., Ahmad, S., Lee, L. J.: Current-generation high-throughput sequencing:

deepening insights into mammalian transcriptomes. *Genes & development* 2009, 23(12):1379-1386.

54. Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., Conesa, A.: Differential expression in RNA-seq: a matter of depth. *Genome research* 2011, 21(12):2213-2223.

55. Kwon, M. J., Oh, E., Lee, S., Roh, M. R., Kim, S. E., Lee, Y., Choi, Y.-L., In, Y.-H., Park, T., Koh, S. S.: Identification of novel reference genes using multiplatform expression data and their validation for quantitative gene expression analysis. *PLoS One* 2009, 4(7):e6162.

56. Bär, M., Bär, D., Lehmann, B.: Selection and validation of candidate housekeeping genes for studies of human keratinocytes—review and recommendations. *Journal of Investigative Dermatology* 2009, 129(3):535-537.

57. Urrutia, A. O., Hurst, L. D.: The signature of selection mediated by expression on human genes. *Genome research* 2003, 13(10):2260-2264.

58. Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M.: The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008, 320(5881):1344-1349.

59. Ozsolak, F., Milos, P. M.: RNA sequencing: advances, challenges and opportunities. *Nature reviews genetics* 2011, 12(2):87-98.

60. Sun, B., Ma, L., Yan, X., Lee, D., Alexander, V., Hohmann, L. J., Lorang, C., Chandrasena, L., Tian, Q., Hood, L.: N-glycoproteome of E14. Tg2a mouse embryonic stem cells. *PloS one* 2013, 8(2):e55722.

61. Rugg-Gunn, P. J., Cox, B. J., Lanner, F., Sharma, P., Ignatchenko, V., McDonald, A. C., Garner, J., Gramolini, A. O., Rossant, J., Kislinger, T.: Cell-surface proteomics identifies lineage-specific markers of embryo-derived stem cells. *Developmental cell* 2012, 22(4):887-901.

62. Sun, B., Utleg, A. G., Hu, Z., Qin, S., Keller, A., Lorang, C., Gray, L., Brightman, A., Lee, D., Alexander, V. M.: Glycocapture-assisted global quantitative proteomics (gagQP) reveals multiorgan responses in serum toxicoproteome. *Journal of proteome research* 2013, 12(5):2034-2044.

63. Bloom, J. S., Khan, Z., Kruglyak, L., Singh, M., Caudy, A. A.: Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression

microarrays. *BMC Genomics* 2009, 10(1):221.

64. Reinartz, J., Bruyns, E., Lin, J.-Z., Burcham, T., Brenner, S., Bowen, B., Kramer, M., Woychik, R.: Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Briefings in functional genomics & proteomics* 2002, 1(1):95-104.

Table 1. Summary of included housekeeping (HK) gene studies.

HK-gene Study	Technique	Reference
Warrington	MA	[13]
Hsiao	MA	[15]
Eisenberg_03	MA	[14]
Tu	MA	[16]
Dezsö	MA	[18]
She	MA	[19]
Chang	MA	[20]
Shyamsundar	MA	[21]
Zhu_MA	MA	[22]
Podder	Sequencing_EST	[24]
Zhu_EST	Sequencing_EST	[22]
Reverter	Sequencing_MPSS	[25]
Ramsköld	Sequencing_RNA-seq	[26]
Eisenberg_13	Sequencing_RNA-seq	[27]
Fagerberg	Sequencing_RNA-seq	[23]

Figure legends:

Fig. 1. Venn diagram of the numbers of HK genes compiled from 9 microarray (MA) studies and 6 sequencing studies.

Fig. 2. Probe coverage (PC) of genes exclusively detected by sequencing.

Fig. 3. Chromosomal location of HK genes detected exclusively by sequencing based technique (sequencing genes, PC=0 and PC>0) and MA (MA genes) respectively as well as by commonly identified. Percentage was calculated by the gene count in each category to the total annotated genes in each chromosome.

Fig. 4. (A) Exon count of genes exclusively detected by MA, sequencing (PC=0 and PC>0) and the shared genes, respectively; (B) Total exon length of genes exclusively detected by MA, sequencing (PC=0 and PC>0) and their shared genes, respectively; (C) Total intron length of genes exclusively detected by MA, sequencing (PC=0 and PC>0) and their shared genes, respectively. (D) GC content of genes exclusively detected by MA, sequencing (PC=0 and PC>0) and their shared genes, respectively.

Fig. 5. (A) Abundance of HK genes exclusively detected by sequencing and MA, respectively (PC=0 and PC>0); **(B)** Detection breadth (DB) of genes exclusively detected by MA, sequencing (PC=0 and PC>0), respectively.

Fig. 6. The top-10 enriched gene ontology (GO) clusters (biological processes_FAT) of genes exclusively detected by MA, sequencing (PC > 0 and PC = 0), and the shared genes, respectively. The color represents value of –log P. Red sidebar highlights GO terms uniquely enriched in MA genes.

Figures:

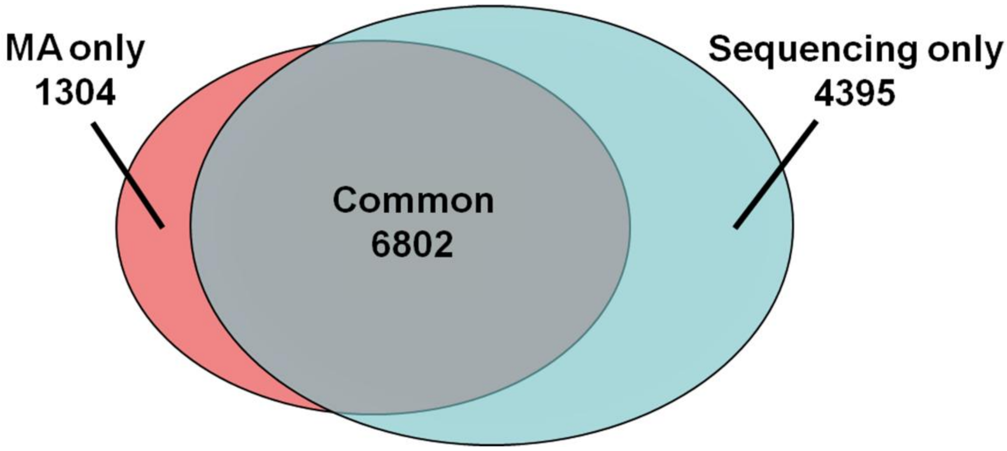


Fig. 1

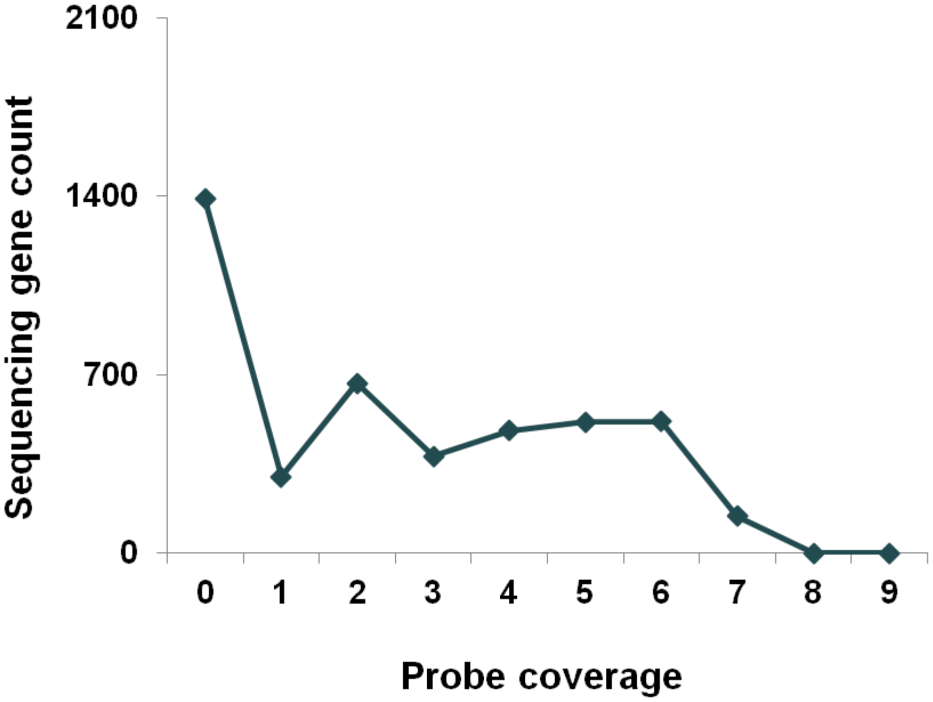
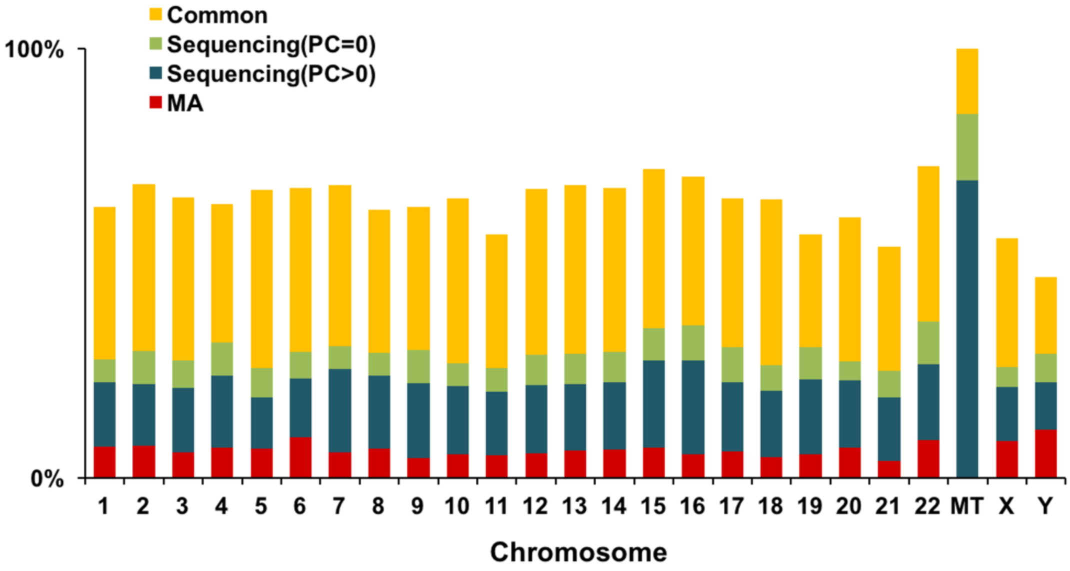

2100
1400
700
0
Sequencing gene count
0
1
2
3
4
5
6
7
8
9
Probe coverage

Fig. 2



Common
Sequencing(PC=0)
Sequencing(PC>0)
MA
100%
0%
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
MT
X
Y
Chromosome

Fig. 3

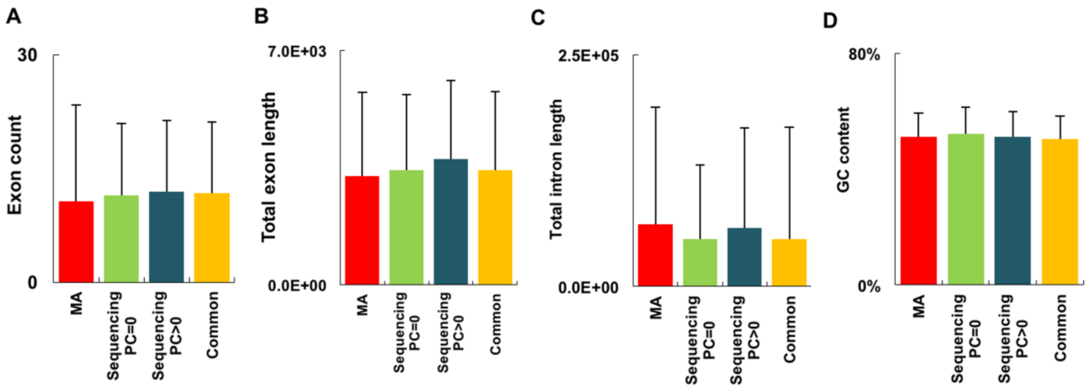

A
30
0
Exon count
MA
Sequencing PC=0
Sequencing PC>0
Common
B
7.0E+03
0.0E+00
Total exon length
MA
Sequencing PC=0
Sequencing PC>0
Common
C
2.5E+05
0.0E+00
Total intron length
MA
Sequencing PC=0
Sequencing PC>0
Common
D
80%
0%
GC content
MA
Sequencing PC=0
Sequencing PC>0
Common

Fig. 4

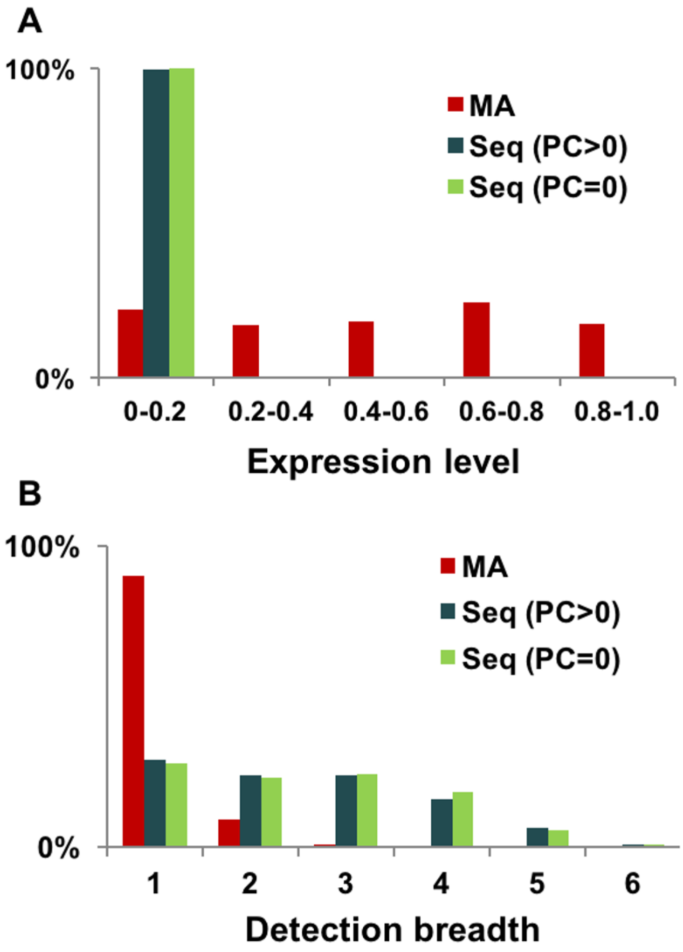

A
100%
0%
MA
Seq (PC>0)
Seq (PC=0)
0-0.2
0.2-0.4
0.4-0.6
0.6-0.8
0.8-1.0
Expression level
B
100%
0%
MA
Seq (PC>0)
Seq (PC=0)
1
2
3
4
5
6
Detection breadth

Fig. 5

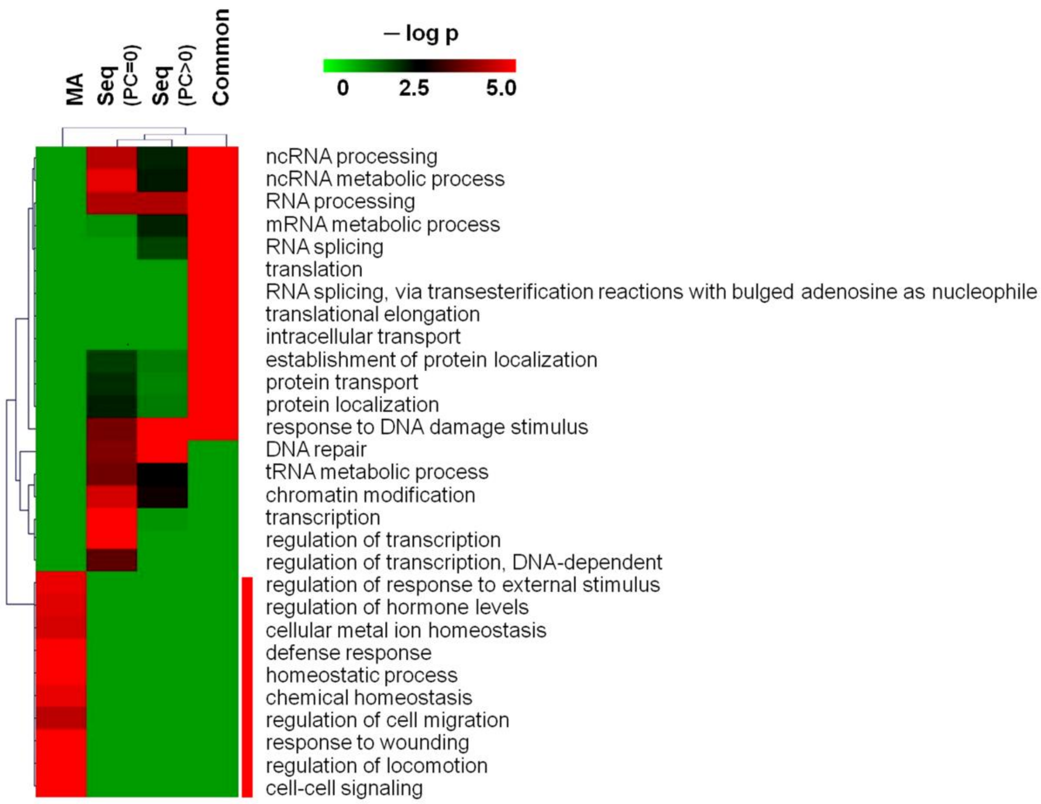

MA
Seq (PC=0)
Seq (PC>0)
Common
− log p
0
2.5
5.0
ncRNA processing
ncRNA metabolic process
RNA processing
mRNA metabolic process
RNA splicing
translation
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
translational elongation
intracellular transport
establishment of protein localization
protein transport
protein localization
response to DNA damage stimulus
DNA repair
tRNA metabolic process
chromatin modification
transcription
regulation of transcription
regulation of transcription, DNA-dependent
regulation of response to external stimulus
regulation of hormone levels
cellular metal ion homeostasis
defense response
homeostatic process
chemical homeostasis
regulation of cell migration
response to wounding
regulation of locomotion
cell-cell signaling

Fig. 6