Published in final edited form as:

*IEEE Int Conf Bioinform Biomed Workshops*. 2011 November ; 2011: 50–57. doi:10.1109/BIBMW.2011.6112354.

# Evaluation of Normalization Methods for RNA-Seq Gene Expression Estimation

**Po-Yen Wu**[1], **John H. Phan**[2], **Fengfeng Zhou**[3], and **May D. Wang**[2,*]

[1]Department of Electrical and Computer Engineering, Georgia Institute of Technology

[2]The Wallace H. Coulter Biomedical Engineering Department, Georgia Institute of Technology and Emory University

[3]Research Center for Biomedical Information Technology, Institute of Biomedical and Health Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

## Abstract

Statistical inferences on RNA-Seq data, e.g., detecting differential gene expression, are meaningful only after proper normalization. However, there is no consensus for choosing a normalization procedure from among the many existing procedures. We evaluated several RNA-Seq normalization procedures by (1) correlating estimated RNA-Seq expression values to those of microarrays, (2) examining the concordance of stable and differential gene detection between the platforms, and (3) applying the procedures to simulated RNA-Seq data. Results suggested that RNA-Seq normalization procedures have little effect on both inter-platform gene expression correlation as well as inter-platform concordance of genes detected as stably or differentially expressed. However, the results of simulated analysis suggested that some normalization procedures are more robust to changes in distribution of differentially expressed genes. These results may provide guidance for selecting RNA-Seq normalization procedures.

## 1. Introduction

Next-Generation Sequencing (NGS) technology has become a major platform for studying both genomics and transcriptomics [1]. RNA-Seq for quantifying RNA expression, one of the major applications of NGS technology, has received increased attention because of its potential to replace microarray technology. Some of the perceived benefits of RNA-Seq over microarrays include (1) improved dynamic range of expression detection and (2) the ability to detect a wide variety of RNA forms, e.g., small RNAs and splice variants, among others [2, 3]. Analogous to microarrays, normalization of RNA-Seq data to obtain quantitative and comparable RNA expression values is an important step [4–6]. Several experimental factors in the sequencing pipeline, e.g., library preparation, sequencing depth, and base calling, etc., can introduce biases in downstream RNA-Seq analysis. The purpose of the normalization step is to detect and calibrate such biases. However, it is unclear (1) how existing RNA-Seq normalization procedures differ when applied to the same data and (2) what NGS factors

*Corresponding Author, Contact information for the corresponding author: maywang@bme.gatech.edu, Phone: 404-385-2954, Fax: 404-894-4243, Address: Suite 4106, UA Whitaker Building, 313 Ferst Drive, Atlanta, GA 30332, USA.

affect normalization performance in terms of correlation with “true” reference RNA expression and detection of stable and differential gene expression. Using two NGS datasets, two microarray datasets (as “true” references) and simulated datasets, we compared several existing procedures for RNA-Seq normalization and evaluated them in terms of detecting stably and differentially expressed genes (SEGs and DEGs).

Current RNA-Seq normalization procedures consist of both simple and sophisticated scaling methods based on procedures for microarray normalization. Most of these procedures are simple global normalization procedures that use constant scaling factors for all genes within a single sequencing sample. We investigated some existing RNA-Seq normalization procedures including “Reads Per Kilobase of exon model per Million mapped reads (RPKM)” [7], “Trimmed Mean of M values (TMM)” [8], “Relative Log Expression (RLE)” [9], and “Quantiles” [10]. RPKM adjusts the total number of mapped reads per sample and the length of template transcripts per gene. However, RPKM can be biased by relatively small proportions of highly-expressed genes and, as such, can bias DEG detection [10]. The number of reads expected to map to a gene is not only dependent on the expression level and length of the gene, but also on the composition of the sampled RNA population. Normalization procedures such as TMM, RLE, and Quantiles attempt to estimate scaling factors between two samples to adjust total RNA output [8]. The TMM method trims log-ratio (M values) and log-average (A values) to find possible sets of stably expressed genes to estimate scaling factors. The RLE method generates a reference library by calculating the geometric mean of each gene across all samples; the median ratio of each sample to the reference is taken as the scaling factor. The Quantiles method uses the ratio of quantiles (e.g., upper quartile) between two samples as the scaling factor. The TMM, RLE, and Quantiles procedures belong to the group of global normalization procedures. The difference between these methods and RPKM is that they consider adjusting total RNA output rather than library size, which can reduce biases caused by highly-expressed genes.

In section 2, we describe RNA-Seq and microarray datasets used in our study as well as preprocessing, normalization and evaluation methods—i.e., correlation, stable and differential gene detection, and simulated data. In section 3, we present the results obtained after evaluating these methods. Lastly, we conclude this work in section 4.

## 2. Methods

Figure 1 illustrates the workflow of this study. We collected microarray and RNA-Seq datasets from online repositories and prepared them using several preprocessing and normalization steps. Our objective was to observe the effect of existing normalization procedures on (1) correlation between two platforms and (2) reproducibility of detecting stable and differential gene expression. We divided the evaluation methods into two categories: comparison with microarray datasets, including correlation analysis and SEG/DEG analysis, and simulation analysis. For correlation analysis, we explored the correlation between all possible pairs of datasets. For SEG/DEG analysis, we used algorithms to find sets of SEGs and DEGs from each dataset, and then observed the effect of normalization procedures on the concordance of these sets between datasets. Lastly, we

simulated datasets containing different RNA-Seq expression distributions to observe their effect on the performance of each normalization procedure.

### 2.1. Data acquisition

In this study, all datasets contain both kidney and liver samples. We obtained two NGS datasets (SRP000225 [11] and ERP000546 [12]) from the Sequence Read Archive (SRA) [13] and two microarray datasets (accession numbers: GSE11045 [11] and GSE3526 [14]) from the Gene Expression Omnibus [15]. Table 1 summarizes these datasets.

SRP000225 and GSE11045 belong to the same study and are composed of three technical replicates for each tissue (kidney or liver) and platform (microarray or NGS) combination. Data were acquired using Illumina Genome Analyzer and Affymetrix GeneChip HG U133 Plus 2.0 Arrays. These two datasets are denoted as NGS_1 and MA_1.

The ERP000546 study contains transcription profiling data from Illumina's Human BodyMap 2.0 project. This project sequenced 16 human tissues, including kidney and liver, using the Illumina HiSeq 2000 platform. The dataset includes both single-end and paired-end sequencing data for each biological sample. This dataset is denoted as NGS_2.

The GSE3526 study used Affymetrix GeneChip HG U133 plus 2.0 Arrays to profile the gene expression of 353 biological samples representing 65 tissues. Each tissue has three to nine biological replicates. Among them, there are four liver samples, four kidney cortex samples, and four kidney medulla samples. This dataset is denoted as MA_2.

### 2.2. NGS and microarray data preprocessing

We used caCORRECT and Robust Multichip Average (RMA) to calculate microarray probeset expression in the log-base two scale [16, 17]. To directly compare expression values between NGS and microarray technologies, we aligned reads from the NGS datasets to microarray probeset sequences. We focused on probeset expression rather than gene expression to avoid the complexity of mapping probesets to genes. We used an alignment reference composed of target sequences from the Affymetrix GeneChip HG U133 Plus 2.0 Array with an additional padding of 100 base pairs at both the 3' and 5' ends. We obtained padding sequences from the original exemplar or consensus mRNA transcripts of the same microarray chip. We used BLAT [18] to find the location of target sequences in the exemplar/consensus mRNA transcripts. Padding is necessary because we do not want to ignore reads aligned to the boundary of target sequences.

We then filtered probesets based on number of aligned short sequences. Probesets with low expression (i.e., with only a small number of short sequences aligned) may contribute to noise in the gene expression estimation. We discarded a probeset if any of its expression values across all samples in the dataset falls below 5, 10, or 30 aligned short sequences. Table 2 lists the number of probesets remaining in each dataset after quality filtering.

We calculated expression values for RNA-Seq data by considering the deepest coverage of all base pairs within the target sequence region of each probeset. We used bwa [19] to align the sequences, resulting in around 3.3% and 11.2% of reads uniquely aligning to target

sequence regions for the NGS_1 and NGS_2 datasets, respectively. We transformed all NGS expression values to the log-base two scale in order to more closely match microarray gene expression values.

### 2.3. Correlation analysis

We used Pearson correlation to examine cross-platform gene expression concordance for six combinations of four datasets to better understand the influence of normalization procedures. The factors we explored include normalization procedures as well as quality filtering thresholds. We also used a test for equality of multiple correlation coefficients [20] to determine if differences between correlations were statistically significant.

### 2.4. Analysis of stably expressed genes

Assuming a dataset has multiple samples {T_1,T_2,…,T_m}, and each sample has expression values {E_(T_i,1),E_(T_i,2),…,E_(T_i,n)} for all probesets. Given a threshold C > 0, a probeset “x” is claimed to be stably expressed if

$$\max \left| (E_{T_i,x} - E_{T_j,x}) \right| \leq C \text{ where} i, j \in \{1, 2, \ldots, m\} \text{and} i \neq j \quad (1)$$

We ranged threshold C from 0.1 to 1.9, which corresponds to fold-changes of 1.07 to 3.73.

We applied equation (1) to all four datasets and obtained SEG sets. The following sets are defined from the original SEG sets (named as in Table 1): NGS_1∩MA_1, NGS_1∩MA_2, NGS_1∩MA_1∩MA_2 for NGS_1 dataset, and NGS_2∩MA_1, NGS_2∩MA_2, NGS_2∩MA_1∩MA_2 for NGS_2 dataset. Based on the assumption that microarrays are the true reference, we calculated the proportion of SEG calls that are simultaneously supported by NGS and microarray datasets to those supported by microarray datasets only. The factors we investigated include normalization procedures as well as quality filtering thresholds.

### 2.5. Analysis of differentially expressed genes

Assuming a dataset as described in section 2.4. Given a threshold D > 0, a probeset “x” is claimed to be differentially expressed if

$$\min \left| (E_{T_i,x} - E_{T_j,x}) \right| \geq D \text{ where } i \in \text{indexes of tissue } 1 \; j \in \text{indexes of tissue } 2 \text{ and } i \neq j$$

(2)

We ranged threshold D from 1.0 to 3.0, which corresponds to fold-changes of 2 to 8. The evaluation process is the same as that of SEG analysis.

### 2.6. Simulation analysis

We assessed the robustness of normalization procedures to various distributions using simulated RNA-Seq datasets. Table 3 summarizes seven expression distributions. We directly simulated raw counts of aligned short sequences so that we can eliminate errors introduced from the sequencing and alignment steps. Using both realistic and hypothetical distributions, we can identify pros and cons of each normalization procedure. The underlying assumption is that these distributions represent true absolute expression levels that normalization procedures should be able to quantify.

## 3. Results and discussion

### 3.1. Correlation analysis

Figure 2 demonstrates the concordance of probeset expression between each pair of four datasets. Intra-platform comparisons generally have higher correlation. For NGS data, increasing the quality filtering threshold decreases the correlation coefficient.

An important result of this study is that there is no difference between normalization procedures in terms of correlation coefficients (i.e., we obtained a p-value of 1 using the test for equality of multiple correlation coefficients [20] with a null hypothesis that there is no significant difference among all correlation coefficients). This is a reasonable conclusion since correlation coefficient is unaffected by changes to data scale. We used scaling normalization procedures. Thus, the differences among these methods cannot be detected via correlation coefficient.

### 3.2. Analysis of stably expressed genes

We compared sets of SEGs identified in datasets NGS_1, MA_1, NGS_2, and MA_2. Stable gene expression is biologically significant because it implies that a gene is functionally essential. Such stably expressed genes are widely used as references in transcriptome studies [21–23]. The distribution of overlap percentage between NGS (with quality filtering of 30) and microarray SEG sets is illustrated in Figure 3.

The overlap between SEGs identified from NGS_1 and SEGs identified from the microarray datasets is consistent regardless of the microarray dataset (MA_1, MA_2, or MA_1 · MA_2) (Figure 3a). Results are similar for NGS_2 (Figure 3b). Although microarray data is notoriously noisy, there is high inter-platform concordance of biological information generated from microarrays [24]. This supports our use of microarrays as “true” references for assessing NGS normalization procedures. The raw count of the aligned short sequences seems to be the worst expression estimation procedure, identifying the lowest proportion of shared SEGs with both microarray datasets. The RPKM procedure did not perform consistently on different NGS datasets, with ~14% lower overlap of SEGs in NGS_2 compared to NGS_1. The other three procedures, i.e. TMM, RLE, and Quantiles, consistently identified SEGs when applied to either NGS_1 or NGS_2.

### 3.3. Analysis of differentially expressed genes

We used fold-change to detect DEGs from four datasets and compared the concordance of these DEG sets. The overlap of DEGs between NGS (with quality filtering of 30) and microarray datasets is illustrated in Figure 4.

The performance of all five methods (Raw, RPKM, TMM, RLE, Quantiles) varied depending on which NGS dataset was used. The key observation from Figure 4 is that TMM, RLE, and Quantiles performed similarly. However, the overlap between DEGs varied depending on the dataset. In contrast to SEG analysis, the raw count of NGS alignment performed fairly well for NGS_1 and NGS_2 in terms of concordance with microarrays.

### 3.4. Simulation analysis

Using microarray data as a reference, RNA-Seq normalization procedures appeared to have little to no effect on gene expression correlation and on concordance of SEG/DEG sets. Thus, based on our limited study of four datasets, we cannot draw any conclusions about the relative performances of normalization procedures. However, the assumption that microarrays can be used as a reference may be problematic. Noise in microarray datasets (especially in such small sample datasets), can hinder the detection of SEGs or DEGs. We simulated several RNA-Seq expression distributions so that we can compare the results of RNA-Seq normalization methods to a true, expected result rather than to a questionable microarray result.

The distributions of simulated datasets are listed in Table 3. In Figure 5, row 1 we computed the fold-change between data2 and data1 and expected a 10-fold-change for some of the probesets. TMM and RLE detected the correct fold-changes, but RPKM and Quantiles did not. In row 2 of Figure 5, we computed the fold-change between data3 and data1. Again, TMM and RLE were able to detect the correct 3-fold-change, whereas RPKM and Quantiles were biased. Row 3 is an extreme and unrealistic case; it assumes that the data is uniformly distributed. TMM was the only method to correctly compute differential expression values for this case. Row 4 is a case in which the data contains many zero expression values. The only failed procedure in this case was RPKM. Row 5 compared data6 to data1. Data6 was generated from data1 with additional random noise for all probesets. Since the noise is small, all normalization procedures performed well. Row 6 is a similar case to Row 2 but with only 5% of probesets up- or down-regulated instead of 20%. Since the number of differentially expressed genes is small and the multiplier is only 3-fold, all procedures worked well in this case.

These simulated data results suggest that TMM is the most robust procedure because it correctly computed differential gene expression in a variety of RNA-Seq distributions. The RPKM and Quantiles procedures were susceptible to data distributions with either a few highly-expressed genes or many differentially expressed genes. RLE is also a robust procedure; the only case in which it failed was the extreme case of uniformly distributed expression values.

## 4. Conclusion

Normalization methods to quantify gene expression from RNA-Seq data directly influence downstream analysis. We explored some existing RNA-Seq normalization procedures, including RPKM, TMM, RLE, and Quantiles, to assess their performance in terms of correlation with microarrays, reproducibility of SEGs and DEGs, and tolerance to different RNA-Seq expression distributions. In this study, we treated microarrays as “true” references. Results suggested that normalization has no effect on correlation between RNA-Seq and microarray data. All procedures generated the same Pearson correlation coefficients. In analyzing the overlap of SEGs between RNA-Seq and microarray datasets, RPKM performance changed slightly depending on the NGS dataset while TMM/RLE/Quantiles performed identically regardless of the NGS or microarray dataset. In analyzing the overlap of DEGs between RNA-Seq and microarray datasets, the choice of NGS or microarray dataset affected performance, but the TMM/RLE/Quantiles methods performed identically. Using simulated realistic and hypothetical RNA-Seq expression distributions, we observed that TMM worked well in all cases, and RLE failed only in an extreme case. Quantiles and RPKM were both affected by the distribution of DEGs.

To conclude, detecting the differences among RNA-Seq normalization methods depends on the evaluation criteria. When using microarrays as a “true” reference, the differences among methods are difficult to detect. However, careful analysis using simulated NGS datasets revealed that some of these RNA-Seq normalization methods are sensitive to the distribution of DEGs. Thus, DEG distribution is an important factor when choosing a normalization method for RNA-Seq data analysis.

## Acknowledgments

This work has been supported by grants from the Parker H. Petit Institute for Bioengineering and Bioscience (IBB), National Institutes of Health (NIH) (Bioengineering Research Partnership R01CA108468, Center for Cancer Nanotechnology Excellence U54CA119338), National Cancer Institute (NCI); Georgia Cancer Coalition (Distinguished Cancer Scholar Award to MDW), and Georgia Research Alliance; Hewlett-Packard and Microsoft Research. This work was also supported in part by National S&T Major Project of China (Grant No. 2009ZX03006-001-01) and Key Basic Research Program of Shenzhen The authors would like to thank Dr. Todd Stokes and all workshop reviewers for valuable and helpful comments.

## References

1. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010 Jan.11:31–46. [PubMed: 19997069]
2. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nature reviews. Genetics. 2011; 12:87.
3. Wang Z, et al. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan.10:57–63. [PubMed: 19015660]
4. Pradervand S, et al. Impact of normalization on miRNA microarray expression profiling. RNA. 2009 Mar.15:493–501. [PubMed: 19176604]
5. Lee S, et al. Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. Nucleic Acids Res. 2011 Jan.39:e9. [PubMed: 21059678]
6. Roberts A, et al. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 2011 Mar 16.12:R22. [PubMed: 21410973]
7. Mortazavi A, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008 Jul.5:621–628. [PubMed: 18516045]

8. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010; 11:R25. [PubMed: 20196867]
9. Bolstad, BM., et al. Quality Assesement of Affymetrix GeneChip Data. In: Gentleman, R., et al., editors. Bioinformatics and Computational Biology Solutions using R and Bioconductor. New York: Springer; 2005. p. 33-48.
10. Bullard JH, et al. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010; 11:94. [PubMed: 20167110]
11. Marioni JC, et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008 Sep.18:1509–1517. [PubMed: 18550803]
12. Illumina bodyMap2 transcriptome. 2011. Available: http://www.ncbi.nlm.nih.gov/sra?term=ERP000546
13. Wheeler DL, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2008 Jan.36:D13–D21. [PubMed: 18045790]
14. Roth RB, et al. Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. Neurogenetics. 2006 May.7:67–80. [PubMed: 16572319]
15. Barrett T, et al. NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res. 2011 Jan.39:D1005–D1010. [PubMed: 21097893]
16. Moffitt R, et al. caCORRECT2: Improving the accuracy and reliability of microarray data in the presence of artifacts. BMC Bioinformatics. 2011; 12:383. [PubMed: 21957981]
17. Irizarry RA, et al. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003 Feb 15.31:e15. [PubMed: 12582260]
18. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002 Apr.12:656–664. [PubMed: 11932250]
19. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15.25:1754–1760. [PubMed: 19451168]
20. Paul SR. Test for the equality of several correlation coefficients. Canadian Journal of Statistics. 1989; 17:217–227.
21. Klein C, et al. Expression stability of putative reference genes in equine endometrial, testicular, and conceptus tissues. BMC Res Notes. 2011; 4:120. [PubMed: 21486450]
22. Narsai R, et al. Defining reference genes in Oryza sativa using organ, development, biotic and abiotic transcriptome datasets. BMC Plant Biol. 2010; 10:56. [PubMed: 20353606]
23. Stamova BS, et al. Identification and validation of suitable endogenous reference genes for gene expression studies in human peripheral blood. BMC Med Genomics. 2009; 2:49. [PubMed: 19656400]
24. Li Z, et al. Microarray platform consistency is revealed by biologically functional analysis of gene expression profiles. BMC Bioinformatics. 2009; 10:S12. [PubMed: 19811677]

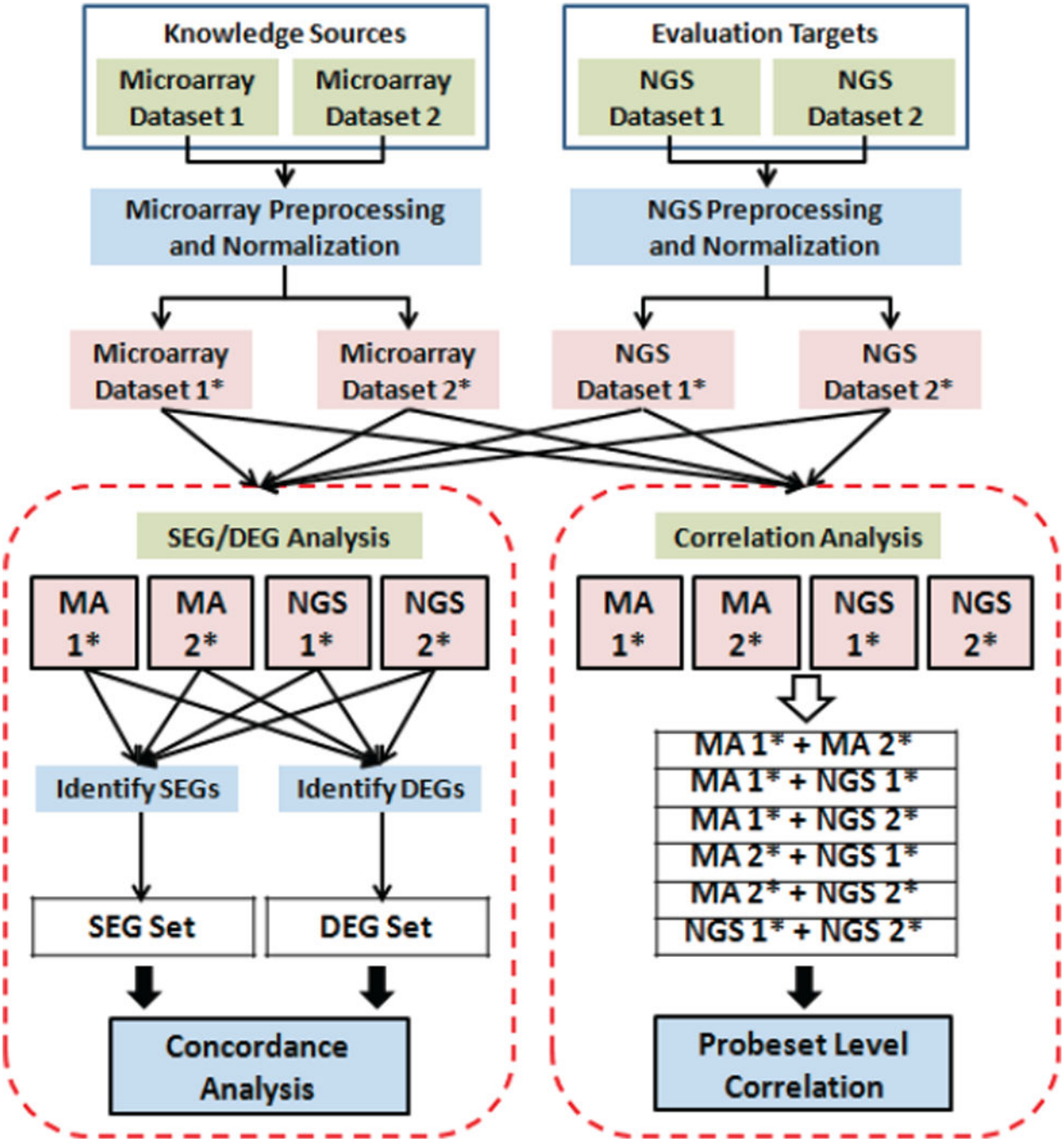


**Figure 1.**

Workflow for evaluating RNA-Seq normalization procedures. The evaluation methods include SEG and DEG analysis (concordance of gene sets between datasets) and probeset level correlation analysis. Abbrev.: SEG – Stably Expressed Genes; DEG – Differentially Expressed Genes; MA – MicroArray.

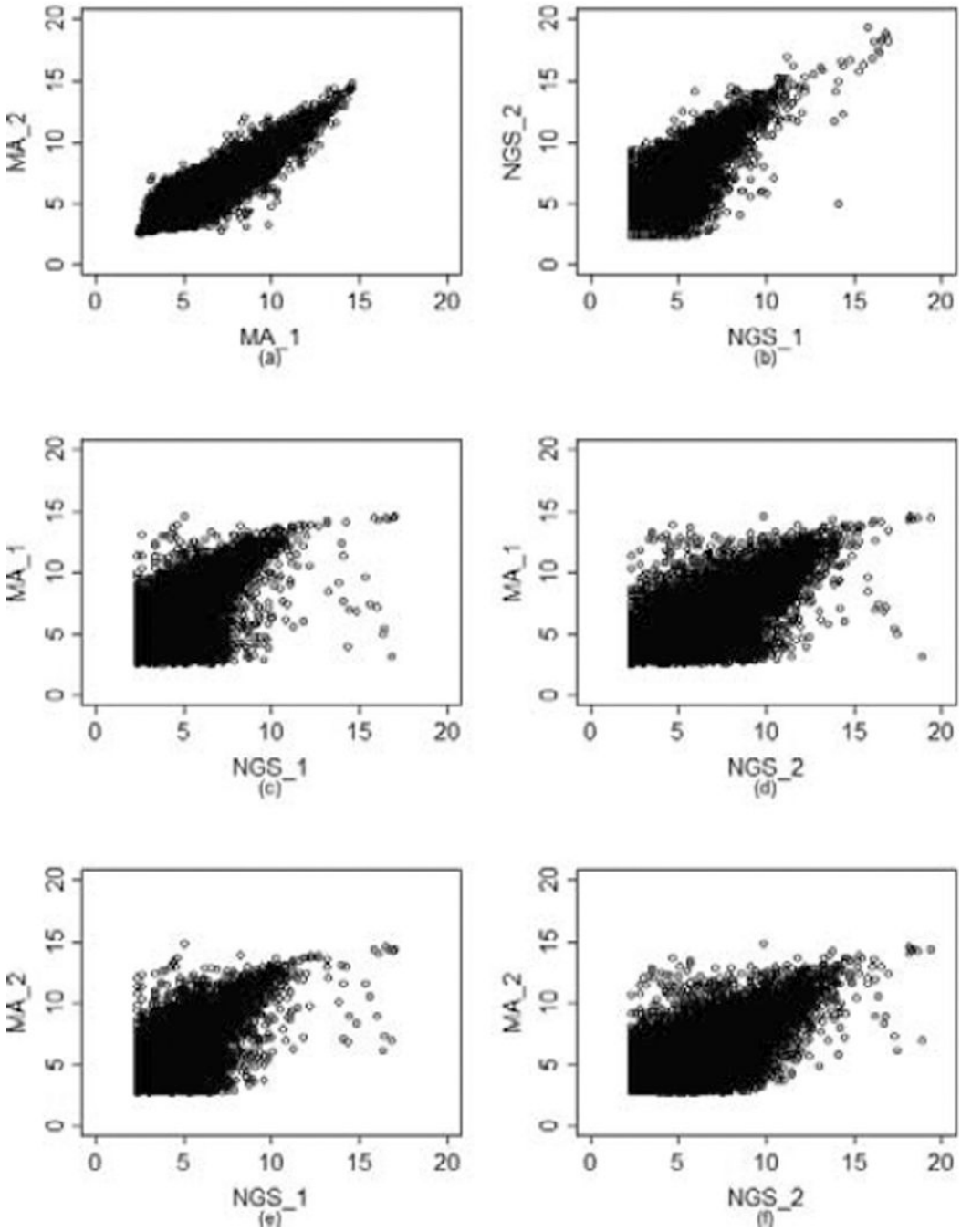


**Figure 2.**

Scatter plots show the relationship between probeset expression from kidney samples for each pair of four datasets (NGS_1, NGS_2, MA_1, and MA_2). Expression values For NGS datasets have been summed across samples and quality filtered with a threshold of 5. All plots use probesets that intersect between pairs of datasets after filtering. Pearson correlation coefficients: (a) 0.921, (b) 0.763, (c) 0.611, (d) 0.686, (e) 0.602, (f) 0.618.

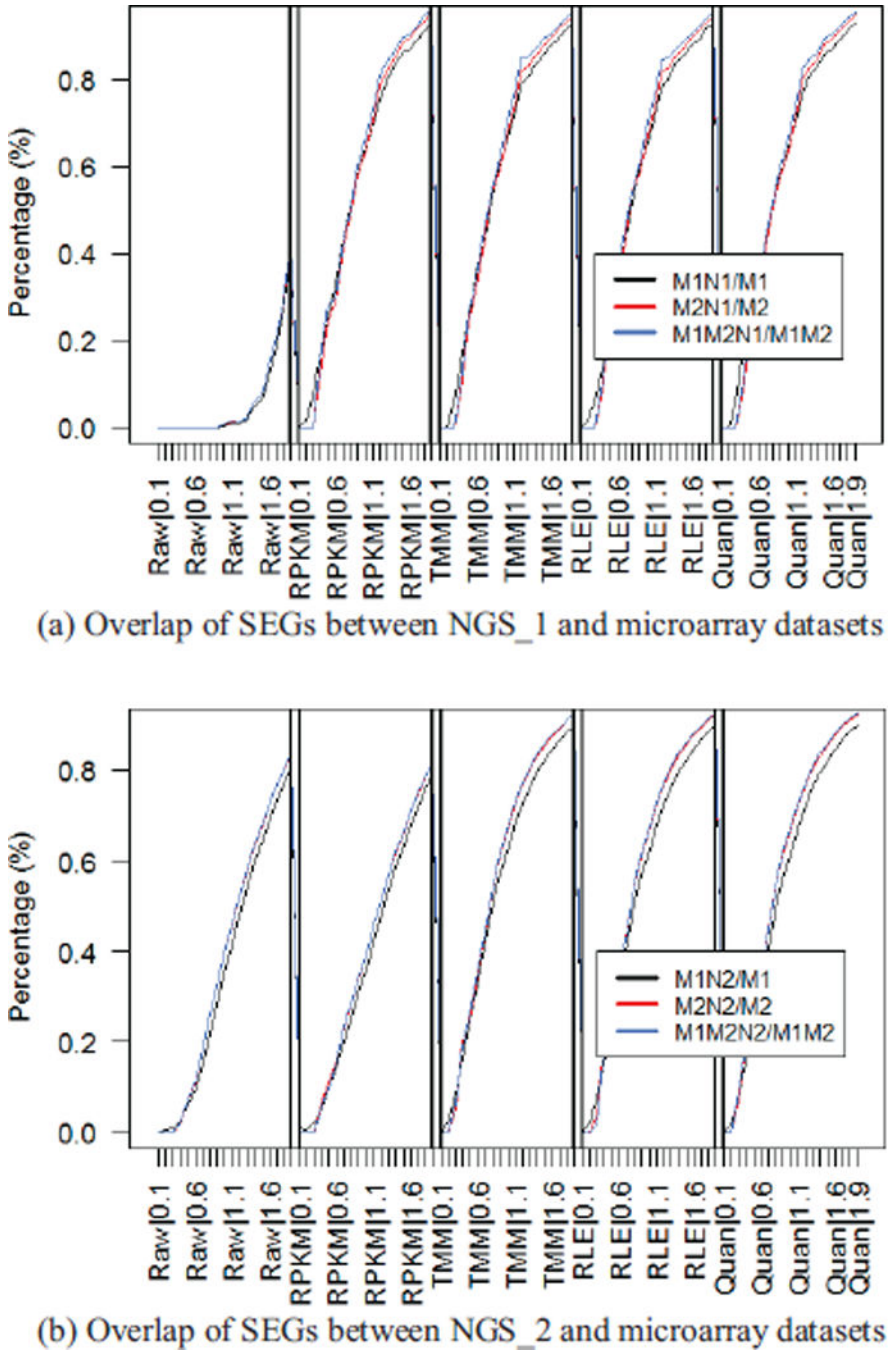


**Figure 3.**

Overlap of SEGs between NGS and microarray datasets using five NGS normalization procedures: raw alignment count, RPKM, TMM, RLE, and Quantiles. The thresholds for SEG detection are $\mathbf{C} \in \{\mathbf{0.1, 0.2, \ldots, 1.9}\}$. Three combinations of datasets are used: M1Nx/M1, M2Nx/M2, and M1M2Nx/M1M2 (where × is either 1 or 2). The value M1 is the number of SEGs in MA_1, and M1Nx is the number of SEGs in both MA_1 and NGS_x (x=1 or 2). The value M2 is the number of SEGs in MA_2, and M1M2 is the number of

SEGs in both MA_1 and MA_2. The values M2Nx and M1M2Nx are defined similarly to M1Nx.

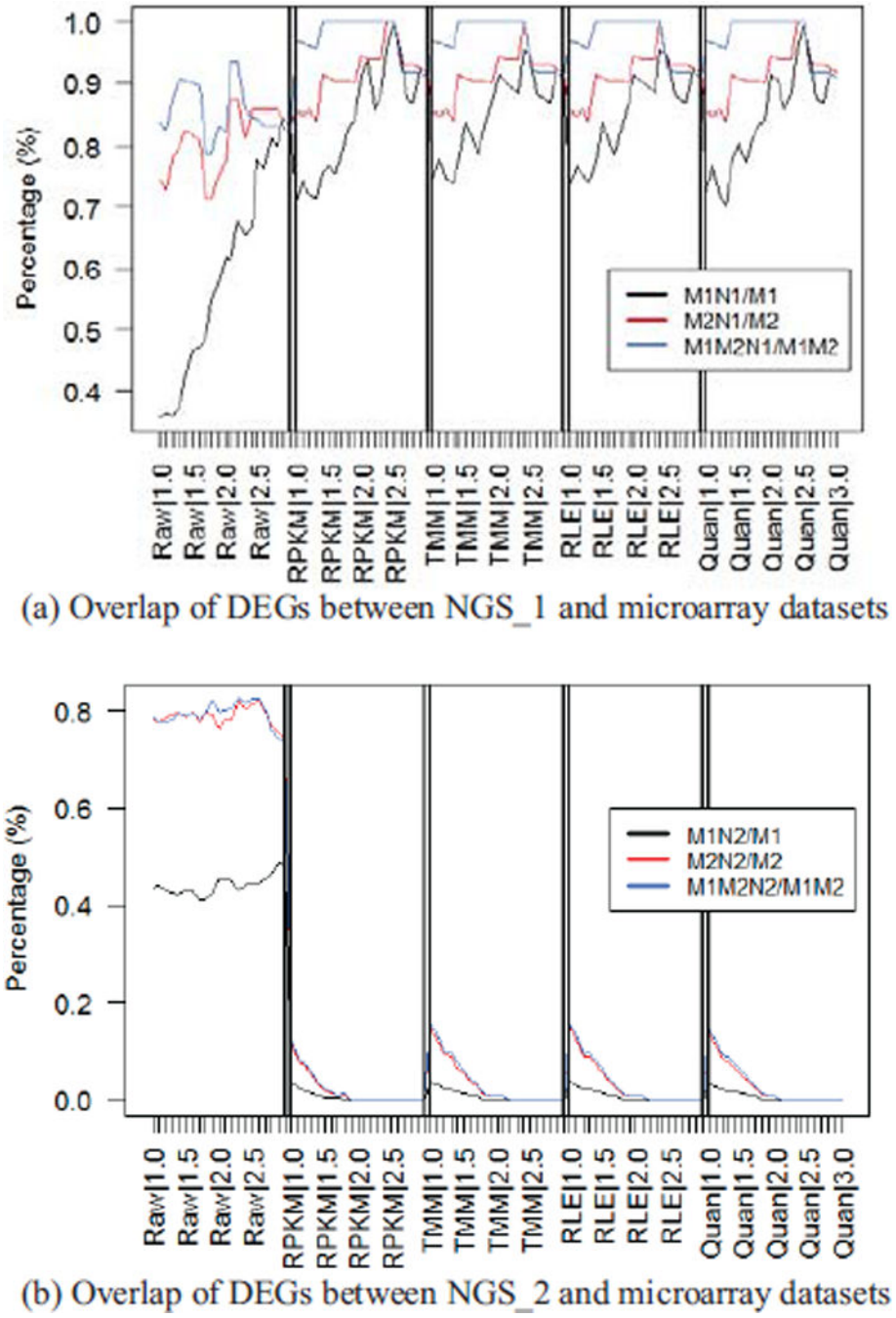


**Figure 4.**

Overlap of DEGs between NGS and microarray datasets using five NGS normalization procedures: raw alignment count, RPKM, TMM, RLE, and Quantiles. The thresholds for DEG detection are **D** ∈ {**1.0, 1.1, …, 3.0**}. Three combinations of data sets are used: M1Nx/M1, M2Nx/M2, and M1M2Nx/M1M2 (where × is either 1 or 2). The value M1 is the number of DEGs in MA_1, and M1Nx is the number of DEGs in both MA_1 and NGS_x (x=1 or 2). The value M2 is the number of DEGs in MA_2, and M1M2 is the number of

DEGs in both MA_1 and MA_2. The values M2Nx and M1M2Nx are defined similarly to M1Nx.

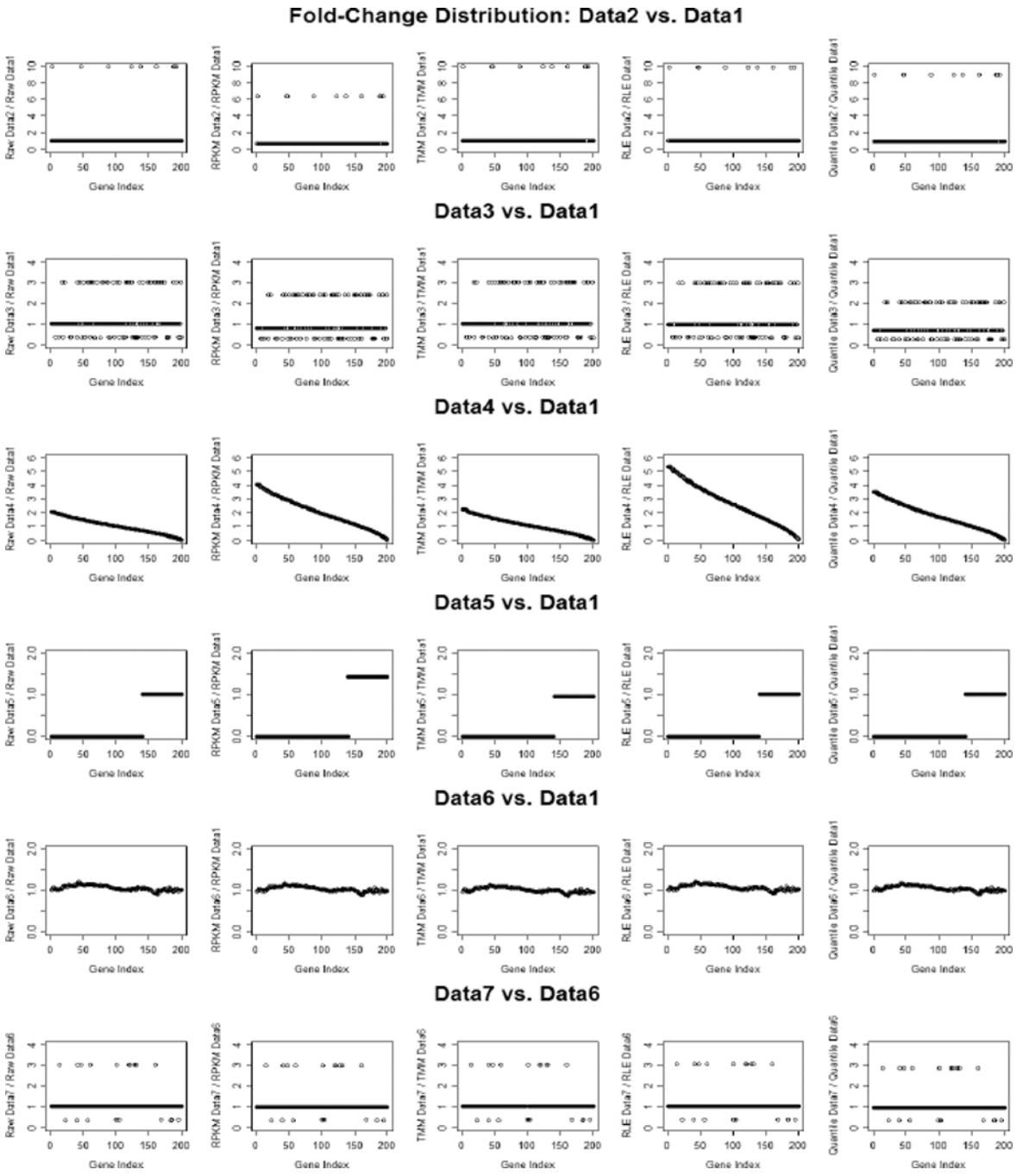


**Figure 5.**

Examining the performance of several RNA-Seq normalization methods using various simulated distributions of DEGs. The left-most plot of each row is the ground truth. Columns 2 to 5 are results for the RPKM, TMM, RLE, and Quantiles methods DEG distributions are described in Table 3.

**Table 1**

Summary of 2 NGS and 2 Microarray Datasets

| **Technology** | **Microarray** | | **Next-Gen Sequencing** | |
|---|---|---|---|---|
| **Accession #** | GSE11045 | GSE3526 | SRP000225 | ERP000546 |
| **Replicates** | Technical | Biological | Technical | Technical |
| **# of Sample {K,L}** | {3,3} | {8,4} | {3,3} | {2,2} |
| **Name** | MA_1 | MA_2 | NGS_1 | NGS_2 |

* K – kidney and L – liver

**Table 2**

Number of Probesets Remaining after Quality Filtering

| Quality Filtering Threshold | 5 | 10 | 30 |
|---|---|---|---|
| NGS_1 Dataset | 9521 | 3989 | 702 |
| NGS_2 Dataset | 24924 | 20609 | 13238 |

* The unit of thresholds is the number of aligned short sequences.

* Original dataset has 54675 probesets.

**Table 3**

Simulated RNA-Seq Expression Distributions

| | |
|---|---|
| **1)** | Sample 200 probesets from sorted NGS_1 data (Threshold-30, total kidney expression) with equal distance |
| **2)** | Randomly select 5% of probesets from (1) and increase expression values by 10-fold |
| **3)** | Randomly select 40% of probesets from (1). Increase expression of half of these probesets by 3-fold, decrease the other half by 3-fold |
| **4)** | Uniform distribution of median expression of (1) |
| **5)** | Set the lowest 70% probesets in (1) to 0 |
| **6)** | Generate random numbers following the histogram of (1) with a window size of 40 |
| **7)** | Randomly select 10% of probesets from (6). Increase expression of half of these probesets by 3-fold, decrease the other half by 3-fold |