Title

A Hypothesis Testing Based Method for Normalization and Differential Expression Analysis of RNA-Seq Data.

Journal

PLoS ONE, 12(1)

Authors

Zhou, Yan
Wang, Guochang
Zhang, Jun
et al.

Publication Date

2017

DOI

10.1371/journal.pone.0169594

PLOS ONE

RESEARCH ARTICLE

A Hypothesis Testing Based Method for Normalization and Differential Expression Analysis of RNA-Seq Data

Yan Zhou[1], Guochang Wang[2], Jun Zhang[1], Han Li[3]*

1 College of Mathematics and Statistics, Institute of Statistical Sciences, Shenzhen University, Shenzhen, China, **2** College of Economics, Jinan University, Guangzhou, China, **3** College of Economics, Shenzhen University, Shenzhen, China

* hli@szu.edu.cn

Citation: Zhou Y, Wang G, Zhang J, Li H (2017) A Hypothesis Testing Based Method for Normalization and Differential Expression Analysis of RNA-Seq Data. PLoS ONE 12(1): e0169594. doi:10.1371/journal.pone.0169594

Editor: Rogerio Margis, Universidade Federal do Rio Grande do Sul, BRAZIL

Received: September 8, 2016

Accepted: December 18, 2016

Published: January 10, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. Data are also from the Marioni JC study whose authors may be contacted at http://genome.cshlp.org/content/18/9/1509/suppl/DC1.

Funding: This work was supported by the Tianyuan Fund for Mathematics (No. 11526143), the Doctor Start Fund of Guangdong Province (No. 2016A030310062 (85118-000043)) and the Natural Science Foundation of SZU (No. 836-00008303) to Yan Zhou. This work was also supported by the National Science Foundation of

Abstract

Next-generation sequencing technologies have made RNA sequencing (RNA-seq) a popular choice for measuring gene expression level. To reduce the noise of gene expression measures and compare them between several conditions or samples, normalization is an essential step to adjust for varying sample sequencing depths and other unwanted technical effects. In this paper, we develop a novel global scaling normalization method by employing the available knowledge of housekeeping genes. We formulate the problem from the hypothesis testing perspective and find an optimal scaling factor that minimizes the deviation between the empirical and the nominal type I error. Applying our approach to various simulation studies and real examples, we demonstrate that it is more accurate and robust than the state-of-the-art alternatives in detecting differentially expression genes.

Introduction

In recent years, next-generation sequencing methods, for instance, ChIP-seq and RNA-seq, due to their distinct advantages in increasing specificity and sensitivity of gene expression, they have become a poular choice in biological studies. Such sequence-based methods have evoked a wide range of novel applications, for instance, splicing variants [1, 2] and single nucleotide polymorphisms [3]. Specifically, RNA-seq has become an attractive alternative to microarrays in the inference of differential expression (DE) between several conditions or tissues, for it gives more accurate detection and measure of gene expression.

We first map the RNA-seq reads to the reference genome and then summarize them as "counts". That is, we use a count number to measure the expression level of each gene. Under different conditions/tissues, the experiments will result in different total read counts, that is, sequencing depths. In order to make the expression levels of genes comparable and further conduct differential expression analysis, normalization is a crucial step in data processing. The normalization step aims to adjust the systematic technical effects and reduce the noise on the data as well.

China (Nos. 11501248 and 11601094) to Guochang Wang and by the Tianyuan Fund for Mathematics (No. 11626160) to Han Li. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Considering the essential difference in technology between microarray and RNA-seq, we can not normalize RNA-seq data with the normalization methods of microarray data directly. A conventional way of RNA-seq analysis is to standardize the data between samples by scaling their total number of reads to a common value. More sophisticated normalizaiton methods could be divided into two groups, which are referred to as the library size concept adjustment and distribution of read counts adjustment. In the first group, several researchers have developed some normalization methods such as modifying the mean expression of a gene with a global factor [4–7]. For instance, Hoen et al. [8] used the square root of scaled counts to analysis LONGSAGE-seq data, and Mortazavi et al. [9] modified sequencing data to reads per kilobase per million mapped (RPKM). Robinson et al. [10] proposed a scale normalization method (TMM), which is a weighted trimmed mean of log-ratios between the test sample and the reference sample. However, TMM method could not normalize the data satisfactorily well for asymmetric data, especially, for large proportion of DE genes. Zhou et al. [11] used an iterating median fold changes to estimate the scale factor and showed that it is more rubust than the TMM method for asymmetric data. In the second group, the standard procedure is to first compute the proportion of each gene's reads relative to the total number of reads in each library. Assuming there are similarities between the distributions of read counts, the methods [12] [13] match the distributions of all genes across all libraries, either on a single quantile or on all the quantiles. Nontheless, the RNA repertoires may change diversely under different experimental conditions, thus the proportions of gene expressions are not comparable in such case. Some authors proposed to use the housekeeping genes as pivot points and match the distributions of those housekeeping genes, instead of all of genes, for inter-sample normalization. Bullard et al. [13] used a single housekeeping gene in normalization. Chen et al. [14] proposed a method to select a subset of housekeeping genes by analyzing experimentally related GO terms and the stability of gene expression.

The key of normalization problem is to choose an appropriate metric of expression to compare across samples. We propose a novel normalization method by exploiting the knowledge of housekeeping genes. We address the problem from the hypothesis testing perspective, by matching the observed and the nominal false discovery rate. The estimated normalization scaling factor is expected to be stable for different confidence level in the hypothesis testing. Thus it can normalize the samples without trimming the data and avoids the problem of the TMM like methods.

The remainder of the paper is organized as follows. In Section 2, we propose a hypothesis testing based method for normalization and detection of DE genes. Subsequently we carry out extensive simulation studies in Section 3. In Section 4, we evaluate the merits of our approach by applying it to a liver and kidney dataset, and demonstrate that it outperforms the alternatives. Finally, some conclusions and suggestions are made.

Materials and Methods

A hypothesis testing based normalization scaling factor method

We propose a new normalization procedure, called hypothesis testing based normalization (HTN), to reduce the bias of normalization by employing the available knowledge of housekeeping genes. We first introduce some notations. Let Y_{gk} be the observed count and μ_{gk} be the true expression level of gene g in library k, where $k = 1, 2$ and $g = 1, \ldots, G$. The length of gene g is denoted by L_g and the total number of reads for library k is denoted by N_k. Assuming the observed count is proportional to the product of the true expression level and the gene length,

the expected value of Y_{gk} is formulated as

$$E[Y_{gk}] = \frac{\mu_{gk} L_g}{S_k} N_k, \tag{1}$$

where $S_k = \sum_{g=1}^{G} \mu_{gk} L_g$ is the total RNA expression of sample k. For two samples or libraries, we test

$$\mathrm{H}_{0g} : \mu_{g1} = \mu_{g2} \quad vs \quad \mathrm{H}_{1g} : \mu_{g1} \neq \mu_{g2} \quad \text{for all } g. \tag{2}$$

Under Eq (1), the above test is equivalent to

$$\mathrm{H}_0 : E[Y_{g1}] = \left(\frac{S_2}{S_1} \times \frac{N_1}{N_2}\right) E[Y_{g2}] \quad vs \quad \mathrm{H}_1 : E[Y_{g1}] \neq \left(\frac{S_2}{S_1} \times \frac{N_1}{N_2}\right) E[Y_{g2}] \quad \text{for all } g. \tag{3}$$

Let $c = S_2/S_1$ be the scaling factor of sample 2 relative to sample 1. Assuming that the counts mapping to a gene are Poisson-distributed, that is, $Y_{gk} \sim Pois(\lambda_{gk})$, the test could be specified as

$$\mathrm{H}_{0g} : \lambda_{g1} = c\frac{N_1}{N_2}\lambda_{g2} \quad vs \quad \mathrm{H}_{1g} : \lambda_{g1} \neq c\frac{N_1}{N_2}\lambda_{g2} \quad \text{for all } g. \tag{4}$$

Conditioning on $Y_{g1} + Y_{g2} = n_g$, we can derive that Y_{g1} follows a binomial distribution, that is,

$$P(Y_{g1} \mid Y_{g1} + Y_{g2} = n_g) = \frac{n_g!}{Y_{g1}!(n_g - Y_{g1})!} p_0^{Y_{g1}} (1 - p_0)^{n_g - Y_{g1}}, \tag{5}$$

where $p_0 = \lambda_{g1}/(\lambda_{g1} + \lambda_{g2}) = (cN_1/N_2)/(1 + cN_1/N_2)$. The p-value for testing H_{0g} is then calculated as

$$\begin{aligned} p_g(c) &= P\left(\left|Y_{g1} - c\frac{N_1}{N_2}(n_g - Y_{g1})\right| \geq \left|y_{g1} - c\frac{N_1}{N_2}y_{g2}\right|\right) \\ &= P\left(\left|\left(1 + c\frac{N_1}{N_2}\right)Y_{g1} - c\frac{N_1}{N_2}n_g\right| \geq \left|\left(1 + c\frac{N_1}{N_2}\right)y_{g1} - c\frac{N_1}{N_2}n_g\right|\right) \end{aligned} \tag{6}$$

where y_{g1}, y_{g2} are the observed counts of gene g in these two samples, respectively, and $n_g = y_{g1} + y_{g2}$. Note that $p_g(c)$ is a function of unknown c. Once c is determined, we could calculate the p-values for all genes and hence determine which genes are differentially expressed.

In our method, we are supposed to have a set of housekeeping genes in priori, which could be reported in published studies or selected based on certain biological information, for example, the GO terms of the genes [14]. Assume we have m housekeeping genes in total and denote the set of housekeeping genes as H. Given the true value of c, the p-values of housekeeping genes are supposed to follow a uniform distribution on (0, 1). Therefore given the significance level α, the false discovery rate of those genes is supposed to be around the nominal level if c is correctly specified. In other words, we find the optimal value of c, denoted as $\hat{c}$, by minimizing the following objective function

$$\left|\frac{1}{m}\sum_{g \in H} I(p_g(c) < \alpha \mid H_0, c) - \alpha\right|. \tag{7}$$

Note that theoretically $\hat{c}$ does not depend on the chosen α. In practice, we observe that $\hat{c}$ is almost the same for varying α. To reduce the arbitrariness of choosing α, we set the final value of $\hat{c}$ as its mean value when $\alpha = 0.1, 0.2, \ldots, 0.9$.

Simulation studies

In this section, we assess the performance of the proposed method by a number of simulation studies. To evaluate the overall effectiveness of HTN, we also compare it with recent methods, including library size, TMM [10], IMM [11], Bull [13] and NHKS [14]. Both Bull and NHKS employ the information of housekeeping genes to determine the global scaling factor. In simulation studies, we have no prior biological information of the genes. Chen et al. [14] suggested to use a statistic called coefficient of variation, which measures the stability of gene expression, to select the most stable housekeeping genes. As in their studies [13, 14], we select a single and 15 housekeeping genes for Bull and NHKS, respectively. Note that for the sample with a single point, we can not calculate its coefficient of variation statistic. Therefore, we will not compare the methods with Bull and NHKS in such case.

In the simulation studies, we generate a synthetic data according to the method described in Robinson et al. [10]. We set different values for the number of genes expressed uniquely to each sample, the proportion, the magnitude and the direction of DE genes between samples under two conditions. We randomly draw data from a given empirical distribution of real counts. We set the expectation of Poison distribution from the sampled read counts by dividing the sum S_k and multiplying a specified library size N_k. With the given mean, we randomly draw data from the corresponding Poisson distribution. Some DE genes are inserted in the data, therefore, we use different statistics to rank the genes and calculate the number of false discoveries [15, 16] for each ranking. In this simulation, we consider two cases: no-repeat sample for each condition and repeat samples for each condition. In each case, we have 500 housekeeping genes by default. We replicate the simulation studies for 100 times and report the average performance of those normalization methods.

In Study 1, we simulate data from only one sample for each condition. We consider two conditions and each is at a rate of 0.1 and 0.5 DE genes at a 1.5-fold level, respectively, and 90% of DE genes are higher in the second condition. In both conditions, let the expression of 10% of genes equal to zero in the first sample and the expression of the corresponding genes in the second sample not equal to zero. Fig 1 shows the scaling factor for each p-value cutoff in the simulation, which demonstrates that the scaling factor is stable for any p-value cutoff. Fig 2 shows M versus A plots for different rates of DE genes, and the scales of the HTN normalization and the TMM normalization. From the left panel of Fig 2, that two scaling factors of normalization are very close for 10% differential expression of total genes. However, as shown in the right panel of Fig 2, when the rate of differentially expressed genes increases to 50%, the red line (HTN) is much closer to the center of non-DE genes than the blue line (TMM). It suggests that HTN gives a more accurate estimate of the normalization factor in this substantive asymmetric setting, that is, for large DE rate.

For no-repeat sample in Study 1, we compare the false discovery rate (FDR) of all normalization methods with different numbers of selected genes. The FDR curves are shown in Fig 3 for DE genes at the rates of 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6, given 1.5-fold level, respectively. From upper panels of Fig 3, we can see that HTN, TMM and IMM have almost the same performance and they are much better than other normalization methods. However, when the rate of DE genes is larger than 0.3, HTN outperforms other methods. Therefore, we can draw the conclusion that HTN performs robustly well for varying rates of DE genes, and has better performance than other methods in the case of a large rate of DE genes.

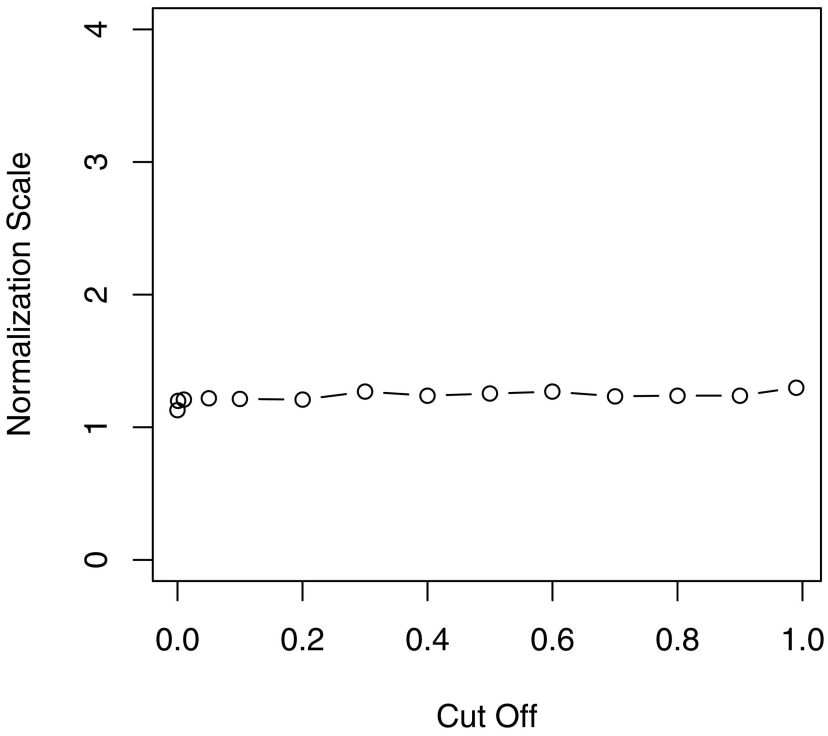

Fig 1. The scaling value for each p-value cutoff in Study 1.

doi:10.1371/journal.pone.0169594.g001

In addition, we further check the robustness of HTN with respect to the signal strength of housekeeping genes. In the above simulation setting, when the rate of DE genes equals 0.4 at a 1.5-fold level, we consider two scenarios: (1) a varying number of housekeeping genes from 50 to 1000; and (2) a varying rate of housekeeping genes that are actually DE genes, which are randomly drawn from all of DE genes. As shown in Fig 4, the numbers of false discovery genes are almost the same in those cases, indicating that HTN is indeed quite robust.

In Study 2, we consider the replicate samples for each condition with different rates of DE genes and compare the proposed method with several popular methods. Here, we consider the performance of the following methods: length-normalized count (Cloonan et al. [17]), Poisson exact test [7] with library size, TMM [10], IMM [11], Bull [13], NHKS [14] and HTN normalization. The essence of virtual length [1] and RPKM [9] are the same as library size normalization and we do not compare them here. Fig 5 shows the false discovery curves of those methods when the genes have different rates of DE genes. The left panel of Fig 5 shows that the FDRs of HTN are similar as those of TMM and IMM with Poisson likelihood ratio statistic or Poisson exact statistic, when the DE rate equals 0.1. However, as the DE rate increases to 0.5, HTN outperforms the alternatives with a lower false discovery rate.

Application to real examples

We apply the proposed HTN method to two real data sets, including several technical replicates of a liver and kidney RNA source [5] (S1 File) and the mouse embryoid bodies versus embryonic stem cells dataset [17], and compare it with other methods. We download human housekeeping genes from [18](S2 File), which is described in [19], and then use the biomaRt package [20] in Bioconductor [21] to match them to the Ensembl gene identifiers. Robinson et al. [10] has also analyzed those real data. For the first real application, Chen et al. [14]

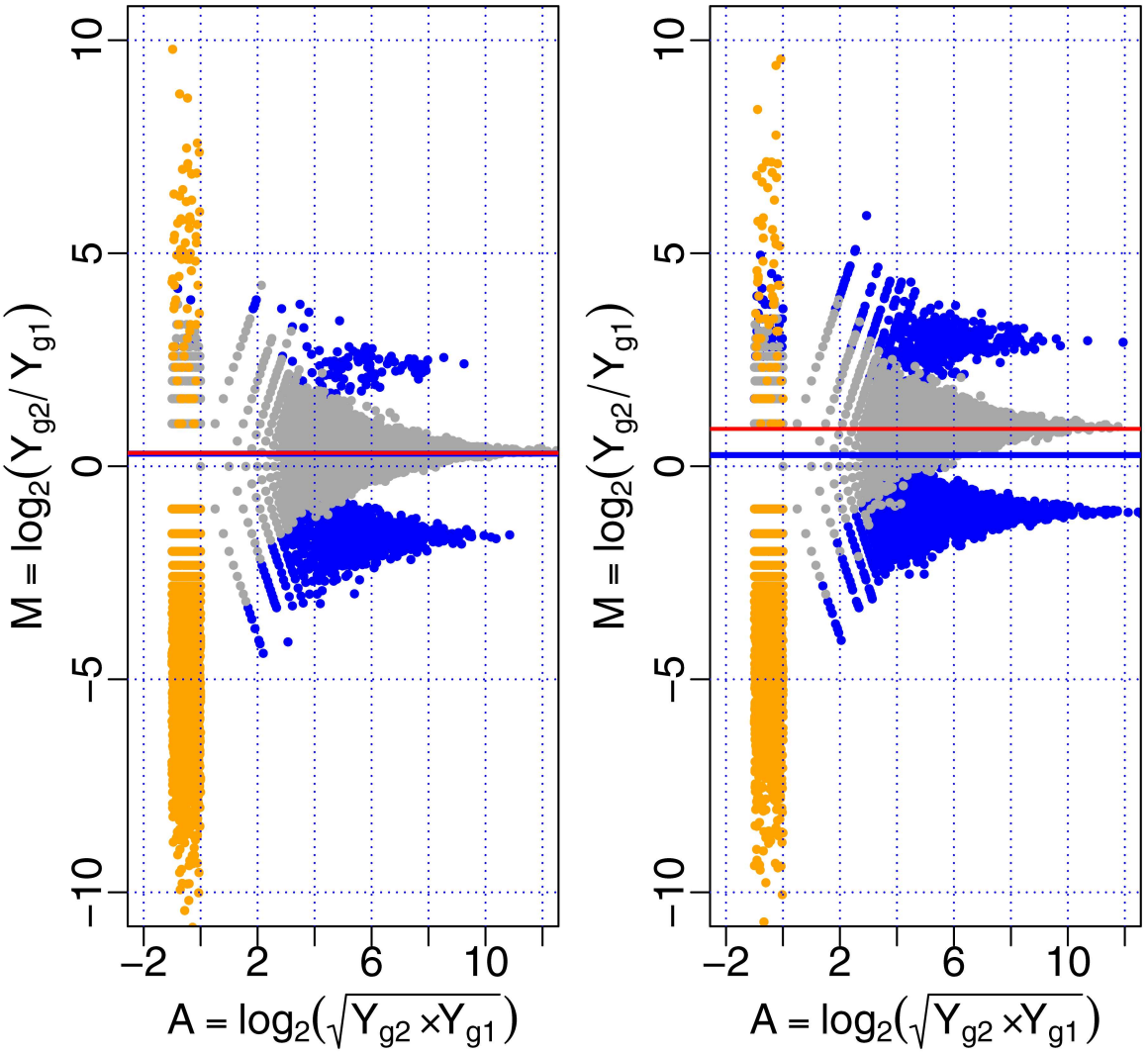

Fig 2. M versus A plots of different rates of DE genes. The left panel and right panel are the MA plots for DE genes at a rate of 0.1 and 0.5, respectively. The blue line is the scale of TMM normalization and the red line is the scale of HTN normalization.

doi:10.1371/journal.pone.0169594.g002

specified 15 housekeeping genes for the liver and kidney dataset normalization in their study. Those genes are also found in the above housekeeping genes, thus we directly use them for Bull and NHKS in this example. For the second real application, given that there is no replicate data in each condition, we will not compare the methods with Bull and NHKS.

We use the exact Poisson statistic to obtain $p-values$ by testing two different conditions and regard the genes as differentially expressed between liver and kidney if their $p-value$ is smaller than 0.0001. Table 1 shows the number of DE genes reported by different normalization methods. From Table 1, we can see that HTN detects 8083 DE genes, 46% of which are significantly higher in liver. The total DE genes and the ratio of DE genes significantly higher in liver (or kidney) by using HTN are similar to those of TMM and IMM. Note that the library size normalization method and NHKS report a much larger number of DE genes that are significantly higher in kidney, while Bull reports more significant genes in liver, and this leads to a larger number of total DE genes in their results. For housekeeping genes, there are 330 DE genes reported by HTN, which is also similar to the results of TMM (329) and IMM (329).

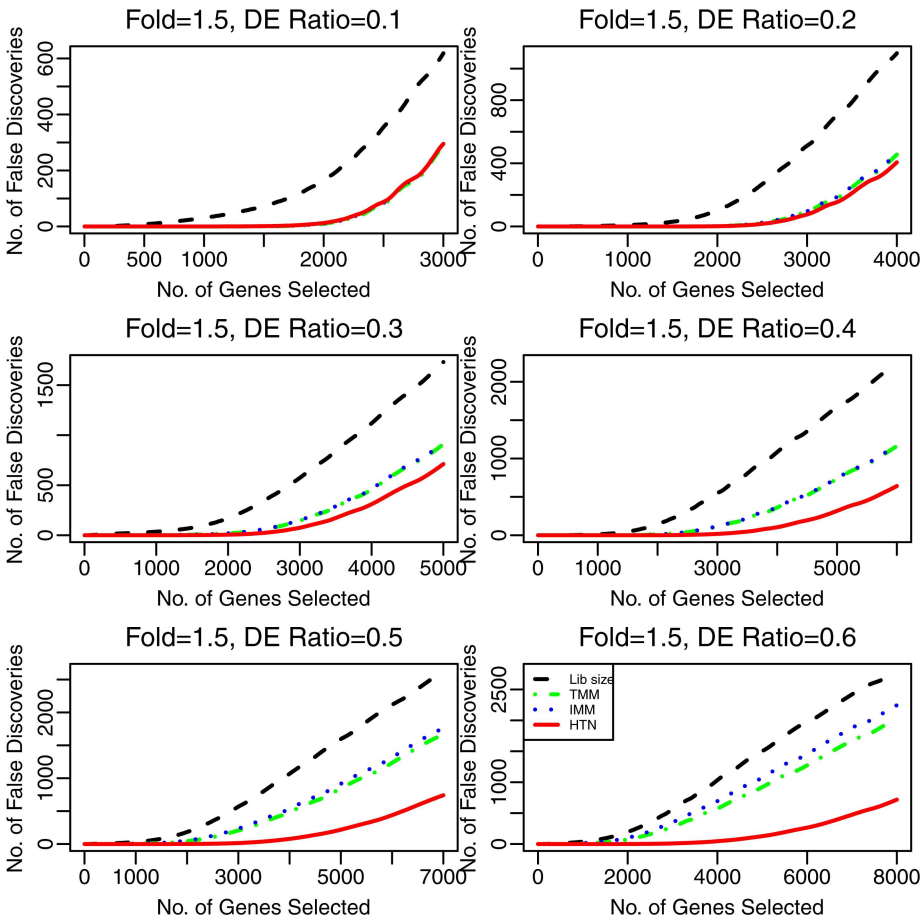

Fig 3. The panels are the false discovery number of test for DE genes at the rates of 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6, respectively.

doi:10.1371/journal.pone.0169594.g003

However, there are more DE genes out of 538 housekeeping genes for the other three normalization methods, which suggests a much larger number of false positive than that of HTN.

The second dataset is comparing mouse embryoid bodies versus embryonic stem cells, which is downloaded from [17], sequenced on the SOLiD system. In this dataset, there are 19005 genes in total, 495 of which are "housekeeping" genes as we know [22]. We get $p - value$ for each gene by using the amended sage.test function [23]. Table 2 shows the results of DE genes output by different normalization methods. The number of DE genes significantly higher in EB is about 22.5%, which is much lower than that of ES (77.5%) by using the HTN normalization. There are 362 DE genes out of 495 housekeeping genes reported by HTN, which is much lower than that of library size normalization (411), TMM (397), IMM (402). Thus, based on the available knowledge of housekeeping genes, HTN tends to have a lower false discovery rate than alternatives in this case. However, as we note, HPN reports 9896 DE genes in total, which is much more than that of library size normalization (9295), TMM (9328) and IMM (9383).

Conclusion

In order to compare the genes expression and thus to detect differently expressed genes between samples, normalization is a crucial step for downstream analysis. In this paper,

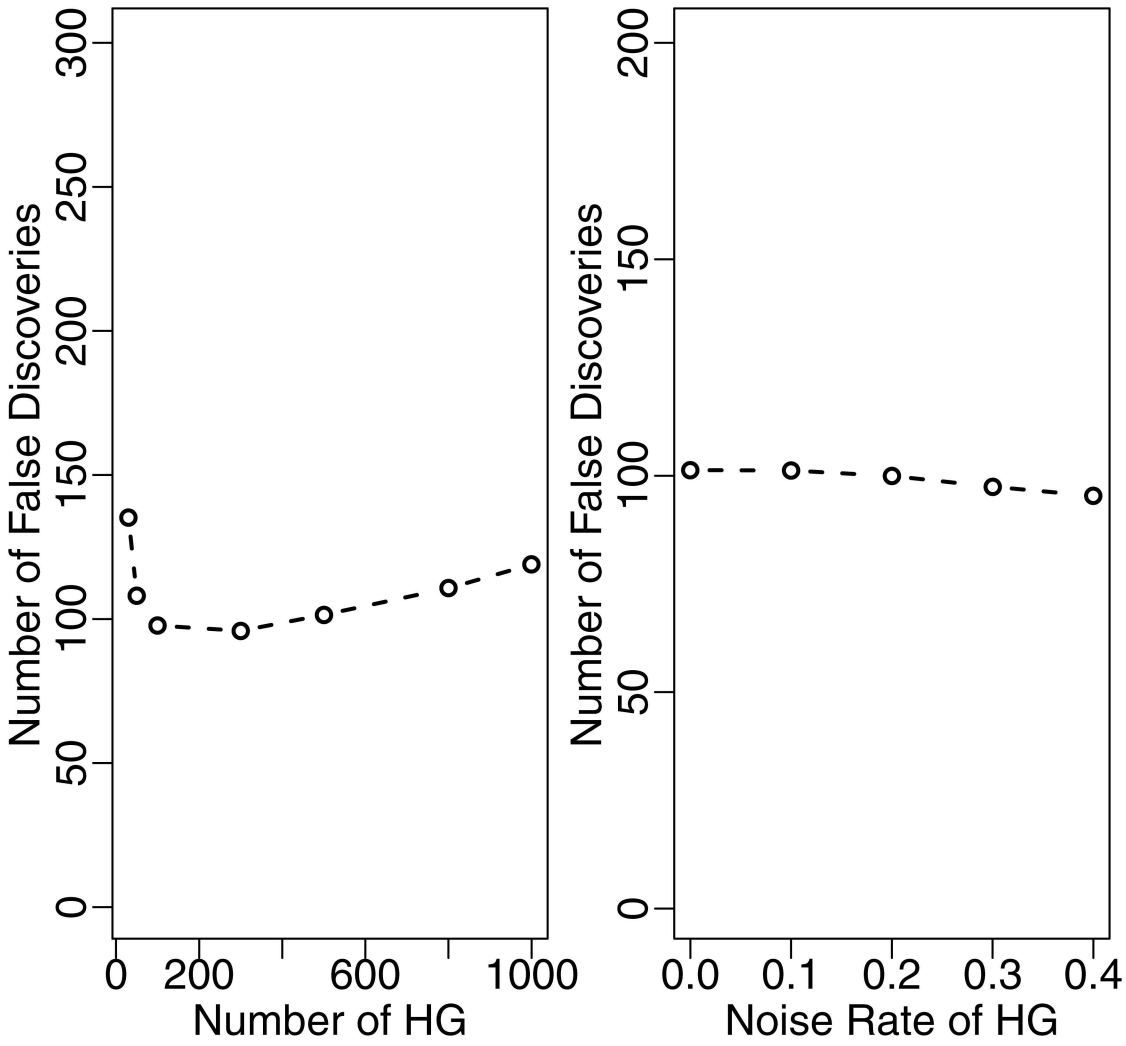

Fig 4. False discovery number for different numbers of housekeeping genes and different rates of noise in housekeeping genes with HTN method.

doi:10.1371/journal.pone.0169594.g004

assuming the information of housekeeping genes is known, we propose a novel normalization method called HTN, which is based on a hypothesis testing, and show it is more effective and robust for normalizing the RNA-seq depth between different samples. The estimated scaling factors between samples can be incorporated into currently used statistical test methods for differential gene expression analysis. The knowledge of housekeeping genes is essential for using our method. To obtain housekeeping genes, users may check the relevant published studies, such as [14, 19].

In the simulation studies, we assess the performance of the proposed method by considering varying ratios of DE genes and varying signal strength of housekeeping genes. We observe that when the ratio is high, the HTN normalization method significantly outperforms the state-of-the-art methods with a lower false discovery rate. The real data analysis also shows that our new method has better performance when judging from the available knowledge of housekeeping genes. Compared with Bull [13] and NHKS [14], which also utilize housekeeping genes, our method seems to be more robust and better, at least as well as, due to that we use all housekeeping genes and the type I error statistic evaluates the overall change of the

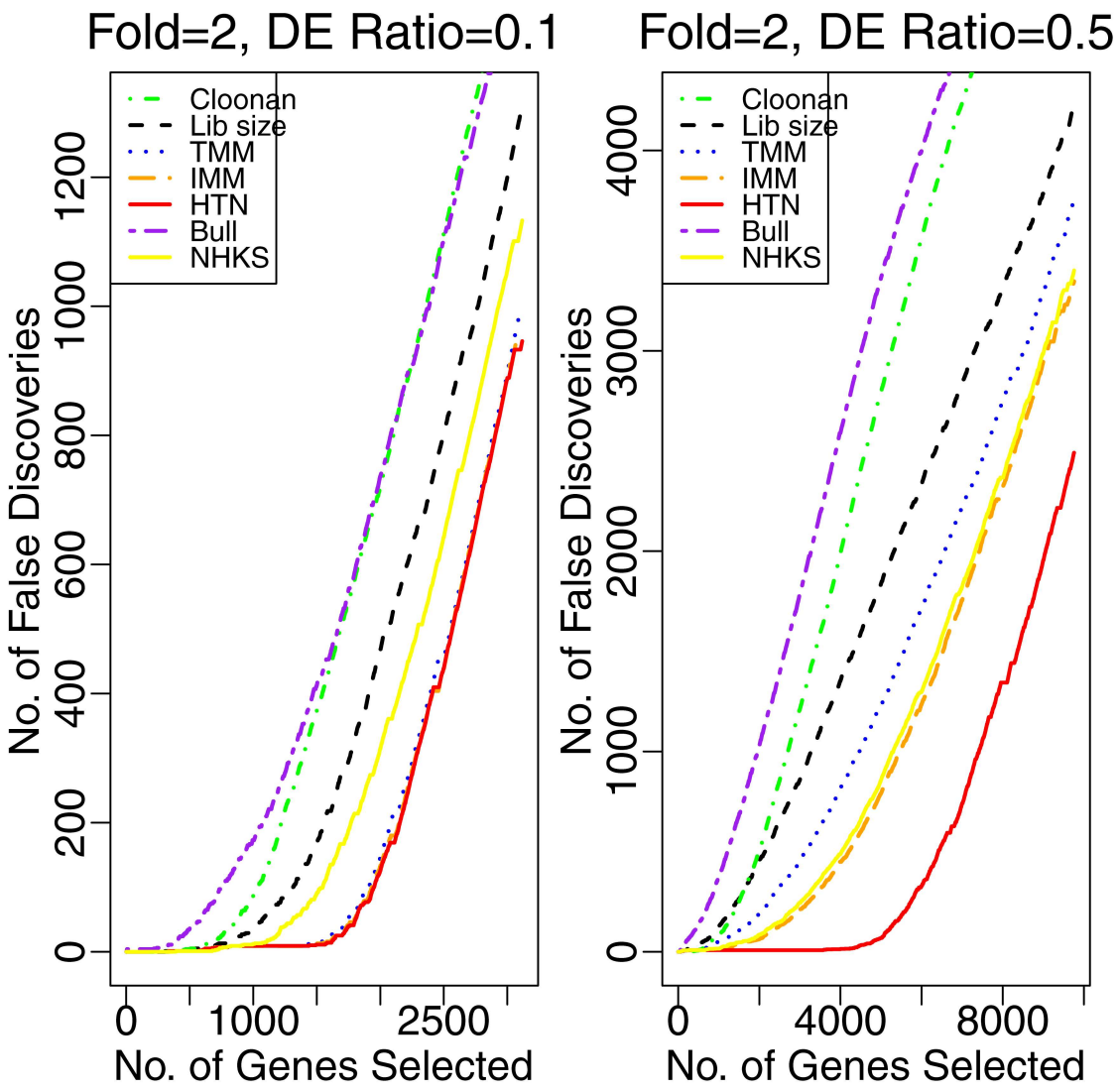

Fig 5. False discovery number for different normalization methods. The left panel and the right panel are the false discovery plots for DE genes at the rates of 0.1 and 0.5, respectively.

doi:10.1371/journal.pone.0169594.g005

Table 1. The number of DE genes between liver and kidney at a cutoff p-value < 10^{-4} for different normalization methods.

	Library size	TMM	IMM	HTN	Bull	NHKS	Overlap
Higher in liver	2082	3759	3797	3680	7248	2836	2082
Higher in kidney	7496	4310	4273	4403	2094	5679	2083
Total	9578	8069	8070	8083	9342	8515	4165
House keeping genes (538)							
Higher in liver	39	120	123	119	287	82	14
Higher in kidney	358	209	206	211	93	276	44
Total	397	329	329	330	380	358	58

doi:10.1371/journal.pone.0169594.t001

Table 2. The number of DE genes between embryoid bodies (EB) and embryonic stem cells (ES) at a cutoff p-value < 10^{-4} for different normalization methods.

	Library size	TMM	IMM	HTN	Overlap
Higher in EB	4441	4156	3941	2227	2189
Higher in ES	4854	5172	5442	7669	4854
Total	9295	9328	9383	9896	7043
House keeping genes (495)					
Higher in EB	279	258	228	104	96
Higher in ES	132	139	174	258	132
Total	411	397	402	362	228

doi:10.1371/journal.pone.0169594.t002

expression of housekeeping genes more efficiently. In conclusion, our empirical studies suggest that the HTN method is a competing alternative for the normalization and differential expression analysis of RNA-seq data.

Supporting Information

S1 File. This is the real data of a liver and kidney RNA source.
(TXT)

S2 File. This is the housekeeping genes of human.
(TXT)

Acknowledgments

The authors thank the Editor and two referees for their constructive comments.

Author Contributions

Formal analysis: YZ GW JZ HL.

Funding acquisition: YZ GW HL.

Methodology: YZ GW JZ HL.

Project administration: YZ.

Software: YZ GW JZ.

Supervision: YZ GW HL.

Writing – original draft: YZ GW JZ HL.

Writing – review & editing: YZ GW JZ HL.

References

1. Wang E.T., Sandberg R., et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008; 456(7221): 470–476. doi: 10.1038/nature07509 PMID: 18978772
2. Sultan M., Schulz M.H., Richard H., et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*. 2008; 321(5891): 956–960. doi: 10.1126/science.1160342 PMID: 18599741
3. Wang X., Sun Q., et al. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS One*. 2008; 3(12): e3839. doi: 10.1371/journal.pone.0003839 PMID: 19052635

4. Bolstad B.M., Irizarry R.A., et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003; 19(2): 185–193. doi: 10.1093/bioinformatics/19.2.185 PMID: 12538238
5. Marioni J.C., Mason C.E., et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008; 18(9): 1509–1517. doi: 10.1101/gr.079558.108 PMID: 18550803
6. Bullard, J.H., Purdom, E.A., et al. Statistical inference in mRNA-Seq: exploratory data analysis and differential expression. *UC Berkeley Division of Biostatistics Working Paper Series*. 2009; paper: 247.
7. Robinson M.D., Smyth G.K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*. 2008; 9(2): 321–332. doi: 10.1093/biostatistics/kxm030 PMID: 17728317
8. AC't Hoen P., Ariyurek Y., et al. Deep sequencing based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res*. 2008; 36(21): e141–e141. doi: 10.1093/nar/gkn705 PMID: 18927111
9. Mortazavi A., Williams B.A., et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008; 5(7): 621–628. doi: 10.1038/nmeth.1226 PMID: 18516045
10. Robinson M.D., Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*. 2010; 11(3): R25. doi: 10.1186/gb-2010-11-3-r25 PMID: 20196867
11. Zhou Y., Lin N., Zhang B. An iteration normalization and test method for differential expression analysis of RNA-seq data. *BioData Mining*. 2014; 7(1): 15. doi: 10.1186/1756-0381-7-15 PMID: 25285156
12. Anders S. and Huber W. Differential expression analysis for sequence count data. *Genome Biology*. 2010; 11(10): R106. doi: 10.1186/gb-2010-11-10-r106 PMID: 20979621
13. Bullard J.H., Purdom E., Hansen K.D., Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*. 2010; 11(1): 94. doi: 10.1186/1471-2105-11-94 PMID: 20167110
14. Chen C.M., Lu Y.L., Sio C.P., Wu G.C., Tzou W.S., Pai T.W. Gene ontology based housekeeping gene selection for RNA-seq normalization. *Methods*. 2014; 67(3):354–363. doi: 10.1016/j.ymeth.2014.01.019 PMID: 24561167
15. Benjamini Y., Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*. 1995; 57(1): 289–300.
16. Storey J.D.The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. *The Annals of Statistics*. 2003; 31(6): 2013–2035. doi: 10.1214/aos/1074290335
17. Cloonan N., Forrest A.R., et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods*. 2008; 5(7): 613–619. doi: 10.1038/nmeth.1223 PMID: 18516046
18. Housekeeping Genes. [http://www.cgen.com/supp_info/Housekeeping_genes.html].
19. Eisenberg E., Levanon E.Y. Human housekeeping genes are compact. *Trends Genet*. 2003; 19(7): 362–365. doi: 10.1016/S0168-9525(03)00140-9 PMID: 12850439
20. Durinck S.M.Y., Kasprzyk A., et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*. 2005; 21(16): 3439–3440. doi: 10.1093/bioinformatics/bti525 PMID: 16082012
21. Gentleman R.C., Carey V.J., et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004; 5(10): R80. doi: 10.1186/gb-2004-5-10-r80 PMID: 15461798
22. Hendrik J.M., De Jonge H.J., et al. Evidence based selection of housekeeping genes. *PLoS One*. 2007; 2(9): e898. doi: 10.1371/journal.pone.0000898 PMID: 17878933
23. CRAN-Package statmod. [http://cran.r-project.org/web/packages/statmod/index.html].