



OPEN

ERgene: Python library for screening endogenous reference genes

Zehua Zeng, Yuzhe Xiong, Wenhuan Guo & Hongwu Du✉

In gene expression analysis, sample differences and experimental operation differences are common, but sometimes, these differences will cause serious errors to the results or even make the results meaningless. Finding suitable internal reference genes efficiently to eliminate errors is a challenge. Aside from the need for high efficiency, there is no package for screening endogenous reference genes available in Python. Here, we introduce ERgene, a Python library for screening endogenous reference genes. It has extremely high computational efficiency and simple operation steps. The principle is based on the inverse process of the internal reference method, and the robust matrix block operation makes the selection of internal reference genes faster than any other method.

Gene expression analysis has become increasingly important in many areas of biological research. The commonly used measurement methods include microarray¹, RT-PCR² and massively parallel sequencing³. However, these measurements also require normalization to reduce the differences between samples. The existing normalization methods include geNorm⁴, Normfinder⁵ and BestKeeper⁶. All three methods start with a limited set of candidate reference genes. Further, geNorm and BestKeeper also calculate a normalization factor. On the one hand, the calculation efficiency of the above methods is not high enough, and the screening process is sometimes complicated. On the other hand, there are currently no available package for screening endogenous reference genes in Python. In order to solve these problems, a new approach is proposed by analyzing the principle of the internal reference method. Using the computational power of the Pandas library in Python, we build a Python library to meet the requirements of normalization and internal reference gene screening.

Results

Screening effect of laboratory gene expression data. We took some tissues from the same location in the brains of two aging mice injected with SHED (Stem cells from human exfoliated deciduous teeth) and two aging mice injected with salt. And gene expression analysis was performed on these tissues to obtain test data. The test data can be found on Github (<https://github.com/Starlitnightly/ERgene/tree/master/example>). First, we analyzed the difference of the sample spectral density in the test data and made a box and a density diagram (Fig. 1a,b). The difference of the sample spectral density generally refers to differences in the spectral density of all gene expressions between each sample, such as the differences between individual mice, or the differences between experiments. If the difference of the sample spectral density is too large, the subsequent analysis will be meaningless. To avoid this, researchers typically use internal reference genes to normalize the data. Therefore, it is very important to look for stable internal reference genes. In the test data, we used the ERgene.FindERG method. Then the candidate internal reference gene *Atp1a3* was found. Using the gene *Atp1a3*, we normalized the test data and made the boxplot in the same way (Fig. 1c,d). By comparing Figs. 1a,c and Fig. 1b,d, we can see that the difference of the sample spectral density has been significantly reduced. Therefore, ERgene has a significant effect on reduce the difference of the sample spectral density when it comes to processing raw lab data.

Screening effect of public datasets. After obtaining good results from the raw lab test data, we selected a dataset in the GEO database that had not been well normalized for verification. The dataset selected was the mouse dataset GSE4786 of Someya⁷. We analyzed the difference of the sample spectral in the experimental data and made a box diagram and a density diagram (Fig. 2a,b). Then, we used the ERgene.FindERG method and found the internal reference gene *1439423_x_at*. Using the gene *1439423_x_at*, we normalized the data and plotted the boxplot in the same way (Fig. 2c,d). By comparing Figs. 2a,c and Fig. 2b,d, we can see that the difference of the sample spectral has been narrowed. The reason for choosing this dataset is that most of the data on GEO

112 Lab, School of Chemistry and Biological Engineering, University of Science and Technology Beijing, Beijing 100083, China. ✉email: hongwudu@ustb.edu.cn

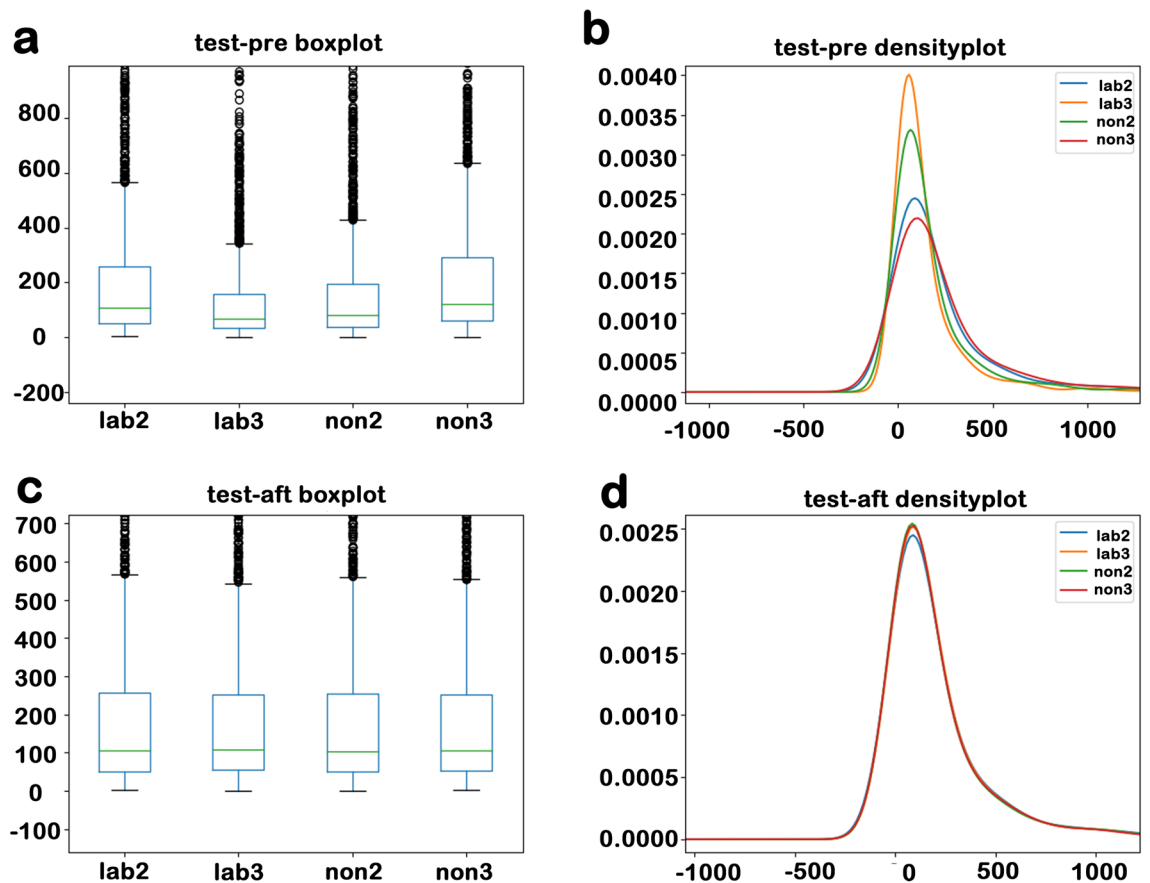


Figure 1. (a) The boxplot of test data before processing. (b) The density plot of the test data before processing. (c) The boxplot of test data after processing. (d) The density plot of the test data after processing. The boxplot's abscissa is the sample, the ordinate is the gene expression, the green line is the median, the blue box line is the quartile, and the black point is the outlier. The density plot's abscissa is the length of the data, the ordinate is the data density, and the lines of different colors represent different samples.

have been normalized. It is not surprising that ERgene had achieved this effect for a dataset that was not so well normalized.

Further confirmed the screening results by literature. In the study of Horison, the main purpose was to investigate the related internal reference genes for aortic lesions associated with bicuspid valve⁸. A total of 12 reference genes ATP5B, ACTB, B2M, CYC1, EIF4A2, GAPDH, SDHA, RPL13A, TOP1, UBC, YWHAZ, and 18S were detected. In his report, geNorm was used to test these 12 genes to determine the most stable single internal reference gene. We used ERgene to analyze the author's raw data. Twenty possible reference genes were obtained (CYC16, CYC11, TOP16, CYC4, TOP11, CYC2, CYC10, CYC7, CYC12, CYC1, CYC9, EIF4A17, CYC8, TOP4, TOP10, TOP2, CYC5, TOP9, CYC3 and TOP7). All the genes that were found can be divided into three families: CYC, TOP, and EIF4A. Only one member in each family should be used for normalization to not bias the results because family members may be coregulated. By comparing these three families with the 12 reference genes obtained in the study, we can see that CYC1 and TOP1 coincide, consistent with the report.

McLoughlin et al. selected a Real-Time PCR Housekeeping Gene Panel in Human Endothelial Colony Forming Cells⁹. A total of 28 candidate internal reference genes were screened out by geNorm (RPL37, RPS29, RPL9, VIM, NDUFB3, ATP51, RPL31, RPS27, CTGF, NDUPB4, ATP5J, RPS6, ACTB, ATP5F1, RPL27A, PGAM4, RPS10, RPL30, HSPA8, RPL13, RPL19, NDUFB8, ATP5L, UBC, VWHAC, PRDX1, GAPDH and B2M). And six stable reference genes (RPL13, RPL31, RPL37, RPL30, RPS6, and RPL19) were verified by experiments. We used ERgene to analyze the authors' original dataset GSE125792. When the depth was set to 2, we obtained twenty candidate reference genes (RPL19, RPS4X, RPL13, TUBA1A, SF3B5, RPS3a, RPL9, RPL39L, LDHA, RPS8, RPL31, FTL, RPS3, RPL22, PINLYP, CAPG, UQCRH, RPS5, RPSAP58, and RPL36). When the depth was set to 3, seven candidate reference genes (RPL13, FTL, RPS3a, SF3B5, RPS4X, TUBA1A, and RPL19) were obtained. RPL13, the most stable reference gene determined by experiments, ranks high in our algorithm results.

The overlap rate of genes in sample pairs. ERgene screened 20 candidate internal reference genes from the two groups of samples in the dataset, and the computational depth represented the number of samples selected. When the computational depth reached 3, the 20 candidates were selected from the three pairs of sam-

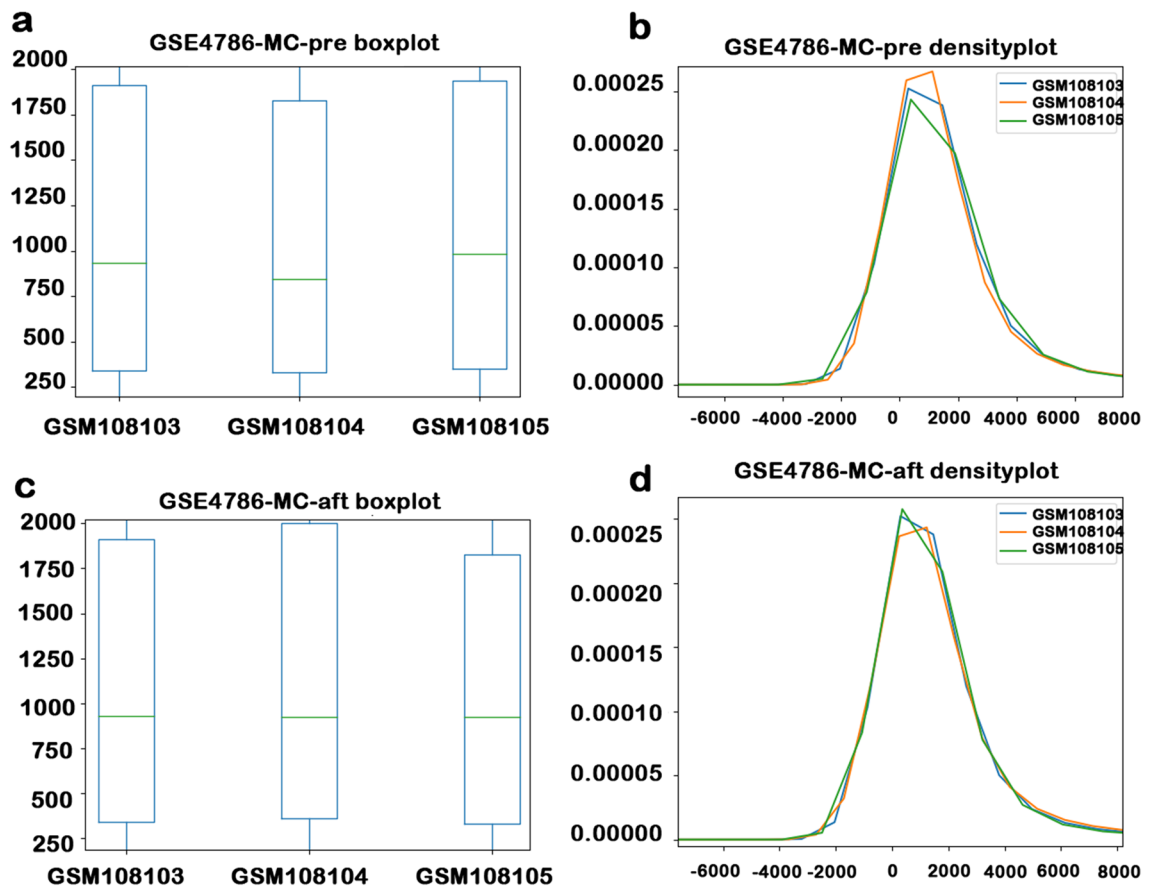


Figure 2. (a) The boxplot of GSE4786-MC before processing. (b) The density plot of the GSE4786-MC before processing. (c) The boxplot of GSE4786-MC after processing. (d) The density plot of the GSE4786-MC after processing. The boxplot's abscissa is the sample, the ordinate is the gene expression, the green line is the median, the blue box line is the quartile, and the black point is the outlier. The density plot's abscissa is the length of the data, the ordinate is the data density, and the lines of different colors represent different samples.

ples (1, 2) (1, 3) and (2, 3), and then take the intersection. Here, we selected 12 samples of the dataset GSE125792 used by McLoughlin et al.⁹, calculated the 20 internal reference genes screened by 66 pairs of samples, and then showed the overlap of samples through the form of Upsetplot (Fig. 3). There are 10 genes in Fig. 3 with sample coverage of more than 80%. The result means that when we increase the computational depth, the duplication rate of the selected candidate internal reference gene is still more than 50%. Ten genes are included in RPL13 (probe ID: ASHG40056316; platform: GPL21827). RPL13 was experimentally confirmed as a stable candidate internal reference gene⁹.

Discussion

ERgene makes up for the fact that python library do not have a right method for screening reference genes, and geNorm is embedded in qbase + or Excel 2003. Normfinder is a source for the R language or Excel 2003. The application of these methods is troublesome and not particularly friendly to Python users. On the Python platform, the user only needs to enter three sentences to start filtering the internal reference gene. It seems extremely simple and friendly. The computational efficiency of ERgene is increased by nearly 90% higher than that of Normfinder, which also uses all genes for reference genes (Table 1). Besides, the internal reference genes found by ERgene, NormFinder and geNorm were similar (Table 2).

Although ERgene may not be new in principle, the calculation uses a new formula, which leads to a significant improvement in computing time over that of the complex matrix operations of geNorm and NormFinder. ERgene using each gene as a normalizer, calculates the ratio of each gene pair as done in the geNorm method (formula Eq. (2)). Also, the sigma squared value is equivalent to that of geNorm (formula Eq. (3)). NormFinder does not use candidate reference genes, but uses all genes to search for internal reference genes, thus, candidate instability can be avoided to a certain extent. The total number of genes tested was 1968. When the computational depth was set to 2 (screening internal reference genes with two samples) it only took the 4.58 s to obtain the possible internal reference genes because there is no complicated exponentiation. When the computational depth is set to 3 or larger, the efficiency begins to decline. When the computational depth set too large, there may be no result. Because when the computational depth is 3 or larger, the screening will select the internal reference genes from two different sample combinations, and then take the intersection from the screening results for all combinations.

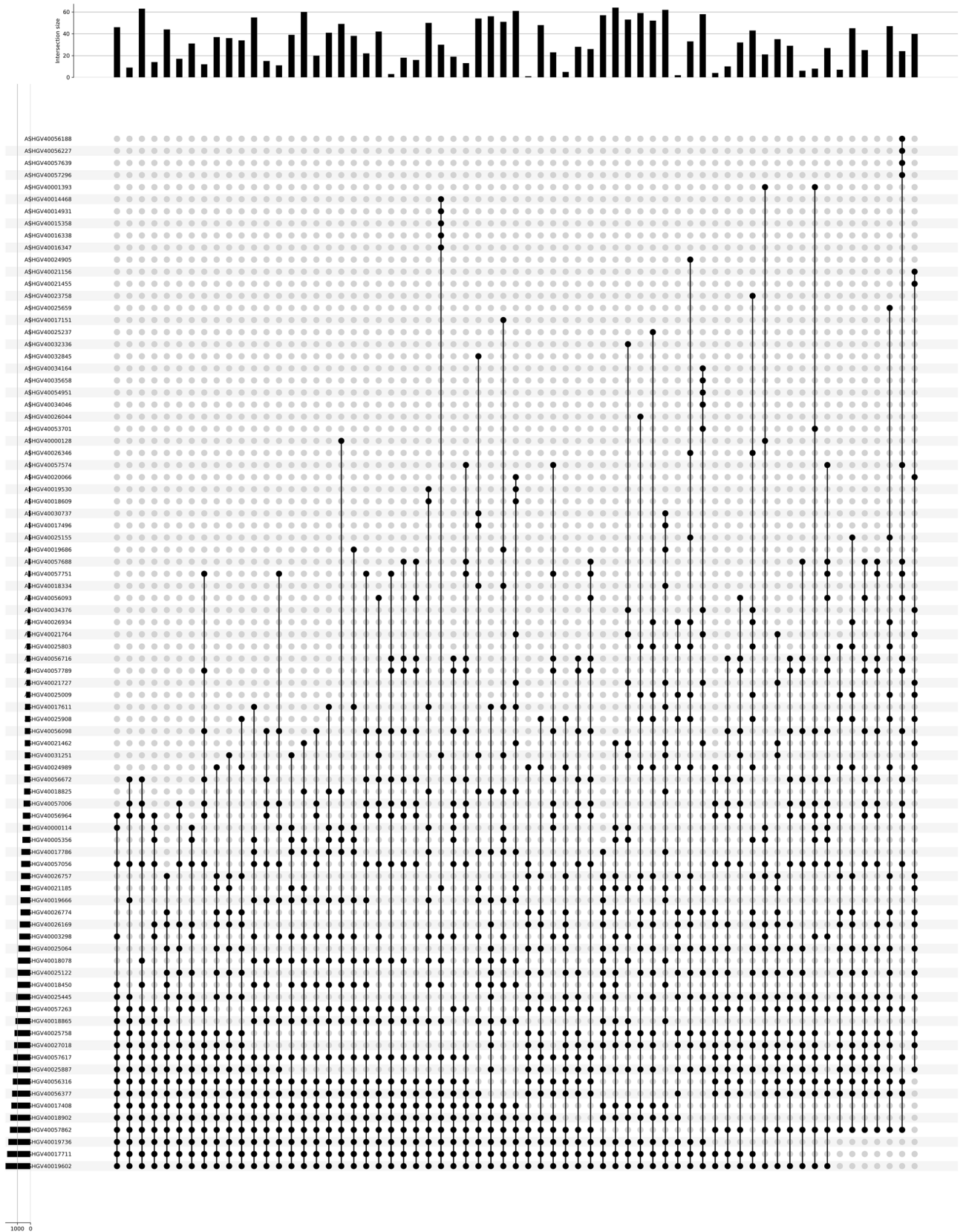


Figure 3. The upsetplot of 66 sample pairs overlap (GSE125792, 12 samples). The abscissa represents the sample pair, and the ordinate represents the appearance of the candidate internal reference genes. The height of the column in the upper bar chart represents the number of the sample pairs. In the upper bar chart, the height of the column represents the ordinal number of the sample pair. The higher the ordinal number, the higher the column.

The Number of genes	2 samples		3 samples		4 samples	
	Normfinder	ERgene	Normfinder	ERgene	Normfinder	ERgene
100	0.1 s	0.1 s	0.5 s	0.66 s	1 s	1.37 s
500	35 s	0.48 s	42 s	1.35 s	44 s	2.67 s
1000	6 min	1.05 s	6 min 11 s	3.20 s	6 min 40 s	10.89 s
2000	55 min	4.58 s	55 min	10.95 s	56 min	28.47 s

Table 1. Normfinder versus ERgene in computational time.

	Normfinder & geNorm	ERgene
Test dataset (No probe conversion was performed)	P47754, Q68FG2 , A0A0G2JDX4, Q9WVA2, P26443, Q8BU30, Q8R1Q8, Q8CBG6, Q8OUW2, A0A494BAX5, Q6PIC6	Q6PIC6 , A3KGU7, Q03265, P63260, O08553, Q62261, P63101, Q68FG2 , P05064, P17182
Horrison dataset	ATP5B, ACTB, B2M, CYCI , EIF4A2, GAPDH, SDHA, RPL13A, <i>TOP1</i> , UBC, YWHAZ	<i>CYC16</i> , <i>CYC11</i> , <i>TOP16</i> , <i>CYCA</i> , <i>TOP11</i> , <i>CYC2</i> , <i>CYC10</i> , <i>CYC7</i> , <i>CYC12</i> , CYCI , <i>CYC9</i> , EIF4A17, <i>CYC8</i> , <i>TOP4</i> , <i>TOP10</i> , <i>TOP2</i> , <i>CYC5</i> , <i>TOP9</i> , <i>CYC3</i> , <i>TOP7</i>
McLoughlin dataset	<i>RPL37</i> , <i>RPS29</i> , RPL9 , VIM, NDUFB3, ATP51, RPL31 , <i>RPS27</i> , CTGF, NDUFB4, ATP5J, <i>RPS6</i> , ACTB, ATP5FI, <i>RPL27A</i> , PGAM4, RPSIO, <i>RPL30</i> , HSPA8, RPL13 , RPL19 , NDUFB8, ATP5L, UBC, VWHAC, PRDX1, GAPDH, B2M	RPL19 , <i>RPS4X</i> , RPL13 , TUBA1A, SF3B5, <i>RPS3a</i> , RPL9 , <i>RPL39L</i> , LDHA, <i>RPS8</i> , RPL31 , FTL, <i>RPS3</i> , <i>RPL22</i> , PINLYP, CAPG, UQCRH, <i>RPS5</i> , <i>RPSAP58</i> , <i>RPL36</i>

Table 2. Comparison of internal reference genes found (the genes in bold are identical; the genes in italics are in the same family) (The test data are not converted by a probe).

The geNorm algorithm has the unique advantage of identifying the most stable reference gene from a tested set of candidate reference genes in each sample. Bestkeeper calculates all kinds of unique Bestkeeper indexes based on the genes of the housekeeper. The amount of calculation is larger than that of geNorm, but the accuracy is improved. Both algorithms require researchers to provide genes in advance, and their applications are limited. However, ERgene directly searches and analyzes all genes according to existing samples, without the need for candidate reference genes; thus its application scope is greatly improved.

Normfinder constructed a mathematical model. It first synthesized a stable value for screening based on the intra-group variation and inter-group variation of all genes, which were improved over those of geNorm and Bestkeeper. This algorithm was excellent. ERgene used the expression multiple during internal standard normalization in the calculation of intra-group variation. And the expression multiple referred to the expression multiple of a gene relative to the internal reference gene. For example, in sample 1, if gene1 was the internal reference, gene2 should be about three times as much as gene1 in sample 2. Then the internal reference gene was screened out by the magnitude of expression multiple changes between different groups. The principle of internal reference was more consistent in ERgene than in Normfinder.

ERgene also provides a processing method for internal reference data, which is not the optimal internal reference processing method but only uses a single gene provided by ERgene. FindERG calculate the normalization factor, and the verification effect is better than other exist methods for the same experimental group. According to the MIQE guidelines¹⁰, it is not acceptable to normalize a single internal reference gene unless the investigator provides clear evidence to the reviewer to confirm its invariable expression under the above experimental conditions. Several studies⁴ have demonstrated the problem of using a single reference gene and recommend using at least two stably expressed reference genes. And the clustering algorithm for normalizing computing factors is most incisive in geNorm, so it is a good choice to use ERgene to screen out internal reference genes and then use geNorm or Normfinder to calculate the normalized factors for normalization processing.

Features and methods

Algorithms and mathematical descriptions. *Principle.* Based on internal standard method, the ratio of each gene expression quantity to the other gene expression quantity was calculated as a relative correction factor, and the calculated results were presented in the form of matrix. The ERgene algorithm inverts this process. Sample 1 calculates the relative correction factor F of each gene, and sample 2 repeats the process. The differences between the results of the two samples were compared to obtain the range ΔF of the relative correction factor between each gene. σ^2 was calculated for the range of variation of each gene, then sort the results of variation from smallest to largest. The program will return the top 20 genes as a result.

When the depth is greater than two samples, for example, a depth of three samples will be selected according to the combined counting method. Three samples (1, 2), (1, 3) and (2, 3) will be selected to obtain the internal reference genes, and then the intersection will be obtained.

Optimization. When the gene dataset is too large, block calculation is adopted. Every 1000 genes are taken as a block. When all the blocks have been computed, the results are combined and sorted by sorting them from smallest to largest. The program will return the top 20 genes as a result.

Mathematical description. Let sample 1 be $x_1 = [A_1, A_2, A_3, \dots, A_n]^T$, where A_i is the expression of the i -th gene in sample 1. The relative correction factor matrix F_1 is

$$F_1 = [x_1 \div A_1, x_1 \div A_2, x_1 \div A_3, \dots, x_1 \div A_n] \\ = \begin{bmatrix} \frac{A_1}{A_1} & \dots & \frac{A_n}{A_1} \\ \vdots & \ddots & \vdots \\ \frac{A_n}{A_1} & \dots & \frac{A_n}{A_n} \end{bmatrix} \quad (1)$$

Similarly, let sample 2 be $x_2 = [B_1, B_2, B_3, \dots, B_n]^T$, where B_i is the expression of the i -th gene in sample 2. The relative correction factor matrix F_2 is

$$F_2 = \begin{bmatrix} \frac{B_1}{B_1} & \dots & \frac{B_n}{B_1} \\ \vdots & \ddots & \vdots \\ \frac{B_n}{B_1} & \dots & \frac{B_n}{B_n} \end{bmatrix} \quad (2)$$

The relative factor change amplitude matrix ΔF is

$$\Delta F = F_1 - F_2 = \begin{bmatrix} \frac{A_1}{A_1} - \frac{B_1}{B_1} & \dots & \frac{A_n}{A_1} - \frac{B_n}{B_1} \\ \vdots & \ddots & \vdots \\ \frac{A_n}{A_1} - \frac{B_n}{B_1} & \dots & \frac{A_n}{A_n} - \frac{B_n}{B_n} \end{bmatrix} \quad (3)$$

The variance vector σ^2 of the magnitude of change in relative factors for each gene is

$$\sigma^2 = \left[\frac{\sum (\Delta F_{1i} - \frac{\sum(\Delta F_{1i})}{n})^2}{n} \quad \dots \quad \frac{\sum (\Delta F_{ni} - \frac{\sum(\Delta F_{ni})}{n})^2}{n} \right]^T \quad (4)$$

Function description. *ERgene.FindERG(data, depth)*. This function is used to screen internal reference genes. The parameter data are in DataFrame format, where the first column is the gene ID, and the other column is the expression level of each gene in the sample. The depth of the parameter refers to the number of samples to be selected for internal reference genes. For example, a depth of 2 is used for screening samples 1 and 2. A depth of 3 means sample 1 and sample 2, sample 1 and sample 3, and sample 2 and sample 3 are screened separately, and then the intersection is removed. The greater the depth, the fewer the number of internal reference genes screened, which cannot be fewer than 2 or more than the number of samples. And users can compare the results at different depths. The speed of calculation depends on the depth of calculation. Genes in calculation results may come from the same family, and only one family member should be used in normalization. When the calculation depth is larger than 2, an Upsetplot will be generated to show the overlap rate of the candidate internal reference genes generated by each pair of samples (Fig. 3).

ERgene.normalizationdata(data, ERGname). This function is used for the standardization of a single internal reference gene. The parameter data are in DataFrame format, where the first column is the gene ID, and the second column is the expression level of each gene in the sample. The parameter ERGname is the name of the internal reference gene to be processed. The computation speed is accelerated by using the multi-threaded matrix operation of Pandas, making the computation speed faster.

Data availability

Raw test data are available at <https://github.com/Starlitnightly/ERgene/tree/master/example>. The GEO datasets analyzed during the current study are available in the Gene Expression Omnibus repository, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4786>, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE125792>.

Code availability

Source code is available for academic non-commercial research purposes. Links to code and documentation are provided at <https://github.com/Starlitnightly/ERgene>.

Received: 29 May 2020; Accepted: 14 September 2020

Published online: 29 October 2020

References

- Schena, M., Shalon, D., Davis, R. & Brown, P. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N. Y.)* **270**, 467–470 (1995).
- Fink, L. *et al.* Real-time quantitative RT-PCR after laser-assisted cell picking. *Nat. Med.* **4**, 1329–1333 (1998).
- Rogers, Y. & Venter, J. C. Massively parallel sequencing. *Nature* **437**, 326–327 (2005).
- Vandesompele, J. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, h31–h34 (2002).
- Andersen, C. L., Jensen, J. L. & Ørntoft, T. F. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res.* **64**, 5245–5250 (2004).

6. Paul, M. W., Tichopad, A., Prgomet, C. & Neuvians, T. P. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: bestkeeper-excel-based tool using pair-wise correlations. *Biotechnol. Lett.* **26**, 509–515 (2004).
7. Someya, S., Yamasoba, T., Weindruch, R., Prolla, T. A. & Tanokura, M. Caloric restriction suppresses apoptotic cell death in the mammalian cochlea and leads to prevention of presbycusis. *Neurobiol. Aging* **28**, 1613–1622 (2007).
8. Harrison, O. J., Moorjani, N., Torrens, C., Ohri, S. K. & Cagampang, F. R. Endogenous reference genes for gene expression studies on bicuspid aortic valve associated aortopathy in humans. *PLoS ONE* **11**, e164329 (2016).
9. McLoughlin, K. J., Pedrini, E., MacMahon, M., Guduric-Fuchs, J. & Medina, R. J. Selection of a real-time PCR housekeeping gene panel in human endothelial colony forming cells for cellular senescence studies. *Front. Med.* **6**, 33 (2019).
10. Bustin, S. A. *et al.* *The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments* (Oxford University Press, Oxford, 2009).

Acknowledgements

We thank James Bruner for his proofread and native processing of this manuscript. This work was supported by Hebei Provincial Department of Science and Technology (No.19942410G).

Author contributions

Z.Z. conceived and designed ERgene, implemented it in Python, tested ERgene, analyzed the data, and prepared all figures. W.G. provided the test data by mouse experiment. Z.Z. and H.D. wrote the main manuscript text. Y.X. reviewed and modified the full text. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020