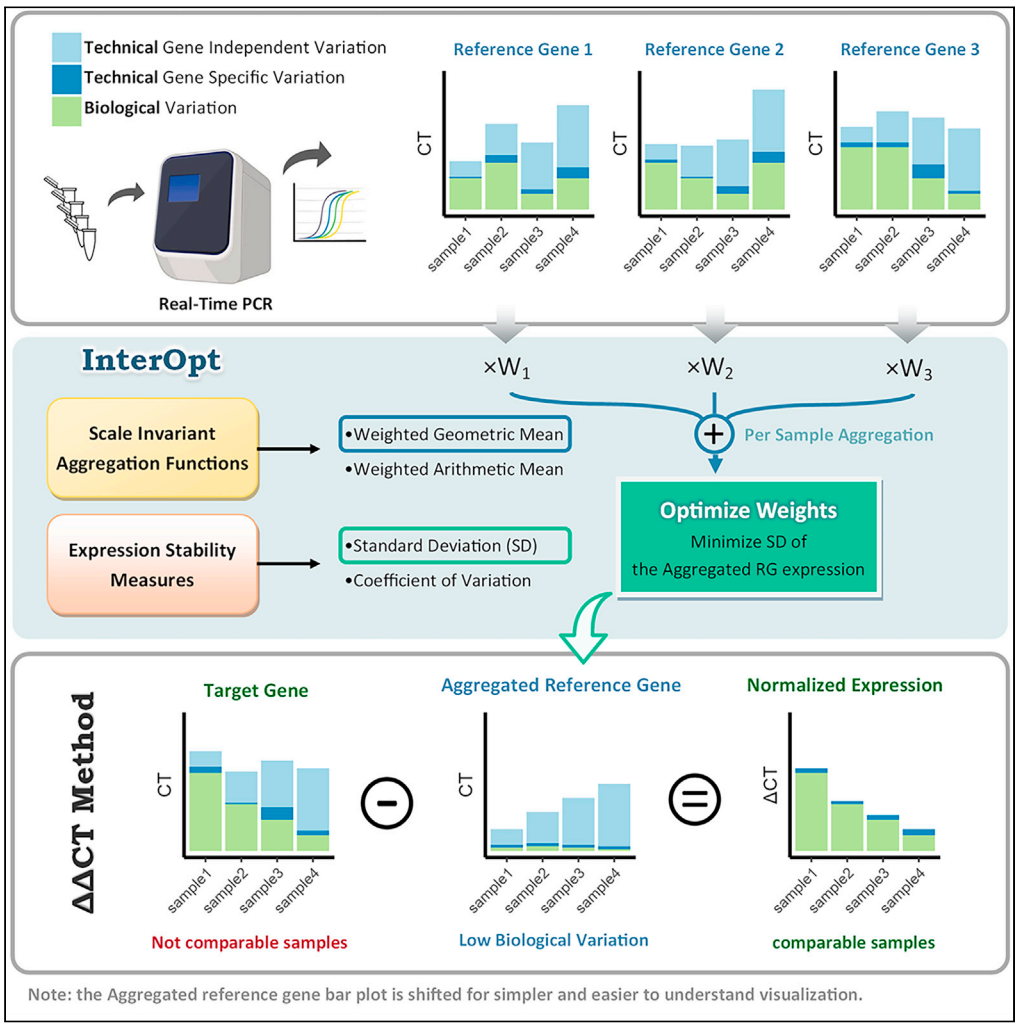
iScience

CellPress
OPEN ACCESS

Article

InterOpt: Improved gene expression quantification in qPCR experiments using weighted aggregation of reference genes



Note: the Aggregated reference gene bar plot is shifted for simpler and easier to understand visualization.

Adel Salimi, Saeid Rahmani, Ali Sharifi-Zarchi

asharifi@sharif.edu

Highlights

Theoretical assumptions behind conventional expression stability measures

InterOpt improves qPCR data normalization by optimizing reference genes aggregation

InterOpt can be easily utilized in combination with usual ΔΔCT method

InterOpt is fast and works in different levels of variability and sample sizes

Salimi et al., iScience 26, 107945
October 20, 2023
https://doi.org/10.1016/j.isci.2023.107945

iScience

CellPress
OPEN ACCESS

Article

InterOpt: Improved gene expression quantification in qPCR experiments using weighted aggregation of reference genes

Adel Salimi,[1,3] Saeid Rahmani,[1,2,3] and Ali Sharifi-Zarchi[1,4,*]

SUMMARY

qPCR is still the gold standard for gene expression quantification. However, its accuracy is highly dependent on the normalization procedure. The conventional method involves using the geometric mean of multiple study-specific reference genes (RGs) expression for cross-sample normalization. While research on selecting stably expressed RGs is extensive, scant literature exists regarding the optimal approach for aggregating multiple RGs into a unified RG. In this paper, we introduce a family of scale-invariant functions as an alternative to the geometric mean aggregation. Our candidate method (weighted geometric mean minimizing standard deviation) demonstrated significantly better results compared to other proposed methods. We provide theoretical and experimental support for this finding using real data from solid tumors and liquid biopsies. Moreover, the closed form and regression-based solution enable efficient computation and straightforward adoption on various platforms. All the proposed methods have been implemented within an easy-to-use R package with graphics processing unit (GPU) acceleration.

INTRODUCTION

Reverse transcription quantitative PCR (RT-qPCR) is one of the most utilized techniques to quantify RNA molecules. Despite high-throughput methods such as RNA sequencing (RNA-seq) are widely used for expression quantification, qPCR is still the primary molecular diagnostic test for clinical and research purposes due to its high specificity and sensitivity, low cost, and reproducibility.[1,2] There are, however, different sources of technical errors that make the accuracy and power of qPCR highly dependent on the normalization procedure.

The qPCR method uses repeated cycles of DNA amplification to measure the expression of target gene(s) in a given sample. The amount of the target region approximately doubles during each amplification cycle. The *Cycle Threshold* (CT) value is defined as the first cycle the amount of amplified target region exceeds a fixed threshold.[3] This raw CT value is affected by two sources of variation: biological and technical.[4] To accurately measure expression alternation of a target gene among different conditions, we need to minimize these variations. For this purpose, cross-sample normalization is performed to make the expression levels of a target gene comparable among different samples, which leads to statistically authentic results. The most usual way of cross-sample normalization employs a *reference gene* (RG) and normalizes the expression of each target gene by subtracting the expression of RG from it. The assumption behind these techniques is that the RG expression is unaffected in different conditions.[5] Choosing an unstable or differentially expressed RG could lead to contradictory results. For example, if an RG has expression alternation between the treatment vs. control groups, a correlated target gene would show lower or no significant expression change after normalization with this RG. An example is provided in Ghanbari et al.[6]

An optimal RG is a gene with minimal biological variation among different samples, which is also highly expressed in all samples. There are housekeeping genes (e.g., Glyceraldehyde 3-phosphate dehydrogenase GAPDH) that are widely used for this purpose. Several studies, however, have shown that those so-called housekeeping genes have high variation of expression in specific tissues or diseases.[7,8] Moreover, some studies question the existence of any housekeeping gene with aforementioned conditions.[9] This challenge is even more apparent in non-coding RNA studies, particularly when using circulating microRNAs (miRNAs) as cancer biomarkers.[10,11] Therefore, using multiple and study-specific RGs is highly recommended in the standard guidelines.[4,12]

Using multiple RGs in a qPCR experiment raises several challenges, including the need to aggregate expression levels of multiple RGs into a single reference value for normalizing the other genes. This single value can be considered as the expression level of a virtual RG. Additionally, we need to measure the expression *stability* of a real or virtual RG which is maximized when that RG shows zero expression variation among different conditions. Although the latter challenge is well studied,[5,13,14] there is a paucity of literature about the former one which is the main focus of this paper.

[1]Computer Engineering Department, Sharif University of Technology, Tehran 11155-1639, Tehran, Iran
[2]School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran 19538-33511, Tehran, Iran
[3]These authors contributed equally
[4]Lead contact
*Correspondence: asharifi@sharif.edu
https://doi.org/10.1016/j.isci.2023.107945

The standard deviation (SD) of CT values is one of the main stability measures of an RG. In the case of 2 candidate RGs with equal SD, the one with a higher expression level is preferable. The reason is genes with low abundance have a higher chance of not being detected due to technical errors. Accordingly, the coefficient of variation (CV) is another widely used measure for an RG's stability, which is the SD of expression divided by mean expression. Both SD and CV are vastly used to measure the stability of an RG in the literature, but no theoretical explanation has been provided.[14–16]

In this work, we first elucidate the theoretical assumptions behind SD and CV as RG expression stability measures. Next, by introducing the family of scale-invariant functions, we explain the reason arithmetic and geometric mean functions can be used for the aggregation of multiple RGs. Then by providing mathematical solutions for the optimization of the weighted version of geometric and arithmetic mean functions, we present four novel weighted aggregation methods to optimally minimize the SD and CV of a virtual RG (see STAR Methods). Each proposed method is defined by an aggregation function (geometric or arithmetic mean) and a measure (SD or CV) to be minimized. They are named as geom(sd), geom(cv), arith(sd) and arith(cv) and we call them weighting methods. The usual unweighted geometric mean is abbrevated as *geom* throught the paper. To evaluate these methods, we utilized qPCR array datasets. qPCR array datasets with a high number of genes or other RNA molecules can be normalized without RGs. This feature enabled us to design a benchmarking pipeline using which we could calculate stability measures of different combinations of weighted RGs on the normalized data and compare the weighting methods. Finally, we chose the best method (geom(sd)) and showed the weights of the weighted geometric mean optimized on SD of logarithm of expression can be calculated solely from the raw CT values without being affected by the gene-independent technical variations. Experimental evaluations show that this method can also be used in low-sample-size conditions.

Related works

Various RG expression stability methods have been previously proposed, and they are being highly utilized for the selection process of RGs and their stability assessment.

GeNorm is an iterative method that uses pairwise variation to measure the stability of RGs. In each iteration, the candidate with the worst stability score is removed. Then, the procedure is repeated until only 2 genes remain. The method's assumption is that all input genes should have low expression variation.[5]

NormFinder is another widely used tool that suggests a mathematical model that separates technical and biological variations and then eliminates the technical biases to find the RG with the lowest biological variation.[13] NormiRazor offers a graphics processing unit (GPU)-based implementation of GeNorm, NormFinder, and BestKeeper to examine the stability of high number of RG combinations in parallel.[17]

To our knowledge there has previously been only one study that has approached the problem of aggregating multiple RGs. This method uses weighted geometric mean and follows a heuristic approach to define each RG's weight by the ratio of their SD.[18] We have named it geom(sd_r), and it has been evaluated along with our proposed weighting methods.

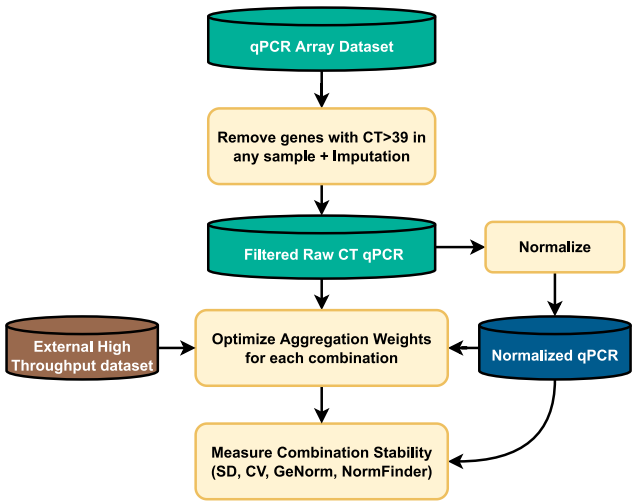


Figure 1. Benchmark workflow to evaluate and compare the weighting methods in terms of stability measures utilizing qPCR array datasets

SD: Standard Deviation, CV: Coefficient of Variation

Figures on the left and right side are for the breast cancer and liver cancer, respectively. RG_d is the weighted/unweighted mean of a combination of d miRNAs.

(A) Mean stability of all combinations of two miRNAs in different weighting methods (the lower is better).

(B) Boxplots for the stability of all combinations of two miRNAs in different weighting methods and the tile figures show paired Wilcoxon test between the stability of different weighting methods on raw CT values. Colored tiles indicate that the row weighting method had significantly ($p < 0.01$) lower stability than the column one.

Figure 1. *Continued*

(C) Sample size analysis: for each sample size, the SD of each combination of two miRNAs is calculated and averaged. This process is repeated 20 times, and the error bars show the standard deviation of the repeats. The weights were calculated based on raw CT values. A statistical comparison in each sample size is provided in Figures S9 and S10.

(D) The number of reference genes effect on stability. For each number of reference genes, the SD of different combinations of miRNAs was calculated. SD: Standard Deviation, Normalized: weights were optimized on the normalized data, raw CT: weights were optimized on the raw CT values, geom: usual geometric mean. Also for complete figures of all stability measures refer to the Figures S5–S8.

RESULTS

Following the benchmark workflow presented in Figure 1, our evaluation of the proposed methods was carried out in three scenarios which are different based on the data type that the aggregation weights were calculated from.

- Raw CT values of the qPCR array: shows the effect of qPCR technical variations on the performance of weighting methods.
- Normalized qPCR array: exhibits how much the weighting method could lower the biological variation if the data had very small to no technical variation. The mean CT of all expressed miRNAs (CT < 35) is used as the normalization factors.
- An external biologically compatible RNA-seq dataset: shows if the weights could be calculated from a separate high-throughput dataset.

In all three scenarios the stability measures (SD, CV, GeNorm, and NormFinder) were calculated on the normalized qPCR array, but the weights of the weighting methods were optimized on the aforementioned data types.

Stability comparison of weighting methods

Here we utilized two qPCR array datasets with different variability levels across samples: a breast cancer tissue dataset and a liver cancer plasma dataset with median expression SD of 1.6 and 2.77, respectively, among their miRNAs. The SD of the log2 of the normalization factors was considered as an estimate of the technical variation caused by RNA abundance of the samples which were 0.45 and 0.72 for the breast and liver cancer datasets, respectively. We calculated the mean stability of all combinations of two miRNAs aggregated by different wighting methods. As illustrated in Figures 2A and 2B, on both datasets the geom(sd) method outperforms all other methods in terms of the aggregated stability, specifically when the aggregation weights are calculated based on the raw CT values. The geom(cv) method is the second-best method on the breast cancer dataset, but it has the worst stability on the liver cancer dataset. By contrast, the arith(sd) method performs as well as geom(sd) in the liver dataset. The difference between the normalized and raw CT results suggests that geom(cv) and arith(cv) are sensitive to technical variations in raw CT values and thus not suitable for datasets with high technical variation. All methods tested outperform the usual geometric mean. Tables 1 and 2 show each stability measure separately. Expectedly, geom(cv) had a lower CV than others, but high numbers on other measures have made its overall stability worse than others in the liver dataset.

Table 1. Comparison between different weighting methods based on four stability measures on the breast cancer dataset

Base	weighting method	SD	CV	GeNorm	NormFinder	Stability
	arith	0.833 ± 0.244	0.654 ± 0.303	0.593 ± 0.082	0.081 ± 0.024	4.004 ± 4.625
	geom	0.765 ± 0.19	0.531 ± 0.172	0.586 ± 0.073	0.075 ± 0.018	2.518 ± 3.412
	geom(rand)	0.826 ± 0.243	0.588 ± 0.247	0.595 ± 0.073	0.08 ± 0.024	3.579 ± 4.323
Raw CT	arith(cv)	0.724 ± 0.173	0.504 ± 0.153	0.574 ± 0.091	0.07 ± 0.017	1.787 ± 3.288
Raw CT	geom(cv)	0.723 ± 0.174	**0.48 ± 0.127**	0.568 ± 0.097	0.07 ± 0.017	1.57 ± 3.236
Raw CT	arith(sd)	0.712 ± 0.163	0.52 ± 0.172	0.575 ± 0.087	0.069 ± 0.016	1.755 ± 3.231
Raw CT	geom(sd_r)	0.726 ± 0.165	0.5 ± 0.144	0.578 ± 0.078	0.071 ± 0.016	1.836 ± 3.065
Raw CT	geom(sd)	**0.703 ± 0.158**	0.487 ± 0.137	**0.567 ± 0.092**	**0.068 ± 0.015**	**1.396 ± 3.076**
Normalized	arith(cv)	0.715 ± 0.172	0.495 ± 0.149	0.571 ± 0.092	0.07 ± 0.017	1.607 ± 3.282
Normalized	geom(cv)	0.711 ± 0.172	0.473 ± 0.125	0.564 ± 0.098	0.069 ± 0.017	1.371 ± 3.227
Normalized	arith(sd)	0.708 ± 0.164	0.514 ± 0.167	0.572 ± 0.089	0.069 ± 0.016	1.648 ± 3.248
Normalized	geom(sd_r)	0.717 ± 0.164	0.494 ± 0.142	0.572 ± 0.084	0.07 ± 0.016	1.637 ± 3.114
Normalized	geom(sd)	0.699 ± 0.158	0.483 ± 0.137	0.564 ± 0.095	0.068 ± 0.015	1.293 ± 3.12

At each column, the minimum number is specified in a bold style and the second minimum is underlined (the rows with Base:Normalized are not considered in this styling). The Base column determines what type of data the weights are optimized on. Arith: arithmetic mean, geom: geometric mean. (sd): weights are optimized to minimize the standard deviation of ct or normalized ct data,(cv): weights are optimized to minimize coefficient of variation of $2^{(-ct)}$ or $2^{(-\text{normalized ct})}$, (rand): weights are randomly sampled from a uniform distribution.

Table 2. Comparison between different weighting methods based on four stability measures on the liver cancer dataset

Base	weighting method	SD	CV	GeNorm	NormFinder	Stability
	arith	1.909 ± 0.769	1.272 ± 0.531	1.22 ± 0.28	0.427 ± 0.174	4.036 ± 3.926
	geom	1.707 ± 0.608	1.041 ± 0.397	1.201 ± 0.255	0.38 ± 0.137	2.805 ± 3.162
	geom(rand)	1.841 ± 0.728	1.105 ± 0.456	1.225 ± 0.25	0.409 ± 0.164	3.43 ± 3.589
Raw CT	arith(cv)	1.525 ± 0.638	0.891 ± 0.331	1.137 ± 0.287	0.34 ± 0.143	1.691 ± 3.3
Raw CT	geom(cv)	1.72 ± 0.82	**0.888 ± 0.325**	1.186 ± 0.269	0.382 ± 0.182	2.416 ± 3.673
Raw CT	arith(sd)	1.431 ± 0.563	0.992 ± 0.435	**1.073 ± 0.335**	0.318 ± 0.126	1.435 ± 3.352
Raw CT	geom(sd_r)	1.514 ± 0.586	0.955 ± 0.384	1.12 ± 0.307	0.337 ± 0.132	1.748 ± 3.273
Raw CT	geom(sd+)	**1.428 ± 0.549**	0.937 ± 0.384	1.088 ± 0.323	**0.317 ± 0.123**	**1.337 ± 3.18**
Normalized	arith(cv)	1.466 ± 0.619	0.837 ± 0.326	1.102 ± 0.312	0.327 ± 0.139	1.268 ± 3.323
Normalized	geom(cv)	1.603 ± 0.765	0.827 ± 0.319	1.151 ± 0.282	0.357 ± 0.171	1.803 ± 3.571
Normalized	arith(sd)	1.398 ± 0.566	0.943 ± 0.41	1.061 ± 0.339	0.311 ± 0.127	1.18 ± 3.36
Normalized	geom(sd_r)	1.492 ± 0.591	0.948 ± 0.387	1.11 ± 0.313	0.332 ± 0.133	1.634 ± 3.324
Normalized	geom(sd+)	1.404 ± 0.554	0.92 ± 0.386	1.085 ± 0.319	0.312 ± 0.125	1.218 ± 3.22

At each column, the minimum number is specified in a bold style and the second minimum is underlined (the rows with Base:normalized are not considered in this styling). The Base column determines what type of data the weights are optimized on. Arith: arithmetic mean, geom: geometric mean. (sd): weights are optimized to minimize the standard deviation of ct or normalized ct data,(cv): weights are optimized to minimize coefficient of variation of $2^{(-ct)}$ or $2^{(-normalized\ ct)}$, (rand): weights are randomly sampled from a uniform distribution.

Sample size analysis

We analyzed the effect of sample size on the performance of the weighting methods. Each sub-sampling is repeated 20 times, and the mean SD of all RG_2 combinations is presented in Figure 2C. The weights were calculated from the raw CT values, and confidence intervals are provided. As expected, increasing the sample size improves all methods. Specifically geom(sd+), the improved version of geom(sd) for small sample sizes, shows significantly better results in lower samples sizes, and as the number of samples decreases the improvement gap between geom(sd+) and geom(sd) increases. On the breast cancer dataset, the only methods which outperformed the usual geometric mean for the low sample size of 10 were geom(sd+) and geom(sd_r), and as samples increased, although geom(sd+) kept improving, geom(sd_r) stayed still. On the liver cancer dataset, all methods except geom(cv) performed better than the usual geometric mean independent of the sample size.

Number of RGs

Consider RG_k is the weighted/unweighted mean of a combination of k RGs. In order to figure out how the number of RGs affects the geom(sd) method, an iterative approach was used. Starting with $k = 2$ first, we evaluated all combinations of RG_2 in terms of SD. Then at each iteration k, the top 400 combinations of RG_k with the lowest SD were crossed with the remaining genes to build the combinations of RG_{k+1}. This process was repeated till the number of RGs reached 20. Figure 2D shows the relation between the number of RGs (k) and the mean SD of RG_k combinations. Until $k = 6$ the geom(sd) weighting method outperforms usual geometric mean. However, when the weights are calculated based on the raw CT values, an over-fitting pattern appears. This result suggests that the geom(sd) weighting method is only applicable to up to five or six RGs.

Weights from external dataset

Here we analyze whether external high-throughput data could be used to calculate the aggregation weights of RGs. The Cancer Genome Atlas (TCGA) breast cancer miRNA expression dataset was obtained as a biologically compatible external dataset for the qPCR array breast cancer dataset. Moreover, the qPCR array dataset is evaluated in three different sample sizes, 20, 30, and 106; just like what we saw in Figure 2, geom(sd+) outperformed other methods (Figure 3). The paired Wilcoxon test reveals that calculating geom(sd+) weights from the miRNA-seq data had better stability results compared to the raw CT with 20 samples but performed on par with the 30-sample case (Figure 3C). On the other hand, the arith(sd) and arith(cv) methods were utterly off by a large margin regarding stability for the miRNA-seq case.

Experimental validation

To demonstrate the utility of the weighting method in a real experiment, we applied it to qPCR data of breast cancer tissue containing expressions of 3 internal controls (miR-16-5p, miR-361-5p, and RNU48) and one target miRNA (miR-21-5p). Figure 4 illustrates how normalizing the expression of the well-known miRNA miR-21-5p using a weighted geometric mean of the internal controls yielded a significant differential expression. In contrast, the usual geometric mean showed no significant change.

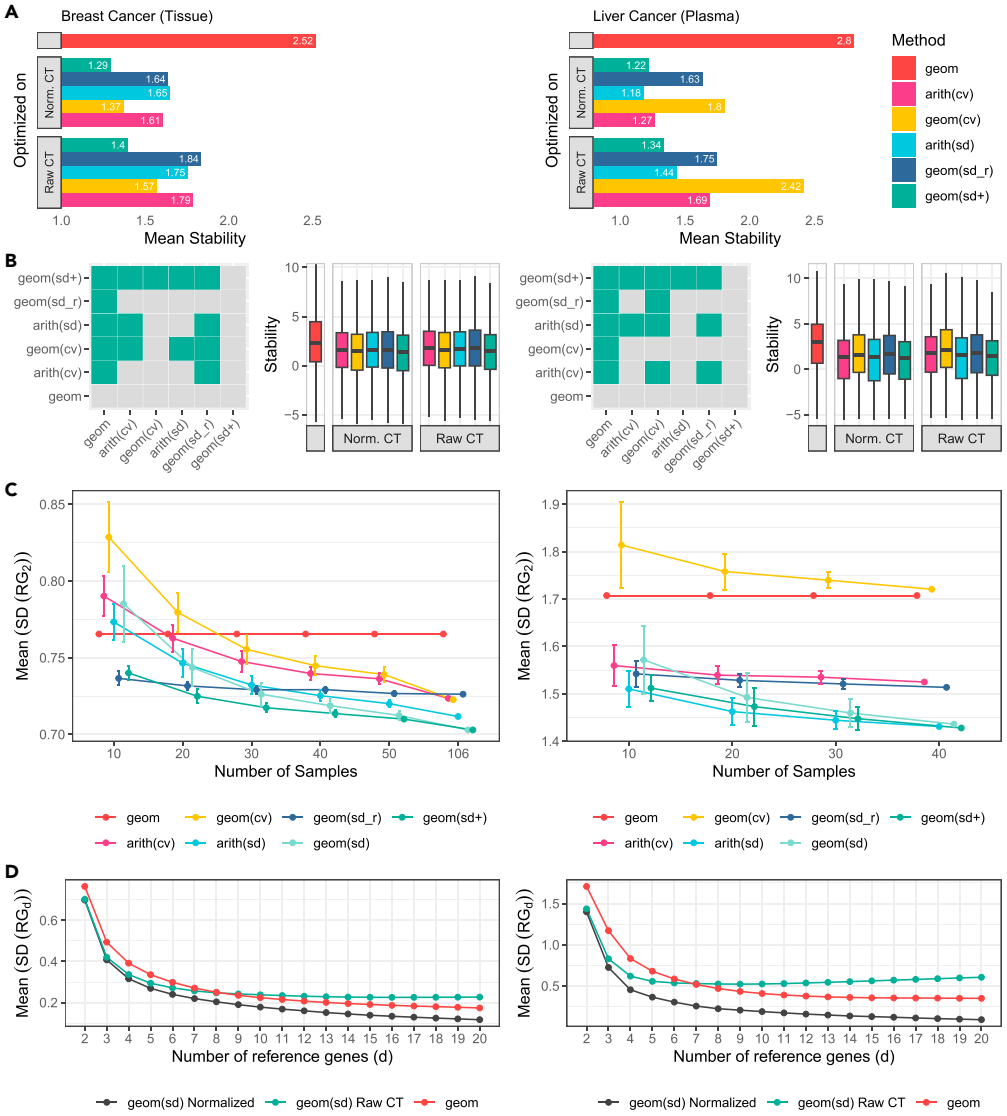


Figure 2. Stability comparison of different weighting methods

The estrogen positive samples of the TCGA BRCA are used as the external RNA-seq dataset.

(A) Mean stability of all combinations of two miRNAs in different weighting methods.

(B) Boxplots for the stability of all combinations of two miRNAs in different weighting methods.

(C) Paired Wilcoxon test between stability scores of different weighting methods. Cells with $p < 0.01$ indicate that the row's weighting method had significantly lower stability scores than the column one (lower is better). Raw CT: weights were optimized on the raw CT values of the entire 106 sample breast cancer qPCR array. n:x means a subset of x samples was taken and an average score of repeating the sub-sampling 20 times was considered. Also for complete figures of all stability measures refer to the Figures S1–S4.

DISCUSSION

In this paper the aggregation of multiple RGs is introduced as an optimization problem. This optimization is formulated in four combinations of stability measures (SD or CV) as the objective functions and weighted mean (geometric or arithmetic) as aggregation functions. The geom(sd) method showed significantly better results compared to other methods in both low- and high-variability conditions as well as different numbers of samples. We also mathematically showed that weights of geom(sd) method are independent of the noise caused by the RNA abundance in different samples, which may justify its superiority over other methods (see STAR Methods). Furthermore, its closed form and regression-based solution allow fast running time and straightforward implementation in various platforms. We have also highlighted how a significant upregulation of miR-21-5p could be overlooked in a real-world case if

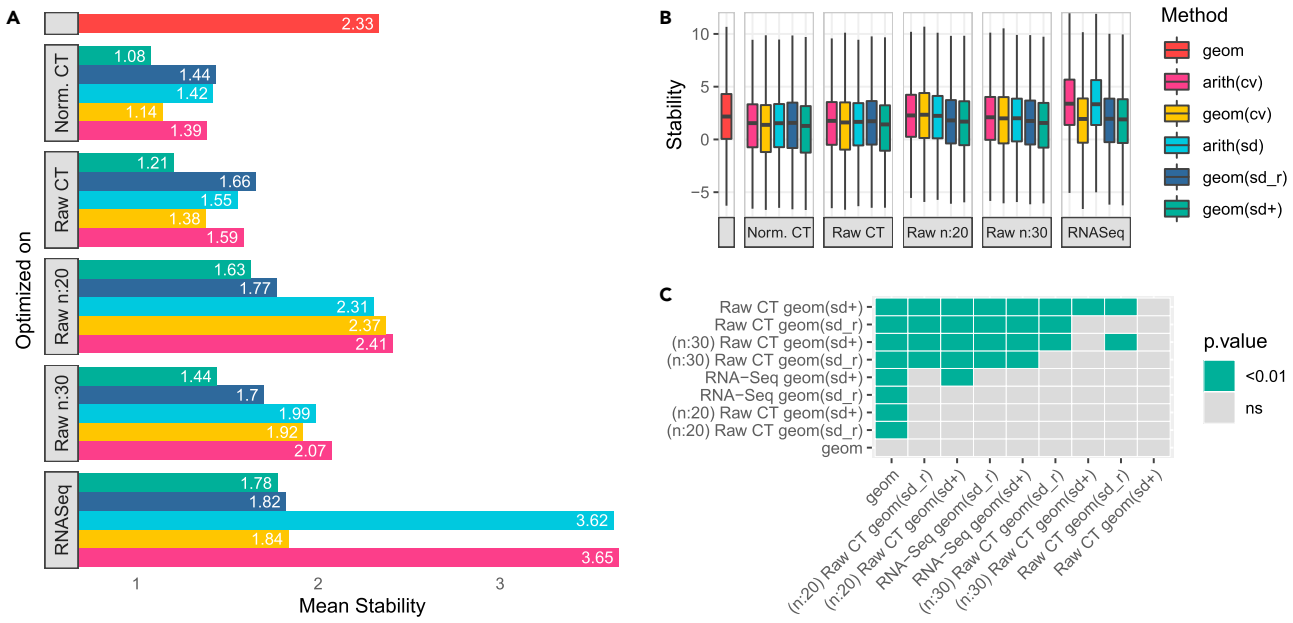


Figure 3. Comparison between external RNA-seq data and raw CT qPCR data with different sample sizes for weights optimization

the non-weighted geometric mean was applied instead. miR-21-5p is a well-known upregulated gene in cancer as it is involved in cell growth and proliferation.[19]

Through optimization of stability measures, our proposed aggregation methods provide either a more stable virtual RG or, at worst, an equal level of stability compared to the single-input RGs. It is also noteworthy that the application of the usual geometric mean could result in a less stable virtual RG (supplemental information of Andersen et al.[13]) specifically when there is a positive covariance between the RGs expression. The geom(sd_r) method which uses the reverse of each RG's SD as aggregation weights is also subject to this problem due to not considering the covariance between RGs.[18] In Equation 25 we demonstrated that in order to minimize the SD of the aggregated virtual RG for a combination of two RGs, their covariance must be taken into account.

This study found that, when enough samples are available, raw CT values are preferable to external high-throughput data for optimizing the aggregation weights of RGs. In our experiment, we used compatible large-sample-size TCGA breast cancer miRNA-seq data, yet a subset of 30 samples from the qPCR dataset showed better results. This can be a consequence of the distribution shift caused by the platform difference or batch effect. Considering these findings, in Figure 5 a workflow for the use cases of the InterOpt tool is presented. There are two main scenarios in which this tool would be useful. The first and most common scenario is when the RGs are preselected, or the experiment is already done. Then based on the number of samples and availability of high-throughput data with similar biological conditions, the weights would be either calculated based on the raw CT values or the normalized high-throughput dataset. In the second scenario this tool can also be used to choose the best weighted combination of RGs from a qPCR experiment of common RGs or a high-throughput dataset. This use case is more suitable for situations where there is no consensus on the best combination of RGs for a particular biological condition. It is worth noting that to have a persistent and reliable result while using an external high-throughput dataset, the similarity of the clinical and pathological characteristics of samples with the qPCR study is highly recommended for choosing the RGs or calculating the weights of the preselected RGs.

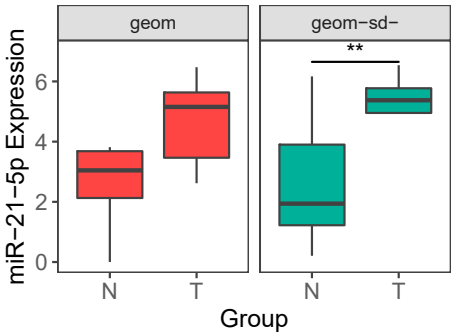


Figure 4. hsa-miR-21-5p expression normalized with the proposed weight method geom(sd) and normal geometric mean of three internal controls: U48, hsa-miR-16-5p, and hsa-miR-361-5p

**: t test p value < 0.01.

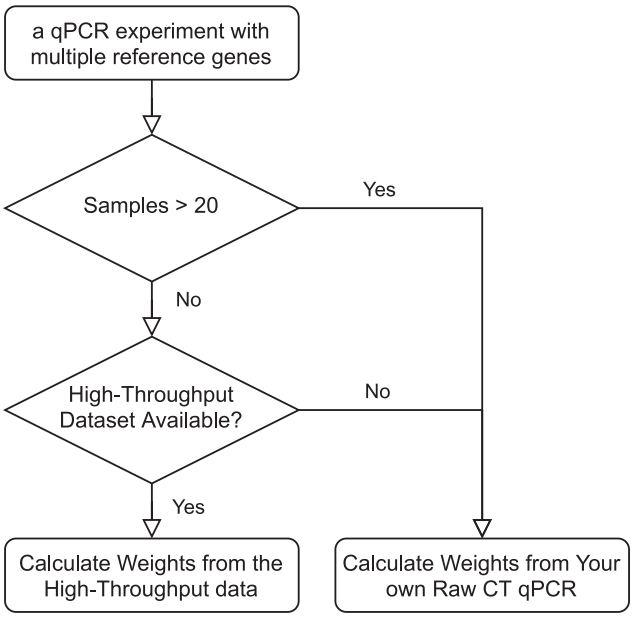


Figure 5. Recommended usage of the proposed method

Due to the effect of gene expression distribution assumption on the RG stability measure, examining other distributions (like beta distribution) and providing better measures are suggested for further improvements. Moreover, we introduced the family of scale-invariant functions as a necessary condition for aggregating multiple RGs. This family of functions can also be explored for more stable members in this line of research.

Limitations of the study

A limitation of the proposed method is the number of RGs. As described in the "Number of RGs" section, aggregating more than 5 RGs by geom(sd) method does not have the expected benefits compared to the regular geometric mean. But it is worth noting that in most cases no more than 3 RGs are quantified. Moreover, the number of samples for evaluating the hsa-miR-21-5p expression in the experimental validation phase was low (12 pairs).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - The criteria of optimal reference gene
  - Gaussian expression distribution leads to CV as RG stability measure
  - Log-normal expression distribution leads to SD as RG stability measure
  - Geometric and arithmetic Mean as aggregation functions
  - Optimizing weighted geometric/arithmetic Mean
  - Benchmark
  - High-throughput and qPCR datasets
  - Implementation
  - Weighting methods mathematical solutions
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107945.

ACKNOWLEDGMENTS

Authors are thankful of Prof. Seyed Javad Mowla from Tarbiat Modares University for his contribution to this work. Also authors thank *greg*, a user from the math.stackexchange.com website for his assistance with some of mathematical solutions.

AUTHOR CONTRIBUTIONS

Conceptualization, A.S., S.R., and A.S.Z.; Methodology, A.S., S.R., and A.S.Z.; Investigation, A.S., S.R., and A.S.Z.; Writing – Original Draft, A.S. and S.R.; Writing – Review & Editing, A.S., S.R., and A.S.Z.; Supervision, A.S.Z.; Visualization, A.S.; Software, A.S. and S.R.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 11, 2023
Revised: August 30, 2023
Accepted: September 13, 2023
Published: September 20, 2023

REFERENCES

1. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., and Erlich, H.A. (1988). Primer-directed enzymatic amplification of dna with a thermostable dna polymerase. Science *239*, 487–491.
2. Garibyan, L., and Avashia, N. (2013). Research techniques made simple: polymerase chain reaction (pcr). J. Invest. Dermatol. *133*, e6.
3. Acinas, S.G., Sarma-Rupavtarm, R., Klepac-Ceraj, V., and Polz, M.F. (2005). Pcr-induced sequence artifacts and bias: insights from comparison of two 16s rrna clone libraries constructed from the same sample. Appl. Environ. Microbiol. *71*, 8966–8969.
4. Taylor, S.C., Nadeau, K., Abbasi, M., Lachance, C., Nguyen, M., and Fenrich, J. (2019). The ultimate qpcr experiment: producing publication quality, reproducible data the first time. Trends Biotechnol. *37*, 761–774.
5. Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. (2002). Accurate normalization of real-time quantitative rt-pcr data by geometric averaging of multiple internal control genes. Genome Biol. *3*. RESEARCH0034.
6. Ghanbari, S., Salimi, A., Rahmani, S., Nafissi, N., Sharifi-Zarchi, A., and Mowla, S.J. (2021). mir-361-5p as a promising qrt-pcr internal control for tumor and normal breast tissues. PLoS One *16*, e0253009.
7. Altenberg, B., and Greulich, K.O. (2004). Genes of glycolysis are ubiquitously overexpressed in 24 cancer classes. Genomics *84*, 1014–1020.
8. Guo, C., Liu, S., and Sun, M.-Z. (2013). Novel insight into the role of gapdh playing in tumor. Clin. Transl. Oncol. *15*, 167–172.
9. Zhang, Y., Li, D., and Sun, B. (2015). Do housekeeping genes exist? PLoS One *10*, e0123691.
10. Marabita, F., De Candia, P., Torri, A., Tegnér, J., Abrignani, S., and Rossi, R.L. (2016). Normalization of circulating microrna expression data obtained by quantitative real-time rt-pcr. Briefings Bioinf. *17*, 204–212.
11. Rice, J., Roberts, H., Rai, S.N., and Galandiuk, S. (2015). Housekeeping genes for studies of plasma microrna: A need for more precise standardization. Surgery *158*, 1345–1351.
12. Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., et al. (2009). The miqe guidelines: M inimum i nformation for publication of q uantitative real-time pcr e xperiments.
13. Andersen, C.L., Jensen, J.L., and Ørntoft, T.F. (2004). 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.
14. Sundaram, V.K., Sampathkumar, N.K., Massaad, C., and Grenier, J. (2019). Optimal use of statistical methods to validate reference gene stability in longitudinal studies. PLoS One *14*, e0219440.
15. Li, C., Xu, J., Deng, Y., Sun, H., and Li, Y. (2019). Selection of reference genes for normalization of cranberry (vaccinium macrocarpon ait.) gene expression under different experimental conditions. PLoS One *14*, e0224798.
16. Li, L., Li, N., Fang, H., Qi, X., and Zhou, Y. (2020). Selection and validation of reference genes for normalisation of gene expression in glehnia littoralis. Sci. Rep. *10*, 7374–7412.
17. Grabia, S., Smyczynska, U., Pagacz, K., and Fendler, W. (2020). Normirazor: tool applying gpu-accelerated computing for determination of internal references in microrna transcription studies. BMC Bioinf. *21*, 425–516.
18. Qureshi, R., and Sacan, A. (2013). A novel method for the normalization of microrna rt-pcr data. BMC Med. Genom. *6*, S14.
19. Wang, H., Tan, Z., Hu, H., Liu, H., Wu, T., Zheng, C., Wang, X., Luo, Z., Wang, J., Liu, S., et al. (2019). microrna-21 promotes breast cancer proliferation and metastasis by targeting lztfl1. BMC Cancer *19*, 738–813.
20. Jansen, M. (2016). microrna expression profiling of response to first-line aromatase inhibitor therapy. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE78870.
21. Shen, J., Wang, A., Wang, Q., Gurvich, I., Siegel, A.B., Remotti, H., and Santella, R.M. (2013). Exploration of genome-wide circulating microrna in hepatocellular carcinoma: Mir-483-5p as a potential biomarker. Cancer Epidemiol. Biomarkers Prev. *22*, 2364–2373.
22. Díaz-Francés, E., and Rubio, F.J. (2013). On the existence of a normal approximation to the distribution of the ratio of two independent normal random variables. Stat. Pap. *54*, 309–323.
23. Bengtsson, M., Ståhlberg, A., Rorsman, P., and Kubista, M. (2005). Gene expression profiling in single cells from the pancreatic islets of langerhans reveals lognormal distribution of mrna levels. Genome Res. *15*, 1388–1392.
24. McCall, M.N., McMurray, H.R., Land, H., and Almudevar, A. (2014). On non-detects in qpcr data. Bioinformatics *30*, 2310–2316.
25. Mestdagh, P., Van Vlierberghe, P., De Weer, A., Muth, D., Westermann, F., Speleman, F., and Vandesompele, J. (2009). A novel and universal method for microrna rt-qpcr data normalization. Genome Biol. *10*, R64.
26. Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The cancer genome atlas pan-cancer analysis project. Nat. Genet. *45*, 1113–1120.
27. Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I., et al. (2016). Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. Nucleic Acids Res. *44*, e71.
28. Venables, W.N., and Ripley, B.D. (2002). Modern Applied Statistics with S, 4th ed. (Springer).
29. Chen, Y., Wiesel, A., Eldar, Y.C., and Hero, A.O. (2010). Shrinkage algorithms for mmse covariance estimation. IEEE Trans. Signal Process. *58*, 5016–5029.

30. Bickel, P.J., and Levina, E. (2008). Covariance regularization by thresholding. Ann. Statist. *36*, 2577–2604.
31. Fletcher, R. (2005). On the barzilai-borwein method. In Optimization and control with applications (Springer), pp. 235–256.
32. Burdakov, O., Dai, Y.-H., and Huang, N. (2019). Stabilized barzilai-borwein method. arXiv. https://doi.org/10.48550/arXiv.1907.06409.
33. RStudio Team (2022). RStudio: Integrated Development Environment for R (RStudio, PBC.).
34. R Core Team (2019). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing Vienna).
35. Lee, K., and You, K. (2021). CovTools: Statistical Tools for Covariance Analysis. R package version 0.5.4.
36. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag).
37. Wilke, C.O. (2020). Cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'.
38. Xiao, N. (2018). Ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for 'ggplot2'.
39. Constantin, A.-E., and Patil, I. (2021). Ggsignif: R Package for Displaying Significance Brackets for 'ggplot2'. PsyArxiv.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
human breast tumor and adjacent normal qPCR	our previous paper, Ghanbari et al.[6]	https://doi.org/10.1371/journal.pone.0253009
breast cancer (tissue) qPCR array	Maurice Jansen, ERASMUS MC - CANCER INSTITUTE	GEO: GSE78870
liver cancer (plasma) qPCR array	Shen et al.[21]	GEO: GSE50013
breast cancer miRNA-Seq	TCGA	TCGA: BRCA miRNA-Seq
Software and algorithms		
InterOpt	This paper	https://github.com/asalimih/InterOpt
TCGAbiolinks v1.4.0	Colaprico A et al.[27]	https://doi.org/10.18129/B9.bioc.TCGAbiolinks RRID:SCR_017683
R v3.6.1	R Core Team	https://www.R-project.org/ RRID:SCR_001905
R Studio	RStudio Team	http://www.rstudio.com RRID:SCR_000432
CovTools v0.5.4	Kyoungjae Lee and Kisung You	https://github.com/kisungyou/CovTools
ggplot2 v3.3.5	Hadley Wickham	https://ggplot2.tidyverse.org/ RRID:SCR_014601
cowplot v 1.0.0	Claus O. Wilke	https://github.com/wilkelab/cowplot RRID:SCR_018081
ggsci v2.9	Nan Xiao	https://github.com/nanxstats/ggsci
ggsignif v0.6.3	Ahlmann-Eltze Constantin and Indrajeet Patil	https://github.com/const-ae/ggsignif RRID:SCR_023047
MASS v7.3-51.4	W. N. Venables and B. D. Ripley	https://github.com/cran/MASS RRID:SCR_019125
nondetects v2.14.0	McCall et al.[24]	https://doi.org/10.18129/B9.bioc.nondetects RRID:SCR_001702

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Ali Sharifi-Zarchi (asharifi@sharif.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The qPCR array datasets (GEO: GSE78870 [20] and GEO: GSE50013 [21]) are publicly available at the Gene Expression Omnibus (GEO) and the imputed preprocessed form are provided inside the InterOpt package (https://github.com/asalimih/InterOpt). The TCGA breast cancer miRNA-Seq dataset can be accessed from Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). The human breast tumor and adjacent normal qPCR data is from our previous paper[6] and is only provided upon request.
- All the proposed methods are publicly available as an R package called InterOpt which can be accessed at https://github.com/asalimih/InterOpt and an online web application is provided at interopt.ir. The code to reproduce all the results and figures is available at https://github.com/asalimih/InterOpt-paper
- Any additional information and tools used in this paper are available from the lead contact upon request

METHOD DETAILS

This section begins with a review of the stability criteria, SD and CV for an optimal RG by modeling the normalization procedure and gene expression distribution. Next, we show why arithmetic and geometric mean could be used as aggregation functions for multiple RGs. We propose solutions to optimize the weighted version of those functions based on different stability criteria. We present a benchmarking pipeline to evaluate the proposed weighting methods in different biological situations as well as number of samples and at last some implementation details are explained.

It is worth mentioning that in this study, a gene consists of both coding and none-coding genes, and by gene expression in a qPCR study, we mean RNA concentration which is affected by both gene transcription and degradation.

The criteria of optimal reference gene

A widely used application of gene expression quantification is differential expression analysis. In which genes expression are compared among samples using fold change or ratio:

$$I = \frac{y_{i,a}}{y_{i,b}} \quad \text{(Equation 1)}$$

Here $y_{i,a}$ is the expression of gene i in sample a and $y_{i,b}$ is the expression of gene i in sample b. We aim to find the biological variations however, the CT values of a qPCR experiment also comprise technical variations. One of the primary sources of technical variation is the different amounts of initial RNA concentration at the start of the qPCR process for each sample. We can model this effect as a coefficient for each sample's different genes:

$$I' = \frac{\alpha_a y_{i,a}}{\alpha_b y_{i,b}} \quad \text{(Equation 2)}$$

Here $\alpha_a y_{i,a}$ and $\alpha_b y_{i,b}$ are the raw measured concentration of gene i in samples a and b accordingly and α_a and α_b represent the technical variation as coefficients. In order to remove this technical variation, each gene expression is divided by an RG (y_r) which is also affected by the technical variation:

$$\frac{\alpha_a y_{i,a}}{\alpha_b y_{i,b}} = \frac{\frac{\alpha_a y_{i,a}}{\alpha_a y_{r,a}}}{\frac{\alpha_b y_{i,b}}{\alpha_b y_{r,b}}} = \frac{y_{i,a}}{y_{i,b}} \frac{y_{r,b}}{y_{r,a}} \approx \frac{y_{i,a}}{y_{i,b}} \quad \text{(Equation 3)}$$

To find the true ratio of the target gene, the ratio of the RG expression in different samples should be close to 1. Z is a continuous random variable with probability density function $f_Z(z;\theta)$ representing this ratio. Hence the objective can be defined as maximizing the probability density of Z in the proximity of 1:

$$\arg\max_{\theta} f_Z(z=1;\theta) \equiv \arg\max_{\theta} \left\{ \lim_{\epsilon \to 0} \int_{1-\epsilon}^{1+\epsilon} f_Z(z;\theta) dz \right\} \quad \text{(Equation 4)}$$

θ represents the parameters of the probability density function.

Gaussian expression distribution leads to CV as RG stability measure

One of the common distributions to model gene expression is the Gaussian distribution. If we model the expression of an RG r by a Gaussian distribution with mean μ and standard deviation σ, the distribution of the ratio of the gene in two different samples would be Equation 6.[22]

$$\frac{y_{r,b} \sim \mathcal{N}(\mu, \sigma^2)}{y_{r,a} \sim \mathcal{N}(\mu, \sigma^2)} = Z \quad \text{(Equation 5)}$$

$$f_Z(z) = \frac{\mu(z+1).e^{-\frac{\frac{\mu^2}{\sigma^2}(z-1)^2}{2(z^2+1)}}}{\sigma\sqrt{2\pi}\left(\sqrt{z^2+1}\right)^3} \cdot \left[\Phi\left(\frac{\mu(z+1)}{\sigma\sqrt{z^2+1}}\right) - \Phi\left(-\frac{\mu(z+1)}{\sigma\sqrt{z^2+1}}\right)\right] + \frac{1}{(z^2+1)\pi} e^{-\frac{\mu^2}{\sigma^2}} \quad \text{(Equation 6)}$$

The definition of coefficient of variation of a gene is defined as standard deviation of gene expression divided by the mean expression:

$$\mu = \frac{\sum_{i=1}^{n} y_{r,i}}{n}, \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n} \left(y_{r,i} - \mu\right)^2}{n}}, \quad \mathrm{CV} = \frac{1}{\delta} = \frac{\sigma}{\mu} \quad \text{(Equation 7)}$$

Where n is the number of samples. According to Equation 4 the goal is to maximize f_z in $z = 1$. Therefore f_z can be rewritten as a function of δ:

$$\delta = \frac{\mu}{\sigma}, z = 1 \rightarrow f_z(z=1,\delta) = \frac{\delta}{2\pi\sqrt{2}}\left[\Phi\left(\sqrt{2}k\right) - \Phi\left(-\sqrt{2}k\right)\right] + \frac{1}{2\pi}e^{-\delta^2}$$
$$= \frac{\delta}{2\pi\sqrt{2}}\int_0^{\delta\sqrt{2}} 2e^{-\frac{t^2}{2}}dt + \frac{1}{2\pi}e^{-\delta^2} \quad \text{(Equation 8)}$$

Now we can show that $\frac{\partial f_z(z=1,\delta)}{\partial\delta}$ is always a positive value:

$$\frac{\partial f_z(z=1,\delta)}{\partial\delta} = \frac{1}{2\sqrt{2}\pi}\int_0^{2k\sqrt{2}} e^{-\frac{t^2}{2}}dt + \frac{2\sqrt{2}\delta}{2\sqrt{2}\pi}e^{-\delta^2} - \frac{2}{2\pi}\delta e^{-\delta^2}$$
$$= \frac{1}{2\sqrt{2}\pi}\int_0^{2k\sqrt{2}} e^{-\frac{t^2}{2}}dt \geq 0 \quad \text{(Equation 9)}$$

Therefore maximization of $\frac{\mu}{\sigma}$ is equivalent to maximizing $f_Z(z=1;\mu,\sigma)$ or in other words minimizing the coefficient of variation (CV) of the Gaussian distribution. This implies CV can be used as a measure of stability, given that the distribution of the RG expression follows a Gaussian distribution.

Log-normal expression distribution leads to SD as RG stability measure

Another previously suggested distribution to model gene expression is log-normal distribution.[23] As the expression of a gene is always a positive number, this distribution has some benefits compared to the Gaussian distribution.

$$\log\left(y_{r,a}\right), \log\left(y_{r,b}\right) \sim \mathcal{N}(\mu,\sigma^2) \quad \text{(Equation 10)}$$

$$Z' = \log(Z) = \log\left(y_{r,a}\right) - \log\left(y_{r,b}\right) \rightarrow Z' \sim \mathcal{N}(0,2\sigma^2) \quad \text{(Equation 11)}$$

Now we can rewrite Equation 4 in terms of Z':

$$\arg\max_{\eta}\int_{-\epsilon}^{\epsilon} f_{Z'}(z',\eta)dz' = \arg\max_{\sigma}\int_{-\epsilon}^{+\epsilon}\mathcal{N}(0,2\sigma^2) = \arg\min_{\sigma}\sigma \quad \text{(Equation 12)}$$

In conclusion, assuming gene expression follows a log-normal distribution, an RG with lower SD of the logarithm of expression is more stable.

Geometric and arithmetic Mean as aggregation functions

We model gene independent technical variations in the form of scaling operations in each sample. As explained in Equation 3, an RG should preserve these variations. Therefore the aggregated virtual RG of a set of d RGs should also follow the same rule:

$$f(\alpha y_1, \alpha y_2, \ldots, \alpha y_d) = \alpha f(y_1, y_2, \ldots, y_d) \quad \text{(Equation 13)}$$

We call the Equation 13 family of functions, scale-invariant functions. Arithmetic and geometric mean are members of this family of functions. Their weighted counterparts are also scale-invariant functions and their weights can be optimized depending on the objective.

$$\text{weighted geometric mean}: RG_d^{(j)} = \prod_{i=1}^{d} y_{i,j}^{w_i}$$
$$\text{weighted arithmetic mean}: RG_d^{(j)} = \sum_{i=1}^{d} w_i \cdot y_{i,j} \quad \text{(Equation 14)}$$
$$\sum_{i=1}^{d} w_i = 1$$

Here $RG_d^{(j)}$ is a weighted mean of a combination of d RGs in the sample j. $y_{i,j}$ is the expression value of i-th RG (among the d RGs) in the sample j and w_i is the i-th RG weight. Throughout the paper, RG_d is also referred as a virtual RG resulted from aggregation of d RGs.

Optimizing weighted geometric/arithmetic Mean

Weighted geometric and arithmetic mean aggregation functions are parametric and the weights can be optimized based on an objective function. Taking together all combinations of arithmetic and geometric mean as the aggregation functions and SD and CV as stability measures we introduce the following weighting methods:

- geom(sd): Aggregation of d RGs using weighted geometric mean, optimized on SD of logarithm of RG_d.
- geom(sd+): Improved version of geom(sd) for small sample size datasets
- arith(cv): Aggregation of d RGs using weighted arithmetic mean, optimized on CV of RG_d.
- geom(cv): Aggregation of d RGs using weighted geometric mean, optimized on CV of RG_d.
- arith(sd): Aggregation of d RGs using weighted arithmetic mean, optimized on SD of logarithm of RG_d.

Each of these weighting methods and their optimization solutions are described in the "weighting methods mathematical solutions" section.

Benchmark

In order to evaluate and compare the weighting methods on real data, we acquired qPCR array datasets and designed the benchmark workflow depicted in Figure 1. First lowly expressed and genes with none-detects in more than 4 samples are removed. The remaining none-detects were imputed using the *nondetects* R package.[24] Next, depending on the weighting method, for any combination of d genes, the Equation 14 is calculated, and the resulting virtual RG's stability is measured based on SD, CV, GeNorm and NormFinder. The individual stability measures are then converted to their corresponding standard z-scores, and their average is taken to obtain the aggregated Stability measure.

The weight optimization of the weighting methods is executed on three different types of data. Raw CT values, normalized CT and a normalized external high-throughput dataset.

High-throughput and qPCR datasets

Two qPCR array datasets (GEO: GSE78870 and GSO: GSE50013) were obtained from the Gene Expression Omnibus (GEO). The GSE78870 [20] contained the expression of 768 miRNAs in 106 primary breast cancer specimens, and GSE50013 [21] contains the expression of 762 (258 detectable) miRNAs in the plasma of 20 patients with hepatocellular carcinoma, as well as 20 healthy donors. To normalize the qPCR array datasets, we have adopted global normalization.[25] In this method the mean expression value of all expressed microRNAs (CT $>$ 35) in each given sample is used as a normalization factor. The Cancer Genome Atlas (TCGA) is a comprehensive cancer genomics program that includes molecular datasets for different types of cancer tissues.[26] Using the TCGAbiolinks package,[27] the breast cancer miRNA-Seq profile of 1097 tumor samples were obtained from the TCGA portal . The count matrix was normalized in count per million (CPM).

For the experimental validation, we used the qPCR dataset from our previous study containing 12 pairs of breast cancer and adjacent normal tissue.[6]

Implementation

All the proposed weighting methods were implemented in an easy to use open source R package called InterOpt which is available at https://github.com/asalimih/InterOpt. geom(sd) and arith(cv) were implemented based on their closed form solutions. The pseudo-inverse in Equation 35 was performed using the *ginv* function from the MASS R package.[28] For geom(cv) a stabilized version of the Barzilai-Borwein gradient method was utilized. It requires less computation and greatly speeds up the convergence compared to other gradient methods. Since no solution for the arith(sd) method was given, an exhaustive search through weights was utilized as an alternative approach.

Running the proposed benchmark for thousands of combinations is not trivial on a single CPU core. Thus we utilized a CUDA accelerated implementation of stability measures (GeNorm, NormFinder) called NormiRazor.[17] The following modifications are applied to the original code:

- Integration of aggregation weights and the capability to handle combinations of $d > 3$ RGs.
- To comply with our normalization method for the qPCR array datasets, NormFinder only uses elements with CT $<$ 35 for calculating each sample mean.
- In Genorm iterations, each gene is only compared with the top 10 stable genes with least SD. This removes the influence of genes with high overall variation on Genorm score.[14]

Weighting methods mathematical solutions

In the following sections the optimization solutions for each weighting method is described in details.

geom(sd) solution 1

This section determines the optimal weighted geometric mean to minimize the SD of the logarithm of the aggregated RG. The geometric mean is equivalent to the arithmetic mean in the logarithmic space; therefore the optimization problem would be as follows:

$$\underset{w_1,w_2,\ldots,w_d}{\arg\min}\,\mathrm{SD}\left(\log\left(\prod_{i=1}^{d} y_{i,j}^{w_i}\right)\right), \quad \text{subject to} \sum_{i=1}^{d} w_i = 1 \tag{Equation 15}$$

In Equation 15, d is the number of RGs, and $y_{i,j}$ is the expression of the i-th RG in sample j. By applying logarithm, production converts to summation and Equation 15 can be rewritten as follows:

$$\begin{aligned} x_{i,j} = \log\left(y_{i,j}\right) \Rightarrow \mathrm{SD}\left(\sum_{i=1}^{d} w_i x_{i,j}\right) &= \sum_{j=1}^{n}\left(\left(\sum_{i=1}^{d} w_i x_{i,j}\right) - \frac{1}{n}\left(\sum_{k=1}^{n}\left(\sum_{i=1}^{d} w_i x_{i,k}\right)\right)\right)^2 \\ &= \sum_{j=1}^{n}\left(\left(\sum_{i=1}^{d} w_i x_{i,j}\right) - \left(\sum_{i=1}^{d} w_i \sum_{k=1}^{n}\left(\frac{x_{i,k}}{n}\right)\right)\right)^2 \\ &= \sum_{j=1}^{n}\left(\sum_{i=1}^{d} w_i\left(x_{i,j} - \sum_{k=1}^{n}\left(\frac{x_{i,k}}{n}\right)\right)\right)^2 \\ &= \sum_{j=1}^{n}\left(\sum_{i=1}^{d} w_i \tilde{x}_{i,j}\right)^2 \end{aligned} \tag{Equation 16}$$

Here $\tilde{x}$ is the mean centered version of x. Equation 16 can be interpreted as the cost function of classic linear regression with MSE cost function:

$$\underset{w_1,w_2,\ldots,w_d}{\arg\min} \sum_{j=1}^{n}\left(\sum_{i=1}^{d} w_i \tilde{x}_{i,j}\right)^2 \equiv \sum_{i=1}^{d} w_i \tilde{X}_i = 0, \quad \text{subject to} \sum_{i=1}^{d} w_i = 1 \tag{Equation 17}$$

To get rid of constraints of w_i, we can rewrite the Equation 17 as Equation 18:

$$\begin{aligned} \sum_{i=1}^{d} w_i \tilde{X}_i &= \sum_{i=1}^{d-1} w_i \tilde{X}_i + w_d \tilde{X}_d \\ &= \sum_{i=1}^{d-1} w_i \tilde{X}_i + \left(1 - \sum_{i=1}^{d-1} w_i\right)\tilde{X}_d \\ &= \sum_{i=1}^{d-1} w_i (\tilde{X}_i - \tilde{X}_d) + \tilde{X}_d \end{aligned}$$

$$\sum_{i=1}^{d-1} w_i (\tilde{X}_d - \tilde{X}_i) - \tilde{X}_d = 0 \quad \equiv \quad \sum_{i=1}^{d-1} w_i G_i = \tilde{X}_d \tag{Equation 18}$$

Here $\tilde{X}_i$ is the mean centered expression of RG_i and $G_i = \tilde{X}_d - \tilde{X}_i$. The solution of Equation 18 linear regression comes in closed form:

$$W_{1..(d-1)} = \left(G^T G\right)^{-1} G^T \tilde{X}_d, \quad w_d = 1 - \sum_{i=1}^{d-1} w_i, \quad G = \begin{bmatrix} \tilde{X}_d - \tilde{X} \\ \tilde{X}_d - \tilde{X}_2 \\ \vdots \\ \tilde{X}_d - \tilde{X}_{d-1} \end{bmatrix} \tag{Equation 19}$$

$W_{1..(d-1)}$ is a $1 \times (d-1)$ matrix consisting of the first $(d-1)$ elements of W.

geom(sd) solution 2

We found a solution to minimize the SD of the weighted geometric mean of multiple RGs. Here, we solve this problem in the context of random variables. Suppose X_1 and X_2 are two random variables representing the logarithm of two RGs expression. The variance of the weighted arithmetic mean of them would be:

$$\mathrm{var}\left(\frac{w_1X_1+w_2X_2}{w_1+w_2}\right) = \frac{w_1^2}{(w_1+w_2)^2}\mathrm{var}(X_1) + \frac{w_2^2}{(w_1+w_2)^2}\mathrm{var}(X_2) + 2\frac{w_1\cdot w_2}{(w_1+w_2)^2}\mathrm{cov}(X_1,X_2) \quad \text{(Equation 20)}$$

w_1 and w_2 are the weights of X_1 and X_2. To find the minimum of this equation, we set the derivative to zero with respect to each of the weights:

$$\frac{d\left\{\mathrm{var}\left(\frac{w_1X_1+w_2X_2}{w_1+w_2}\right)\right\}}{dw_1} = \frac{2w_1(w_1+w_2)^2 - 2w_1^2(w_1+w_2)}{(w_1+w_2)^4}\mathrm{var}(X_1) + \frac{-2w_2^2(w_1+w_2)}{(w_1+w_2)^4}\mathrm{var}(X_2) + 2\frac{w_2(w_1+w_2)^2 - 2w_1w_2(w_1+w_2)}{(w_1+w_2)^4}\mathrm{cov}(X_1,X_2) = 0 \quad \text{(Equation 21)}$$

$$\frac{d\left\{\mathrm{var}\left(\frac{w_1X_1+w_2X_2}{w_1+w_2}\right)\right\}}{dw_1} = \left(2w_1(w_1+w_2) - 2w_1^2\right)\mathrm{var}(X_1) + \left(-2w_2^2\right)\mathrm{var}(X_2) + 2(w_2(w_1+w_2) - 2w_1w_2)\mathrm{cov}(X_1,X_2) = 0 \quad \text{(Equation 22)}$$

After several steps of simplification, we have the Equation 23:

$$\begin{cases} \frac{d\left\{\mathrm{var}\left(\frac{w_1X_1+w_2X_2}{w_1+w_2}\right)\right\}}{dw_1} = 0 \quad \rightarrow w_1w_2\mathrm{var}(X_1) - w_2^2var(X_2) + \left(w_2^2 - w_1w_2\right)\mathrm{cov}(X_1,X_2) = 0 \\ \frac{d\left\{\mathrm{var}\left(\frac{w_1X_1+w_2X_2}{w_1+w_2}\right)\right\}}{dw_2} = 0 \quad \rightarrow -w_1^2\mathrm{var}(X_1) + w_1w_2\mathrm{var}(X_2) + \left(w_1^2 - w_1w_2\right)\mathrm{cov}(X_1,X_2) = 0 \end{cases}$$

$$\begin{aligned} \left(w_1^2+w_1w_2\right)\mathrm{var}(X_1) + \left(-w_2^2 - w_1w_2\right)\mathrm{var}(X_2) + \left(w_2^2 - w_1^2\right)\mathrm{cov}(X_1,X_2) &= 0 \\ w_1\mathrm{var}(X_1) - w_2\mathrm{var}(X_2) + (w_2 - w_1)\mathrm{cov}(X_1,X_2) &= 0 \\ \frac{\mathrm{var}(X_2) - \mathrm{cov}(X_1,X_2)}{\mathrm{var}(X_1) - \mathrm{cov}(X_1,X_2)} &= \frac{w_1}{w_2} \end{aligned} \quad \text{(Equation 23)}$$

Next if we assume that the sum of w_1 and w_2 is equal to 1 then the closed-form solution of w_1 and w_2 would be as follows:

$$w_1 = \frac{\mathrm{var}(X_2) - \mathrm{cov}(X_1,X_2)}{\mathrm{var}(X_1)+\mathrm{var}(X_2) - 2\mathrm{cov}(X_1,X_2)}, \quad w_2 = \frac{\mathrm{var}(X_1) - \mathrm{cov}(X_1,X_2)}{\mathrm{var}(X_1)+\mathrm{var}(X_2) - 2\mathrm{cov}(X_1,X_2)} \quad \text{(Equation 24)}$$

These equations can also be rewritten this way:

$$p_1 = (\mathrm{var}(X_1) - \mathrm{cov}(X_1,X_2))^{-1}, \quad p_2 = (\mathrm{var}(X_2) - \mathrm{cov}(X_1,X_2))^{-1}$$
$$w_1 = \frac{p_1}{p_1+p_2}, \quad w_2 = \frac{p_2}{p_1+p_2} \quad \text{(Equation 25)}$$

The expression values obtained from qPCR (CT values) are subject to technical and biological variations. Here we model the technical variation as an additive random variable called F. So we replace X_1 and X_2 with $X_1 + F$ and $X_2 + F$ in Equation 24. As X_1 and X_2 are in logarithmic space, adding F to them is like applying a random coefficient to each of the samples' true expressions. Now we can simplify the equation as follows:

$$\begin{aligned} w_1 &= \frac{\mathrm{var}(X_2+F) - \mathrm{cov}(X_1+F,X_2+F)}{\mathrm{var}(X_1+F)+\mathrm{var}(X_2+F) - 2\mathrm{cov}(X_1+F,X_2+F)} \\ &= \frac{\mathrm{var}(X_2)+\mathrm{var}(F) - \mathrm{cov}(X_1,X_2) - \mathrm{var}(F)}{\mathrm{var}(X_1)+\mathrm{var}(F)+\mathrm{var}(X_2)+\mathrm{var}(F) - 2\mathrm{cov}(X_1,X_2) - 2\mathrm{var}(F)} \\ &= \frac{\mathrm{var}(X_2) - \mathrm{cov}(X_1,X_2)}{\mathrm{var}(X_1)+\mathrm{var}(X_2) - 2\mathrm{cov}(X_1,X_2)} \end{aligned} \quad \text{(Equation 26)}$$

The same steps could be applied to w_2. As you can see, the result is the same as Equation 24. This shows the robustness of geom(sd) and geom(sd+) method to gene-independent technical variations of qPCR raw CT values.

geom(sd+)

Equation 25 only requires to estimate the variance and covariance of the RGs expression. This enabled us to enhance geom(sd) by utilizing specialized covariance matrix estimation methods for small sample size datasets. After comparing covariance estimation methods for different sample sizes (n), we employed a hybrid approach that uses the oracle approximation shrinkage method for $n < 15$,[29] soft thresholding for $15 \leq n < 85$, and hard thresholding for $n \geq 85$.[30] This hybrid method is named geom(sd+) throughout this paper.

arith(cv)

Suppose W is a $1 \times d$ matrix of RGs weights and Y is a $d \times n$ matrix containing the expression of d RGs in n samples. Here the goal is to minimize the CV of the weighted arithmetic mean of the RGs. This optimization problem is demonstrated in Equation 27:

$$\arg\min_{W} \frac{\mathrm{SD}(WY)}{\mathrm{Mean}(WY)}, \quad \sum_{i=1}^{d} w_i = 1 \qquad \text{(Equation 27)}$$

Equation 27 can be written as:

$$\arg\min_{W} \frac{\sqrt{\frac{1}{n} \| WY - \frac{1}{n} WY\mathbb{1}_n \|_2^2}}{\frac{1}{n} WY\mathbb{1}_n}, \quad \sum_{i=1}^{d} w_i = 1 \qquad \text{(Equation 28)}$$

where $\mathbb{1}_n$ is a $n \times 1$ matrix of ones. To get rid of the constraint, an unconstrained vector x is introduced and used to construct a column vector w, which satisfies the constraint.

$$w = \frac{x}{\mathbb{1}_n^T x} \quad \Rightarrow \quad \mathbb{1}_n^T w = \frac{\mathbb{1}_n^T x}{\mathbb{1}_n^T x} \doteq 1 \qquad \text{(Equation 29)}$$

Then for algebraic convenience, some auxiliary variables are defined:

$$\begin{aligned}
&J = \mathbb{1}_n \mathbb{1}_n^T \\
&C = I - \frac{1}{n} J \quad \text{(Centering Matrix)} \\
&w = W^T \quad \text{(column vector constructed from } x\text{)} \\
&u = Y^T w \quad \Rightarrow du = Y^T dw \\
&z = Cu \quad \Rightarrow dz = CY^T dw \\
&\alpha = \mathbb{1}_d^T x \quad \Rightarrow d\alpha = \mathbb{1}_d^T dx \\
&\beta = \mathbb{1}_n^T u \quad \Rightarrow d\beta = \mathbb{1}_n^T du = \mathbb{1}_n^T Y^T dw \\
&w = \alpha^{-1} x \quad \Rightarrow dw = \alpha^{-1} dx - x\alpha^{-2} d\alpha \\
&\qquad\qquad \Rightarrow dw = \alpha^{-1} \left(I - w\mathbb{1}_d^T \right) dx
\end{aligned} \qquad \text{(Equation 30)}$$

Note that $C^T = C = C^2$ and $\beta = \mathbb{1}_n^T Y^T w = w^T Y \mathbb{1}_n = WY\mathbb{1}_n$. These properties will be used in several of the steps below. First the vector appearing in the numerator is simplified using the new variables:

$$\left(WY - \frac{1}{n} WY\mathbb{1}_n\mathbb{1}_n^T \right)^T = \left(Y^T w - \frac{1}{n} JY^T w \right) = Cu = z \qquad \text{(Equation 31)}$$

The objective function is called φ, and we start by differentiating its square:

$$\begin{aligned}
\varphi^2 &= n\beta^{-2}z^T z \\
2\varphi\, d\varphi &= 2n\beta^{-2}z^T dz - 2n\beta^{-3}z^T z\, d\beta \\
d\varphi &= n\varphi^{-1}\beta^{-3}z^T(\beta\, dz - z\, d\beta) \\
&= n\varphi^{-1}\beta^{-3}z^T\left(\beta\, CY^T - z\mathbb{1}_n^T Y^T\right) dw \\
&= n\varphi^{-1}\alpha^{-1}\beta^{-3}z^T\left(\beta\, CY^T - z\mathbb{1}_n^T Y^T\right)\left(I - w\mathbb{1}_d^T\right) dx \\
\frac{\partial\varphi}{\partial x} &= n\varphi^{-1}\alpha^{-1}\beta^{-3}\left(I - \mathbb{1}_d w^T\right)\left(\beta\, YC - Y\mathbb{1}_n z^T\right)z
\end{aligned} \tag{Equation 32}$$

The gradient is set to zero:

$$\left(\mathbb{1}_d w^T\right)\left(\beta\, YC - Y\mathbb{1}_n z^T\right) z = I\left(\beta\, YC - Y\mathbb{1}_n z^T\right) z \tag{Equation 33}$$

Next, z is eliminated in favor of w.

$$\begin{aligned}
\left(\mathbb{1}_d w^T\right)\left(\beta YC - Y\mathbb{1}_n w^T YC\right) CY^T w &= \left(\beta YC - Y\mathbb{1}_n w^T YC\right) CY^T w \\
\left(\mathbb{1}_d w^T\right)\left(\beta I - Y\mathbb{1}_n w^T\right) YCY^T w &= \left(\beta I - Y\mathbb{1}_n w^T\right) YCY^T w \\
\left(\beta\mathbb{1}_d w^T - \mathbb{1}_d w^T Y\mathbb{1}_n w^T\right) YCY^T w &= \left(\beta I - Y\mathbb{1}_n w^T\right) YCY^T w \\
0 &= \left(\beta I - Y\mathbb{1}_n w^T\right) YCY^T w \\
Y\mathbb{1}_n w^T \sigma YCY^T w &= \beta\,\sigma YCY^T w \\
\left((Y\mathbb{1}_n)w^T\right)\sigma v &= \beta\sigma v \\
Bv &= \beta v
\end{aligned} \tag{Equation 34}$$

The last line is an eigenvalue equation. Since the matrix B is rank-1, there is only one non-trivial eigenvector, which surprisingly allows for a closed-form solution to the problem.

$$\begin{aligned}
v &= Y\mathbb{1}_n \text{ (eigenvector of } B) \\
\left(YCY^T\right)w &= Y\mathbb{1}_n \\
W^T = w &= \left(YCY^T\right)^+ Y\mathbb{1}_n + \left(I - \left(YCY^T\right)^+ YCY^T\right)q
\end{aligned} \tag{Equation 35}$$

Where C is the Centering Matrix $\left(I - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\right)$, H^+ denotes the pseudo-inverse of H, I is the identity matrix, and q is an arbitrary vector.

geom(cv)

The matrix product form of weighted geometric mean function is $\exp(W \ln(Y))$ So if $Y' = \ln(Y)$, then the optimization problem can be formualted as:

$$\arg\min_W \frac{\mathrm{SD}\left(\exp\left(WY'\right)\right)}{\mathrm{Mean}\left(\exp\left(WY'\right)\right)}, \quad \sum_{i=1}^{d} w_i = 1 \tag{Equation 36}$$

The Equation 36 can be written as:

$$\arg\min_W \frac{\sqrt{\frac{1}{n}\left\|\exp\left(WY'\right) - \frac{1}{n}\exp\left(WY'\right)\mathbb{1}_n\right\|_2^2}}{\frac{1}{n}\exp\left(WY'\right)\mathbb{1}_n}, \quad \sum_{i=1}^{d} w_i = 1 \tag{Equation 37}$$

where $\mathbb{1}_n$ is a $n\times 1$ matrix of ones. To get rid of the constraint, an unconstrained vector x is introduced and used to construct a column vector w, which satisfies the constraint.

$$w = \frac{x}{\mathbb{1}_n^T x} \quad \Rightarrow \quad \mathbb{1}_n^T w = \frac{\mathbb{1}_n^T x}{\mathbb{1}_n^T x} \doteq \mathbb{1}_n \tag{Equation 38}$$

Then for algebraic convenience, some auxiliary variables are defined:

$$
\begin{aligned}
& j = \mathbb{1}_n \mathbb{1}_n^T \\
& C = I - \frac{1}{n} J \qquad \text{(Centering Matrix)} \\
& w = W^T \qquad \text{(column vector constructed from } x\text{)} \\
& Q = Y' \circ \mathbb{1}_d y^T \qquad \circ : \text{Hadamard Product} \\
& u = exp\left(Y'^T w\right) \qquad \Rightarrow du = Y'^T \circ exp\left(Y'^T w\right) \mathbb{1}_d^T dw = Q^T dw \\
& z = Cu \qquad \Rightarrow dz = CQ^T dw \\
& \alpha = \mathbb{1}_d^T x \qquad \Rightarrow d\alpha = \mathbb{1}_d^T dx \\
& \beta = \mathbb{1}_n^T u \qquad \Rightarrow d\beta = \mathbb{1}_n^T du = \mathbb{1}_n^T Q^T dw \\
& w = \alpha^{-1} x \qquad \Rightarrow dw = \alpha^{-1} dx - x\alpha^{-2} d\alpha \\
& \qquad\qquad \Rightarrow dw = \alpha^{-1} \left(I - w \mathbb{1}_d^T\right) dx
\end{aligned}
\qquad \text{(Equation 39)}
$$

Note that $C^T = C = C^2$ and $\beta = \mathbb{1}_n^T \exp\left(Y'^T w\right) = \exp\left(w^T Y'\right) \mathbb{1}_n = \exp\left(WY'\right) \mathbb{1}_n$. These properties will be used in several of the steps below. First the vector appearing in the numerator is simplified using the new variables:

$$
\left(\exp(WY') - \frac{1}{n} \exp(WY') \mathbb{1}_n \mathbb{1}_n^T\right)^T = \left(\exp\left(Y'^T w\right) - \frac{1}{n} J \exp\left(Y'^T w\right)\right) = Cu = z \qquad \text{(Equation 40)}
$$

The objective function is called φ, and we start by differentiating its square:

$$
\begin{aligned}
\varphi^2 &= n\beta^{-2} z^T z \\
2\varphi \, d\varphi &= 2n\beta^{-2} z^T dz - 2n\beta^{-3} z^T z \, d\beta \\
d\varphi &= n\varphi^{-1} \beta^{-3} z^T \left(\beta \, dz - z \, d\beta\right) \\
&= n\varphi^{-1} \beta^{-3} z^T \left(\beta \, CQ^T - z \mathbb{1}_n^T Q^T\right) dw \\
&= n\varphi^{-1} \alpha^{-1} \beta^{-3} z^T \left(\beta \, CQ^T - z \mathbb{1}_n^T Q^T\right) \left(I - w \mathbb{1}_d^T\right) dx \\
\frac{\partial \varphi}{\partial x} &= n\varphi^{-1} \alpha^{-1} \beta^{-3} \left(I - \mathbb{1}_d w^T\right) \left(\beta \, QC - Q \mathbb{1}_n z^T\right) z
\end{aligned}
\qquad \text{(Equation 41)}
$$

Contrary to the previous method, there is no closed form solution here. However, the gradient with respect to x where $W^T = \frac{x}{\mathbb{1}_n^T x}$ can be obtained.

$$
\varphi(x) = \left(n\beta^{-2} z^T z\right)^{1/2} \qquad \text{(Equation 42)}
$$

$$
g(x) = n\varphi^{-1} \alpha^{-1} \beta^{-3} \left(I - \mathbb{1}_d w^T\right) \left(\beta \, QC - Q \mathbb{1}_n z^T\right) z \qquad \text{(Equation 43)}
$$

Where $g(x)$ is the gradient of the objective function with respect to x. To optimize it we used a gradient-descend method called Barzilai-Borwein.[31] In Practice, the original Barzilai-Borwein method could diverge from the optimal point. To handle this inconvenience, the step size is cautiously controlled using the stabilized Barzilai-Borwein method:[32]

Initialize

$$
x_0 = random
$$

First step

$$
\begin{aligned}
g_0 &= g(x_0) \\
x_1 &= x_0 - \left(\frac{0.05 \, \varphi(x_0)}{g_0^T g_0}\right) g_0 \\
k &= 1
\end{aligned}
$$

Subsequent steps

$$g_k = g(x_k)$$

$$\text{Step size: } \alpha_k = min\left\{\frac{(x_k - x_{k-1})^T(g_k - g_{k-1})}{(g_k - g_{k-1})^T(g_k - g_{k-1})}, \frac{\Delta}{\| g_k \|}\right\}, \Delta > 0$$

$$x_{k+1} = x_k - \alpha_k g_k$$

$$k = k + 1$$

Stop when $g_k \approx 0$.

arith(sd)

Unlike previous methods no mathematical solution was provided for this optimization problem. Thus, based on the constraint that weights sum to 1, a numerical procedure in which an exhaustive search through weights with a precision of 0.01 was used.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were executed in the RStudio integrated development environment[33] and R language v3.6.1.[34] Paired Wilcoxon rank sum test was carried out for stability comparison between weighting methods with a p.value significance level of 0.01. Covariance estimation of geom(sd+) method was carried out by the CovTools v0.5.4 package.[35] Figures were produced using ggplot2 v3.3.5,[36] cowplot v1.0.0,[37] ggsci v2.9[38] and ggsignif v0.6.3[39] packages.