Journal of Biometrics & Biostatistics

Brown and Dinu, J Biom Biostat 2015, 6:4
DOI: 10.4172/2155-6180.1000252

Review Article Open Access

Ratios and Housekeeper Normalization

Justin R Brown* **and Valentin Dinu**

Department of Biomedical Informatics, Arizona State University, USA

Abstract

A common practice in gene expression studies is to use 'housekeepers', i.e., genes expected to be expressed at relatively constant levels across experimental conditions, to normalize data. The process is to divide an expression value by some composite of one or more stable housekeepers to remove the effect of processing and nuance variables. Despite its reverence and widespread use, we argue that this approach is fundamentally flawed on multiple levels. The outcome of housekeeper normalization is a set of ratio variables which are not amenable to many standard statistical tests. There are no universal housekeeper genes and even within specific cohorts proposed housekeeper genes often fail to replicate. Furthermore, there is also no single agreed upon algorithm for performing housekeeper normalization or agreement regarding what constitutes a good housekeeper. We urge researchers to consider the use of alternative methodologies in their research.

Keywords: Gene expression; Housekeeper normalization; Ratio variables

Introduction

Housekeeper normalization is a process commonly used in gene expression [1,2]. The basic process of housekeeper normalization is to divide a gene expression value by some composite of one or more housekeepers, i.e., constitutive genes that are believed to have a relatively constant level of expression across experimental conditions. The idea behind housekeeper normalization is that the process will help remove or control for sources of variation within an experiment in order to help identify differences between groups or samples [1-3]. Countless sources of variation have been proposed by researchers including but not limited to differences in samples, sample degradation, total amount of RNA, variability in processing steps such as reverse transcription in RT-PCR and others [1,3].

Despite the popularity of housekeeper normalization, the process has a number of highly deleterious flaws. Some problematic issues such as the lack of universal housekeepers, changing roles of housekeepers in different cohorts, no clear standard for what constitutes a good housekeeper and vagueness in the explanation of how housekeeping normalization is mathematically calculated by researchers are issues which have been addressed by researchers in the past. Multiple papers specifically focus on finding good housekeeper genes for use within a given research area [1,3-5] while others focus on finding universal housekeepers [6]. Nonetheless, despite such attempts, careful inspection of literature across multiple conditions or disease states can often find multiple conflicting reports on the validity of many housekeepers as well as the variability in housekeepers across samples, tissue types and physiological states is well documented [7,8]. It has also been documented that experimental conditions often change the expression level of housekeeper genes [9,10]. These findings have led to questions as to whether or not housekeeper genes actually exist in higher organisms [11]. For example, it has been noted that genes such as GAPDH and Beta-Actin which were once thought to be good universal housekeepers in fact perform quite poorly in many conditions and that the expression of these genes is are known to vary significantly across conditions [1,11].

Given the concern about the validity and reliability of housekeeper genes, it has become more common for researchers to use experimental data driven methods for picking housekeepers rather than selecting from a set of "universal housekeepers" [11]. One consistent problem with defining good housekeeper(s) is that there is no clear standard as to how to best pick a housekeeper. A plethora of methods have been used by researchers to screen for viable housekeepers including: comparison of geometric means, standard deviations, coefficient of variation (CV), Pearson product moment correlation coefficients, rank order statistics, regression based methods and more [1,6,12]. Even in the presence of consistent methodologies, there is no clear agreement as to what values of a given test would make for a good housekeeper. Additionally, the use of multiple housekeepers is quite common and there is no consistent agreement about how many housekeepers are needed or how precisely to aggregate multiple housekeepers into a single composite measure. In short, it is almost impossible for a reader to truly know what exactly was done when a paper says they selected and normalized to a set of housekeepers.

In addition to these commonly noted problems, we believe that perhaps an even larger problem is the fact that the process of housekeeper normalization produces a set of ratio variables and mathematical coupling. Statisticians have known for over a century that performing analyses on ratio variables can be very dangerous while often leading to biased results and spurious effects [13]. Additionally, in many cases involving ratios, the null hypothesis of a regression model that the correlation is equal to zero in the population is no longer valid.

Correlations and Regression Coefficient Estimates Among Genes Normalized to a Common Set of Housekeepers

While there is likely variation amongst researchers as to how to physically normalize a set of genes once a set of housekeepers is selected, the basic idea is to divide a gene by some composite value of the housekeeper(s); usually a mean.

***Corresponding author:** Justin R Brown, Department of Biomedical Informatics, Arizona State University, 13212 East Shea Boulevard, Scottsdale, AZ 85259, USA, Tel: +1 480-884-0220; E-mail: Jrbrown2@live.com

Received September 24, 2015; **Accepted** October 12, 2015; **Published** October 19, 2015

Citation: Brown JR, Dinu V (2015) Ratios and Housekeeper Normalization. J Biom Biostat 6: 252. doi:10.4172/2155-6180.1000252

$$\text{Normalized Gene Expression Value}=\frac{\log_2\text{GeneExpression}}{\dfrac{\log_2\text{Hk}_1+\log_2\text{Hk}_2+\ldots\log_2\text{Hk}_n}{n}}$$

Equation 1: Possible Formula for Housekeeper Normalization using n Housekeepers (Hk_i)

It is apparent form equation 1 that the resulting normalized gene expression value is a ratio of some gene expression value to a composite housekeeper value which we will denote as H_v. Since every gene of interest in a dataset will be normal to H_v the relationship among every gene in the dataset is immediately biased. In 1897 Karl Pearson derived the expected correlation between two ratio variables. Briefly, for two genes x and y and a mean value for a set of housekeepers H_v, after normalizing x and y by applying Equation 1 we would have two ratio variables: $x_{norm} = x/ H_v$ and $y_{norm} = y/H_v$. Pearson noted that the correlation between ratio variables depends on the correlation between the variables independently – x, y and H_v in our example – as well as the variances of each variable. If x, y and h ,all have equal variances and in the population, y and h are all uncorrelated with one another, then the correlation between x_{norm} and y_{norm} would be approximately 0.5! Furthermore the expected correlation between x_{norm} and H_v will be approximately $\frac{-1}{\sqrt{2}}$ or ~-0.7071.

Given that genes of interest should change based on some grouping variable such as disease state (e.g., benign or malignant/cancerous) while housekeepers are expected to be relatively constant or at a minimum not correlated with a grouping variable, it follows that on average we would expect little if any correlation between housekeepers and a gene of interest. It is precisely this lack of correlation between the genes of interest and housekeepers that produces artificially high correlations among completely unrelated variables. The second we normalize a set of genes to a common set of housekeepers, the interpretation of any correlation we compute instantly changes.

At best one could argue that ratio variables simply change the standard interpretation of the null hypothesis. It is common statistical practice to construct a standard null hypothesis stating that the expected correlation is equal to zero or that there is no effect in the population. However, with ratio variables, the expected correlation when there is no relationship is quite different; 0.5 to be precise. Additionally, the research question changes in that the analysis is no longer about variables x and y but about composite variables x_{norm} and y_{norm}. In the biomedical research literature, it is very rare to see gene expression data subject to housekeeper normalization being discussed with respect to composite variables. Researchers almost always refer to the component variables and not the composite.

Karl Pearson suggested more than a century ago that correlations among ratio variables are often “spurious” correlations [13]. However, we tend to agree with others such as Neyman who point out that calling the correlation between composite variables “spurious” is not quite accurate because there is nothing wrong with Pearson’s original equations for calculating correlations. Rather the problem is the “method of study” and the act of creating the composite variables that is the problem [14-17].

Making any statistical inferences under the assumption that a correlation of zero is a standard null hypothesis that assumes there is no relationship among gene expression values normalized with common housekeepers is statistically inaccurate because the expected correlation is no longer zero even if there is no correlation in the population. While Neyman was technically correct in suggesting that this really only changes the interpretation and null hypothesis [14], it should be noted that usually researchers make conclusions based off of an incorrect null hypothesis both mathematically and theoretically. Results reported based on an incorrect null hypothesis, whether statistically significant or statistically non-significant, are in fact spurious effects from the perspective that the reported results are not real and are simply artifacts of incorrect statistical analysis.

Gene expression data using housekeeper normalization is so highly confounded with the housekeepers used and the housekeeper normalization process, one can only make claims about the composite variables. Although the composite variables when normalized to the same set of housekeepers under the same mathematical formulas may be reproducible in limited and carefully controlled circumstances, it is unclear what the interpretation of such variable is. Given that the set of housekeepers used in normalization varies quite precipitously from study to study along with the actual mathematical calculations for computing normalization composite variables, it is very difficult to see how one can make comparisons across studies.

It should also be noted that gene expression data is commonly log base 2 transformed in order to make the data more linear; and thus applicable to the use of standard statistical models. However, even after log transformation, care needs to be exercised to ensure distributional assumptions are met and to ensure that the data does not still exhibit properties of an outlier prone distribution. Data necessary to assess a sample’s distributional properties such as equality of variances and normality is rarely if ever given. Distributional properties are critical in determining the accuracy of continuous variables and without distributional information the expected correlation is among a set of ratio variables is almost impossible. In light of this, it is practically impossible to reassess results from previous studies without access to the raw data.

In addition to correlation coefficients, the use of ratio variables also creates many problems for other statistical tests such as regression. Among the most prominent problems created by the use of ratio variables in regression modeling is the effect on the variances [17]. When creating ratio variables, if the component variables are homoscedastic (have equal variances) creating ratios often has the tendency to lead to heteroscedastic (unequal variances). Assumptions regarding equality of variances is one of the most core principles underlying ordinary least squares regression and violations have the ability to dramatically bias regression estimates by either inflating or deflating the estimate [18].

For gene expression studies, the issue of homoscedastic variances is complicated because of the common distributional properties of gene expression values. Researchers noting the effects of variance structures after creating ratio variables primarily focused on the case where the variables were homoscedastic to begin with [17]. Raw intensity values for gene expression microarrays are commonly on a multiplicative scale rather than a linear scale and to compensate, the data is almost always log2 transformed. Unfortunately, log2 transformation does not always produce variables which meet test for normality and homoscedastic variances. As a result, predicting the distributional properties of resultant ratio variables is an additional concern.

Astute statisticians will note that tests for homoscedastic variances should be commonly performed and that there are a number of estimating procedures beyond ordinary least squares such as empirical covariance estimators or “sandwich estimators” which will produce accurate estimates with heteroscedastic variances [19]. We believe that using such methods would only be a band-aid on a larger problem and would ignore the plethora of other pitfalls surrounding ratio variables. Over a century of scientific literature has thoroughly documented the hazards and warned against the use of ratio variables across a range of statistical tests. Researchers are well intentioned for using housekeeper normalization to control external sources of variability and improve the accuracy of results. However, the resulting ratio variables are doing more harm than good. We advocate researchers consider alternative methods for normalization and exercise due diligence when interpreting housekeeper normalized research studies [20,21].

Discussion and Conclusion

The processing and analysis of gene expression data is a complex process. As researchers have pointed out for many years, there are a significant number of external variables which need to be addressed in order to obtain accurate results. A common approach taken in the field is to use housekeeper normalization. While the goal behind the method is laudable, the problem is that the mathematical computation creates a set of ratio variables. More than a century of statistical literature consistently illustrates the mathematical problems and lack of interpretability of using ratio variables; irrespective of the fields of study. The correlation among ratio variables is biased, null hypotheses are not interpretable and fundamental assumptions of OLS regression models are violated when ratio variables are treated as normal, independent and continuous. As a result of the well documented problems with ratio variables, the lack of clarity as to how housekeepers are chosen, mathematical details of how housekeeper normalization is computed, we have to suggest that the process be entirely abandoned. The mathematical properties of ratio variables formed by the housekeeper normalization process at best biases any estimates, and at worst makes one's ability to draw association conclusions and compare across studies impossible. We strongly suggest researchers consider alternative methodologies in their studies.

References

1. Dheda K, Huggett JF, Bustin SA, Johnson MA, Rook G, et al. (2004) Validation of housekeeping genes for normalizing RNA expression in real-time PCR. Biotechniques 37: 112-114, 116, 118-119.
2. Karge WH, Schaefer EJ, Ordovas JM (1998) Quantification of mRNA by polymerase chain reaction (PCR) using an internal standard and a nonradioactive detection method. Methods Mol Biol 110: 43-61.
3. Peltier HJ, Latham GJ (2008) Normalization of microRNA expression levels in quantitative RT-PCR assays: identification of suitable reference RNA targets in normal and cancerous human solid tissues. RNA 14: 844-852.
4. Kok J, Roelofs R, Giesendorf B, Pennings J, Waas, E, et al. (2005) Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. Laboratory Investigation 85: 154-159.
5. Nicot N, Hausman JF, Hoffmann L, Evers D (2005) Housekeeping gene selection for real-time RT-PCR normalization in potato during biotic and abiotic stress. J Exp Bot 56: 2907-2914.
6. Lee S, Jo M, Lee J, Koh SS, Kim S (2007) Identification of novel universal housekeeping genes by statistical analysis of microarray data. J Biochem Mol Biol 40: 226-231.
7. Suzuki T, Higgins PJ, Crawford DR (2000) Control selection for RNA quantitation. Biotechniques 29: 332-337.
8. Chen X, Cheung ST, So S, Fan ST, Barry C, et al. (2002) Gene expression patterns in human liver cancers. Mol Biol Cell 13: 1929-1939.
9. Schmittgen TD, Zakrajsek BA (2000) Effect of experimental treatment on housekeeping gene expression: validation by real-time, quantitative RT-PCR. J Biochem Biophys Methods 46: 69-81.
10. Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, et al. (1999) Housekeeping genes as internal standards: use and limits. J Biotechnol 75: 291-295.
11. Stafford P (2008) Methods in Microarray Normalization, CRC Press.
12. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res 29: 2549-2557.
13. Pearson K (1897) Mathematical contributions to the theory of evolution–On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London 60: 489-497.
14. Neyman J (1952) Lectures and conferences on mathematical statistics and probability. (2nd edition), U.S. Department of Agriculture,Washington, DC.
15. Yule E (1910) On the interpretation of correlations between indices or ratios. Journal of the Royal Statistical Society 73: 644-647.
16. Long SB (1979) The continuing debate over the use of ratio variables: Facts and fiction. In Sociological methodology 1980, San Francisco.
17. Khu E, Meyer J (1955) Correlation and Regression Estimates when the Data are Ratios. Econometrica 23: 400-416.
18. Cohen J, Cohen P, West S, Aiken L (2003) Applied Multivariate Regression for the Behavioral Sciences. (3rdedn), Routledge Academic.
19. Morel JG, Bokossa MC, Neerchal NK (2003) Small Sample Correction for the Variance of GEE Estimators. Biometrical Journal 4: 395-409.
20. Meyers R, Montgomery D, Anderson-Cook C (2009) Response Surface Methodology: Process and Product Optimization Using Designed Experiments. Wiley Publishing.
21. Montgomery D (2008) Introduction to Statistical Quality Control. Wiley Publishing.