PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Proportionality: A Valid Alternative to Correlation for Relative Data

David Lovell[1]*, Vera Pawlowsky-Glahn[2], Juan José Egozcue[3], Samuel Marguerat[4], Jürg Bähler[5]

1 Queensland University of Technology, Brisbane, Australia, **2** Dept. d'Informàtica, Matemàtica Aplicada i Estadística. U. de Girona, España, **3** Dept. Applied Mathematics III, U. Politécnica de Catalunya, Barcelona, Spain, **4** MRC Clinical Sciences Centre, Imperial College London, United Kingdom, **5** Research Department of Genetics, Evolution and Environment, University College London, United Kingdom

* David.Lovell@qut.edu.au

Citation: Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J (2015) Proportionality: A Valid Alternative to Correlation for Relative Data. PLoS Comput Biol 11(3): e1004075. doi:10.1371/journal.pcbi.1004075

Editor: Roland L. Dunbrack Jr., Fox Chase Cancer Center, UNITED STATES

Received: June 29, 2014

Accepted: December 8, 2014

Published: March 16, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: DL was funded by the Commonwealth Scientific and Industrial Research Organisation (www.csiro.au). VPG was funded by the Spanish Ministry of Education, Culture and Sports under a Salvador de Madariaga grant (Ref. PR2011-0290) and by the Spanish Ministry of Economy and Competitiveness under the project METRICS Ref. MTM2012-33236. JJE was funded by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under project Ref: 2009SGR424. SM was funded by the UK Medical Research Council. JB was

Abstract

In the life sciences, many measurement methods yield only the relative abundances of different components in a sample. With such relative—or *compositional*—data, differential expression needs careful interpretation, and correlation—a statistical workhorse for analyzing pairwise relationships—is an inappropriate measure of association. Using yeast gene expression data we show how correlation can be misleading and present proportionality as a valid alternative for relative data. We show how the strength of *proportionality* between two variables can be meaningfully and interpretably described by a new statistic ϕ which can be used instead of correlation as the basis of familiar analyses and visualisation methods, including co-expression networks and clustered heatmaps. While the main aim of this study is to present proportionality as a means to analyse relative data, it also raises intriguing questions about the molecular mechanisms underlying the proportional regulation of a range of yeast genes.

Author Summary

Relative abundance data is common in the life sciences, but appreciation that it needs special analysis and interpretation is scarce. Correlation is popular as a statistical measure of pairwise association but should not be used on data that carry only relative information. Using timecourse yeast gene expression data, we show how correlation of relative abundances can lead to conclusions opposite to those drawn from absolute abundances, and that its value changes when different components are included in the analysis. Once all absolute information has been removed, only a subset of those associations will reliably endure in the remaining relative data, specifically, associations where pairs of values behave proportionally across observations. We propose a new statistic ϕ to describe the strength of proportionality between two variables and demonstrate how it can be straightforwardly used instead of correlation as the basis of familiar analyses and visualization methods.

funded by a Wellcome Trust Senior Investigator Award (grant #095598/Z/11/Z). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

This is a *PLOS Computational Biology* Methods paper.

Introduction

Relative abundance measurements are common in molecular biology: nucleic acids typically have to be provided at a set concentration for sequencing or microarray analysis; sequencing methods report a large but finite total of reads, of which any particular sequence is a proportion. Sometimes, researchers are interested in the relative abundance of different components. Other times, they have to make do with relative abundance to gain insight into the system under study. Whatever the case, data that carry only *relative* information need special treatment.

Awareness is growing [1, 2, 3] but it is not yet widely appreciated that common analysis methods—including correlation—can be very misleading for data carrying only relative information. *Compositional data analysis* [4] (CoDA) is a valid alternative that harks back to Pearson's observation [5] of *'spurious correlation'*, i.e., while statistically independent variables X, Y, and Z are not correlated, their ratios X/Z and Y/Z must be, because of their common divisor. (Note: this differs from the logical fallacy that "correlation implies causation".)

Proportions, percentages and parts per million are familiar examples of compositional data; the fact that the representation of their components is constrained to sum to a constant (i.e., 1, 100, 10^6) emphasizes that the data carry only relative information. Note that compositional data do not necessarily have to sum to a constant; what *is* essential is that only the *ratios* of the different components are regarded as informative.

Correlation—Pearson, Spearman or other—leads to meaningless conclusions if applied to compositional data because its value depends on which components are analyzed [4]. Problems with correlation can also be demonstrated geometrically (Fig. 1): the bivariate joint distribution of relative abundances says nothing about the distribution of absolute abundances that gave rise to them. Thus, relative data is also problematic for mutual information and other distributional measures of association. To further illustrate how correlation can be misleading we applied it to absolute and relative gene expression data in fission yeast cells deprived of a key nutrient [6].

How then can we make sound inferences from relative data? We show how *proportionality* provides a valid alternative to correlation and can be used as the basis of familiar analyses and visualizations. We conclude by putting this analysis strategy in perspective, discussing challenges, caveats and issues for further work, as well as the biological questions raised in this study.

Results

Data on absolute mRNA abundance

Our results are based on data from Marguerat *et al.* [6] on the absolute levels of gene expression (i.e., mRNA copies per cell) in fission yeast after cells were deprived of a key nutrient (Fig. 2). Unlike many experiments where researchers ensure (or assume) cells produce similar amounts of mRNA across conditions [7], this experiment ensured cells produced very different amounts so as to illustrate the merits of absolute quantification (S1 Fig.). Total abundance may vary dramatically in other experimental settings—such as in comparing diseased and normal tissues, tissues at different stages of development, or microbial communities in different environments.

To illustrate the key points of this paper, we worked with positive data only (i.e., we excluded records with any zero or **NA** values): measurements of 3031 components (i.e., mRNAs) at 16 time points. Furthermore, we applied analysis methods (specifically, correlation) to the absolute abundance data *without* transformation (e.g., taking logarithms) because we believe this approach yields useful insights and simplifies the presentation of the central ideas of this paper (see [8] and S1 Supporting Information).

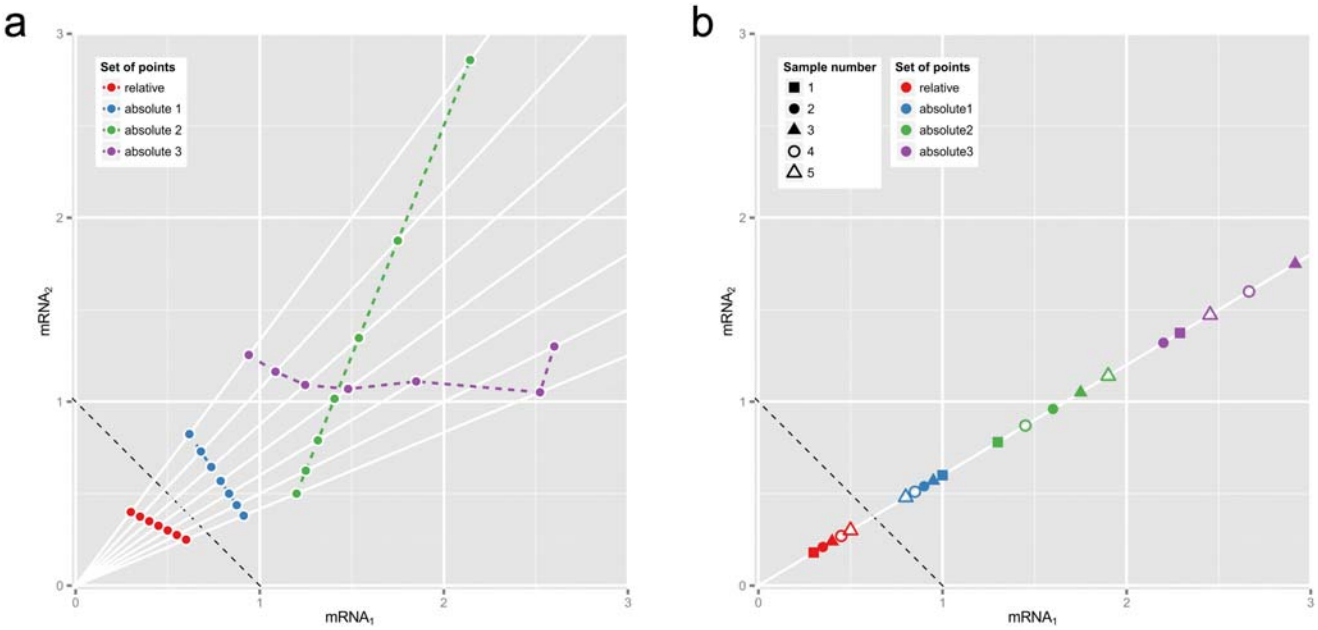

Fig 1. Why correlations between relative abundances tell us absolutely nothing. These plots show two hypothetical mRNAs that are part of a larger total. (a) Seven pairs of relative abundances ($mRNA_1$/total, $mRNA_2$/total) are shown in red, representing the two mRNAs in seven different experimental conditions. The dotted reference line shows ($mRNA_1$ + $mRNA_2$)/total = 1.) Rays from origin through the red points show *absolute* abundances that could have given rise to these relative abundances, e.g., the blue, green or purple sets of points (whose Pearson correlations are –1, +1 and 0.0 respectively). (b) Relative abundances that are proportional must come from equivalent absolute abundances. Here the blue, green or purple sets of point pairs have the same proportionality as the pairs of relative abundances in red, though not necessarily the same order or dispersion.

doi:10.1371/journal.pcbi.1004075.g001

Challenges in interpreting "differential expression"

Before looking at issues with pairs of components, it is important to note that interpreting differences in the relative abundance of a single component can be challenging.

Tests for differential expression are popular for analyzing relative data in bioscience. Much attention has been given to dealing with small numbers of observations and large numbers of tests, but comparatively little to ". . .the commonly believed, though rarely stated, assumption that the absolute amount of total mRNA in each cell is similar across different cell types or experimental perturbations" [7].

The relationship between the relative and absolute abundance of a component can be understood in terms of fold change over time. When total absolute abundance of mRNA stays constant, fold changes in both absolute and relative abundance of each mRNA are equal. When total absolute abundance varies, fold changes in absolute and relative abundances of each mRNA are no longer equal and can change in *different* directions. Between 0 and 3 hours there were 1399 yeast mRNAs whose absolute abundance *decreased*, and whose relative abundance *increased*. Clearly, mRNAs are being expressed differently, but to describe them as "under- or over-expressed" is too simplistic—here lies the interpretation challenge (see S1 Supporting Information).

Correlations between relative abundances tell us absolutely nothing

While "differential expression" of relative abundances is challenging to interpret, in the absence of any other information or assumptions, correlation of relative abundances is just wrong. We

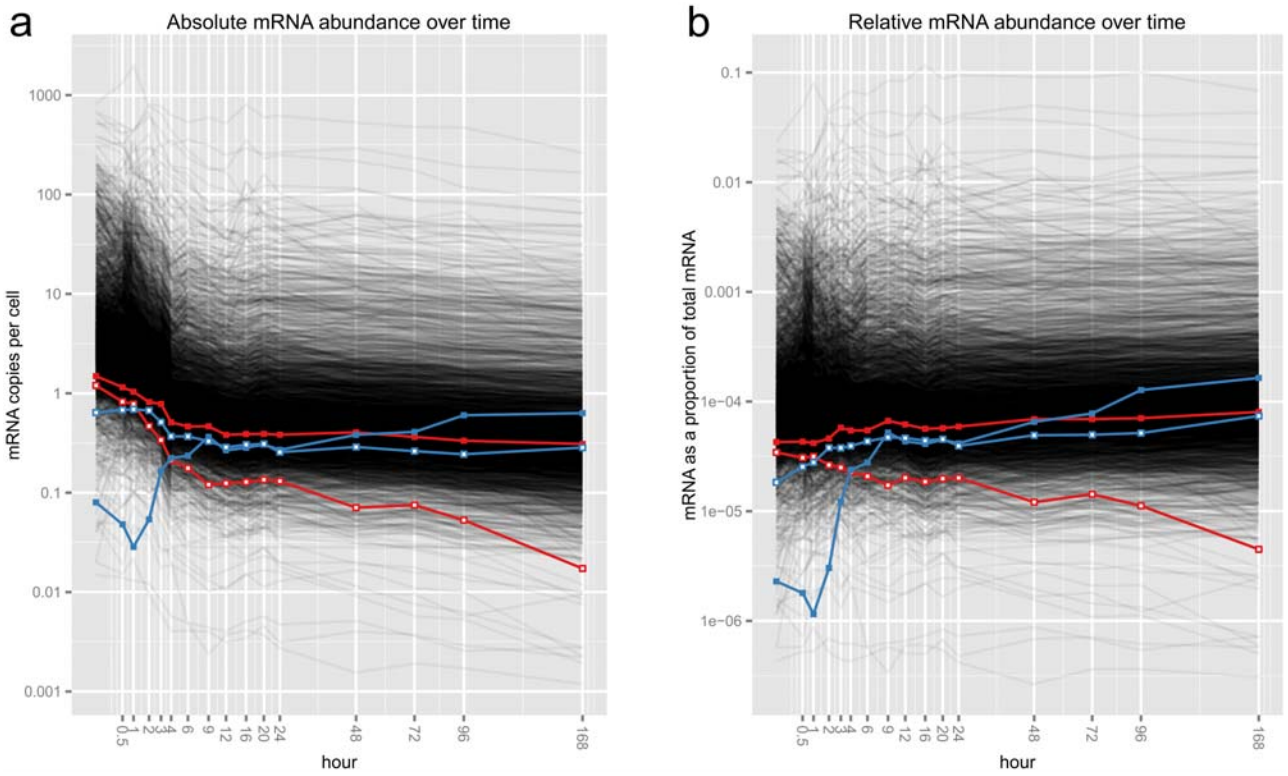

Fig 2. Fission yeast gene expression data of Marguerat *et al.* (a) Absolute and (b) relative abundances of 3031 yeast mRNAs over a 16-point time course. *y*-axes are scaled logarithmically; *x*-axes are on a square-root scale for clarity. Each grey line represents the expression levels of a particular mRNA. The red and blue pairs of mRNAs are discussed later in this paper.

doi:10.1371/journal.pcbi.1004075.g002

stress *in the absence of any other information or assumptions* to highlight the common assumption of constant absolute abundance of total mRNA across all experimental conditions. If this assumption holds, and all the mRNAs comprising that total are considered, the relative abundance of each kind of mRNA will be proportional to its absolute abundance, and analyses of correlation or "differential expression" of the relative values will have clear interpretations. The revisitation of this assumption [7] should raise alarm bells about the inferences drawn from many gene expression studies.

Fig. 1(a) shows why correlation between relative abundances tells us nothing about the relationship between the absolute abundances that gave rise to them: the perfectly correlated relative abundances could come from *any* set of absolute abundance pairs that lie on the rays from the origin. This many-to-one mapping means that other measures of statistical association (e.g., rank correlations or mutual information) will not tell us anything either when applied to purely relative data.

But is this problem just a theoretical construct? A rare issue? Consider the red mRNA pair in Fig. 2: while their *absolute* abundances over time are strongly positively correlated, if someone (inappropriately) used correlation to measure the association between the *relative* abundances of these two mRNAs they would form the opposite view (Fig. 3(a)); correlation between the blue mRNA pair in Fig. 2 is similarly misleading (S2 Fig.). What of the other 4.5 million pairs of mRNAs? Fig. 3(b) summarizes all discrepancies between correlations of absolute abundance, and correlations of relative abundance, showing clearly that the apparent correlations of relative abundances tell a very different story from those of the absolute data. So how *should* we go about analyzing these relative data?

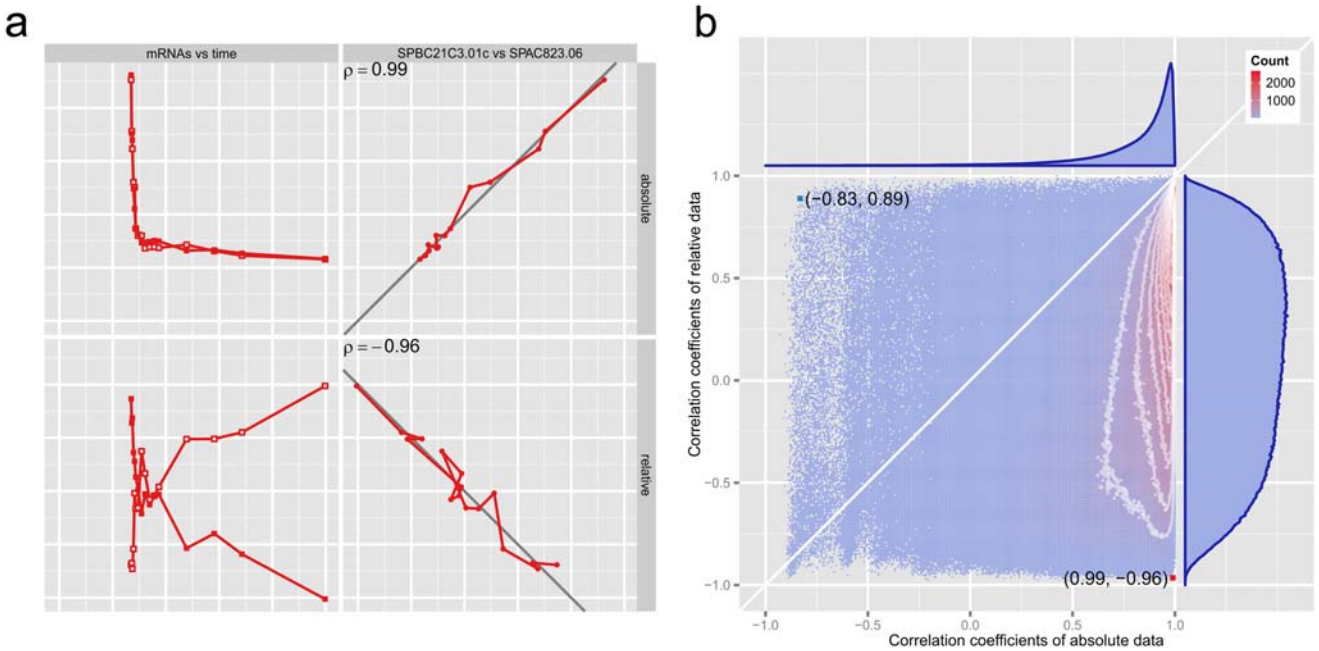

Fig 3. Correlations between relative abundances bear no relationship to the corresponding correlations between absolute abundances. (a) The pair of mRNAs labeled in red in Fig. 2, shown on a linear scale. Values have been scaled and translated to have zero mean and unit variance. Upper panels show absolute abundances; the lower show relative abundances. The left panels show mRNA values over time; the right show the value of one mRNA plotted against the other at each time point. The correlation between the relative abundances is almost the complete opposite of that between the absolute abundances of this pair of mRNAs. (b) 2D histogram of the sample correlation coefficient observed for the relative abundances of a given pair of mRNAs, against the correlation observed for the absolute abundances of that same pair, over all pairs. The red and blue points correspond to the red and blue pairs of mRNA in Fig. 2. White contour lines are shown at intervals of 100 counts. The top marginal histogram shows that the absolute abundances of most pairs are very strongly correlated. The right marginal histogram shows "the negative bias difficulty" [4].

doi:10.1371/journal.pcbi.1004075.g003

Principles for analyzing relative data

CoDA theory provides three principles [4, 9]:

1. Scale invariance: analyses must treat vectors with proportional positive components as representing the same composition (e.g., (2, 3, 4) is equivalent to (20, 30, 40))
2. Subcompositional coherence: inferences about subcompositions (subsets of components) should be consistent, regardless of whether the inference is based on the subcomposition or the full composition.
3. Permutation invariance: the conclusions of analyses must not depend on the order of the components.

Correlation is not subcompositionally coherent: its value depends on which components are considered in the analysis, e.g., if you deplete the most abundant RNAs from a sample [10] and use correlation to measure association between relative abundances, you get different correlations to the undepleted sample (S3 Fig.).

Proportionality is meaningful for relative data

Proportionality obeys all three principles for analyzing relative data. If relative abundances x and y are proportional across experimental conditions i, their *absolute* abundances must be in

proportion:

$$\frac{x_i}{t_i} \propto \frac{y_i}{t_i} \Rightarrow x_i \propto y_i$$

where t_i is the total abundance in condition i (Fig. 1(b)).

We proposed a "goodness-of-fit to proportionality" statistic ϕ to assess the extent to which a pair of random variables (x, y) are proportional [11]. ϕ is related to *logratio variance* [4], var(log(x/y)), and is zero when x and y behave perfectly proportionally. However, when x and y are not proportional, ϕ has both a clear geometric interpretation and a meaningful scale, addressing concerns raised about logratio variance [3]: the closer ϕ is to zero, the stronger the proportionality. We consider "strength" of proportionality (goodness-of-fit) rather than *testing the hypothesis of proportionality* because it allows us to *compare* relationships between different pairs of mRNAs (S1 Supporting Information).

We calculated ϕ for the relative abundances of all pairs of mRNAs and compared it to the correlations between their absolute abundances (S4 Fig.): clearly, the absolute abundances of most mRNA pairs are strongly positively correlated; far fewer are also strongly proportional. Focusing on these strongly proportional mRNAs, we extracted the 424 pairs with $\phi < 0.05$. We graphed the network of relationships between these mRNAs (S5 Fig.), an approach similar to gene co-expression network [12] or weighted gene co-expression analysis [13] but founded on proportionality and therefore valid for relative data. The network revealed one cluster of 96, and many other smaller clusters of mRNAs behaving proportionally across conditions. Using ϕ as a dissimilarity measure, we formed heatmaps of the three largest clusters (S6 and S7 Figs.) similar to the method of Eisen *et al.* [14] but, again, using proportionality not correlation.

Discussion

This paper does not deny pairwise statistical associations between absolute abundances. What it does say is that once all the absolute information has been removed, only a subset of those associations will reliably endure in the remaining relative data, specifically, associations where values behave proportionally across observations.

Other approaches to compositional data in the molecular biosciences

Other researchers have recognized the compositional nature of molecular bioscience data, including [15] as discussed in [16]. Strategies have been proposed to ameliorate spurious correlation in the analysis of relative abundances [2, 3]. We contend that there is no way to salvage a coherent interpretation of correlations from relative abundances without additional information or assumptions; our argument is based on Fig. 1.

ReBoot [2] attempts to establish a null distribution of correlations against which bootstrapped estimates of correlations can be compared. Aitchison articulates problems with this approach [4, p.56–58]. SparCC [3] injects additional information by assuming the number of different components is large and the true correlation network is sparse. This equates to assuming "that the average correlations [between absolute abundances] are small, rather than requiring that any particular correlation be small" [3, Eq.14]. This means the expected value of the total absolute abundance will be constant (as the sum of many independently distributed amounts). We are concerned with situations where that assumption cannot be made, or where the aim is to describe associations between relative amounts.

Caution about correlation

We are also keen to raise awareness that correlation (and other statistical methods that assume measurements come from real coordinate space) should not be applied to relative abundances. This is highly relevant to gene coexpression networks [12]. Correlation is at the heart of methods like Weighted Gene Co-expression Network Analysis [13] and heatmap visualization [14]. These methods are potentially misleading if applied to relative data. This concern extends to methods based on mutual information (e.g., relevance networks [17]) since, as Fig. 1 shows, the bivariate joint distribution of relative abundances (from which mutual information is estimated) can be quite different from the bivariate joint distribution of the absolute abundances that gave rise to them.

Measures of association produce results regardless of the data they are applied to—it is up to the analyst to ensure that the measures are appropriate to the data. Currently, there are many gene co-expression databases available that provide correlation coefficients for the relative expression levels of different genes, generally from multiple experiments with different experimental conditions (see e.g., [18]). As far as we are aware, none of the database providers explicitly address whether absolute levels of gene expression were constant across experimental conditions. If the answer to this question is "no", we would not recommend these correlations be used for the reasons demonstrated in this paper. If the answer is "yes" we still advocate caution in applying correlation to absolute abundances for reasons discussed in S1 Supporting Information.

Results in relation to genome regulation in fission yeast

While the main aim of this study is to present and illustrate principles for analyzing relative abundances, it has also uncovered intriguing biological insight with respect to gene regulation.

The largest cluster of proportionally regulated mRNAs (96 genes, S1 Supporting Information) was highly enriched for mRNAs down-regulated as part of the core environmental stress response [19], including 66 mRNAs that encode ribosomal proteins, and the remaining mRNAs also associated with roles in protein translation, such as ribosome biogenesis, rRNA processing, tRNA methyltransferases and translation elongation factors. The absolute levels of these mRNAs decrease after removal of nitrogen [6]. The notable coherence in biological function among the mRNAs in this cluster is higher than typically seen when correlative similarity metrics for clustering are applied (e.g., [19]). These 96 mRNAs show remarkable proportionality to each other over the entire timecourse (S8 Fig.), and maintain near constant ratios across all conditions (S9 Fig.). Given the huge energy invested by yeast cells for protein translation (most notably ribosome biogenesis [20, 21], it certainly makes sense for cells to synchronize the expression of relevant genes such that translation is finely tuned to nutritional conditions.

Evidently, numerous ribosomal proteins and RNAs function together in the ribosome, demanding their coordinated expression; more surprisingly, multiple other genes, with diverse functions in translation, show equally pronounced proportional regulation across the timecourse. These findings raise intriguing questions as to the molecular mechanisms underlying this proportional regulation, suggesting sophisticated, coordinated control of numerous mRNAs at both transcriptional and post-transcriptional levels of gene expression.

Challenges and future work

While proportionality and the ϕ-statistic provide a valid alternative to correlation for relative data, there are still some challenges in their application. First is the treatment of zeroes, for which there is currently no simple general remedy [22]. Second, and related, is the fact that "many things that we measure and treat as if they are continuous are really discrete count data,

even if only at the molecular extremes" [23] and count data is not purely relative—the count pair (1, 2) carries different information than counts of (1000, 2000) even though the relative amounts of the two components are the same. Correspondence analysis [24], or methods based on count distributions (e.g., logistic regression and other generalized linear models) may provide ways forwards.

Methods

Reproducing this research

All data and code [25] needed to reproduce the analyses and visualizations set out in this paper are contained in the Supporting Information, along with additional illustrations and detailed explanations.

Measuring proportionality

The "goodness-of-fit to proportionality" statistic ϕ can be used to assess the extent to which a pair of random variables (x, y) are proportional [11]. Aitchison [4] proposed *logratio variance*, $\mathrm{var}(\log(x/y))$, as a measure of association for variables that carry only relative information. When x and y are exactly proportional $\mathrm{var}(\log(x/y)) = 0$, but when x and y are not exactly proportional, "it is hard to interpret as it lacks a scale. That is, it is unclear what constitutes a large or small value. . . (does a value of 0.1 indicate strong dependence, weak dependence, or no dependence?)" [3]. Logratio variance can be factored into two more interpretable terms:

$$\begin{aligned}\mathrm{var}(\log(x/y)) &= \mathrm{var}(\log x - \log y) \\ &= \mathrm{var}(\log x) + \mathrm{var}(\log y) - 2\mathrm{cov}(\log x, \log y)\end{aligned} \tag{1}$$

$$\begin{aligned}&= \mathrm{var}(\log x) \cdot \left(1 + \frac{\mathrm{var}(\log y)}{\mathrm{var}(\log x)} - 2\sqrt{\frac{\mathrm{var}(\log y)}{\mathrm{var}(\log x)}}\frac{\mathrm{cov}(\log x, \log y)}{\sqrt{\mathrm{var}(\log x)\mathrm{var}(\log y)}}\right) \\ &= \mathrm{var}(\log x) \cdot (1 + \beta^2 - 2\beta|r|) \\ &\triangleq \mathrm{var}(\log x) \cdot \phi(\log x, \log y)\end{aligned} \tag{2}$$

where β is the *standardized major axis* estimate [26] of slope of random variables $\log y$ on $\log x$, and r the correlation between those variables. The first term in Equation 2, $\mathrm{var}(\log x)$, is solely about the magnitude of variation at play and has nothing to do with y. The second term, ϕ, describes the degree of proportionality between x and y, and forms the basis of our analysis of the relationships between relative values. Other non-negative functions of β and r that are zero when x and y are perfectly proportional could be formed; this is described in more detail in S1 Supporting Information, as well as why ϕ is preferable to an hypothesis testing approach. There is no need to calculate β or r to assess strength of proportionality; they simply provide a clear geometric interpretation of ϕ; in practice, one can use the relationship $\phi(\log x, \log y) = \mathrm{var}(\log(x/y))/\mathrm{var}(\log x)$.

Alternative measures of proportionality

The ϕ statistic is a measure of goodness-of-fit to proportionality that combines two quantities of interest: β, the slope of the line best describing the relationship between random variables $\log x$ and $\log y$; and r, whose magnitude estimates the strength of the linear relationship between $\log x$ and $\log y$. "Goodness-of-fit" describes how well a statistical model fits a set of observations and is a familiar concept in regression, including linear and generalised linear

models, but note that ϕ—specifically the slope (β) of the standardized major axis—is motivated by *allometry* rather than regression modeling. We are interested in assessing whether two variables are directly proportional, rather than *predicting* one from the other: "use of regression would often lead to an incorrect conclusion about whether two variables are isometric or not" [26, p.265]. Note also that ordinary least squares regression fits are not symmetric: in general, the slope of y regressed on x is different to the slope of x regressed on y [27].

While goodness-of-fit measures for regression may not generally be appropriate for assessing proportionality, Zheng [28] explores the *concordance correlation coefficient* ρ_c [29] which could be modified to provide an alternative measure of proportionality defined as

$$\rho_p(\log x,\log y) \triangleq \frac{2\mathrm{cov}(\log x,\log y)}{\mathrm{var}(\log x)+\mathrm{var}(\log y)}$$

and related to var(log(x/y)) by the terms in Equation 1. This "proportionality correlation coefficient" ranges from −1 (perfect reciprocality) to +1 (perfect proportionality) and lacks the clear geometric interpretation of ϕ.

Centered logratio (clr) representation

We have used $\phi(\log x, \log y)$ to emphasize the relationship between ϕ and logratio variance. However to ensure that the ϕ values for component pair (i, j) are on the same scale (i.e., comparable to) the ϕ values for component pair (m, n), it is necessary to use the *centered logratio* (clr) transformation instead of just the logarithm (S1 Supporting Information). The clr representation of composition $\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_D)$ is the logarithm of the components after dividing by the geometric mean of $\mathbf{x}$:

$$\mathrm{clr}(\mathbf{x}) = \left(\log\frac{x_1}{g_m(\mathbf{x})}, \cdots, \log\frac{x_i}{g_m(\mathbf{x})}, \cdots, \log\frac{x_D}{g_m(\mathbf{x})},\right)$$

ensuring that the sum of the elements of clr($\mathbf{x}$) is zero. Note that dividing all components in a composition by a constant (i.e., the geometric mean $g_m(\mathbf{x})$) does not alter the *ratios* of components.

Using ϕ to form co-expression networks and clustered heatmaps

Gene co-expression networks [12, 13] are generally based on a pairwise distance or dissimilarity matrix which is often a function of correlation and thus not appropriate for relative data. Proportionality is appropriate, but ϕ does not satisfy the properties of a *distance*—most obviously, it is not symmetric unless $\beta = 1$:

$$\begin{aligned}\phi(\log x,\ \log y) &= 1+\beta^2-2\beta|r|\\ \phi(\log y,\ \log x) &= 1+\frac{1}{\beta^2}-2\frac{1}{\beta}|r|.\end{aligned}$$

We are most interested in pairs of variables where β and r are near 1 and want to preserve the link between $\phi(\log x, \log y)$, β and r. Hence, our approach to forming a dissimilarity matrix is simply to work with $\phi(\log x_i, \log x_j)$ where $i < j$, in effect, the lower triangle of the matrix of ϕ values between all pairs of components. This symmetrised form of ϕ was then used to lay out a network of the 145 mRNAs that were involved in 424 pairwise relationships with $\phi < 0.05$. We used the symmetrised form of ϕ as the basis of the cluster analysis and heatmap expression pattern display (e.g., S10 Fig.) described by Eisen *et al.* [14].

Supporting Information

S1 Fig. Total abundance of yeast mRNAs in copies per cell over the 16-point time course. Times 0 and 3 are highlighted for further study.
(EPS)

S2 Fig. The pair of mRNAs labeled in blue in Fig. 2, shown on a linear scale. Values have been scaled and translated to have zero mean and unit variance. Upper panels show absolute abundances; the lower show relative abundances. The left panels show mRNA values over time; the right show the value of one mRNA plotted against the other at each time point. As with Fig. 3, the correlation between the relative abundances is almost the complete opposite of that between the absolute abundances of this pair of mRNAs.
(EPS)

S3 Fig. A 2D histogram of the correlation coefficient observed for the relative abundances of a given pair of mRNAs in a sample where the ten most abundant mRNAs have been removed, against the correlation coefficient observed for the relative abundances of that same pair, over all pairs. White contour lines are shown at intervals of 100 counts. While the distribution of the correlation coefficient pairs lies more on the diagonal than in the preceding figure, it is clear that correlation of relative abundances is sensitive to what is in (or out of) the total, i.e., correlation is *not* subcompositionally coherent.
(TIFF)

S4 Fig. A 2D histogram of $\phi(\mathrm{clr}(x_i), \mathrm{clr}(x_j))$ for the relative abundances of a given pair (i, j) of mRNAs, against the correlation coefficient observed for the absolute abundances of that same pair, over all pairs. The red and blue points correspond to the red and blue pairs of mRNA in Fig. 2. White contour lines are shown at intervals of 100 counts and the top marginal histogram is the same as in S2(b) Fig. The few mRNA pairs that are strongly proportional (within the red rectangle) are also strongly positively correlated. However, the converse is not true: strong positive correlation between mRNAs does not imply that they are strongly proportional.
(TIFF)

S5 Fig. A graph of the proportionality relationships between the 424 pairs of mRNAs with $\phi(\mathrm{clr}(x_i), \mathrm{clr}(x_j)) < 0.05$.
(EPS)

S6 Fig. Heatmap visualisation of the 96 mRNA cluster seen in S5 Fig.
(EPS)

S7 Fig. Heatmap visualisation of two smaller mRNA clusters seen in S5 Fig.
(EPS)

S8 Fig. The relative abundances of each of the mRNAs from the 96 mRNA cluster seen in S5 Fig. over time. The geometric mean at each timepoint is shown in blue.
(EPS)

S9 Fig. Each of the mRNAs from the 96 mRNA cluster seen in S5 Fig. divided by the geometric mean of the mRNAs at each timepoint.
(EPS)

S10 Fig. Heatmap visualisation of the 66 pairs of mRNAs with $\phi(\mathrm{clr}(x_i), \mathrm{clr}(x_j)) < 0.025$. The hierarchical clustering of these components is cut into six colour-coded groups, shown at the left edge of the heatmap.
(EPS)

S1 Supporting Information. The detailed and reproducible analysis reported in this paper. This PDF file is the output obtained by executing SupplementaryInfo.Rnw from S2 Supporting Information. In addition to all the figures and results in the manuscript it provides additional detail and information for those interested in understanding more about compositional data analysis and the analyses we have conducted.
(PDF)

S2 Supporting Information. R code and data to reproduce this paper's analysis. This Zip file contains SupplementaryInfo.Rnw, the Sweave source which is executed to analyse the contents of the ./data folder and present the results in S1 Supporting Information.
(ZIP)

Author Contributions

Conceived and designed the experiments: JB SM. Performed the experiments: JB SM. Analyzed the data: DL JJE VPG. Wrote the paper: DL VPG JJE SM JB. Developed the data analysis method: DL VPG JJE.

References

1. van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D, et al. (2003) Monitoring global messenger RNA changes in externally controlled microarray experiments. EMBO Reports 4: 387–393. doi: 10.1038/sj.embor.embor798 PMID: 12671682
2. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, et al. (2012) Microbial co-occurrence relationships in the human microbiome. PLoS Comput Biol 8: e1002606. doi: 10.1371/journal.pcbi.1002606 PMID: 22807668
3. Friedman J, Alm EJ (2012) Inferring correlation networks from genomic survey data. PLoS Comput Biol 8: e1002687. doi: 10.1371/journal.pcbi.1002687 PMID: 23028285
4. Aitchison J (1986) The statistical analysis of compositional data. Chapman & Hall, Ltd. doi: 10.1007/978-94-009-4109-0
5. 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. doi: 10.1098/rspl.1896.0076
6. Marguerat S, Schmidt A, Codlin S, Chen W, Aebersold R, et al. (2012) Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. Cell 151: 671–683. doi: 10.1016/j.cell.2012.09.019 PMID: 23101633
7. Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, et al. (2012) Revisiting global gene expression analysis. Cell 151: 476–482. doi: 10.1016/j.cell.2012.10.012 PMID: 23101621
8. Pawlowsky-Glahn V, Egozcue JJ, Lovell DR (2014) Tools for compositional data with a total. Statistical Modelling. doi: 10.1177/1471082X14535526
9. Egozcue JJ, Pawlowsky-Glahn V (2011) Basic concepts and procedures. In: Pawlowsky-Glahn V, Buccianti A, editors, Compositional Data Analysis: Theory and Applications, Chichester, UK: John Wiley & Sons, Ltd. pp. 12–27.
10. O'Neil D, Glowatz H, Schlumpberger M (2001) Ribosomal RNA depletion for efficient use of RNASeq capacity. In: Current Protocols in Molecular Biology, John Wiley & Sons, Inc. doi: 10.1371/journal.pcbi.1002687
11. Lovell D, Pawlowsky-Glahn V, Egozcue JJ (2013) Have you got things in proportion? a practical strategy for exploring association in high-dimensional compositions. In: Hron K, Filzmoser P, Templ M, editors, Proceedings of the 5th International Workshop on Compositional Data Analysis. Vorau, Austria, pp. 100–110.
12. López-Kleine L, Leal L, López C (2013) Biostatistical approaches for the reconstruction of gene co-expression networks based on transcriptomic data. Briefings in Functional Genomics 12: 457–467. doi: 10.1093/bfgp/elt003 PMID: 23407269
13. Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology 4. doi: 10.2202/1544-6115.1128 PMID: 16646834

14. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95: 14863–14868. doi: 10.1073/pnas.95.25.14863

15. Vencio R, Varuzza L, de B Pereira C, Brentani H, Shmulevich I (2007) Simcluster: clustering enumeration gene expression data on the simplex space. BMC Bioinformatics 8: 246. doi: 10.1186/1471-2105-8-246 PMID: 17625017

16. Lovell D, Müller W, Taylor J, Zwart A, Helliwell C (2010) Caution! compositions! technical report and companion software (publication—technical). Technical Report EP10994, CSIRO.

17. Butte AJ, Kohane IS (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In: Pacific Symposium on Biocomputing, Stanford University, volume 5. pp. 418–429. PMID: 10902190

18. Obayashi T, Kinoshita K (2011) COXPRESdb: a database to compare gene coexpression in seven model animals. Nucleic Acids Research 39: D1016–D1022. doi: 10.1093/nar/gkq1147 PMID: 21081562

19. Chen D, Toone WM, Mata J, Lyne R, Burns G, et al. (2003) Global transcriptional responses of fission yeast to environmental stress. Molecular Biology of the Cell 14: 214–229. doi: 10.1091/mbc.E02-08-0499 PMID: 12529438

20. Warner JR (1999) The economics of ribosome biosynthesis in yeast. Trends in Biochemical Sciences 24: 437–440. doi: 10.1016/S0968-0004(99)01460-7 PMID: 10542411

21. López-Maury L, Marguerat S, Bähler J (2008) Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. Nature Reviews Genetics 9: 583–593. doi: 10.1038/nrg2398 PMID: 18591982

22. Martín-Fernández JA, Palarea-Albaladejo J, Olea RA (2011) Dealing with zeros. In: Pawlowsky-Glahn V, Buccianti A, editors, Compositional Data Analysis: Theory and Applications, Chichester, UK: John Wiley & Sons, Ltd. pp. 43–58. doi: 10.1002/9781119976462.ch4

23. Bacon-Shone J (2008) Discrete and continuous compositions. In: Daunis-i Estadella J, Martín-Fernández J, editors, Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop. University of Girona.

24. Greenacre M (2011) Compositional data and correspondence analysis. In: Pawlowsky-Glahn V, Buccianti A, editors, Compositional Data Analysis: Theory and Applications, Chichester, UK: John Wiley & Sons, Ltd. pp. 104–113. doi: 10.1002/9781119976462.ch8

25. Ince DC, Hatton L, Graham-Cumming J (2012) The case for open computer programs. Nature 482: 485–488. doi: 10.1038/nature10836 PMID: 22358837

26. Warton DI, Wright IJ, Falster DS, Westoby M (2006) Bivariate line-fitting methods for allometry. Biological Reviews 81: 259–291. doi: 10.1017/S1464793106007007 PMID: 16573844

27. Draper N, Smith H (1998) Applied Regression Analysis. New York: Wiley-Interscience, third edition edition. doi: /10.1002/9781118625590

28. Zheng B (2000) Summarizing the goodness of fit of generalized linear models for longitudinal data. Statistics in Medicine 19: 1265–1275. doi: 10.1002/(SICI)1097-0258(20000530)19:10%3C1265::AID-SIM486%3E3.0.CO;2-U PMID: 10814976

29. Lin LIK (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45: 255–268. doi: 10.2307/2532051 PMID: 2720055