nature methods

MATTERS ARISING

https://doi.org/10.1038/s41592-020-01006-1

Check for updates

Examining microbe–metabolite correlations by linear methods

Thomas P. Quinn [1]✉ and Ionas Erb [2]

ARISING FROM J. T. Morton et al. *Nature Methods* https://doi.org/10.1038/s41592-019-0616-3 (2019)

Analyzing correlative relationships between microbes and metabolites is a timely topic[1–3] but is complicated by the compositional (i.e., relative) nature of the data[4,5]. Recently, Morton et al. proposed a neural network architecture called MMvec to predict metabolite abundances from microbe presence[6]. We do not doubt the usefulness of MMvec but write in defense of simple linear statistics. When used correctly, correlation and proportionality[5,7] can be scale invariant and can outperform MMvec in certain conditions.

Scale invariance is important because we do not want a method that is sensitive to (variant with) changes in technical factors such as sequencing depth (differences in scale). In compositional data analysis, scale invariance is forced by using a log-ratio transformation that normalizes the data with an internal reference[8]. The resulting log ratios are scale invariant, and so analyses of log ratios are scale invariant too. This is also true for multi-omics data, but only if the transformation is performed correctly. Let us consider two possible centered log-ratio (CLR) transformations of multi-omics data, presented here as functions of the input

$$\begin{aligned}\mathcal{A}(\mathbf{u}_i, \mathbf{v}_i) &= \mathrm{clr}([u_{i1}, \ldots, u_{iM}, v_{i1}, \ldots, v_{iN}]) \\ &= \log\left(\frac{[u_{i1}, \ldots, u_{iM}, v_{i1}, \ldots, v_{iN}]}{\sqrt{[M+N]\Pi_j^M u_{ij}\Pi_j^N v_{ij}}}\right) \\ \mathcal{B}(\mathbf{u}_i, \mathbf{v}_i) &= [\mathrm{clr}([u_{i1}, \ldots, u_{iM}]), \mathrm{clr}([v_{i1}, \ldots, v_{iN}])] \\ &= \left[\log\left(\frac{[u_{i1}, \ldots, u_{iM}]}{\sqrt{[M]\Pi_j^M u_{ij}}}\right), \log\left(\frac{[v_{i1}, \ldots, v_{iN}]}{\sqrt{[N]\Pi_j^N v_{ij}}}\right)\right]\end{aligned}$$

for sample i, where $\mathbf{u}_i$ measures 1, …, M microbes and $\mathbf{v}_i$ measures 1, …, N metabolites. Only approach $\mathcal{B}$ is scale invariant. Morton et al. use approach $\mathcal{A}$ in the original paper where they claim that correlation and proportionality underperform MMvec.

Why is approach $\mathcal{B}$ valid, but not approach $\mathcal{A}$? It is because the microbe and metabolite data are generated from two separate sampling processes: they are individually, not jointly, constrained to sum to 1. In other words, the abundance of microbe 1 is limited by the abundance of microbes 2 through M, but is not limited by the abundance of metabolites 1 through N. Consequently, the denominator from approach $\mathcal{A}$ has no meaning. In contrast, the denominators from approach $\mathcal{B}$ have the property that they cancel any constant factor multiplied with their respective numerators. As such, they cancel the implicit sequencing biases that cause the samples to be on different scales. An additional property of these denominators is that they are useful normalization factors themselves[9]: under the assumption that the majority of features are unchanged, approach $\mathcal{B}$ will make the transformed data proportional to the original absolute data and thus performs effective library-size normalization.

We repeated the authors' analysis to measure the F1 score (precision and recall) for the top microbe–metabolite associations using approach $\mathcal{B}$. Figure 1 shows the performance of correlation and proportionality, both of which outperformed MMvec on their simulated benchmark. Interestingly, correlation performed best, suggesting that the 'ground truth' includes power-law relationships between microbes and metabolites (log–linear relationships with slopes other than 1, which could mean, for example, that although an increase in two microbe units associates with a doubling of metabolites, an increase of four units associates with a quadrupling). Because ϕ and ρ are designed for intercept-free linear relationships, these power-law relationships will usually go undetected. Note that, although SPIEC-EASI already implements $\mathcal{B}$ in 'multi-source' mode, it makes a strong assumption that the true ecological association network is sparse[10]. This assumption does not appear to hold true for the simulated data (see ref. [6]). If one instead calculates covariance via a second inversion of the regularized inverse covariance matrix, the model performs well (see QUIC-cov in Fig. 1).

Data sparsity, by which we mean an excess of zero counts, presents a major challenge to microbiome data analysis. For one, a log-ratio transformation fails for a zero entry. Many methods have been proposed to address compositional zeros, including Bayesian imputation strategies[11] and alternative transformations[12]. The simplest approach involves replacing all zeros with a very small number. Every zero-handling strategy has limitations; however, it remains unclear whether a neural network will necessarily perform better. For the simulated microbiome data used in Fig. 1, about 14% of the values are zero (the data are 14% sparse). We increased the sparsity by sampling new counts from an equivalent multinomial distribution where we used the closed counts as parameters at one-twentieth the sequencing depth. This sampling generated new relative data with 71% sparsity, without any change to the corresponding absolute data. Figure 2 shows how simple correlations at 71% sparsity, despite a considerable drop in accuracy, still outperform the MMvec baseline at 14% sparsity. Interestingly, Spearman's rank correlation is the method most impaired by data sparsity, likely because any change to small counts would distort ranks more than parametric covariance estimates.

It is worth noting that neither the precision nor the recall is high for any of these methods. This is consistent with how information is lost when producing compositional counts, especially if undersampling leads to an excess of zero counts. It is also worth noting that CLR-based correlations, by definition, describe how microbes

[1]Applied Artificial Intelligence Institute, Deakin University, Geelong, Victoria, Australia. [2]Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Spain. ✉e-mail: contacttomquinn@gmail.com

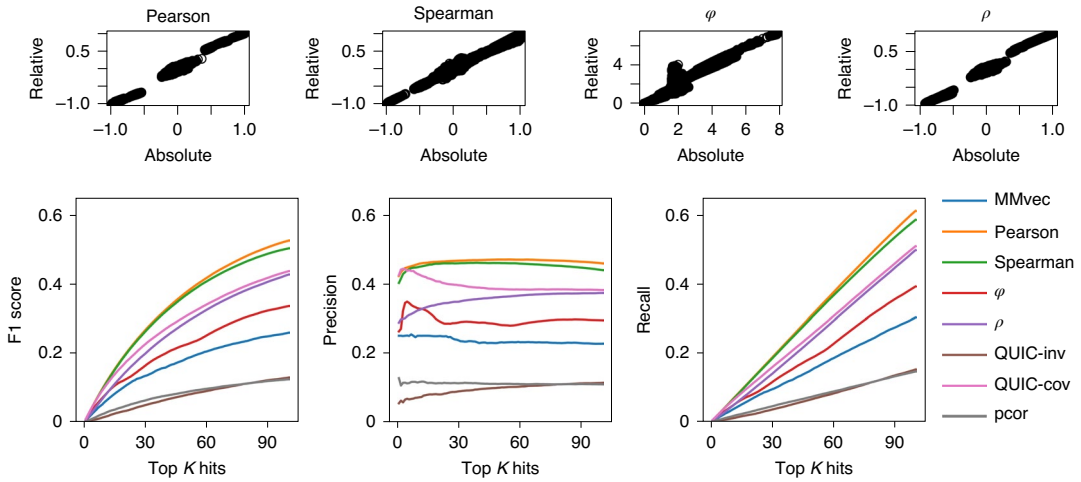

Fig. 1 | Reanalysis of the simulated data from Morton et al. The top panels show agreement between absolute and relative metrics when using approach $\mathcal{B}$. The bottom panels show the updated performances from the simulated data, where QUIC refers to the regularized inverse covariance matrix and pcor refers to partial correlations.

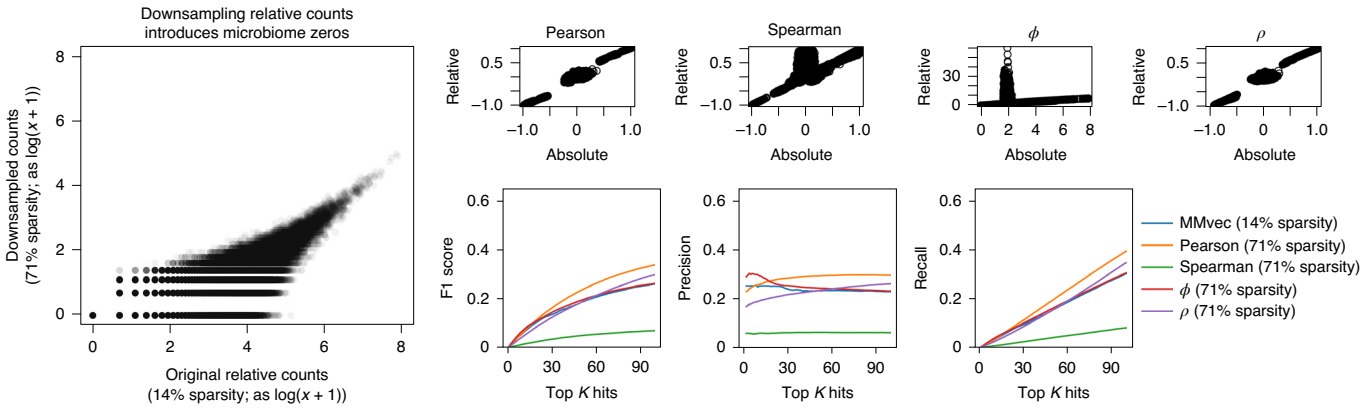

Fig. 2 | Reanalysis of the sparsified simulated data. The leftmost panel shows a log–log plot of the new relative data with 71% sparsity (*y* axis) versus the original relative data with 14% sparsity (*x* axis), confirming successful downsampling. The top right panels show agreement between absolute and relative metrics when using approach $\mathcal{B}$. The bottom right panels show the updated performances from the sparse simulated data.

and metabolites behave relative to their respective sample means. Although the CLR can, under some circumstances, provide a useful normalization of the data, analysts must take care not to forget that the geometric mean is foremost a reference frame[13], a kind of yardstick against which to compare the relative abundances to establish a scale-invariant analysis of the data. If the CLR transform does not perfectly normalize the relative data, then some discrepancies might be seen between the estimates and the true associations[7].

Even when the CLR is not a perfect normalization tool, proportionality is designed to still reveal some linear associations without having to make the relationship between the variables and the reference explicit. On the other hand, CLR-based correlations depend more on the chosen reference because any expression of power laws will necessarily involve that reference. To visualize this, we assumed that the correlation coefficient was high enough to detect a linear relationship between the logarithms of two features $\mathbf{x}$ and $\mathbf{y}$, both having the same reference $\mathbf{r}$. We have the log–linear model

$$\log \frac{\mathbf{y}}{\mathbf{r}} = m \log \frac{\mathbf{x}}{\mathbf{r}} + b + \epsilon$$

(with offset b and error term ϵ). This implies that $\mathbf{y} = e^{b+\epsilon}\mathbf{r}^{1-m}\mathbf{x}^{m}$. From this, we can see how the reference (geometric mean from CLR) influences the relationship between variables when the slope m is not 1.

We do not disagree that neural networks can add value to multi-omics data integration. Their ability to learn nonlinear relationships could improve metabolite prediction by directly modeling complex microbe–metabolite interactions[14]. However, neural networks do not offer a magical solution to the problems of compositional data analysis[15]. They are merely a nested series of transformed linear operators. As such, they may be prone to yielding spurious results whenever a simple linear method would yield spurious results. It seems to us that MMvec's primary advantage is how it handles compositional data, not its neural network architecture per se. For example, the use of a softmax transformation, which is equivalent to an inverse CLR transformation, might imply that the linear operations from previous layers actually occur in CLR coordinates[6].

We conclude by reminding readers that not all problems in biology are solved by adding layers of complexity: sometimes it is sufficient to use the simplest solutions more carefully.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-01006-1.

Received: 17 November 2019; Accepted: 27 October 2020;
Published online: 04 January 2021

References

1. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
2. Tang, Z.-Z. et al. Multi-omic analysis of the microbiome and metabolome in healthy subjects reveals microbiome-dependent relationships between diet and metabolites. *Front. Genet.* **10**, 454 (2019).
3. Yachida, S. et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* **25**, 968–976 (2019).
4. Fernandes, A. D. et al. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**, 15 (2014).
5. Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S. & Bähler, J. Proportionality: a valid alternative to correlation for relative data. *PLoS Comput. Biol.* **11**, e1004075 (2015).
6. Morton, J. T. et al. Learning representations of microbe–metabolite interactions. *Nat. Methods* **16**, 1306–1314 (2019).
7. Erb, I. & Notredame, C. How should we measure proportionality on relative gene expression data? *Theory Biosci.* **135**, 21–36 (2016).
8. Aitchison, J. *The Statistical Analysis of Compositional Data* (Chapman & Hall, 1986).
9. Quinn, T. P., Erb, I., Richardson, M. F. & Crowley, T. M. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics* **34**, 2870–2878 (2018).
10. Kurtz, Z. D. et al. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, e1004226 (2015).
11. Palarea-Albaladejo, J. & Martín-Fernández, J. A. zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics Intel. Lab. Syst.* **143**, 85–96 (2015).
12. Martino, C. et al. A novel sparse compositional technique reveals microbial perturbations. *mSystems* **4**, e00016-19 (2019).
13. Morton, J. T. et al. Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* **10**, 2719 (2019).
14. Le, V., Quinn, T. P., Tran, T. & Venkatesh, S. Deep in the bowel: highly interpretable neural encoder–decoder networks predict gut metabolites from gut microbiome. *BMC Genomics* **21**, 256 (2020).
15. Delgado, R. T., Talebi, H., Khodadadzadeh, M. & van den Boogaart, K. G. On machine learning algorithms and compositional data. in *CoDaWork2019: Proc. 8th Intl Workshop on Compositional Data Analysis* (2019).

Methods

For Fig. 1, we performed a reanalysis of the simulated data by taking the following steps: (1) we loaded in the absolute and relative datasets provided by the authors in the 'results/benchmark_output/CF_sims/data' directory; (2) we replaced all zeros with the minimum non-zero value; (3) we performed a CLR of the microbe and metabolite data separately for each of the absolute and relative datasets; (4) we calculated proportionality (using propr package version 4.2.8) and correlation (using base R version 3.6.3) for each of the absolute and relative datasets; and (5) we measured and plotted the precision and recall of the relative data analysis against the MMvec results using a Python script from the authors. For Fig. 2, we repeated this same procedure but added a new step where we downsampled the relative microbiome data by using a multinomial distribution at one-twentieth the sequencing depth, where the expected proportions were set as the original relative microbiome proportions.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data used in Figs. 1 and 2 are available from https://doi.org/10.5281/zenodo.3610709 and https://doi.org/10.5281/zenodo.3833174, respectively.

Code availability

Scripts used in Figs. 1 and 2 are available from https://doi.org/10.5281/zenodo.3610709 and https://doi.org/10.5281/zenodo.3833174, respectively.

Acknowledgements

I.E. has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825835 (BovReg), Secretaria de Universidades e Investigación del Departamento de Economía y Conocimiento de la Generalidad de Cataluña, 2017 SGR 447 (SGR), Agencia Estatal de Investigación (AEI) and FEDER under Project BFU2017-88264-P (Plan Estatal). I.E. also acknowledges the following CRG funding sources: support of the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership, Centro de Excelencia Severo Ochoa, the CERCA Programme / Generalitat de Catalunya and the European Regional Development Fund (ERDF).

Author contributions

T.P.Q. performed the analysis. T.P.Q. and I.E. drafted the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-01006-1.

Correspondence and requests for materials should be addressed to T.P.Q.

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

nature research

Corresponding author(s): Thomas P. Quinn

Last updated by author(s): Sep 9, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	Data were analyzed in R version 3.6.3 using the software package propr version 4.2.8.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data and scripts used in Figure 1 and Figure 2 are available from https://doi.org/10.5281/zenodo.3610709 and https://doi.org/10.5281/zenodo.3833174, respectively.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study uses public data from https://doi.org/10.1038/s41592-019-0616-3. All available samples were used. Rationale for sample size is provided by the original authors.
Data exclusions	This study uses public data from https://doi.org/10.1038/s41592-019-0616-3. No further samples were excluded. Rationale for data exclusion is provided by the original authors.
Replication	This study uses public data from https://doi.org/10.1038/s41592-019-0616-3. The findings were replicated by simulations.
Randomization	This study uses public data from https://doi.org/10.1038/s41592-019-0616-3. Randomization was determined by the original authors.
Blinding	This study uses public data from https://doi.org/10.1038/s41592-019-0616-3. Blinding was determined by the original authors.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging