Lawrence Berkeley National Laboratory
Recent Work

Title
Reply to: Examining microbe-metabolite correlations by linear methods.

Permalink

Authors
Morton, James T
McDonald, Daniel
Aksenov, Alexander A
et al.

Peer reviewed

Check for updates

Reply to: Examining microbe–metabolite correlations by linear methods

James T. Morton[1,2], **Daniel McDonald**[1,3], **Alexander A. Aksenov**[4,5], **Louis Felix Nothias**[4,5], **James R. Foulds**[6], **Robert A. Quinn**[7], **Michelle H. Badri**[8], **Tami L. Swenson**[9], **Marc W. Van Goethem**[9], **Trent R. Northen**[9,10], **Yoshiki Vazquez-Baeza**[3,11], **Mingxun Wang**[4,5], **Nicholas A. Bokulich**[12,13], **Aaron Watters**[14], **Se Jin Song**[1,3], **Richard Bonneau**[8,14,15,16], **Pieter C. Dorrestein**[4,5] and **Rob Knight**[1,2,17,18] ✉

Quinn and Erb[1] propose to apply a centered log-ratio (CLR) transform before performing correlation analysis and make the case that, when used correctly, correlation and proportionality can outperform MMvec in identifying microbe–metabolite interactions. While this may be an appealing strategy, it is important to note that the correlations estimated from CLR-transformed data will have a fundamentally different interpretation than the true correlations in the environment, namely:

$$\mathrm{Cov}\left(x_i, y_j\right) \neq \mathrm{Cov}\left(\mathrm{clr}(x)_i, \mathrm{clr}(y)_j\right)$$

where x_i and y_j are the absolute abundances for microbe abundances x and metabolite abundances y in taxon i and metabolite j. Because the absolute abundances are often not available, inferring the true correlations between microbes and metabolites is not tractable (Supplementary Note 1). This phenomenon has been extensively studied in refs. [2–4], and one of our recent studies provides the intuition behind this in the case of differential abundance[5]. Because of this discrepancy, we proposed to use co-occurrence probabilities instead of correlation.

We relied on simulated data in the original paper[6] as an artificial ground truth, as is common in the evaluation of omics tools. However, simulated data will always have limitations because of the inability to model unknown features of the real system or because of deliberate simplifications that clarify key points in the model system. Furthermore, it is possible to identify simulations where a proposed model is optimal. In Fig. 1, we used Bayesian Optimization[7] to identify simulations where MMvec was able to accurately estimate the correct parameters and Pearson underperformed. If the appropriate assumptions are satisfied, MMvec can correctly estimate the co-occurrence probabilities with machine precision.

Therefore, a crucial aspect of the MMvec manuscript was to test performance both on simulations and on real data. Performance on real data is the ultimate test of methods, and we recommend that simulated datasets be complemented with experimentally validated datasets where possible. Accordingly, we applied the same proportionality-based scripts described by Quinn and Erb[1] and evaluated them on one of the real datasets we used in the MMvec paper.

A major obstacle to analyzing real-world microbiome and metabolomics data is sparsity. Traditional compositional methods such as the proposed CLR transform cannot automatically deal with zeros and require imputation as a preprocessing step. This imputation adds bias and is impractical for the sparse datasets typically encountered[8,9]. Microbiome and untargeted metabolomics datasets are generally sparse: in large studies, such as the American Gut Project[10], the sparsity for stool samples alone is 99.946%. MMvec was designed to handle sparse data. In the desert biocrust soils dataset (sparsity of 51%; ref. [11]) that was used in the MMvec publication, we observe that MMvec dramatically outperformed the newly proposed linear methods (Fig. 2).

Contrary to the argument by Quinn and Erb[1] regarding the complexity of neural networks, the MMvec model[6] is not much more complex than the proposed regression techniques. It is a simple one-layer neural network, which is in effect a two-stage log–bilinear regression.

Methods similar to MMvec have been successful at the task of learning word co-occurrences. Since Mikolov et al.[12], these models have been designed with an emphasis on practical methods for learning useful word representations at scale, rather than on perfectly modeling the data distribution.

MMvec is only one tool in the arsenal of correlative methods. It is not perfect for every correlation type or dataset and is not a one-size-fits-all solution. However, we have found that MMvec is a powerful discovery tool, as demonstrated by the other real datasets

[1]Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA. [2]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA. [3]Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA, USA. [4]Collaborative Mass Spectrometry Innovation Center, University of California, San Diego, La Jolla, CA, USA. [5]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA, USA. [6]Department of Information Systems, University of Maryland–Baltimore County, Baltimore, MD, USA. [7]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA. [8]Department of Biology, New York University, New York, NY, USA. [9]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [10]DOE Joint Genome Institute, Walnut Creek, CA, USA. [11]Jacobs School of Engineering, University of California, San Diego, La Jolla, CA, USA. [12]The Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA. [13]Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA. [14]Flatiron Institute, Simons Foundation, New York, NY, USA. [15]Department of Computer Science, Courant Institute, New York, NY, USA. [16]Center for Data Science, New York University, New York, NY, USA. [17]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA. [18]Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA, USA. ✉e-mail: rknight@ucsd.edu
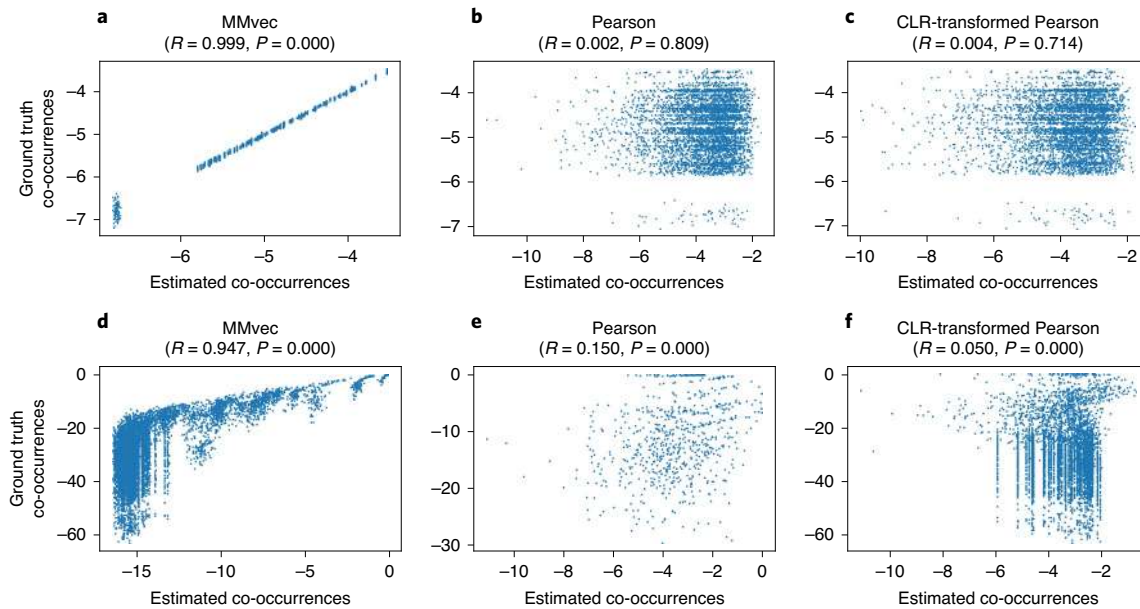
Fig. 1 | A simulation benchmark comparing MMvec to Pearson. Simulations were obtained through Bayesian Optimization[7] to showcase scenarios where MMvec outperforms Pearson. **a–c**, Simulation of a scenario where the microbiome dataset is 99% dense. **d–f**, Simulation of a scenario where the microbiome dataset is 60% dense. All axes are represented on a log scale. Pearson's R is used to measure the agreement between the simulated ground truth co-occurrences and the estimated co-occurrences.
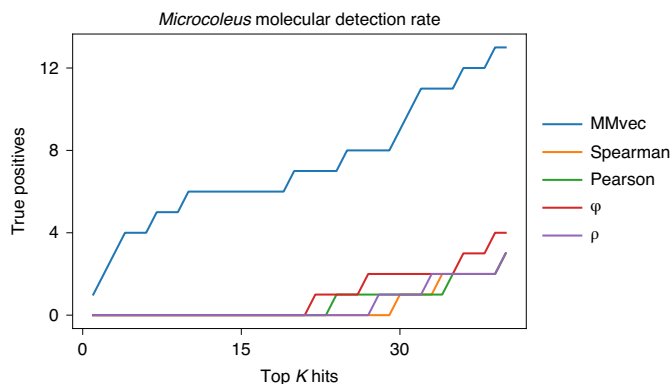

Fig. 2 | Biocrust soils benchmark. A comparison of MMvec to metrics proposed by Quinn and Erb[1]. These proposed metrics include Spearman, Pearson, φ and ρ applied after a CLR transformation[13].

References

1. Quinn, T. P. & Erb, I. Examining microbe–metabolite correlations by linear methods. *Nat. Methods* https://doi.org/10.1038/s41592-020-01006-1 (2020).
2. Aitchison, J. A concise guide to compositional data analysis. http://www.leg.ufpr.br/lib/exe/fetch.php/pessoais:abtmartins:a_concise_guide_to_compositional_data_analysis.pdf (2003).
3. Filzmoser, P. & Hron, K. Correlation analysis for compositional data. *Math. Geosci.* **41**, 905 (2009).
4. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
5. Morton, J. T. et al. Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* **10**, 2719 (2019).
6. Morton, J. T. et al. Learning representations of microbe–metabolite interactions. *Nat. Methods* **16**, 1306–1314 (2019).
7. Nogueira, F. Bayesian Optimization: open source constrained global optimization tool for Python. https://github.com/fmfn/BayesianOptimization (2014).
8. Martín-Fernández, J. A., Barceló-Vidal, C. & Pawlowsky-Glahn, V. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Math. Geol.* **35**, 253–278 (2003).
9. Silverman, J. D., Roche, K., Mukherjee, S. & David, L. A. Naught all zeros in sequence count data are the same. Preprint at *bioRxiv* https://doi.org/10.1101/477794 (2018).
10. McDonald, D. et al. American Gut: an open platform for citizen science microbiome research. *mSystems* **3**, e00031-18 (2018).
11. Swenson, T. L., Karaoz, U., Swenson, J. M., Bowen, B. P. & Northen, T. R. Linking soil biology and chemistry in biological soil crust using isolate exometabolomics. *Nat. Commun.* **9**, 19 (2018).
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. in *Advances in Neural Information Processing Systems* 3111–3119 (2013).
13. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc. B* **44**, 139–160 (1982).

we evaluated in the original article. It is critical that we provide accurate guidance to the community so that scenarios where one method works better than others are well understood. While there may be scenarios where linear methods outperform neural networks, we show that there are scenarios where neural networks outperform linear methods. We appreciate the communication on the topic to the extent that it helps the community better understand the advantages and limitations of the different approaches and prompts the community to continue to innovate in this area.

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/s41592-020-01007-0.

Methods

The simulations were created by using the generative form of MMvec; the microbe and metabolite factor loadings were randomly generated from a normal distribution to parameterize the MMvec parameters. Microbial counts were then drawn from a multinomial logistic normal distribution and fed into MMvec to generate the metabolite counts. To identify scenarios where CLR correlations underperformed in comparison to MMvec, we used Bayesian Optimization to tune the distributions used to generate the simulations.

The CLR-transformed correlations suggested by Quinn and Erb were benchmarked on the desert biocrust soils dataset using the R scripts provided in ref. [1].

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

Data availability

The datasets to reproduce the results presented here can be found at https://github.com/knightlab-analyses/multiomic-cooccurrences.

Code availability

The analysis software to reproduce the results presented here can be found at https://github.com/knightlab-analyses/multiomic-cooccurrences.

Author contributions

J.T.M. performed all analyses and wrote the manuscript. All authors have contributed edits to the manuscript.

Corresponding author(s):	Rob Knight
Last updated by author(s):	9/10/2020

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	Only simulation data was used.
Data analysis	All data analysis scripts can be found here: https://github.com/knightlab-analyses/multiomic-cooccurences

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

The biocrust soils data was retrieved from the supplemental section in Swenson et al

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	We used 500 samples in our simulations, since this is larger than most of the studies that we have analyzed. For the biocrust soils study, there were 19 samples and after filtering there were 466 unique microbial taxa and 85 metabolite features.
Data exclusions	Taxa that appeared in less than 10 samples for each study were removed, since there are fewer samples than degrees of freedom in the model to infer these microbes co-occurrence patterns.
Replication	Replication was not necessary, since the data was simulated, not collected.
Randomization	Randomization was not necessary, since the data was simulated, not collected.
Blinding	Blinding was not necessary, since the data was simulated, not collected.

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