

All hail reproducibility in microbiome research

Jacques Ravel^{1*} and K Eric Wommack^{2*}

At its core, microbiome research relies on complex and sophisticated statistical analyses of large datasets and their associated metadata (e.g., experimental parameters, sample characteristics). Heavy reliance on big data has presented new challenges in communicating the details of complex analyses in a manner sufficient for others to replicate analytical workflows. Reproducibility is a pillar of sound research, and scientific journals need to embrace transparency and make every effort to enable reproducibility through comprehensive and clear reporting of analytical approaches. In this issue of *Microbiome*, a report by Meadow *et al.* [1] on the microbial communities of classroom surfaces sets a new bar for thoroughness in the availability of data, metadata, and analytical resources (code and scripts). It is our hope that this paper will serve as a template for the clever use of publicly available resources and code repositories to enable fully reproducible microbiome research.

“Scientific publications have at least two goals: (i) to announce a result and (ii) to convince readers that the result is correct... papers in experimental science should describe the results and provide a clear enough protocol to allow successful repetition and extension” [2]. Reproducibility and extension are only possible if: data is easily and freely accessible and delivered in format that adheres to international standards; and analysis workflows and scripts are embedded in the publication. Microbiome research is, by its nature, a multi-disciplinary endeavor where experimentalists often work with biostatisticians, mathematicians, computer scientists, or epidemiologists. At times, this multi-disciplinary character can result in a clash of scientific cultures with different approaches to openness, transparency and data release. For example, large sequence datasets and most importantly associated metadata have resulted from our work with epidemiologists [3,4]. However, the notion of releasing data and analysis scripts along with a publication has often been met with great surprise by our epidemiology colleagues. Now that

microbiome research is transitioning from a descriptive and associative science to a translational science that will start impacting lives, we feel the time is right for the community to set standards for complete transparency and full reproducibility. Experimental science suffers each time there is a realization that a high profile report of a scientific finding is not reproducible. Over the long term, news stories of irreproducible science in the popular press can have lasting negative effects on the credibility of the scientific community in general [5]. Without reproducibility, microbiome science will battle to regain credibility and opportunities for scientific advancement will be lost.

Scientific journals should be at the forefront of efforts to ensure that data is accessible prior to publication and made available during the peer review process. Today, fortunately, there are numerous options for data release, such as among others, the NCBI Database of Genotypes and Phenotypes (dbGaP - www.ncbi.nlm.nih.gov/gap) and the Short Read Archive (SRA - www.ncbi.nlm.nih.gov/sra/), options selected by the Human Microbiome Project for example, or other services such as FigShare (www.figshare.com), which was used in the Meadow *et al.* paper [1]. Data deposited into FigShare is permanently archived and redundantly backed up at major universities around the world through the CLOCKSS system (a not-for-profit venture started by libraries and publishers committed to ensuring long-term access to scholarly publications in digital format - www.clockss.org), and a permanent digital object identifier (DOI) is supplied with each dataset. Metadata associated with any dataset should also be made available, and in standard format with controlled ontology. Standards such as the minimum information about a marker gene sequence (MIMARKS) or the minimum information about any (x) sequence (MIxS) [6] are community driven standards that if fully adopted would enhance the long-term scientific use of microbiome datasets.

Data availability is critical but detailed descriptions of the procedures used in the processing of raw data and statistical analyses are equally important for reproducibility. Simply providing scripts and workflow is not enough; data and code have to be understandable to be

* Correspondence: javel@som.umaryland.edu; wommack@dbi.udel.edu

¹Institute for Genome Sciences, University of Maryland School of Medicine, 801 W. Baltimore Street, Baltimore, MD 21201, USA

²Delaware Biotechnology Institute, University of Delaware, 15 Innovation Way, Newark, DE 19711, USA

reproducible. Hence, commenting and versioning is essential and should be included in the publication of scripts. There are several tools available depending on the statistical package or programming language. For example, iPython notebook (www.ipython.org/ipython-doc/dev/interactive/notebook.html) for python scripts enables commenting and tutorials for documenting use cases. Popular tools such as DigiNorm developed by Dr. C. Titus Brown (Michigan State University) use iPython notebook (www.ged.msu.edu/papers/2012-diginorm/) and it is no mistake that the best documented tools often turn out to be more frequently used by microbiome researchers. Statistical analyses in microbiome research increasingly rely on the R statistical language [7]. The R Markdown language simplifies creation of fully-reproducible statistical analysis [8], and has been implemented in packages such as Sweave [9] or knitr [10]. Combined with GitHub (www.github.com), a code versioning repository, scripts can be run and analytical outcomes from reported datasets can be fully reproduced. Dozens of other packages are available for commenting and release of workflow and scripts. Again, Meadow and co-authors [1] used both knitr and GitHub in making their statistical workflow and code publicly available. We applaud the efforts of initiatives such as the Minimum Information About a Bioinformatics investigation (MIABi) [11], which seeks to advance standards for bioinformatics activities that will improve the persistence, reproducibility, and disambiguation of code. Ultimately, these practices will improve transparency and reproducibility. Moving forward *Microbiome* will seek to raise the bar for reproducibility in microbiome research by asking authors to provide easy access to data and code that will ultimately enrich our vibrant and growing research field.

Competing interests

The authors declare that they have no competing interests.

Authors' contribution

JR and KEW wrote the manuscript. Both authors read and approved the final manuscript.

Received: 3 March 2014 Accepted: 3 March 2014

Published: 7 March 2014

References

1. Meadow JF, Altrichter AE, Kembel SW, Moriyama M, O'Connor TK, Wommack AM, Brown GZ, Green JL, Bohannan BJM: **Bacterial communities on classroom surfaces vary with human contact.** *Microbiome* 2014, **2**:7.
2. Mesirov JP: **Computer science. Accessible reproducible research.** *Science* 2010, **327**(5964):415–416.
3. Gajer P, Brotman RM, Bai G, Sakamoto J, Schutte UM, Zhong X, Koenig SS, Fu L, Ma ZS, Zhou X, Abdo Z, Forney LJ, Ravel J: **Temporal dynamics of the human vaginal microbiota.** *Sci Transl Med* 2012, **4**(132):132ra152.
4. Ravel J, Jiang L, Stanley ST, Wilson MR, Decker RS, Read TD, Worsham P, Keim PS, Salzberg SL, Fraser-Liggett CM, Rasko DA: **The complete genome sequence of *Bacillus anthracis* Ames "Ancestor".** *J Bacteriol* 2009, **191**(1):445–446.
5. Editorial Team: **Trouble at the lab.** In *The Economist*. 2013. <http://www.economist.com/node/21588057>.
6. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, Morrison N, Rocca-Serra P, Sterk P, Arumugam M, Bailey M, Baumgartner L, Birren BW, Blaser MJ, Bonazzi V, Booth T, Bork P, Bushman FD, Buttigieg PL, Chain PS, Charlson E, Costello EK, Huot-Creasy H, et al: **Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications.** *Nat Biotechnol* 2011, **29**(5):415–420.
7. R Development Core Team: *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2008.
8. Baumer B, Cetinkaya-Rundel M, Bray A, Loi L, Horton NJ: **R Markdown: integrating a reproducible analysis tool into introductory statistics.** *arXiv.org* 2014, arXiv:1402.1894.
9. Leisch F: **Sweave: Dynamic generation of statistical reports using literate data analysis.** In *Compstat 2002 - Proceedings in Computational Statistics.* Heidelberg: Physica Verlag; 2002:575–580.
10. Xie Y: **knitr: a comprehensive tool for reproducible research in R.** In *Implementing Reproducible Computational Research.* Edited by Stodden V, Leisch F, Peng RD. Boca Raton, FL: Chapman & Hall/CRC The R Series; 2013.
11. Tan TW, Tong JC, Khan AM, de Silva M, Lim KS, Ranganathan S: **Advancing standards for bioinformatics activities: persistence, reproducibility, disambiguation and Minimum Information About a Bioinformatics investigation (MIABi).** *BMC Genomics* 2010, **11**(Suppl 4):S27.

doi:10.1186/2049-2618-2-8

Cite this article as: Ravel and Wommack: All hail reproducibility in microbiome research. *Microbiome* 2014 **2**:8.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

