

Reproducible research

ANSC 595

Today's goals

- Importance of reproducibility in microbiome/data science
- Barriers to reproducibility
- Some tools to overcome those barriers
- Data and Project management
- Introduction to unix

Reproducible research

Reading

- Meadow, J. F., Altrichter, A. E., Kembel, S. W., Moriyama, M., O'Connor, T. K., Womack, A. M., et al. (2014). Bacterial communities on classroom surfaces vary with human contact. *Microbiome*, 2(1), 7–7. <http://doi.org/10.1186/2049-2618-2-7>
- Ravel, J., & Wommack, K. E. (2014). All hail reproducibility in microbiome research. *Microbiome*, 2(1), 8. <http://doi.org/10.1186/2049-2618-2-8>
- Casadevall, A., Ellis, L. M., Davies, E. W., McFall-Ngai, M., & Fang, F. C. (2016). A Framework for Improving the Quality of Research in the Biological Sciences. (Vol. 7). mBio. <http://doi.org/10.1128/mBio.01256-16>
- Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature*, 505(7485), 612–613.
- Noble, W. S. (2009). A quick guide to organizing computational biology projects. *PLoS Computational Biology*, 5(7), e1000424. <http://doi.org/10.1371/journal.pcbi.1000424>

Case study: The Productive Post-doc Part #1

A very productive and confident post-doc in the lab, Jane, has performed some analyses on soil samples that were used for a 16S rRNA microbiome survey. She maintains the contextual data in an excel spreadsheet on her laptop, and backed up on the lab's server. To analyze these data, she routinely copies and pastes the values into a new "clean" tab to try different calculations, but she keeps the original data in the first tab. She also sometimes loads the excel tables into R for other statistical analyses. She keeps her R workflows on a private GitHub repository. In the end, she writes a paper describing her results, including an analysis of the explanatory value of the soil chemistry for microbiome community composition.

Discussion

Critique how Jane's stores and analyzes her data. What aspects of her strategy are reproducible? What aspects of her strategy can be improved?

Case study: The Productive Post-doc Part #2

A student in the lab, John, is a co-author on the paper that Jane is leading. He wants to check her workflow. He cannot access the private workflow on GitHub, and asks for the workflow to be made public. In the mean time, he checks the excel files that are backed up on the lab's servers and notices that a key sample, Sample1, is missing from all of the tabs in Jane's file. He also notices that some of the tabs were not copied correctly from the original data, and this has impacted some of the excel results.

Discussion

1. What should John do?
2. What aspects of this situation may make it difficult for John to decide what to do?

Case study: The Productive Post-doc Part #2

Jane makes the GitHub repo available to John and he reproduces the R workflow. He notices that the p-values that he generates are off from what Jane has reported in the results section of the paper. In fact, where Jane reports a p-value of 0.031, John generates a p-value of 0.31. John now feels very uncomfortable being an author on the paper.

Discussion

1. What would you advise John to do?
2. What steps should be taken before the paper is submitted?
3. Do you think that Jane's errors constitute misconduct or negligence? What is the difference?

Case study wrap-up & comparison to wet-bench

Consider the following scenarios:

1. A lab won't share their modifications to a DNA extraction protocol that they used to generate a fungal ITS leaf survey
2. A post-doc takes all of the freezer stocks of her genetic constructs when she moves to her new faculty position.
3. A graduate student never takes laboratory notes, and instead writes calculations on paper scraps and then discards.

Reproducible research

Pat Schloss Riffomonas Project

http://www.riffomonas.org/reproducible_research/

What does reproducible research look like?

[Aims and scope](#)

[Fees and funding](#)

[Language editing
services](#)

[Copyright](#)

▼ [Preparing your
manuscript](#)

[Research](#)

[Comment](#)

[Correspondence](#)

[Methodology](#)

[Review](#)

[Brief report](#)

[Software](#)

[Meeting Report](#)

[Prepare supporting
information](#)

[Conditions of
publication](#)

[Editorial policies](#)

[Peer-review policy](#)

[Manuscript transfers](#)

[Promoting your
work](#)

Research

Criteria

Availability of data, metadata and analytical scripts

At *Microbiome* we are striving to make reproducibility a priority. Data availability at time of submission is a key aspect to this process as it allows reviewers to fully evaluate your work.

Microbiome follows a strict data release policy ([Research Data Policy Type 4](#)). We require that all datasets on which the conclusions of the paper rely should be available to the reviewers and readers. We ask that authors make sure their datasets are either deposited in publicly available repositories (where available and appropriate) or presented in the main manuscript or additional supporting files whenever possible. Accompanying metadata must be available in the repository or as supporting files to the manuscript. Metadata should be formatted according to the MixS (Minimum Information about any (x) Sequence) standards developed by the Genome Standards Consortium (GSC). Template can be found here: <http://gensc.org/mixs/>). The sample identifiers in the repository must refer to the same sample identifiers used in the manuscript. Please see Springer Nature's information on [recommended repositories](#).

We are also requiring that authors make the code/scripts used for their analysis available as knitr files, iPython Notebooks, or any other formats they might find suitable. Again, this effort encourages transparency and complete reproducibility of your study. A good example is a paper published in *Microbiome* by [Meadow et al.](#)

[Aims and scope](#)[Fees and funding](#)[Language editing services](#)[Copyright](#)

▼ [Preparing your manuscript](#)

[Research](#)[Comment](#)[Correspondence](#)[Methodology](#)[Review](#)[Brief report](#)[Software](#)[Meeting Report](#)[Prepare supporting information](#)[Conditions of publication](#)[Editorial policies](#)[Peer-review policy](#)[Manuscript transfers](#)[Promoting your work](#)

Research

Criteria

Availability of data, metadata and analytical scripts

At *Microbiome* we are striving to make reproducibility a priority. Data availability at time of submission is a key aspect to this process as it allows reviewers to fully evaluate your work.

Microbiome follows a strict data release policy ([Research Data Policy Type 4](#)). We require that all datasets on which the conclusions of the paper rely should be available to the reviewers and readers. **We ask that authors make sure their datasets are either deposited in publicly available repositories (where available and appropriate) or presented in the main manuscript or additional supporting files whenever possible.** Accompanying metadata must be available in the repository or as supporting files to the manuscript. Metadata should be formatted according to the MixS (Minimum Information about any (x) Sequence) standards developed by the Genome Standards Consortium (GSC). Template can be found here: <http://gensc.org/mixs/>). The sample identifiers in the repository must refer to the same sample identifiers used in the manuscript. Please see Springer Nature's information on [recommended repositories](#).

We are also requiring that authors make the code/scripts used for their analysis available as knitr files, iPython Notebooks, or any other formats they might find suitable. Again, this effort encourages transparency and complete reproducibility of your study. A good example is a paper published in *Microbiome* by [Meadow et al.](#)

[Aims and scope](#)[Fees and funding](#)[Language editing services](#)[Copyright](#)

▼ [Preparing your manuscript](#)

[Research](#)[Comment](#)[Correspondence](#)[Methodology](#)[Review](#)[Brief report](#)[Software](#)[Meeting Report](#)[Prepare supporting information](#)[Conditions of publication](#)[Editorial policies](#)[Peer-review policy](#)[Manuscript transfers](#)[Promoting your work](#)

Research

Criteria

Availability of data, metadata and analytical scripts

At *Microbiome* we are striving to make reproducibility a priority. Data availability at time of submission is a key aspect to this process as it allows reviewers to fully evaluate your work.

Microbiome follows a strict data release policy ([Research Data Policy Type 4](#)). We require that all datasets on which the conclusions of the paper rely should be available to the reviewers and readers. **We ask that authors make sure their datasets are either deposited in publicly available repositories (where available and appropriate) or presented in the main manuscript or additional supporting files whenever possible.** **Accompanying metadata must be available in the repository or as supporting files to the manuscript.** Metadata should be formatted according to the MixS (Minimum Information about any (x) Sequence) standards developed by the Genome Standards Consortium (GSC). Template can be found here: <http://gensc.org/mixs/>). The sample identifiers in the repository must refer to the same sample identifiers used in the manuscript. Please see Springer Nature's information on [recommended repositories](#).

We are also requiring that authors make the code/scripts used for their analysis available as knitr files, iPython Notebooks, or any other formats they might find suitable. Again, this effort encourages transparency and complete reproducibility of your study. A good example is a paper published in *Microbiome* by [Meadow et al.](#)

[Aims and scope](#)

[Fees and funding](#)

[Language editing
services](#)

[Copyright](#)

▼ [Preparing your
manuscript](#)

[Research](#)

[Comment](#)

[Correspondence](#)

[Methodology](#)

[Review](#)

[Brief report](#)

[Software](#)

[Meeting Report](#)

[Prepare supporting
information](#)

[Conditions of
publication](#)

[Editorial policies](#)

[Peer-review policy](#)

[Manuscript transfers](#)

[Promoting your
work](#)

Research

Criteria

Availability of data, metadata and analytical scripts

At *Microbiome* we are striving to make reproducibility a priority. Data availability at time of submission is a key aspect to this process as it allows reviewers to fully evaluate your work.

Microbiome follows a strict data release policy ([Research Data Policy Type 4](#)). We require that all datasets on which the conclusions of the paper rely should be available to the reviewers and readers. We ask that authors make sure their datasets are either deposited in publicly available repositories (where available and appropriate) or presented in the main manuscript or additional supporting files whenever possible. Accompanying metadata must be available in the repository or as supporting files to the manuscript. Metadata should be formatted according to the MixS (Minimum Information about any (x) Sequence) standards developed by the Genome Standards Consortium (GSC). Template can be found here: <http://gensc.org/mixs/>). The sample identifiers in the repository must refer to the same sample identifiers used in the manuscript. Please see Springer Nature's information on [recommended repositories](#).

We are also requiring that authors make the code/scripts used for their analysis available as knitr files, iPython Notebooks, or any other formats they might find suitable. Again, this effort encourages transparency and complete reproducibility of your study. A good example is a paper published in *Microbiome* by [Meadow et al.](#)

[Aims and scope](#)

[Fees and funding](#)

[Language editing services](#)

[Copyright](#)

▼ [Preparing your manuscript](#)

[Research](#)

[Comment](#)

[Correspondence](#)

[Methodology](#)

[Review](#)

[Brief report](#)

[Software](#)

[Meeting Report](#)

[Prepare supporting information](#)

[Conditions of publication](#)

[Editorial policies](#)

[Peer-review policy](#)

[Manuscript transfers](#)

[Promoting your work](#)

Research

Criteria

Availability of data, metadata and analytical scripts

At *Microbiome* we are striving to make reproducibility a priority. Data availability at time of submission is a key aspect to this process as it allows reviewers to fully evaluate your work.

Microbiome follows a strict data release policy ([Research Data Policy Type 4](#)). We require that all datasets on which the conclusions of the paper rely should be available to the reviewers and readers. We ask that authors make sure their datasets are either deposited in publicly available repositories (where available and appropriate) or presented in the main manuscript or additional supporting files whenever possible. Accompanying metadata must be available in the repository or as supporting files to the manuscript. Metadata should be formatted according to the MixS (Minimum Information about any (x) Sequence) standards developed by the Genome Standards Consortium (GSC). Template can be found here: <http://gensc.org/mixs/>. The sample identifiers in the repository must refer to the same sample identifiers used in the manuscript. Please see Springer Nature's information on [recommended repositories](#).

We are also requiring that authors make the code/scripts used for their analysis available as knitr files, iPython Notebooks, or any other formats they might find suitable. Again, this effort encourages transparency and complete reproducibility of your study. A good example is a paper published in *Microbiome* by [Meadow et al.](#)

[Aims and scope](#)[Fees and funding](#)[Language editing
services](#)[Copyright](#)

▼ [Preparing your
manuscript](#)

[Research](#)[Comment](#)[Correspondence](#)[Methodology](#)[Review](#)[Brief report](#)[Software](#)[Meeting Report](#)[Prepare supporting
information](#)[Conditions of
publication](#)[Editorial policies](#)[Peer-review policy](#)[Manuscript transfers](#)[Promoting your
work](#)

Research

Criteria

Availability of data, metadata and analytical scripts

At *Microbiome* we are striving to make reproducibility a priority. Data availability at time of submission is a key aspect to this process as it allows reviewers to fully evaluate your work.

Microbiome follows a strict data release policy ([Research Data Policy Type 4](#)). We require that all datasets on which the conclusions of the paper rely should be available to the reviewers and readers. We ask that authors make sure their datasets are either deposited in publicly available repositories (where available and appropriate) or presented in the main manuscript or additional supporting files whenever possible. Accompanying metadata must be available in the repository or as supporting files to the manuscript. Metadata should be formatted according to the MixS (Minimum Information about any (x) Sequence) standards developed by the Genome Standards Consortium (GSC). Template can be found here: <http://gensc.org/mixs/>. The sample identifiers in the repository must refer to the same sample identifiers used in the manuscript. Please see Springer Nature's information on [recommended repositories](#).

We are also requiring that authors make the code/scripts used for their analysis available as knitr files, iPython Notebooks, or any other formats they might find suitable. Again, this effort encourages transparency and complete reproducibility of your study. A good example is a paper published in *Microbiome* by [Meadow et al.](#)



RESEARCH

Open Access

Bacterial communities on classroom surfaces vary with human contact

James F Meadow^{1*}, Adam E Altrichter¹, Steven W Kembel^{1,2}, Maxwell Moriyama^{1,3}, Timothy K O'Connor^{1,4}, Ann M Womack¹, G Z Brown^{1,3}, Jessica L Green^{1,5} and Brendan J M Bohannan¹

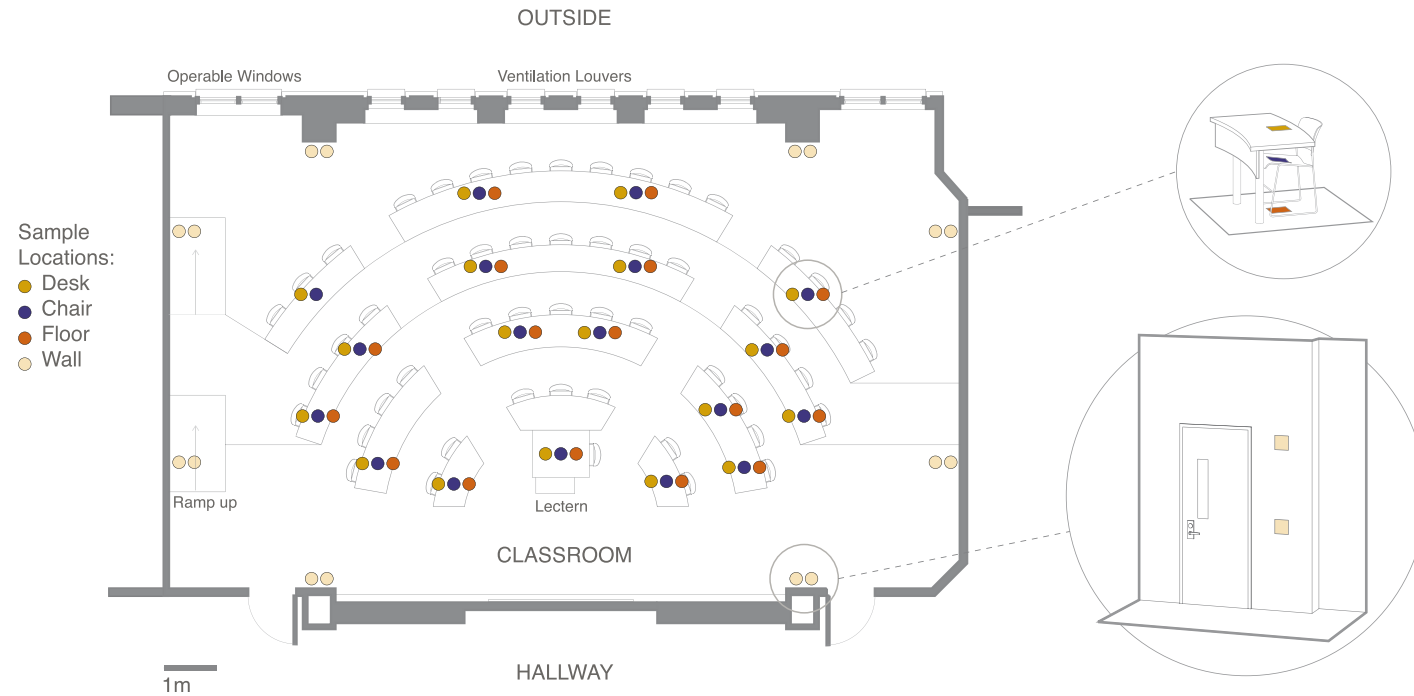


Figure 1 Schematic of sampling design. Four different types of surfaces (desks, chairs, floors and walls) were sampled throughout an amphitheater-style classroom.

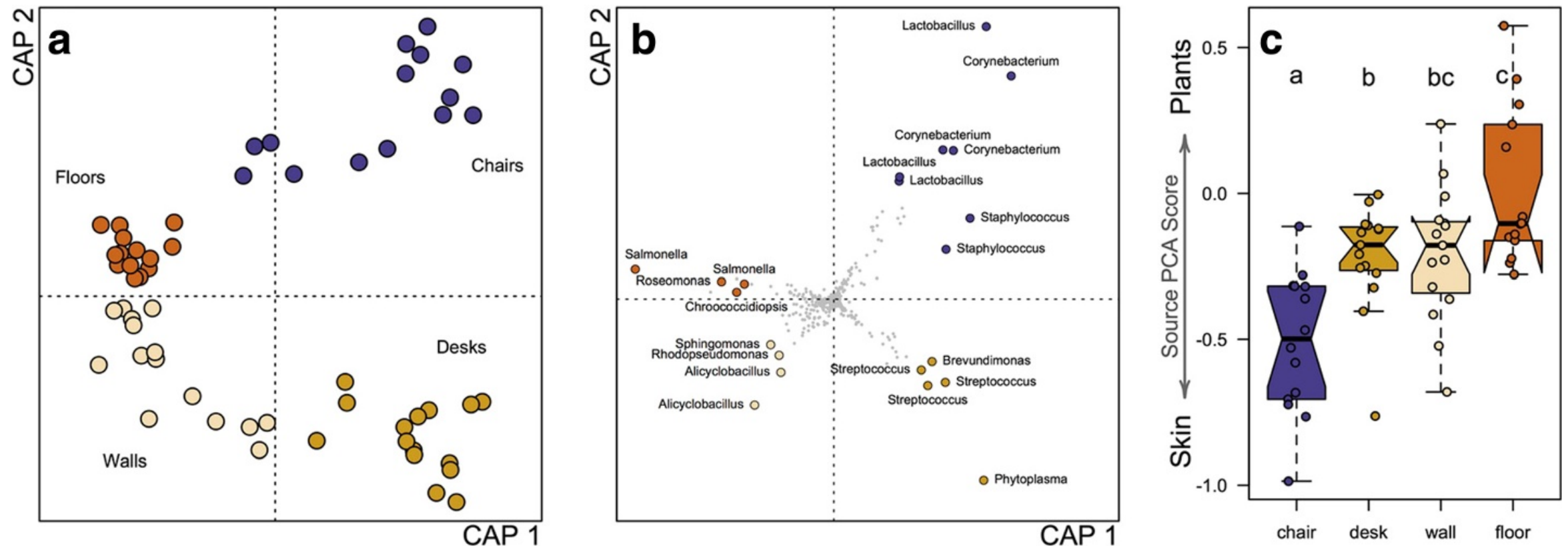


Figure 2 Surfaces harbored significantly different bacterial communities and were linked to differential human contact. (a) Bacterial communities were constrained by four different surface types using distance-based redundancy analysis (DB-RDA; constrained inertia = 11.4%) and were significantly different among types based on Canberra taxonomic distances ($P = 0.001$ from permutational multivariate analysis of variation). **(b)** Bacterial operational taxonomic units (OTUs) from DB-RDA are shown weighting communities in the same four primary directions. The first and second axes from DB-RDA are used in both ordinations (CAP 1 and CAP 2). The strongest ten weighting OTUs for each surface type are highlighted if they were also significant indicator OTUs (all P values < 0.05). **(c)** All samples were compared to potential source environments using principal components analysis (PCA), and the first principal component (37.8% of variance explained) was used as a surrogate for community similarity to either phyllosphere or human skin bacterial communities. Boxplots delineate (from bottom) minimum value, Q1, median (Q2), Q3, maximum value; notches approximate 95% confidence around median value, and outliers fall outside of the quartile range. Letters above each box indicate significant groupings after Tukey's honest significant difference (HSD) test (adjusted P value < 0.05).

Table 1 Closest known isolates related to indicator operational taxonomic units

Greengenes genus	P value	Surface type	Closest 16S NCBI isolate and accession	Isolate source environment	Sequence similarity to isolate (%)
<i>Lactobacillus</i>	0.001*	Chairs	<i>Lactobacillus johnsonii</i> NR_075064.1	Human gut	99
<i>Corynebacterium</i>	0.001*	Chairs	<i>Corynebacterium resistens</i> NR_040999.1	Human infection	99
<i>Corynebacterium</i>	0.001*	Chairs	<i>Corynebacterium confusum</i> NR_026449.1	Human clinical specimens	99
<i>Staphylococcus</i>	0.011*	Chairs	<i>Staphylococcus epidermidis</i> NR_074995.1	Human skin	99
<i>Corynebacterium</i>	0.001*	Chairs	<i>Corynebacterium riegelii</i> NR_026434.1	Human urinary tract	99
<i>Staphylococcus</i>	0.019*	Chairs	<i>Staphylococcus saprophyticus</i> NR_074999.1	Human urinary tract	99
<i>Lactobacillus</i>	0.001*	Chairs	<i>Lactobacillus crispatus</i> NR_074986.1	Human vagina	99
<i>Lactobacillus</i>	0.003*	Chairs	<i>Lactobacillus acidophilus</i> NR_075049.1	Human gut	99
<i>Streptococcus</i>	0.001*	Desks	<i>Streptococcus oralis</i> NR_102809.1	Human oral	99
<i>Streptococcus</i>	0.001*	Desks	<i>Streptococcus salivarius</i> NR_102816.1	Human oral	99
<i>Brevundimonas</i>	0.002*	Desks	<i>Brevundimonas variabilis</i> NR_037106.1	Pond water	99
<i>Streptococcus</i>	0.001*	Desks	<i>Streptococcus intermedius</i> NR_102797.1	Human purulent infection	99
<i>CandidatusPhytoplasma</i>	0.001*	Desks	None**	-	-
<i>Alicyclobacillus</i>	0.001*	Walls	<i>Tumebacillus permanentifrigoris</i> NR_043849.1	Soil	99
<i>Chroococcidiopsis</i>	0.028*	Walls	<i>Halospirulina tapeticola</i> NR_026510.1	Saline aquatic	96
<i>Alicyclobacillus</i>	0.001*	Walls	<i>Tumebacillus permanentifrigoris</i> NR_043849.1	Soil	98
<i>Rhodopseudomonas</i>	0.001*	Walls	<i>Methylobacterium adhaesivum</i> NR_042409	Drinking water	98
<i>Salmonella</i>	0.001*	Floors	<i>Pantoea ananatis</i> NR_103927.1	Phyllosphere	99
<i>Roseomonas</i>	0.001*	Floors	<i>Roseomonas gilardii</i> NR_029061.1	Human blood	99
<i>Roseomonas</i>	0.001*	Floors	<i>Roseomonas frigid aquae</i> NR_044455.1	Water-cooling system	99
<i>Salmonella</i>	0.001*	Floors	<i>Pantoea ananatis</i> NR_103927.1	Phyllosphere	99

All extant operational taxonomic units labeled in Figure 2 (and thus influential in distance-based redundancy analysis, as well as significant indicator taxa for their respective surface type) were related to their closest known bacterial isolate using 16S rRNA sequences in the NCBI Bacteria & Archaea Isolate Database. Source environments are from each isolate's respective published source environment. *Unadjusted *P* value < 0.05. **Closest known isolate 89% similar. NCBI: National Center for Biotechnology Information.

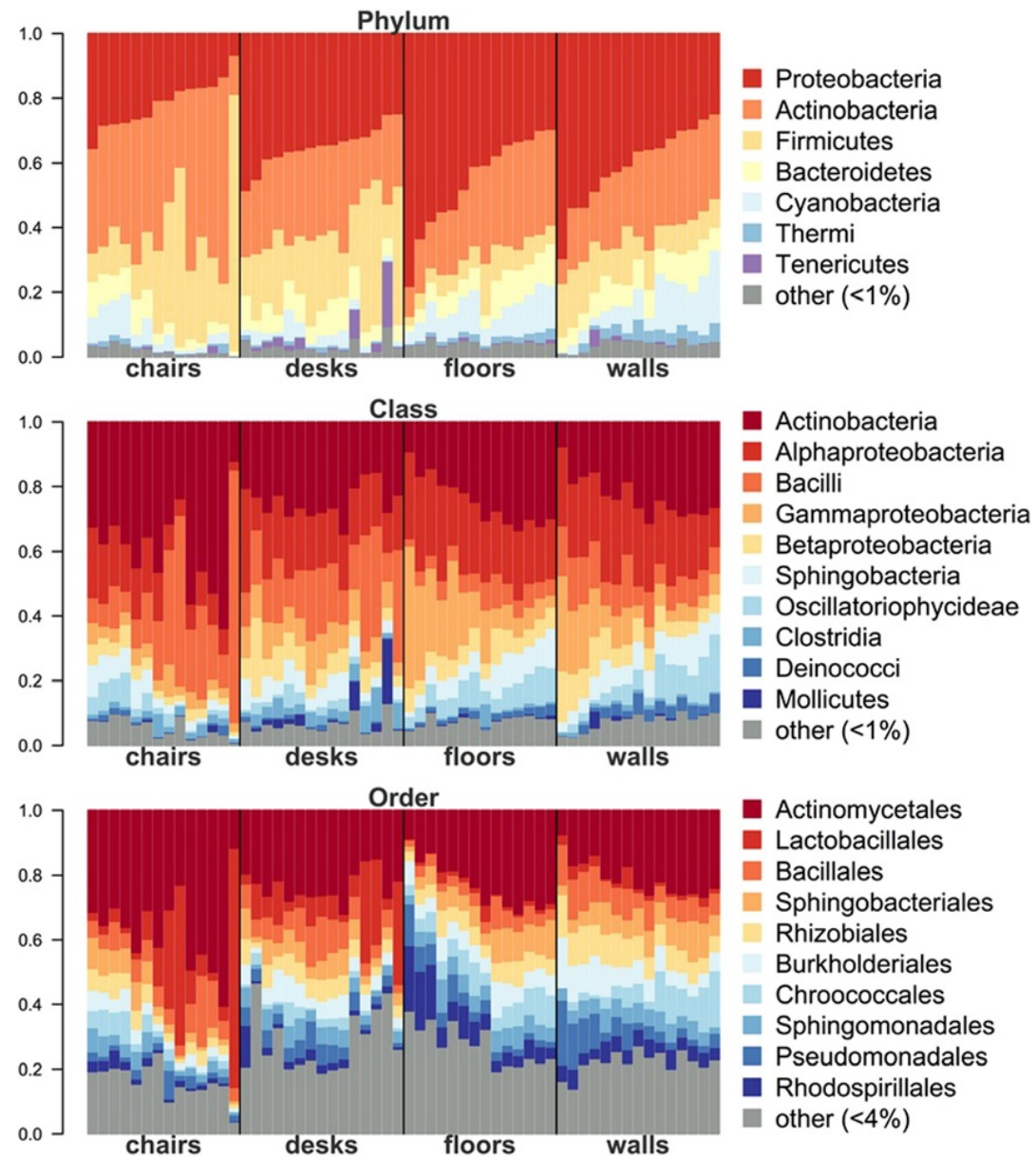
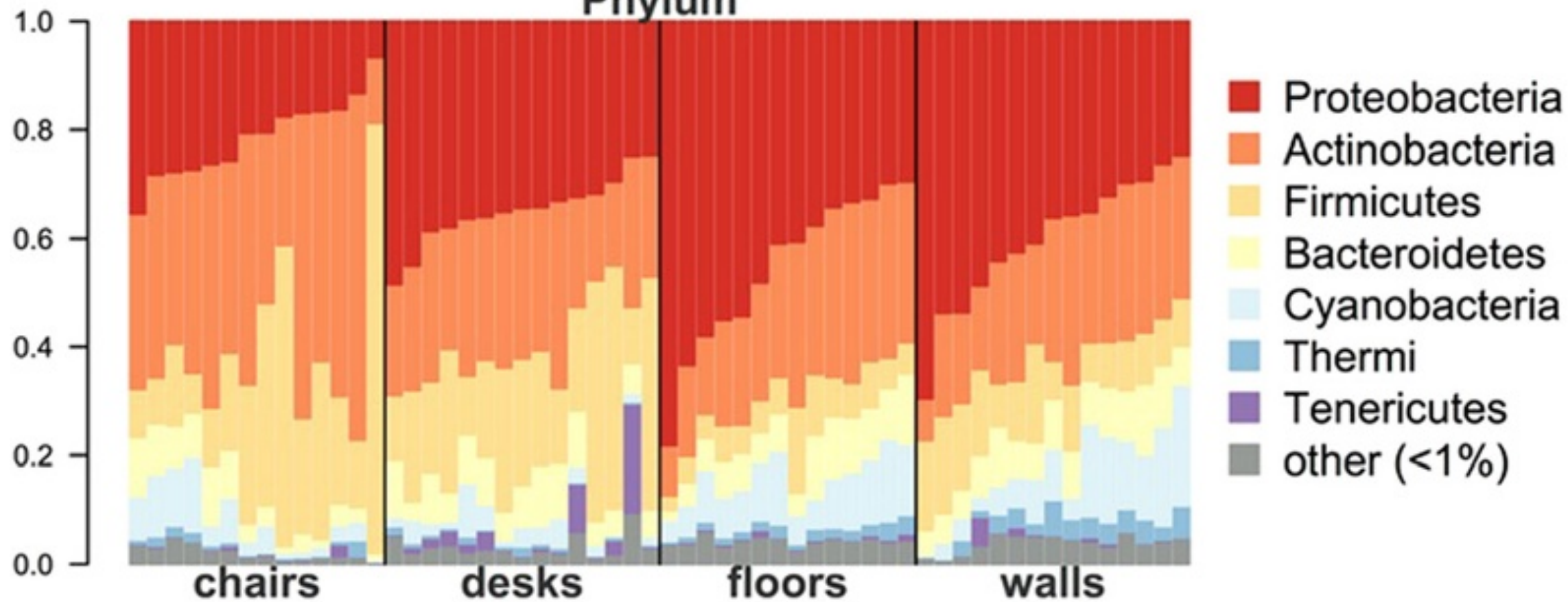
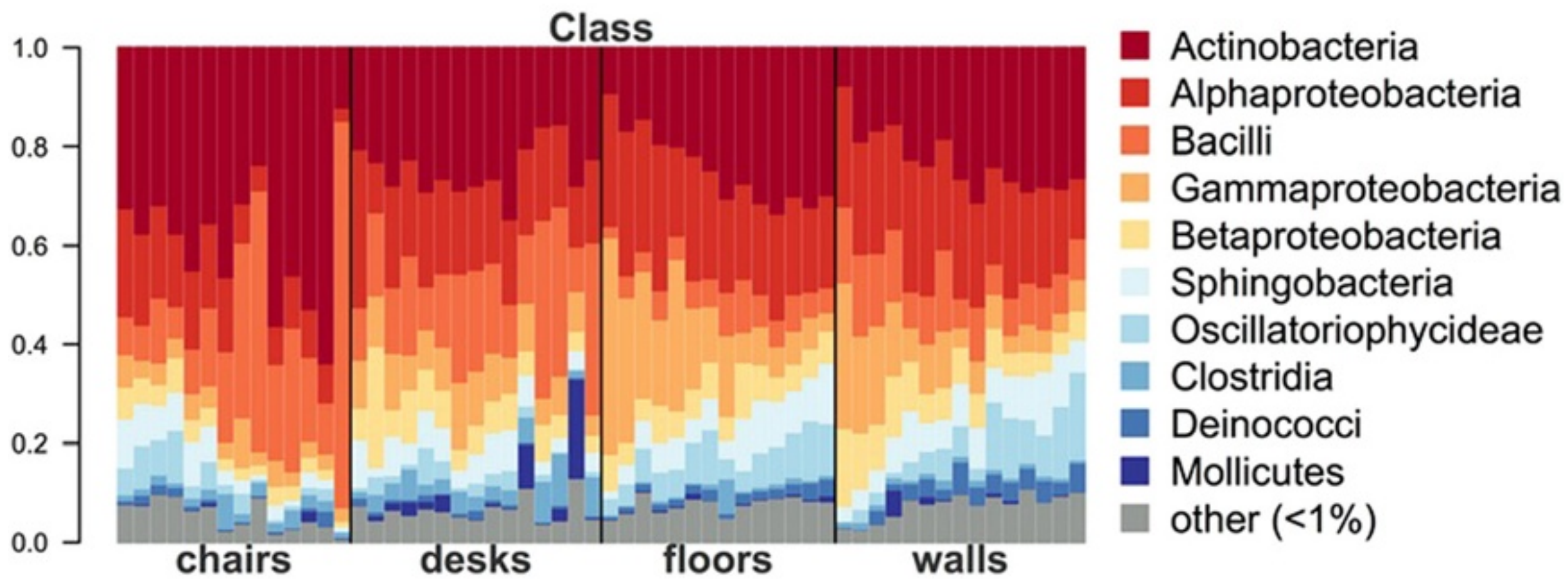
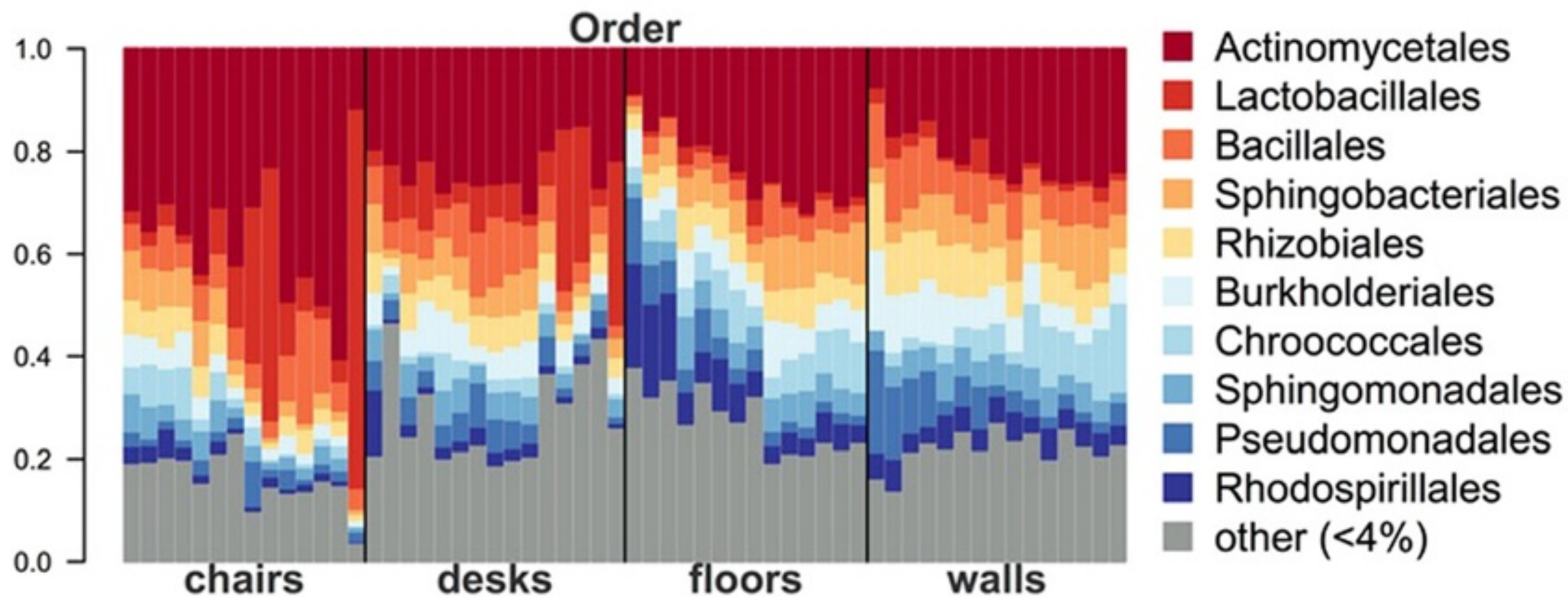


Figure 3 Taxonomic composition of all 58 samples used in this study. Samples are grouped by surface type. All taxonomic groups representing <1% (Phylum and Class) and <4% (Order) of sequences were grouped into 'other'.

Phylum









EDITORIAL

Open Access


All hail reproducibility in microbiome research

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

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.

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.

A Framework for Improving the Quality of Research in the Biological Sciences

 **Arturo Casadevall**,^a Editor in Chief, *mBio*, **Lee M. Ellis**,^b AAM Colloquium Steering Committee Member, **Erika W. Davies**,^c Publishing Ethics Manager, ASM, **Margaret McFall-Ngai**,^d Editor, *mBio*, Senior Editor, *mSystems*, **Ferric C. Fang**,^e Editor in Chief, *Infection and Immunity*

Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA^a; Division of Surgery, Department of Surgical Oncology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA^b; American Society for Microbiology, Washington, DC, USA^c; Pacific Biosciences Research Center, University of Hawaii at Manoa, Honolulu, Hawaii, USA^d; University of Washington School of Medicine, Seattle, Washington, USA^e

ABSTRACT The American Academy of Microbiology convened a colloquium to discuss problems in the biological sciences, with emphasis on identifying mechanisms to improve the quality of research. Participants from various disciplines made six recommendations: (i) design rigorous and comprehensive evaluation criteria to recognize and reward high-quality scientific research; (ii) require universal training in good scientific practices, appropriate statistical usage, and responsible research practices for scientists at all levels, with training content regularly updated and presented by qualified scientists; (iii) establish open data at the timing of publication as the standard operating procedure throughout the scientific enterprise; (iv) encourage scientific journals to publish negative data that meet methodologic standards of quality; (v) agree upon common criteria among scientific journals for retraction of published papers, to provide consistency and transparency; and (vi) strengthen research integrity oversight and training. These recommendations constitute an actionable framework that, in combination, could improve the quality of biological research.

In the second decade of the 21st century, investigators in the biological sciences are making tremendous progress in a wide variety of areas. However, problems in the conduct of science are likely to be responsible for most instances of irreproducible research, including laboratory errors

Suggestion #1: Design Rigorous And Comprehensive Evaluation Criteria To Recognize And Reward High-quality Scientific Research

- Impact factor mania
- these journals often require clean stories
 - sanitize their data
 - overstate the conclusions
- make papers more attractive
- bad science?
- outright misconduct?

Suggestion #2: Require universal training in good scientific practices, appropriate statistical usage, and responsible research practices for scientists at all levels, with training content regularly updated and presented by qualified scientists

- Solid statistical training
 - Misuse of p value
- Experimental design

Suggestion #3: Establish open data as the standard operating procedure throughout the scientific enterprise

- Biologists are now data scientists?
- All data analyzed should be openly available?
 - Is there a down side to this?
- Publishing original data

Suggestion #4: Encourage scientific journals to publish negative data that meet standards of quality

- Barriers to publishing negative results
 - May be a false negative
 - Don't give same reward as a positive result
 - Reputationally
 - Financially
- Benefits of publishing negative results
 - Shine a light on reproducibility
 - Maybe positive results are false positives?

Suggestion #5: Agree upon common criteria among scientific journals for retraction of published papers, to provide consistency and transparency

- Honest or dishonest mistakes

Suggestion #6: Strengthen research integrity oversight and training

- “A finding of misconduct is a career-ending event for most scientists...”
- Misconduct can lead to the public’s distrust in science and medicine
- Research integrity education should occur on all levels, PI to undergrad