

# Topic: Methods for Monitoring Microbes in Built Environments

Anthony Fodor  
Professor  
Department of Bioinformatics and Genomics  
UNC Charlotte



<http://fodorclasses.github.io/gc/gc.pdf>

Gordon Research Conference of the Built Environment  
June 2025



## What is PreMiEr?

The Engineering Research Center (ERC) for Precision Microbiome Engineering (PreMiEr) studies and improves the **microbiomes** of the **built environment**.

It is a National Science Foundation (NSF) funded collaboration between Duke University, North Carolina Agricultural & Technical State University (N.C. A&T), North Carolina State University (NCSU), the University of North Carolina at Chapel Hill (UNC-CH), and the University of North Carolina at Charlotte (UNC Charlotte).

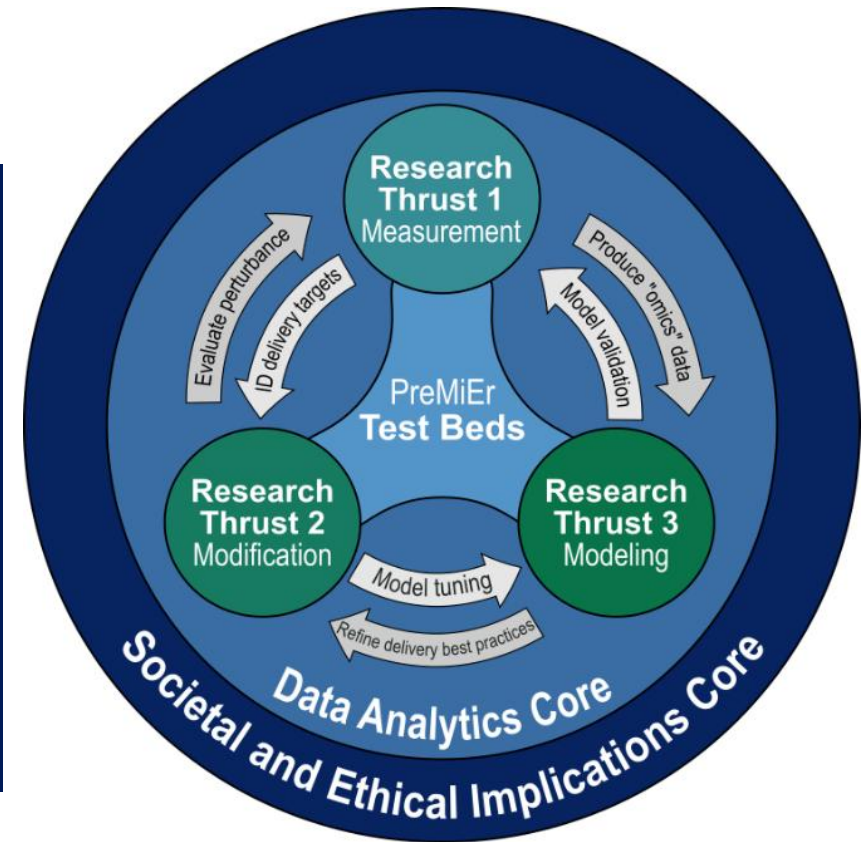


Figure CR-1. Relationship of PreMiEr's research thrusts and cores.



<https://premier-microbiome.org/>

There are important resources, attempts to create standards that everyone should study

---

**nature microbiology**

Consensus Statement

<https://doi.org/10.1038/s41564-025-02035-2>

# **Guidelines for preventing and reporting contamination in low-biomass microbiome studies**

---

Noah Fierer <sup>1</sup>✉, Pok Man Leung <sup>2,3,28</sup>, Rachael Lappan <sup>2,3,28</sup>, Raphael Eisenhofer <sup>4,28</sup>, Francesco Ricci <sup>2,3,28</sup>, Sophie I. Holland<sup>2,3,28</sup>, Nicholas Dragone<sup>1</sup>, Linda L. Blackall <sup>5</sup>, Xiyang Dong <sup>6</sup>, Cristina Dorador<sup>7</sup>, Belinda C. Ferrari <sup>8</sup>, Jacqueline Goordial <sup>9</sup>, Susan P. Holmes <sup>10</sup>, Fumio Inagaki <sup>11</sup>, Tal Korem <sup>12</sup>, Simone S. Li <sup>2</sup>, Thulani P. Makhalanyane <sup>13</sup>, Jessica L. Metcalf <sup>14</sup>, Niranjan Nagarajan <sup>15,16</sup>, William D. Orsi <sup>17</sup>, Erin R. Shanahan<sup>18</sup>, Alan W. Walker <sup>19</sup>, Laura S. Weyrich <sup>20</sup>, Jack A. Gilbert<sup>21,22</sup>, Amy D. Willis <sup>23</sup>, Benjamin J. Callahan <sup>24</sup>, Ashley Shade<sup>25</sup>, Julian Parkhill <sup>26</sup>, Jillian F. Banfield <sup>2,27</sup> & Chris Greening <sup>2,3</sup>✉

Your control strategy should be tied to the purpose of your study ←

Design to avoid the possibility of cross-contamination

There is no “magic” algorithm that will make the contamination problem go away.  
Negative controls are an experimental category in your study and should be powered as such

Reproducibility across studies can build confidence in the rigor of your results

Contamination is not your only problem

## Negative Controls

Three types of negative controls are minimally required to allow adequate monitoring of contaminants throughout sample handling and processing, and provide the ability to detect when and how contaminants are introduced into biological samples. At least one of each type of negative control must be included per sampling, extraction, and amplification batch. Although we would recommend that two negative controls should be used and placed strategically to monitor contaminants from the start to the end of the process (e.g., the first tube should be negative control #1, the last tube should be negative control #2). For larger studies using robotic systems with plates, eight of each type of negative control should be minimally required per study [66].

- (i) *Sampling Blank Controls*. These allow the detection of contaminant DNA introduced during the sampling procedure, including items used to collect the sample, such as swabs, gauze, or drills, and any reagents or preservatives used to store or transport the samples (e.g., media, alcohol, or RNA stabilizer). Material analyzed in sampling blanks should be collected in the same room and at the same time as the biological samples, and should undergo the same laboratory treatment as the biological samples, from collection to sequencing. Although sampling controls will contain DNA from the extraction process, it will allow the researcher to discern which contaminants are specific to the sampling location and equipment versus the laboratory.
- (ii) *DNA Extraction Blank Controls*. These monitor contaminant DNA content in extraction kits, molecular reagents, and the laboratory environment during the DNA extraction process and, as above, should be processed alongside the biological samples from extraction to sequencing.
- (iii) *No-Template Amplification Controls*. These can monitor contaminant DNA present in reagents and the laboratory environment during library preparation and sequencing. All negative controls provide a semi-quantitative estimate of background contaminants and allow researchers to identify contaminants that can be used in downstream subtractive analyses. Finally, it should be noted that negative controls can contain too little DNA to be effectively processed. In these cases, the use of known carrier DNA in blank controls can help to efficiently amplify contaminants [67].

## Opinion

### Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations

Raphael Eisenhofer <sup>1,2,\*</sup> Jeremiah J. Minich,<sup>3</sup> Clarisse Marotz,<sup>4</sup> Alan Cooper,<sup>1,2</sup> Rob Knight,<sup>4,5,6</sup> and Laura S. Weyrich<sup>1,2</sup>

Not all controls are created equal.  
Controls at the beginning of an experiment  
have more power for contaminant detection.

Controls at each step have more power in  
resolving the source of a contaminant.



Your control strategy should be tied to the purpose of your study

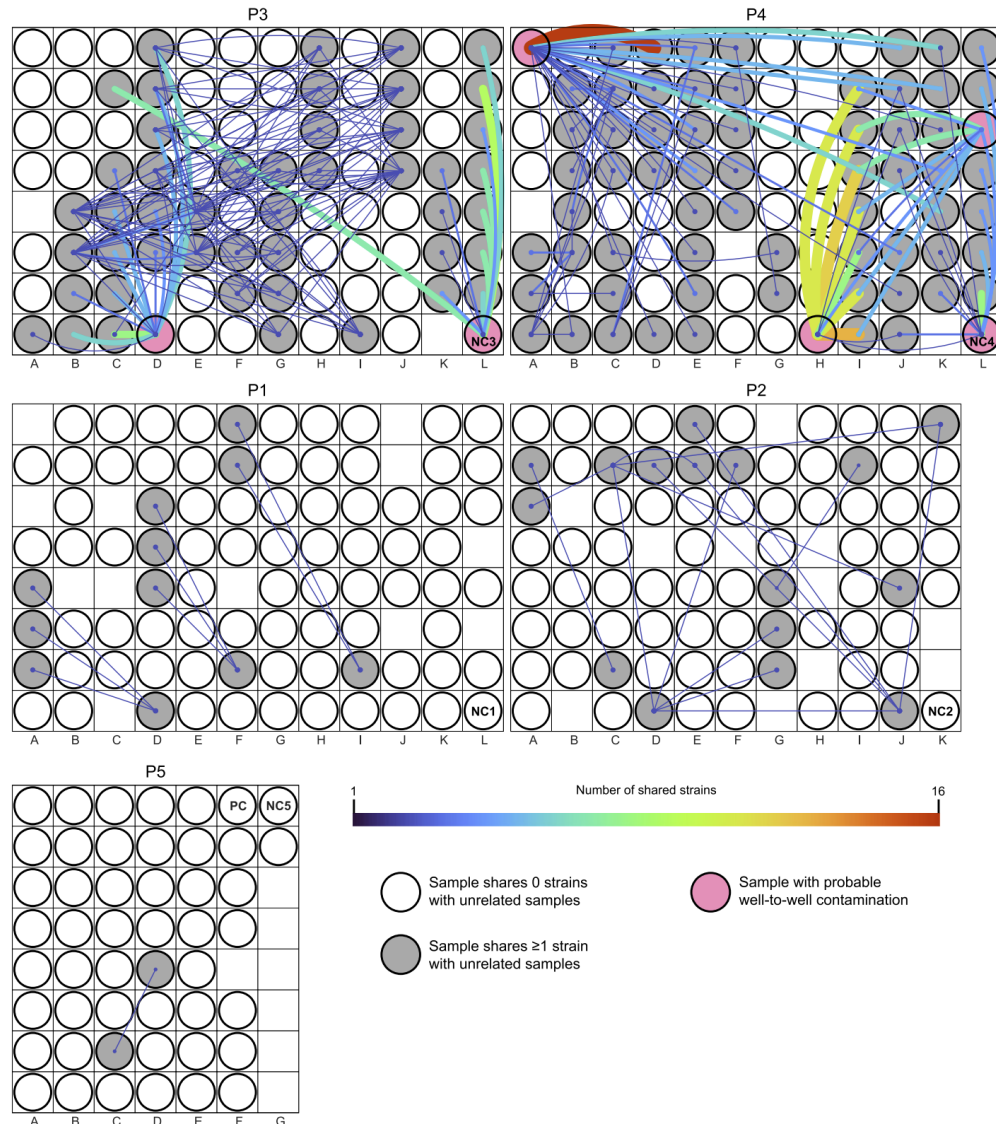
Design to avoid the possibility of cross-contamination ←

There is no “magic” algorithm that will make the contamination problem go away.  
Negative controls are an experimental category in your study and should be powered as such

Reproducibility across studies can build confidence in the rigor of your results

Contamination is not your only problem

By request: Cross well contamination is a sometimes under-appreciated source of artifact in low biomass samples



**Fig. 2** Within-plate strain sharing among unrelated samples. Rectangular areas represent plates (P1-P5) and circles show sample placements within each plate. A line was drawn between unrelated samples if they shared  $\geq 1$  strain. The more strains a sample pair shared, the thicker and brighter the line. If a sample did not share any strains with other unrelated samples, its corresponding circle is colorless. Pink circles represent samples that were likely cross-contaminated

## Using strain-resolved analysis to identify contamination in metagenomics data

Yue Clare Lou<sup>1</sup>, Jordan Hoff<sup>2</sup>, Matthew R. Olm<sup>1,3</sup>, Jacob West-Roberts<sup>4</sup>, Spencer Diamond<sup>2,5</sup>, Brian A. Firek<sup>6</sup>, Michael J. Morowitz<sup>6</sup> and Jillian F. Banfield<sup>2,4,5\*</sup>


Cross-contamination is widespread, especially if 96 well plate automation is used.

Mixing high-abundance and low-abundance samples on the same plate is a recipe for potential disaster...

Make sure case and control samples are not run on separate plates as plate effects may be mistaken for true category effects

Your control strategy should be tied to the purpose of your study

Design to avoid the possibility of cross-contamination

There is no “magic” algorithm that will make the contamination problem go away.   
Negative controls are an experimental category in your study and should be powered as such

Reproducibility across studies can build confidence in the rigor of your results

Contamination is not your only problem



# **Guidelines for preventing and reporting contamination in low-biomass microbiome studies**

With low-biomass samples, the process of using negative controls to differentiate biological signal from contamination is often not straightforward. Therefore, removal of contaminants from sequence data is often challenging to do with absolute certainty, making transparency in the reporting of data and associated analyses even more critical (see Table 2 and Box 1).

It is crucial to deposit both filtered and un-filtered data into public databases  
Including the original data will allow for re-analysis as methods improve

**Ervin, Benjamin - 0222 - MITLL**

 7:49 PM (14 minutes ago)



to me ▼

“Old data, when seen in the light of new ideas, can give us an entirely new insight into a phenomenon; we have an impressive recent example of this in the Bayesian spectrum analysis of nuclear magnetic resonance data, which enables us to make accurate quantitative determinations of phenomena which were not accessible to observation at all with the previously used data analysis by Fourier transforms. When a data set is mutilated (or, to use the common euphemism, ‘filtered’) by processing according to false assumptions, important information in it may be destroyed irreversibly. As some have recognized, this is happening constantly from orthodox methods of detrending or seasonal adjustment in econometrics. However, old data sets, if preserved un mutilated by old assumptions, may have a new lease on life when our prior information advances.”

[https://assets.cambridge.org/97805215/92710/frontmatter/9780521592710\\_frontmatter.pdf](https://assets.cambridge.org/97805215/92710/frontmatter/9780521592710_frontmatter.pdf)

Probability theory The logic of Science – E.T. Jaynes - 2002

Your control strategy should be tied to the purpose of your study

Design to avoid the possibility of cross-contamination

There is no “magic” algorithm that will make the contamination problem go away.

Negative controls are an experimental category in your study and should be powered as such 

Reproducibility across studies can build confidence in the rigor of your results

Contamination is not your only problem

Your control strategy should be tied to the purpose of your study

Design to avoid the possibility of cross-contamination

There is no “magic” algorithm that will make the contamination problem go away.  
Negative controls are an experimental category in your study and should be powered as such

Reproducibility across studies can build confidence in the rigor of your results 

Contamination is not your only problem

# **Guidelines for preventing and reporting contamination in low-biomass microbiome studies**

While it may not be logistically or financially feasible in all studies, if the question of contamination is pressing enough, the best available method of verification is to obtain identical results independent of the laboratory of origin<sup>84,85</sup>.



An integrated analysis suggest strong reproducibility in the existing literature when examining different habitats (hand-level vs. floor) of the built environment



Abeoseh Flemister

An integrated analysis suggest strong reproducibility in the existing literature when examining different habitats (hand-level vs. floor) of the built environment

4 publicly available 16S datasets that have hand-associated vs. floor-associated samples



Abeoseh Flemister

An integrated analysis suggest strong reproducibility in the existing literature when examining different habitats (hand-level vs. floor) of the built environment

4 publicly available 16S datasets that have hand-associated vs. floor-associated samples



Leave one study out –  
train a random forest model on the other 3



Abeoseh Flemister

An integrated analysis suggest strong reproducibility in the existing literature when examining different habitats (hand-level vs. floor) of the built environment

4 publicly available 16S datasets that have hand-associated vs. floor-associated samples



Leave one study out –  
train a random forest model on the other 3



Capture the predictions on the left out 4<sup>th</sup> study with an ROC curve.



Abeoseh Flemister

An integrated analysis suggest strong reproducibility in the existing literature when examining different habitats (hand-level vs. floor) of the built environment

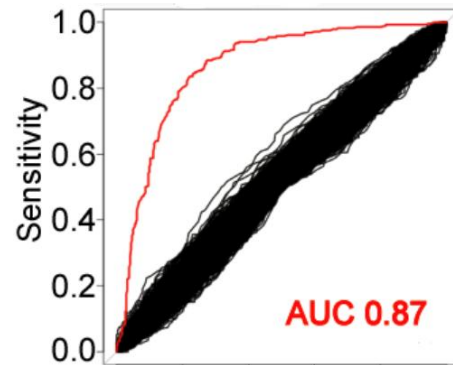
4 publicly available 16S datasets that have hand-associated vs. floor-associated samples



Leave one study out –  
train a random forest model on the other 3



Capture the predictions on the left out 4<sup>th</sup> study with an ROC curve.



Red line – “real data”

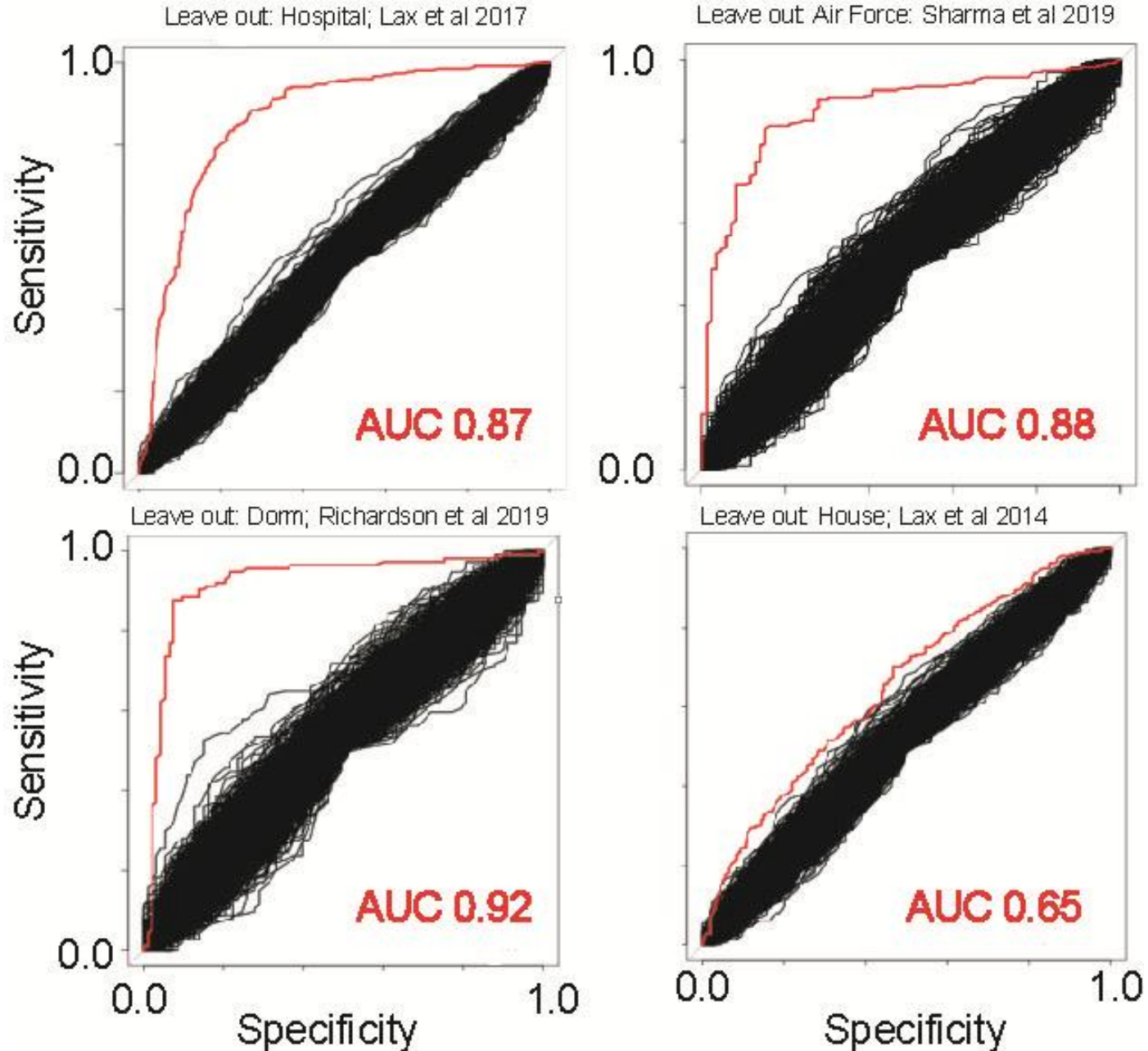
Black lines – model trained on data with labels scrambled



Abeoseh Flemister



An integrated analysis suggest strong reproducibility in the existing literature when examining different habitats (hand-level vs. floor) of the built environment



Abeoseh Flemister

Your control strategy should be tied to the purpose of your study

Design to avoid the possibility of cross-contamination

There is no “magic” algorithm that will make the contamination problem go away.  
Negative controls are an experimental category in your study and should be powered as such

Reproducibility across studies can build confidence in the rigor of your results

Contamination is not your only problem ←

New Results

 [Follow this pre](#)

# **Correction for spurious taxonomic assignments of k-mer classifiers in low microbial biomass samples using shuffled sequences**

Shan Sun,  Anthony A Fodor

doi: <https://doi.org/10.1101/2025.06.18.660363>



Shan Sun

<https://www.biorxiv.org/content/10.1101/2025.06.18.660363v1>

Choose six publicly available cancer datasets (low biomass samples)



<https://www.biorxiv.org/content/10.1101/2025.06.18.660363v1>



Shan Sun

Choose six publicly available cancer datasets (low biomass samples)



Run through the classifier Kraken



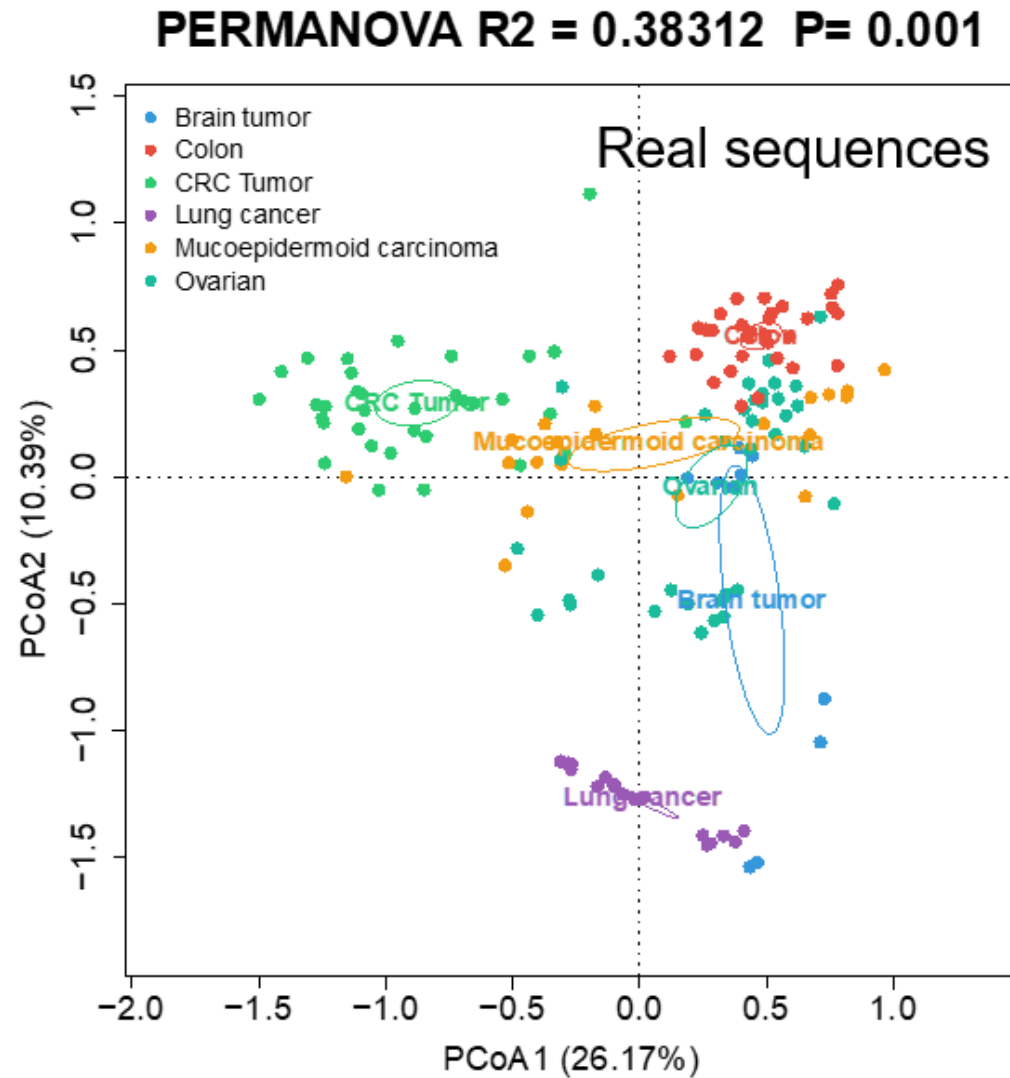
<https://www.biorxiv.org/content/10.1101/2025.06.18.660363v1>



Shan Sun



Non-shuffled sequences produce distinct microbial “signatures” in cancer samples



Shan Sun

Choose six publicly available cancer datasets (low biomass samples)



Run through the classifier Kraken



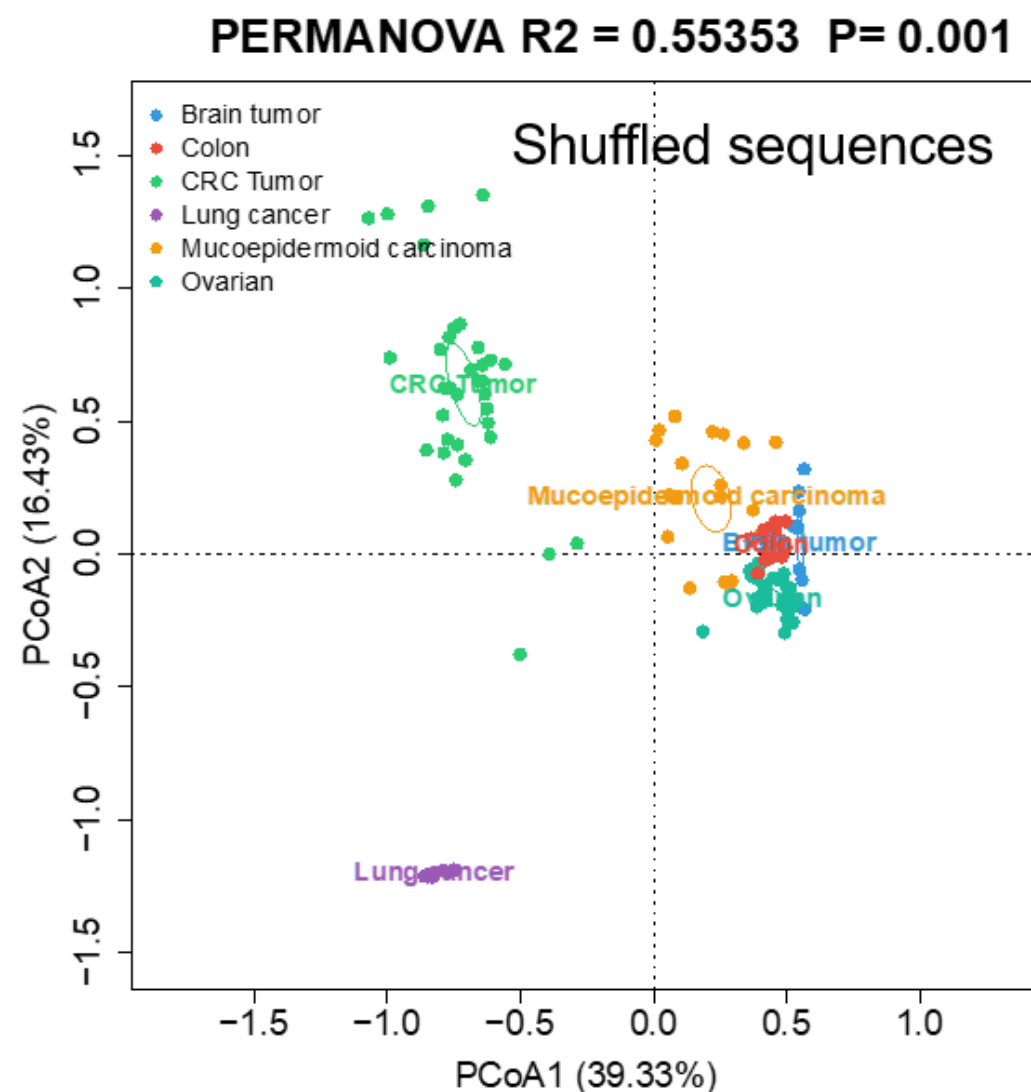
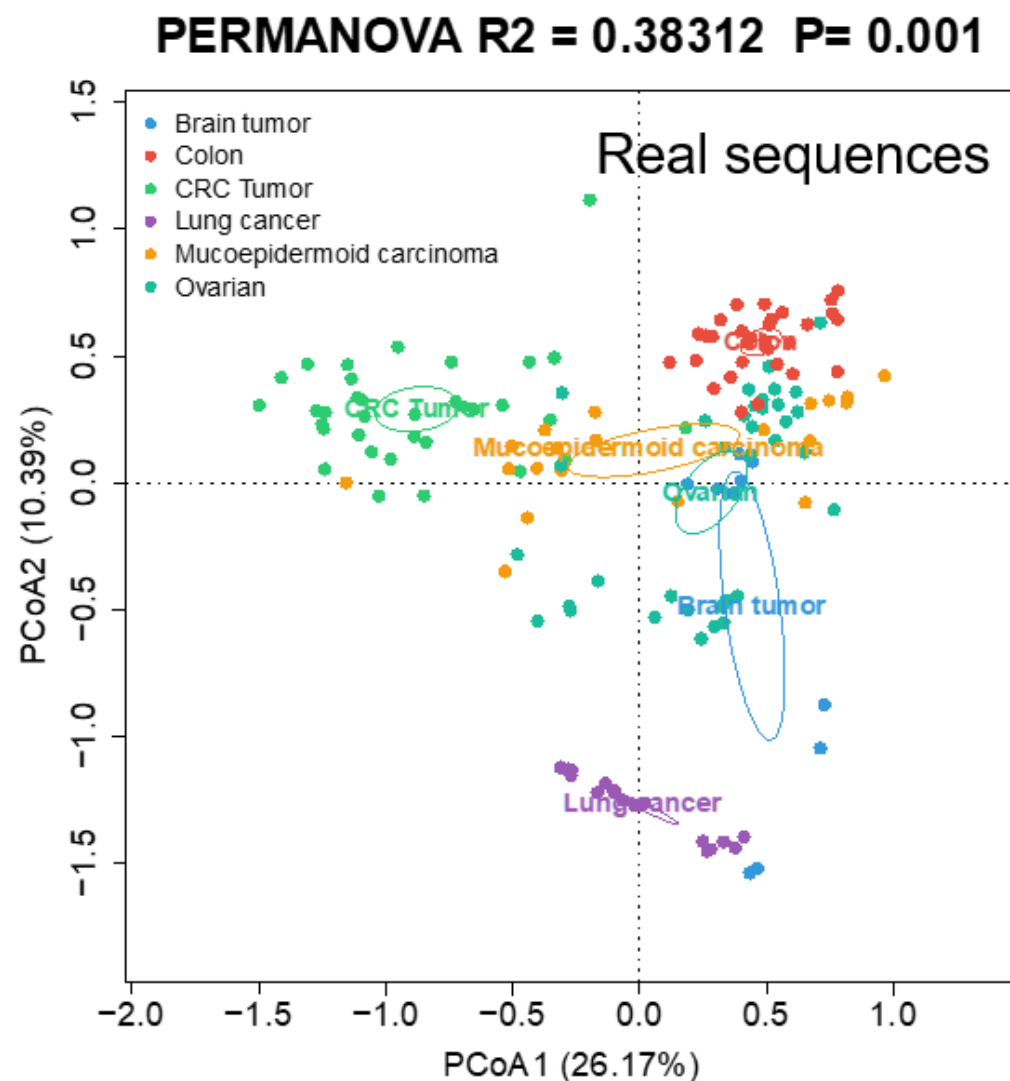
**Shuffle** the sequences.

Run the shuffled sequences through Kraken



Shan Sun

Both shuffled and non-shuffled sequences produce distinct microbial “signatures” in cancer samples



Shan Sun

Conclusions:

Avoid optimism

Design to avoid cross-contamination

Negative controls that are as close as possible in space and time to your original samples can be prioritized. Don't skimp on negative controls.

Share original, un-filtered data

Have a prior that everything in your samples is an artifact and use rigorous statistical inference to convince yourself that your results are real

# Acknowledgements

## UNC-Chapel Hill

Janelle Arthur  
Ian Carroll  
Annie Green Howard  
Marcus Muhlbauer  
Penny Gordon-Larsen  
Jonathan Hansen  
Temitope Keku

## UNC-Charlotte

Roshonda Barner  
Matthew Brown  
Farnaz Fouladi  
Ra'ad Gharaibeh  
Cynthia Gibas  
Jon McCafferty  
Shan Sun  
Michael Sioda  
Kathryn Winglee  
Matthew Brown

## UF- Gainesville

Christian Jobin

## USC

Tanya L Alderete  
Michael Goran  
Emily Noble  
Scott Kanoski

## Vanderbilt University

Martha J. Shrubsole

Kathryn Winglee



Annie Green Howard



Penny Gordon-Larsen



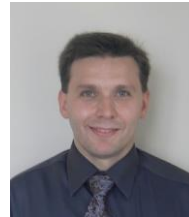
Raad Gharaibeh



Jon  
McCafferty



Marcus Muhlbauer



Janelle Arthur



Christian Jobin







## Rob Knight, Ph.D.

Dr. Knight is the founding director of the Center for Microbiome Innovation and professor of pediatrics, bioengineering, and computer science and engineering at the University of California, San Diego. He is the Wolfe Family Endowed Chair in Microbiome Research and a fellow of both the American Association for the Advancement of Science and the American Academy of Microbiology. He was honored with the 2019 U.S. National Institutes of Health Director's Pioneer Award for his microbiome research and received the 2017 Massry Prize,



## Benjamin Ervin · 3rd

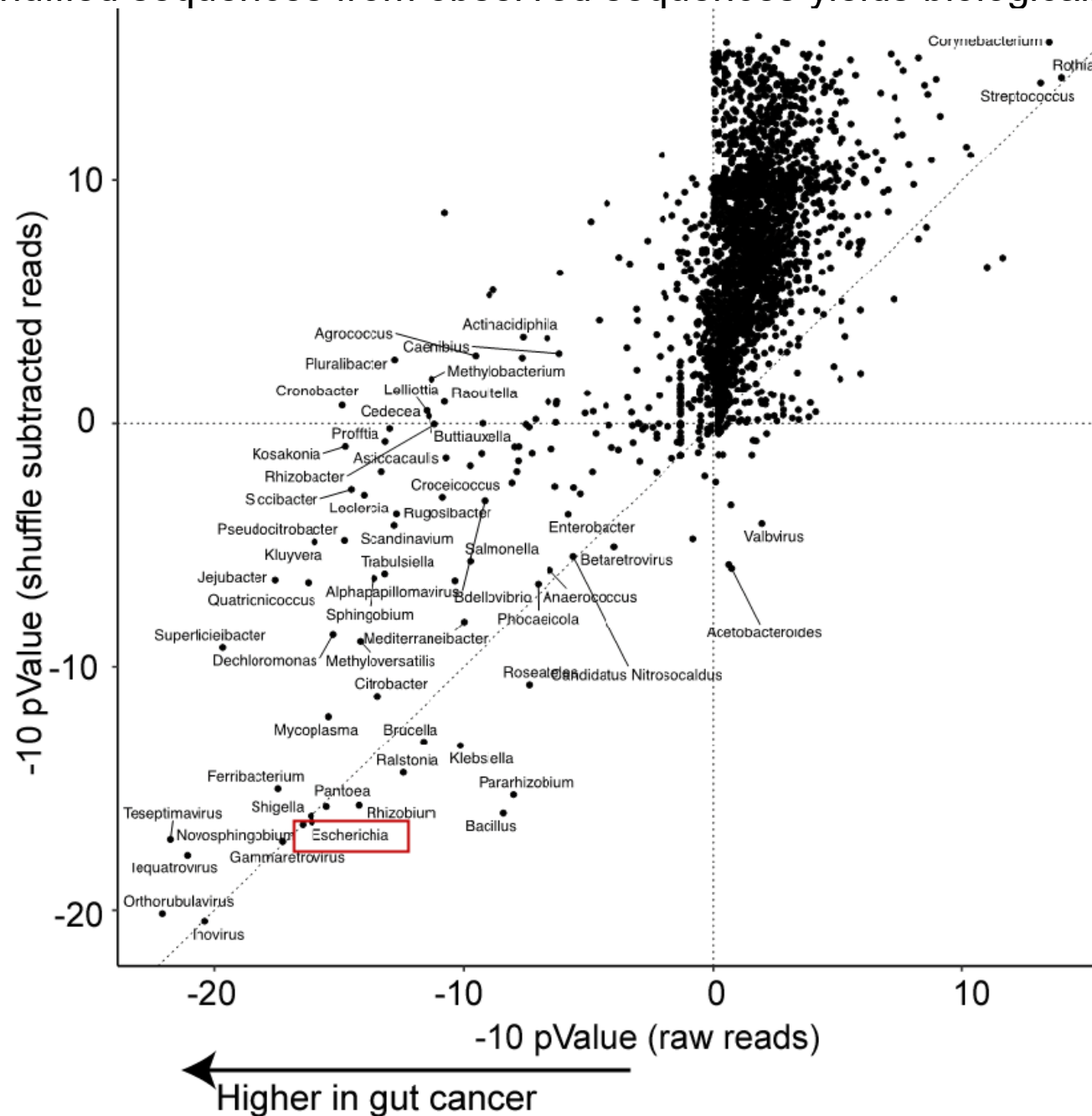
Group Leader at MIT Lincoln Laboratory

Lexington, Massachusetts, United States · [Contact info](#)



Source	Is clearly contaminant	Can be addressed with database searches	Can be solved with long-read requirements	Can be addressed algorithmically	Implications for inference
Non-biological contaminant (e.g. kit contaminant)	Yes. Using a different kit will produce different results.	Probably (especially if a kit has been used in many previous experiments)	Possibly	Possibly (contaminants should be about equally present in all samples irrespective of metadata)	When comparing samples of similar biomass, <b>spurious results are unlikely to be produced.</b>
Biological contaminant (e.g. skin contaminant)	Unclear. If you send a researcher into the built environment, the environment changes. But is that contaminant? Like electrons, measuring the built environment will change the built environment	Possibly (but each person has a distinct skin microbiome!)	No	Possibly (contaminants should be equally present in all samples, but depends on when in the process samples were exposed to human microbiome)	When comparing samples of similar biomass, <b>spurious results are unlikely to be produced.</b>
Cross-	Yes.	No (since each	No	Possibly but only if	<b>Minimal if all blocks</b>

# Subtraction of shuffled sequences from observed sequences yields biologically plausible inference



Shan Sun