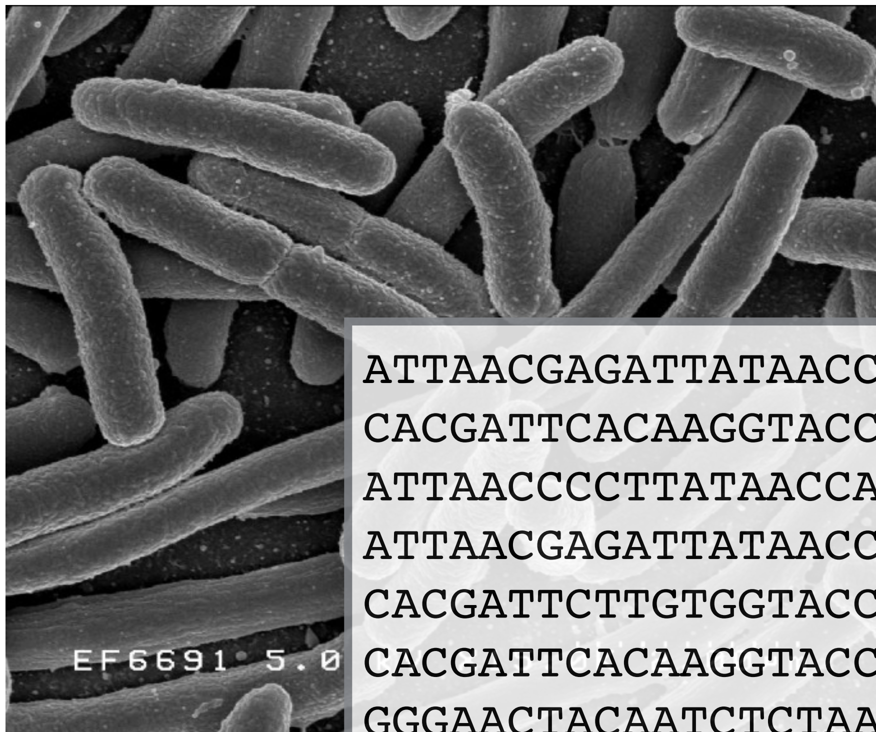
Contamination, controls and **accurate** sequencing-based **measurement** of microbial communities

STAMPS 2023

A Microbial Census




```
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACCCCTTATAACCAGAGTACGAATACCGAACA
ATTAACGAGATTATAACCAGAGAGAGAATACCGAAC
CACGATTCTTGTGGTACCACAAGGTAACATAGCTCC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
GGGAACTACAATCTCTAAGGTGAAGTCTCAGTCTAT
ATTAACGAGATTATAACCAGA
CACGATTCACAAGGTACCACA
ATTAACGAGATTATAACCAGA
```

Lactobacillus crispatus	1300	5	0	882	596
Ureaplasma urealytica	15	0	220	0	0
Gardnerella vaginalis	22	0	1	0	412
Prevotella intermedia	0	0	8	12	0
...	...	...	...	...	...

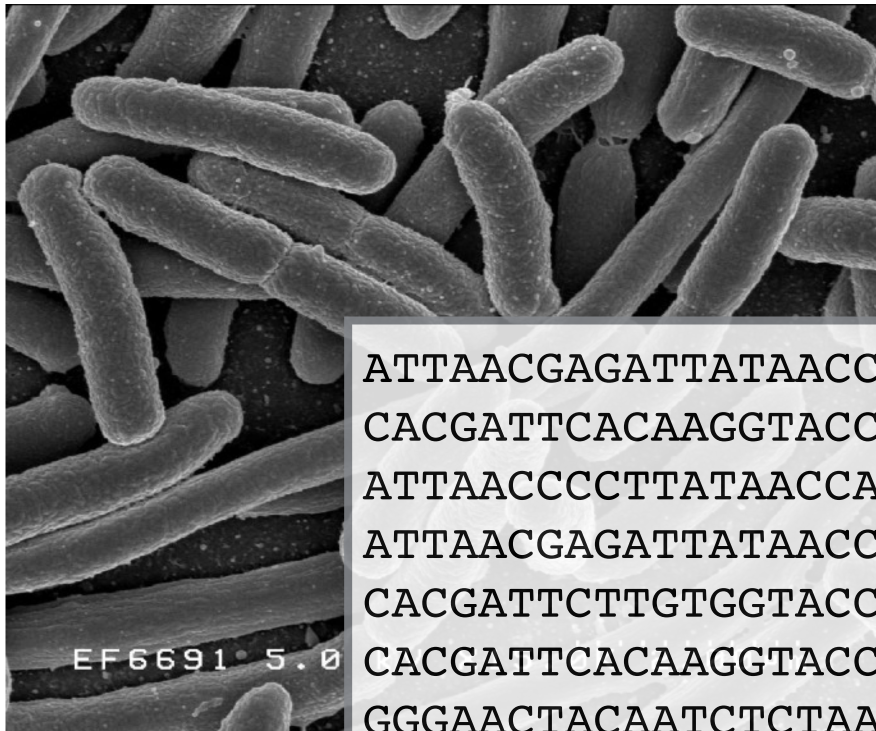
Inference

Visualization

Exploration

A Microbial Census

Marker-gene or Metagenomics Sequencing (MGS)




```
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACCCCTTATAACCAGAGTACGAATACCGAACA
ATTAACGAGATTATAACCAGAGAGAGAATACCGAAC
CACGATTCTTGTGGTACCACAAGGTAACATAGCTCC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
GGGAACTACAATCTCTAAGGTGAAGTCTCAGTCTAT
ATTAACGAGATTATAACCAGA
CACGATTCACAAGGTACCACA
ATTAACGAGATTATAACCAGA
```

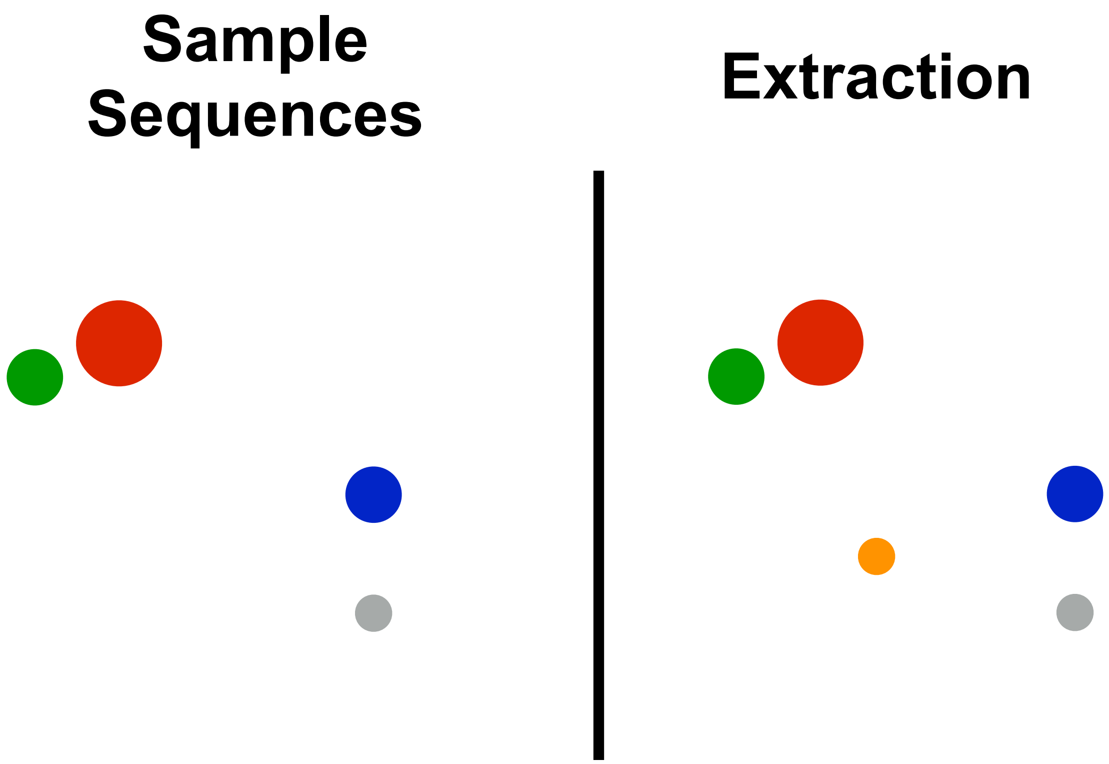
Lactobacillus crispatus	1300	5	0	882	596
Ureaplasma urealytica	15	0	220	0	0
Gardnerella vaginalis	22	0	1	0	412
Prevotella intermedia	0	0	8	12	0
...	...	...	...	...	...

Inference

Visualization

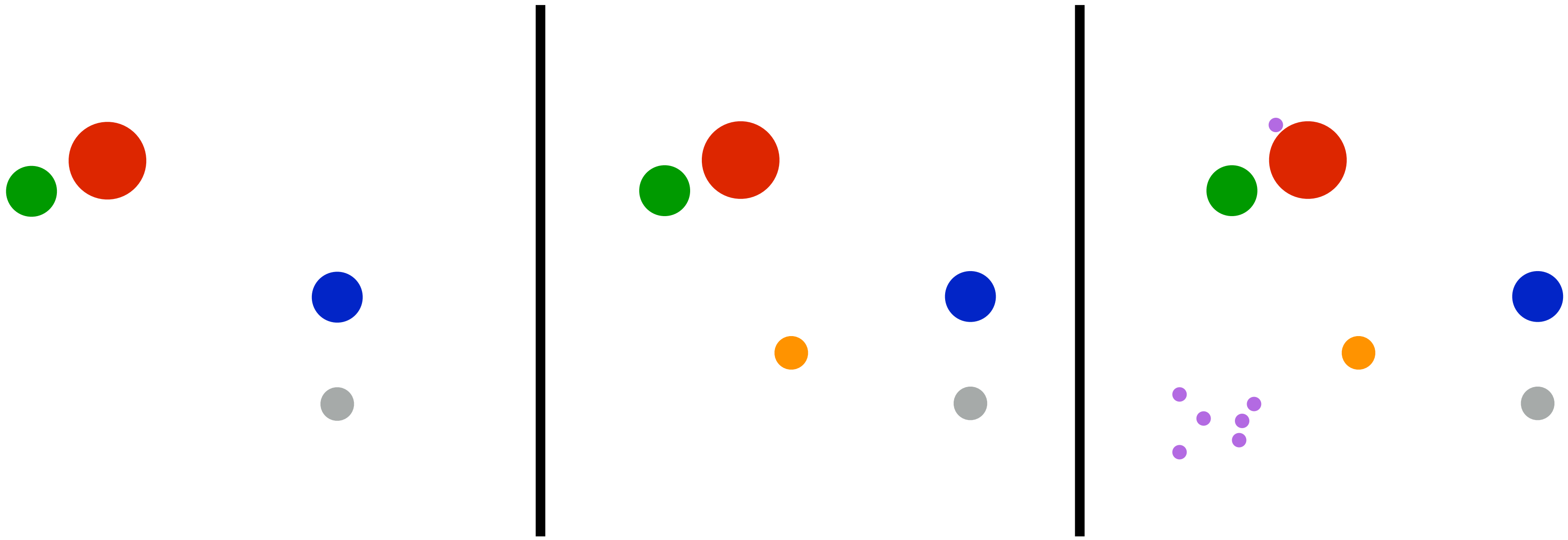
Exploration

MGS: What is really there?

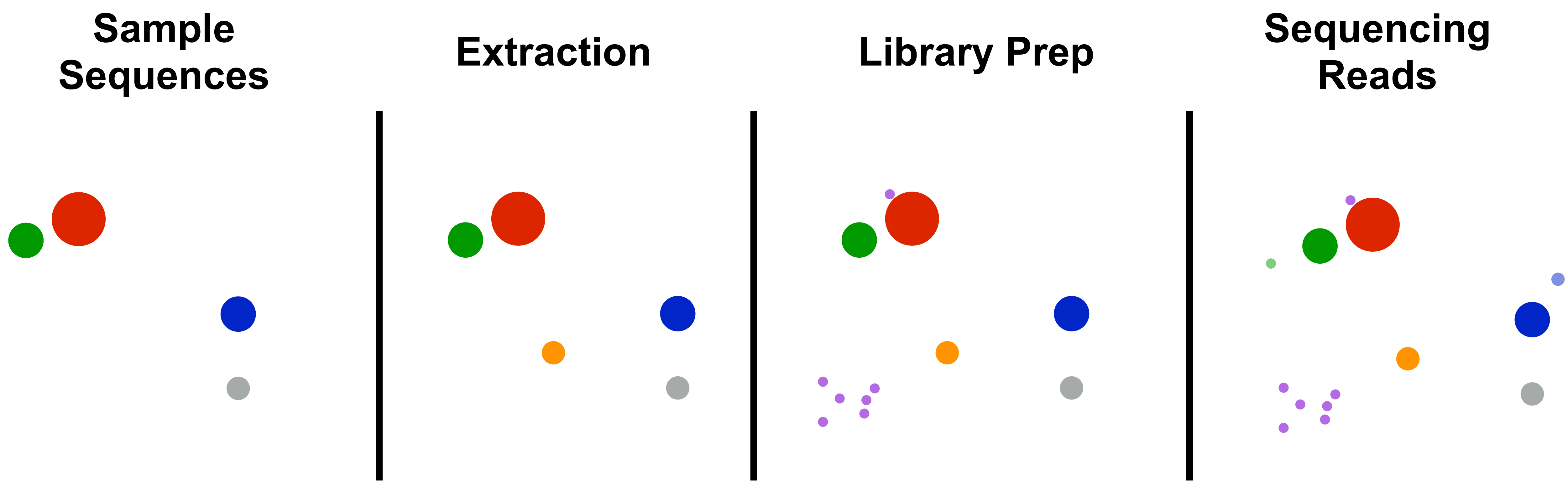


MGS: What is <u>really</u> there?

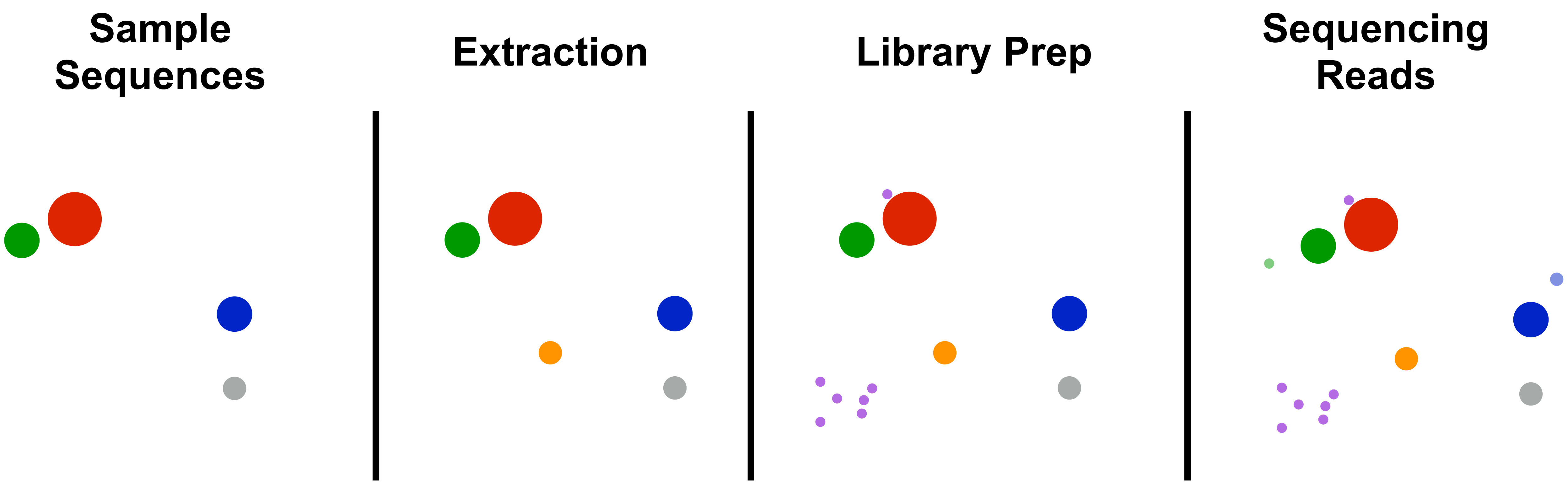
Sample Sequences | **Extraction** | **Library Prep**



MGS: What is really there?

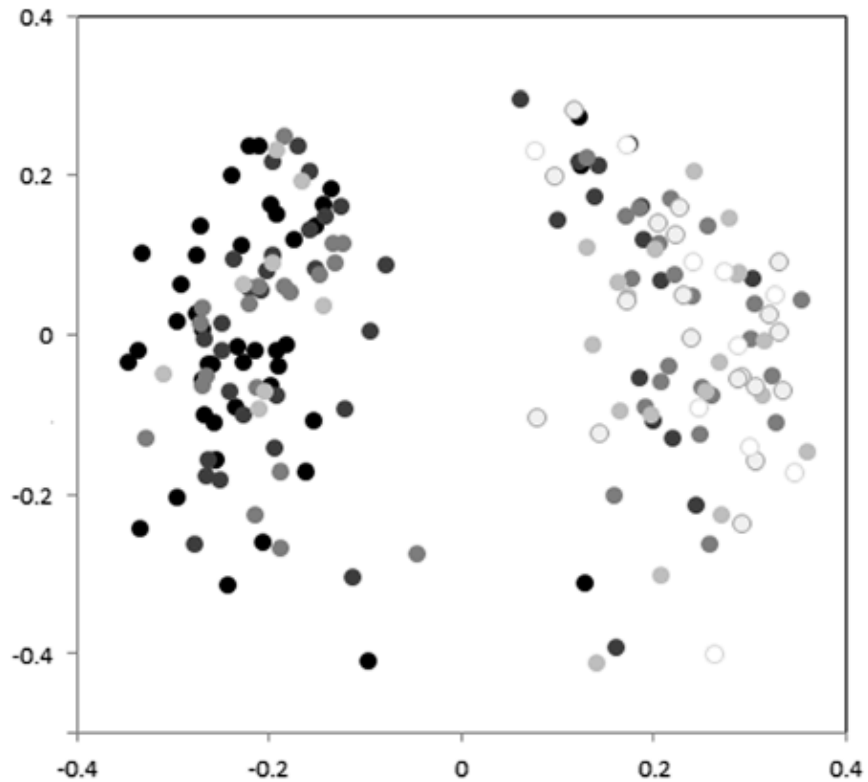


Problem: Contamination



Contaminants — DNA sequences from organisms not truly present in the sample.

Problem: Contamination



a) Full data coloured by age (months)

Figure: Salter, et al. BMC Biology, 2014.

Problem: Contamination

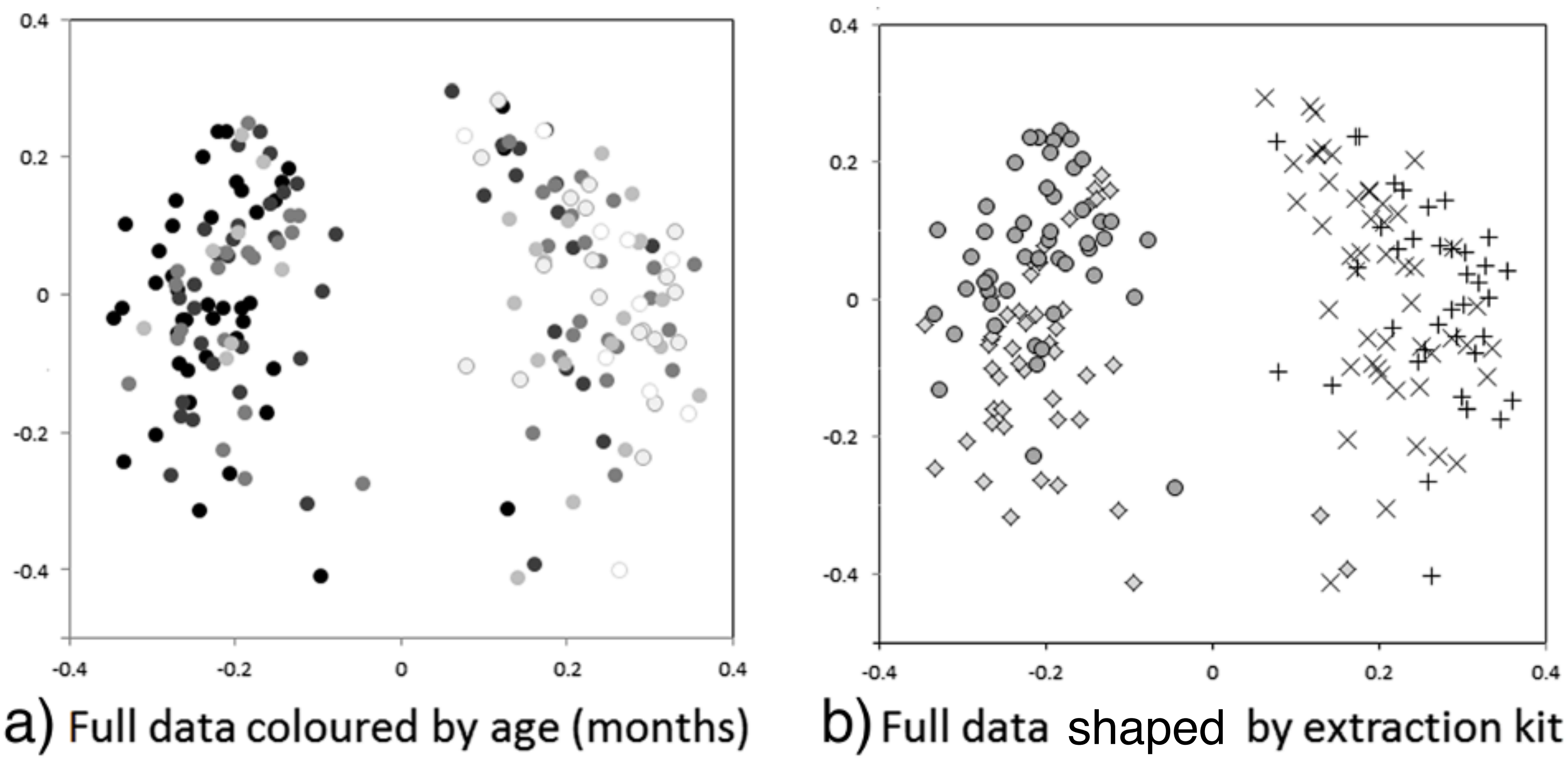
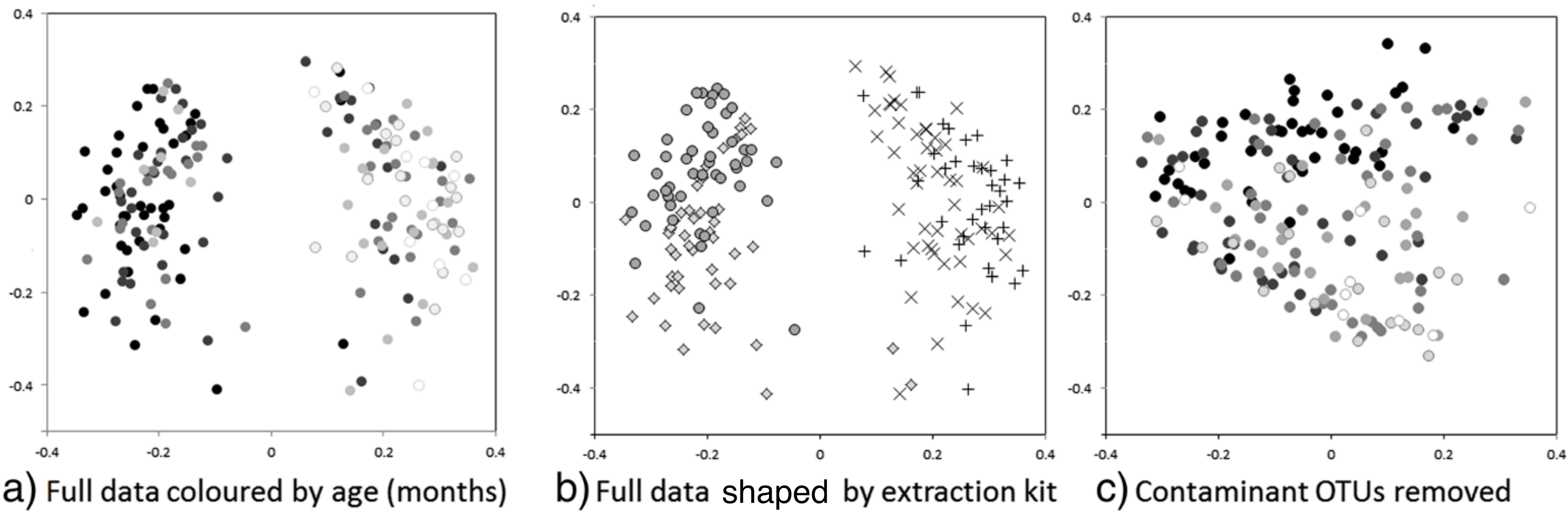


Figure: Salter, et al. BMC Biology, 2014.

Problem: Contamination



Spurious signal driven by contaminants!

Figure: Salter, et al. BMC Biology, 2014.

Problem: Contamination

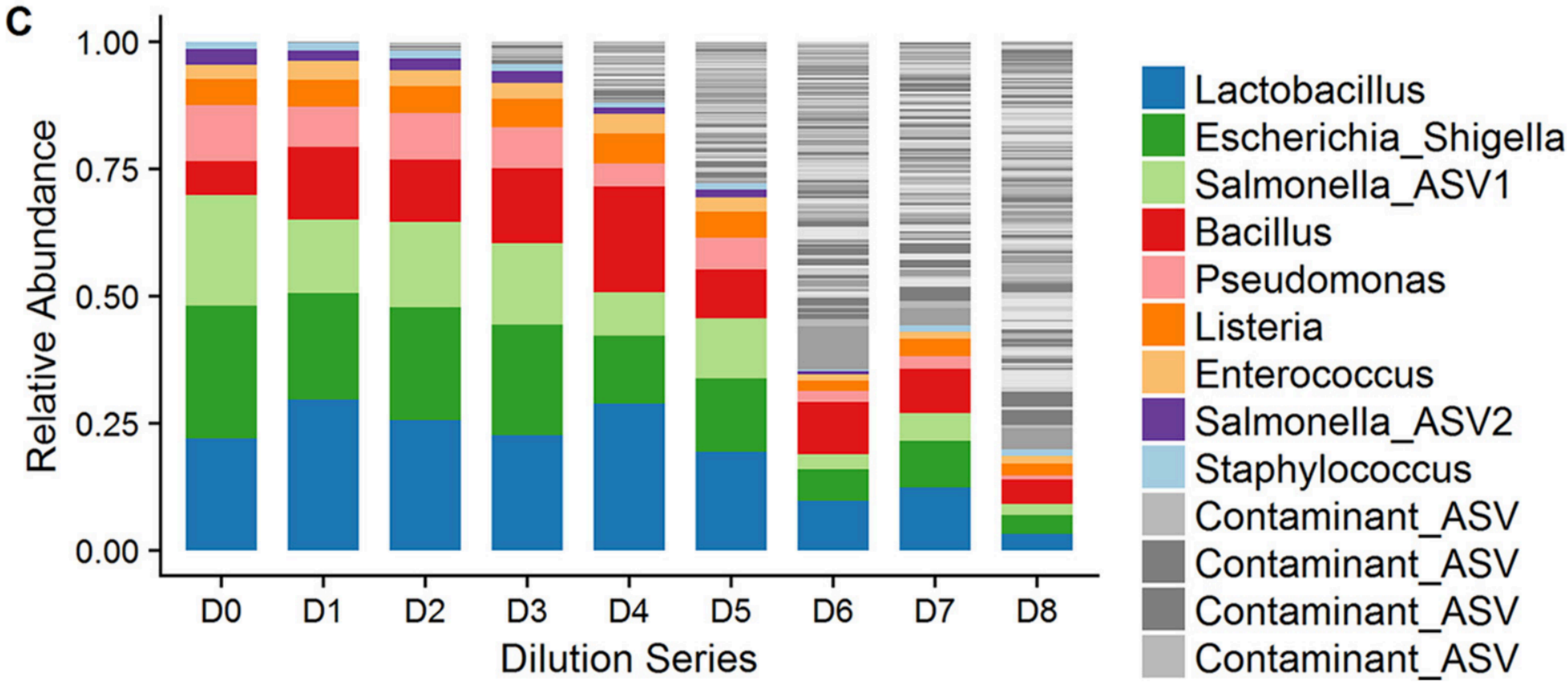


Figure: Karstens, et al. mSystems, 2019.

Problem: Contamination

Science Translational Medicine

Current Issue Firs

RESEARCH ARTICLE | MICROBIOME

The Placenta Harbors a Unique Microbiome

KJERSTI AAGAARD , JUN MA, KATHLEEN M. ANTONY, RADHIKA GANU, JOSEPH PETROSINO, AND JAMES VERSALOVIC Authors Info & Affiliations

SCIENCE TRANSLATIONAL MEDICINE • 21 May 2014 • Vol 6, Issue 237 • p. 237ra65 • DOI: 10.1126/scitranslmed.3008599

Bacteria in Healthy Placentas

Contrary to the prevailing idea of a "sterile" intrauterine environment, Aagaard and coauthors demonstrated the consistent presence of a microbiome in placentas from healthy pregnancies. This microbiome was consistently different from those reported in other parts of the body, including the skin and urogenital tract. The placental microbiome was most similar to that of the oral cavity, but the clinical implications of this finding remain to be explored. In addition, the authors identified associations between the composition of the placental microbiome and a history of remote antenatal infection, as well as preterm birth, raising the possibility that the placental microbiome may play a role in these events.

Problem: Contamination

Article | Published: 31 July 2019

Human placenta has no microbiome but can contain potential pathogens

Marcus C. de Goffau, Susanne Lager, Ulla Sovio, Francesca Gaccioli, Emma Cook, Sharon J. Peacock, Julian Parkhill, D. Stephen Charnock-Jones & Gordon C. S. Smith

Nature **572**, 329–334 (2019) | Cite this a

27k Accesses | **326** Citations | **643** Alt

EDITORIAL | VOLUME 220, ISSUE 3, P213-214, MARCH 01, 2019

De-Discovery of the Placenta Microbiome

Frederic D. Bushman, PhD

Lack of detection of a human placenta microbiome in samples from preterm and term deliveries

Jacob S. Leiby, Kevin McCormick, Scott Sherrill-Mix, Erik L. Clarke, Lyanna R. Kessler, Louis J. Taylor, Casey E. Hofstaedter, Aoife M. Roche, Lisa M. Mattei, Kyle Bittinger, Michal A. Elovitz, Rita Leite, Samuel Parry & Frederic D. Bushman

Microbiome **6**, Article number: 196 (2018) | Cite this article

8898 Accesses | **143** Citations | **110** Altmetric | Metrics

Problem: Contamination

Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA

Mark Kowarsky[a], Joan Camunas-Soler[b], Michael Kertesz[b,1], Iwijn De Vlaminck[b], Winston Koh[b], Wenying Pan[b], Lance Martin[b], Norma F. Neff[b,c], Jennifer Okamoto[b,c], Ronald J. Wong[d], Sandhya Kharbanda[e], Yasser El-Sayed[f], Yair Blumenfeld[f], David K. Stevenson[d], Gary M. Shaw[d], Nathan D. Wolfe[g,h], and Stephen R. Quake[b,c,i,2]

[a]Department of Physics, Stanford University, Stanford, CA 94305; [b]Department of Bioengineering, Stanford University, Stanford, CA 94305; [c]Chan Zuckerberg Biohub, San Francisco, CA 94158; [d]Department of Pediatrics, Stanford University School of Medicine, Stanford University, Stanford, CA 94305; [e]Pediatric Stem Cell Transplantation, Lucille Packard Children's Hospital, Stanford University, Stanford, CA 94305; [f]Division of Maternal–Fetal Medicine, Department of Obstetrics and Gynecology, Stanford University School of Medicine, Stanford University, Stanford, CA 94305; [g]Metabiota, San Francisco, CA 94104; [h]Global Viral, San Francisco, CA 94104; and [i]Department of Applied Physics, Stanford University, Stanford, CA 94305

Contributed by Stephen R. Quake, July 12, 2017 (sent for review April 28, 2017; reviewed by Søren Brunak and Eran Segal)

Blood circulates throughout the human body and contains molecules drawn from virtually every tissue, including the microbes and viruses which colonize the body. Through massive shotgun sequencing of circulating cell-free DNA from the blood, we identified hundreds of new bacteria and viruses which represent previously unidentified members of the human microbiome. Analyzing cumulative sequence data from 1,351 blood samples collected from 188 patients enabled us to assemble 7,190 contiguous regions (contigs) larger than 1 kbp, of which 3,761 are novel with little or no sequence homology in any existing databases. The vast majority of these novel contigs possess coding sequences, and we have validated their existence both by finding their presence in independent experiments and by performing direct PCR amplification. When their nearest neighbors are located in the tree of life, many of the organisms represent entirely novel taxa, showing that microbial diversity within the human body is substantially broader than previously appreciated.

cell-free DNA | microbiome | metagenomics | biological dark matter

the body (18, 19); combining this observation with the average genome sizes of a human, bacterium, and virus (Gb, Mb, and kb, respectively) suggests that approximately 1% of DNA by mass in a human is derived from nonhost origins. Previous studies by us and others have shown that indeed approximately 1% of cfDNA sequences appear to be of nonhuman origin, but only a small fraction of these map to existing databases of microbial and viral genomes (16). This suggests that there is a vast diversity of as yet uncharacterized microbial diversity within the human microbiome and that this diversity can be analyzed through "unmappable" sequencing reads.

We analyzed the cfDNA-derived microbiomes of 1,351 samples from 188 patients in four longitudinally sampled cohorts—heart transplant (HT), 610 samples (76 patients); lung transplant (LT), 460 samples (59 patients); bone marrow transplant (BMT), 161 samples (21 patients); and pregnancy (PR), 120 samples (32 patients)—and discovered that the majority of assembled

Problem: Contamination

Numerous uncharacterized and highly divergent microbes which colonize humans are revealed

Candidate Phyla Radiation in Human Blood?

A. Murat Eren (Meren)

Web Email Twitter Github

Twitter is bad. I mostly follow scientists, and often end up running into interesting findings from other groups that make me want to take a quick look at their data. Although most of our procrastinations don't end up on the blog, sometimes they do: 1, 2, 3. Well, today was one of those days.

ing of circulating cell-free DNA from the blood, we identified

homology in any existing databases. The vast majority of these novel contigs possess coding sequences, and we have validated their existence both by finding their presence in independent experiments and by performing direct PCR amplification. When their nearest neighbors are located in the tree of life, many of the organisms represent entirely novel taxa, showing that microbial diversity within the human body is substantially broader than previously appreciated.

cell-free DNA | microbiome | metagenomics | biological dark matter

mass in a human is derived from nonhost origins. Previous

the human microbiome and that this diversity can be analyzed through "unmappable" sequencing reads.

We analyzed the cfDNA-derived microbiomes of 1,351 samples from 188 patients in four longitudinally sampled cohorts—heart transplant (HT), 610 samples (76 patients); lung transplant (LT), 460 samples (59 patients); bone marrow transplant (BMT), 161 samples (21 patients); and pregnancy (PR), 120 samples (32 patients)—and discovered that the majority of assembled

https://merenlab.org/2017/08/23/CPR-in-blood/

Problem: Contamination

Now what?

Modeling Contaminants

$\mathbf{T = S + C}$, where $\mathbf{C}$ is constant

hence

$\mathbf{f_C = C/(S+C) \sim 1/T}$, where $\mathbf{C \ll S}$

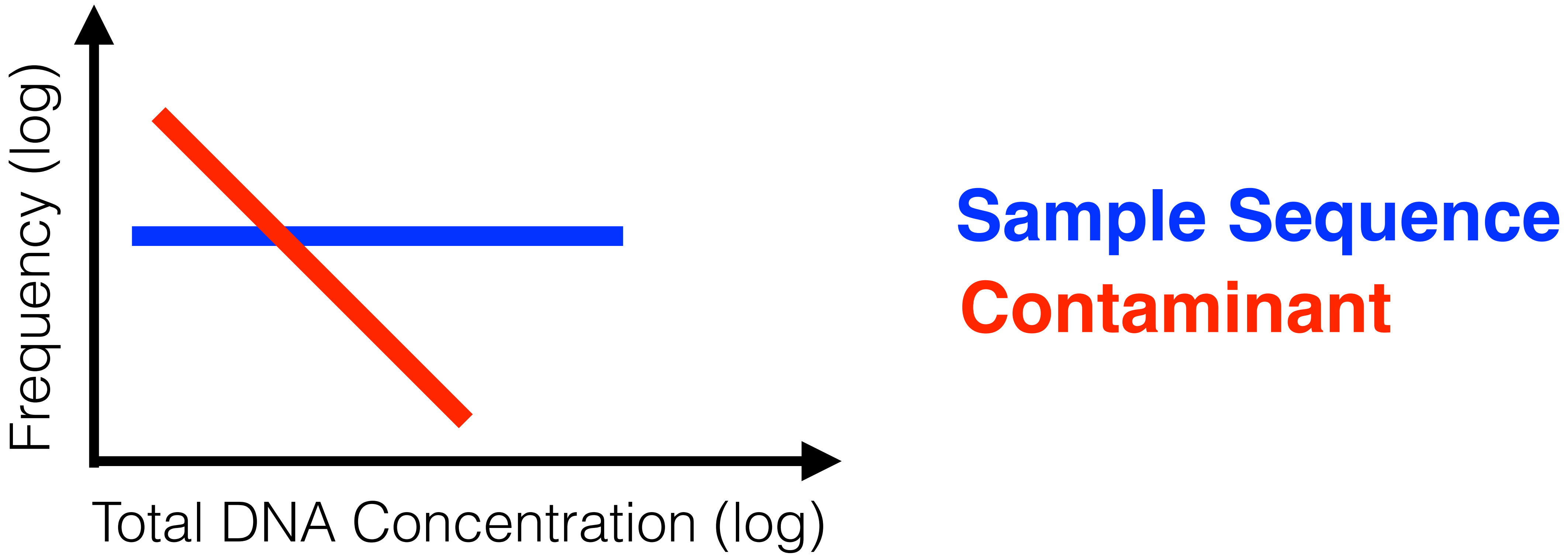
Davis, et al. Microbiome, 2018.

Modeling Contaminants

$$\mathbf{T = S + C}\text{, where } \mathbf{C} \text{ is constant}$$

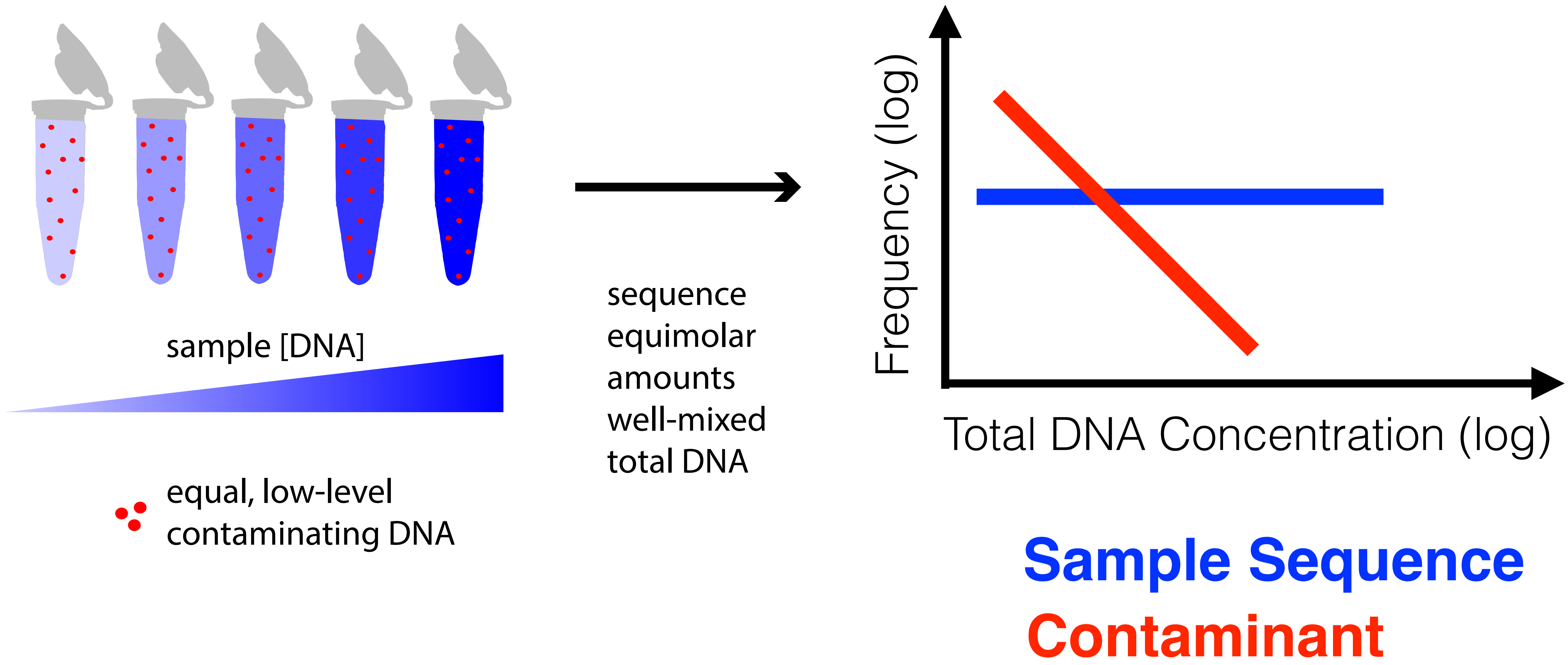
hence

$$\mathbf{f_C = C/(S+C) \sim 1/T}\text{, where } \mathbf{C \ll S}$$



Davis, et al. Microbiome, 2018.

Modeling Contaminants



Davis, et al. Microbiome, 2018.

Decontam Method

Frequency

Input: DNA concentrations,
Feature table w/ abundances.

Output: Score 0 (contaminant) - 1 (non-contaminant),
Binary classification based on threshold.

Decontam Method

Frequency

Input: DNA concentrations,
Feature table w/ abundances.

Output: Score 0 (contaminant) - 1 (non-contaminant),
Binary classification based on threshold.

```
contam <- isContaminant(seqtab, is.neg
```

Decontam Method

Frequency

Input: DNA concentrations,
Feature table w/ abundances.

Output: Score 0 (contaminant) - 1 (non-contaminant),
Binary classification based on threshold.

Prevalence

Input: Categorization of samples as negative controls,
Feature table w/ abundances or presences.

Output: Score 0 (contaminant) - 1 (non-contaminant)
Binary classification based on threshold.

Decontam Method

Needs range of DNA concentrations

Frequency

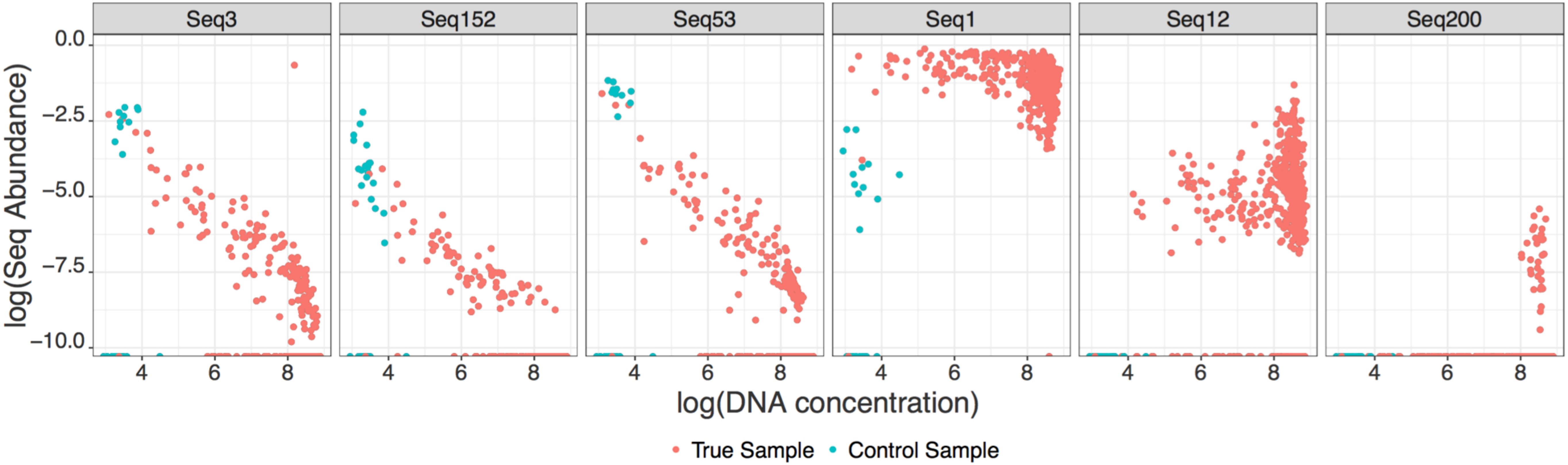
Input: DNA concentrations,
Feature table w/ abundances.
Output: Score 0 (contaminant) - 1 (non-contaminant),
Binary classification based on threshold.

Needs multiple (5+) sequenced negative controls

Prevalence

Input: Categorization of samples as negative controls,
Feature table w/ abundances or presences.
Output: Score 0 (contaminant) - 1 (non-contaminant)
Binary classification based on threshold.

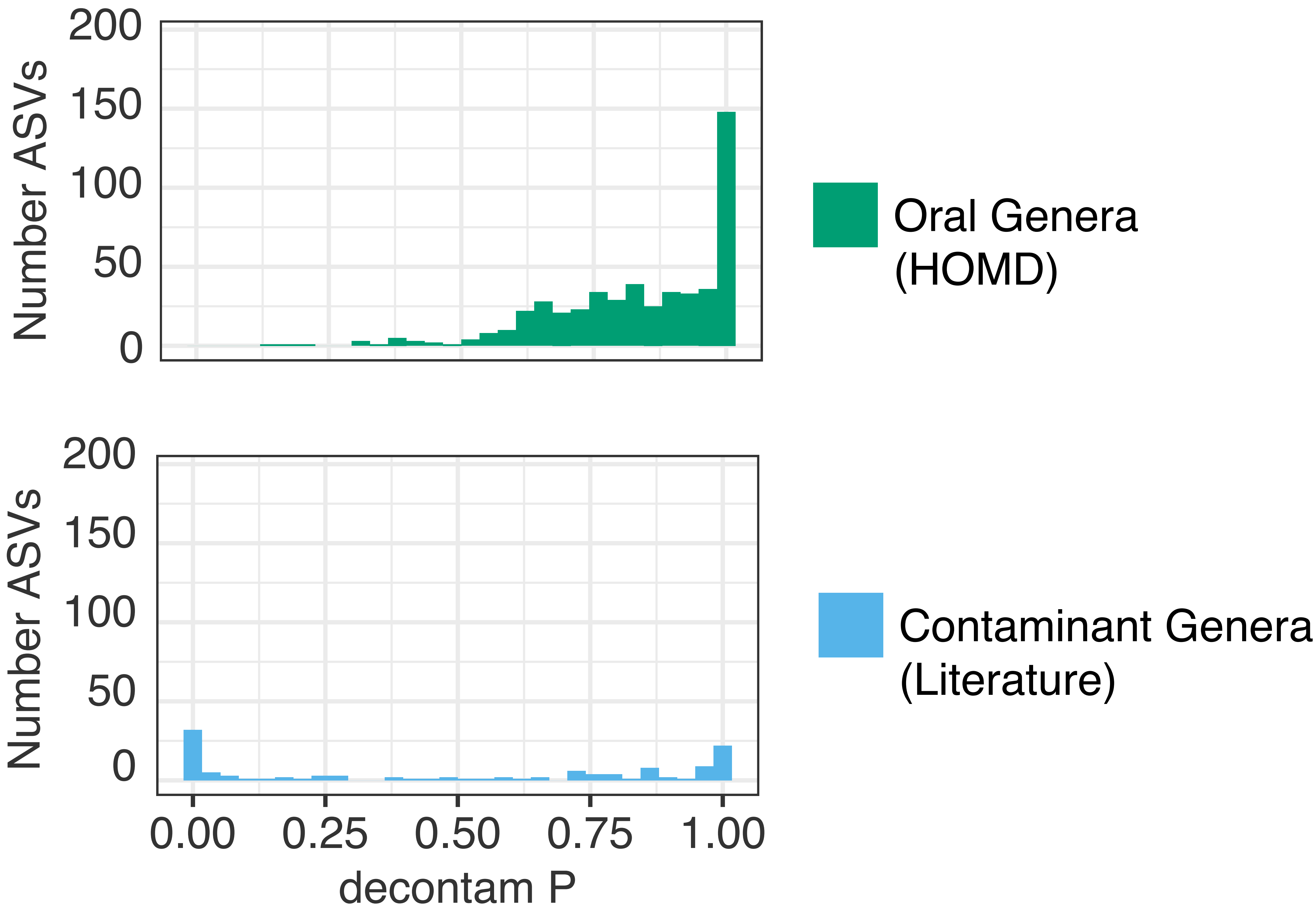
Validating the Model



Davis, et al. Microbiome, 2018.

Validating the Model

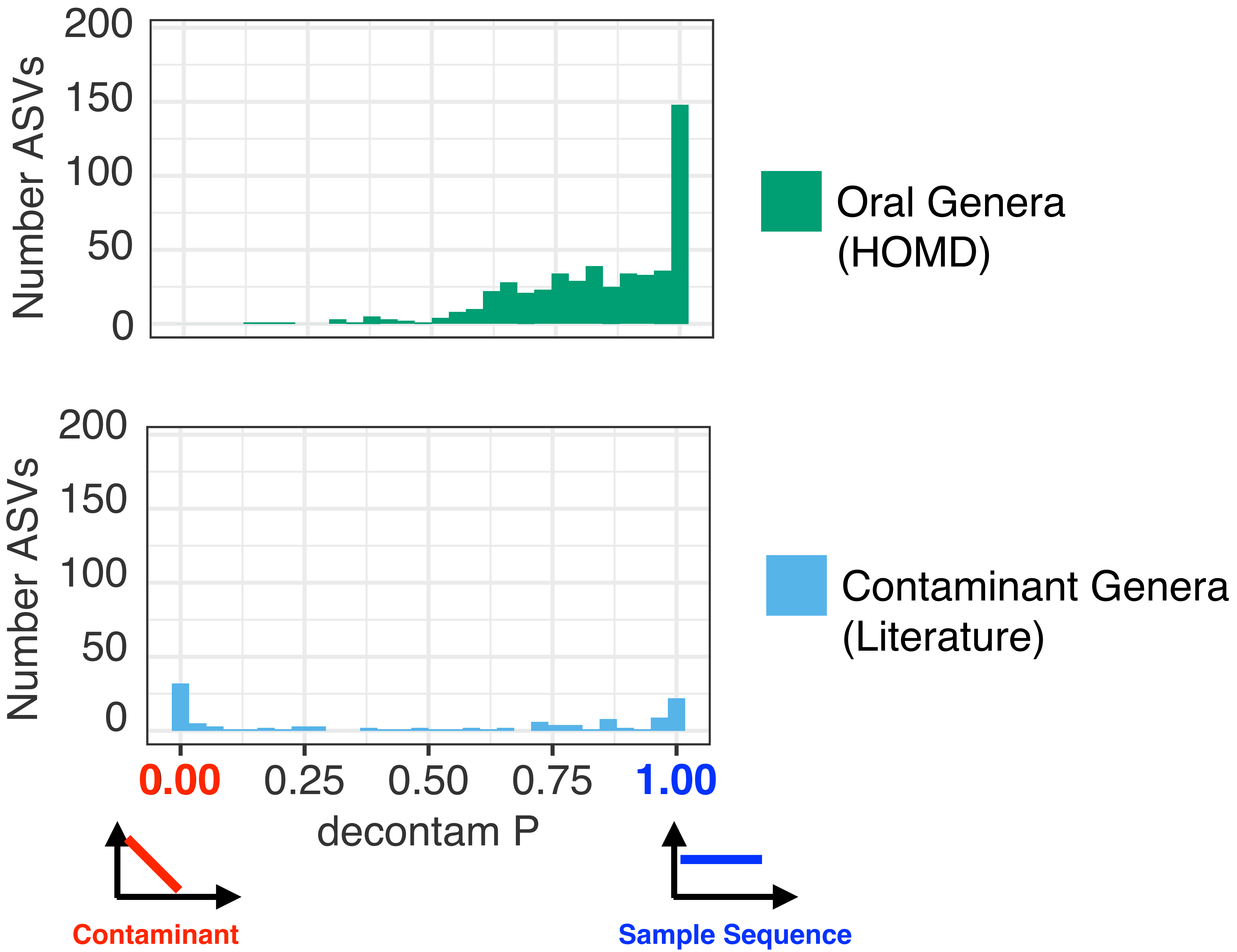
Oral Mucosal Dataset



Davis, et al. Microbiome, 2018.

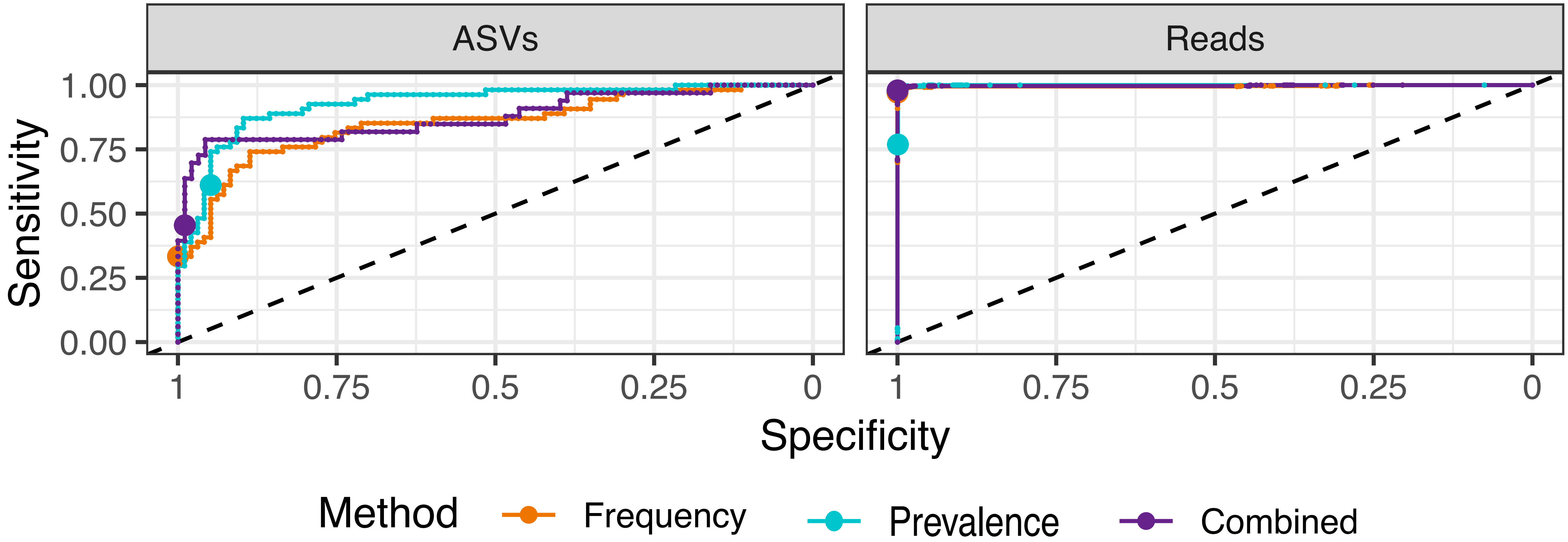
Validating the Model

Oral Mucosal Dataset



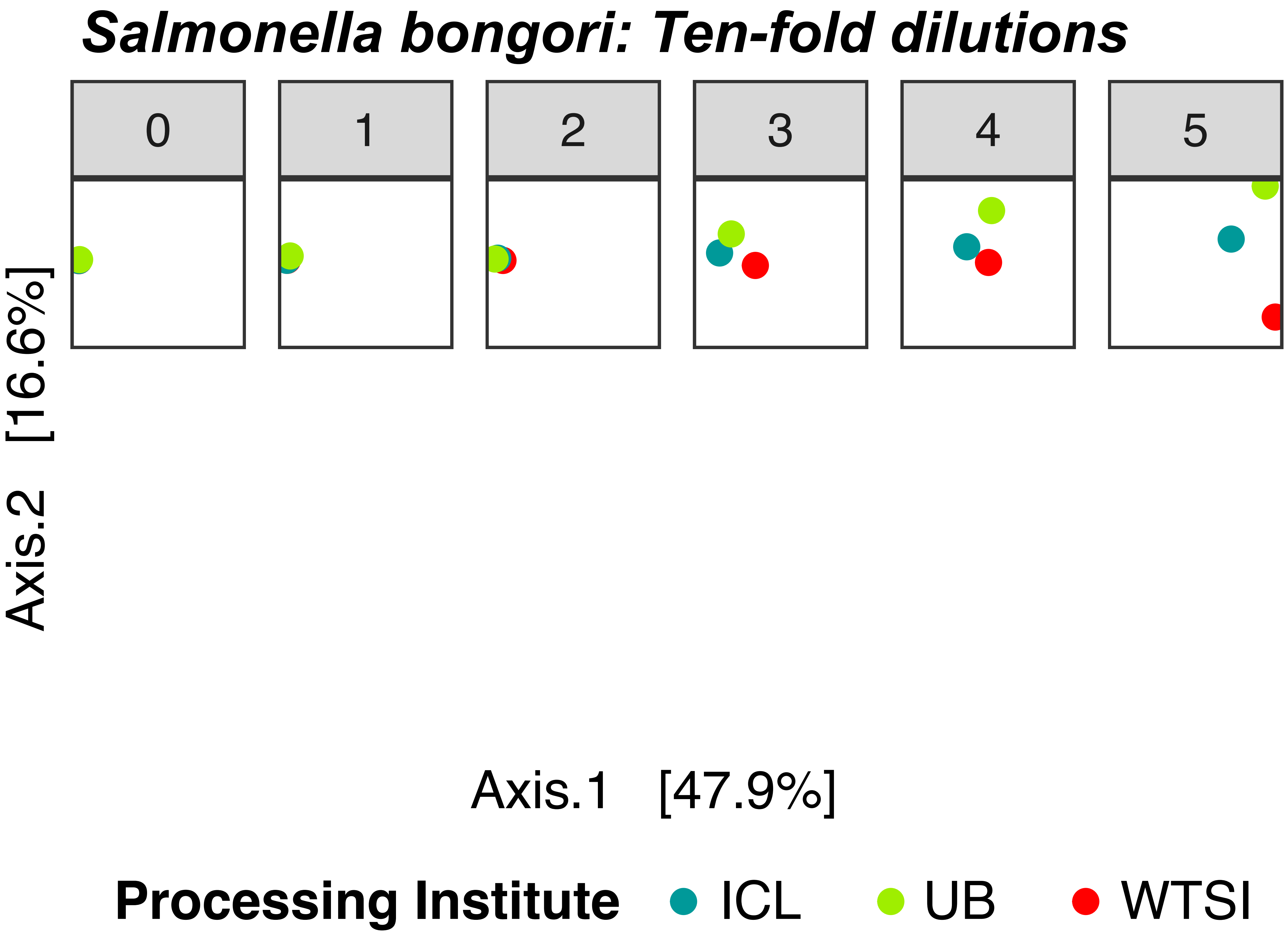
Classification Accuracy

Oral Mucosal Dataset



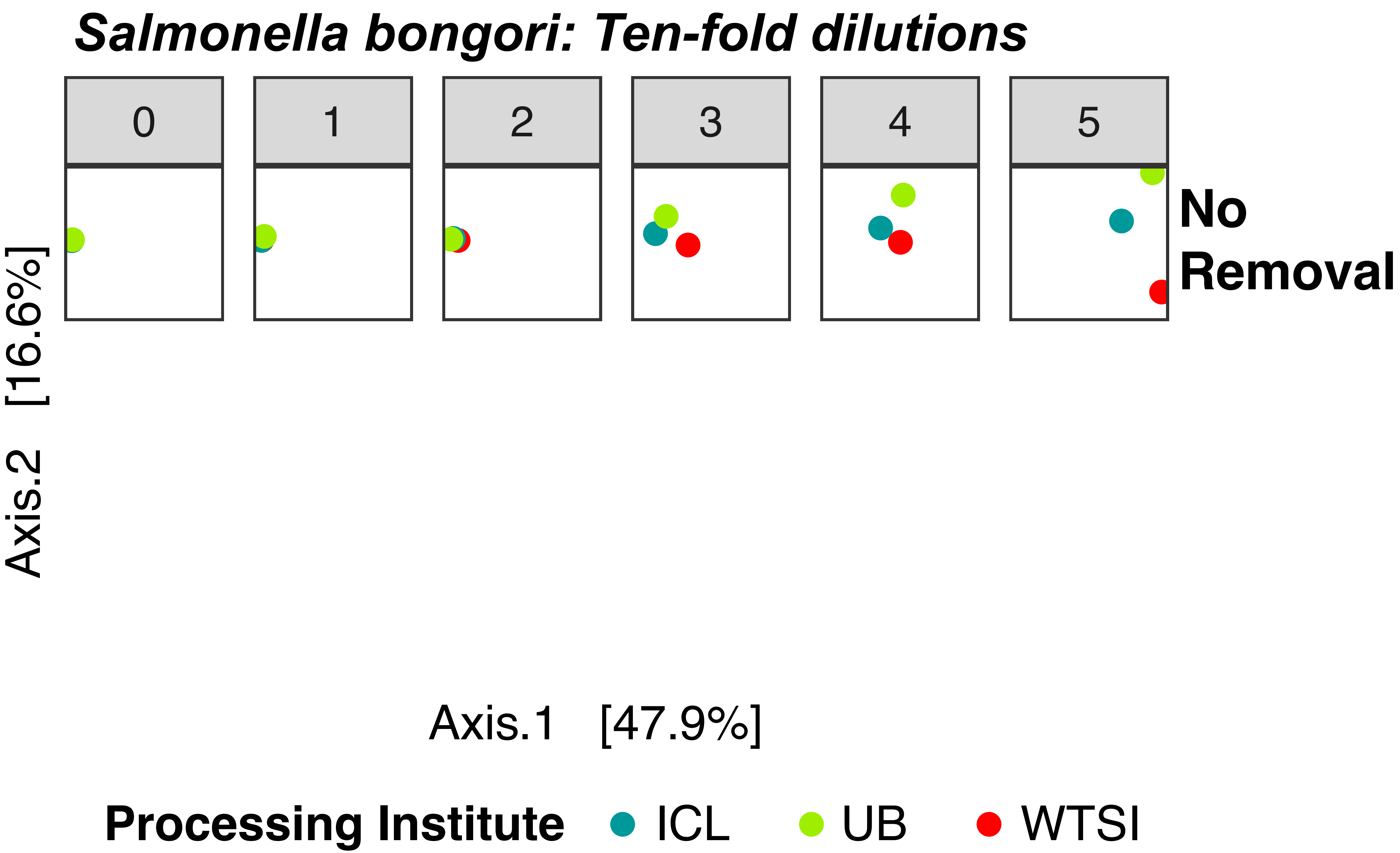
Davis, et al. Microbiome, 2018.

Reducing Technical Variation



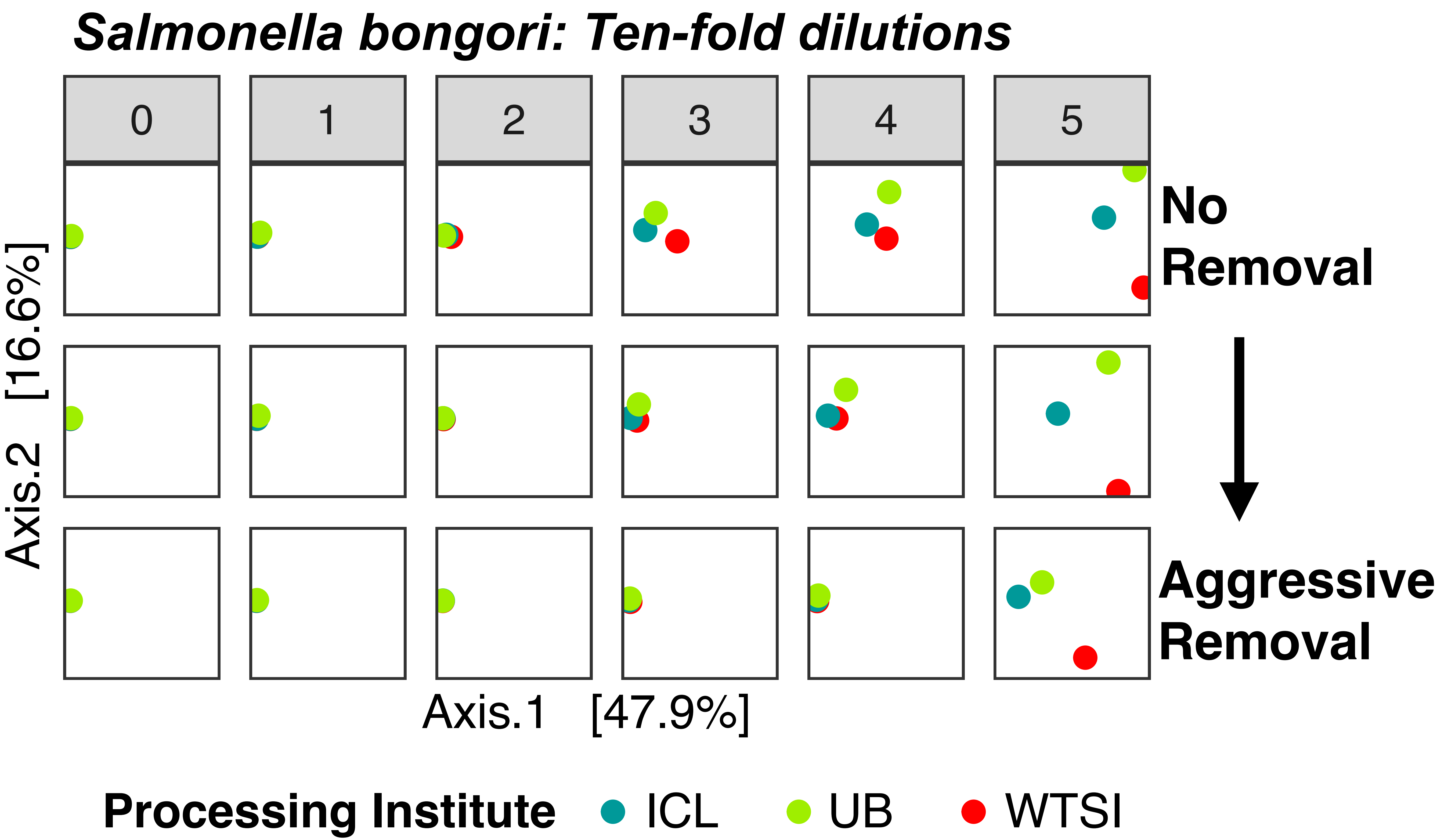
Davis, et al. Microbiome, 2018.

Reducing Technical Variation



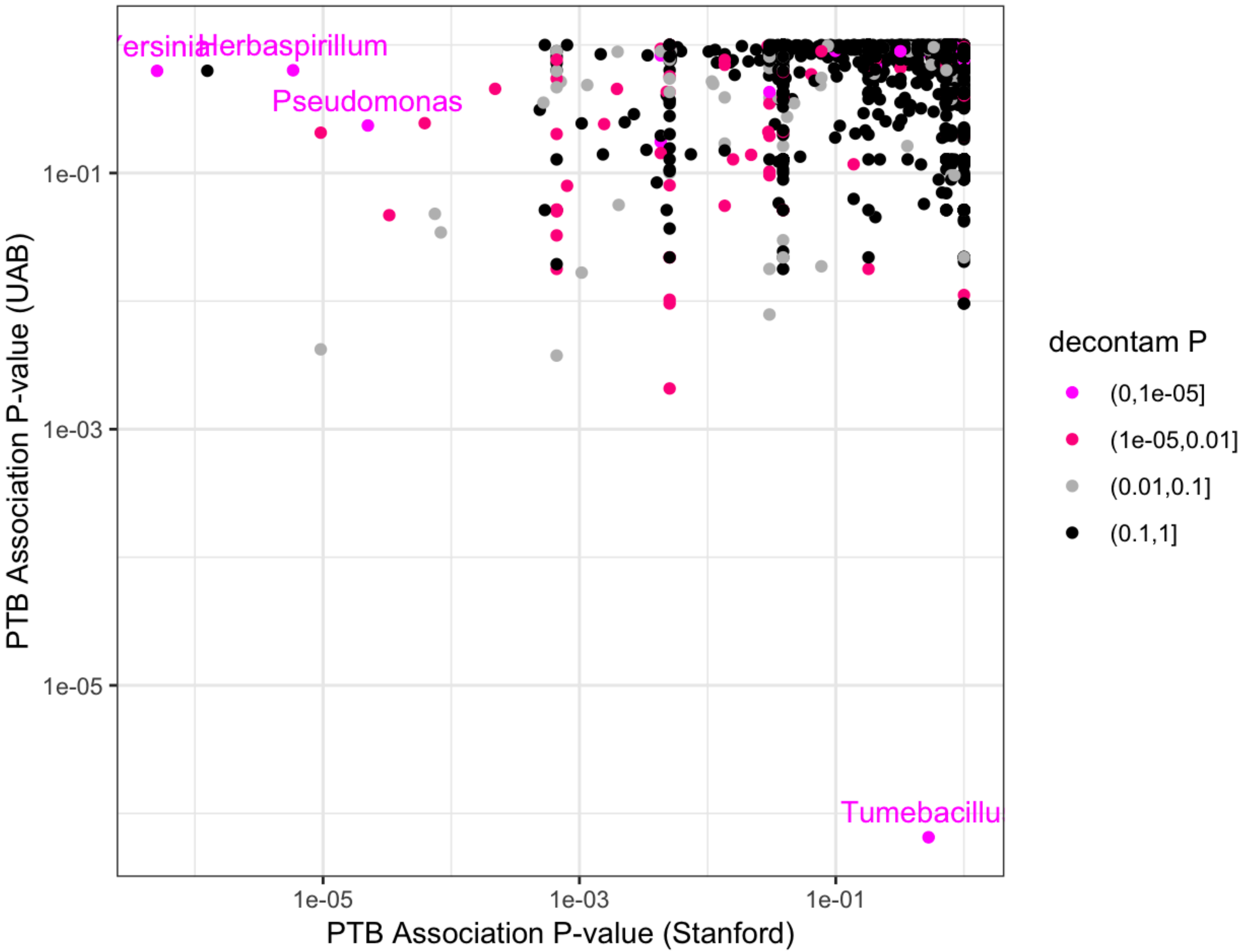
Davis, et al. Microbiome, 2018.

Reducing Technical Variation



Davis, et al. Microbiome, 2018.

Avoiding Spurious Results



Davis, et al. Microbiome, 2018.

Available now…

Methodology | Open Access

Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data

Nicole M. Davis, Diana M. Proctor, Susan P. Holmes, David A. Relman and Benjamin J. Callahan

Microbiome 2018 **6**:226

DECONTAM

- Open-source
- Well documented
- R package
- 16S or shotgun

Decontam Method

Frequency

Input: DNA concentrations,
Feature table w/ abundances.

Output: Score 0 (contaminant) - 1 (non-contaminant),
Binary classification based on threshold.

```
contam <- isContaminant(seqtab, conc, threshold)
```

Decontam Method

<u>Frequency</u>

Input: DNA concentrations,
Feature table w/ abundances.

Output: Score 0 (contaminant) - 1 (non-contaminant),
Binary classification based on threshold.

```
contam <- isContaminant(seqtab, conc, threshold)
```

ASV or OTU table
(or phyloseq object)

Vector of DNA concentrations
(or phyloseq variable name)

Number: 0 to 1
(default 0.5)

Decontam Method

Prevalence

Input: Categorization of samples as negative controls,
Feature table w/ abundances or presences.

Output: Score 0 (contaminant) - 1 (non-contaminant)
Binary classification based on threshold.

Decontam Method

Prevalence

Input: Categorization of samples as negative controls,
Feature table w/ abundances or presences.

Output: Score 0 (contaminant) - 1 (non-contaminant)
Binary classification based on threshold.

```
contam <- isContaminant(seqtab, neg, threshold)
```

Decontam Method

Prevalence

Input: Categorization of samples as negative controls,
Feature table w/ abundances or presences.

Output: Score 0 (contaminant) - 1 (non-contaminant)
Binary classification based on threshold.

```
contam <- isContaminant(seqtab, neg, threshold)
```

ASV or OTU table
(or phyloseq object)

Vector: True if neg control, False otherwise
(or phyloseq variable name)

Number: 0 to 1
(default 0.5)

Decontam Method

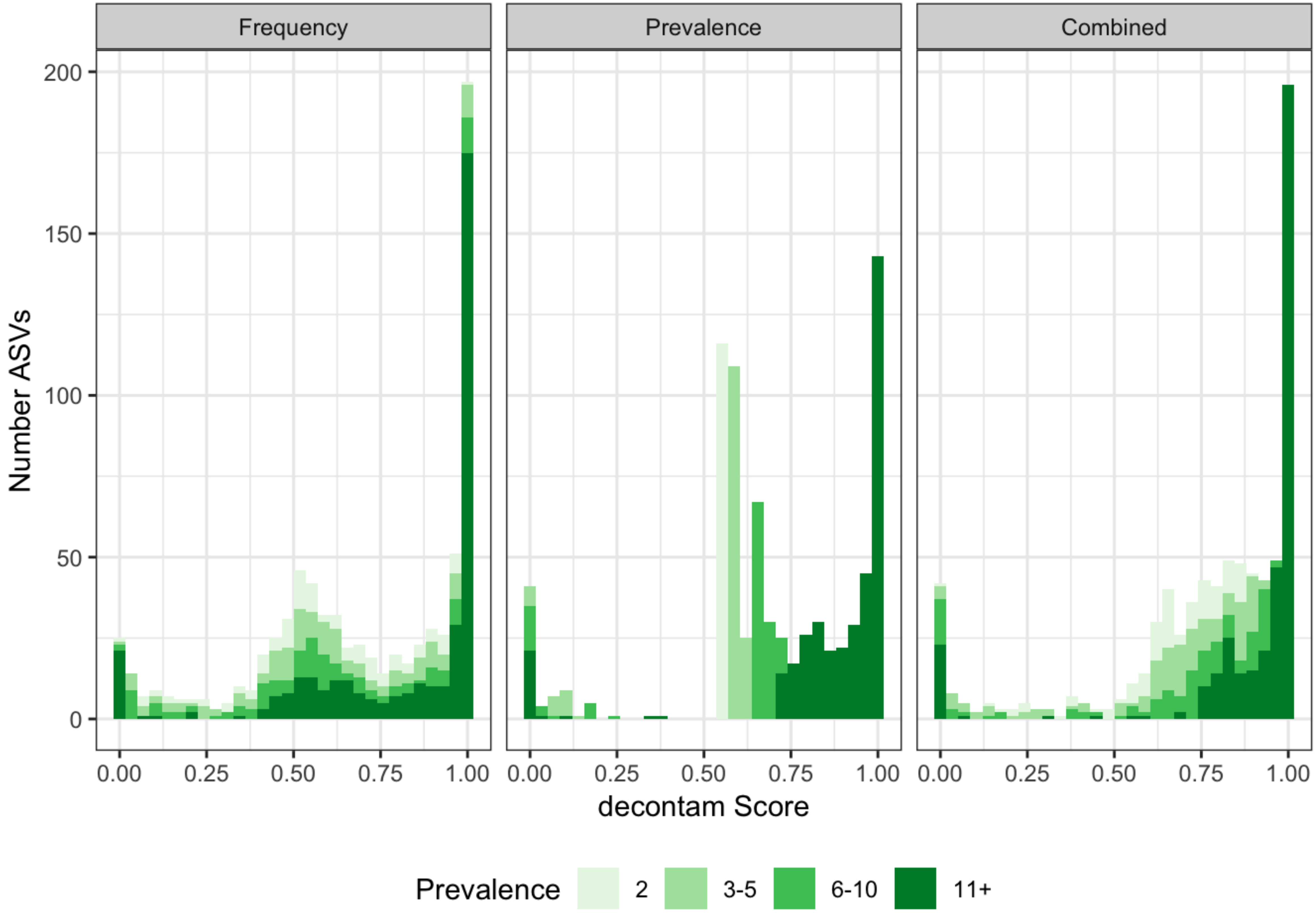
Output

```
> class(contam)

## [1] "data.frame"

> head(contam)

##                  freq prev       p.freq p.prev            p contaminant
## Seq1 0.323002694  549 1.000000e+00     NA 1.000000e+00       FALSE
## Seq2 0.098667396  538 1.000000e+00     NA 1.000000e+00       FALSE
## Seq3 0.003551358  160 1.135975e-18     NA 1.135975e-18        TRUE
## Seq4 0.067588419  519 9.999998e-01     NA 9.999998e-01       FALSE
## Seq5 0.045174743  354 1.000000e+00     NA 1.000000e+00       FALSE
## Seq6 0.040417101  538 1.000000e+00     NA 1.000000e+00       FALSE
```

Score: 0 to 1
(0: contaminant-like,
1: non-contaminant-like)

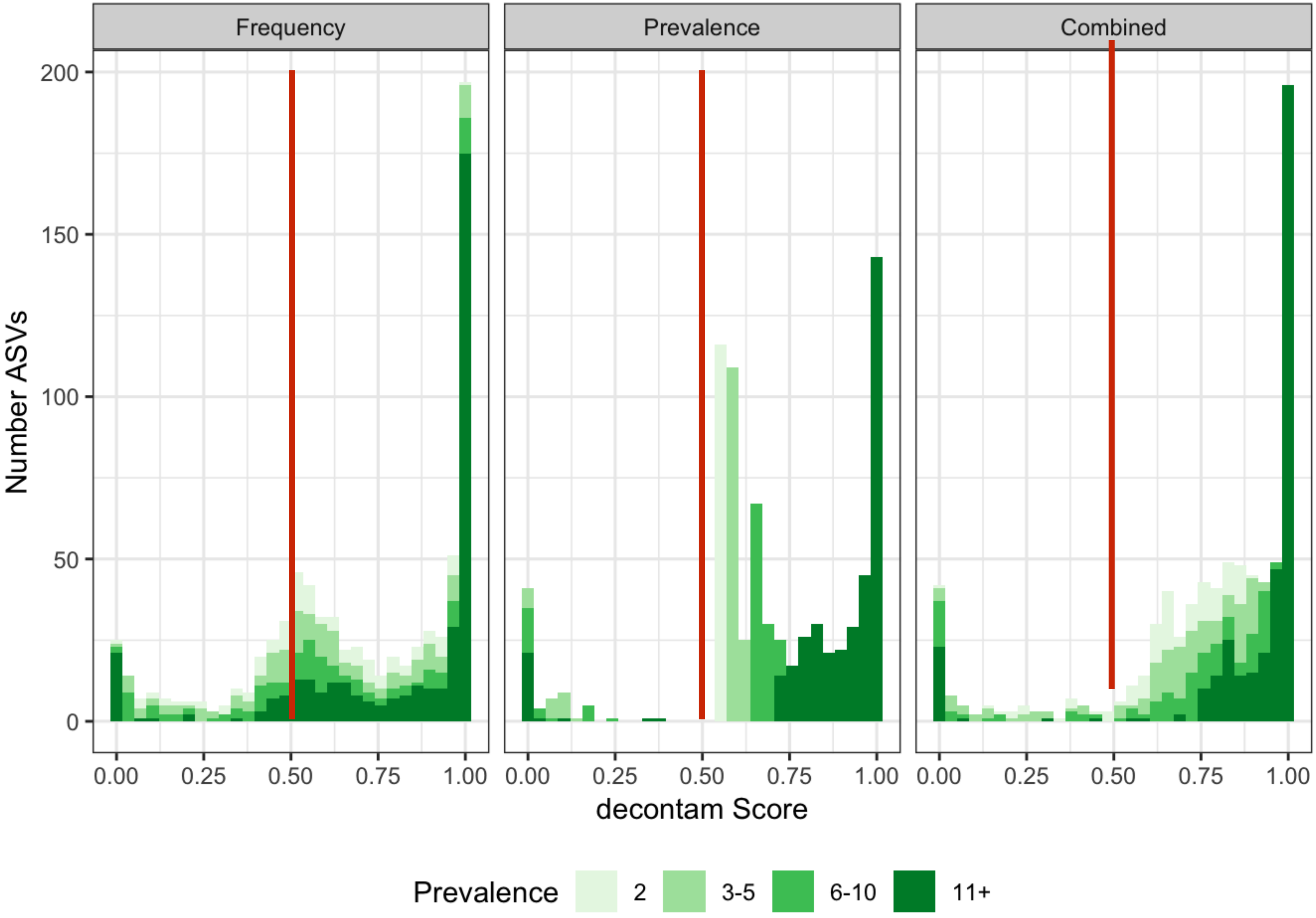
Classification: T/F
(score < threshold)

Beyond defaults



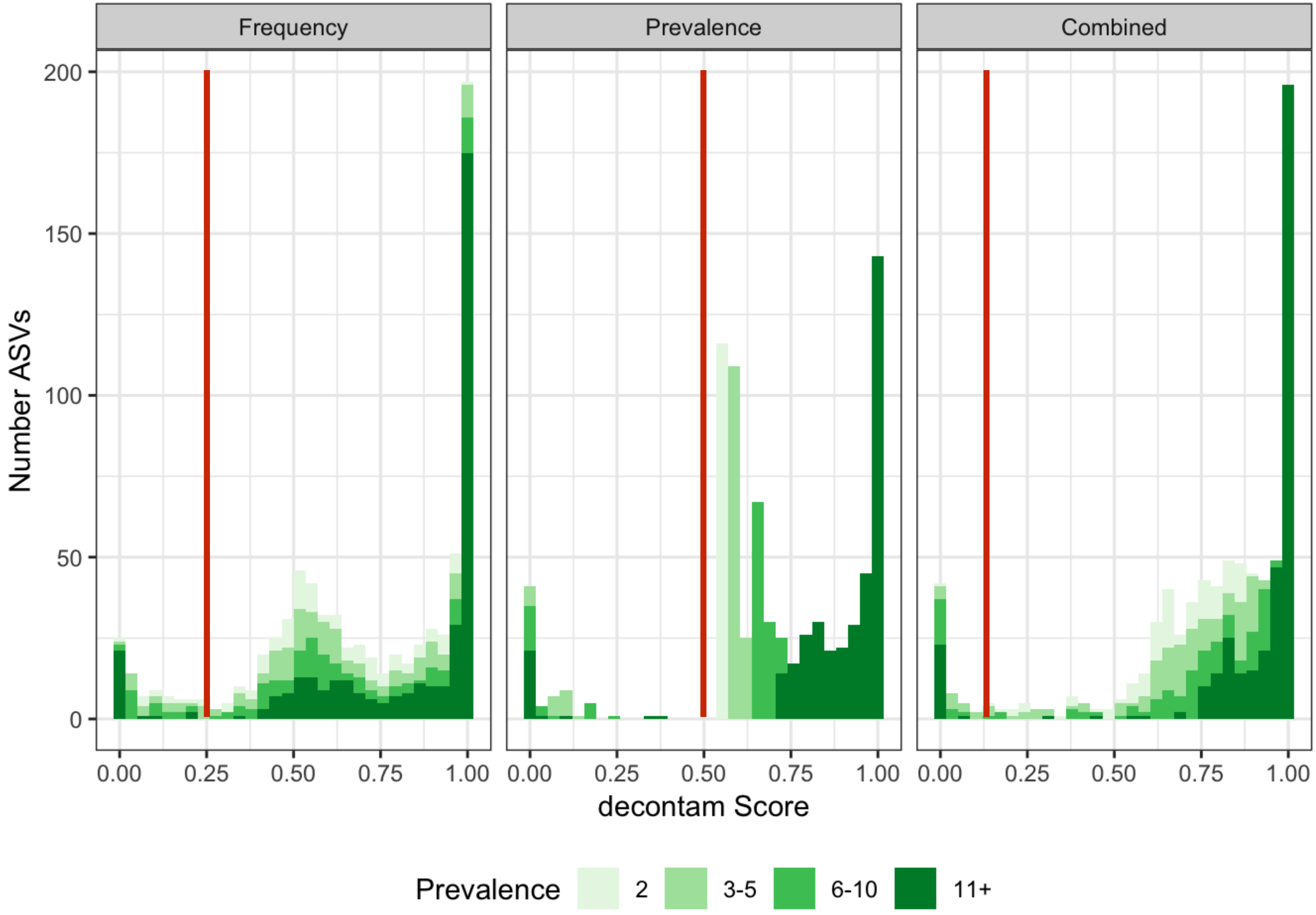
Beyond defaults

Default



Beyond defaults

Better



When to care?

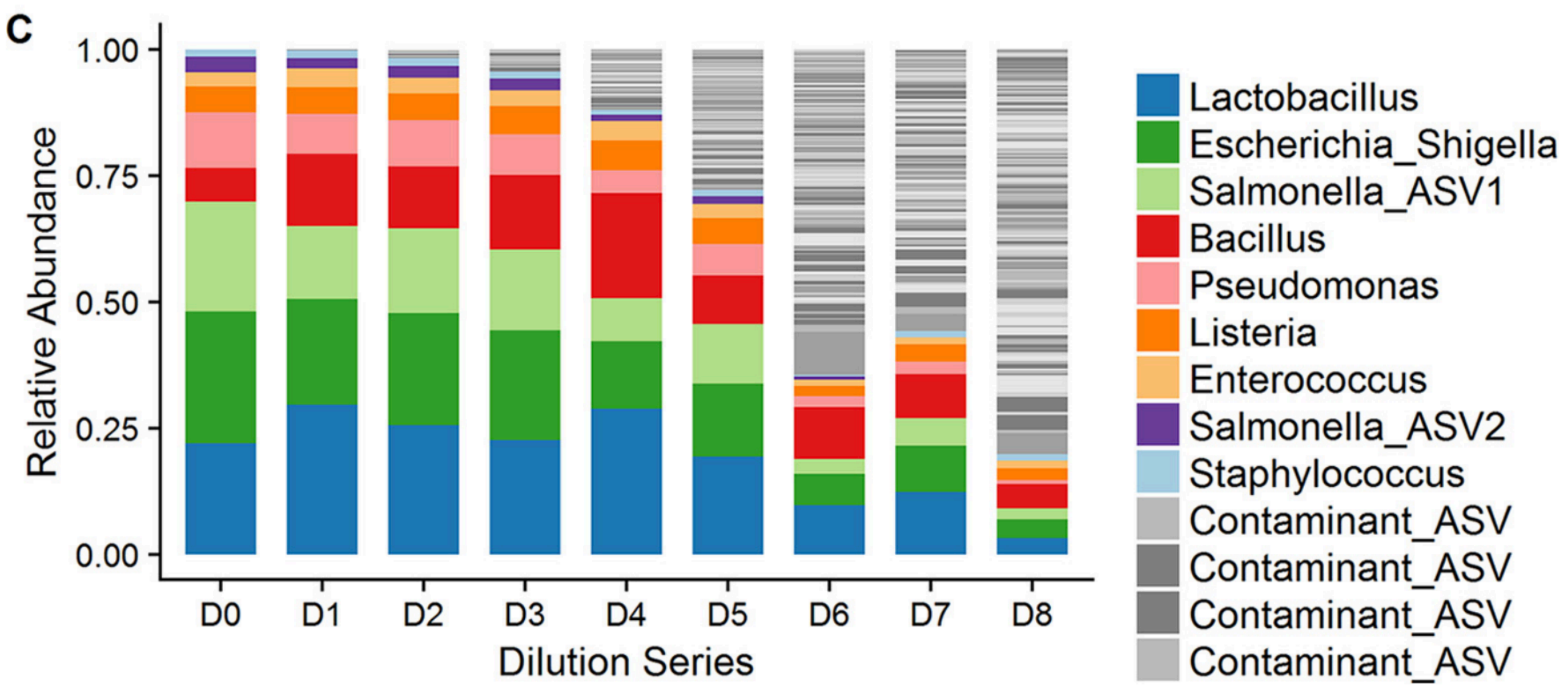
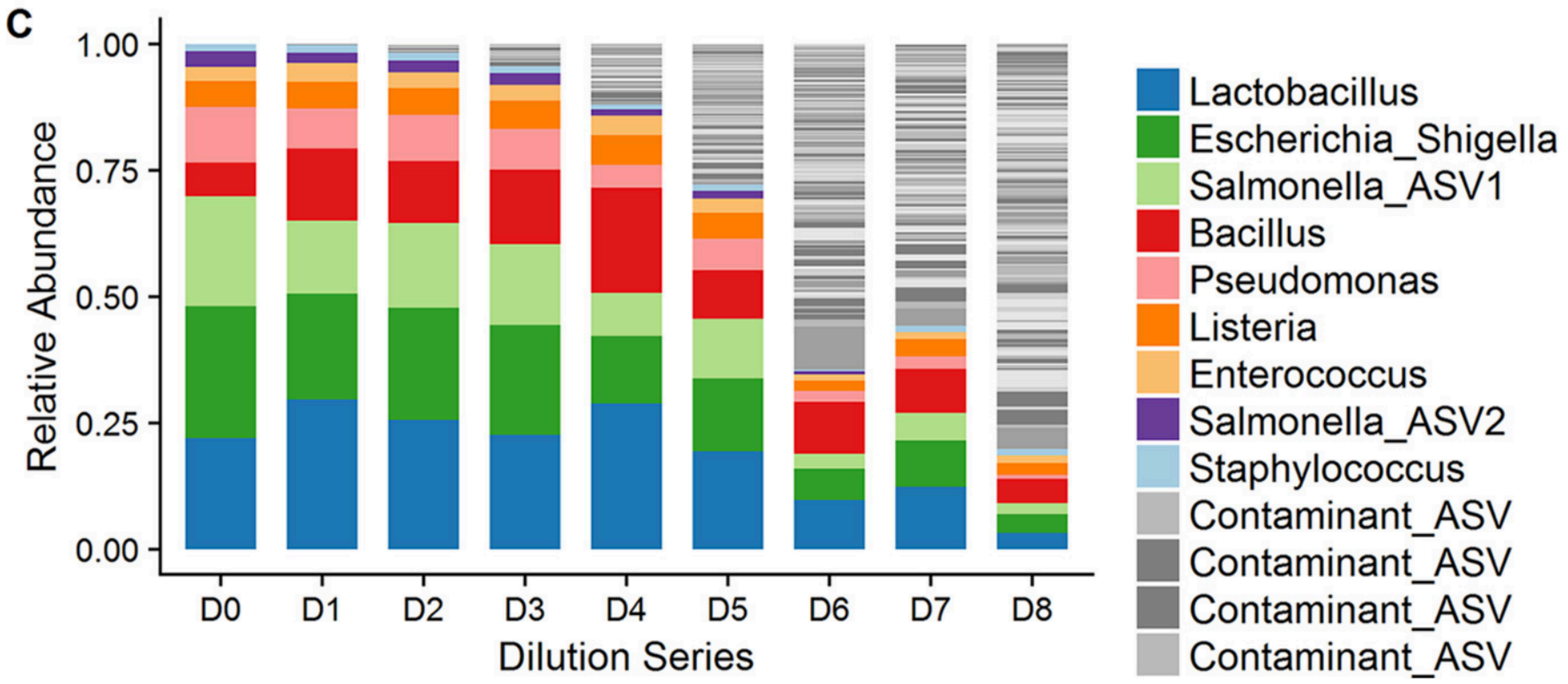


Figure: Karstens, et al. mSystems, 2018.

When to care?



Aerobiome, placenta, internal tissues, rare stuff matters, dry surfaces, parchment…

Figure: Karstens, et al. mSystems, 2018.

Recommendations…

- **There is no substitute for clean lab practices**

Recommendations…

- **There is no substitute for clean lab practices**
- **Sequence** *multiple* full-process negative controls!

Recommendations…

- **There is no substitute for clean lab practices**
- **Sequence** *multiple* full-process negative controls!
- Consider dilution series of a positive control

Recommendations…

- **There is no substitute for clean lab practices**
- **Sequence** *multiple* full-process negative controls!
- Consider dilution series of a positive control
- Record DNA concentrations

Recommendations…

- **There is no substitute for clean lab practices**
- **Sequence** *multiple* full-process negative controls!
- Consider dilution series of a positive control
- Record DNA concentrations
- *In silico* decontamination (at high resolution)

Recommendations…

- **There is no substitute for clean lab practices**
- **Sequence** *multiple* full-process negative controls!
- Consider dilution series of a positive control
- Record DNA concentrations
- *In silico* decontamination (at high resolution)
- Be skeptical of unexpected or implausible taxa

Recommendations…

- **There is no substitute for clean lab practices**
- **Sequence** *multiple* full-process negative controls!
- Consider dilution series of a positive control
- Record DNA concentrations
- *In silico* decontamination (at high resolution)
- Be skeptical of unexpected or implausible taxa
- Report taxa in negative controls

Recommendations…

Sequencing-based evidence of a microbiome in locations previously thought "sterile" is not conclusive on its own!

Recommendations…

Sequencing-based evidence of a microbiome in locations previously thought “sterile” is not conclusive on its own!

What additional evidence could make it convincing?

Growing options

Article | Open Access | Published: 10 November 2022

De novo identification of microbial contaminants in low microbial biomass microbiomes with Squeegee

Yunxi Liu, R. A. Leo Elworth, Michael D. Jochum, Kjersti M. Aagaard & Todd J. Treangen

Nature Communications **13**, Article number: 6799 (2022) | Cite this article

Environmental DNA

Open Access

Dedicated to the study and use of environmental DNA for basic and applied sciences

ORIGINAL ARTICLE | Open Access

microDecon: A highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies

Donald T. McKnight, Roger Huerlimann, Deborah S. Bower, Lin Schwarzkopf, Ross A. Alford, Kyall R. Zenger

First published: 16 May 2019 | https://doi.org/10.1002/edn3.11 | Citations: 68

Article | Published: 16 March 2023

Contamination source modeling with SCRuB improves cancer phenotype prediction from microbiome data

George I. Austin, Heekuk Park, Yoli Meydan, Dwayne Seeram, Tanya Sezin, Yue Clare Lou, Brian A. Firek, Michael J. Morowitz, Jillian F. Banfield, Angela M. Christiano, Itsik Pe'er, Anne-Catrin Uhlemann, Liat Shenhav & Tal Korem

Nature Biotechnology (2023) | Cite this article

Growing options

Article | Open Access | Published: 10 November 2022

De novo identification of microbial contaminants in low microbial biomass microbiomes with Squeegee

Yunxi Liu, R. A. Leo Elworth, Michael D. Jochum, Kjersti M. Aagaard & Todd J. Treangen

Nature Communications **13**, Article number: 6799 (2022) | Cite this article

Environmental DNA

Open Access

Dedicated to the study and use of environmental DNA for basic and applied sciences

ORIGINAL ARTICLE | Open Access

microDecon: A highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies

Donald T. McKnight, Roger Huerlimann, Deborah S. Bower, Lin Schwarzkopf, Ross A. Alford, Kyall R. Zenger

First published: 16 May 2019 | https://doi.org/10.1002/edn3.11 | Citations: 68

Article | Published: 16 March 2023

Contamination source modeling with SCRuB improves cancer phenotype prediction from microbiome data

George I. Austin, Heekuk Park, Yoli Meydan, Dwayne Seeram, Tanya Sezin, Yue Clare Lou, Brian A. Firek, Michael J. Morowitz, Jillian F. Banfield, Angela M. Christiano, Itsik Pe'er, Anne-Catrin Uhlemann, Liat Shenhav & Tal Korem

Nature Biotechnology (2023) | Cite this article

What assumptions are these methods making?
What additional data do these methods require?
When is is appropriate to use these methods?

Available now…

Methodology | Open Access

Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data

Nicole M. Davis, Diana M. Proctor, Susan P. Holmes, David A. Relman and Benjamin J. Callahan

Microbiome 2018 **6**:226

DECONTAM

- Open-source
- Well documented
- R package
- 16S or shotgun

Available now…

Methodology | Open Access

Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data

Nicole M. Davis, Diana M. Proctor, Susan P. Holmes, David A. Relman and Benjamin J. Callahan

Microbiome 2018 **6**:226

DECONTAM

- Open-source
- Well documented
- R package
- 16S or shotgun

*and recently, also via QIIME2!

Package resources

Manuscript
https://doi.org/10.1186/s40168-018-0605-2

Accompanying analyses in R
https://github.com/benjjneb/decontammanuscript

Vignette
https://benjjneb.github.io/decontam/vignettes/decontam_intro.html

Github (and use Issues for support)
https://github.com/benjjneb/decontam

Acknowledgements

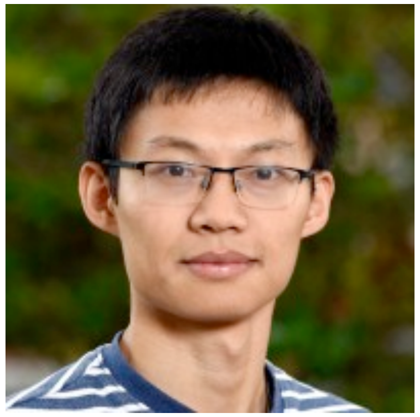
Susan Holmes

Nicole Davis

Diana Proctor

David Relman

And more recent developments…



Caizhi “David” Huang

Jorden Rabasco