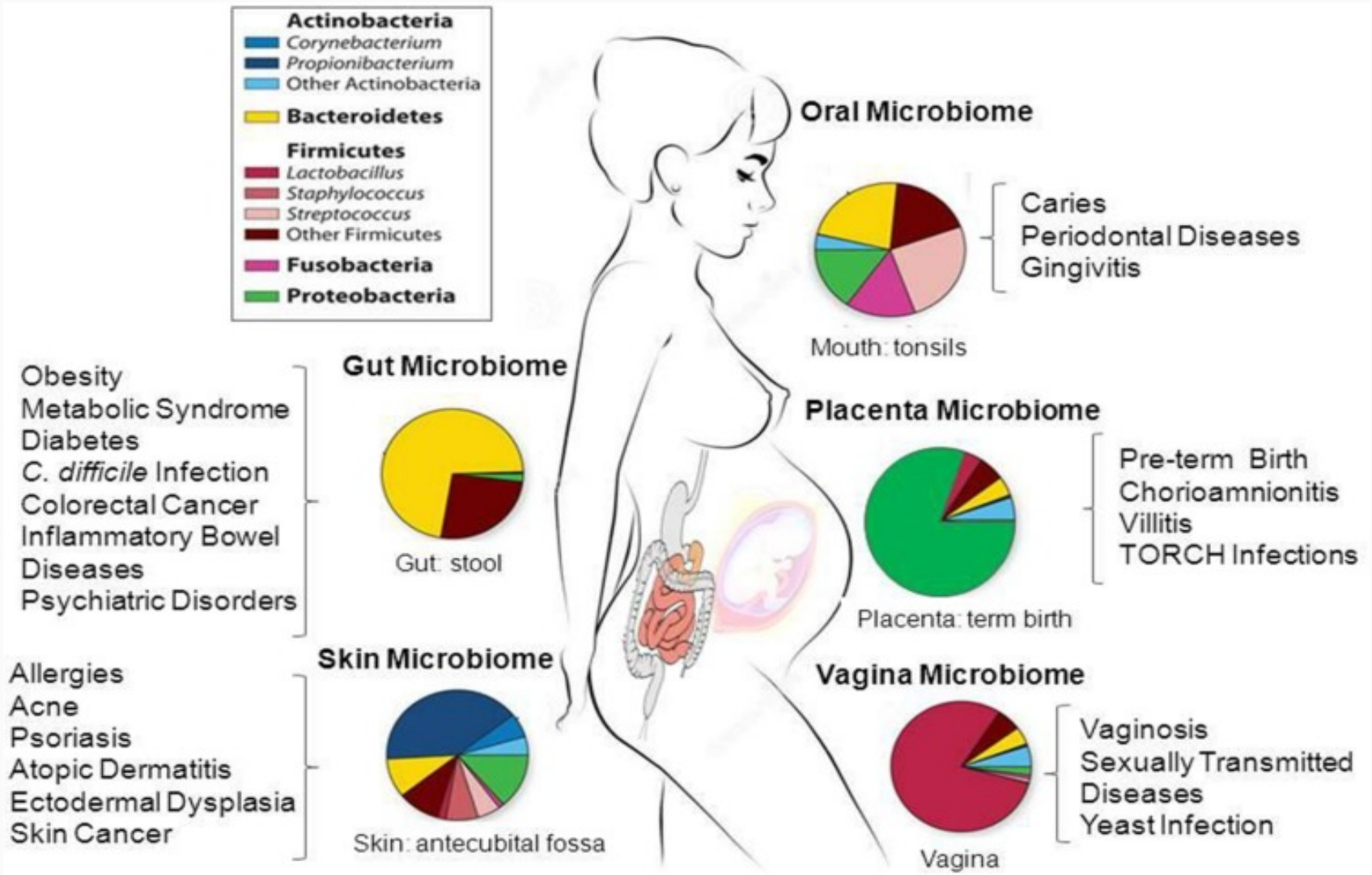
Microbiome data

Current challenges and opportunities

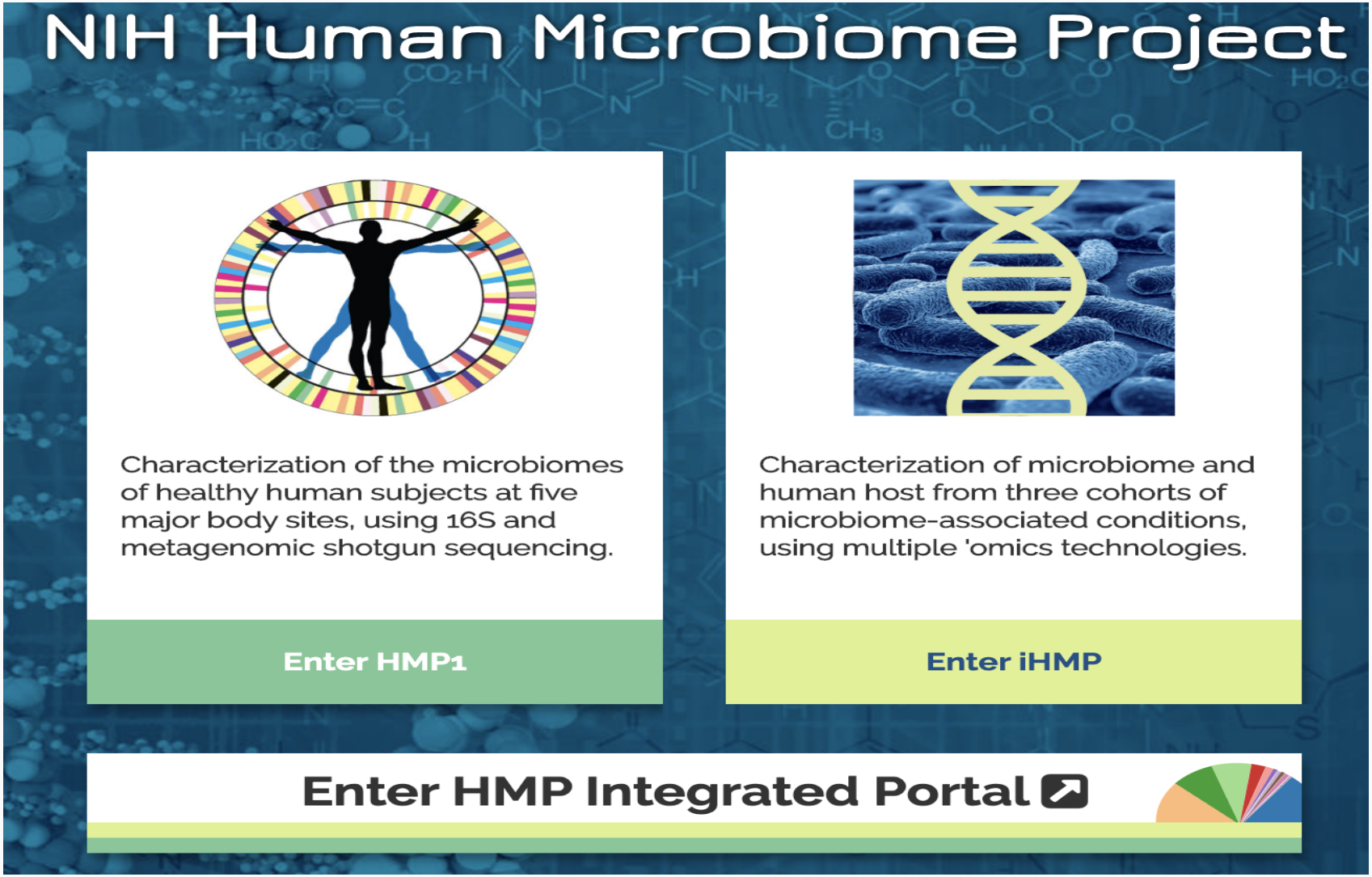
Ekaterina Smirnova, Virginia Commonwealth University
ekaterina.smirnova@vcuhealth.org
https://github.com/katiasmirn

Human microbiome studies



Source: Belizario and Napolitano (2015)

Human microbiome project (HMP)



HMP Bioconductor data packages

HMP16SData

platforms all | rank 84 / 384 | posts 0 | build ok
updated < 3 months | dependencies 116

DOI: 10.18129/B9.bioc.HMP16SData

16S rRNA Sequencing Data from the Human Microbiome Project

HMP Bioconductor data packages

HMP16SData

platforms all | rank 84 / 384 | posts 0 | build ok
updated < 3 months | dependencies 116

DOI: 10.18129/B9.bioc.HMP16SData

16S rRNA Sequencing Data from the Human Microbiome Project

HMP2Data

platforms all | rank 377 / 384 | posts 0 | build ok
updated before release | dependencies 145

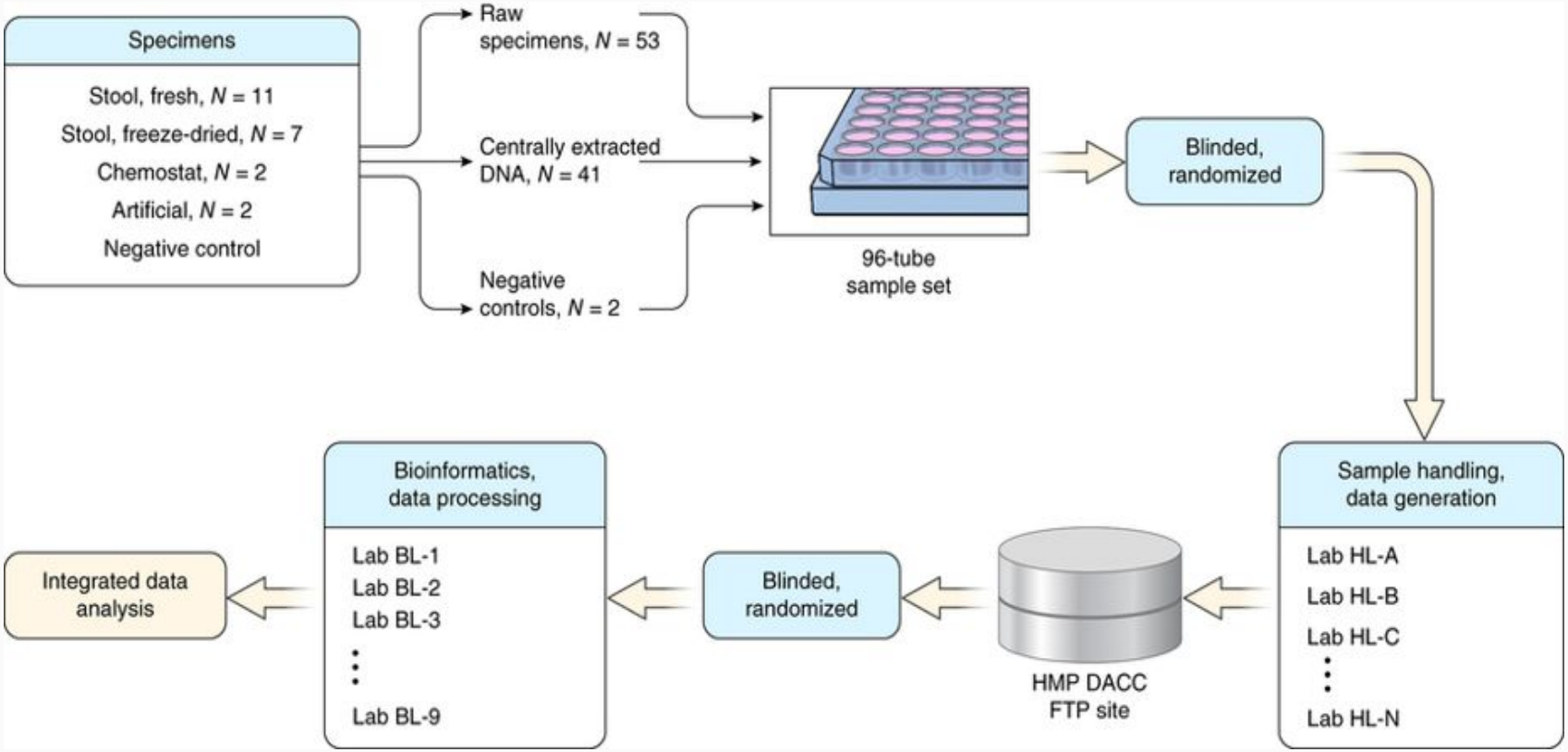
DOI: 10.18129/B9.bioc.HMP2Data

16s rRNA sequencing data from the Human Microbiome Project 2

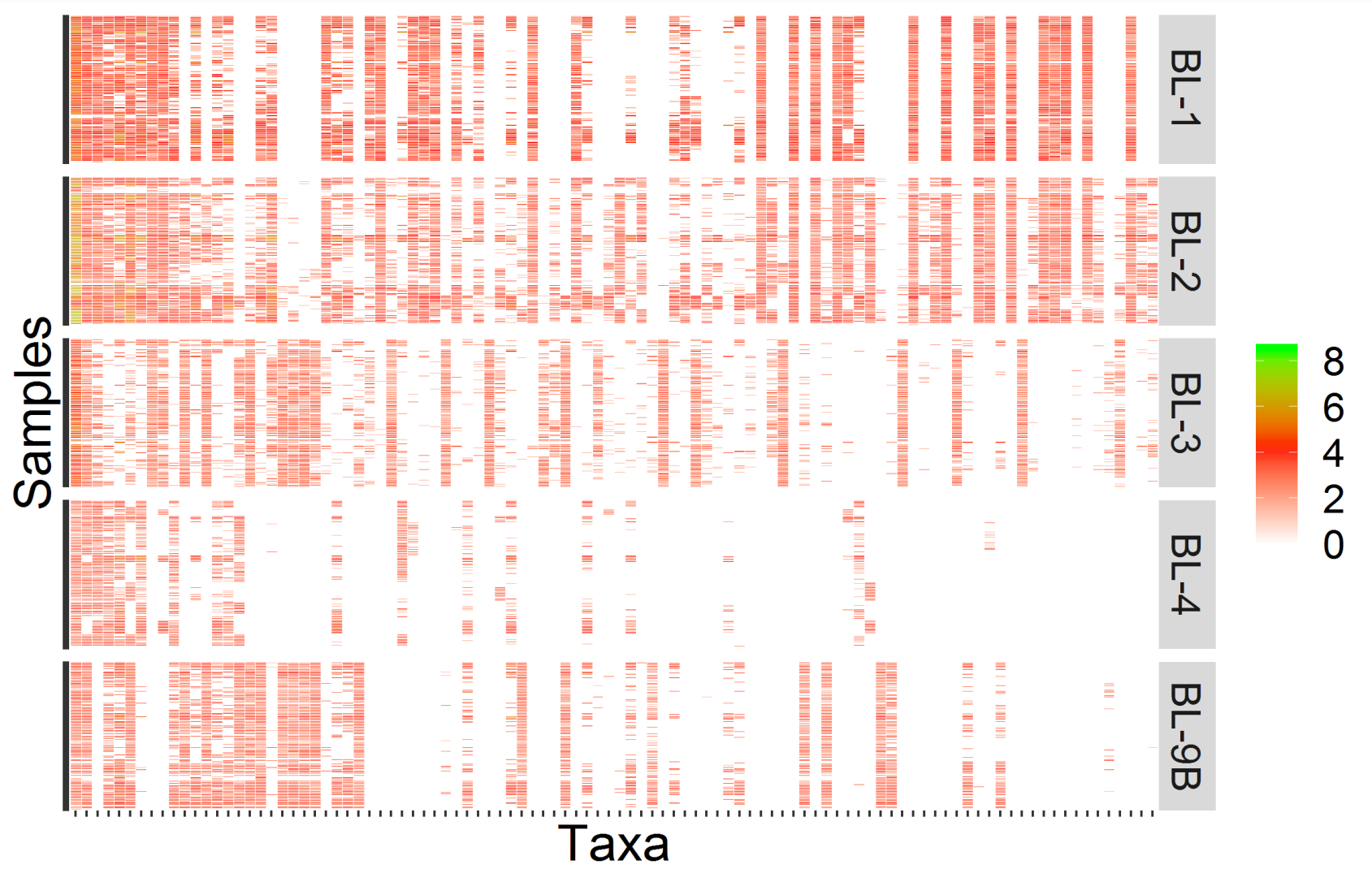
Microbiome studies: major study steps

- Sample collection and study design: medical, ecological (ocean microbiome), etc
- Wet lab: sample handling and processing, DNA extraction protocols
- Dry lab: mapping sequence reads to bacterial organisms, bioinformatics protocols
- Data analysis: statistical methodology for taxa table analysis

Microbiome qualily control study (MBQC)



High lab-to-lab variability



Taxa table quality control

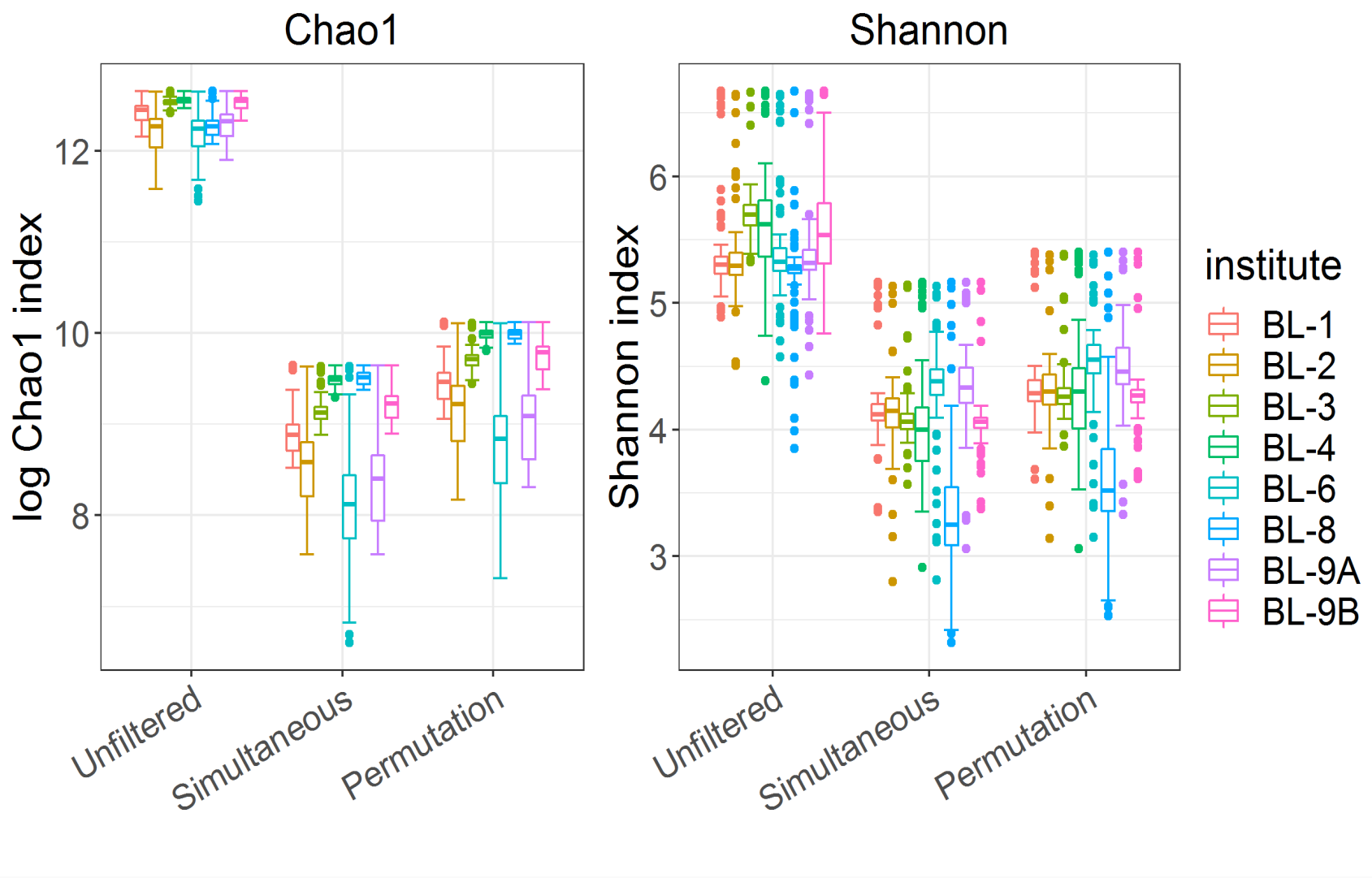
Contaminant taxa removal (e.g. bacteria on lab surfaces, sequencing plates, not expected in sample environment)

- More contamination expected in samples with low DNA concentration
- Taxa observed in negative controls
- Statistical method: Decontam (Davis et al., 2018)

Filtering taxa observed in small number of samples

- Rare (e.g. observed in $< 0.1\%$ of minimal reads) taxa
- Statistical method: PERFect (Smirnova et al., 2018)

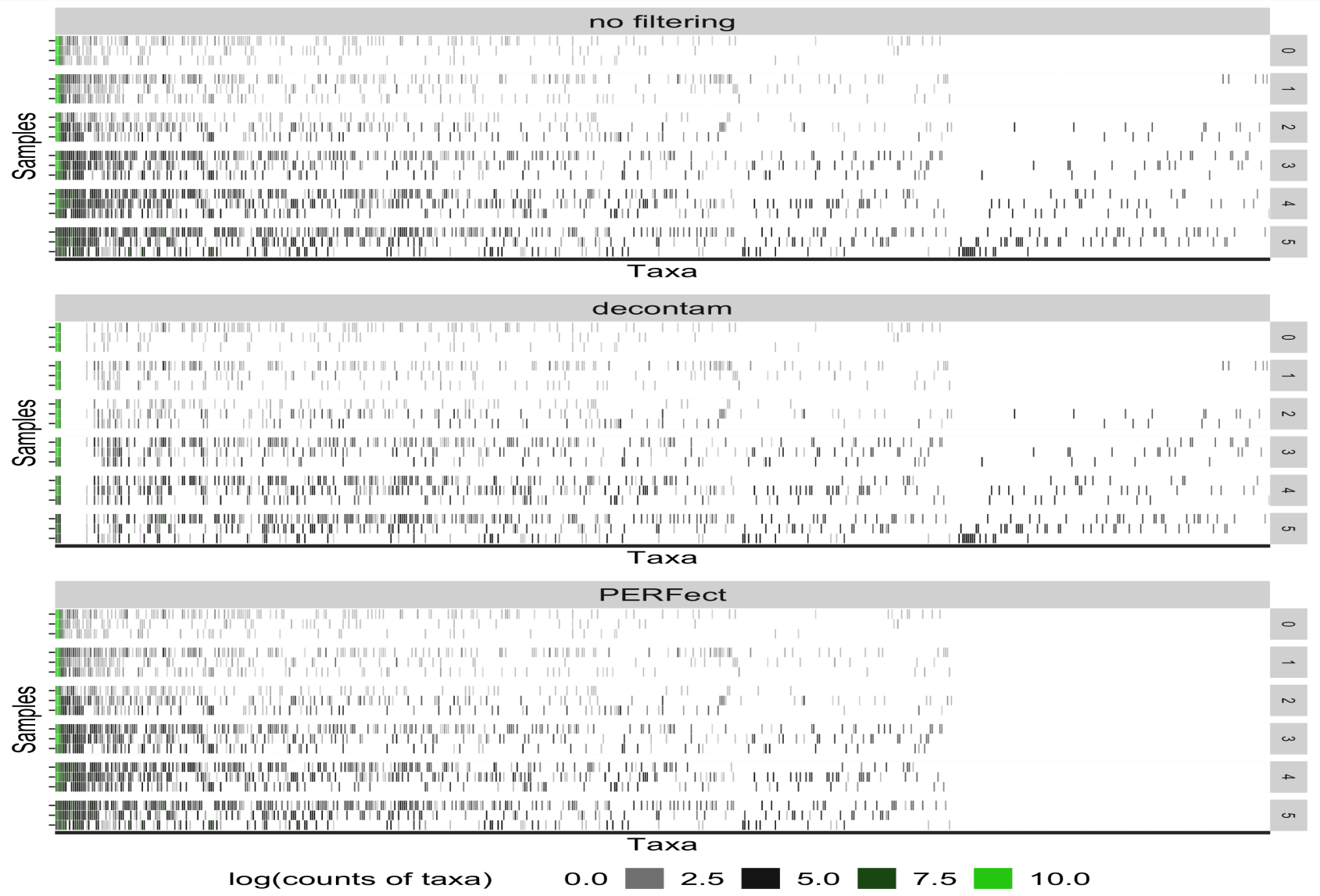
Filtering effects: alpha diversity



Filtering effects: alleviating technical variability

	Unfiltered		Simultaneous		Permutation	
Comparison	Difference	P-values	Difference	P-values	Difference	P-values
BL-1 - BL-2	0.00	0.4990	0.20	0.4214	-0.06	0.4778
BL-1 - BL-3	-10.75	< 0.0001	2.52	0.0074	1.27	0.1307
BL-2 - BL-3	-10.83	< 0.0001	2.35	0.0115	1.33	0.1283
BL-1 - BL-4	-7.22	< 0.0001	3.90	0.0001	0.29	0.4173
BL-2 - BL-4	-7.28	< 0.0001	3.73	0.0001	0.35	0.4088

Filtering versus contamination



reagent and laboratory contamination study, Salter et al, 2014

Methods - Assumptions

decontam frequency

* Contaminant C present in uniform concentration across samples
* Sample S present in varying concentration across samples
* In the limit of $S >> C$

$$f_C = \frac{C}{C+S} \sim \frac{1}{T}$$

$$f_S = \frac{S}{C+S} \sim 1$$

f: frequency; T: Total DNA concentration

PERFect

* Non-contaminant DNAs exist across samples
* DNAs with low abundance are more likely to be noises
* Over half DNAs are noises
* If a taxon is unimportant, it contributes little to total covaraince of the OTU table

Methods

decontam frequency

- Models
  - Contaminant model
    $\log(f_{seq}) \sim -1 \times \log(T)$
  - Non-contaminant model
    $\log(f_{seq}) \sim 1$
  - $R = \frac{SSE_C}{SSE_S}$
    $R \sim F(n-1,\ n-1)$
    n: samples that contain the taxon
- Limitation: Does not work in samples with fewer DNA fragments

PERFect

- Test
  - $FL(J) = 1 - \frac{\text{covariance after removing } J \text{ taxa}}{\text{total covariance}}$
  - Hypothesis:
    - $H_0 : dF_{j+1} = 0$
    - $H_1 : dF_{j+1} > 0$
  - Notation
    - FL: filting loss
    - dF: difference of filtering loss
- Limitation: Not for contaminant taxa with high abundance

Filtering versus contamination

Table 1: Noise taxa identified

	decontam	PERFect
Seq4	Yes	No
Seq5	Yes	No
Seq6	Yes	No
Seq7	Yes	No
Seq8	Yes	No
Seq9	Yes	No
Seq42	NA	Yes
Seq49	NA	Yes
Seq54	NA	Yes
Seq77	NA	Yes
Seq85	NA	Yes
Seq104	NA	Yes

Table 2: Number of noises identified

	decontam	PERFect
No	353	469
Yes	61	166
NA	221	0

Extensions

- Effect of filtering on taxa selection across two groups
- Contaminant removal methods using positive and negative controls data
- Wet Lab-to-lab variability by running MBQC data on same pipeline
- Understanding sources of variability and using this knowledge
- Inform sample processing and bioinformatics protocols
- Alleviate technical variability across labs

Research group

- Bi-weekly “virtual” microbiome research group over zoom
- Thursdays, 4 pm US ET
- Welcome interested researchers to:
  - give a guest lecture
  - join the group and collaborate
  - discuss relevant paper (literature club)