Extracting the

signal from the ***noise***

in high-throughput amplicon data

Stanford Microbiome Symposium - May 2016

The amplicon inference problem

Infer the sample types and abundances {(s, a)} from error-ful amplicon reads {r}.

Amplicon inference problems

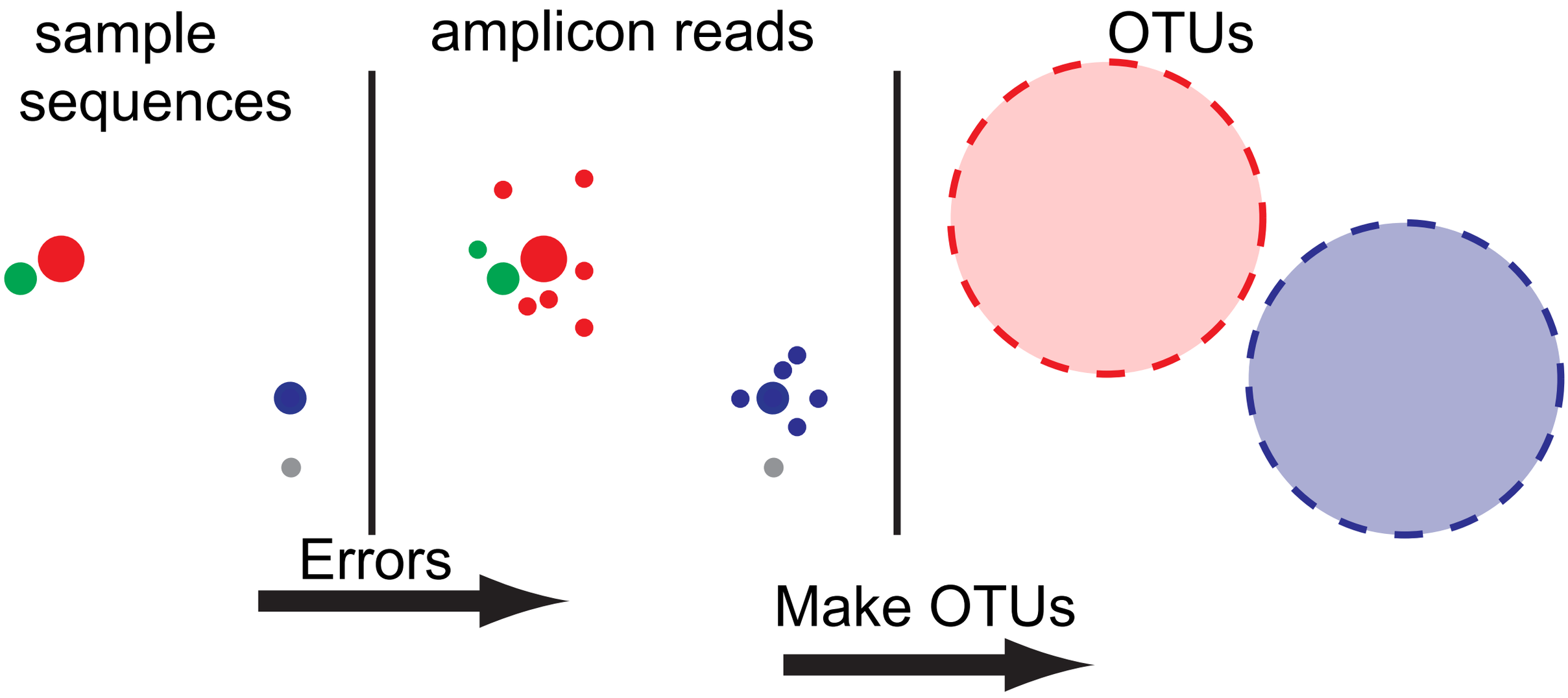
1. **Low resolution**
   - **97%, genus at best**

2. **High false positive rate**
   - **#(OTUs) >> richness**

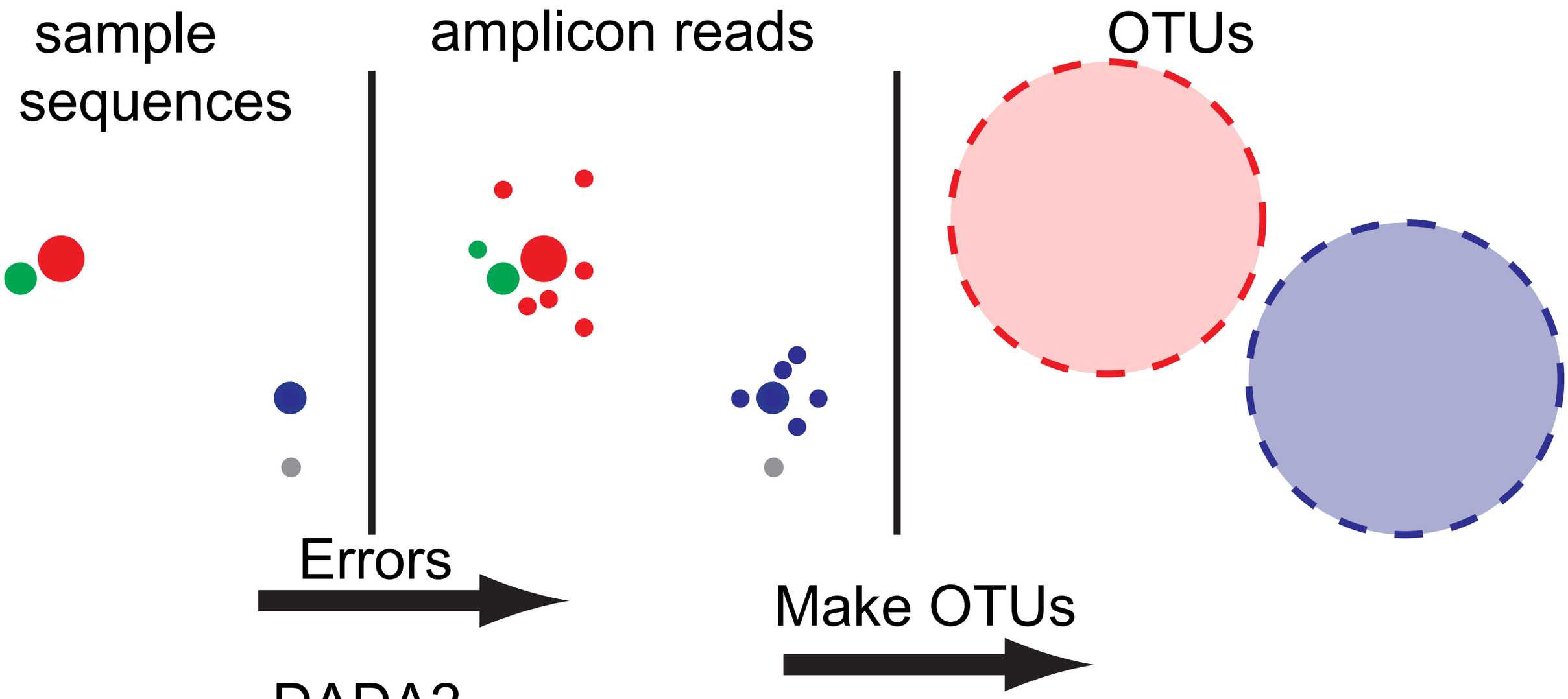
3. **Big data scaling**
   - **time scales super-linearly**

4. **Cross-study comparison**
   - **must reprocess all data together**

DADA2: High resolution



DADA2: High resolution



Callahan, et al. Nature Methods, 2016.

DADA2: Error model

```
s: ATTAACGAGATTATAACCAGAGTACGAATA...
     |                    |
r: ATCAACGAGATTATAACAAGAGTACGAATA...
```

$$p(r|s) = \prod_{i=1}^{L} p(r(i)|s(i), q_r(i), Z)$$

DADA2: Error model

```
s: ATTAACGAGATTATAACCAGAGTACGAATA...
     |                    |
r: ATCAACGAGATTATAACAAGAGTACGAATA...
```

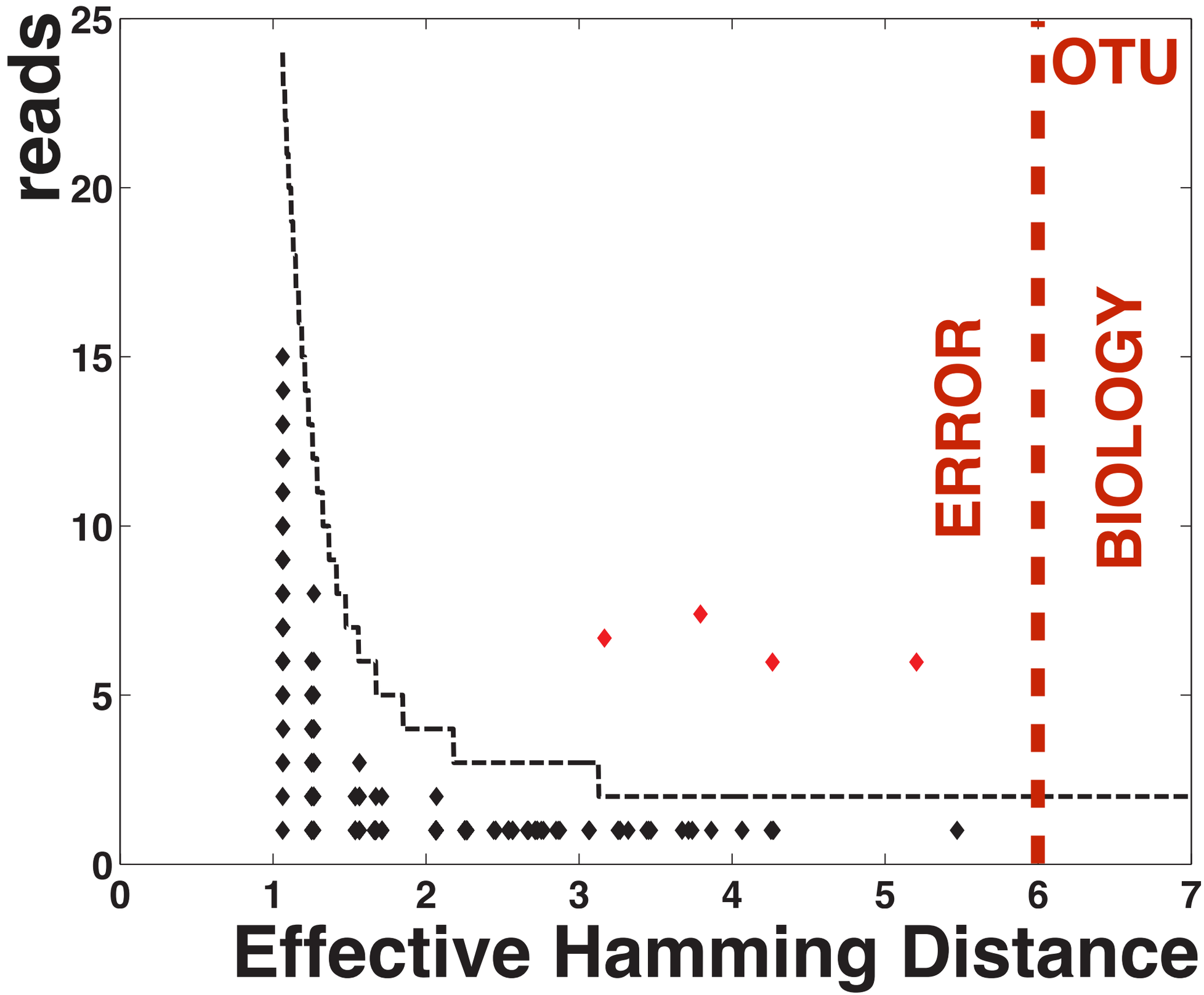
$$p(r|s) = \prod_{i=1}^{L} p(r(i)|s(i), q_r(i), Z)$$

Error rates depend on....

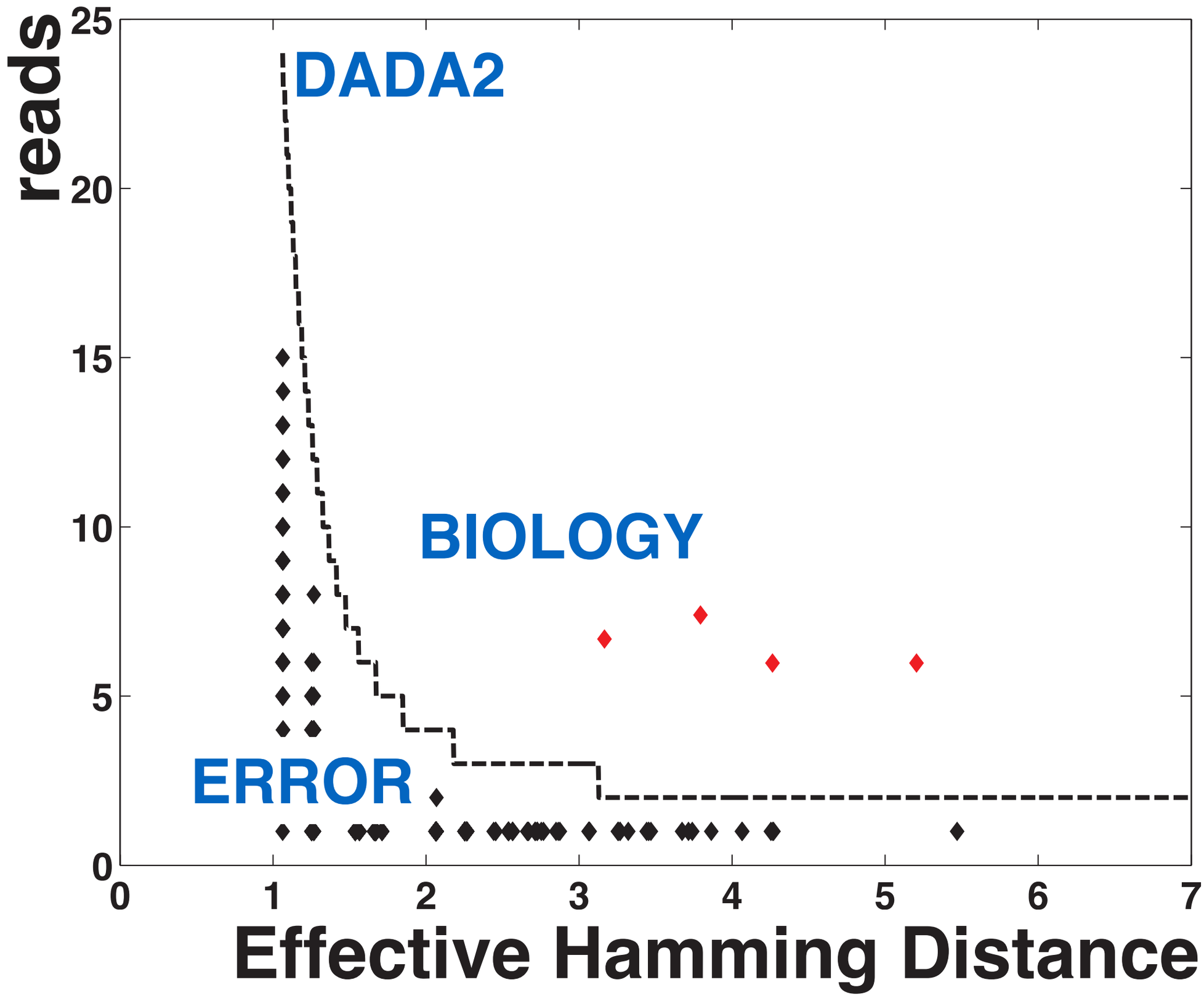
- Substitution (eg. A->C)
- Quality score (eg. Q=30)
- Batch effect (eg. run)

Using more data!

Signal from Noise: OTUs



Signal from Noise: DADA2

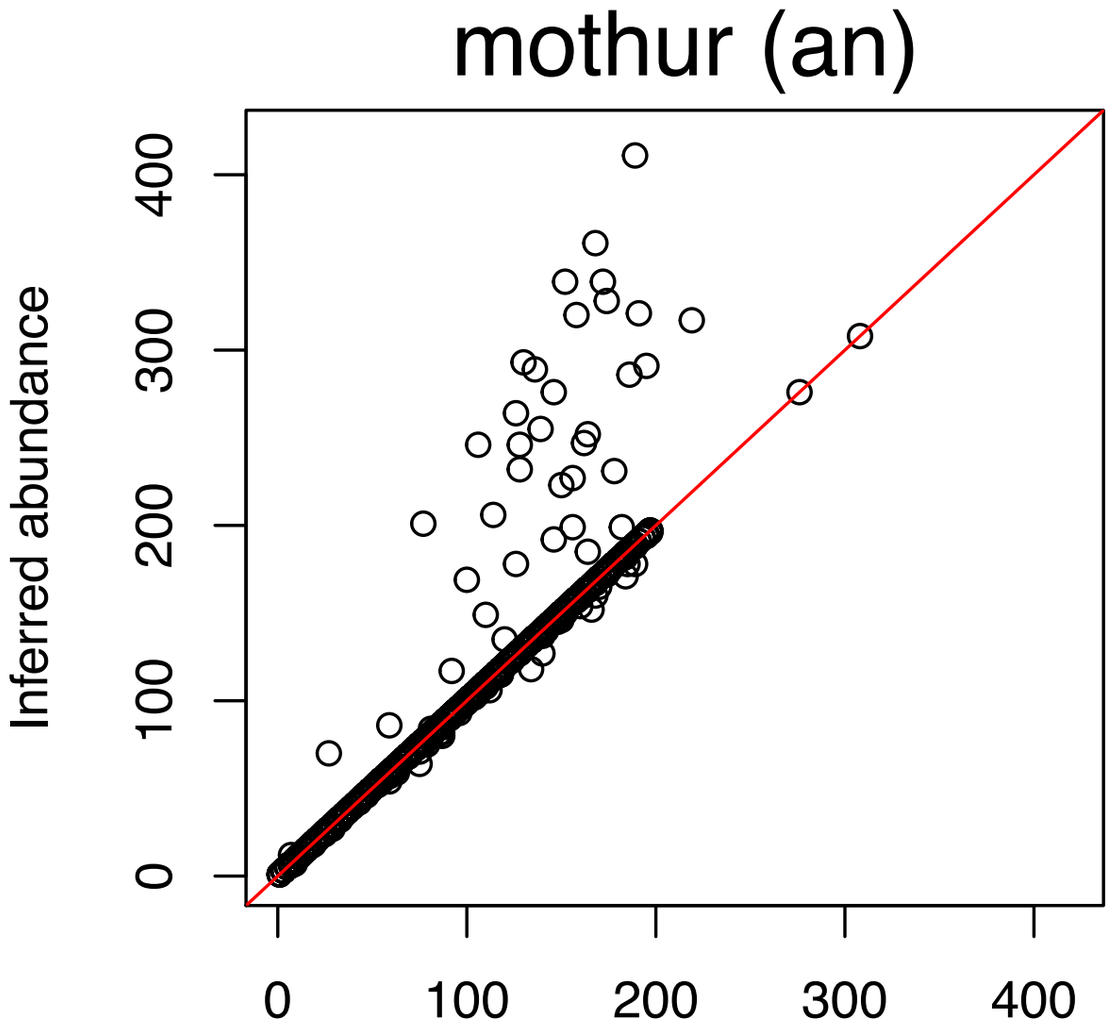


Accuracy

and

Resolution

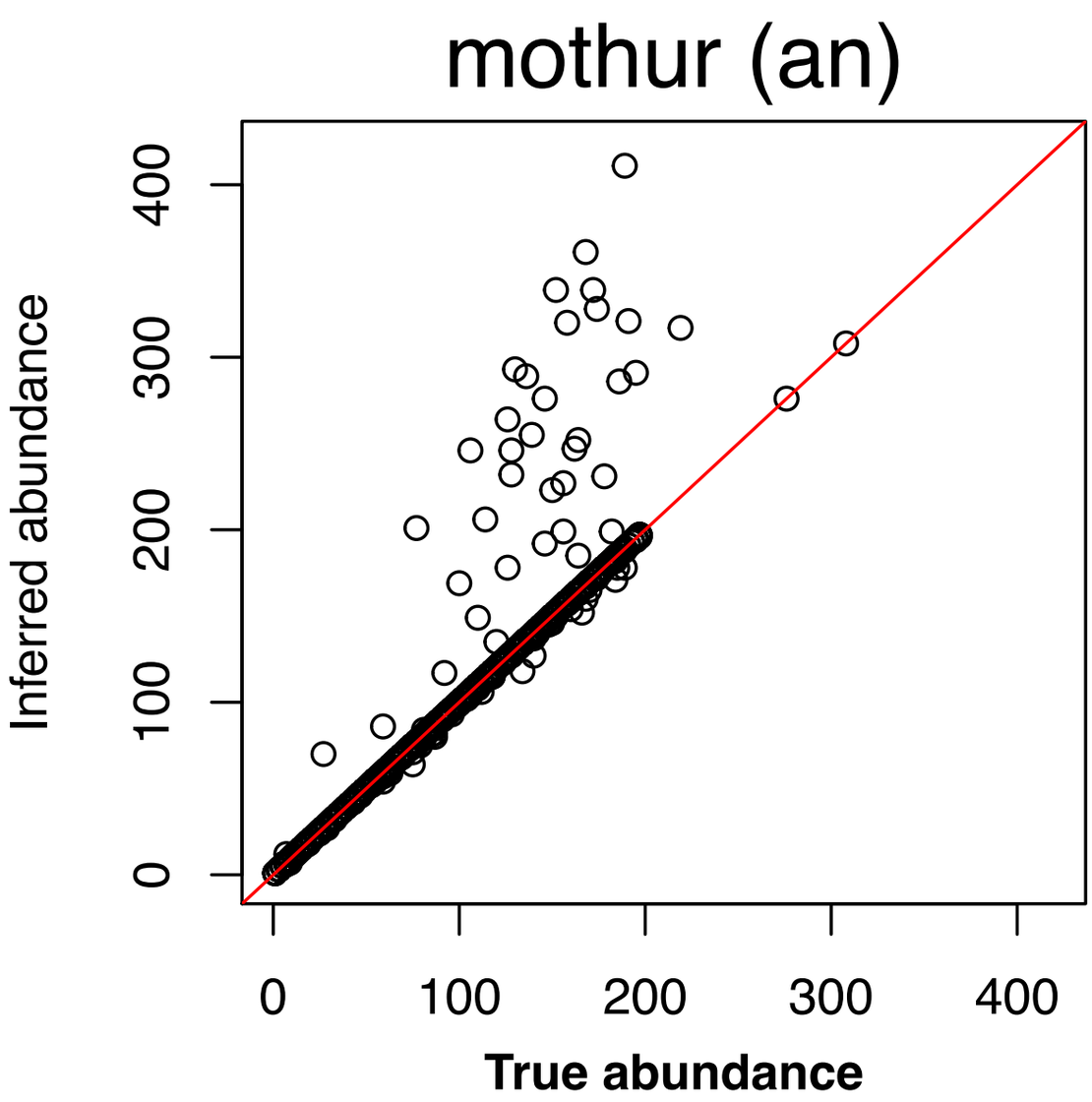
Accuracy: Simulated data



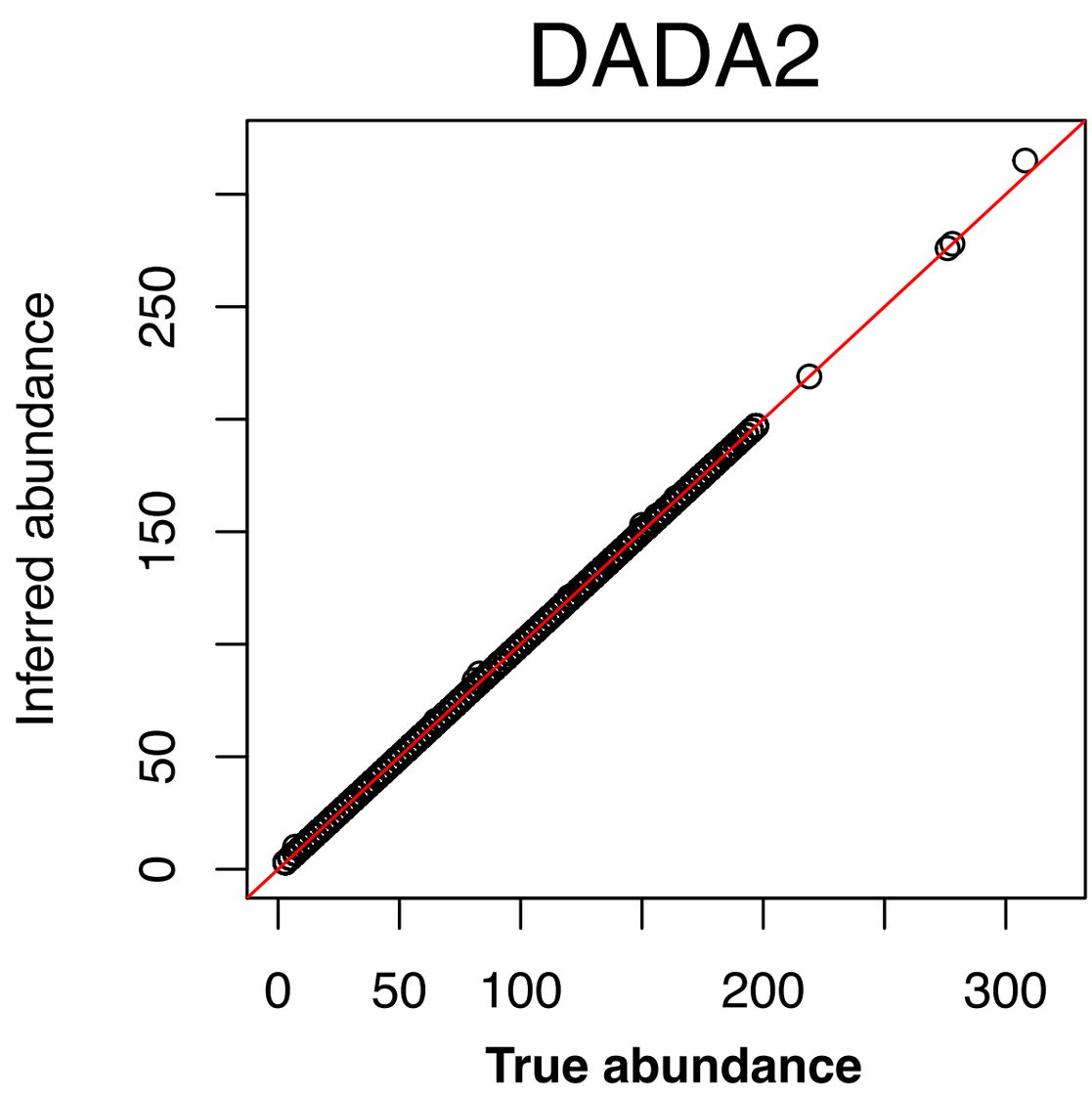
TP: 978
FP: 272
FN: 77
cor: 0.935

Data: Kopylova, et al. mSystems, 2016.

Accuracy: Simulated data



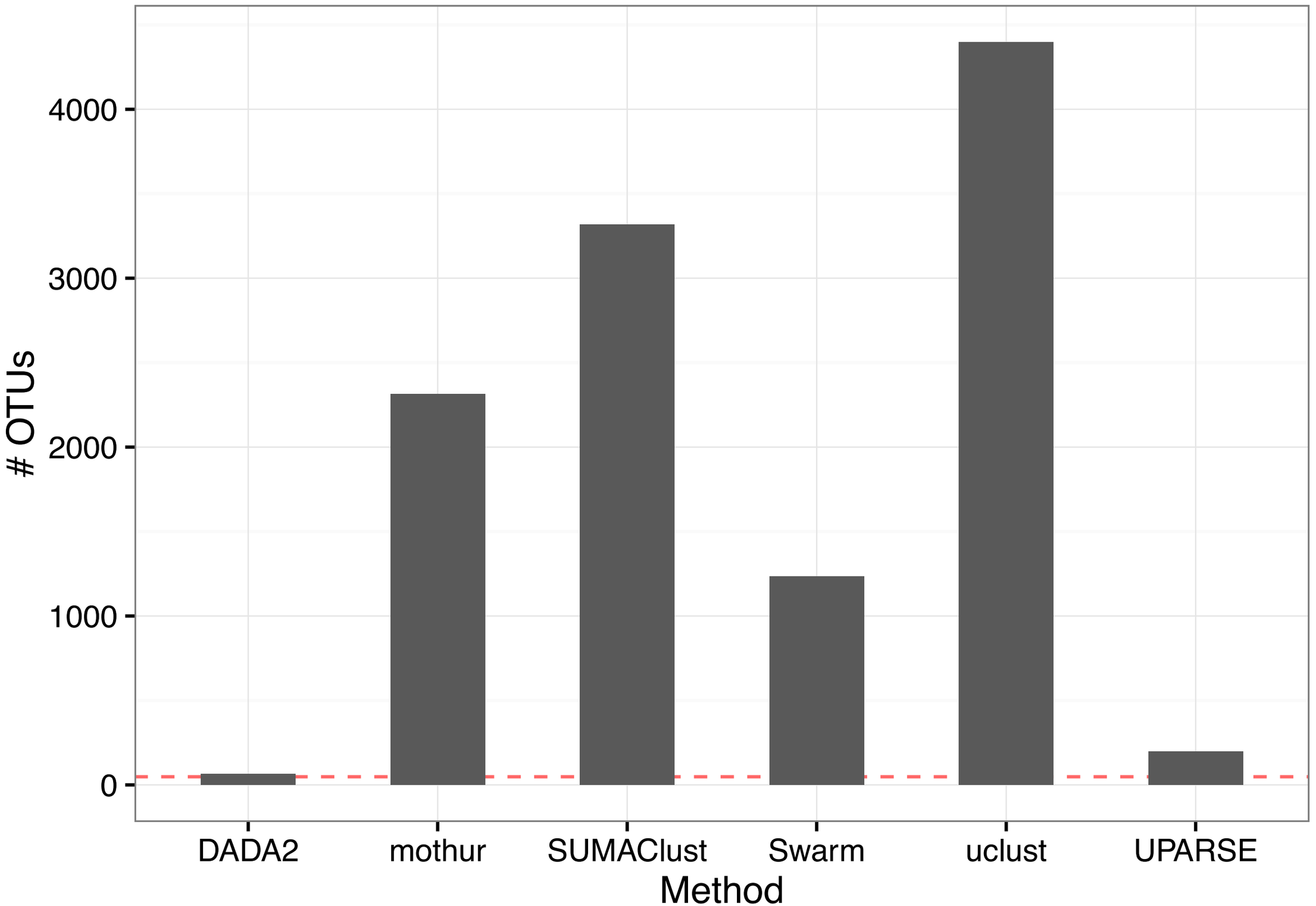
TP: 978
FP: 272
FN: 77
cor: 0.935



TP: 1042
FP: 0
FN: 13
cor: 0.999

Data: Kopylova, et al. mSystems, 2016.

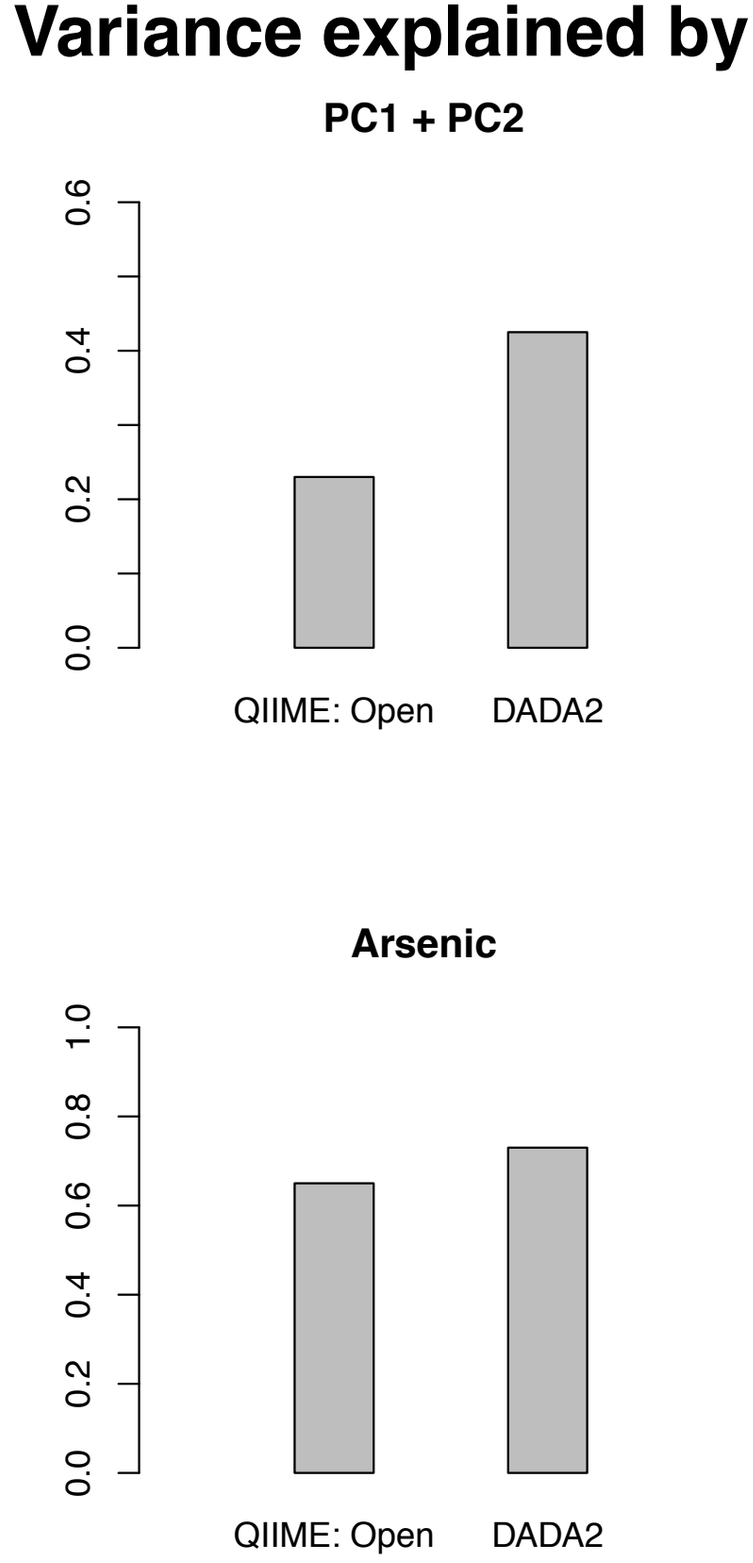
Accuracy: Mock community



Credit: Kopylova, et al. mSystems, 2016.

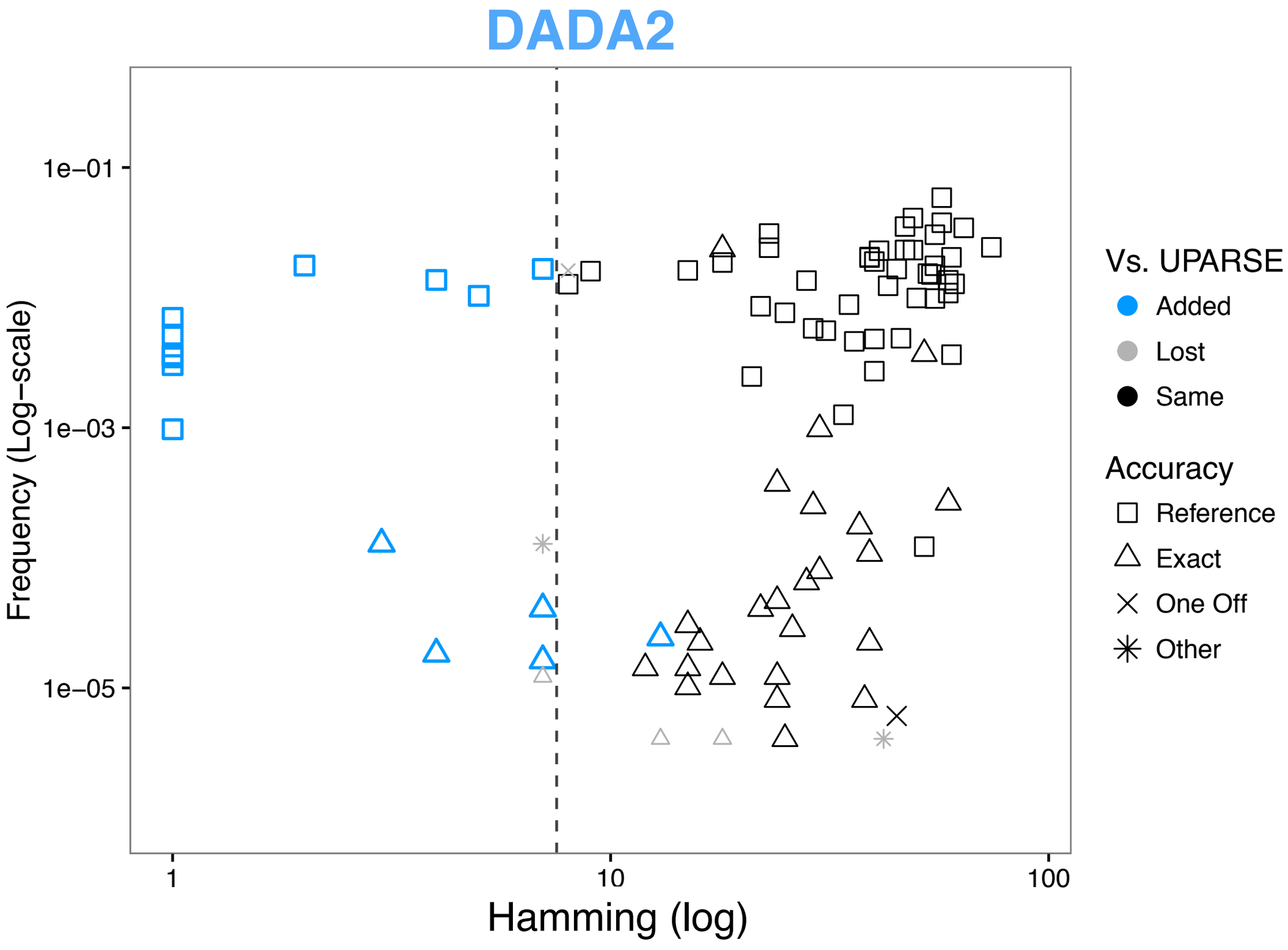
Accuracy: Arsenic treatment

Unpublished ordination removed.



Credit: Dylan Dahan & Gabriel G. Perron

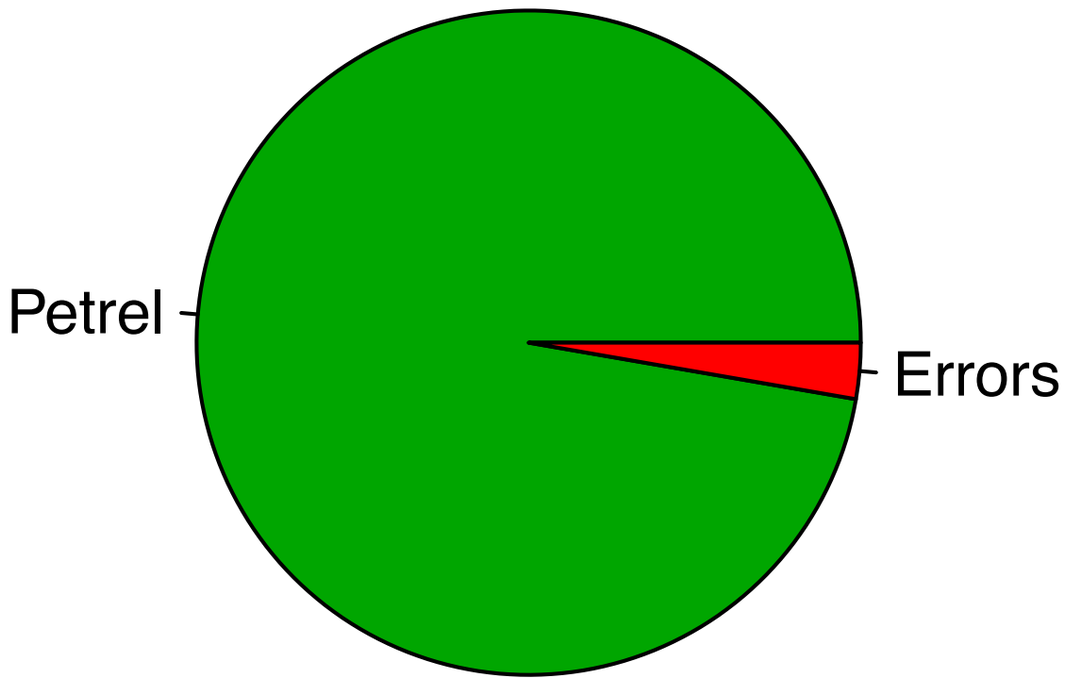
Resolution: Mock Community



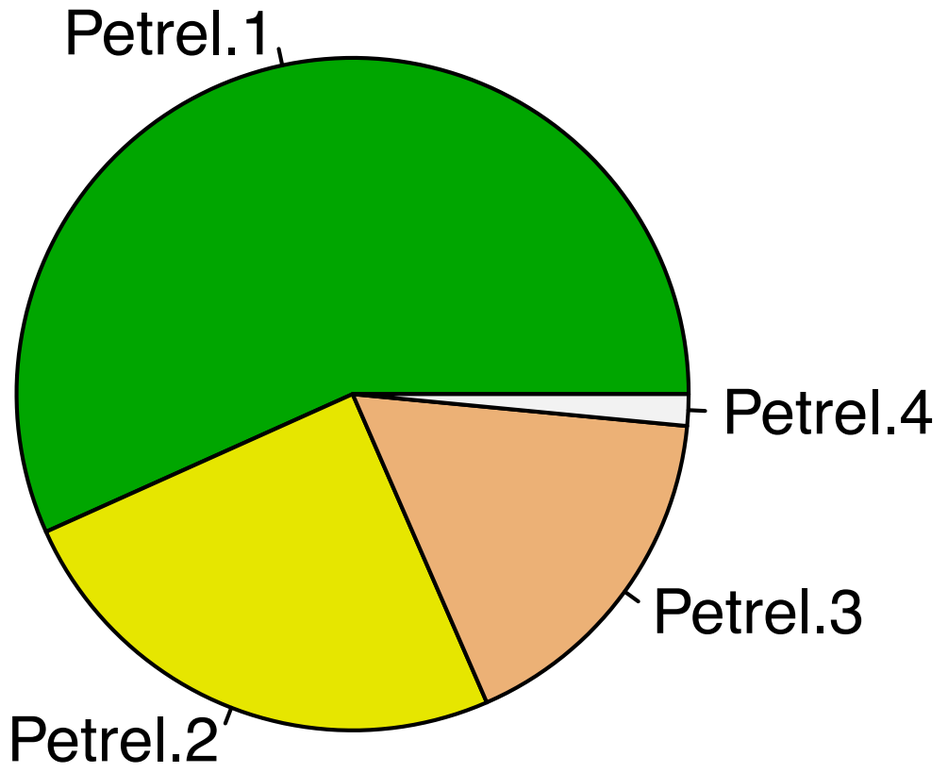
Callahan et al., Nature Methods, 2016.

Resolution: Petrel aDNA

QIIME: De novo

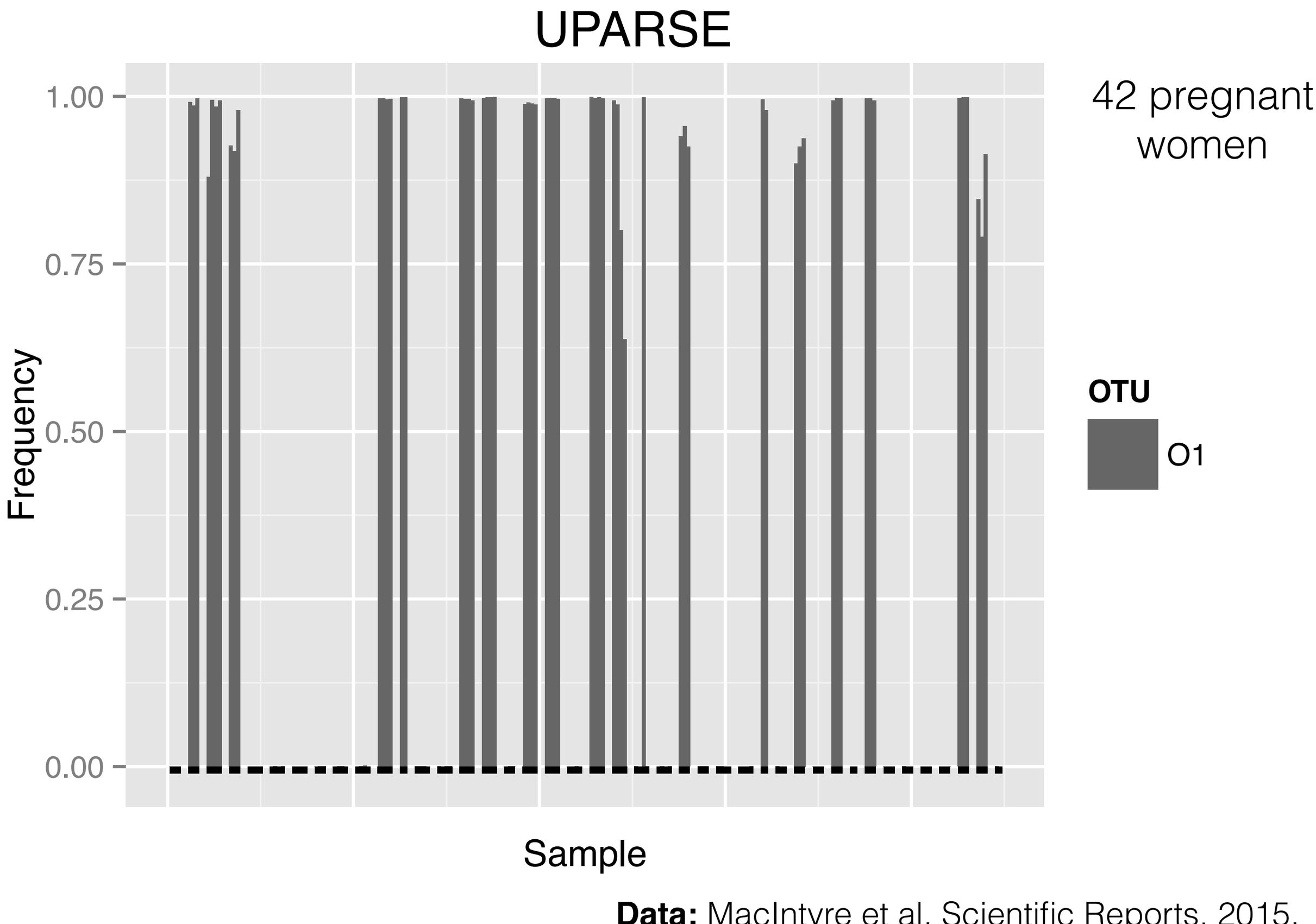


DADA2



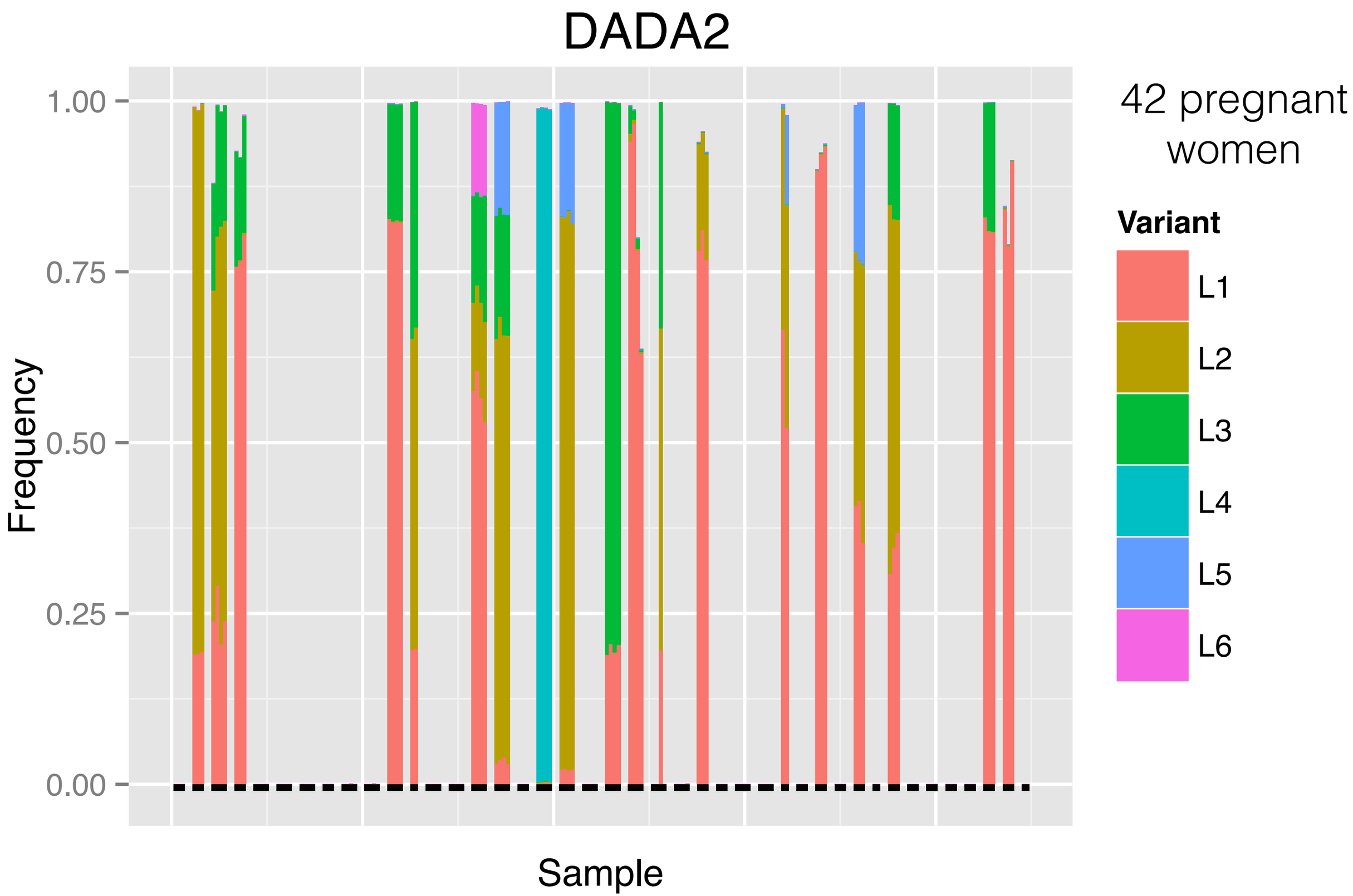
Credit: Kealoha Kinney, Michael Bunce, Andreanna Welch

Resolution: L. crispatus



Data: MacIntyre et al. Scientific Reports, 2015.

Resolution: L. crispatus



Data: MacIntyre et al. Scientific Reports, 2015.

Scaling and Comparison

The sequence is the label

`ATTAACGAGATTATAACCAGAGTACGAATA...`

is a ***consistent label***

`OTU85` is ***not***

The sequence is the label

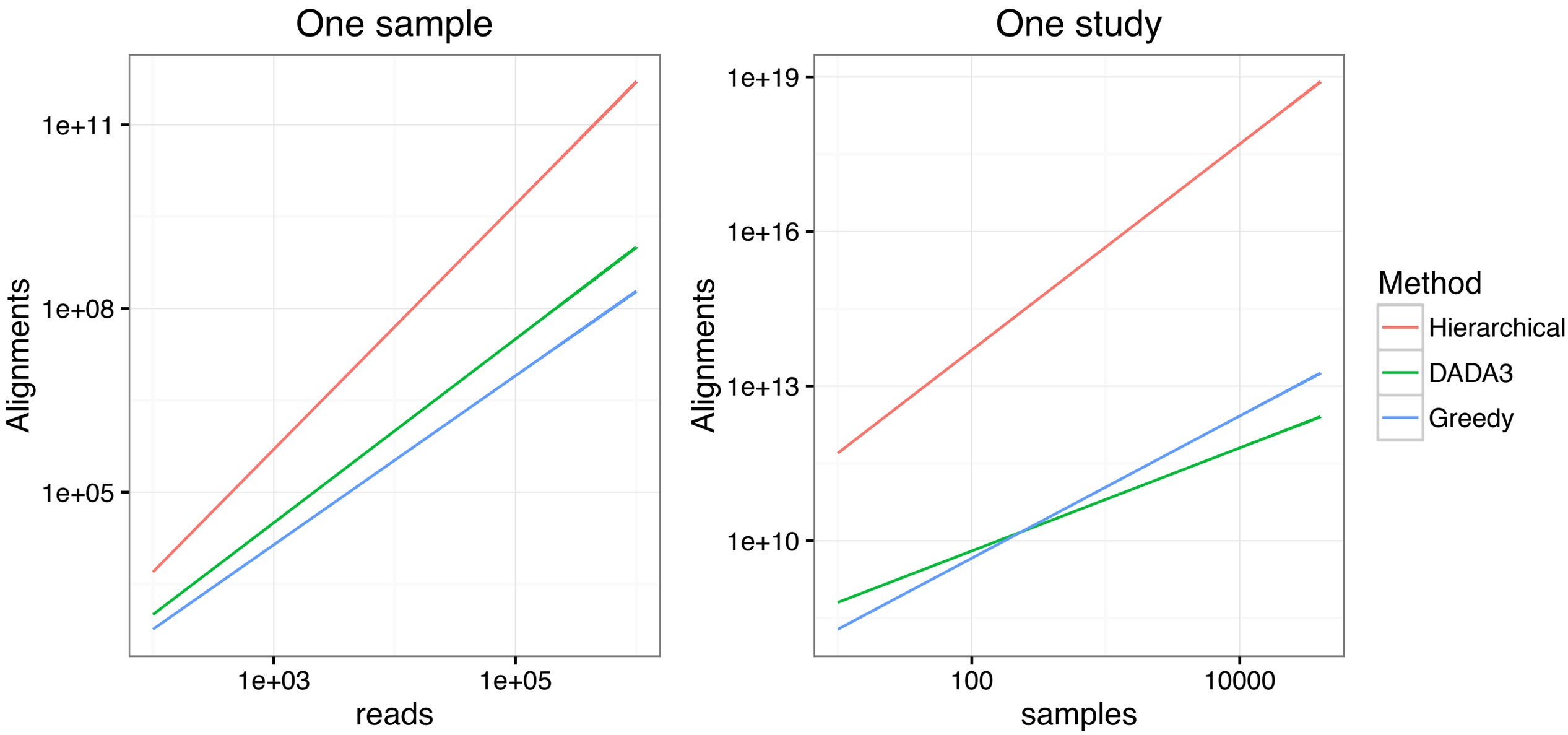
`ATTAACGAGATTATAACCAGAGTACGAATA...`

is a ***consistent label***

`OTU85` is ***not***

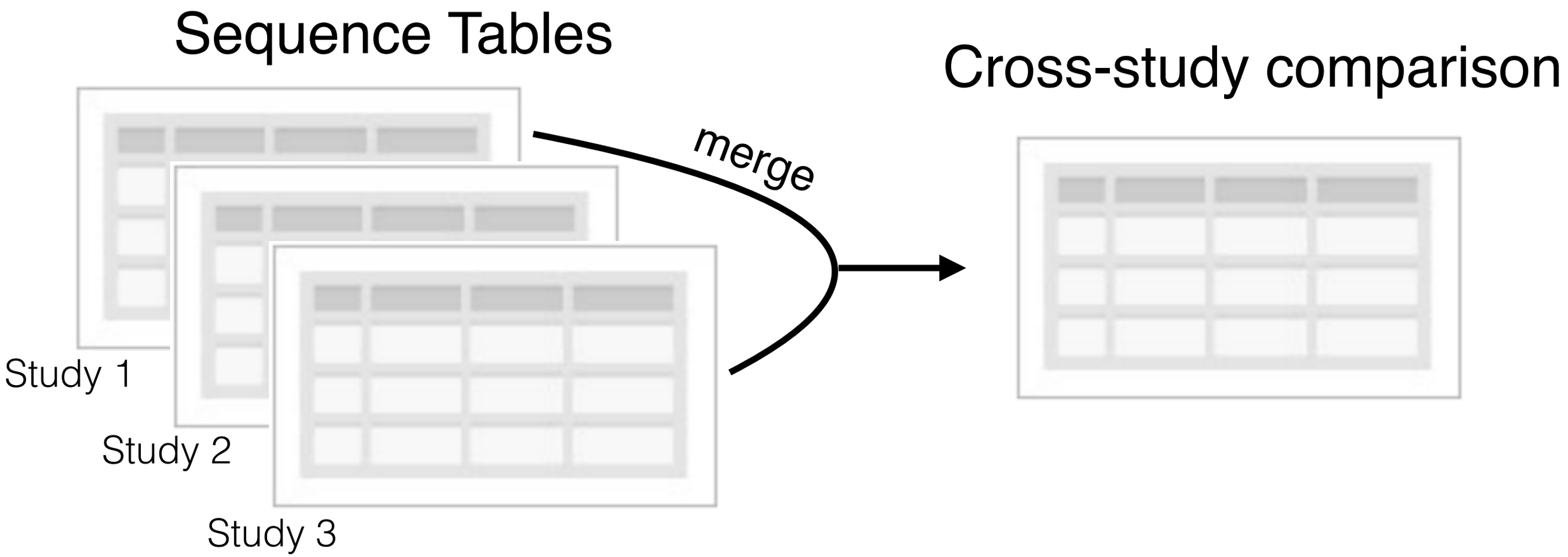
Consistent labels
allow samples to be
independently processed

Consistent Labels: Scaling



Separable processing.
Flat memory requirements.
Task parallelization.

Consistent Labels: Comparison



Eliminates need for joint reprocessing of raw data.

DADA2 is...

- A replacement for OTU picking
- Accurate and high-resolution
- Implemented in an R package
- Open source

DADA2 is...

- A replacement for OTU picking
- Accurate and high-resolution
- Implemented in an R package
- Open source

NATURE METHODS | BRIEF COMMUNICATION

DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson & Susan P Holmes

@bejcal -- https://github.com/benjjneb/dada2

Acknowledgements

Susan Holmes

Joey McMurdie

Michael Rosen

BIO-X

NIH National Institutes of Health

http://benjjneb.github.io/dada2/