

# **The DADA2 Method**

# **The amplicon inference problem**

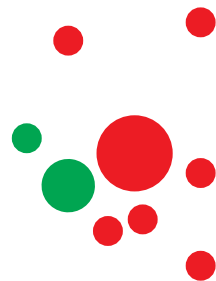
Infer the sample types and abundances  $\{(s, a)\}$   
from error-ful amplicon reads  $\{r\}$ .

sample  
sequences

amplicon reads



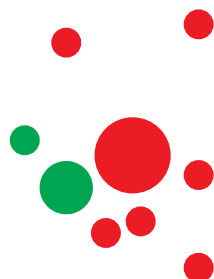
Errors



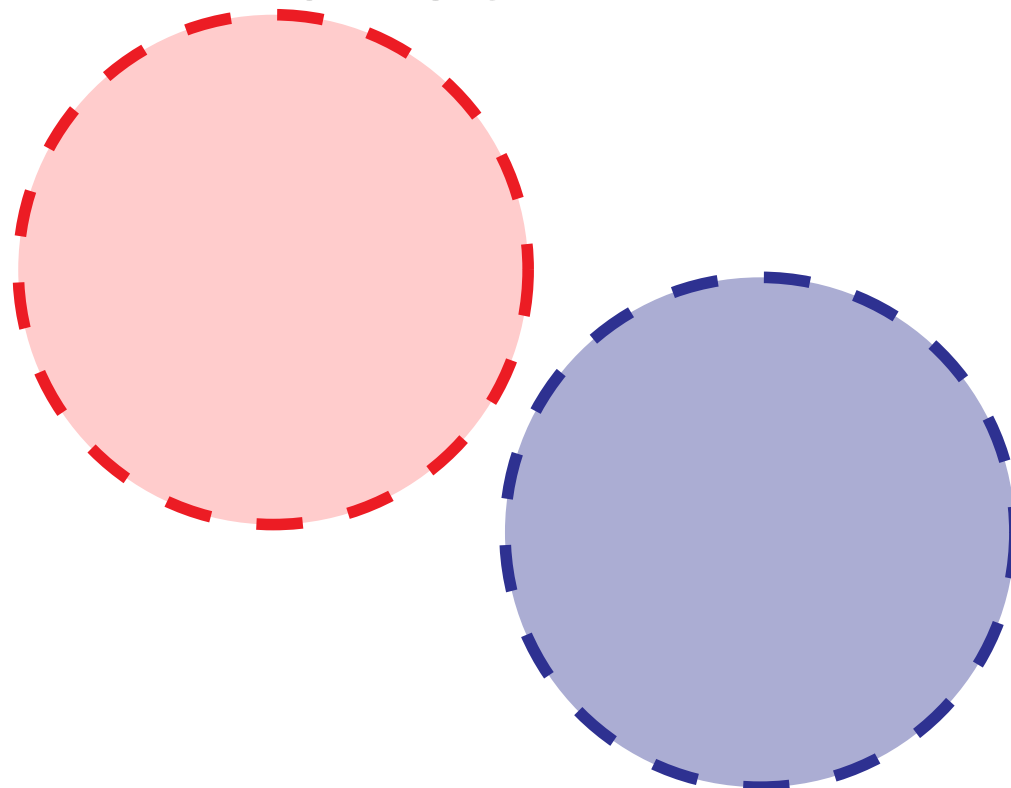
sample  
sequences



amplicon reads



OTUs



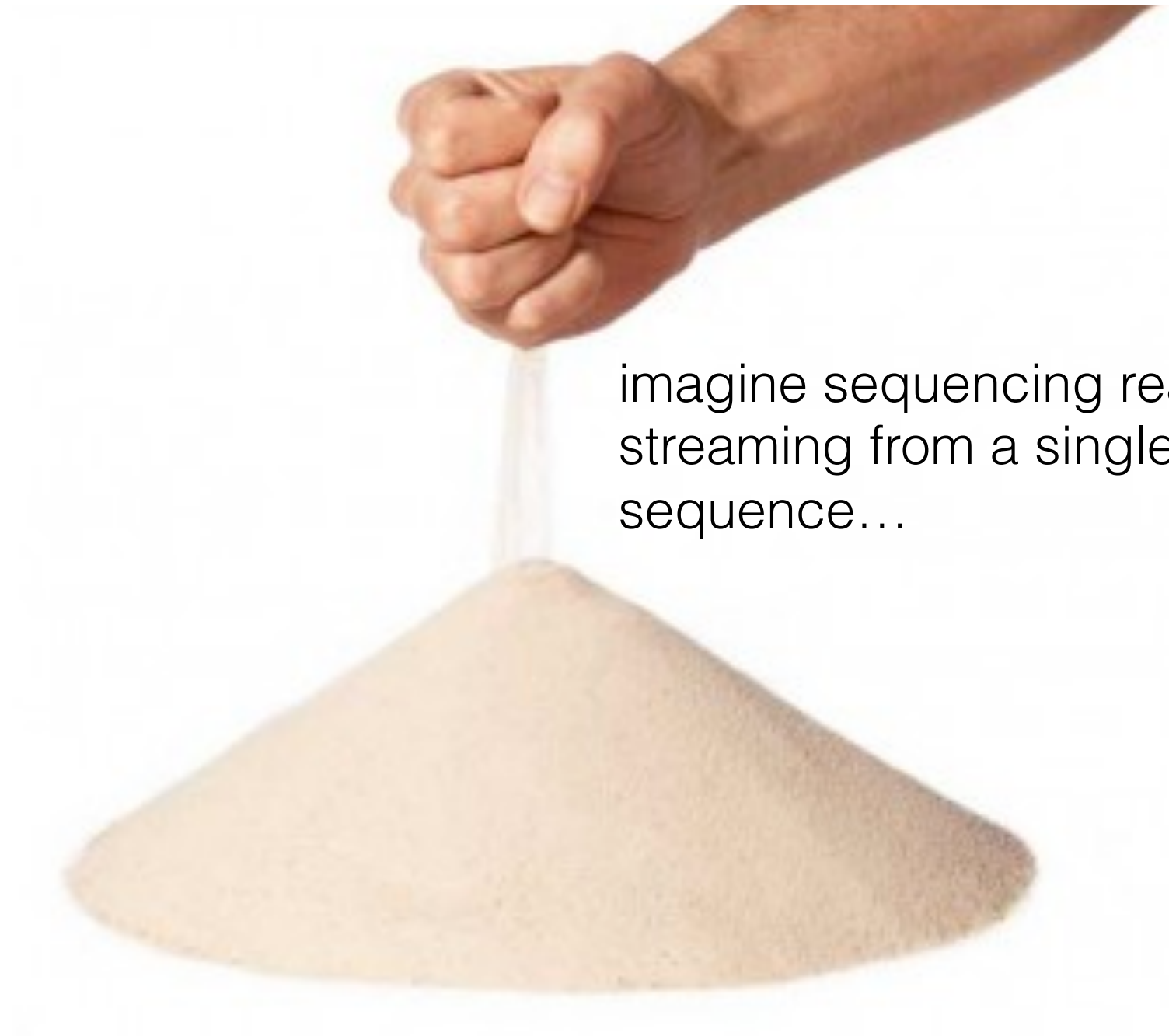
Errors



Make OTUs



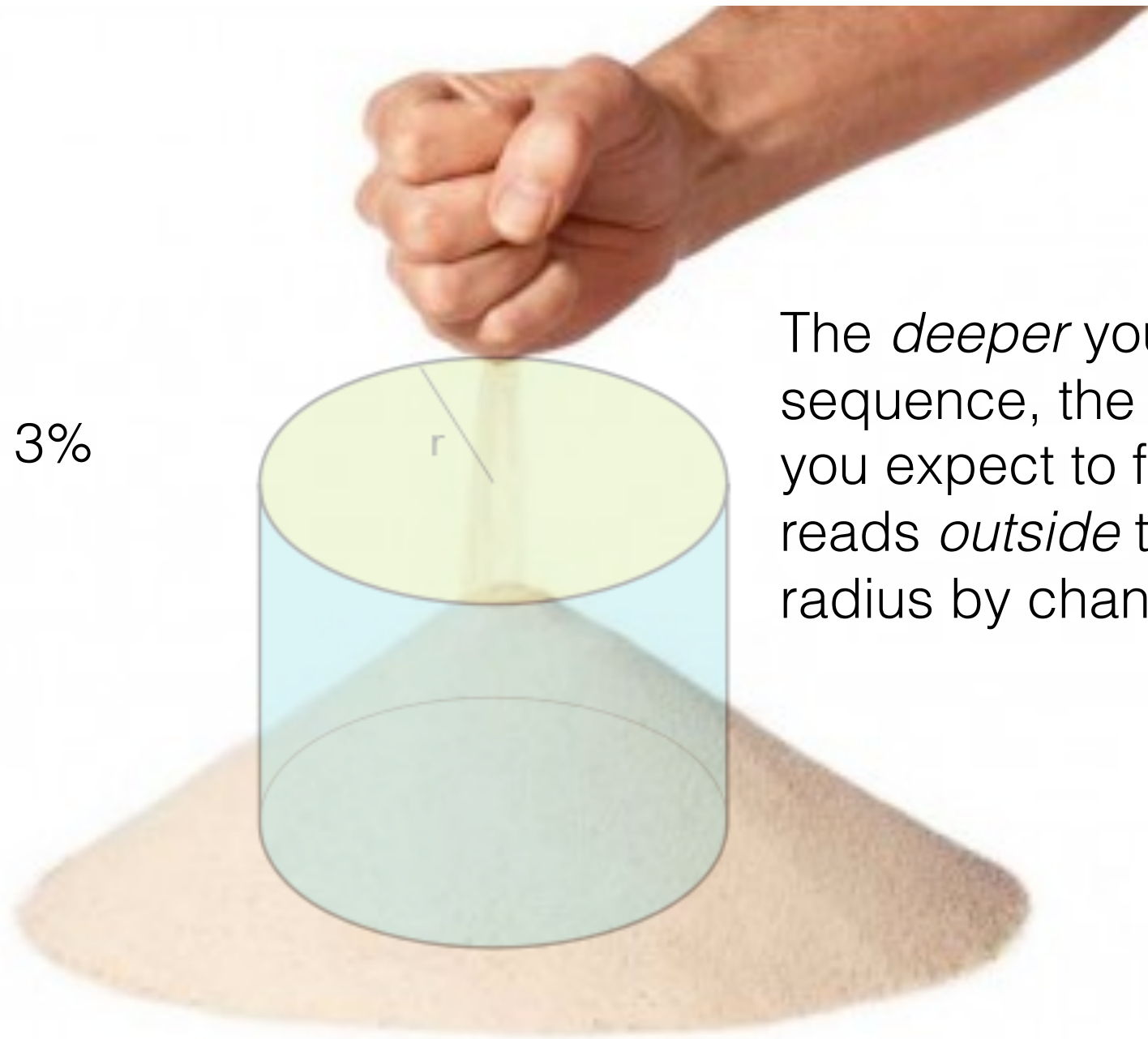
Motivation: Lingering problems with “OTU”



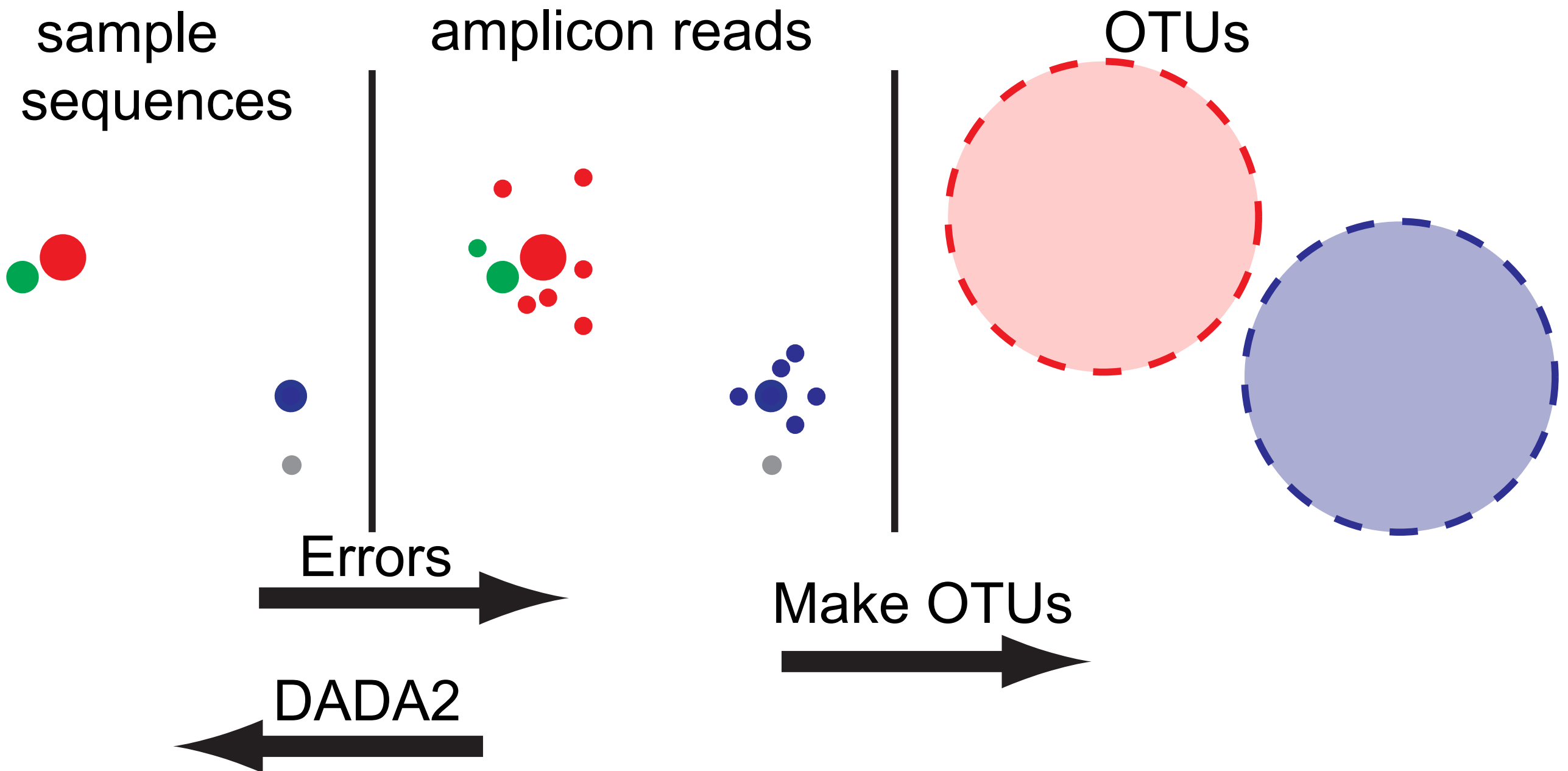
imagine sequencing reads  
streaming from a single true  
sequence...

## Motivation: Lingering problems with “OTU”

$r = 3\%$



The *deeper* you sequence, the more you expect to find reads *outside* the radius by chance.





Amplicon Sequencing. **Exactly.** *Version 1.8*



# Error Model

# An Error Model

---

**s :** ATTAACGAGATTATAACCCAGAGTACGAATA . . .  
          |  |  
**r :** ATCAACGAGATTATAACAAGAGTACGAATA . . .

# An Error Model

---

**s :** ATTAACGAGATTATAACCCAGAGTACGAATA . . .  
          |  |  
**r :** ATCAACGAGATTATAACAAGAGTACGAATA . . .

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

# An Error Model

---

**s :** ATTAACGAGATTATAACCCAGAGTACGAATA . . .  
          |  |  
**r :** ATCAACGAGATTATAACAAGAGTACGAATA . . .

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

**Error process is independent across nucleotides.**

Per-nucleotide transition rate depends on:

- Sample nucleotide
- Read nucleotide
- Read quality at that position
- Batch effect (eg. run)

# Quantifying overabundance

---

Assuming reads are independent:

$$p_A(r|s) = \frac{\sum_{a=a_r}^{\infty} \rho_{pois}(a_s p(r|s), a)}{1 - \rho_{pois}(a_s p(r|s), 0)}$$

# Quantifying overabundance

---

Assuming reads are independent:

$$p_A(r|s) = \frac{\sum_{a=a_r}^{\infty} \rho_{pois}(a_s p(r|s), a)}{1 - \rho_{pois}(a_s p(r|s), 0)}$$

The **abundance p-value**  $p_A$  quantifies the notion that there are too many reads with sequence  $r$  to be explained by sequencing errors.

# Quantifying overabundance

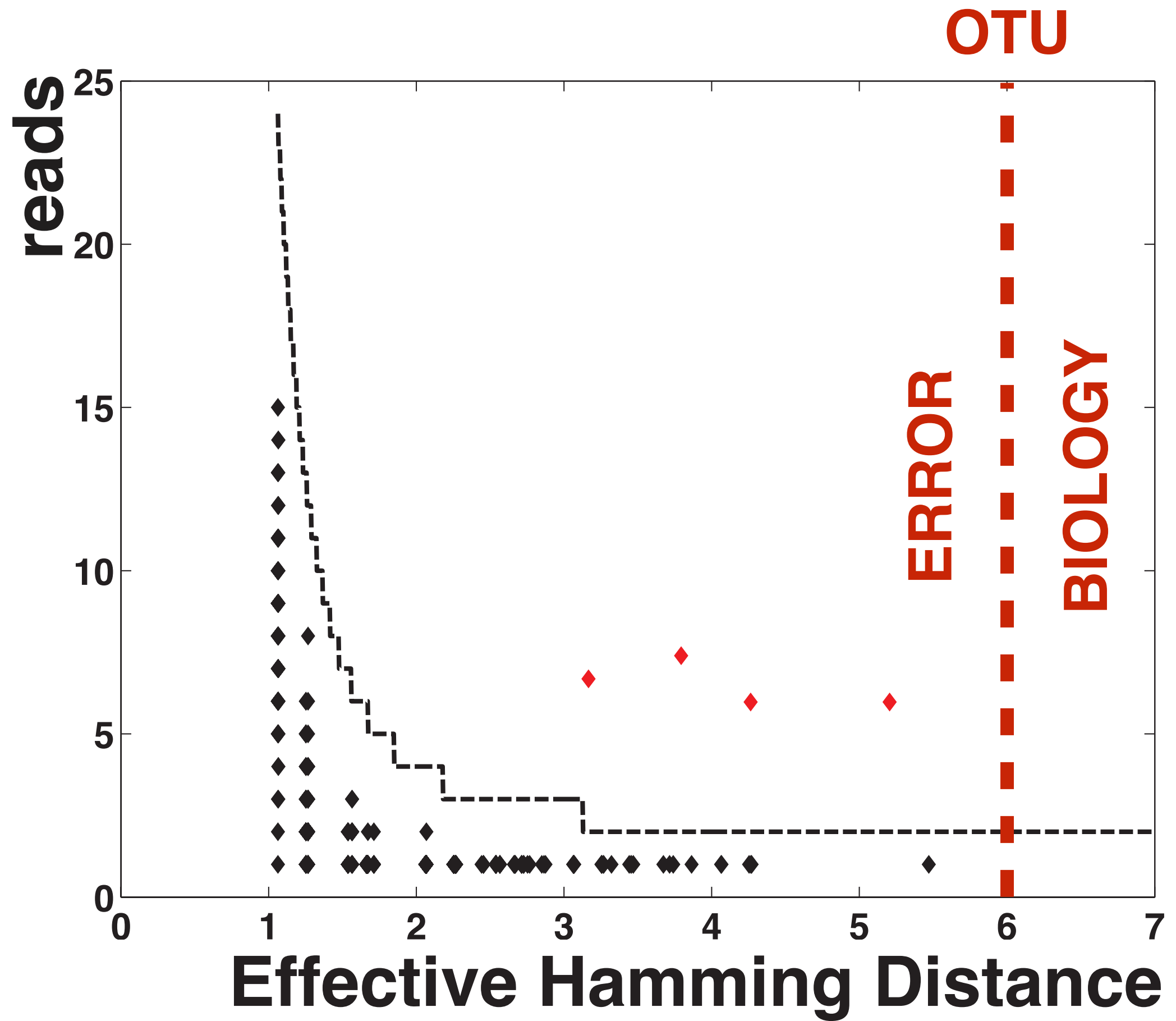
---

Assuming reads are independent:

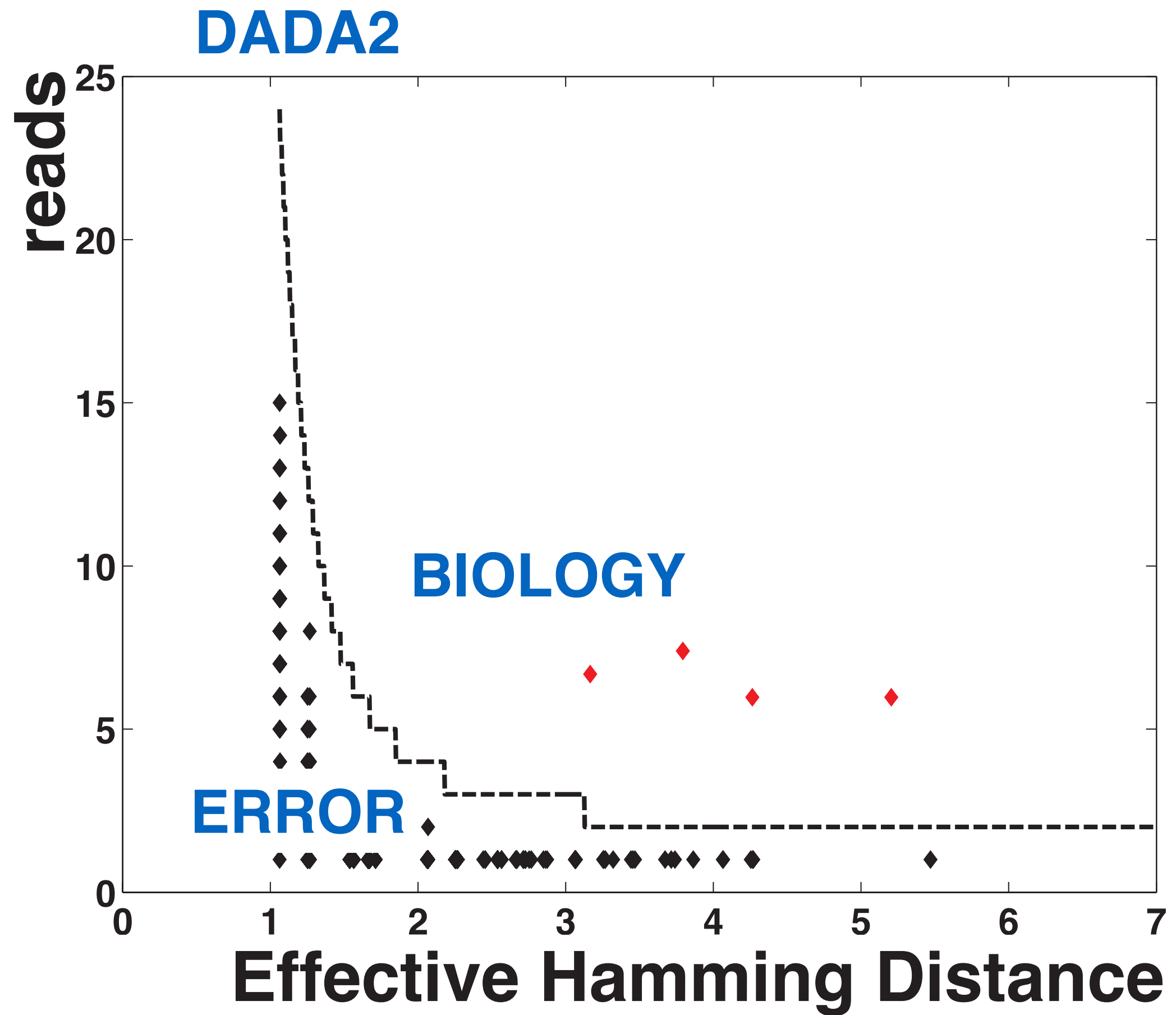
$$p_A(r|s) = \frac{\sum_{a=a_r}^{\infty} \rho_{pois}(a_s p(r|s), a)}{1 - \rho_{pois}(a_s p(r|s), 0)}$$

The **abundance p-value**  $p_A$  quantifies the notion that there are too many reads with sequence  $r$  to be explained by sequencing errors.

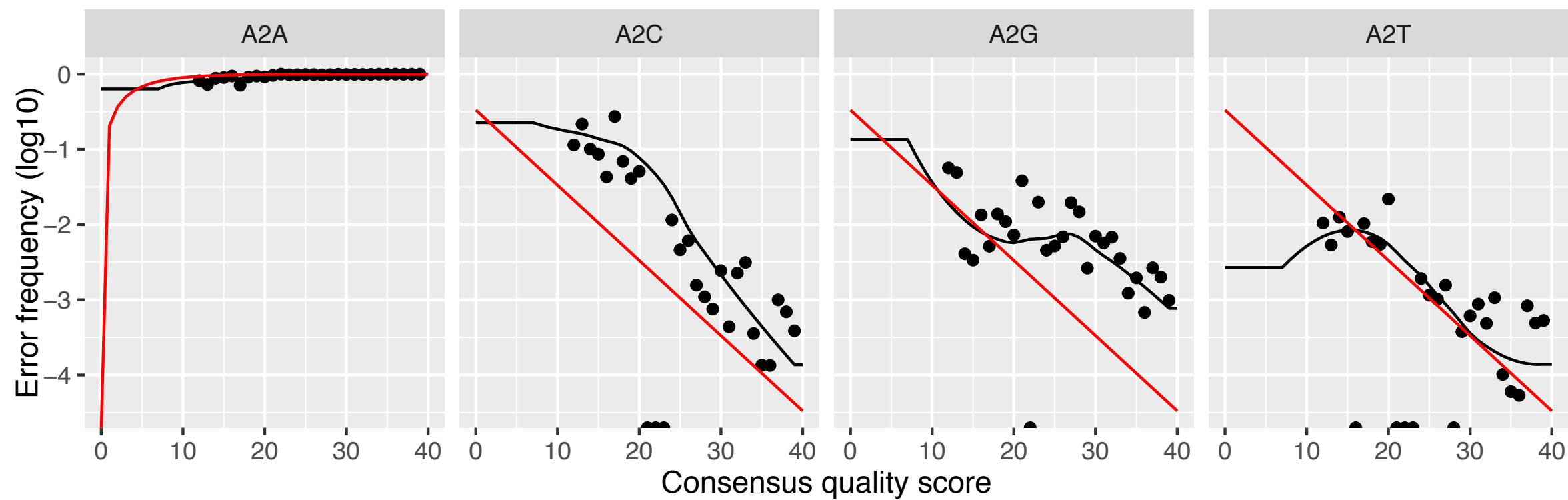
This is our criteria to distinguish biology from errors.



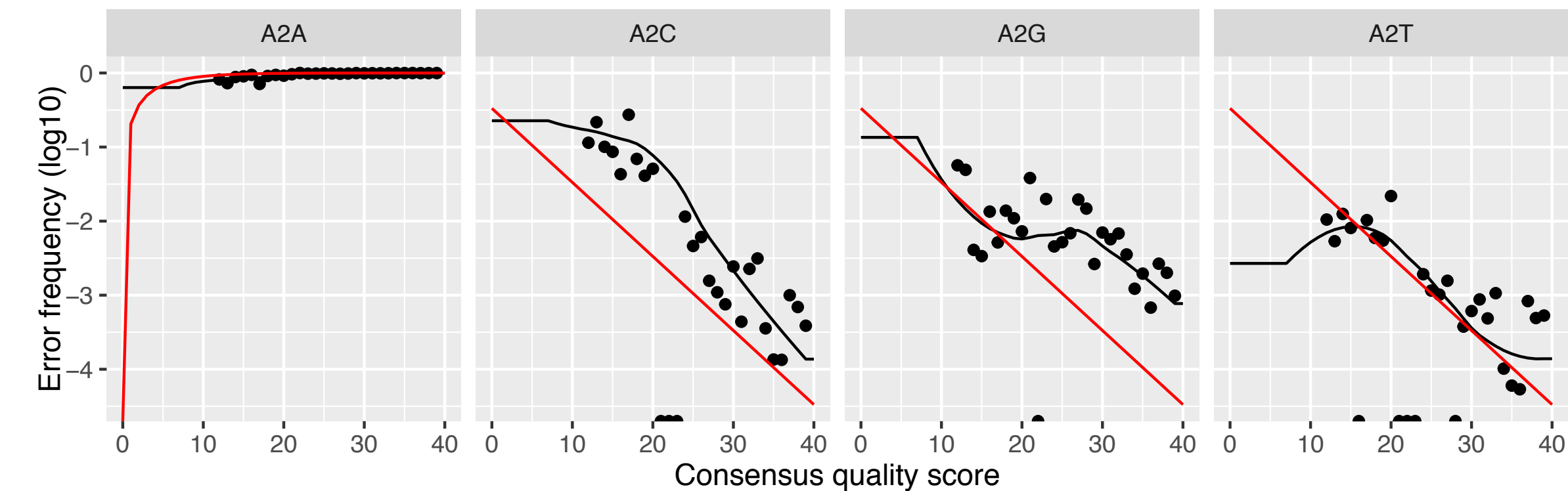




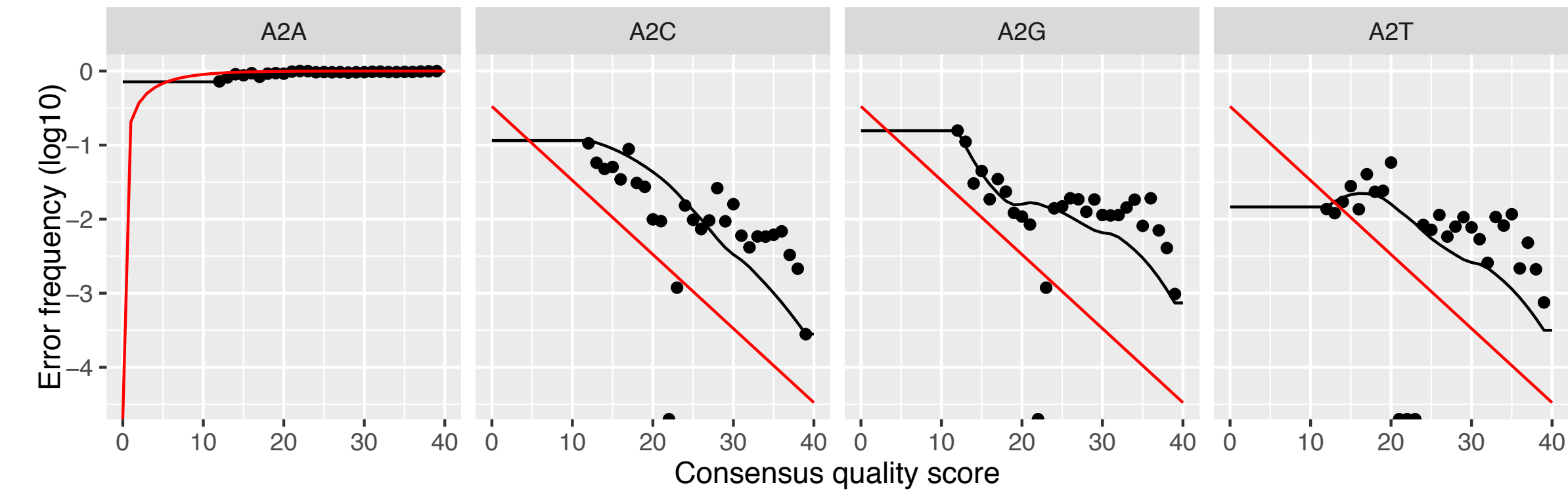
# Learning Errors



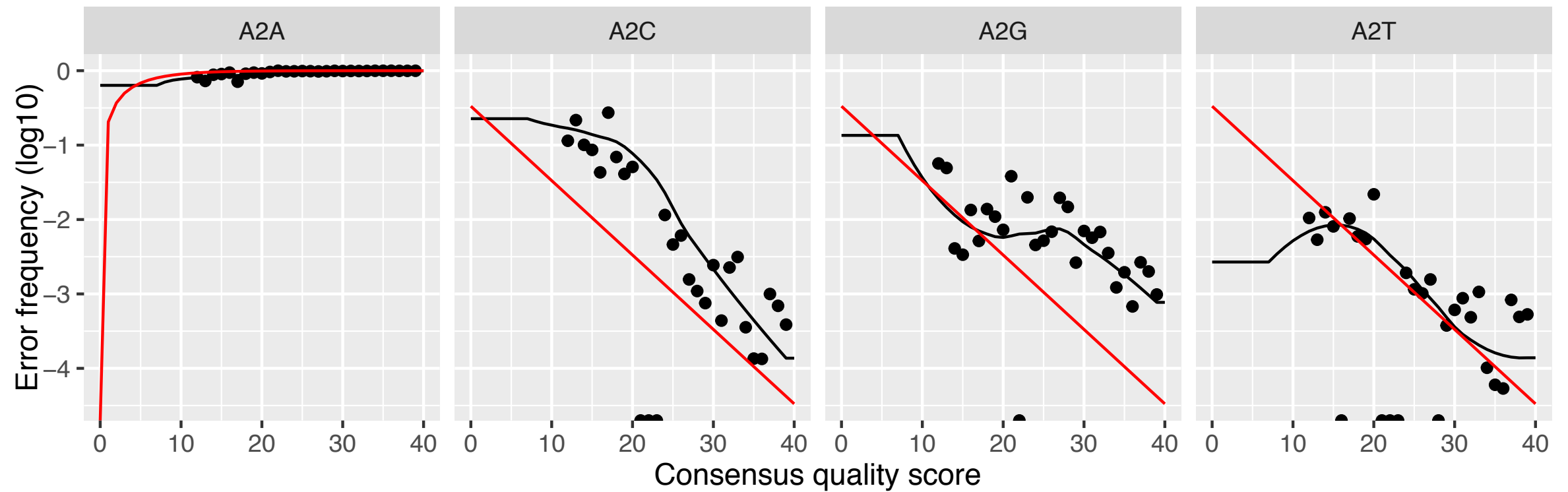
# Study A



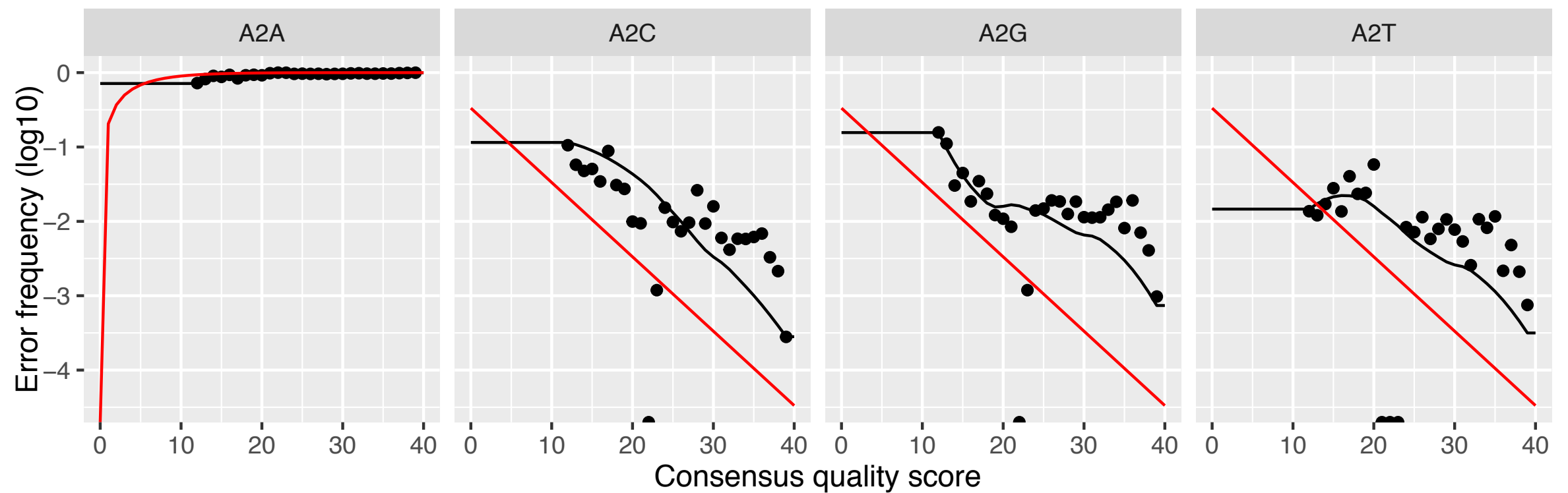
# Study B



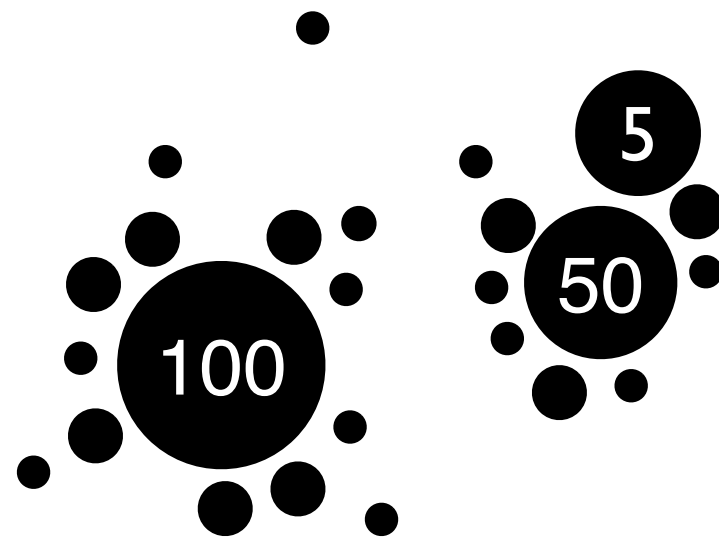
# Study A



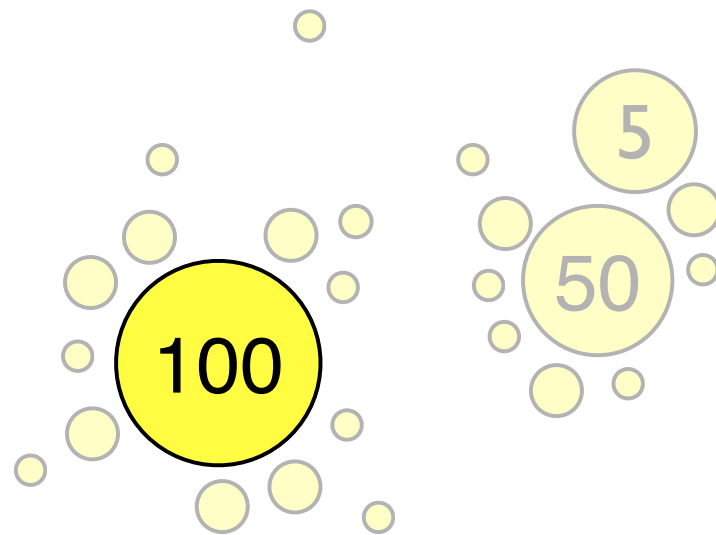
# Study B



**But How?**



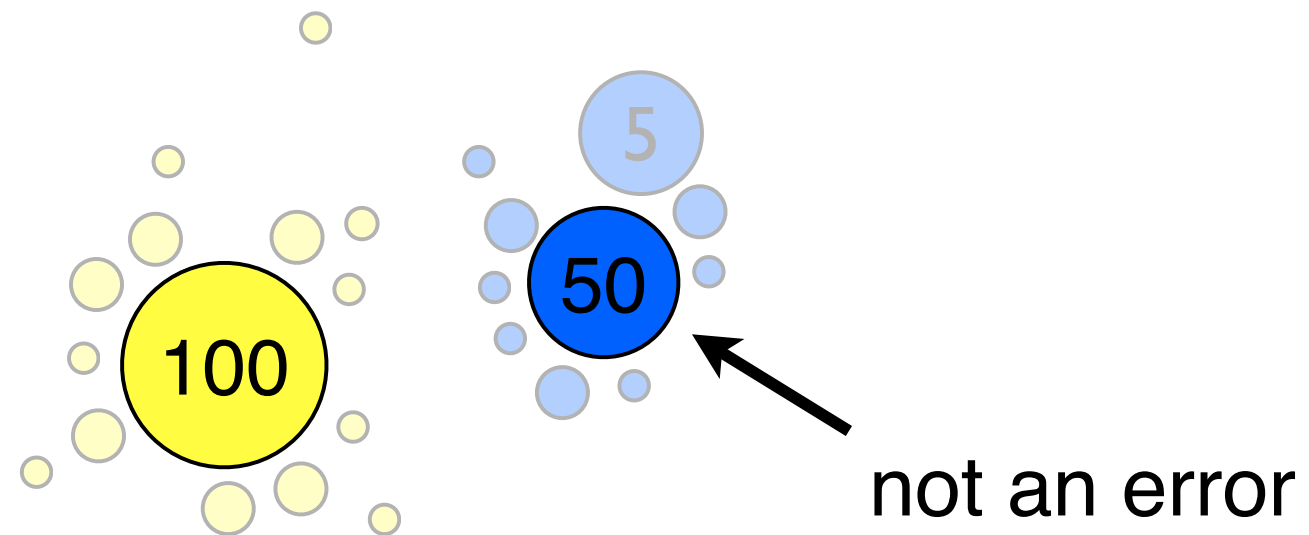
Initial guess: one real sequence + errors



**Infer initial *error model*** under this assumption.

$\text{Pr}(i \rightarrow j) =$

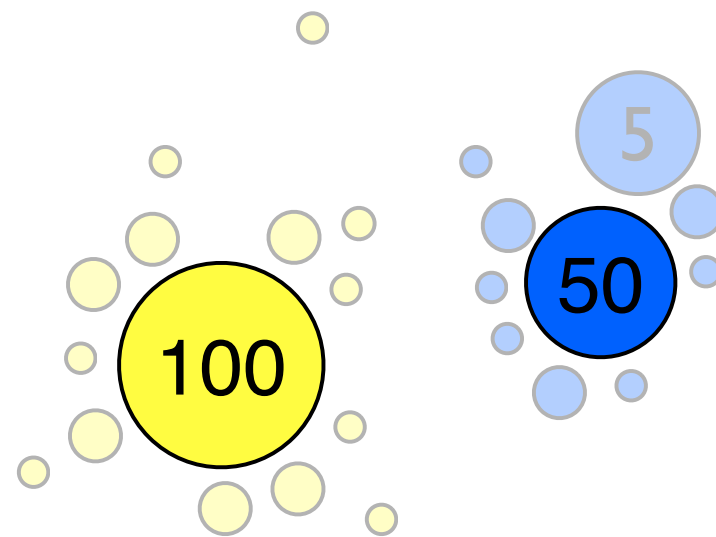
	A	C	G	T
A	0.97	$10^{-2}$	$10^{-2}$	$10^{-2}$
C	$10^{-2}$	0.97	$10^{-2}$	$10^{-2}$
G	$10^{-2}$	$10^{-2}$	0.97	$10^{-2}$
T	$10^{-2}$	$10^{-2}$	$10^{-2}$	0.97



**Reject** unlikely error under model. **Recruit** errors.

	A	C	G	T
A	0.97	$10^{-2}$	$10^{-2}$	$10^{-2}$
C	$10^{-2}$	0.97	$10^{-2}$	$10^{-2}$
G	$10^{-2}$	$10^{-2}$	0.97	$10^{-2}$
T	$10^{-2}$	$10^{-2}$	$10^{-2}$	0.97

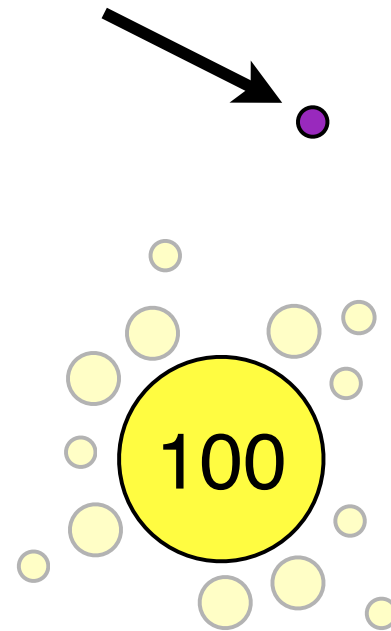




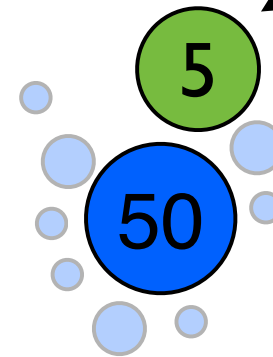
Update the model.

	A	C	G	T
A	0.997	$10^{-3}$	$10^{-3}$	$10^{-3}$
C	$10^{-3}$	0.997	$10^{-3}$	$10^{-3}$
G	$10^{-3}$	$10^{-3}$	0.997	$10^{-3}$
T	$10^{-3}$	$10^{-3}$	$10^{-3}$	0.997

not an error

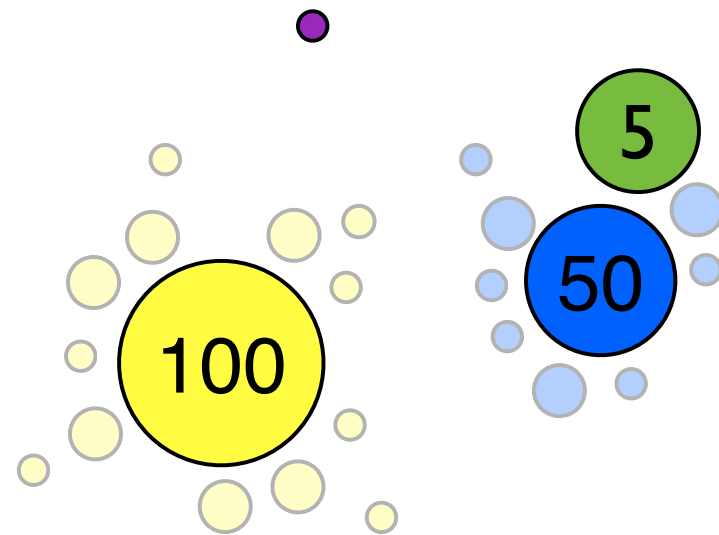


not an error



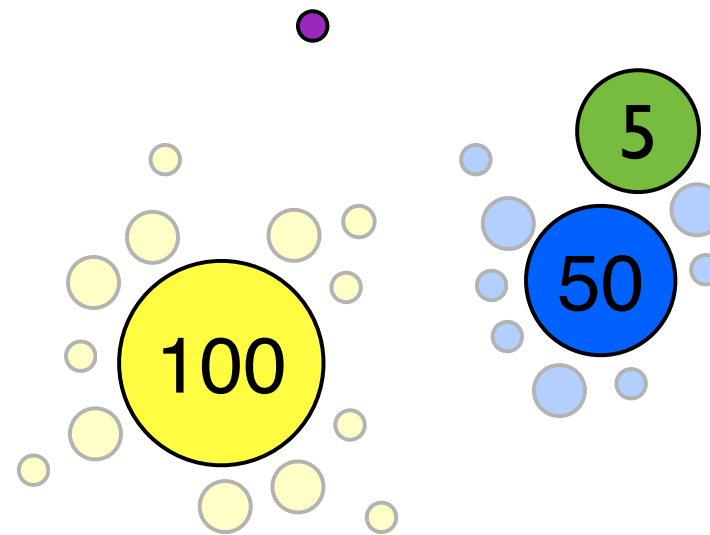
**Reject** more sequences under *new* model

	A	C	G	T
A	0.997	$10^{-3}$	$10^{-3}$	$10^{-3}$
C	$10^{-3}$	0.997	$10^{-3}$	$10^{-3}$
G	$10^{-3}$	$10^{-3}$	0.997	$10^{-3}$
T	$10^{-3}$	$10^{-3}$	$10^{-3}$	0.997



Update model again

	A	C	G	T
A	0.998	$1 \times 10^{-4}$	$2 \times 10^{-3}$	$2 \times 10^{-4}$
C	$6 \times 10^{-5}$	0.999	$3 \times 10^{-6}$	$1 \times 10^{-3}$
G	$1 \times 10^{-3}$	$3 \times 10^{-6}$	0.999	$6 \times 10^{-5}$
T	$2 \times 10^{-4}$	$2 \times 10^{-3}$	$1 \times 10^{-4}$	0.998



**Convergence:** all errors are plausible

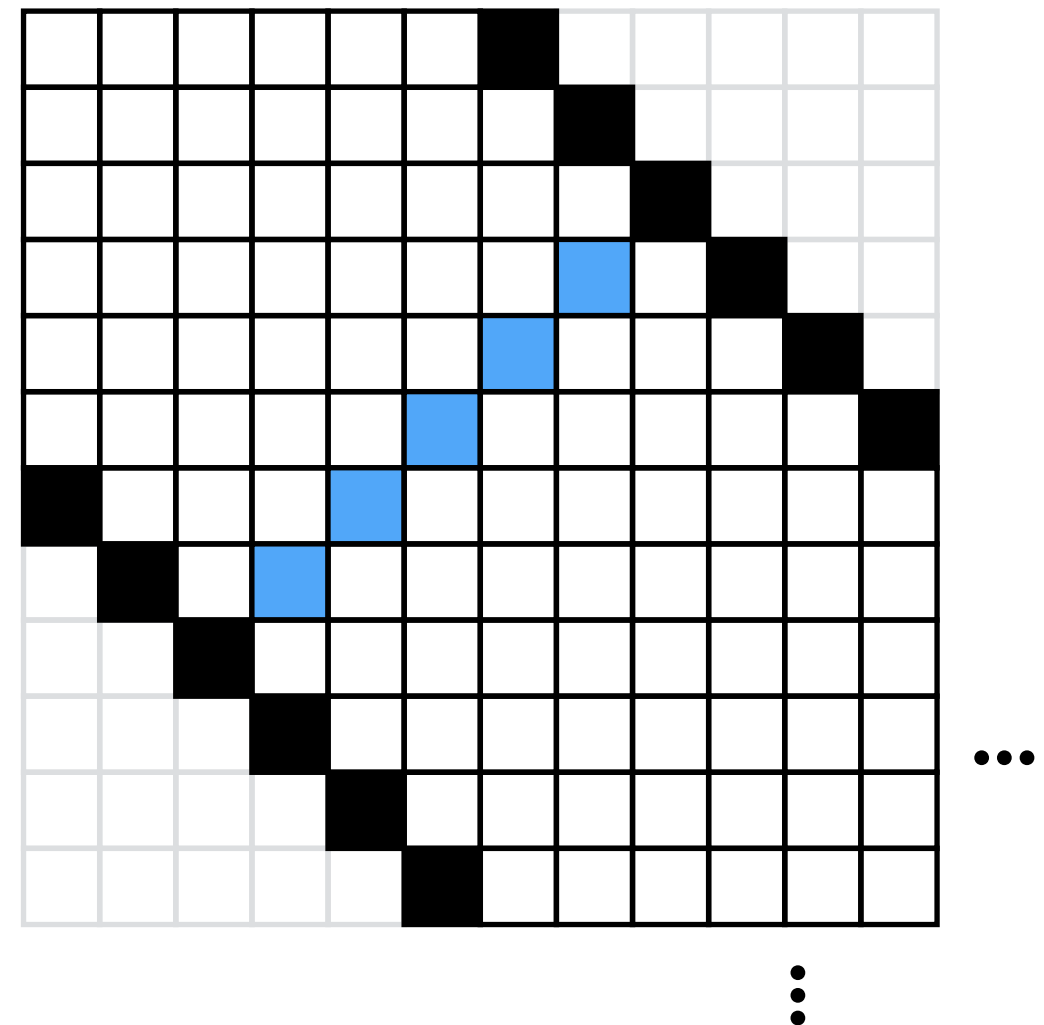
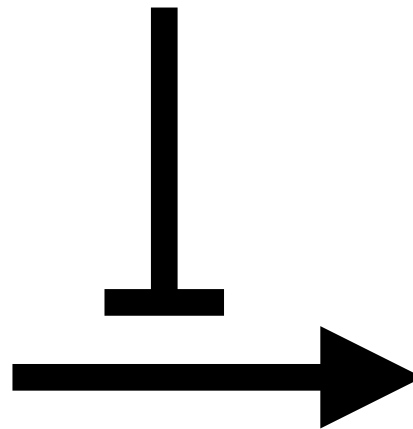
	A	C	G	T
A	0.998	$1 \times 10^{-4}$	$2 \times 10^{-3}$	$2 \times 10^{-4}$
C	$6 \times 10^{-5}$	0.999	$3 \times 10^{-6}$	$1 \times 10^{-3}$
G	$1 \times 10^{-3}$	$3 \times 10^{-6}$	0.999	$6 \times 10^{-5}$
T	$2 \times 10^{-4}$	$2 \times 10^{-3}$	$1 \times 10^{-4}$	0.998

**Speed**

# Efficient Sequence Alignment

$$\text{Kmer-dist}(i,j) > \mathbf{K}^*$$

i: ATTAACGAGATTA...  
j: TTAAACGAGAATA...

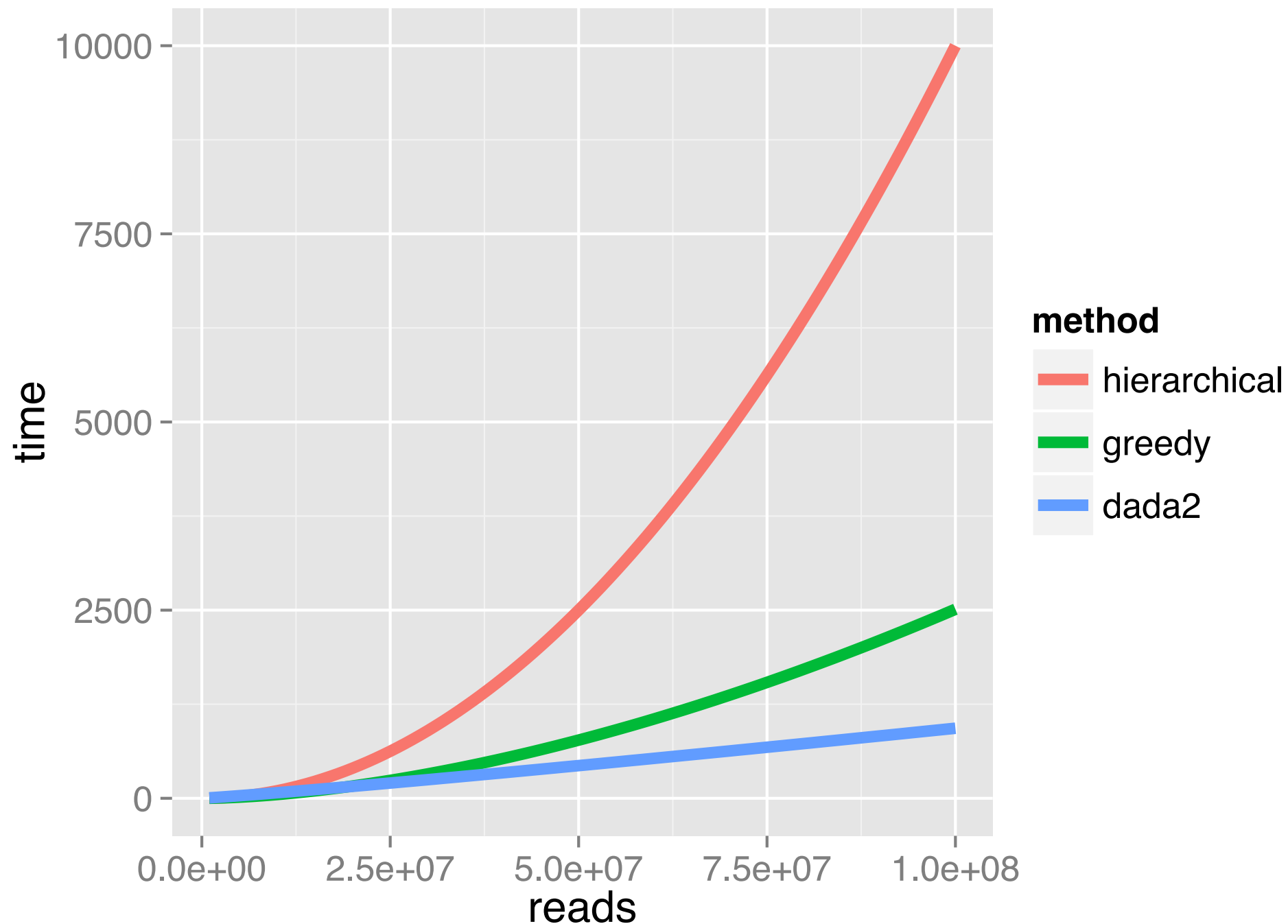


**Vectorized Banded**  
Needleman-Wunsch

**~200X speedup**

# Independent Sample Processing

---



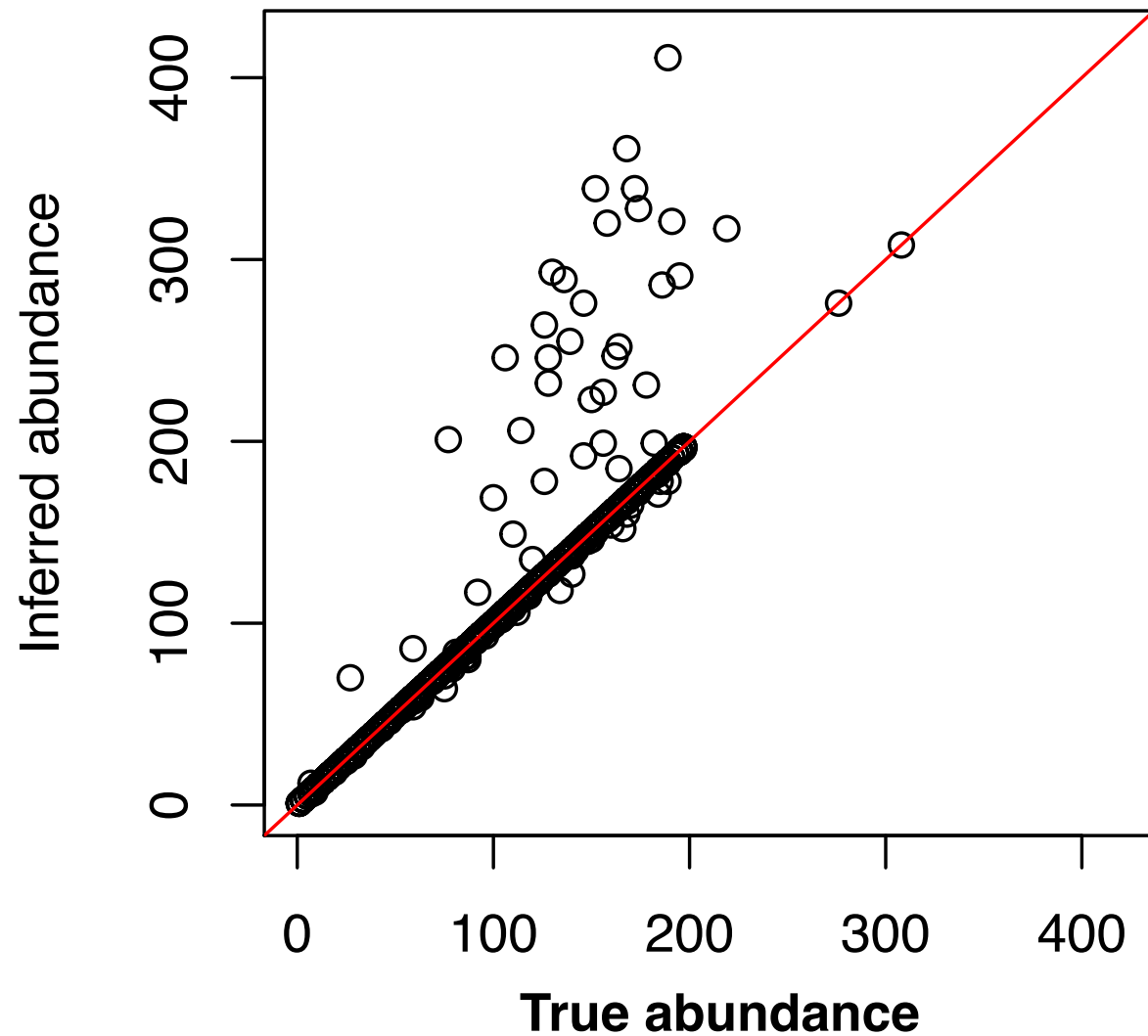
Exactly inferred sequences are comparable  
across samples. Separable processing.  
**De novo sample inference in linear time.**

# **Accuracy** and **Resolution**



# Accuracy: Simulated data

mothur (an)



**TP:** 978

**FP:** 272

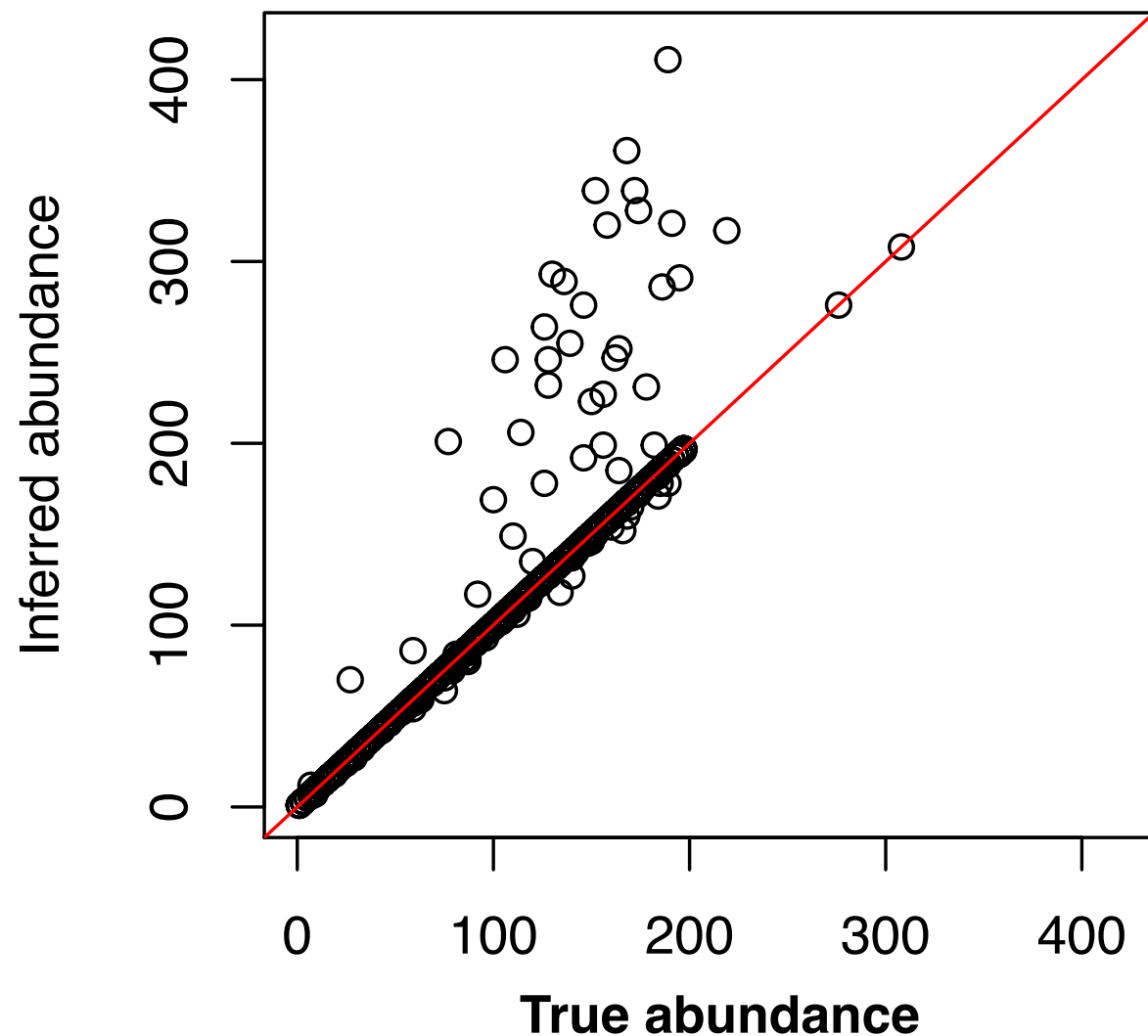
**FN:** 77

**cor:** 0.935

**Data:** Kopylova, et al. mSystems, 2016.

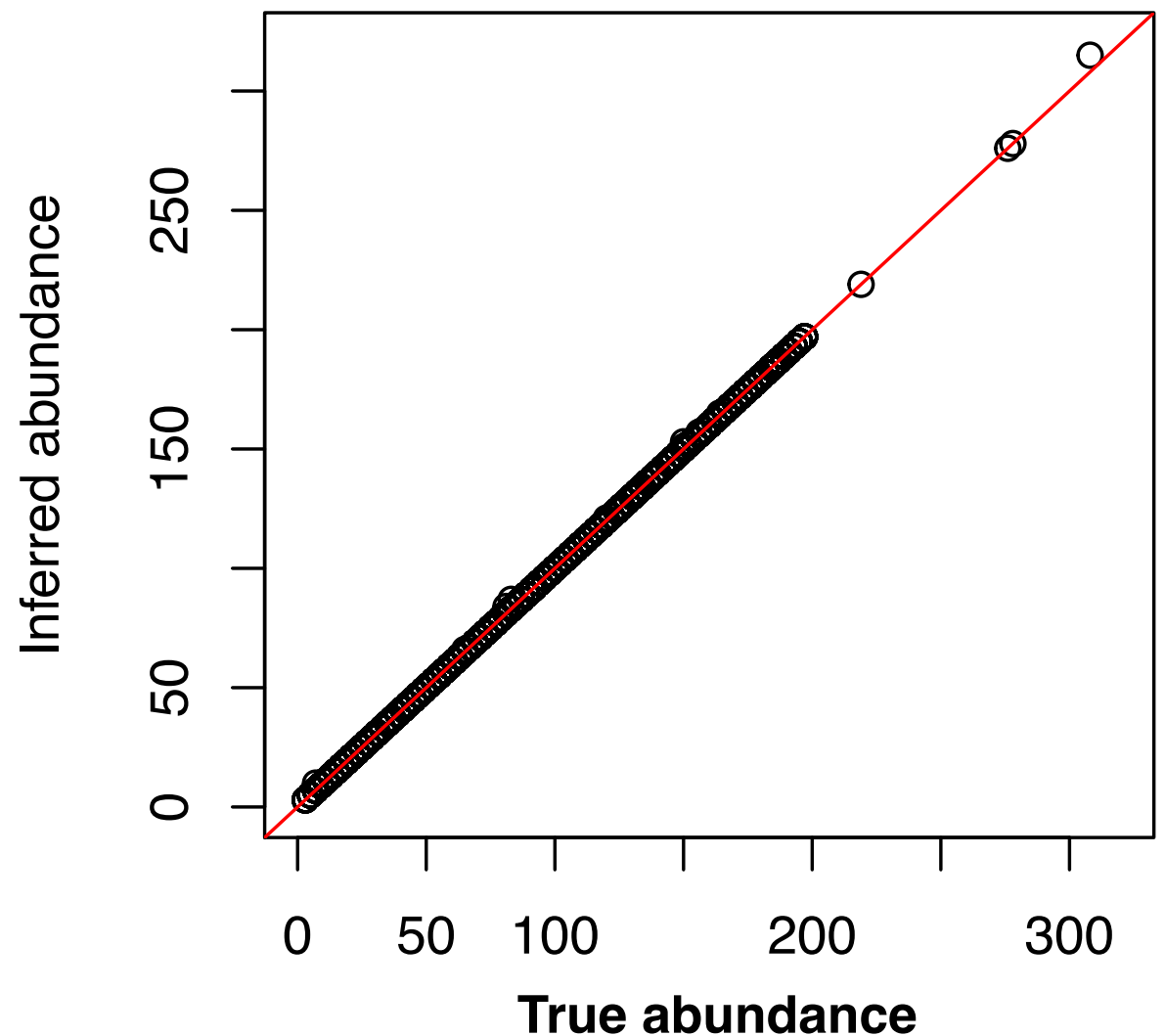
# Accuracy: Simulated data

mothur (an)



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

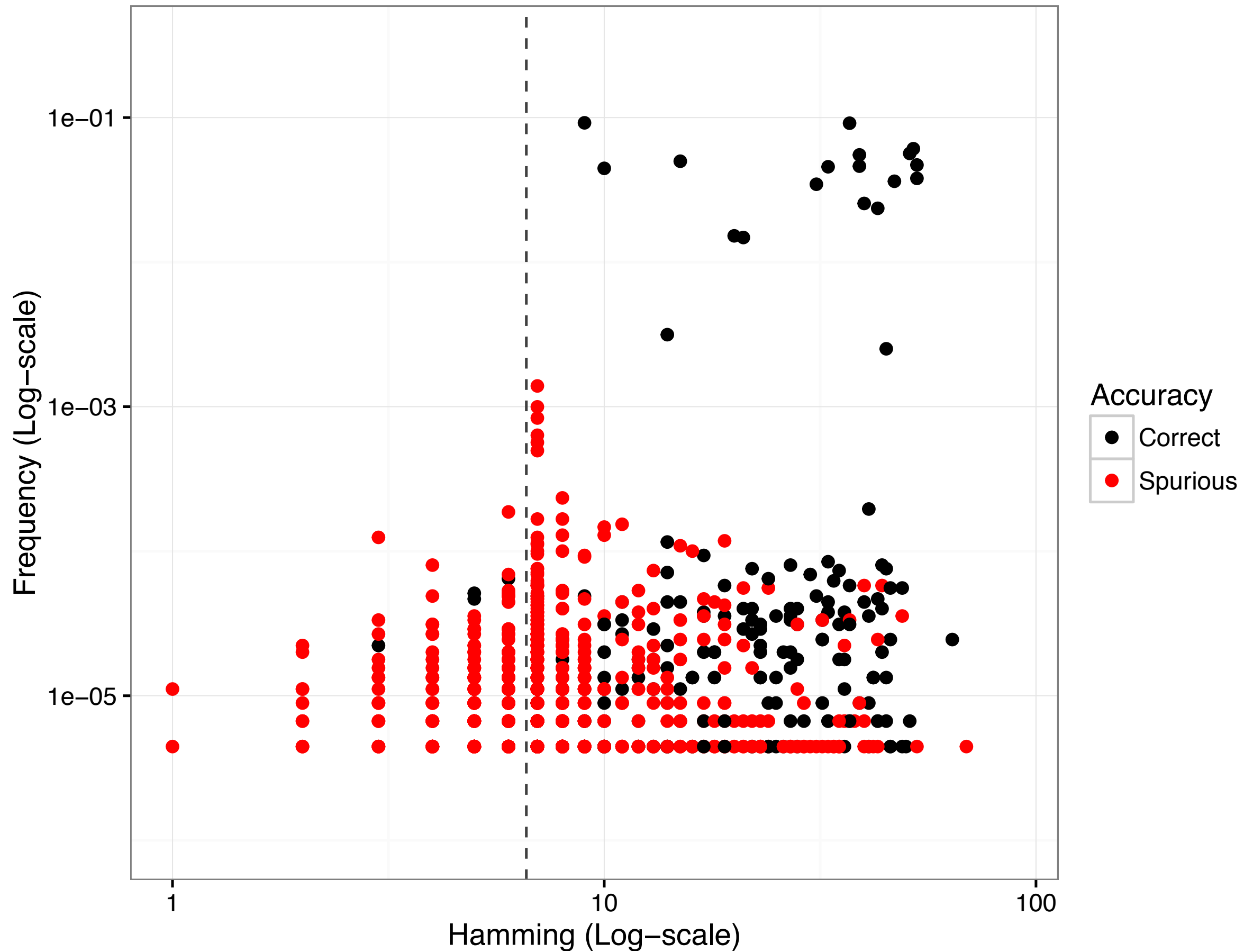
DADA2



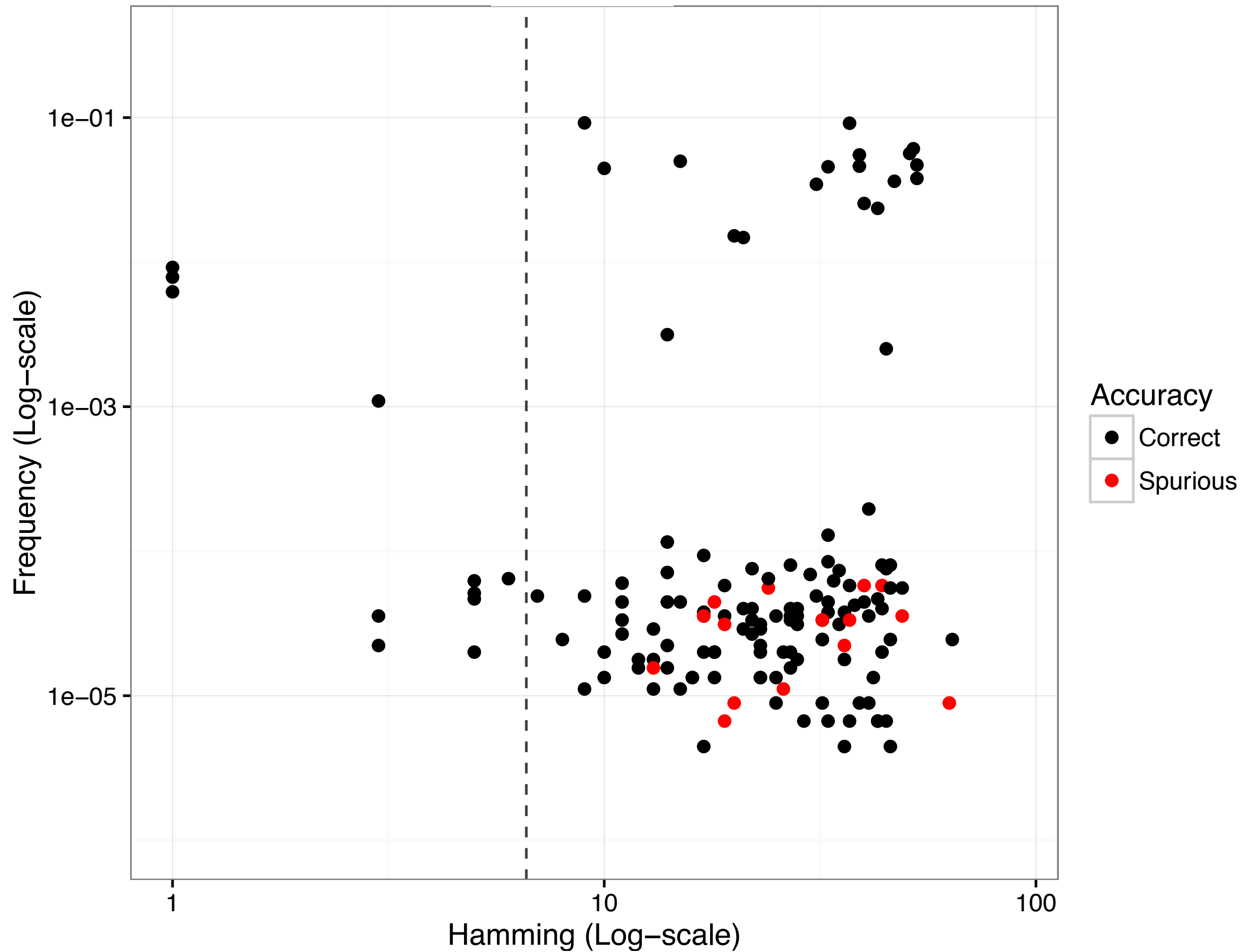
**TP:** 1042  
**FP:** 0  
**FN:** 13  
**cor:** 0.999

**Data:** Kopylova, et al. mSystems, 2016.

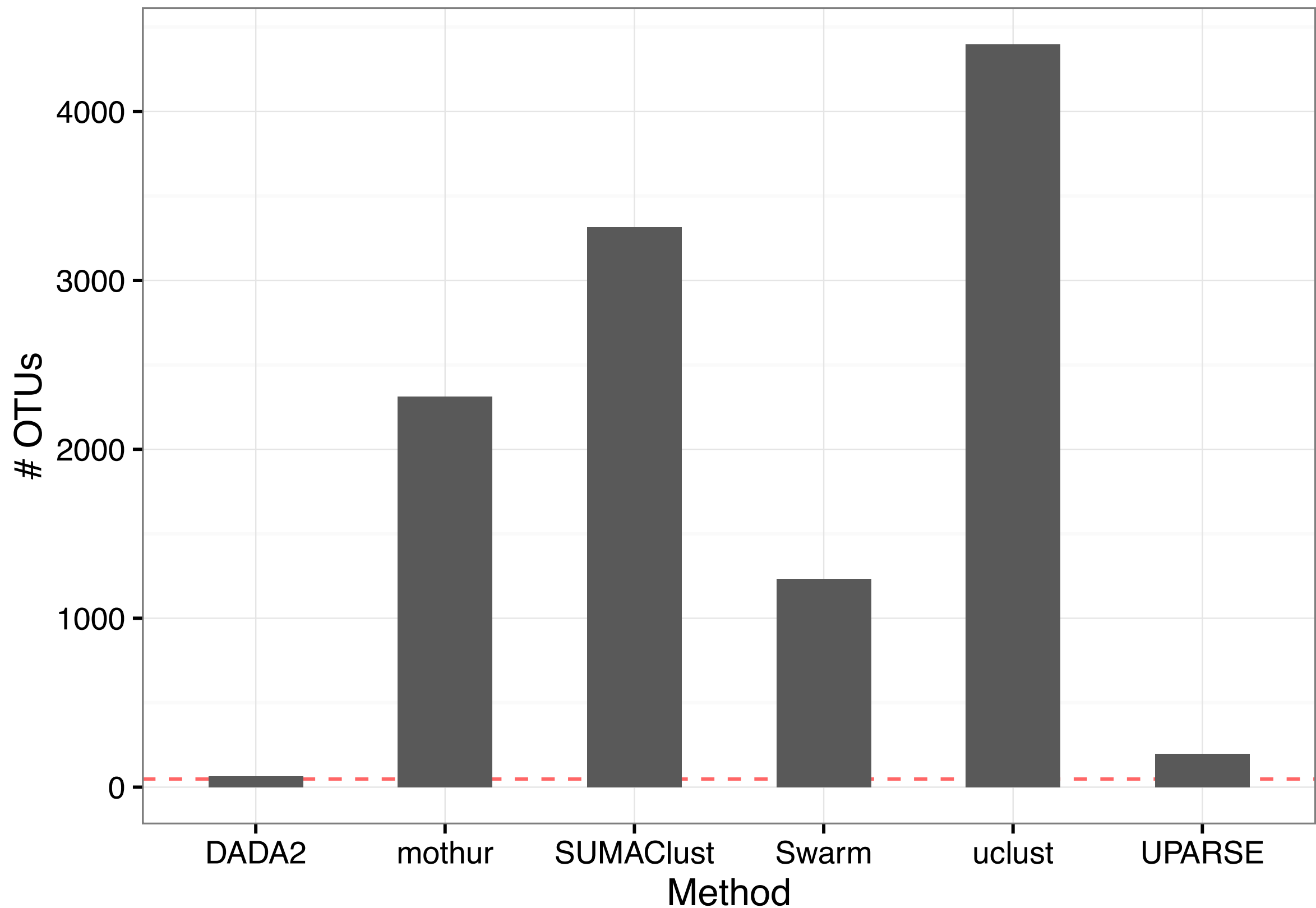
# QIIME1/uclust



# DADA2

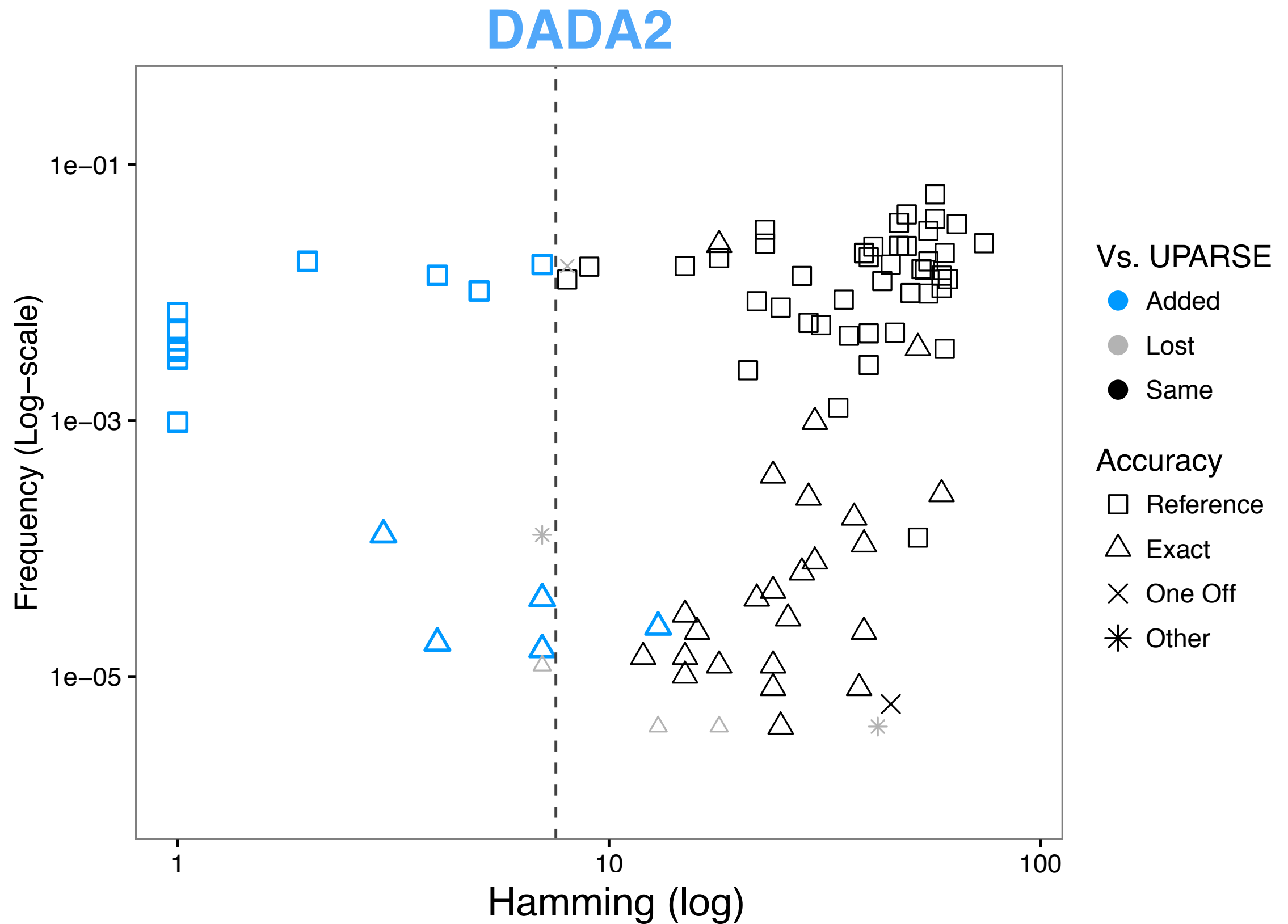


# Accuracy: Mock community



**Credit:** Kopylova, et al. mSystems, 2016.

# Resolution: Mock Community

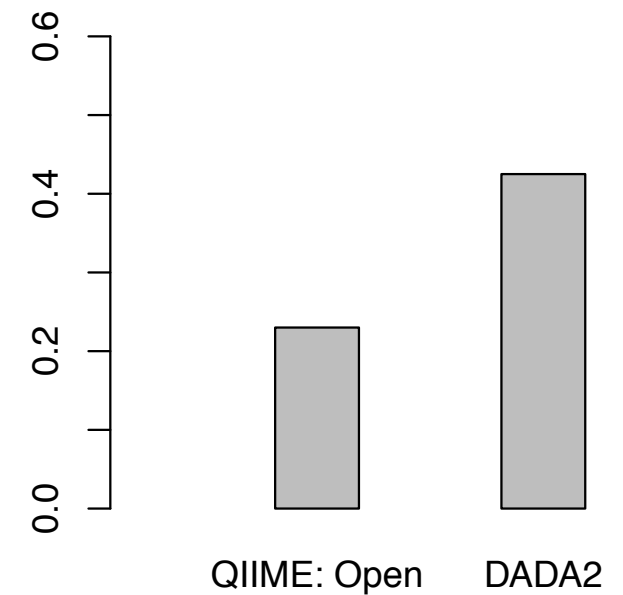


# Accuracy: Arsenic treatment

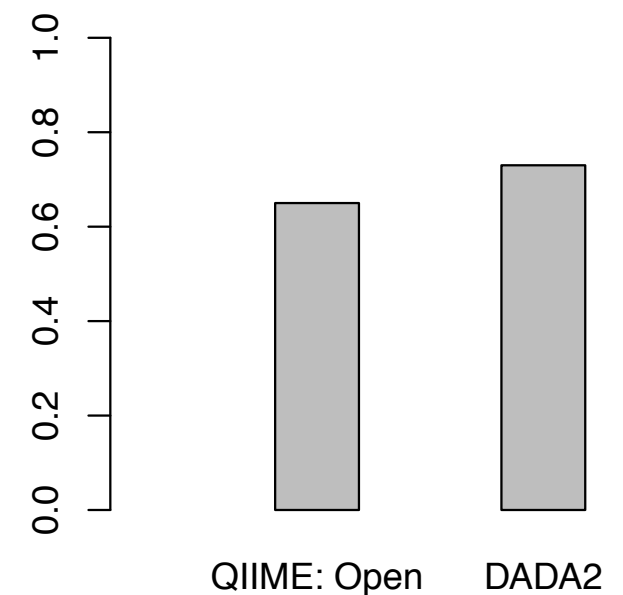
**Unpublished  
ordination  
removed.**

## Variance explained by

PC1 + PC2

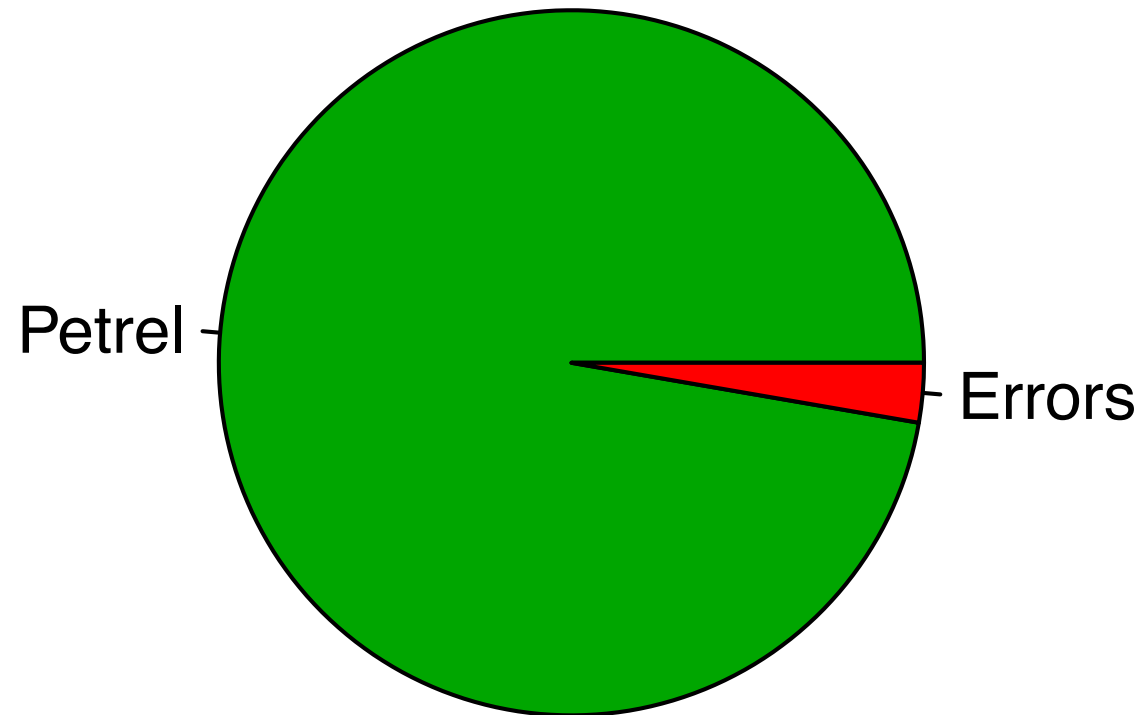


**Arsenic**

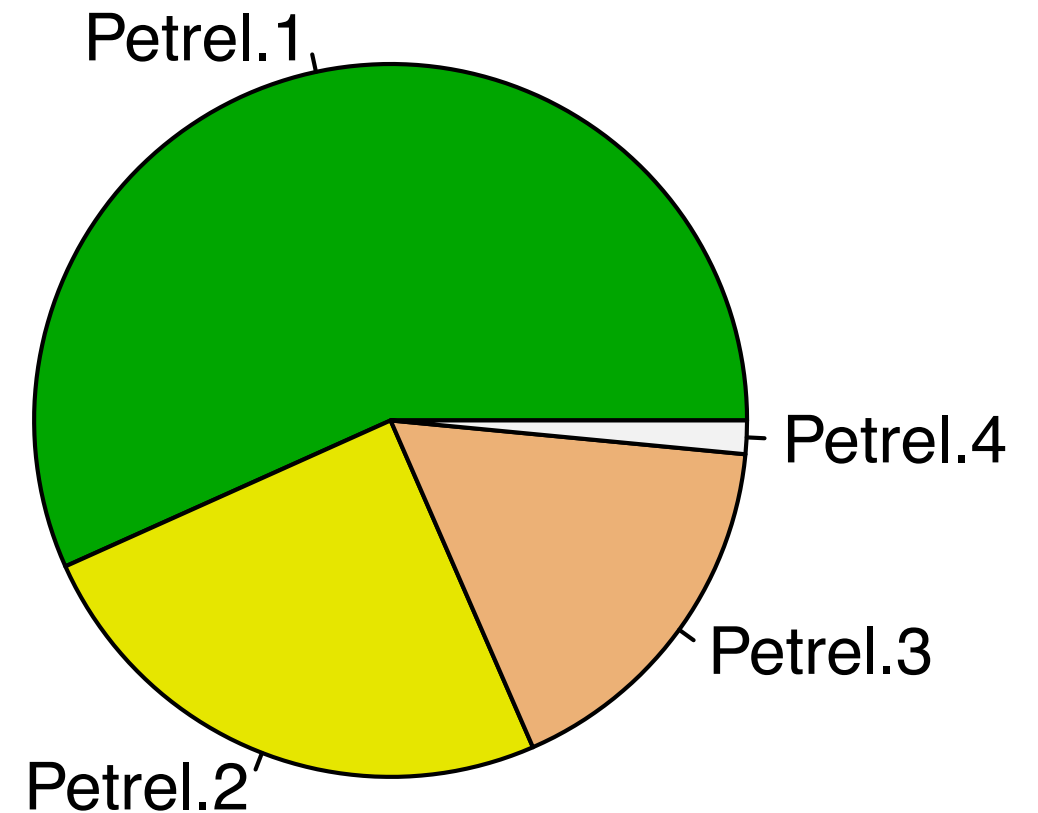


# Resolution: Petrel aDNA

**QIIME: De novo**

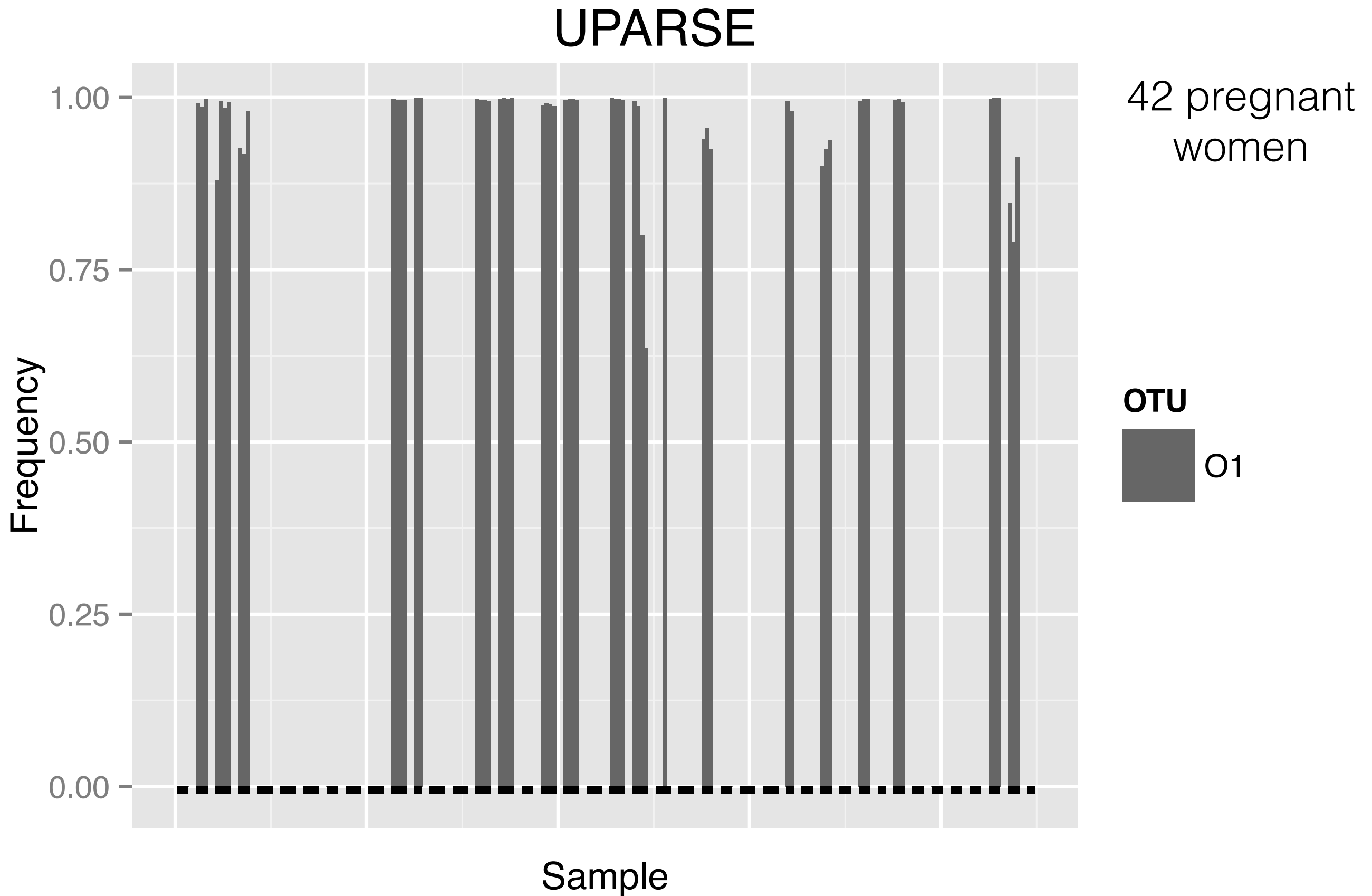


**DADA2**



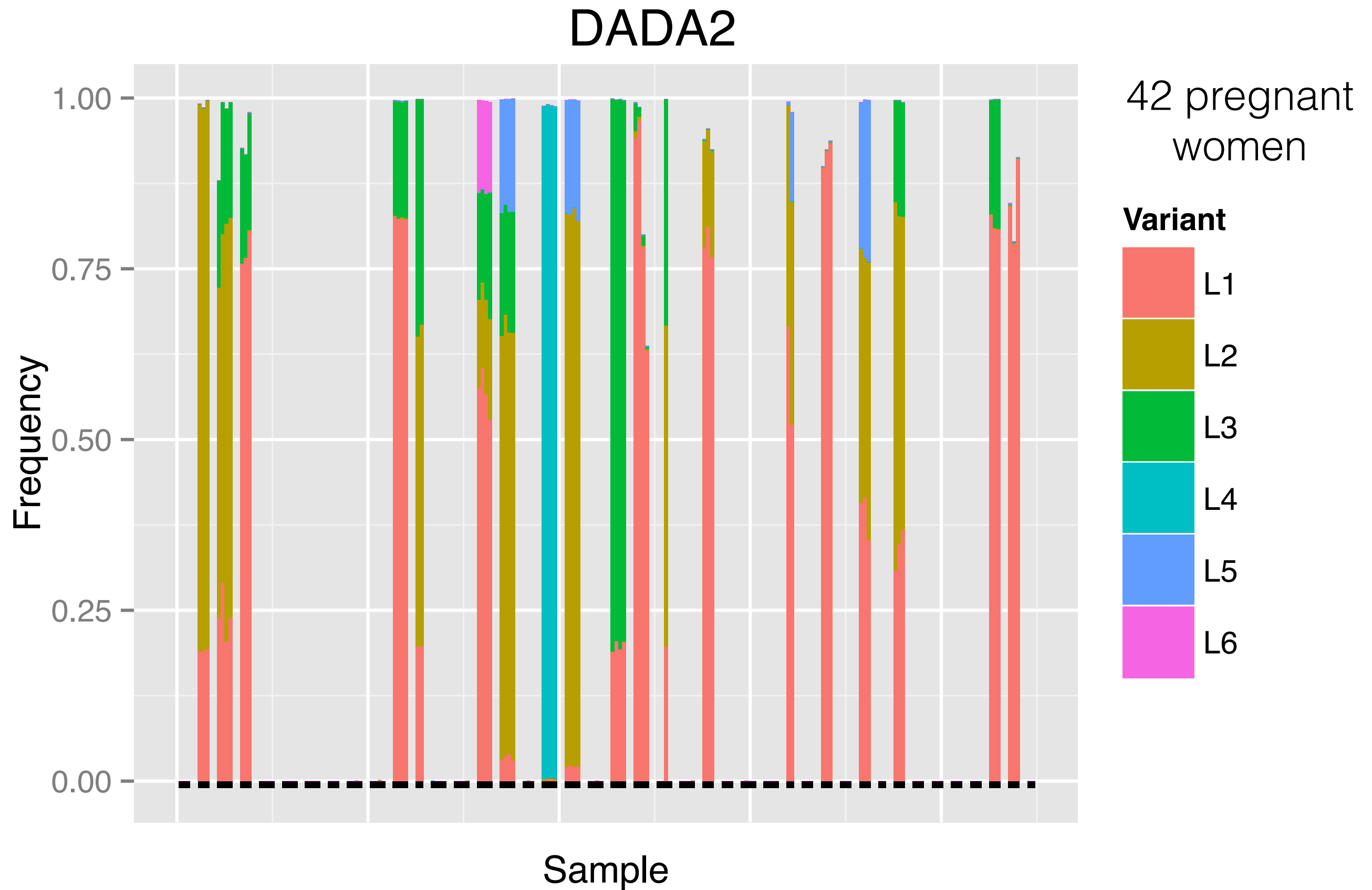


# Resolution: *L. crispatus*



**Data:** MacIntyre et al. Scientific Reports, 2015.

# Resolution: *L. crispatus*



**Data:** MacIntyre et al. Scientific Reports, 2015.

# Acknowledgements

---



Susan Holmes



Joey McMurdie



Michael Rosen



National Institutes of Health

**<https://benjjneb.github.io/dada2/>**