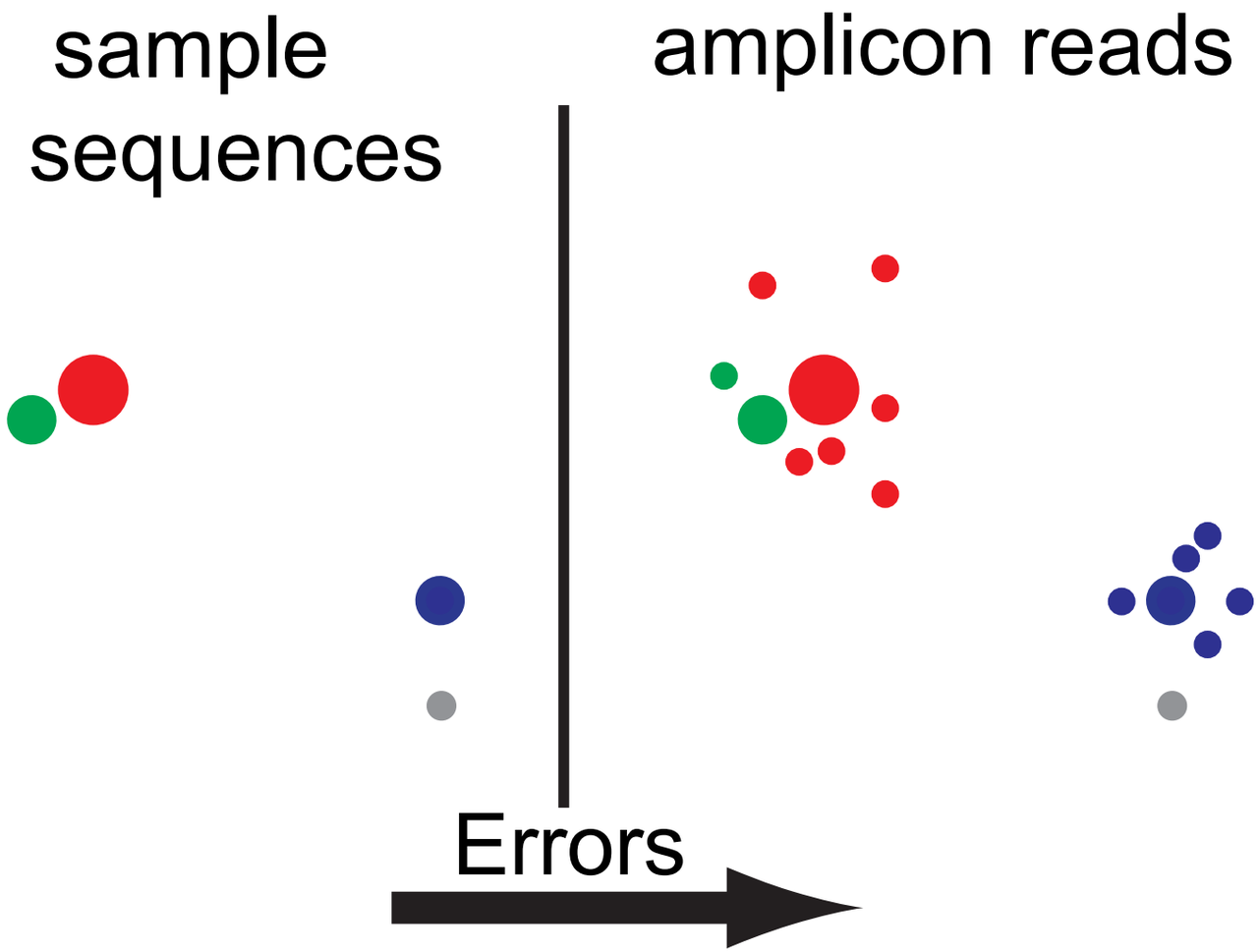
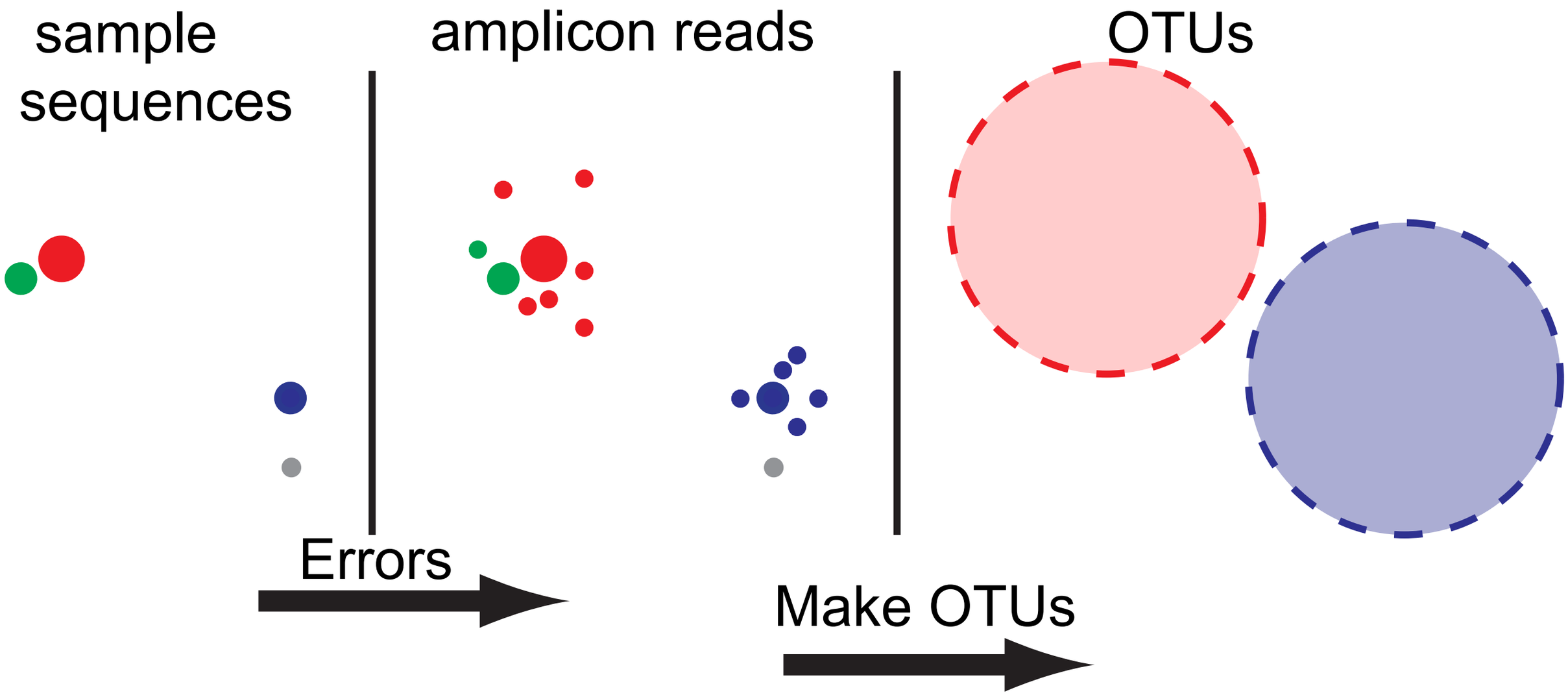
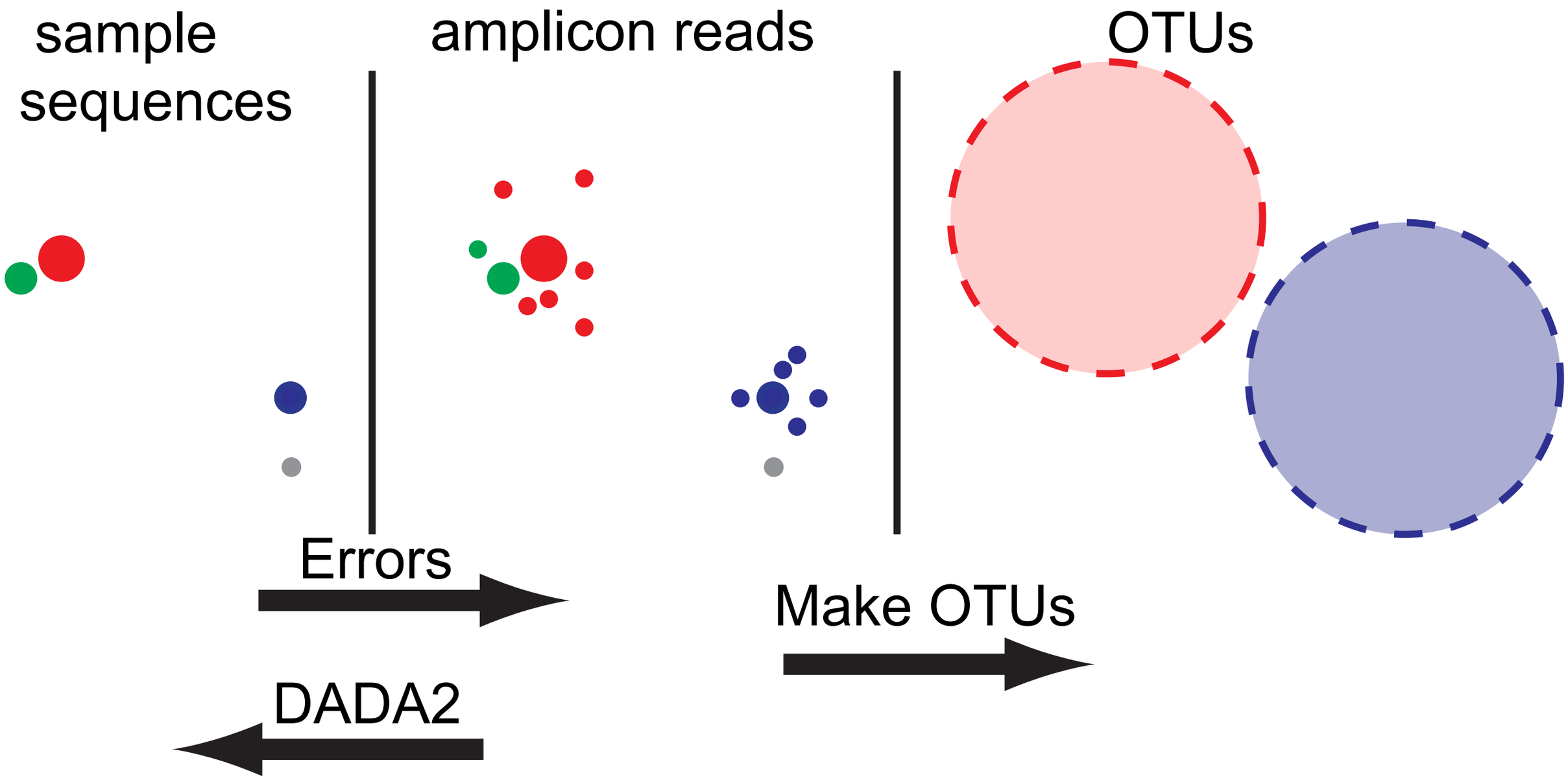
The DADA2 Method

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



Callahan, et al. Nature Methods, 2016.


DA2
Amplicon Sequencing. Exactly. Version 1.10

Error Model

An Error Model

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

An Error Model

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

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

An Error Model

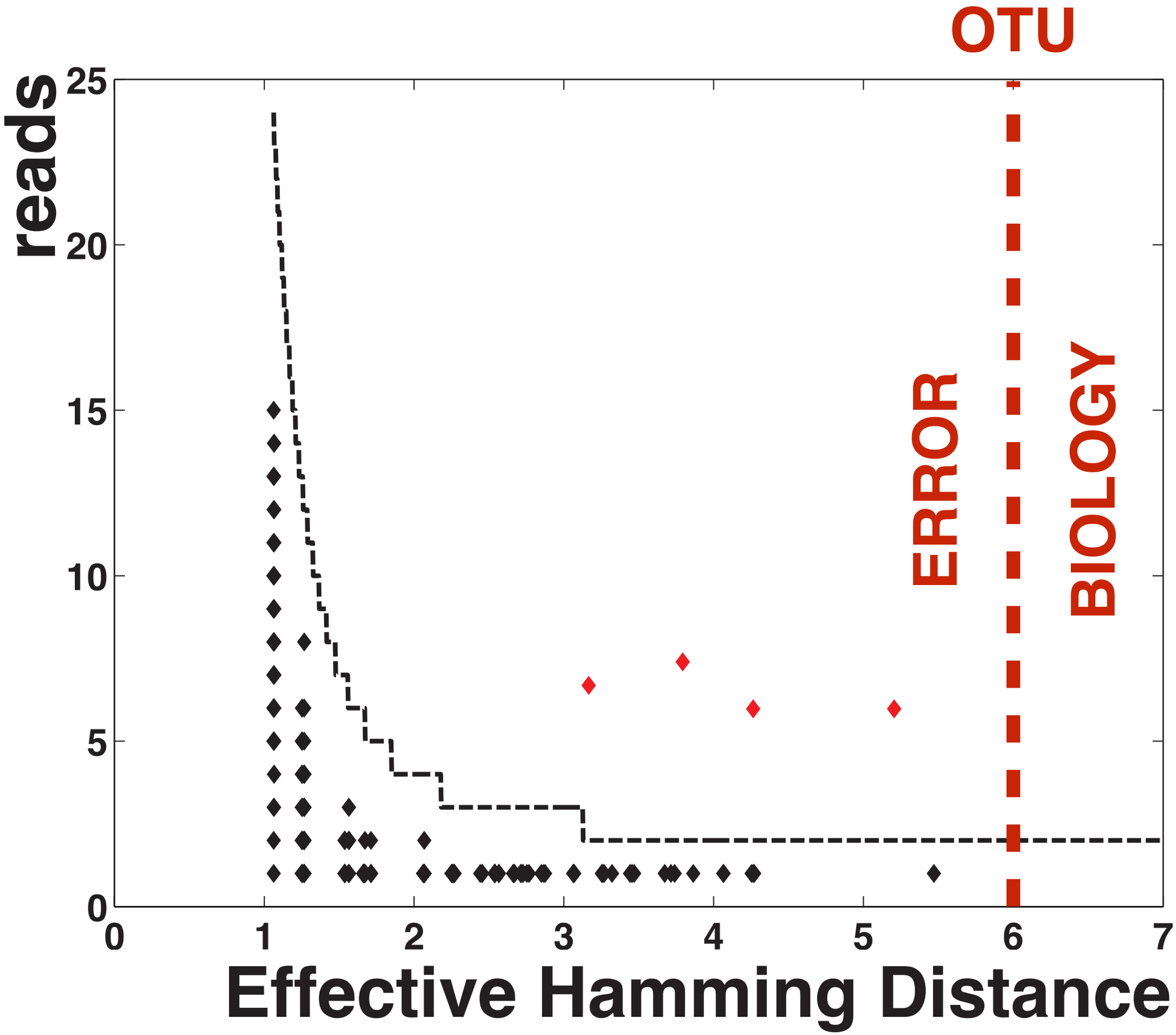
```
s: ATTAACGAGATTATAACCAGAGTACGAATA...
     |                     |
r: ATCAACGAGATTATAACAAGAGTACGAATA...
```

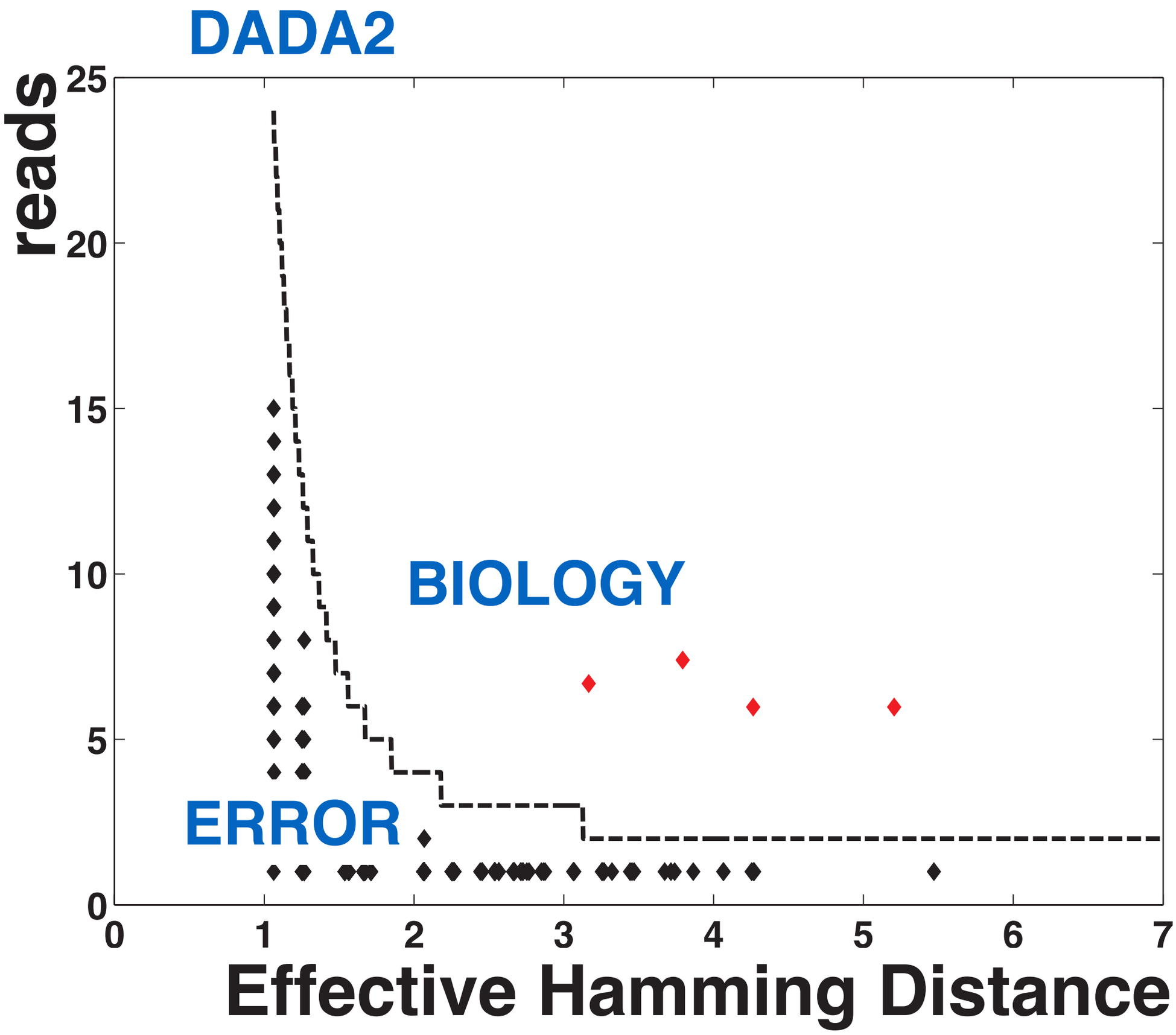
$$p(r|s) = \prod_{i=1}^{L} p(r(i)|s(i), q_r(i), Z)$$

Error process is independent across nucleotides.

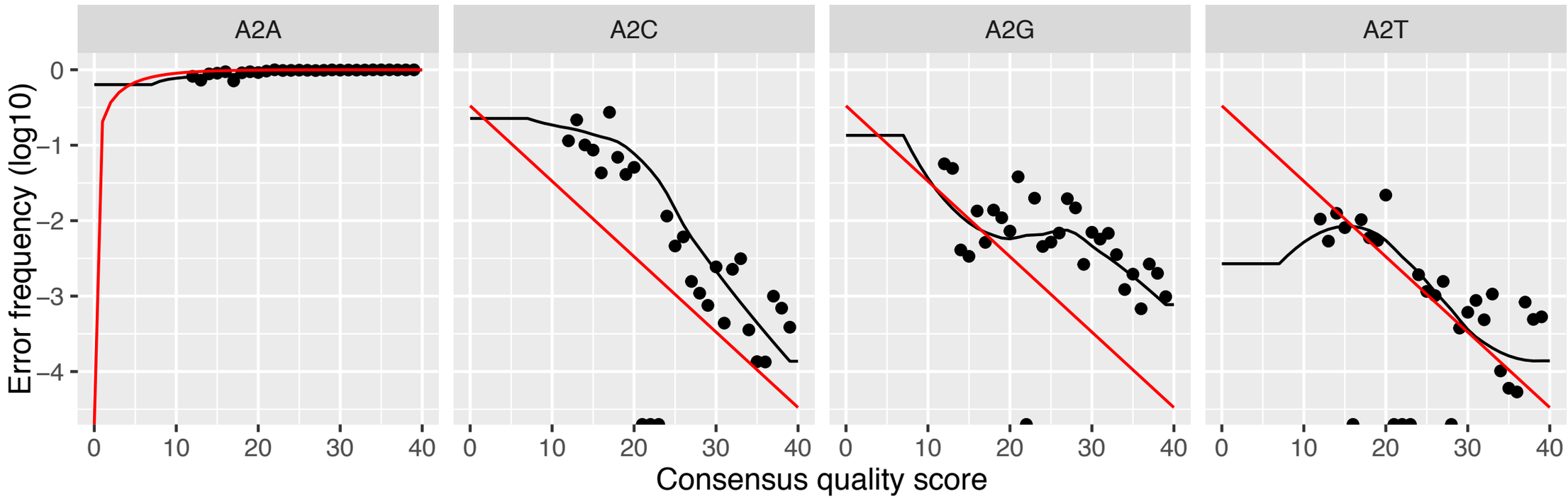
Per-nucleotide transition rate depends on:

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

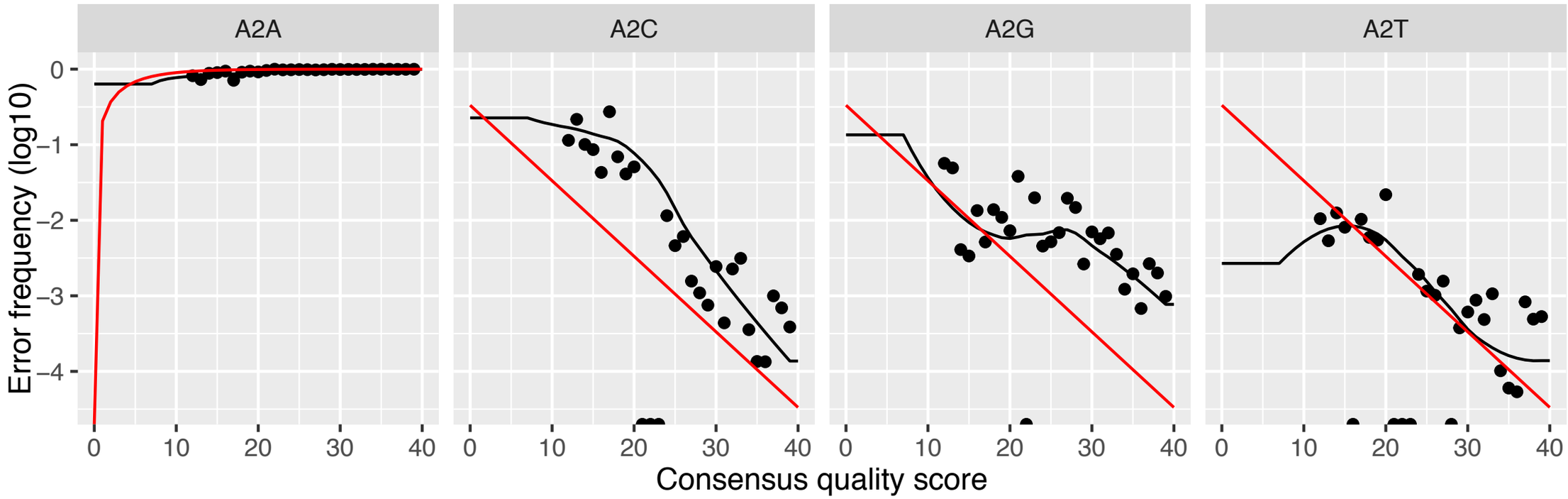

OTU
ERROR
BIOLOGY
reads
25
20
15
10
5
0
0
1
2
3
4
5
6
7
Effective Hamming Distance


DADA2
reads
25
20
15
10
5
0
BIOLOGY
ERROR
0
1
2
3
4
5
6
7
Effective Hamming Distance

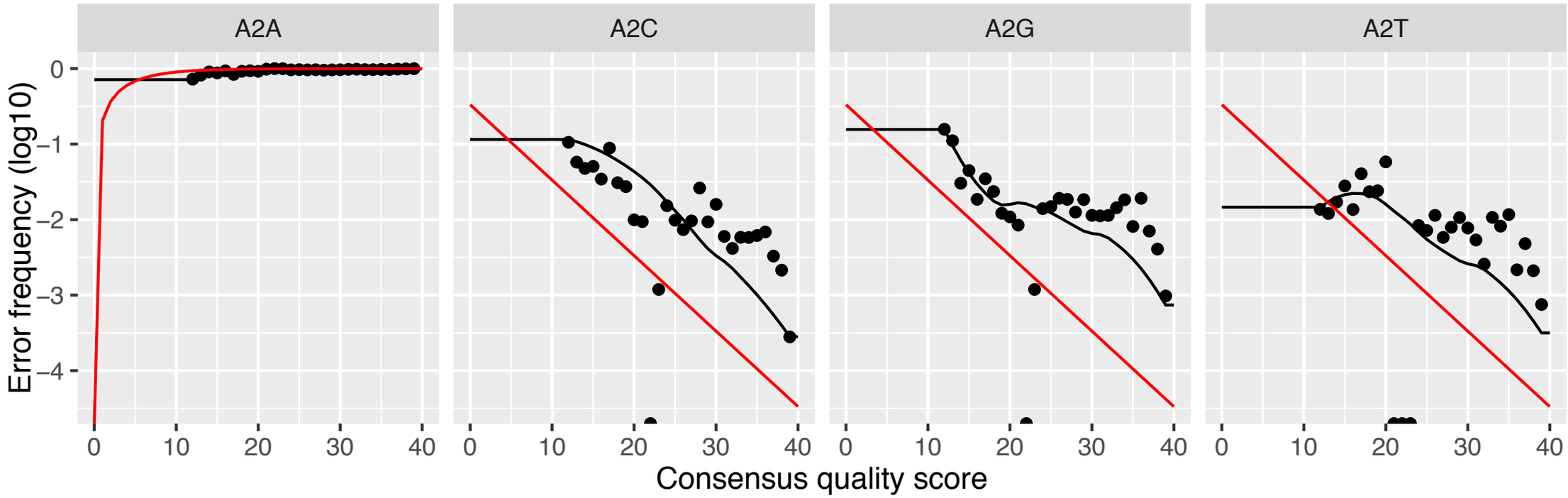
Learning Errors


A2A
A2C
A2G
A2T
Error frequency (log10)
0
−1
−2
−3
−4
0
10
20
30
40
Consensus quality score

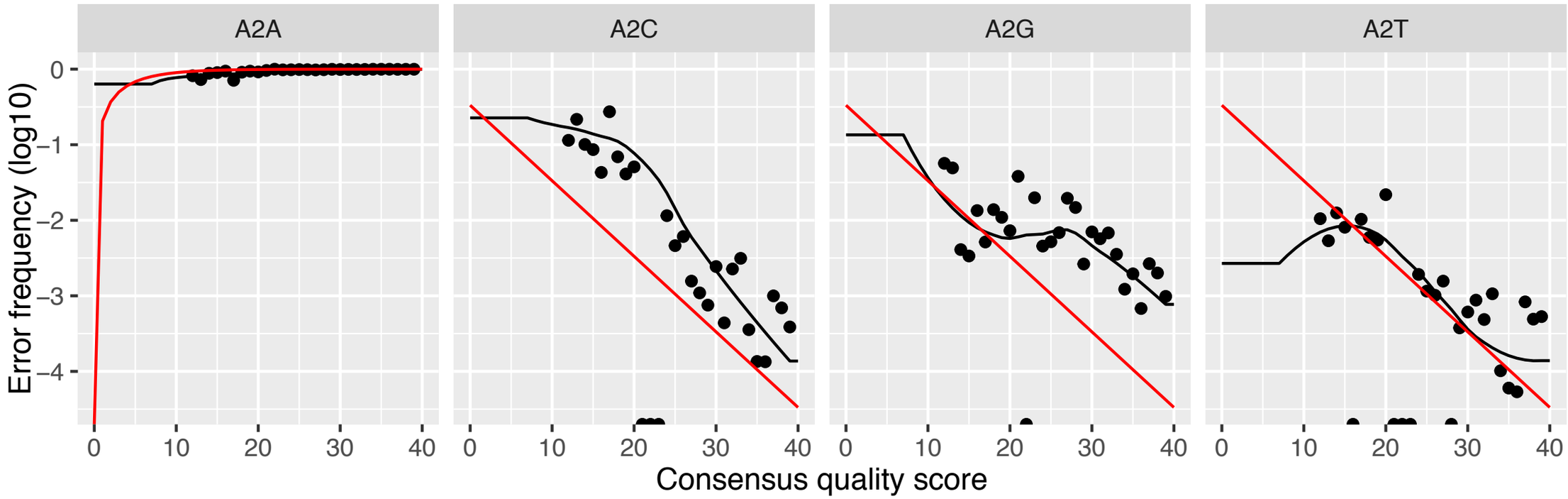
Study A



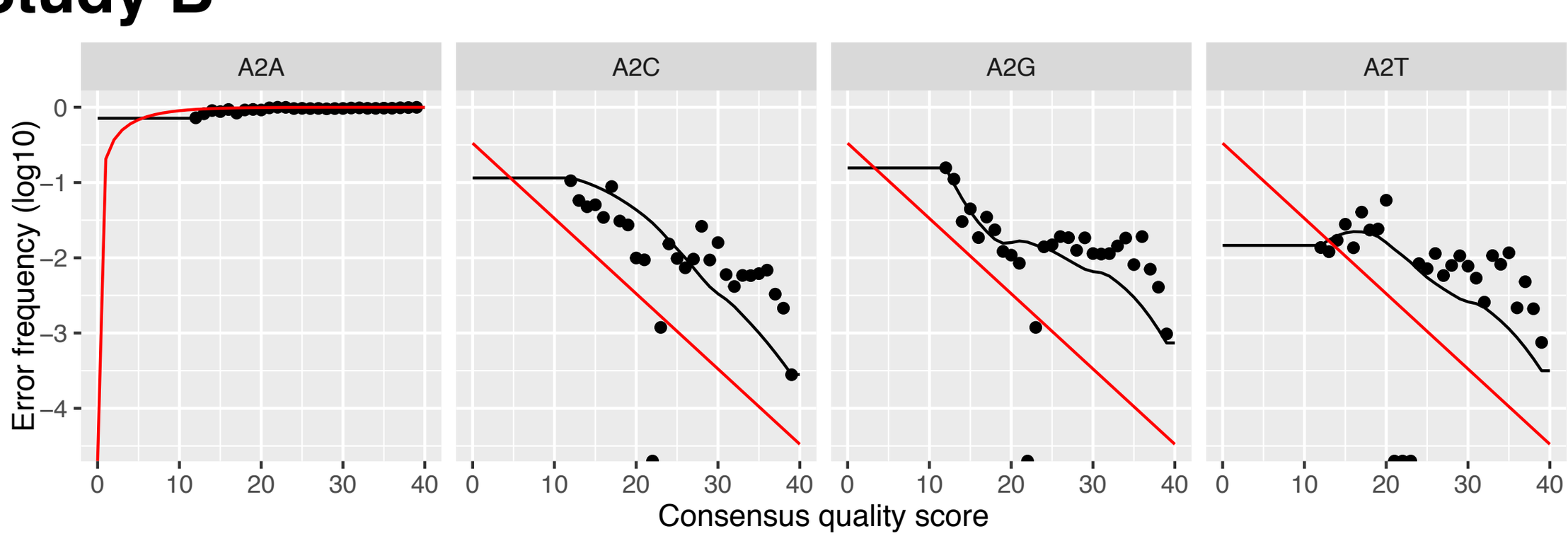
Study B



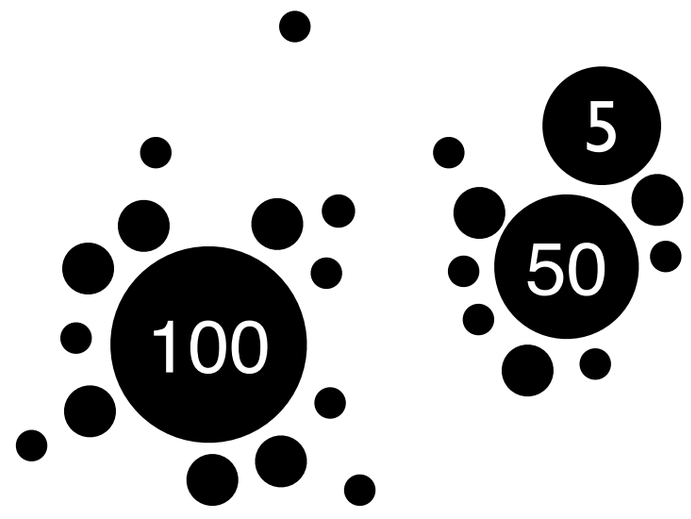
Study A



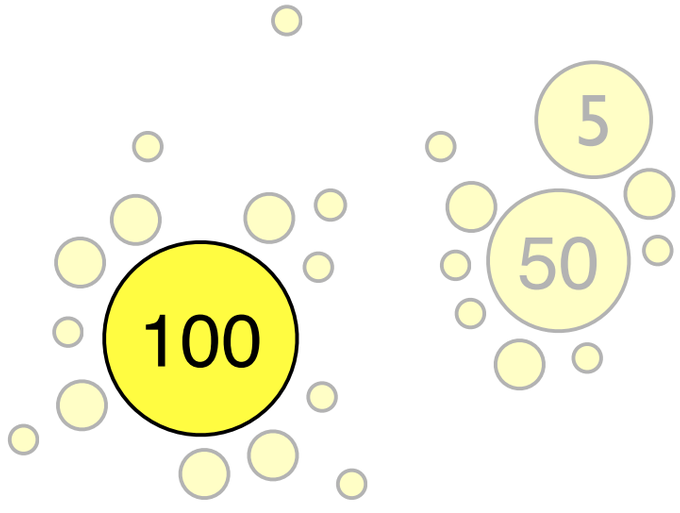
Study B



But How?



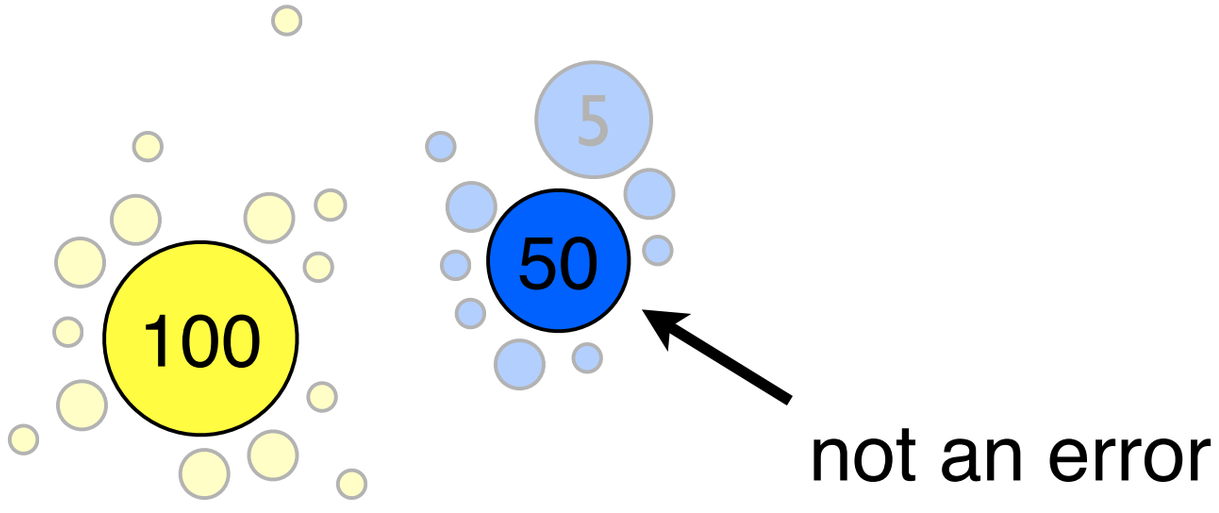
Initial guess: one real sequence + errors



Infer initial *error model* under this assumption.

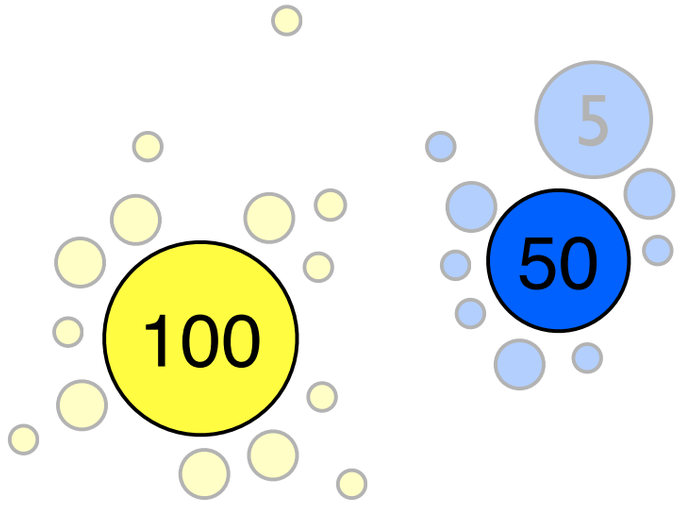
$\Pr(i \rightarrow j) =$

	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



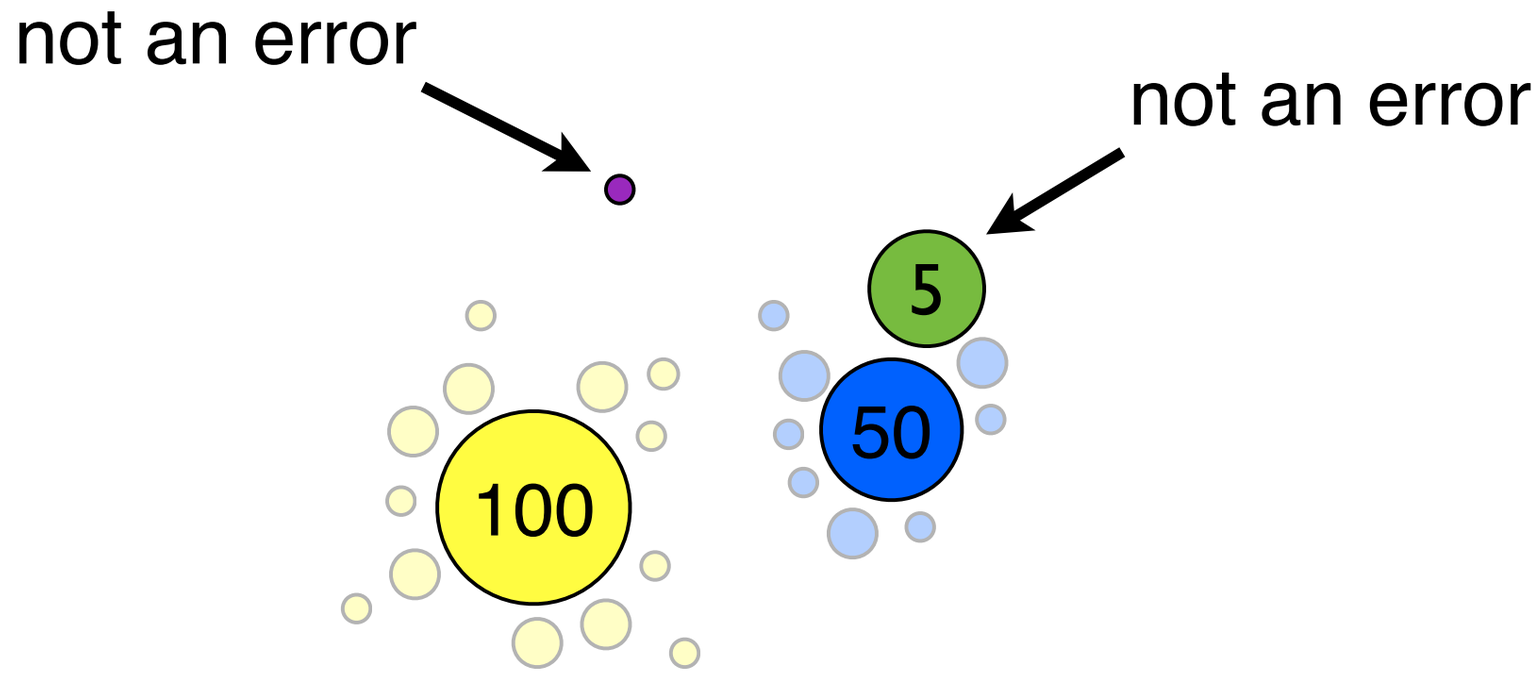
Reject unlikely error under model. **Recruit** errors.

	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



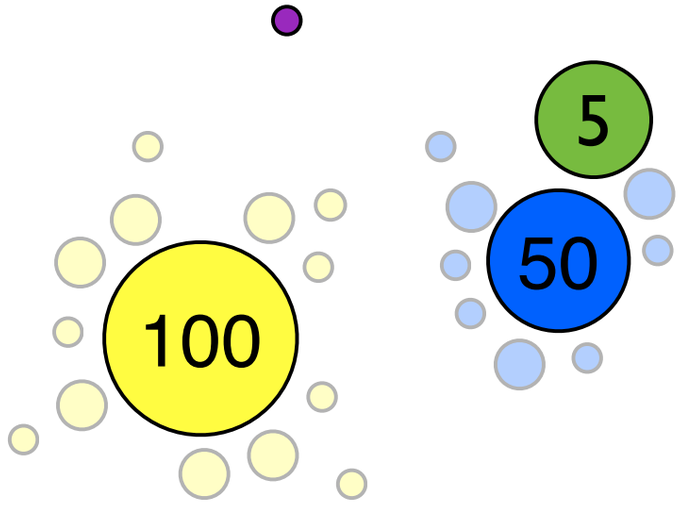
Update the model.

	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



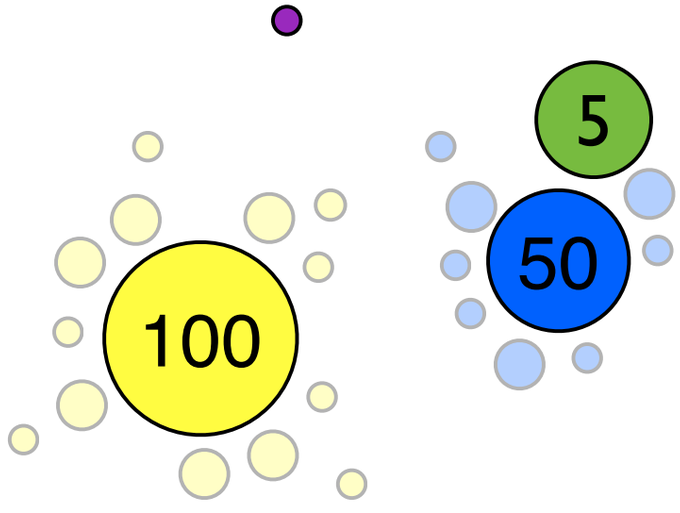
Reject more sequences under ***new*** model

	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	$1x10^{-4}$	$2x10^{-3}$	$2x10^{-4}$
C	$6x10^{-5}$	0.999	$3x10^{-6}$	$1x10^{-3}$
G	$1x10^{-3}$	$3x10^{-6}$	0.999	$6x10^{-5}$
T	$2x10^{-4}$	$2x10^{-3}$	$1x10^{-4}$	0.998



Convergence: all errors are plausible

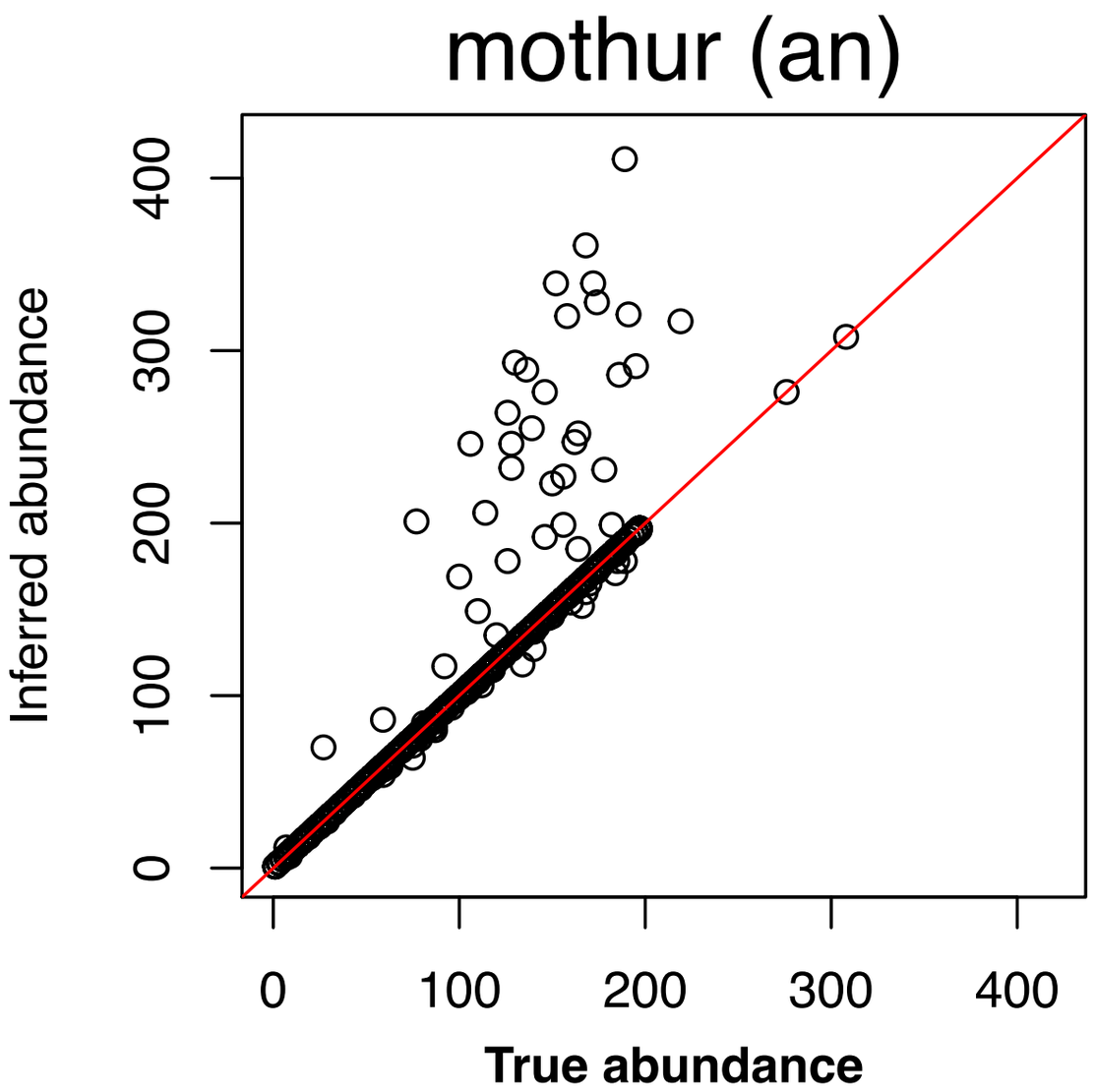
	A	**C**	**G**	**T**
A	0.998	$1x10^{-4}$	$2x10^{-3}$	$2x10^{-4}$
C	$6x10^{-5}$	0.999	$3x10^{-6}$	$1x10^{-3}$
G	$1x10^{-3}$	$3x10^{-6}$	0.999	$6x10^{-5}$
T	$2x10^{-4}$	$2x10^{-3}$	$1x10^{-4}$	0.998

Accuracy

and

Resolution

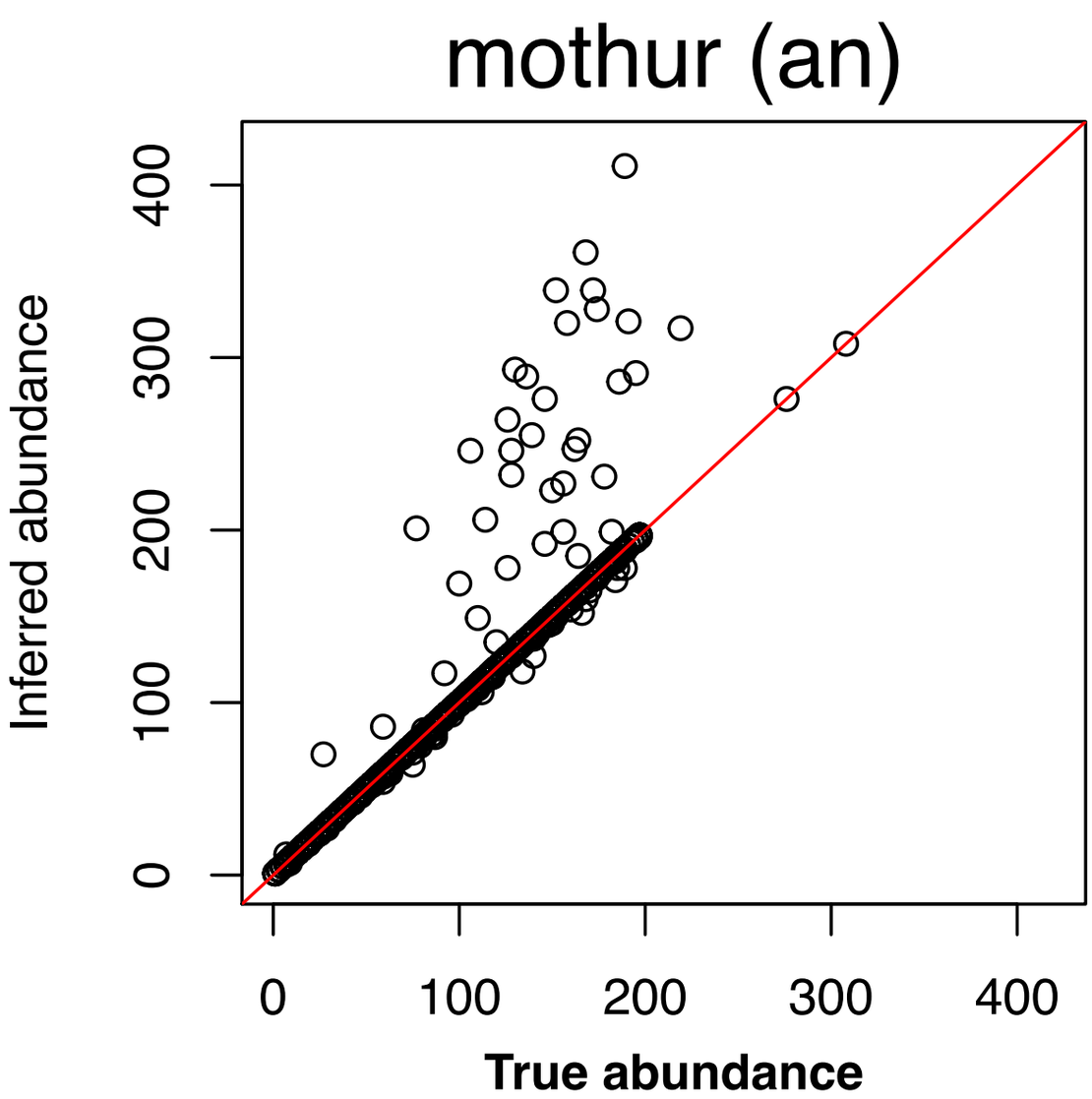
Accuracy: Simulated data



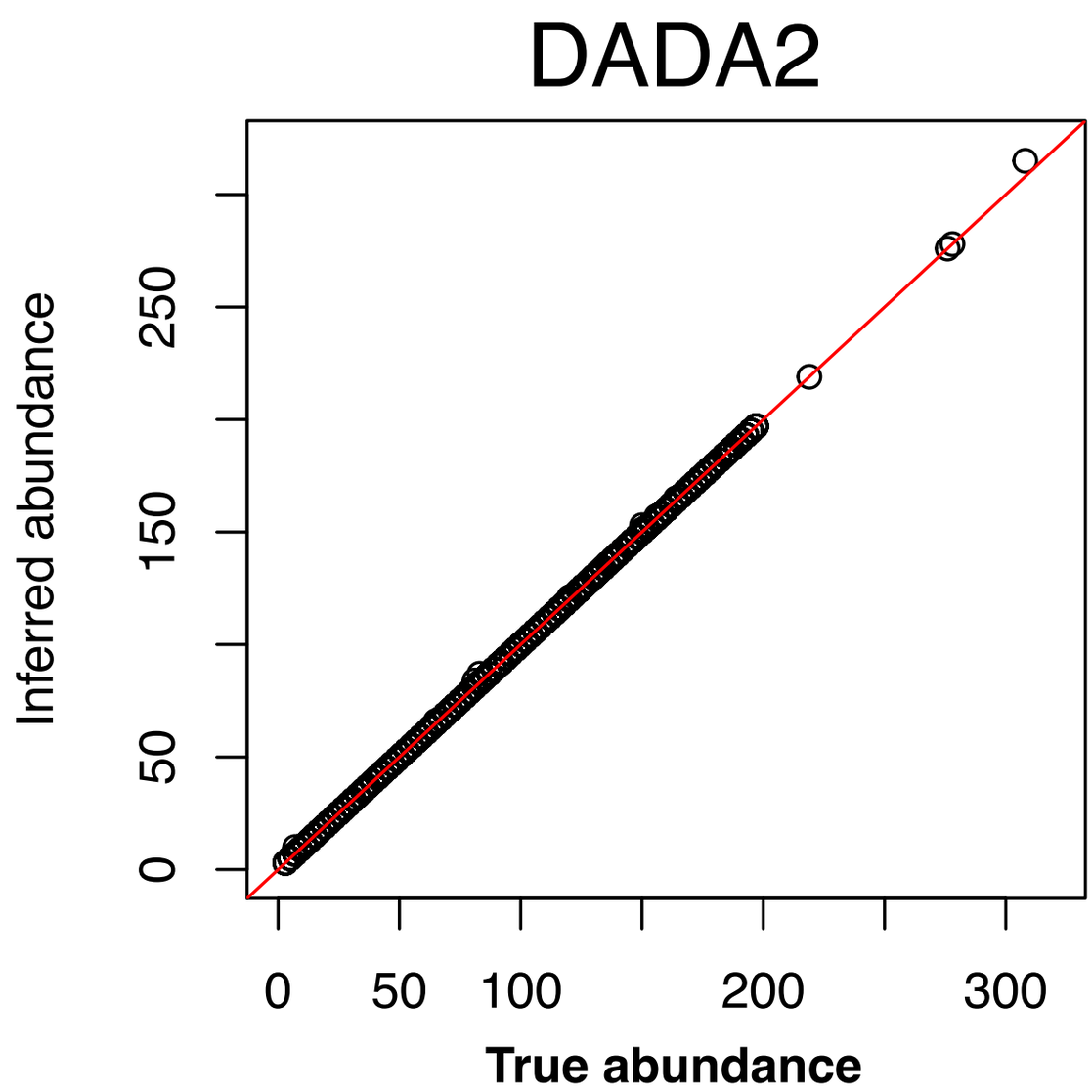
TP: 978
FP: 272
FN: 77
cor: 0.935

Data: Kopylova, et al. mSystems, 2016.

Accuracy: Simulated data



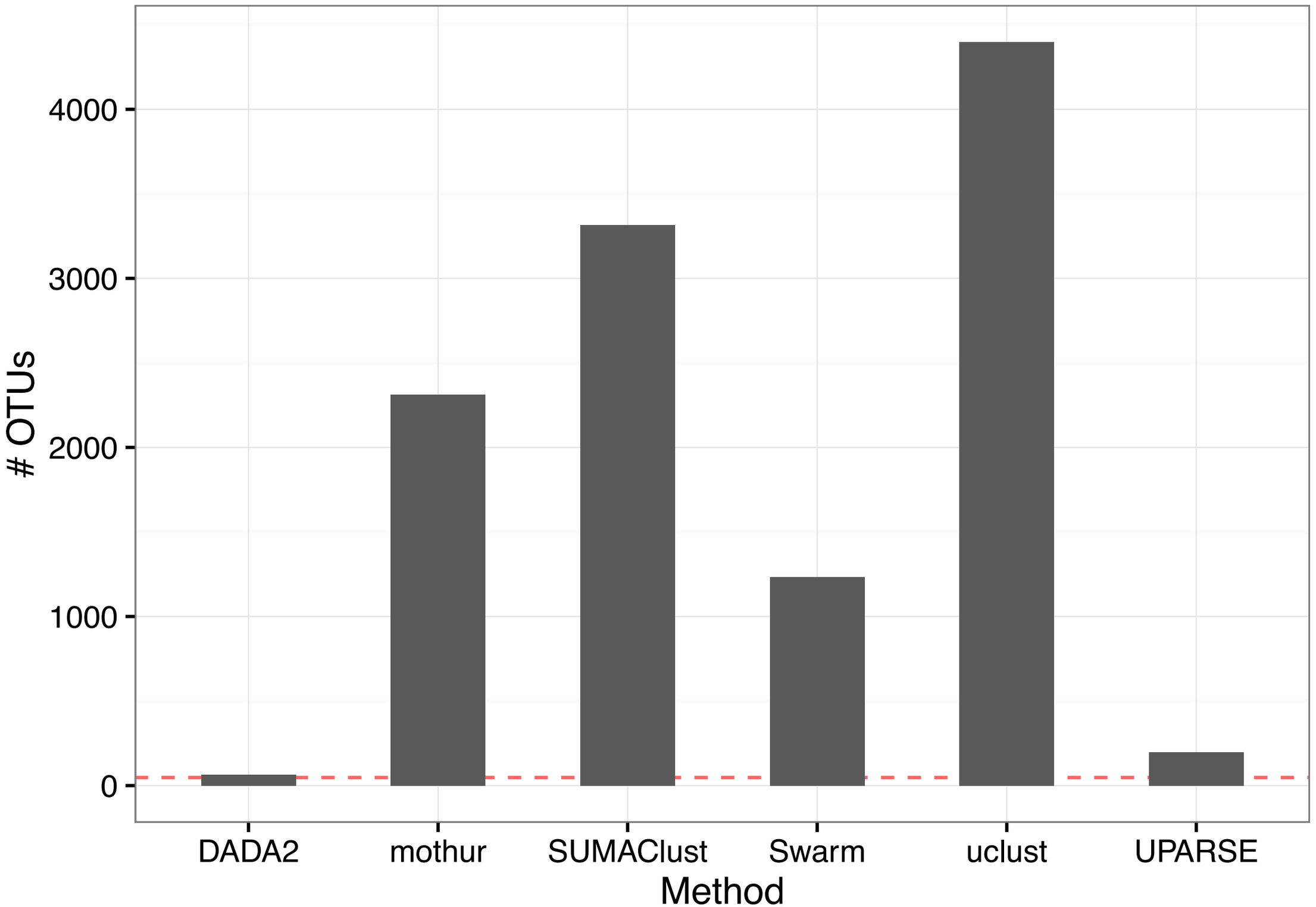
TP: 978
FP: 272
FN: 77
cor: 0.935



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

Data: Kopylova, et al. mSystems, 2016.

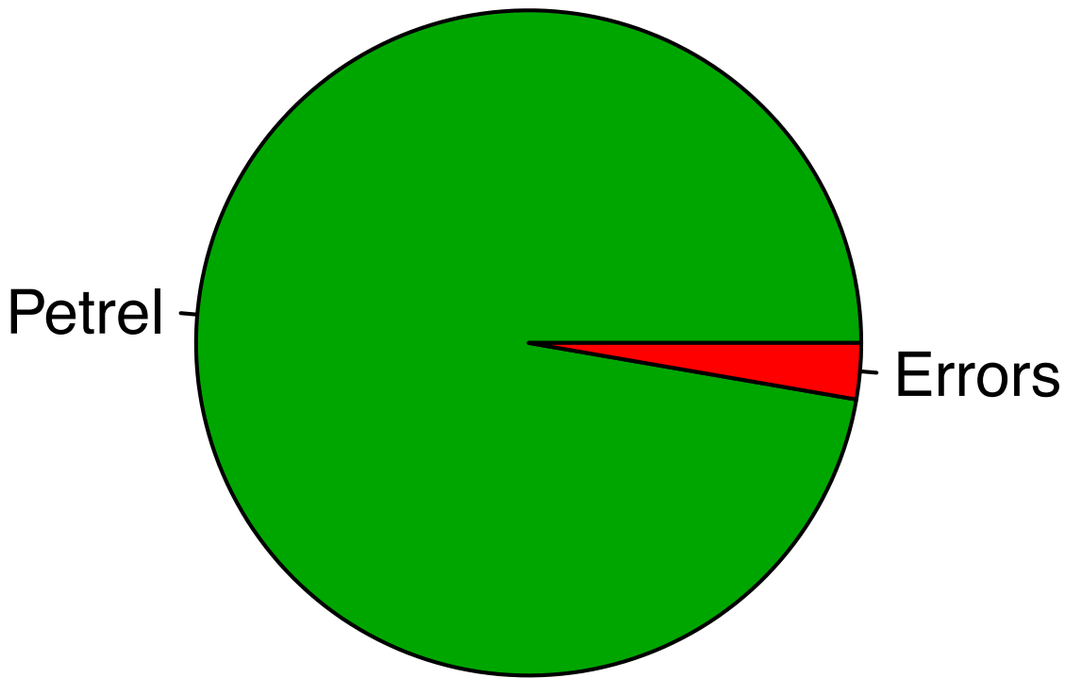
Accuracy: Mock community



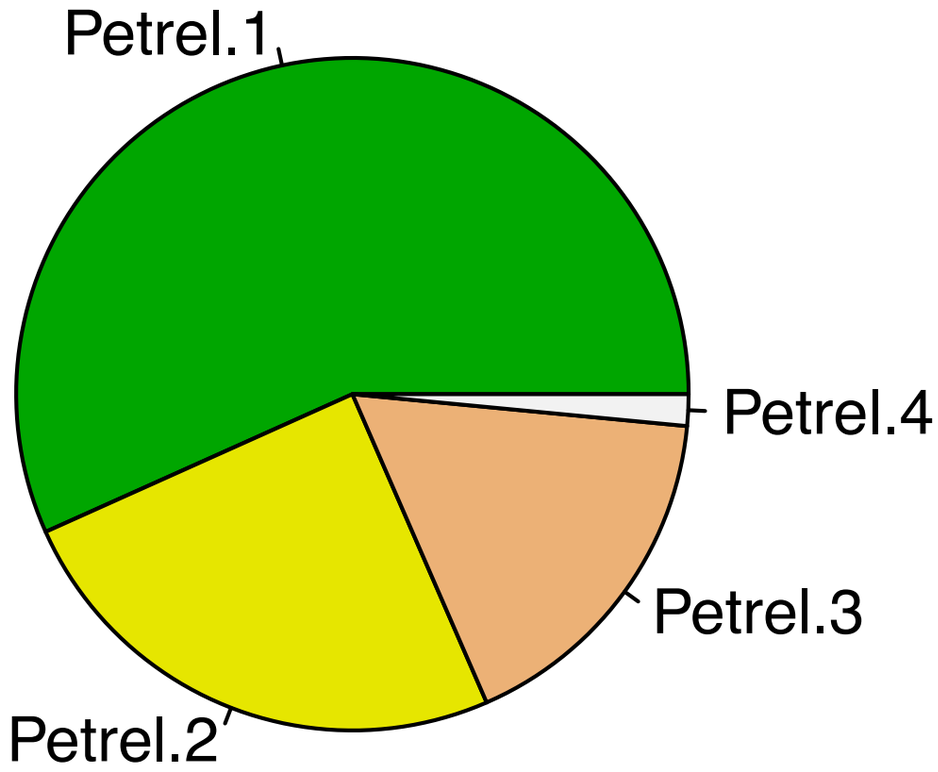
Credit: Kopylova, et al. mSystems, 2016.

Resolution: Petrel aDNA

QIIME: De novo



DADA2

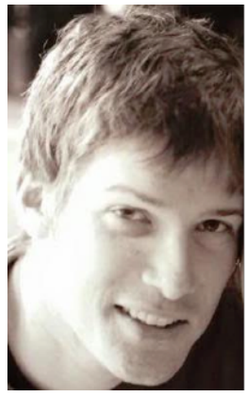


Credit: Kealoha Kinney, Michael Bunce, Andreanna Welch

Acknowledgements

Susan Holmes

Joey McMurdie



Michael Rosen

https://benjjneb.github.io/dada2/