LECTURE III: DENOISING & CLUSTERING

ALEXANDER EILER
DEPARTMENT OF BIOSCIENCES - AQUA
alexander.eiler@ibv.uio.no
https://www.mn.uio.no/ibv/english/people/aca/alexaei/

ILLUMINA AMPLICON READS

- Illumina MiSeq sequencing is the current state of the art technology for amplicon sequencing because of its high yield relative to cost.
- The error rate (~1%) of Illumina sequencing is so high that the chance of capturing true reads is small (about 1% for a 450 base fragment, not adjusted for improvements due to overlapping reads).
- This problem is not new, and has been discussed ever since high-throughput sequencing was first applied to taxonomical markers.

PROPOSED SOLUTIONS

- For many years now, the common practice was to solve this by UPGMA- style clustering at a fixed sequence distance (97% similarity).
- Sequences were clustered into operational taxonomic units based on sequence similarity or dissimilarity cutoffs.

UPMG

Single linkage clustering.
Advantages:

- reduces the impact of clustering parameters on the resulting OTUs by avoiding arbitrary global clustering thresholds and input sequence ordering dependence.

UPARSE

UPARSE implements a greedy algorithm that performs OTU clustering

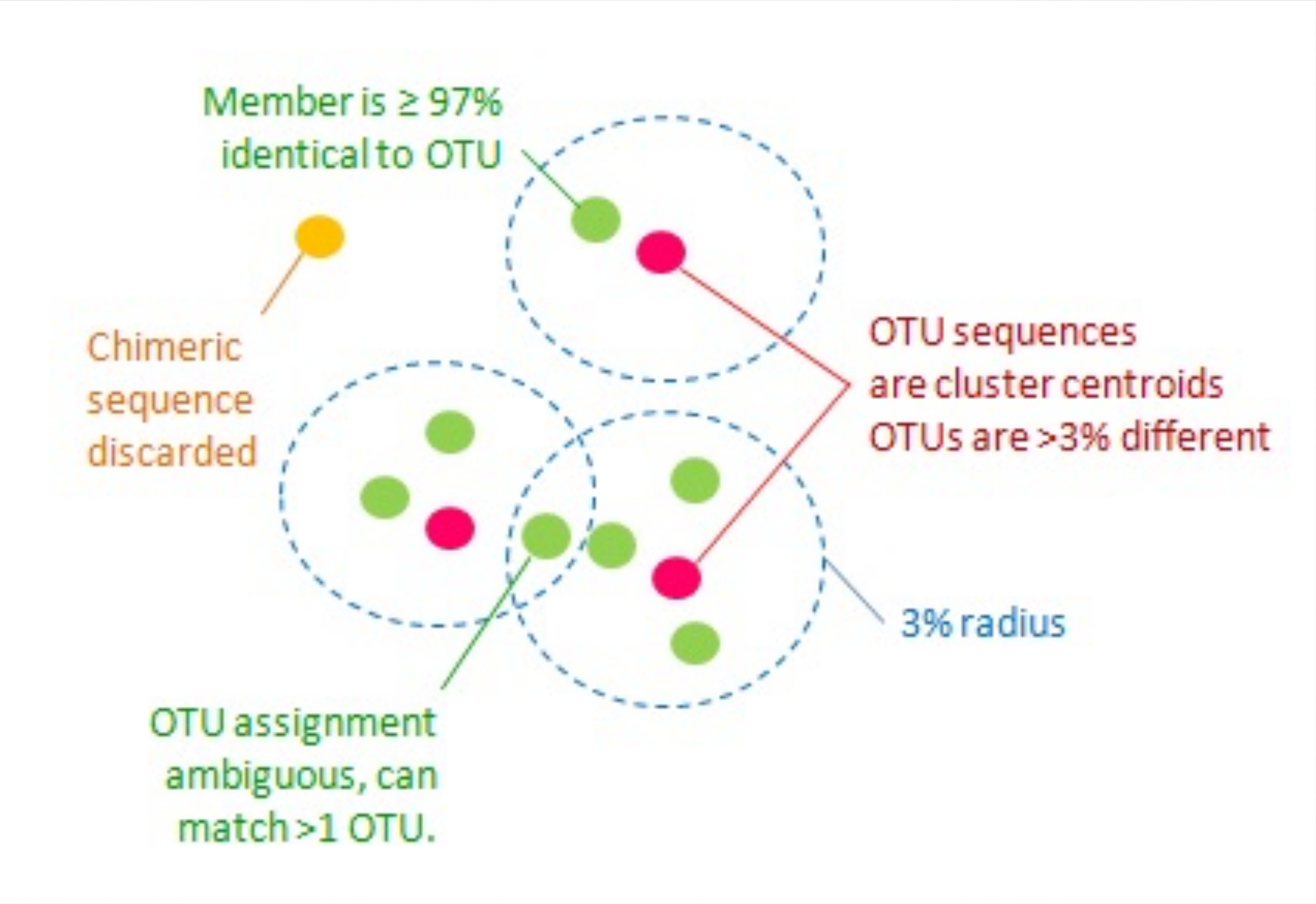
Advantages:

- Fast and greedy.

Disadvantages:

- Greedy clustering methods suffer from two fundamental problems.
  1. They use an arbitrary fixed global clustering threshold. As lineages evolve at variable rates, no single cut-off value can accommodate the entire tree of life.
  2. The input order of amplicons strongly influences the clustering results. Previous centroid selections are not re-evaluated as clustering progresses, which can generate inaccurately formed OTUs

UPARSE



https://drive5.com/usearch/manual/uparseotu_algo.html

SWARM

Swarm is a de novo clustering algorithm based on an unsupervised single-linkage-clustering method.

Advantages:

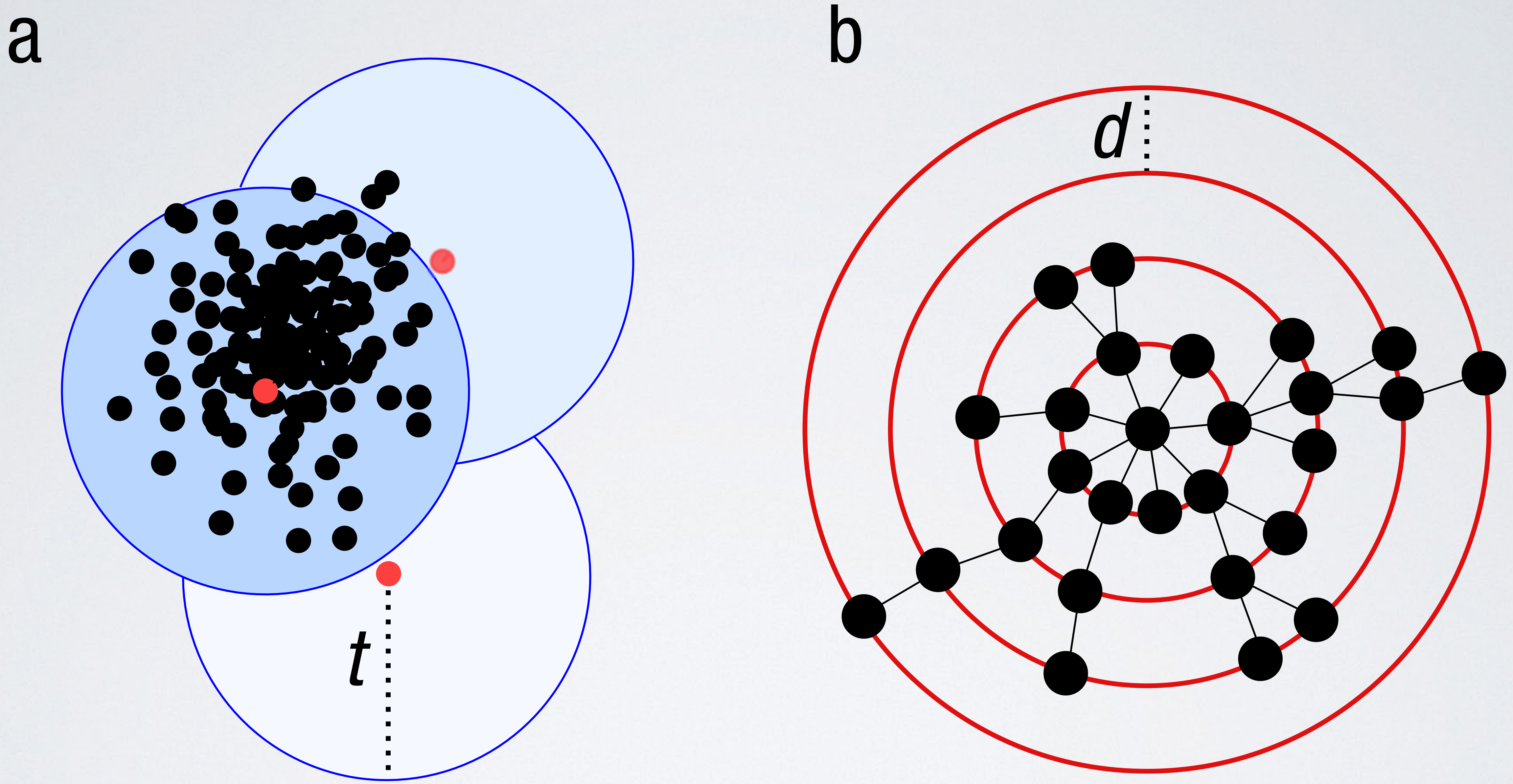
- reduces the impact of clustering parameters on the resulting OTUs by avoiding arbitrary global clustering thresholds and input sequence ordering dependence.

SWARM

Swarm builds OTUs in two steps:

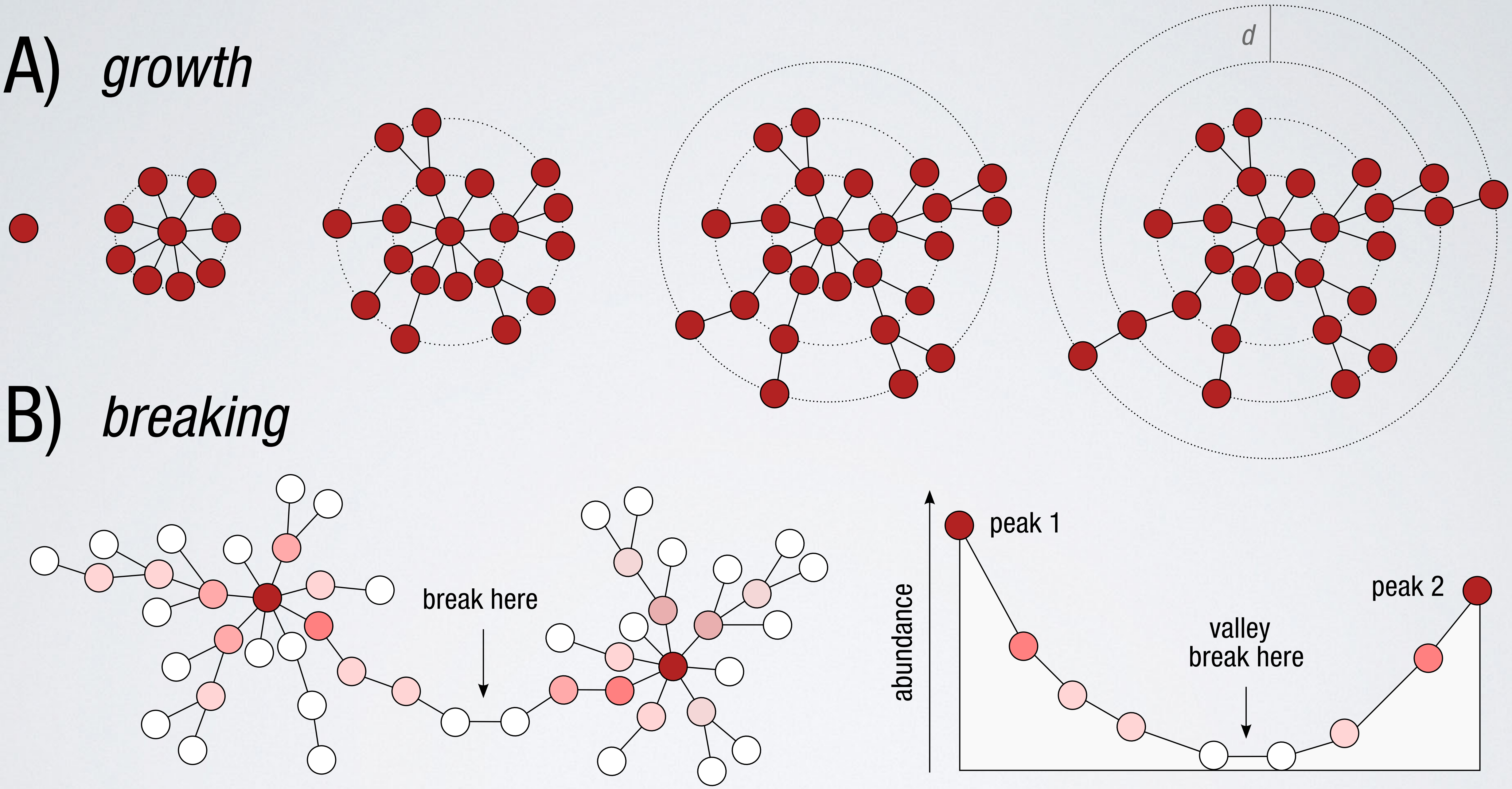
1. an initial set of OTUs is constructed by iteratively agglomerating similar amplicons
2. amplicon abundance values are used to reveal OTUs' internal structures and to break them into sub-OTUs, if necessary.

SWARM



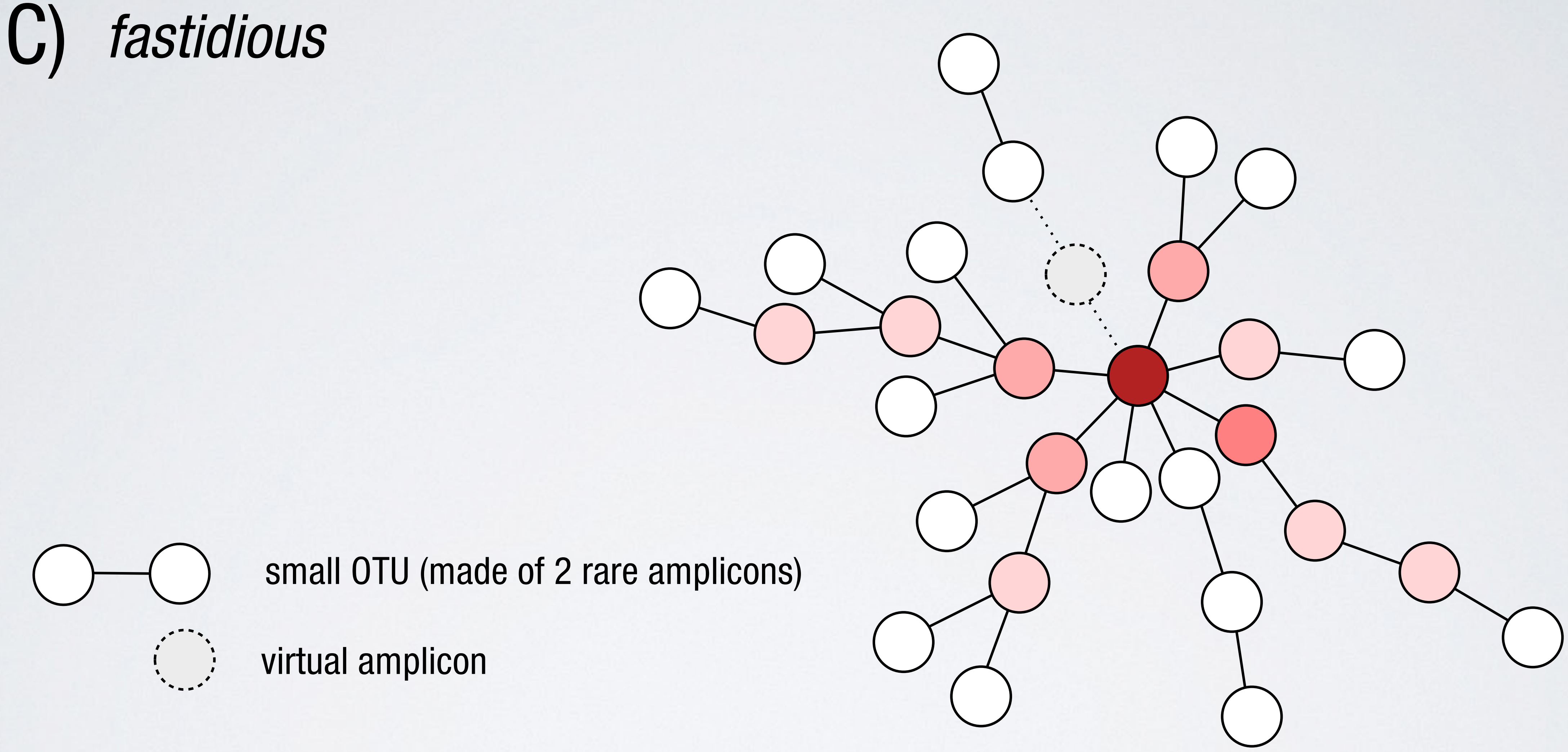
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4178461/pdf/peerj-02-593.pdf

SWARM



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4690345/pdf/peerj-03-1420.pdf

SWARM

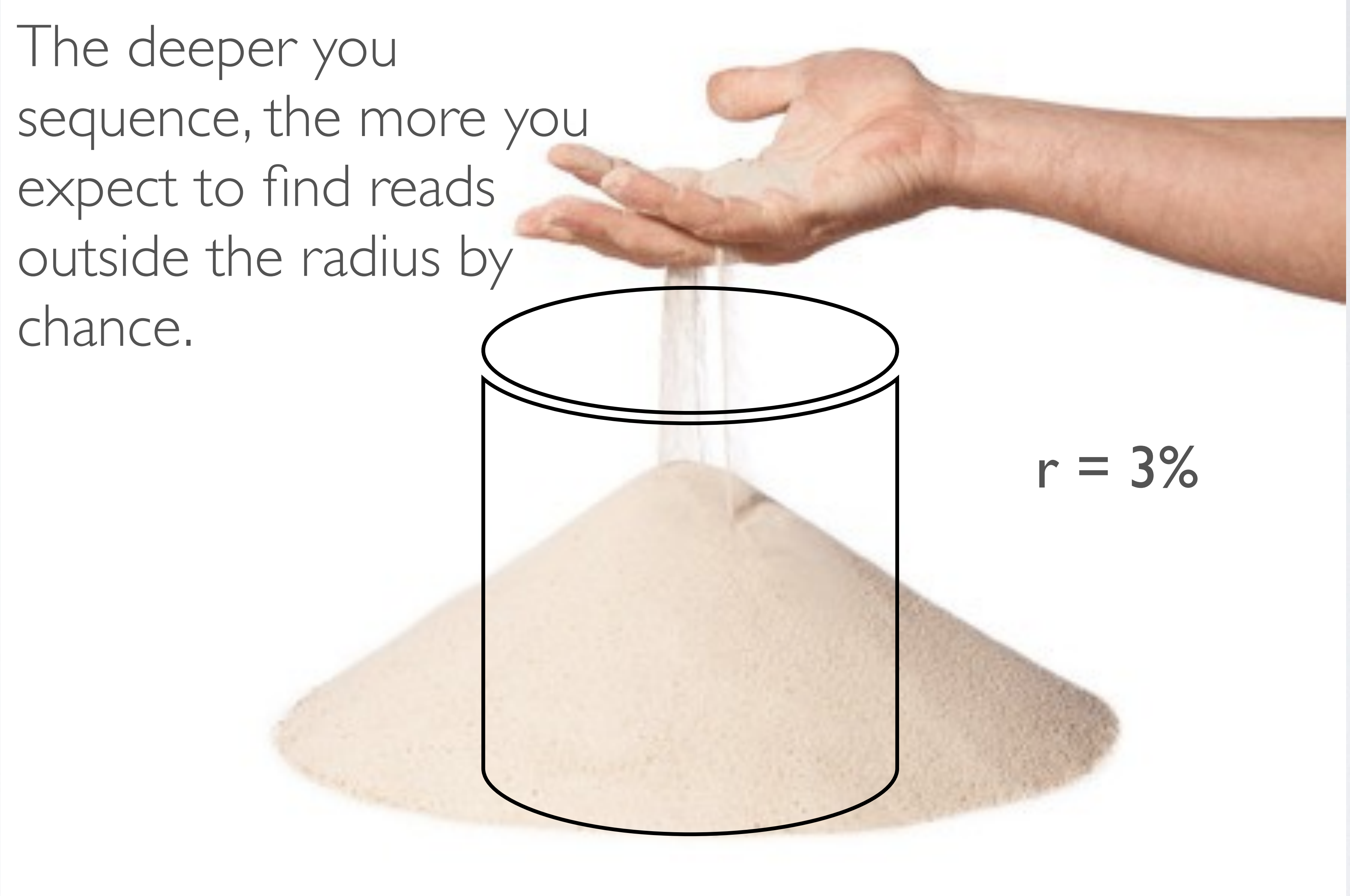


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4690345/pdf/peerj-03-1420.pdf

LINGERING PROBLEMS WITH "OTU"

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

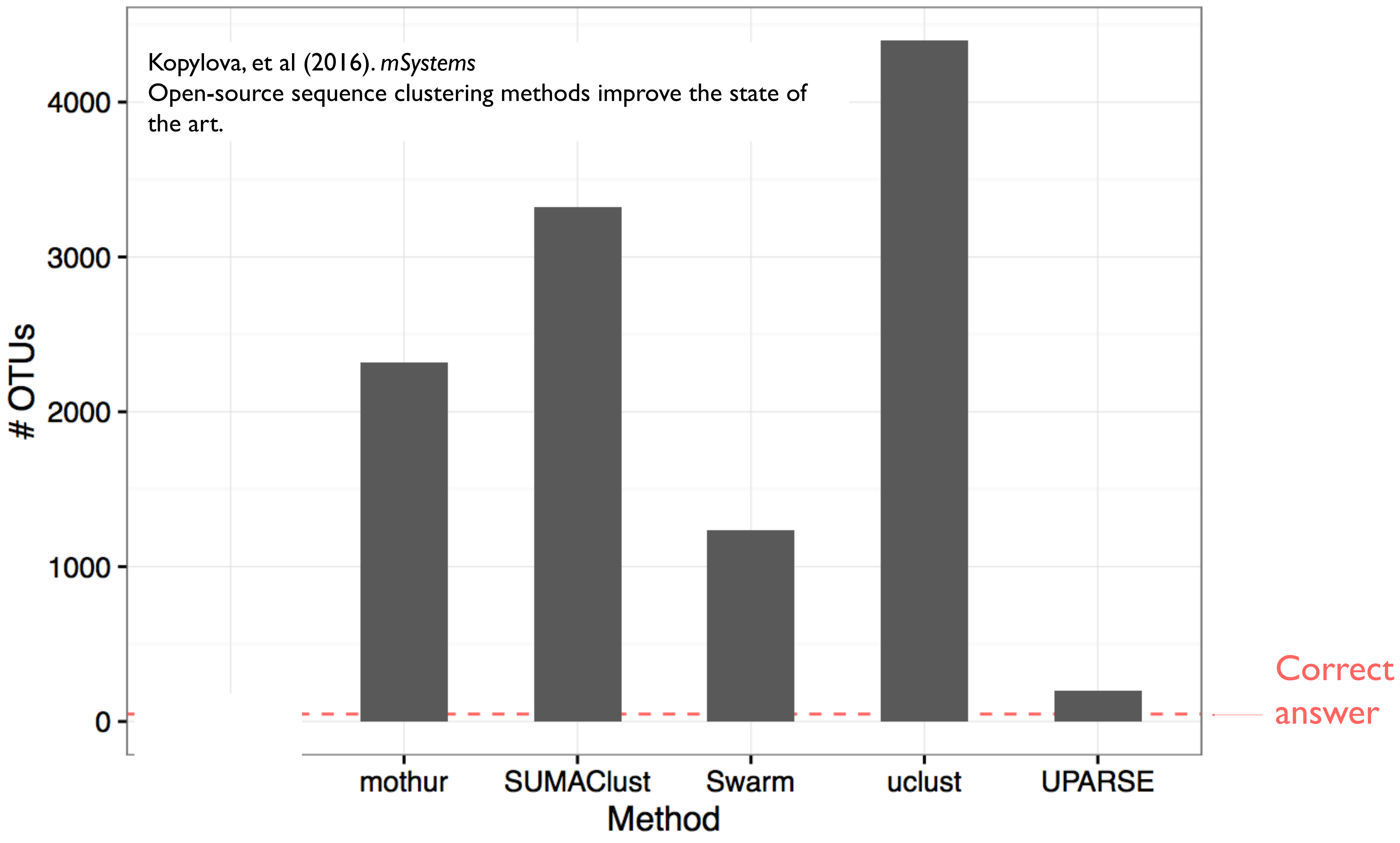
LINGERING PROBLEMS WITH "OTU"



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

LINGERING PROBLEMS WITH "OTU"

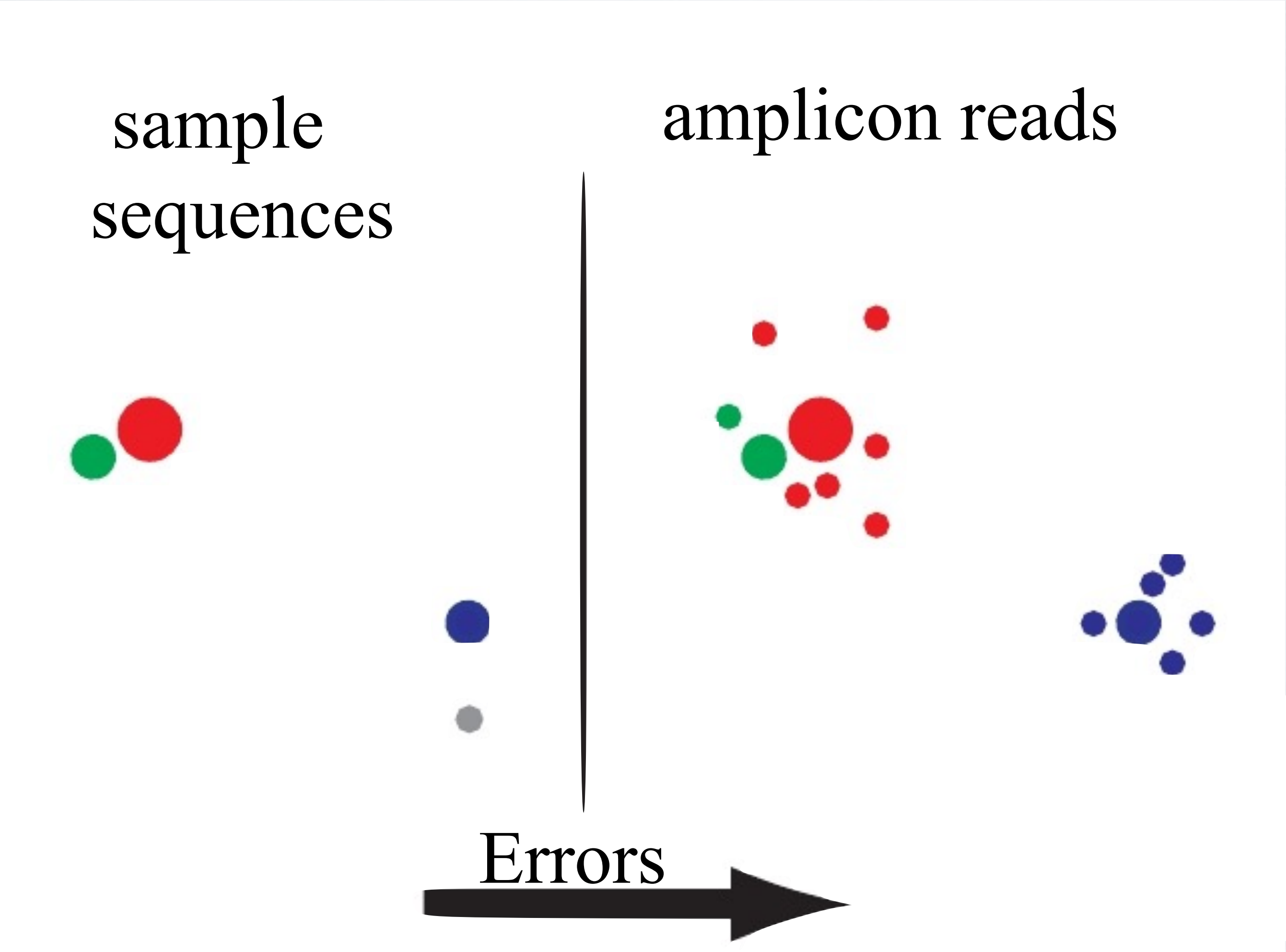
Typical "OTU" performance on validation data ("mock community")

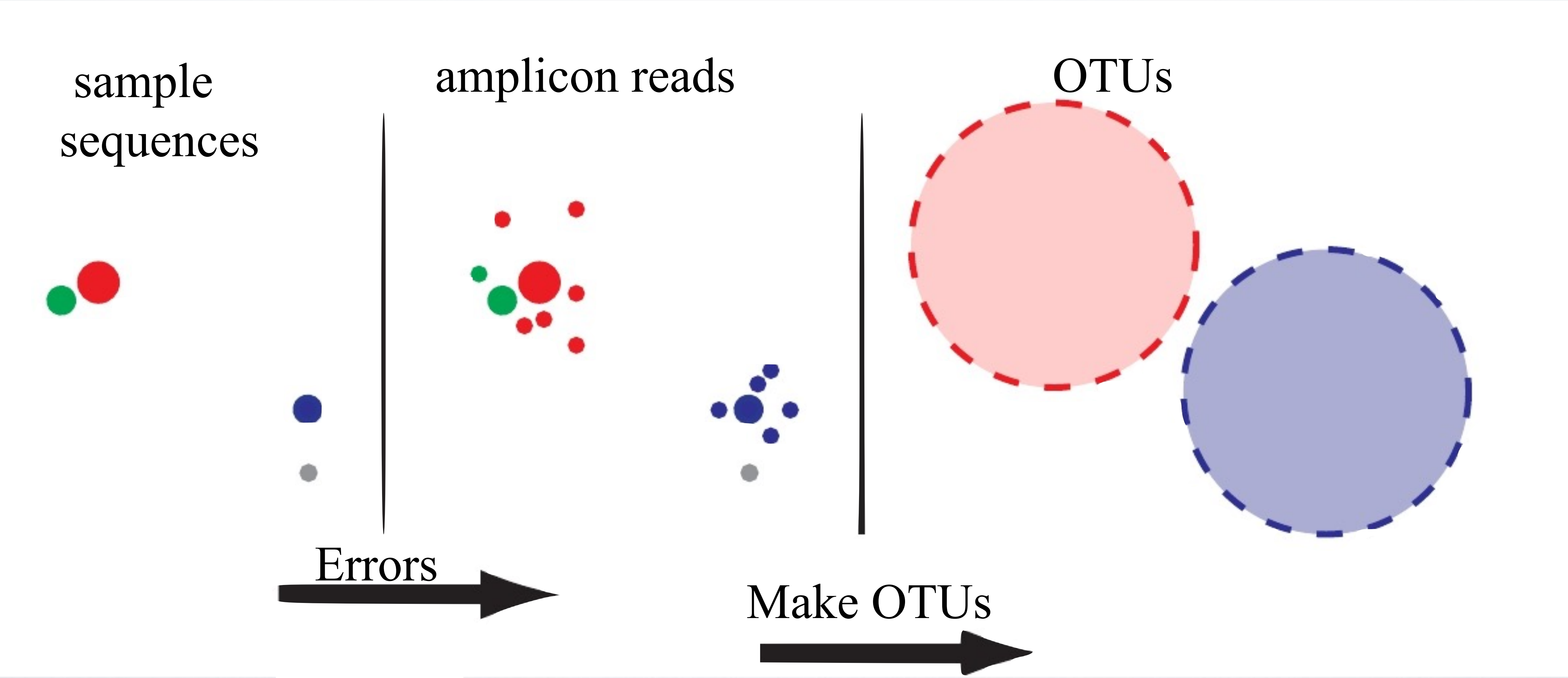


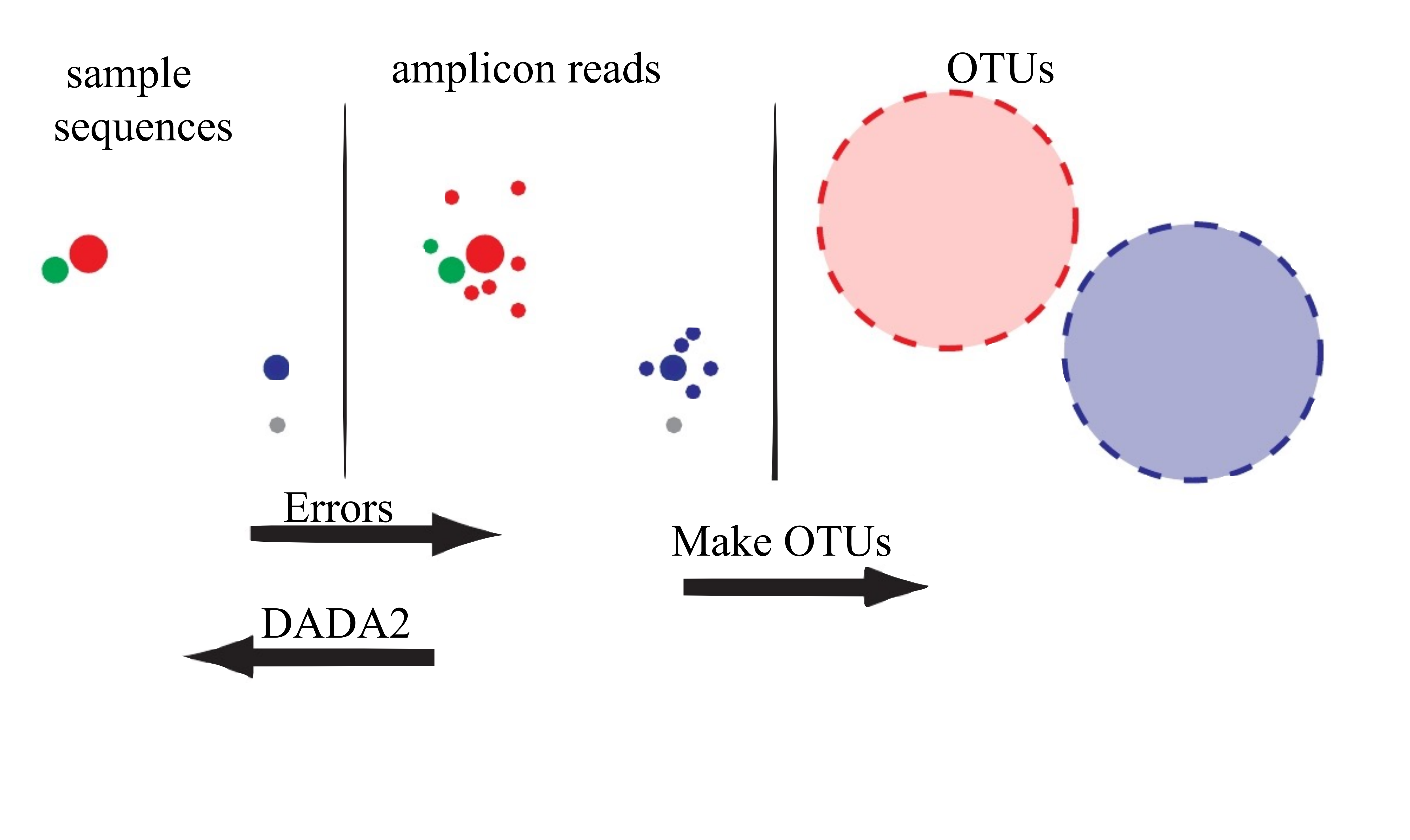
Kopylova, et al (2016). *mSystems*
Open-source sequence clustering methods improve the state of the art.

Correct answer

http://benjjneb.github.io/dada2/R/SotA.html

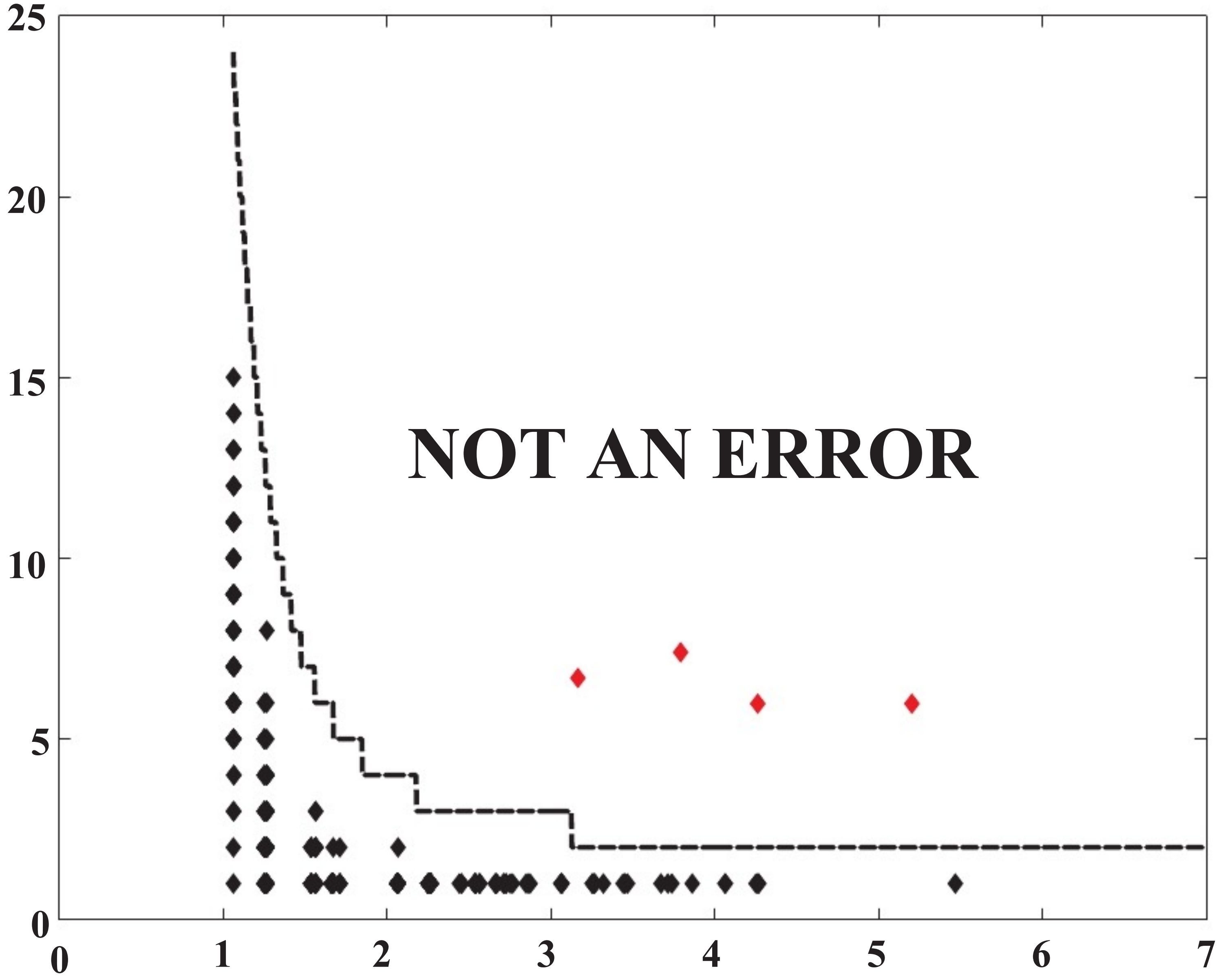

sample sequences
amplicon reads
Errors


sample sequences
amplicon reads
OTUs
Errors
Make OTUs


sample sequences
amplicon reads
OTUs
Errors
Make OTUs
DADA2

The shape of amplicon sequencing errors

counts, unique sequence

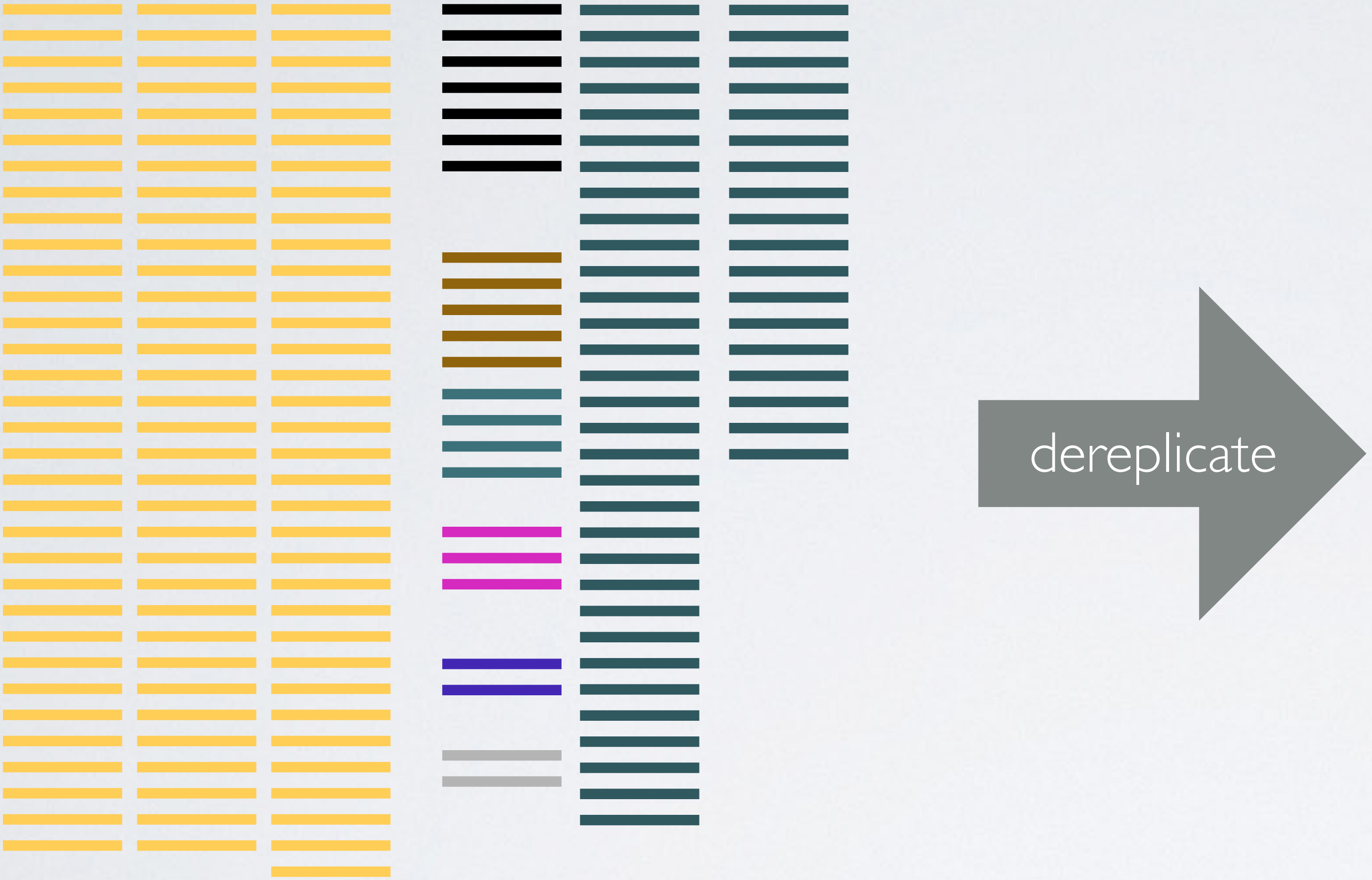


Effective Hamming Distance (number of substitutions from parent/reference)

Slide adapted from Benjamin Callahan

DADA2

“raw” reads

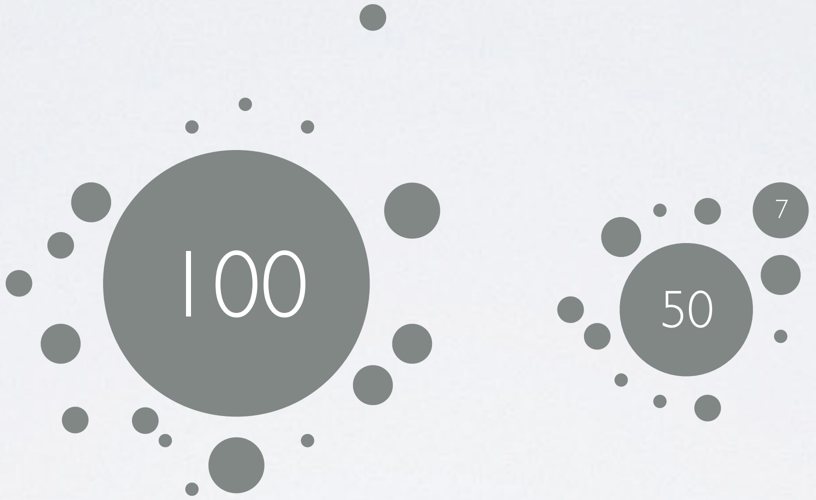


Input:

unique sequences	abundance	mean-Q
▬	100	32
▬	50	32
▬	7	20
▬	5	. . .
▬	4	. . .
▬	3	. . .
▬	2	. . .
▬	2	. . .

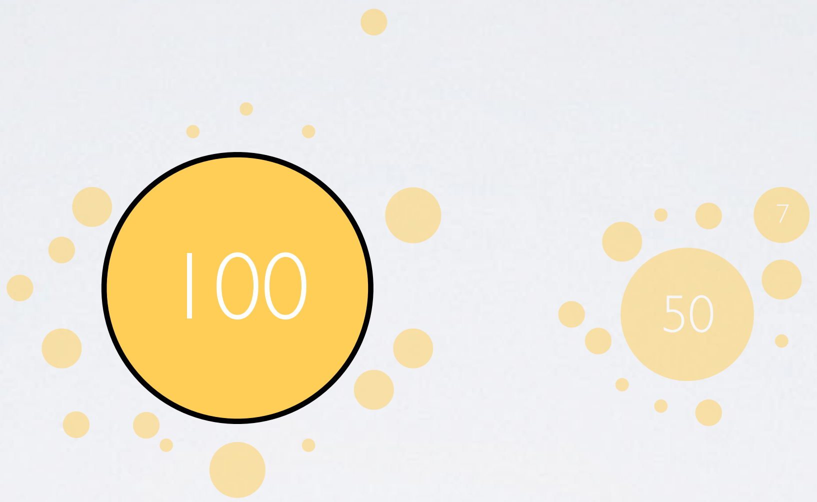
DADA2

Initial guess: one real sequence + errors



DADA2

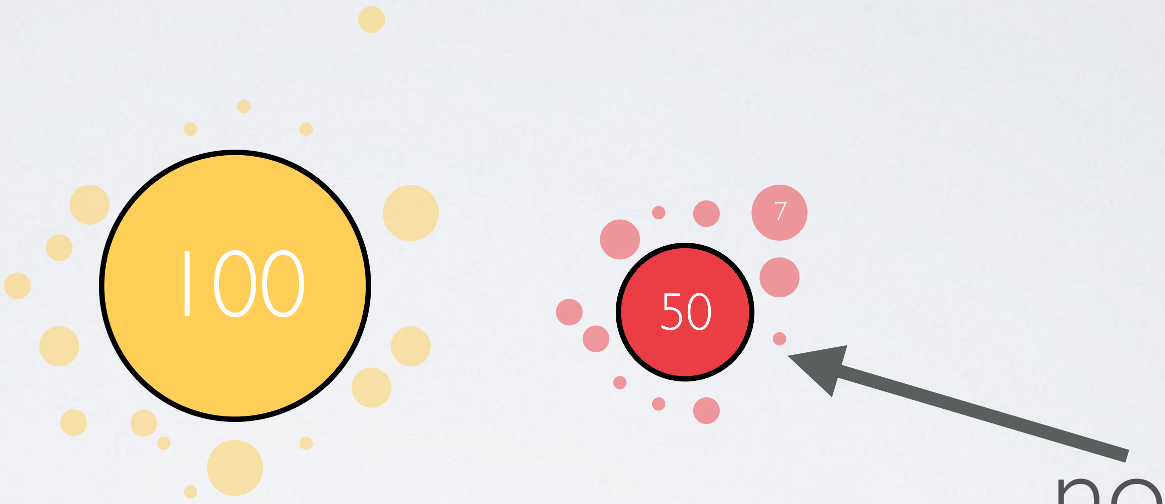
Infer initial error model under this assumption.



$$\Pr(i \rightarrow j) = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & 0.97 & 10^{-2} & 10^{-2} & 10^{-2} \\ C & 10^{-2} & 0.97 & 10^{-2} & 10^{-2} \\ G & 10^{-2} & 10^{-2} & 0.97 & 10^{-2} \\ T & 10^{-2} & 10^{-2} & 10^{-2} & 0.97 \end{array}$$

DADA2

Update the model.



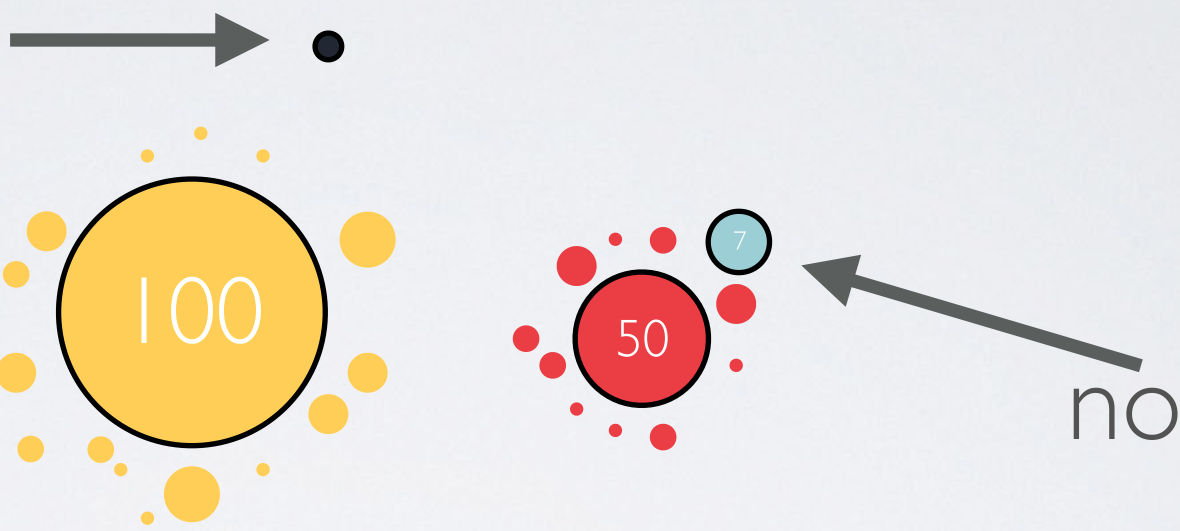
not an error

$$\Pr(i \rightarrow j) = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & 0.997 & 10^{-3} & 10^{-3} & 10^{-3} \\ C & 10^{-3} & 0.997 & 10^{-3} & 10^{-3} \\ G & 10^{-3} & 10^{-3} & 0.997 & 10^{-3} \\ T & 10^{-3} & 10^{-3} & 10^{-3} & 0.997 \end{array}$$

DADA2

Update model again

not an error



not an error

$\Pr(i \rightarrow j) =$

	A	C	G	T
A	0.998	1×10^{-4}	2×10^{-3}	2×10^{-4}
C	6×10^{-5}	0.998	3×10^{-4}	1×10^{-3}
G	1×10^{-4}	1×10^{-4}	0.998	6×10^{-5}
T	2×10^{-4}	2×10^{-3}	1×10^{-4}	0.998

DADA2 ASSUMPTIONS

DADA2 Error Model:

- Errors independent b/w different sequences
- Errors independent b/w sites within a sequence
- Sequence i is produced from parent sequence j with probability equal to the product of site-wise substitution probabilities:

$$\lambda_{j \to i} = \prod_{l=0}^{L} p(j(l) \to i(l), q(l))))$$

- Each substitution probability depends on original nt, substituting nt, and quality score at position in i

DADA2 ASSUMPTIONS

DADA2 Abundance Model:

- Errors are independent across reads
- Abundance of reads w/ sequence i produced from more-abundant parent sequence j is poisson distributed
- Expectation of abundance equals error rate, λj→i, multiplied by the abundance of sequence j
- i has count greater than or equal to one
- "Abundance p-value" for sequence i is thus:

$$p_A(j \rightarrow i) = \sum_{a=a_1}^{\infty} \rho_{pois}(n_j\lambda_{j\rightarrow i}, a)/(1 - \rho_{pois}(n_j\lambda_{j\rightarrow i},0)$$

- "Probability of seeing an abundance of sequence i that is equal to or greater than observed value, by chance, given sequence j." (Bonferroni-corrected)
- A low pA indicates there are more reads of sequence i than can

APPLICATIONS

- Any amplicon-seq data, not just 16S rRNA or even microbiome
- Sequence variants unique to an individual host
- Sequence variants associated with a clinical outcome
- Improved meta-genomic inference (e.g. PiCRUST)
- Mitigate ambiguity of representative genome(s) to use
- Detecting pathogens (special cases)