Clustering

Anders K. Krabberød
University of Oslo - Department of Biosciences
a.k.krabberod@ibv.uio.no

Illumina seq. and amplicon data

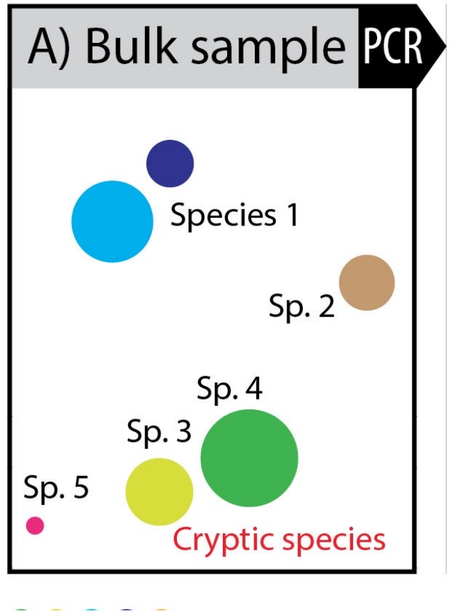
- Illumina HiSeq and MiSeq sequencing has been used extensively for amplicons sequencing in the last years
  - High yields with relative low cost
  - HiSeq (and NovaSeq) short fragments 2*150bp
  - MiSeq a bit longer reads 2*300bp
- Low error rate, but due to the high number of sequences generated errors are bound to occur in the library
  - (About 1% pr 1000 nt, not adjusted for improvements due to overlapping reads.)
- How to deal with errors, has been an ongoing discussion for high-throughput data for a quite some time

Why cluster?

- Reasons for clustering of metabarcoding data:
  - Reduce the effect of sequencing error
  - Reduce other sources of “noise”, PCR artefacts, intragenomic variation

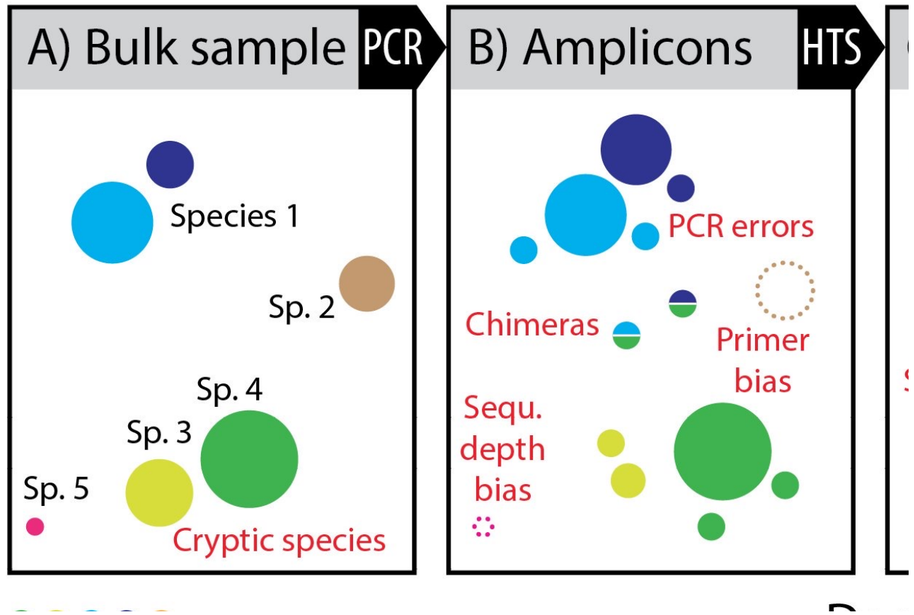
Why cluster?

- Reasons for clustering of metabarcoding data:
  - Reduce the effect of sequencing error
  - Reduce other sources of “noise”, PCR artefacts, intragenomic variation



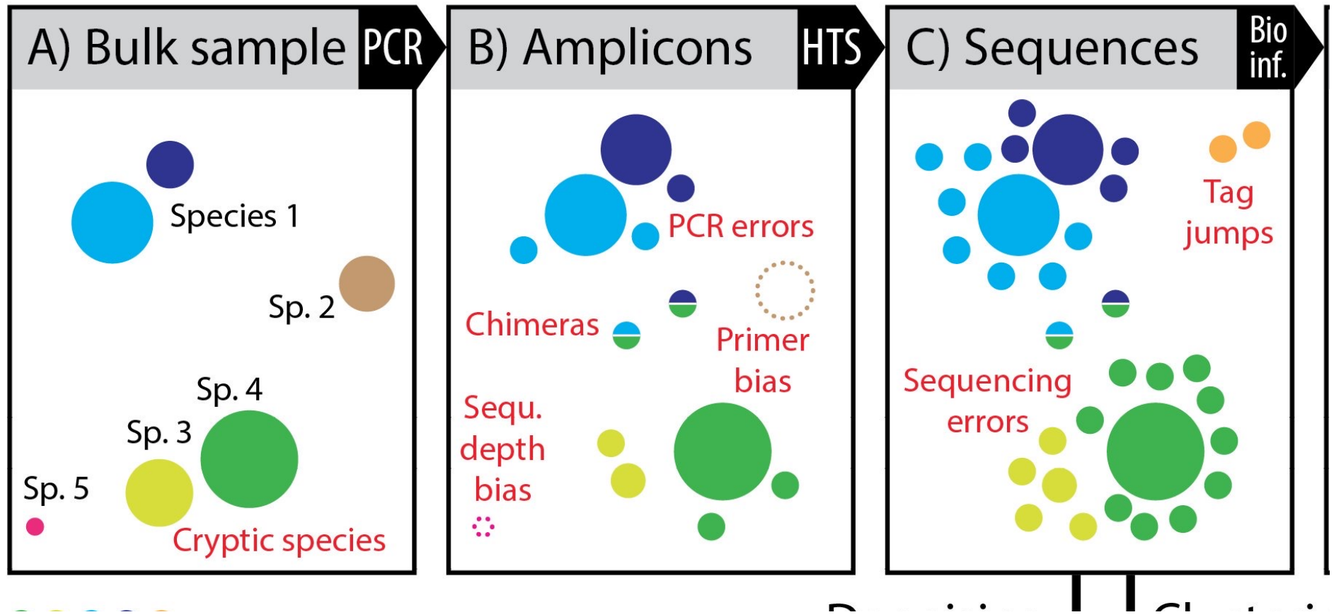
Why cluster?

- Reasons for clustering of metabarcoding data:
  - Reduce the effect of sequencing error
  - Reduce other sources of “noise”, PCR artefacts, intragenomic variation



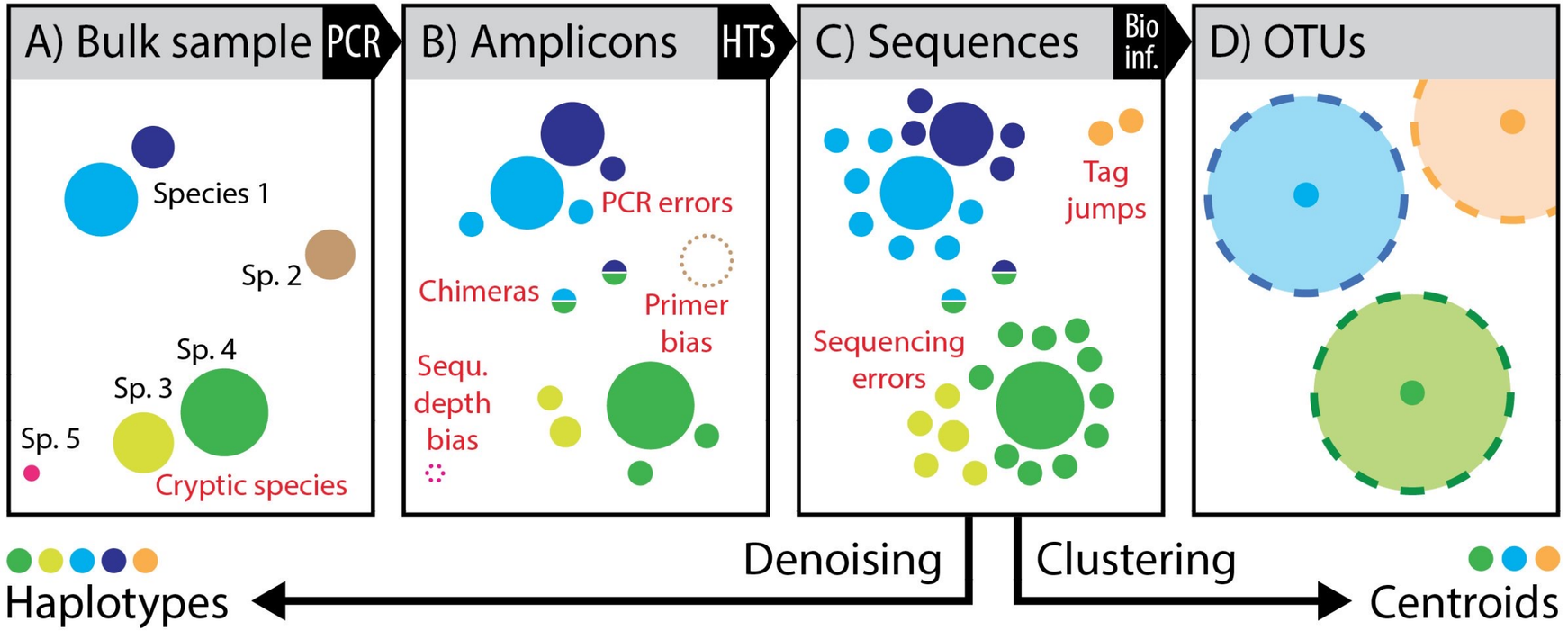
Why cluster?

- Reasons for clustering of metabarcoding data:
  - Reduce the effect of sequencing error
  - Reduce other sources of “noise”, PCR artefacts, intragenomic variation

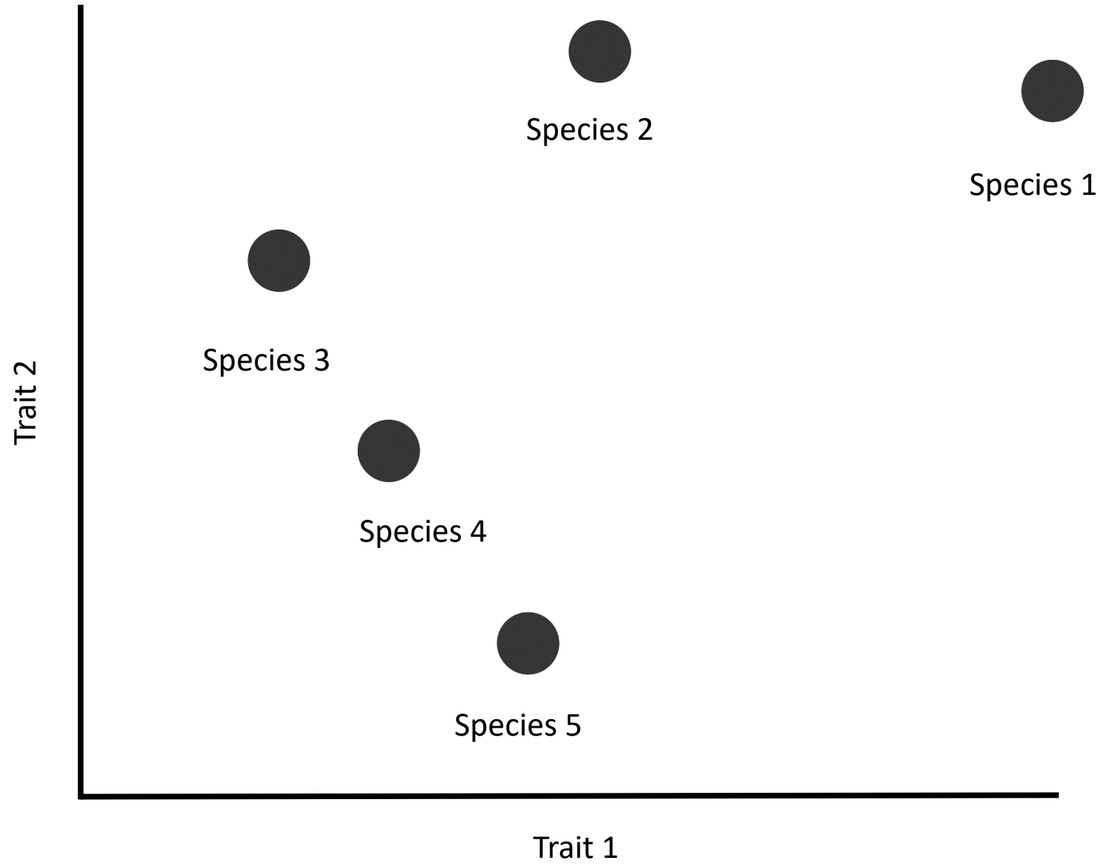


Why cluster?

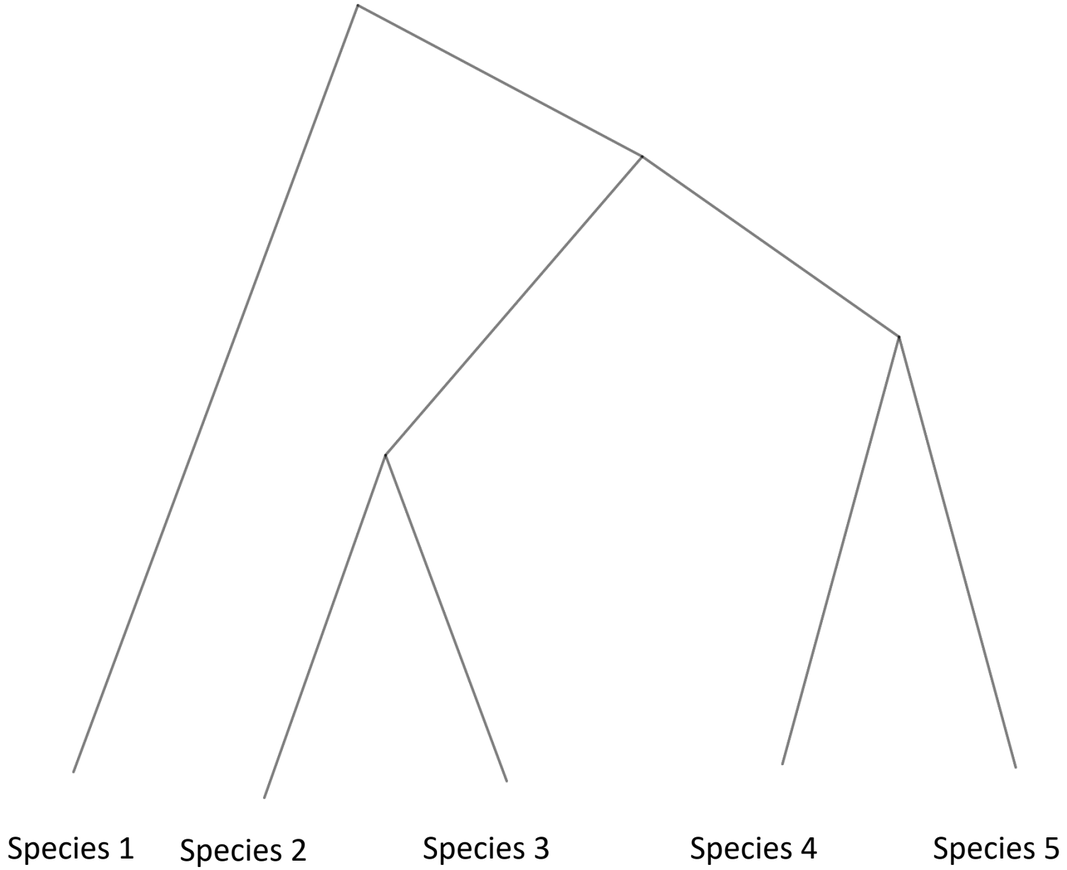
- Reasons for clustering of metabarcoding data:
  - Reduce the effect of sequencing error
  - Reduce other sources of “noise”, PCR artefacts, intragenomic variation

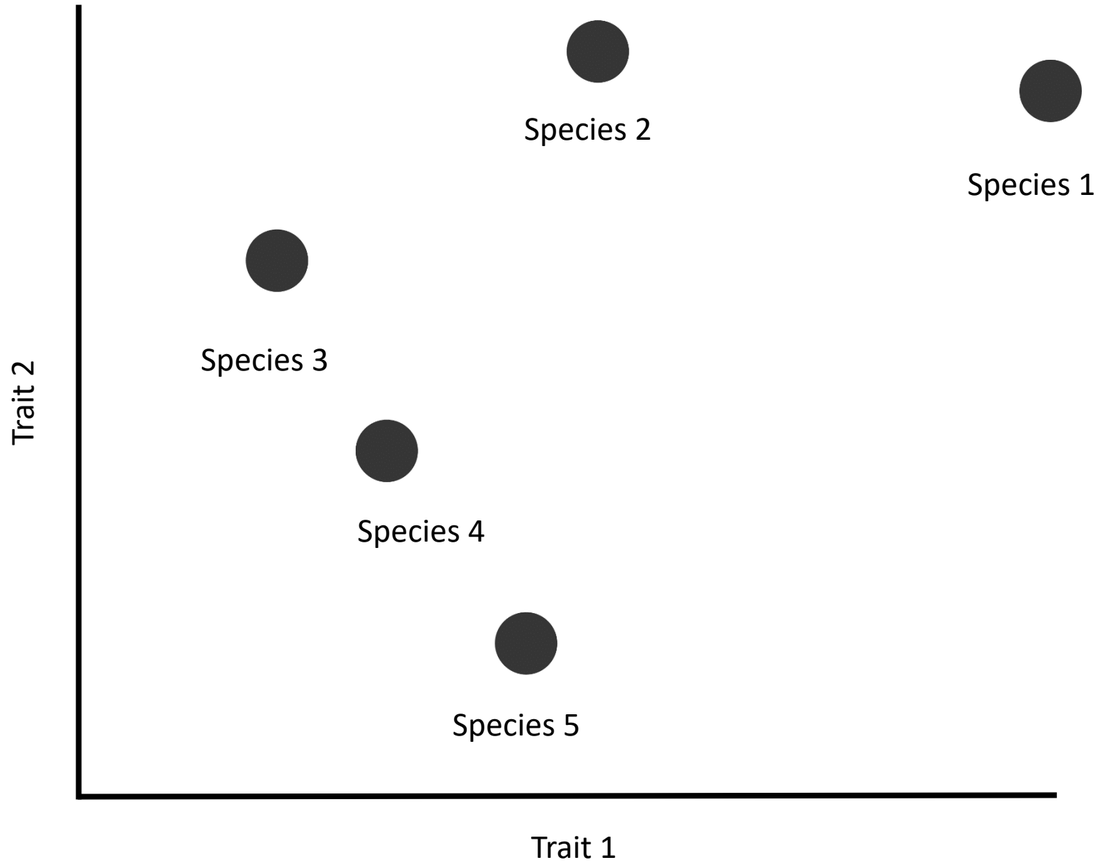


The ideal picture

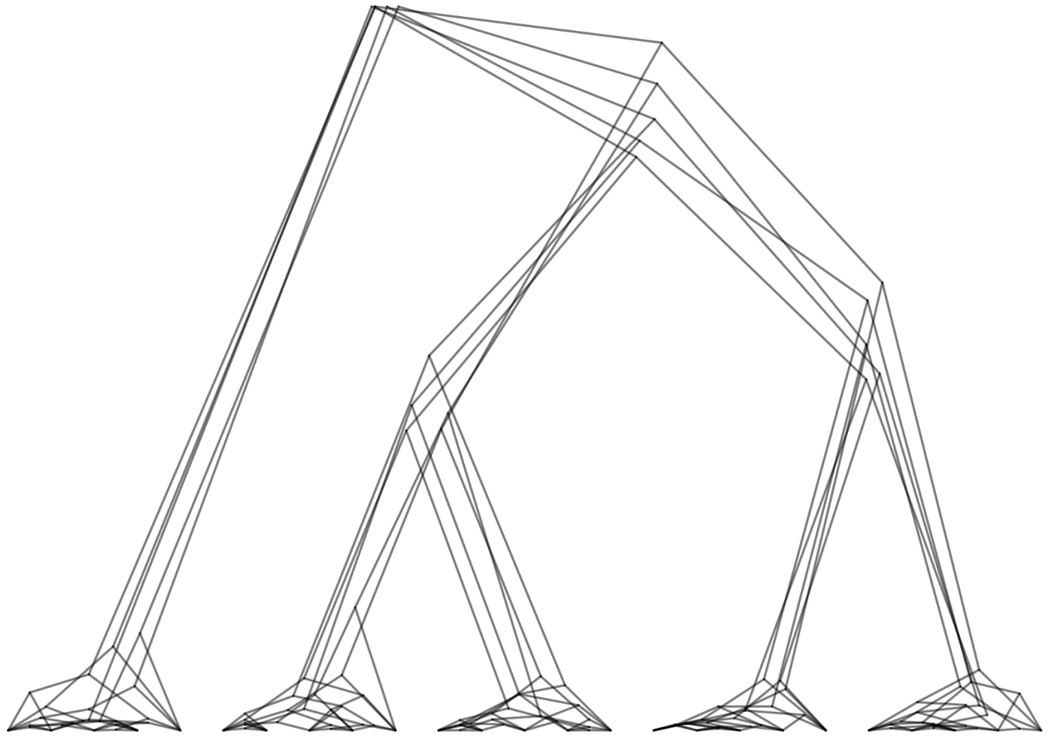


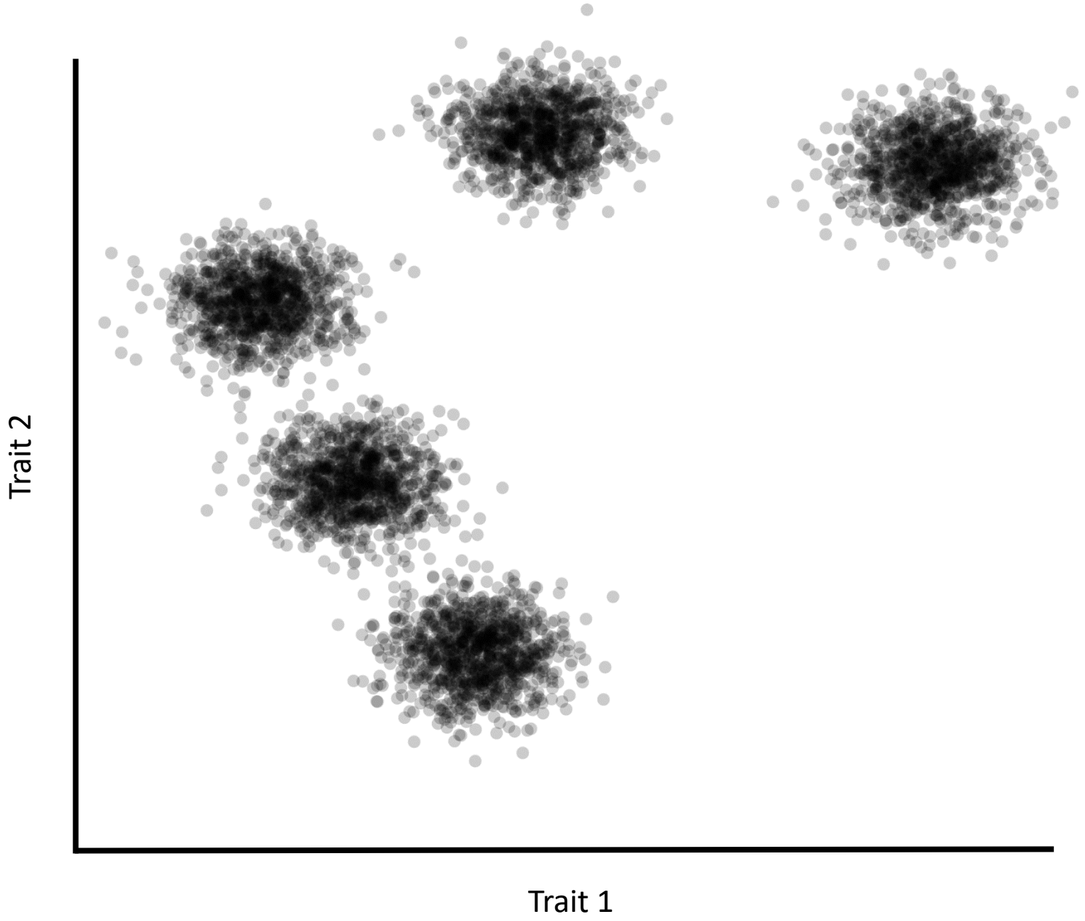
The ideal picture



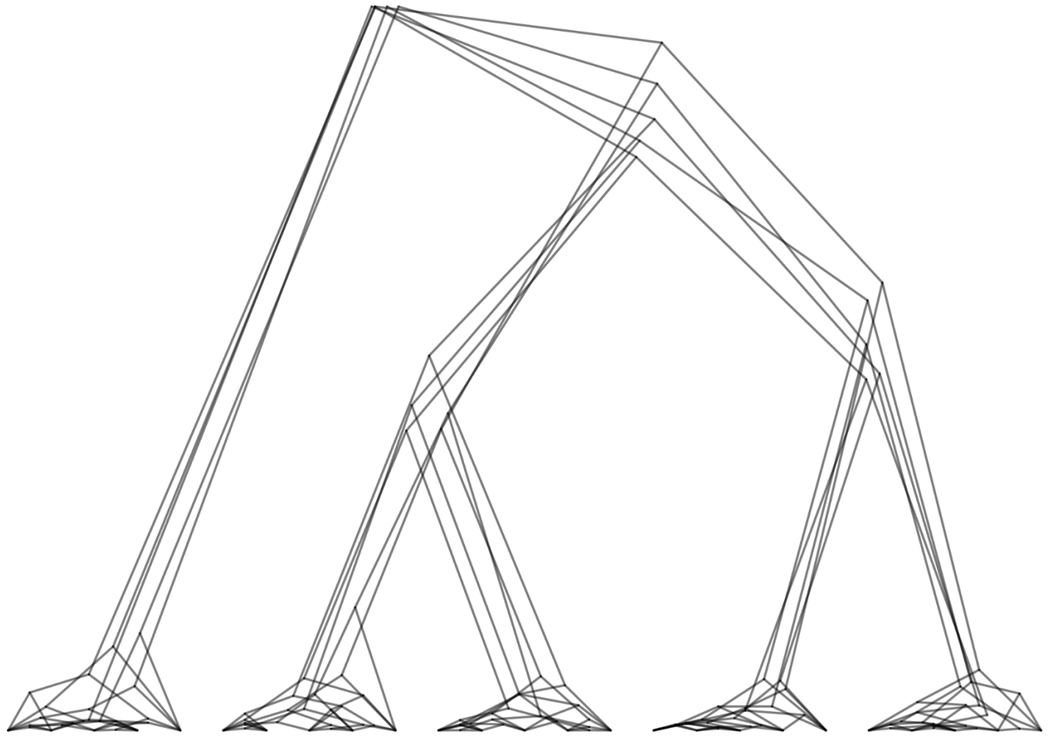


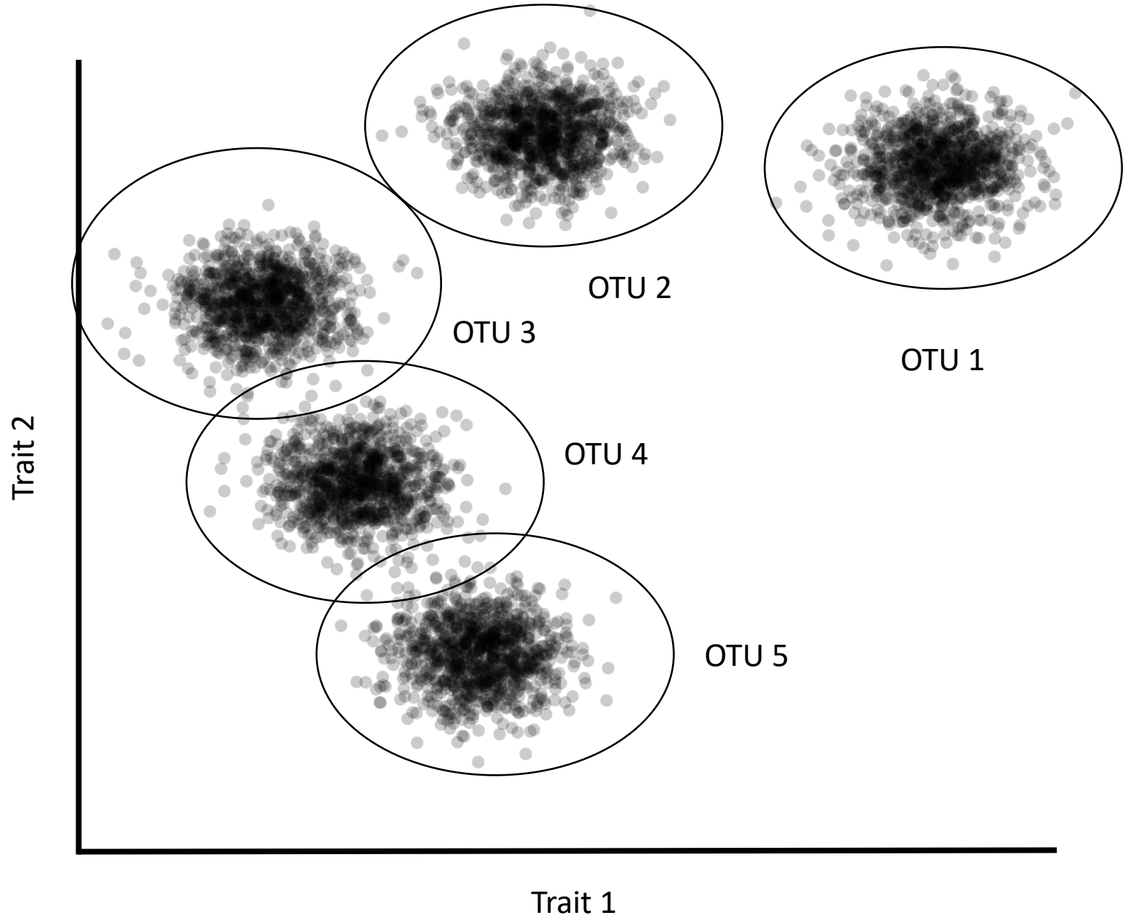
The reality





The reality



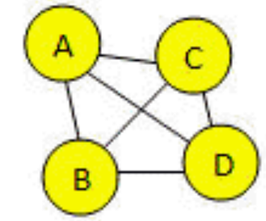


Amplicons, OTUs, and species

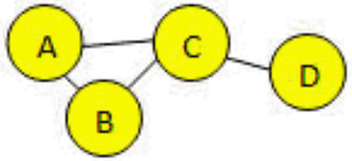
- Clusters of similar amplicons is called OTUs (operational taxonomic units)
  - Proxies for species, ecotypes, populations or another functional unit
  - Clustered based on often based on dissimilarity cut-off
  - But what cut-off to choose?
  - And how to calculate the similarity/dissimilarity?
- One solution is UPGMA-style clustering.
  - I.e. clustering at a fixed sequence distance
  - For instance 97%, which is often used for 18S V4

Clustering types

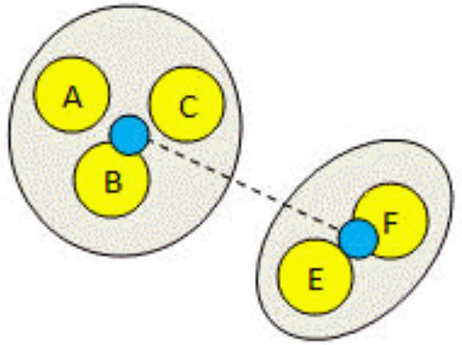
- **Complete linkage** means that *all* pairs of sequences in a cluster must be closer than the threshold.
- **Single linkage** means that a sequence should be included in a cluster if the distance to any other sequence is below the threshold.
- **Average linkage** (similar to UPGMA) distance between the “average sequence” for a cluster with the other cluster (the average is in fact calculated over all pairs).
- With average linkage, clusters tend to be larger than maximum linkage and smaller than minimum linkage.



Maximum distance
Complete linkage



Minimum distance
Single linkage



Average linkage

https://drive5.com/usearch/manual8.1/linkage.html

UPARSE - (U/VSEARCH)

- UPARSE-OTU takes sequences in order of decreasing abundance as input. This means that OTU centroids tend to be selected from the more abundant reads, and hence are more likely to be correct biological sequences.

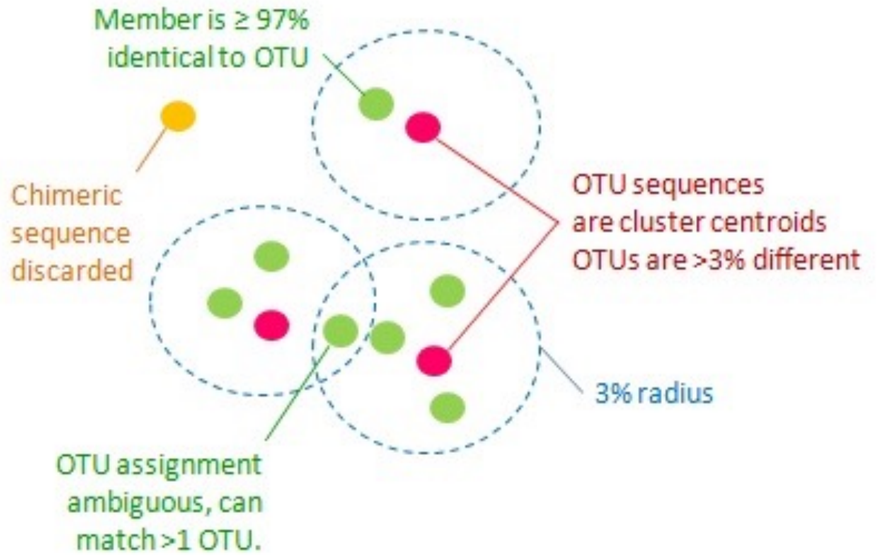
1. All pairs of OTU sequences should have pair-wise sequence identity for a set value (typically 97%)
2. An OTU sequence should be the most abundant within a 97% neighbourhood.
3. Chimeric sequences should be discarded.
4. All non-chimeric input sequences should match at least one OTU with >= 97% identity.

Published: 18 August 2013

UPARSE: highly accurate OTU sequences from microbial amplicon reads

Robert C Edgar

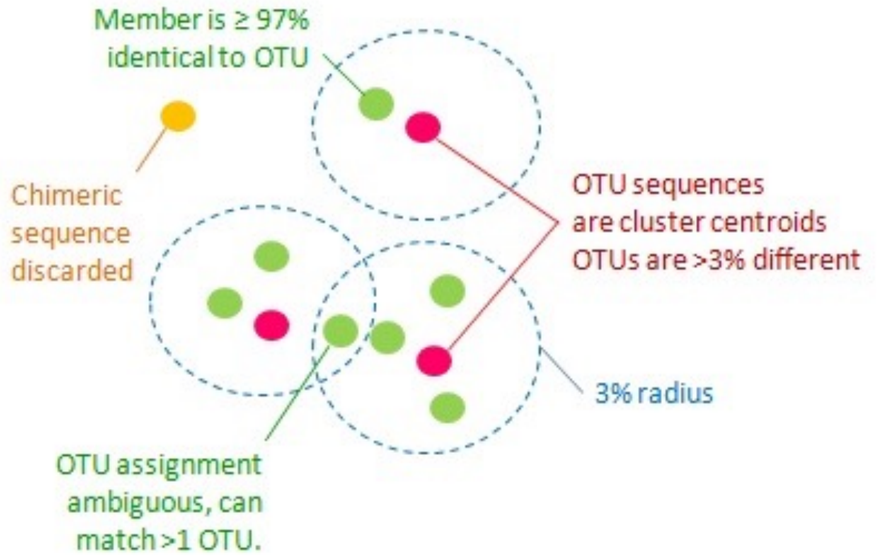
Nature Methods **10**, 996–998(2013) | Cite this article



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

UPARSE

- UPARSE implements a greedy algorithm that performs OTU clustering
- Advantages
  - Fast and greedy
  - Similarity threshold is relatively easy to interpret (although not necessarily easy to decide…)
- Disadvantages
  - Arbitrary fixed global clustering threshold. But different lineages evolve at different rate, which means that there is no single cut-off value for the entire tree of life
  - The input order of the amplicons strongly influences the clustering results. Centroid selections are not re-evaluated during the clustering process, this might lead to inaccurately formed OTUs.



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

SWARM

PeerJ

Swarm: robust and fast clustering method for amplicon-based studies

Frédéric Mahé[1,2,3], Torbjørn Rognes[4,5], Christopher Quince[6], Colomban de Vargas[1,2] and Micah Dunthorn[3]

- Fast and exact, two-phased, agglomerative, unsupervised (de novo) single-linkage-clustering algorithm.
- Advantage
  - No global (and arbitrary) clustering threshold
  - The result is not dependent on the input sequence order
- SWARM builds OTUs in two steps
  - An initial set of OTUs is constructed by iteratively agglomerating similar amplicons
  - Amplicon aundances are used to break clusters into sub-OTUs

Greedy cluster vs. SWARM

PeerJ

Swarm: robust and fast clustering method for amplicon-based studies

Frédéric Mahé[1,2,3], Torbjørn Rognes[4,5], Christopher Quince[6], Colomban de Vargas[1,2] and Micah Dunthorn[3]

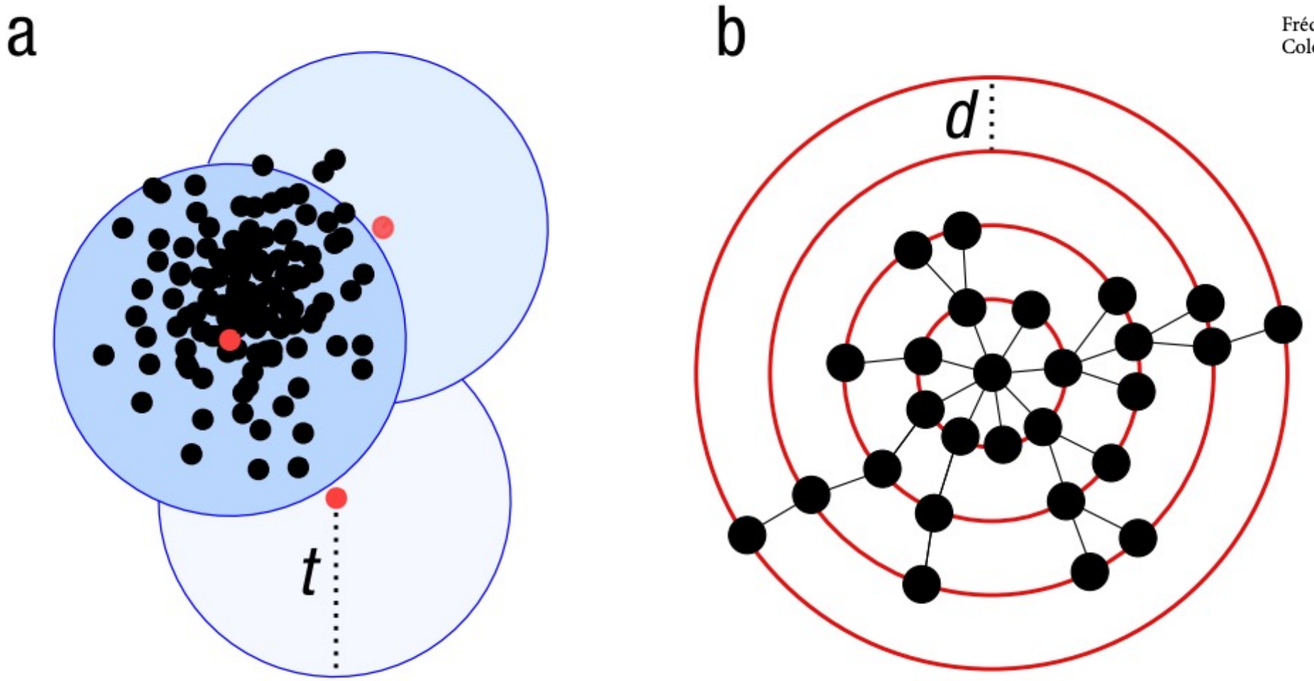
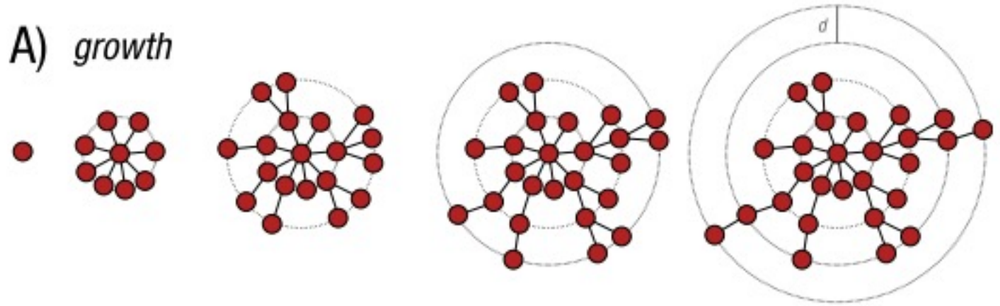


Figure 1 Schematic view of the greedy clustering approach and comparison with swarm. (A) Visualization of the widely used greedy clustering approach based on centroid selection and a global clustering threshold, *t*, where closely related amplicons can be placed into different OTUs. (B) By contrast, Swarm clusters iteratively by using a small user-chosen local clustering threshold, *d*, allowing OTUs to reach their natural limits.

SWARM

- (A) Swarm clusters amplicons iteratively by using a small user-chosen local threshold, d, allowing OTUs to grow to their natural limits, where no other amplicons can be added.



PeerJ

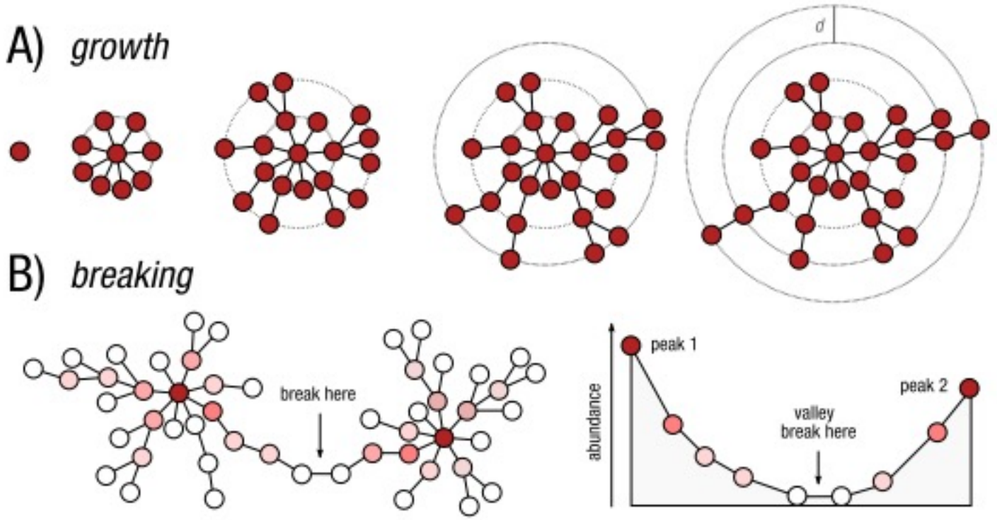
Swarm v2: highly-scalable and high-resolution amplicon clustering

Frédéric Mahé[1], Torbjørn Rognes[2,3], Christopher Quince[4], Colomban de Vargas[5,6] and Micah Dunthorn[1]

[1] Department of Ecology, Technische Universität Kaiserslautern, Kaiserslautern, Germany
[2] Department of Informatics, University of Oslo, Oslo, Norway
[3] Department of Microbiology, Oslo University Hospital, Rikshospitalet, Oslo, Norway
[4] Warwick Medical School, University of Warwick, Warwick, United Kingdom
[5] UMR 7144, EPEP-Évolution des Protistes et des Écosystèmes Pélagiques, Station Biologique de Roscoff, CNRS, Roscoff, France
[6] UMR7144 Station Biologique de Roscoff, Sorbonne Universités, UPMC Univ Paris 06, Roscoff, France

SWARM

- (A) Swarm clusters amplicons iteratively by using a small user-chosen local threshold, d, allowing OTUs to grow to their natural limits, where no other amplicons can be added.
- (B) Swarm takes into account the abundance of each amplicon to produce higher resolution clusters, by not allowing the formation of amplicon chains. The darker the red, the higher the abundance.



PeerJ

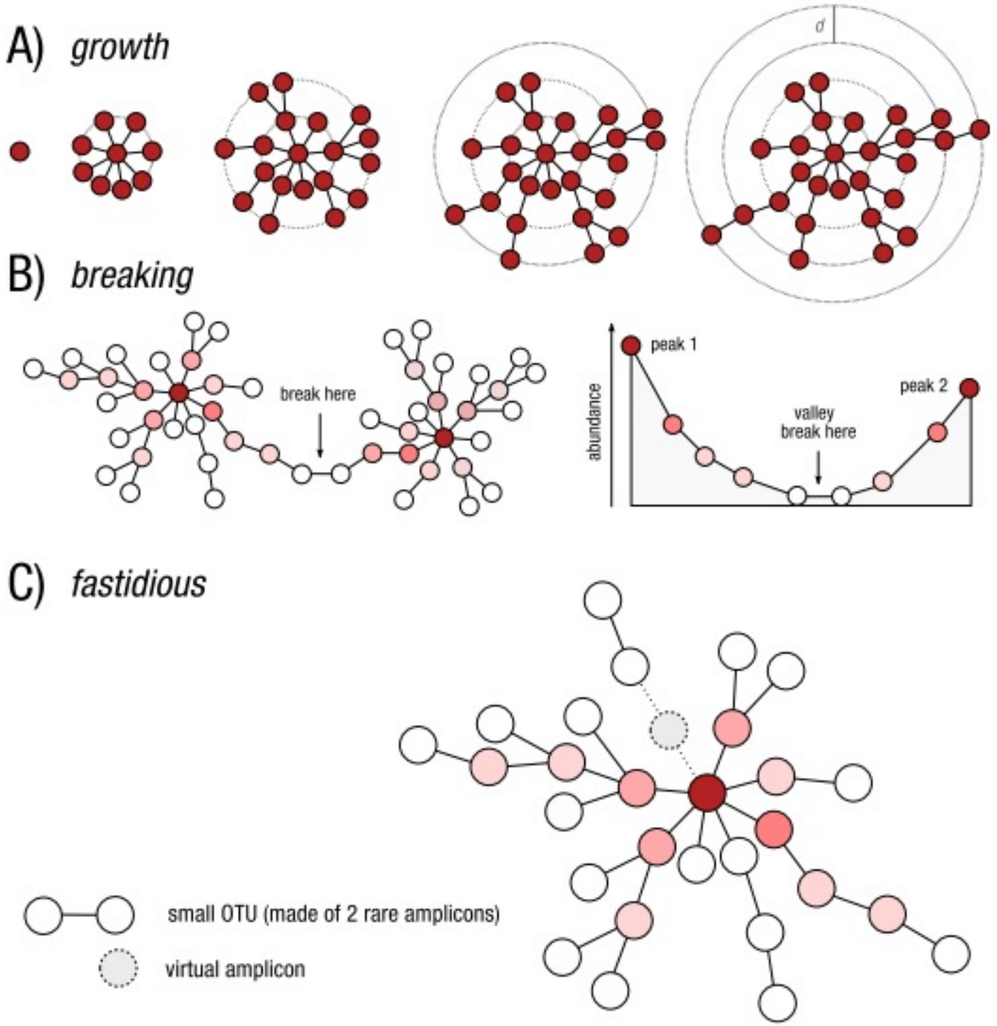
Swarm v2: highly-scalable and high-resolution amplicon clustering

Frédéric Mahé[1], Torbjørn Rognes[2,3], Christopher Quince[4], Colomban de Vargas[5,6] and Micah Dunthorn[1]

[1] Department of Ecology, Technische Universität Kaiserslautern, Kaiserslautern, Germany
[2] Department of Informatics, University of Oslo, Oslo, Norway
[3] Department of Microbiology, Oslo University Hospital, Rikshospitalet, Oslo, Norway
[4] Warwick Medical School, University of Warwick, Warwick, United Kingdom
[5] UMR 7144, EPEP-Évolution des Protistes et des Écosystèmes Pélagiques, Station Biologique de Roscoff, CNRS, Roscoff, France
[6] UMR7144 Station Biologique de Roscoff, Sorbonne Universités, UPMC Univ Paris 06, Roscoff, France

SWARM

- (A) Swarm clusters amplicons iteratively by using a small user-chosen local threshold, d, allowing OTUs to grow to their natural limits, where no other amplicons can be added.
- (B) Swarm takes into account the abundance of each amplicon to produce higher resolution clusters, by not allowing the formation of amplicon chains. The darker the red, the higher the abundance.
- (C) The fastidious option avoids under-grouping (e.g., the production of small OTUs such as singletons and doubletons) by postulating the existence of virtual linking amplicons to graft smaller OTUs onto larger ones.



PeerJ

Swarm v2: highly-scalable and high-resolution amplicon clustering

Frédéric Mahé[1], Torbjørn Rognes[2,3], Christopher Quince[4], Colomban de Vargas[5,6] and Micah Dunthorn[1]

[1] Department of Ecology, Technische Universität Kaiserslautern, Kaiserslautern, Germany
[2] Department of Informatics, University of Oslo, Oslo, Norway
[3] Department of Microbiology, Oslo University Hospital, Rikshospitalet, Oslo, Norway
[4] Warwick Medical School, University of Warwick, Warwick, United Kingdom
[5] UMR 7144, EPEP - Évolution des Protistes et des Écosystèmes Pélagiques, Station Biologique de Roscoff, CNRS, Roscoff, France
[6] UMR7144 Station Biologique de Roscoff, Sorbonne Universités, UPMC Univ Paris 06, Roscoff, France

DADA2

- An update of DADA (Rosen et al., 2012)
- **Divisive Amplicon Denoising Algorithm**
  - Originally made for 454 data.
- **DADA2** Infers ”amplicon sequence variants from Illumina-scale amplicon data data without imposing the arbitrary dissimilarity thresholds that define molecular OTUs”

DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan[1], Paul J McMurdie[2], Michael J Rosen[3], Andrew W Han[2], Amy Jo A Johnson[2] & Susan P Holmes[1]

Callahan et al., Nat.Meth. (2016); Callahan et al., ISMEj (2017)

DADA2

- Advantages (according to themselves, aka the selling point…)
  - **Resolution**: DADA2 infers exact amplicon sequence variants (ASVs) from amplicon data, resolving biological differences of even 1 or 2 nucleotides.
  - **Accuracy**: DADA2 reports fewer false positive sequence variants than other methods report false OTUs.
  - **Comparability**: The ASVs output by DADA2 can be directly compared between studies, without the need to reprocess the pooled data.
  - **Computational Scaling**: The compute time of DADA2 scales linearly sample number, and memory requirements are essentially flat.

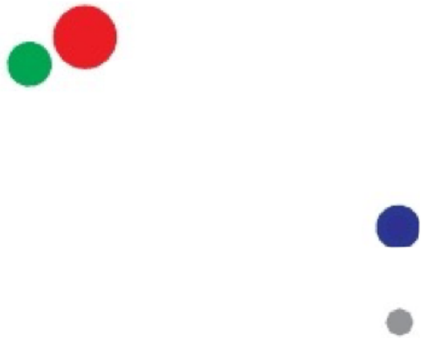
Callahan et al., Nat.Meth. (2016); Callahan et al., ISMEj (2017)

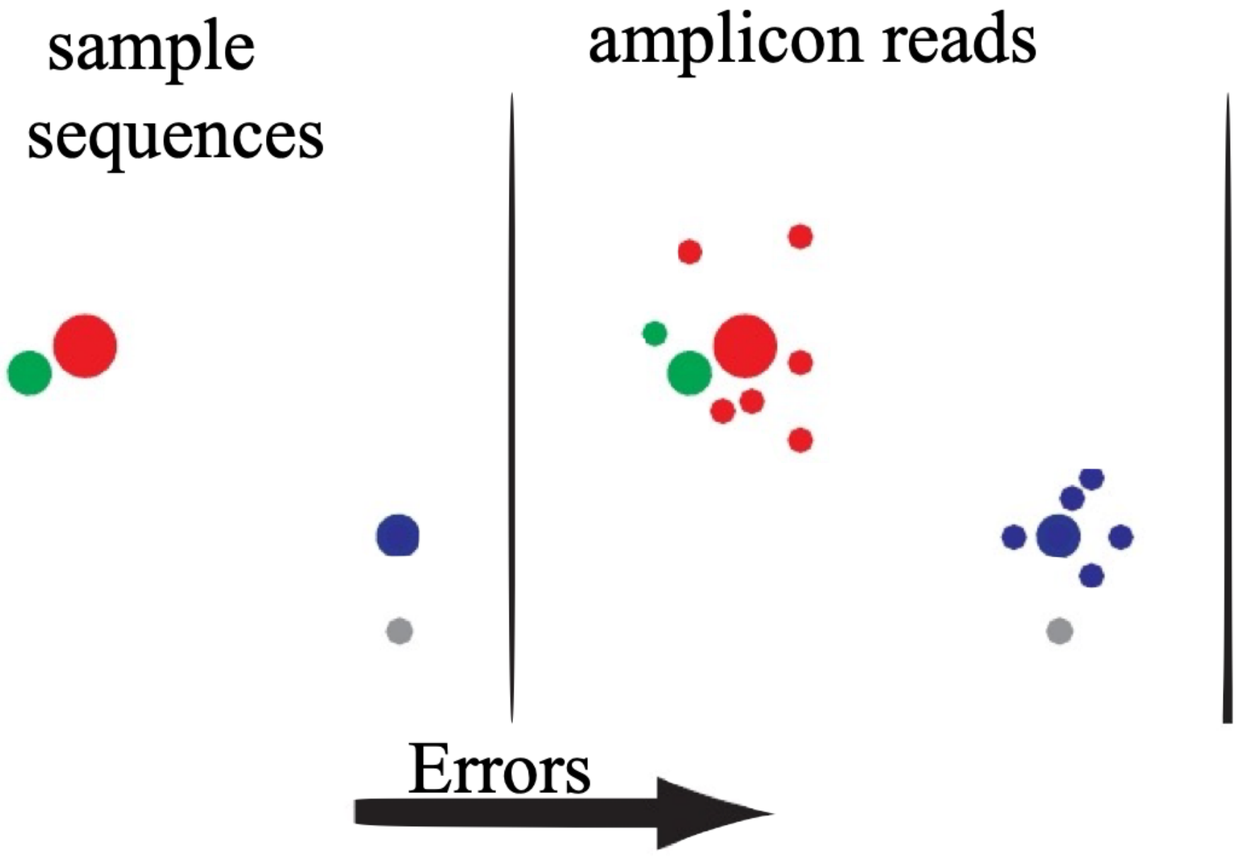
DADA2

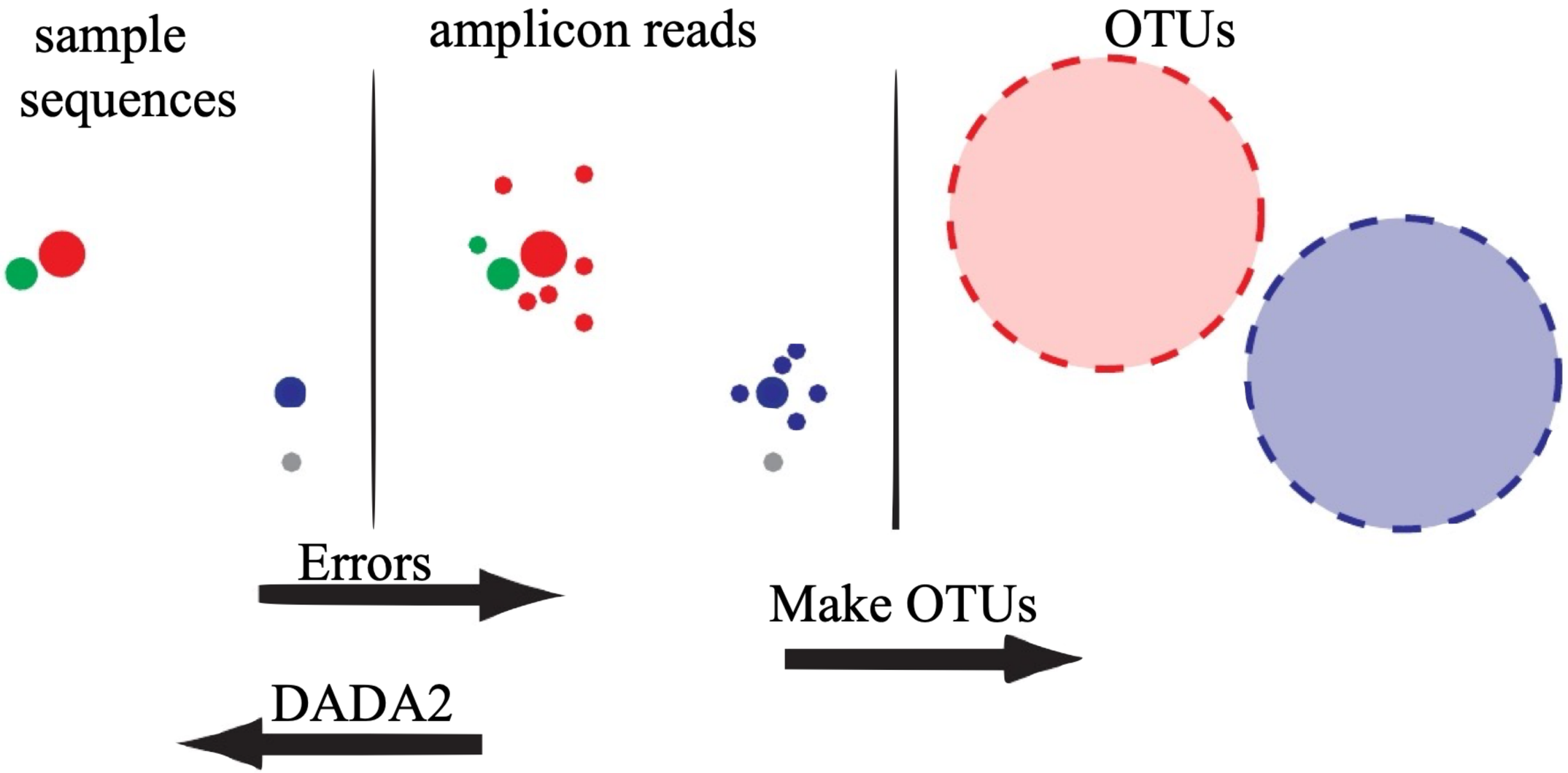
- The error model incorporates quality information, which is usually ignored by other methods (after filtering).
- The error model incorporates quantitative abundances, whereas most other methods use abundance ranks (if they use abundance at all).
- The error model have different error rates for different substitutions,
  - eg. A->C, is different from A->G whereas other methods merely count the mismatches.
- DADA2 can parameterize its error model from the data itself, rather than relying on previous datasets that may or may not reflect the PCR and sequencing protocols used in your study.

Callahan et al., Nat.Meth. (2016); Callahan et al., ISMEj (2017)

sample sequences




sample
sequences
amplicon reads
Errors

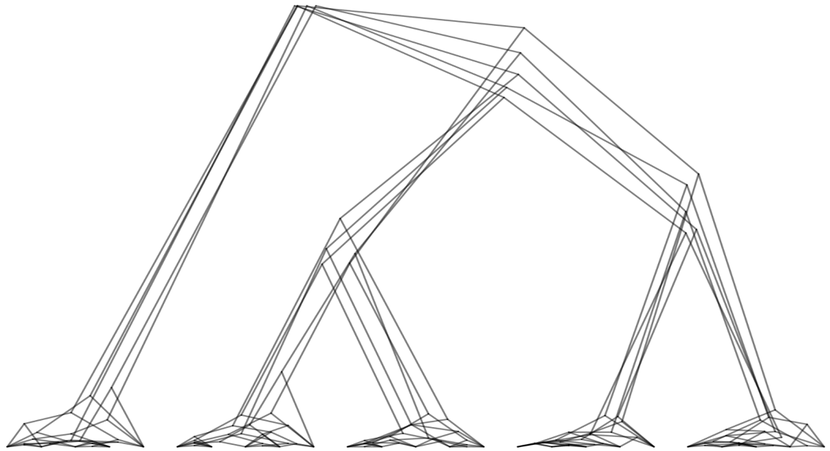

sample sequences
amplicon reads
OTUs
Errors
Make OTUs
DADA2

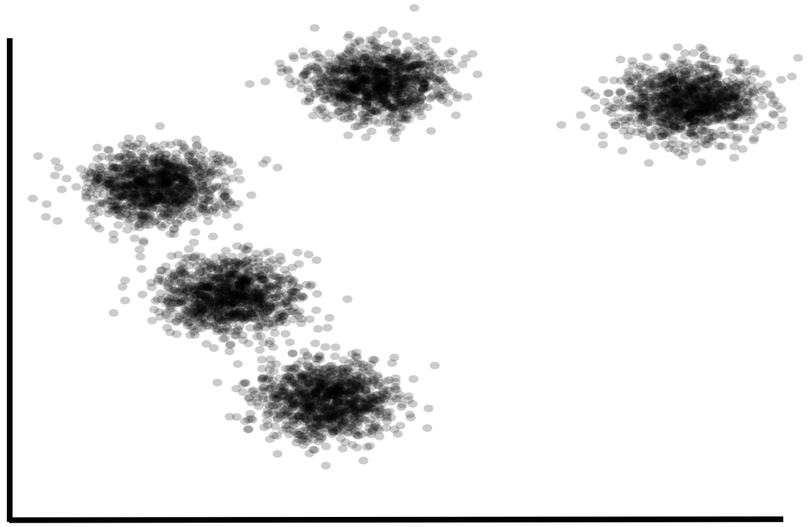
DADA2

- The core denoising algorithm in the DADA2 is built on a model of the errors in Illumina-sequenced amplicon reads.
- **However:** Clustering might still be needed since there are other sources of noise.
- And depends on the marker sequenced
  - SSU (18S) has multiple copies in many eukaryotes
  - Intragenomic variation might be around the “usuall” cut-off used for vsearch/usearch type clustering (i.e. 2-3% difference).
  - ITS is highly variable

Do you still need to cluster?

- Clustering might still be needed since there are other sources of “noise”.
  - I.e. even though an inferred read is a “true sequence” with a discernible abundance it might not by itself represent a species.





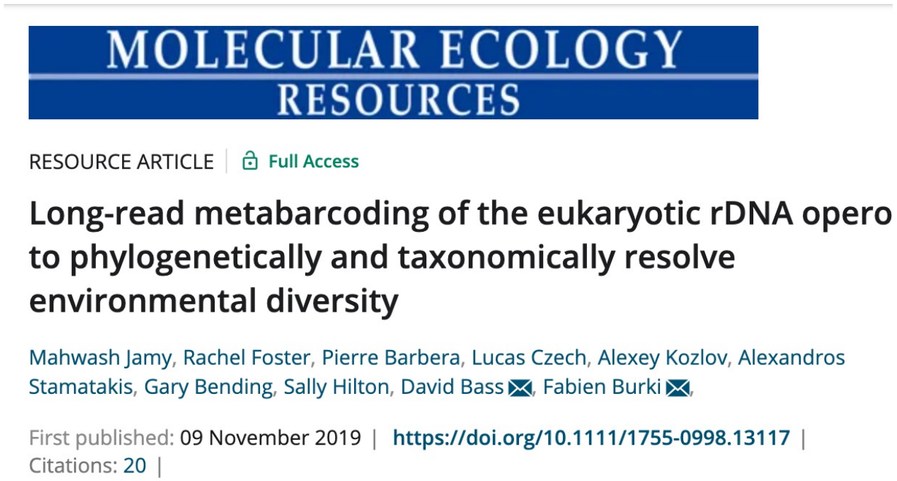
The output

- For our purpose, the important part is that the output is a table with the ***read*** abundance for ***OTUs*** across the samples.
- Keep in mind that these are *Operational Taxonomic Units* **not** necessarily the same as species
- *Read abundance* does not equal organismal abundance

	Sample_1	Sample_2	Sample_3	Sample_4	Sample_5	Sample_6	etc…
OTU_1	1	563	87	459	43	0	
OTU_2	213	7	5	3	68	8	
OTU_3	15	6	88	95	9	6	
OTU_4	762	5	43	1	0	0	
….etc							

Long read amplicon sequencing

- Recent development in long read sequencing technologies:
  - PacBio HiFi reads are 99.9% accurate and up to 15 kbp long!
  - Still not as many reads as Illumina sequencing, but there is no need for denoise.
  - The much longer reads have a higher phylogenetic signal



MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE | Full Access

Long-read metabarcoding of the eukaryotic rDNA opero to phylogenetically and taxonomically resolve environmental diversity

Mahwash Jamy, Rachel Foster, Pierre Barbera, Lucas Czech, Alexey Kozlov, Alexandros Stamatakis, Gary Bending, Sally Hilton, David Bass, Fabien Burki,

First published: 09 November 2019 | https://doi.org/10.1111/1755-0998.13117 | Citations: 20 |

https://doi.org/10.1111/1755-0998.13117

DADA2

- Resources:
- The developer has a nice tutorial
  - https://benjjneb.github.io/dada2/
  - https://benjjneb.github.io/dada2/tutorial.html
- DADA2 for ITS
  - https://benjjneb.github.io/dada2/ITS_workflow.html

DADA2

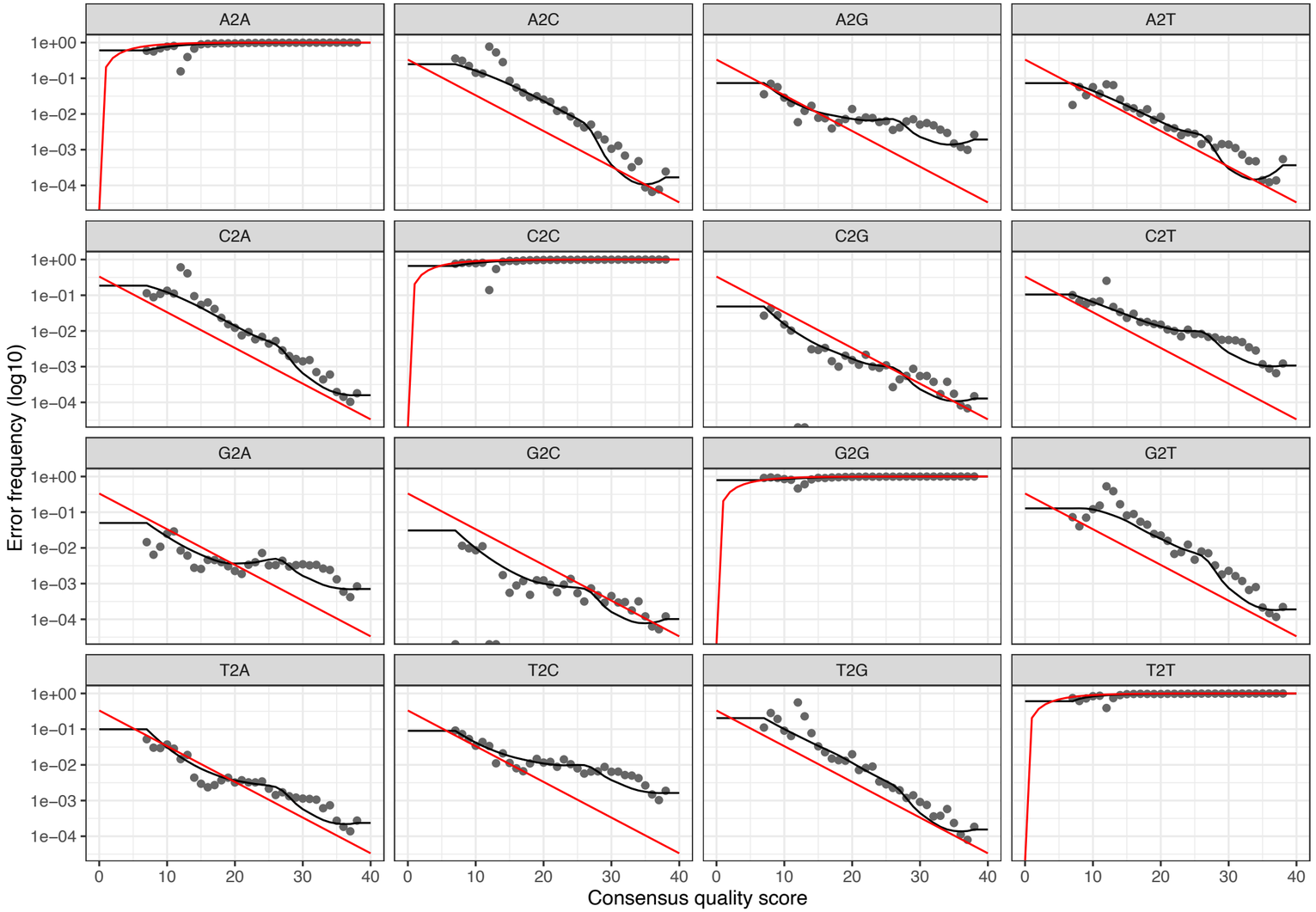
- **Build error model**
  - Assumption: Errors occurs independently within a read
  - Assumption: Errors occurs independently between reads
  - The rate at which an amplicon read with sequence *i* is produced from sample sequence *j* is the product of the transition probabilities of the aligned nucleotides:

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

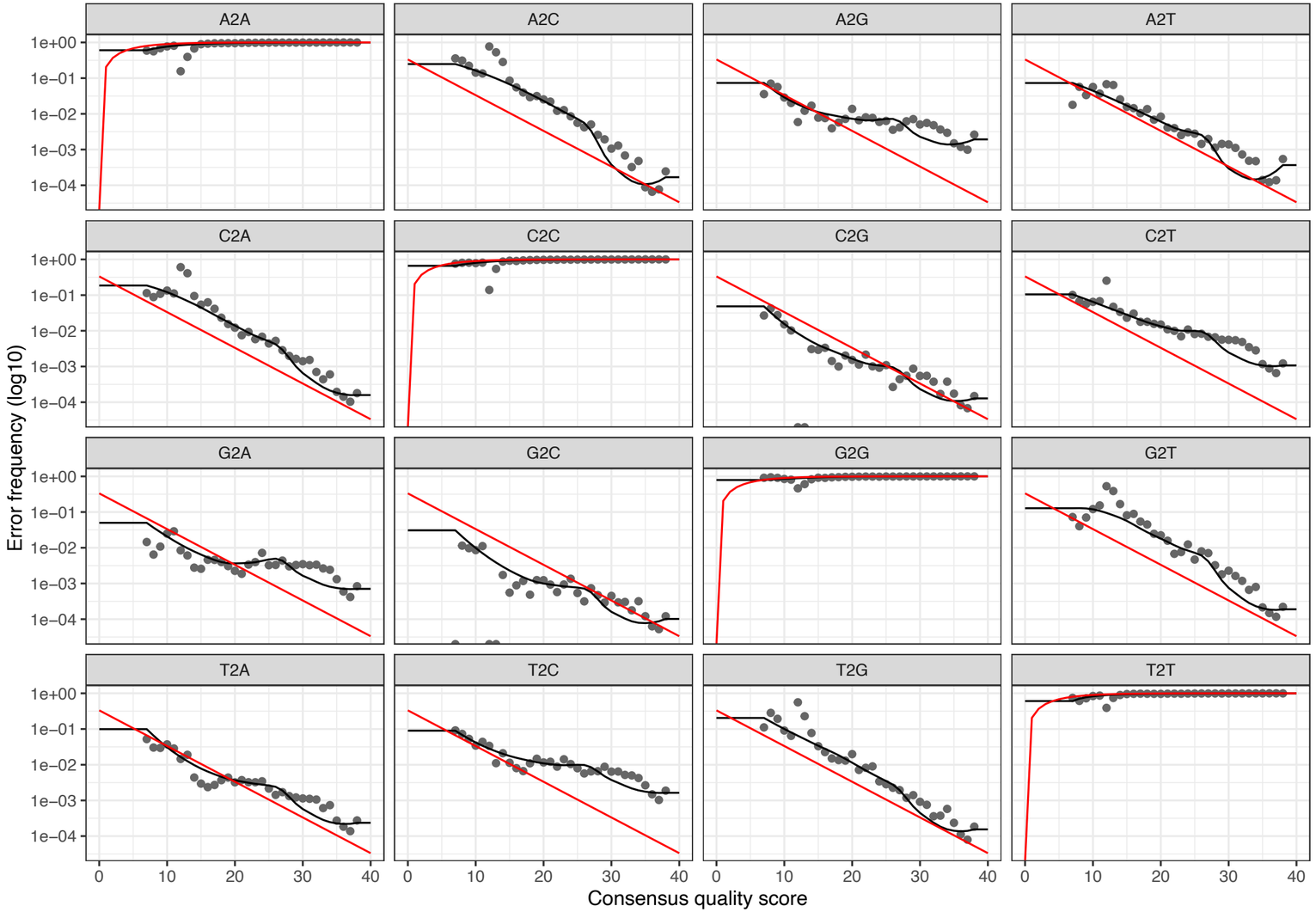
  - The transition probability between aligned nucleotides depend on the original nucleotide, substituting nucleotide, and associated quality score, for example, p(A→C, 35).
  - After sequence alignment, the error rate *λji* is calculated and stored.

Callahan et al., (2016)

DADA2 - Error model



DADA2 - Error model



DADA2

- **The abundance p-value.**
  - The number of amplicon reads with sequence *i* that will be produced from sample sequence *j* is Poisson distributed with expectation equal to an error rate *λji* multiplied by the expected reads of sample sequence *j*

$$p_A(j \to i) = \frac{1}{1 - \rho_{\text{pois}}(n_j \lambda_{ji}, 0)} \sum_{a=a_i}^{\infty} \rho_{\text{pois}}(n_j \lambda_{ji}, a)$$

  - A low pA indicates that there are more reads of sequence *i* than can be explained by errors introduced during the amplification and sequencing of *nj* copies of sample sequence *j*.

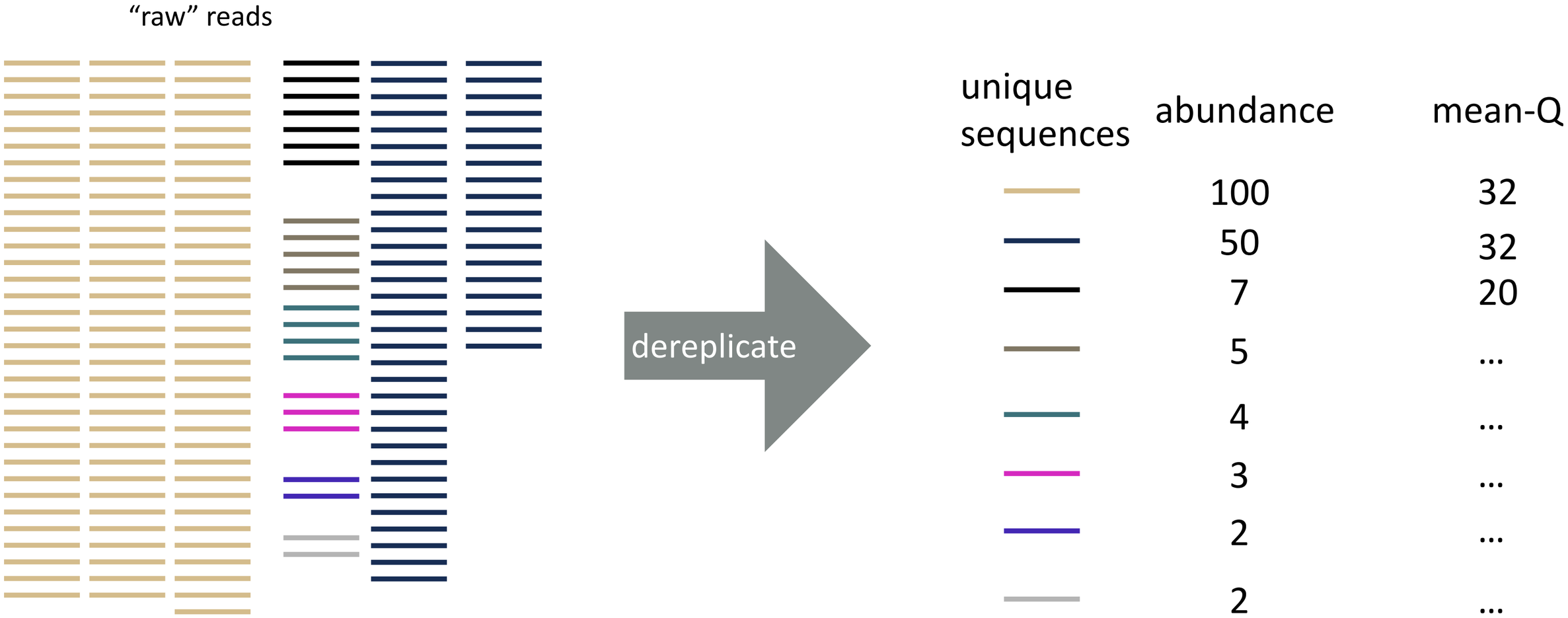
DADA2

- **The divisive partitioning algorithm.**
  - Amplicon reads with the same sequence are grouped into unique sequences with an associated abundance and consensus quality profiles (aka dereplicated).
  - The algorithm is initiated by placing all sequences in a single partition with the most abundant as the centre.
  - All unique sequences are compare to the centre
  - Calculate error rates
  - Calculate abundance p-value

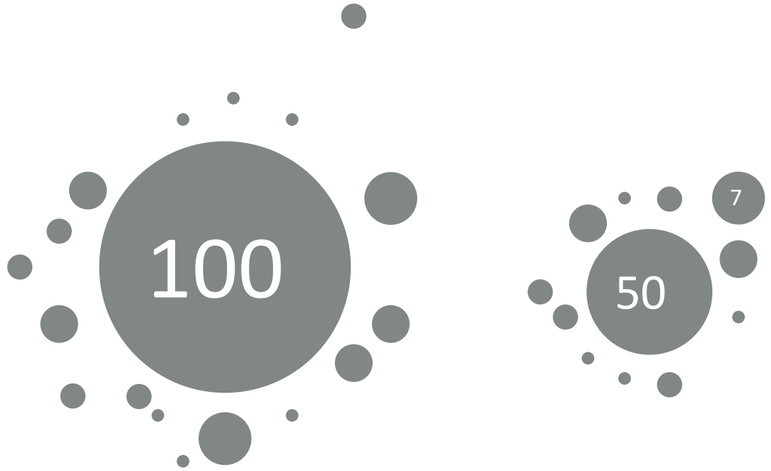
DADA2

- **The divisive partitioning algorithm.**
  - If the smallest p-value falls below a threshold (OMEGA_A= 1e-40 (default)) a new partition is formed
  - After the new partition is formed, every unique sequence is allowed to join the partition most likely to have produced it
  - Repeat until all unique sequences are consistent with being produced by amplicon sequencing the center of their partition.

DADA2 - dereplication



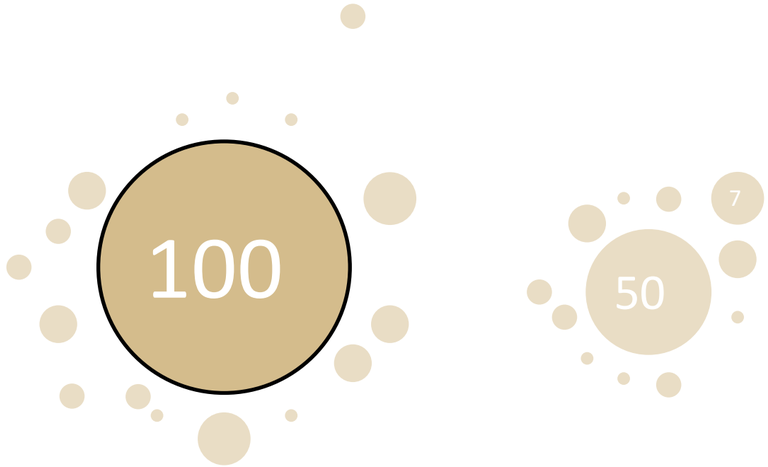
DADA2 - error model



Initial guess: one real sequence + errors

Borrowed from several online lectures, all citing:
“provided by Benjamin Callahan, not necessarily with permission"

DADA2

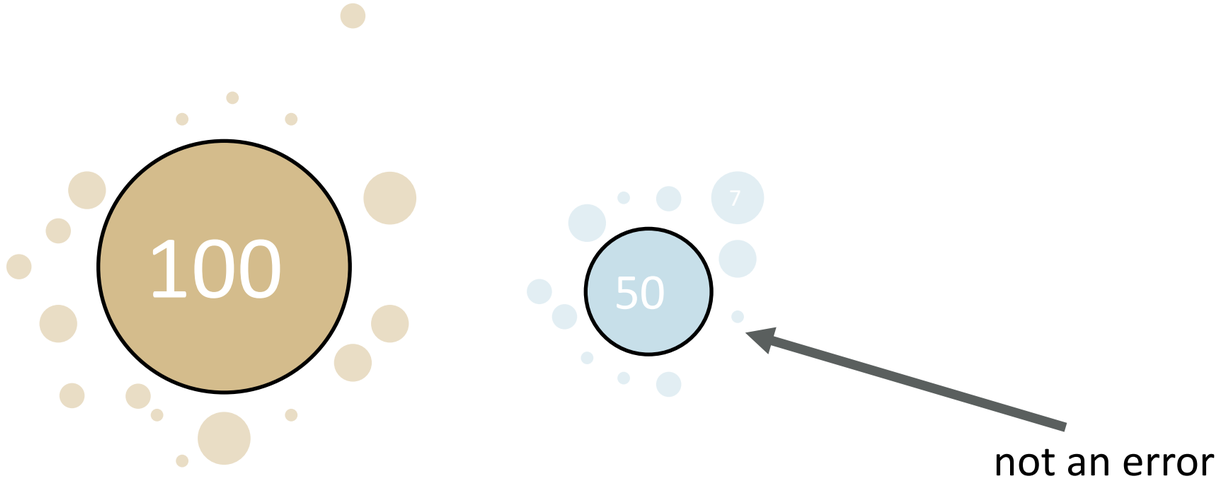


Infer initial error model under this assumption

$Pr(i \rightarrow j) =$

	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

DADA2

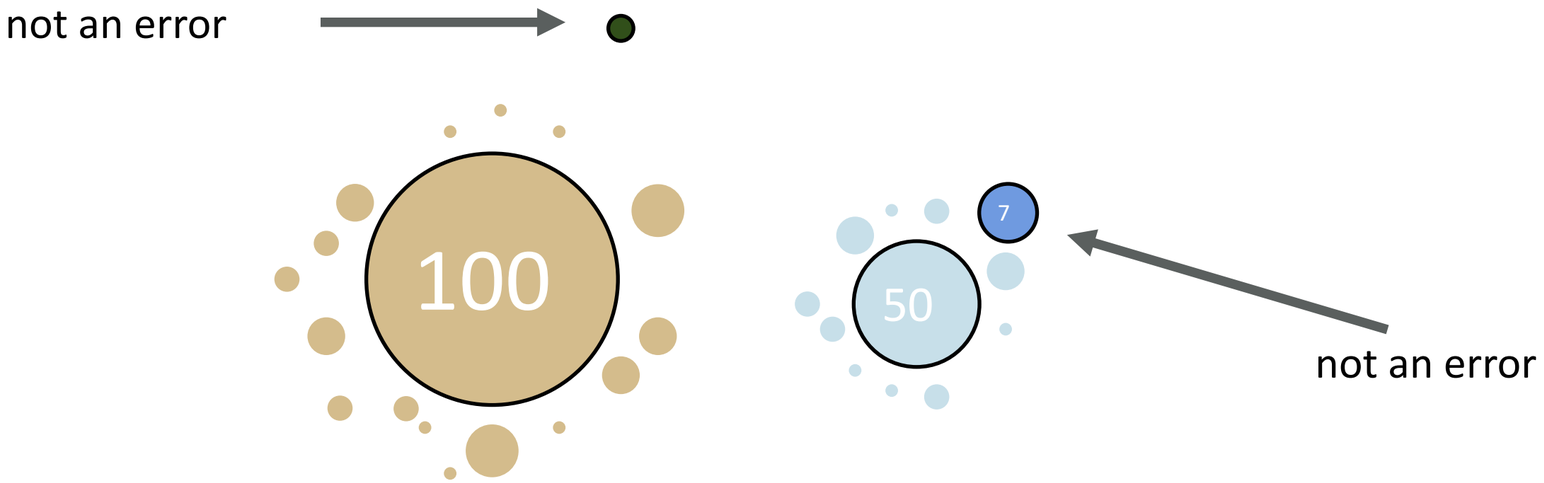


Update the model.

$Pr(i \rightarrow j) =$

	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

DADA2



Update model again

$Pr(i \rightarrow j) =$

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

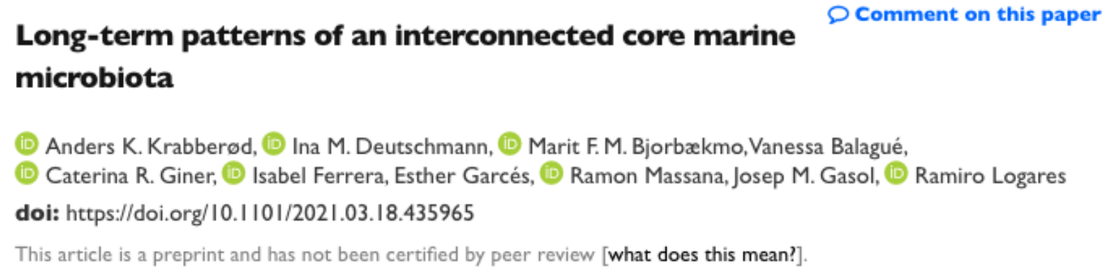
DADA2 - in R

- Main steps

Step	Function	Explanation
1	filterAndTrim()	Filters and trims an input fastq file(s) (can be compressed) based on several user-definable criteria
2	learnErrors()	Error rates are learned by alternating between sample inference and error rate estimation until convergence.
3	derepFastq()	A custom interface to FastqStreamer for dereplicating amplicon sequences from fastq or compressed fastq files,
4	dada()	The dada function takes as input dereplicated amplicon sequencing reads and returns the inferred composition of the sample (or samples).
5	mergePairs()	This function attempts to merge each denoised pair of forward and reverse reads, rejecting any pairs which do not sufficiently overlap or which contain too many (>0 by default) mismatches in the overlap region.
6	makeSequenceTable()	This function constructs a sequence table (analogous to an OTU table) from the provided list of samples.
7	removeBimeraDenovo()	screen for and remove chimeras
8	assignTaxonomy()	assignTaxonomy implements the RDP Naive Bayesian Classifier algorithm described in Wang et al. Applied and Environmental Microbiology 2007,

DADA2 – Examples

- If you want to try DADA2 with some real data:
- Scripts and setup on Github https://github.com/krabberod/BIO9905MERG1_V21
- Dataset for the run-through:
  - Selected samples from Blanes Bay Marine Observatory (BBMO) near Barcelona
  - Mini-time series: January, April, July and October for 2004 and 2005.
  - Subsample of a larger dataset: Preprint on BioRxiv:
    - doi.org/10.1101/2021.03.18.435965



Comment on this paper

Long-term patterns of an interconnected core marine microbiota

Anders K. Krabberød, Ina M. Deutschmann, Marit F. M. Bjorbækmo, Vanessa Balagué, Caterina R. Giner, Isabel Ferrera, Esther Garcés, Ramon Massana, Josep M. Gasol, Ramiro Logares

doi: https://doi.org/10.1101/2021.03.18.435965

This article is a preprint and has not been certified by peer review [what does this mean?].