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

NATURE METHODS | VOL.13 NO.7 | JULY 2016

Jeongsup Moon

0 Table of contents

DADA2 Amplicon Sequencing. Exactly. Version 1.18

Metagenome analysis

FOOD

We're Adding These Refreshing Probiotic Drinks to Our Routine ASAP—Here's Why

Dec. 7, 2021 SPONSORED

https://www.purewow.com/food/kevita-probiotic-drinks-wellness-routine

Metagenome analysis

FOOD

We're Adding These Refreshing P

BEAUTY

Why Postbiotics Work Better In Topical Skin Care Products Than Their Biotic Precursors

mbg Beauty Director
By Alexandra Engler

https://www.mindbodygreen.com/articles/why-you-should-use-postbiotics-not-probiotics-in-skin-care

Metagenome analysis

FOOD

We're Addin

P BEAUTY

A Why Pos

De Care Pro

mbg Beauty

By Alexandr

Boss Dog launches 11th product, probiotic-infused dog kibbles

in

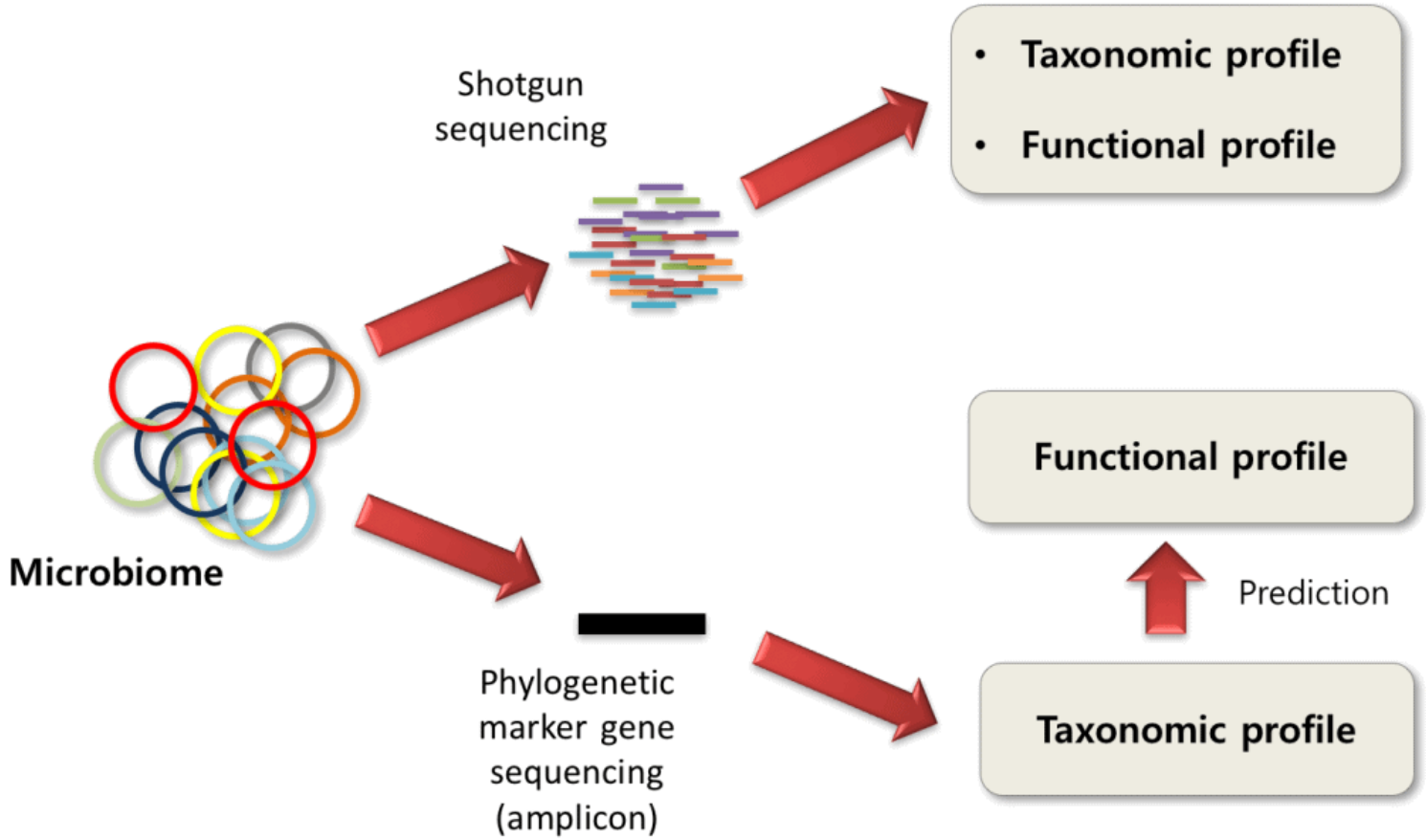
's

https://www.petfoodprocessing.net/articles/14616-boss-dog-launches-11th-product-probiotic-infused-dog-kibbles

Metagenome analysis

Q. What is metagenome?

Q. How?

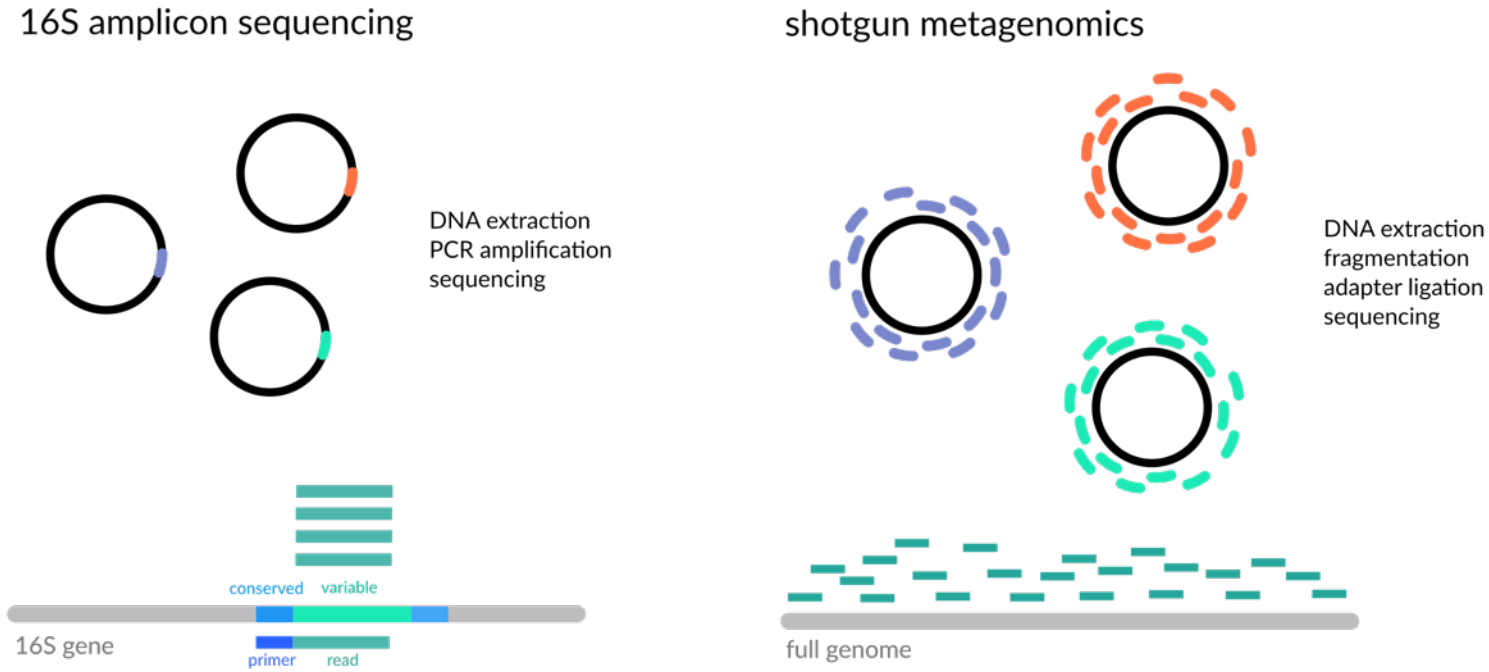


https://help.ezbiocloud.net/microbiome-basics/#microbiome-bioinformatics

Metagenome analysis

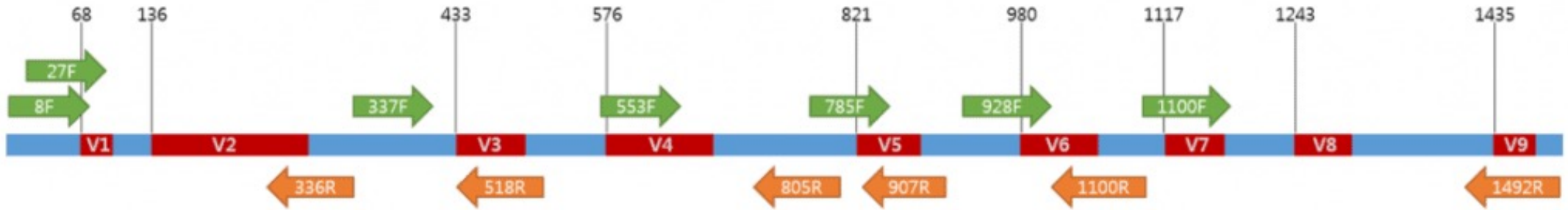
Q. What is metagenome?

Q. How?



https://github.com/Gibbons-Lab/isb_course_2021/blob/main/docs/16S/assets/16S.png

Metagenome analysis



V1-V3 ~510 bp for Roche 454

V3-V4 ~428 bp for MiSeq PE

V3-V5 ~548 bp for Roche 454

V4 ~252 bp for HiSeq

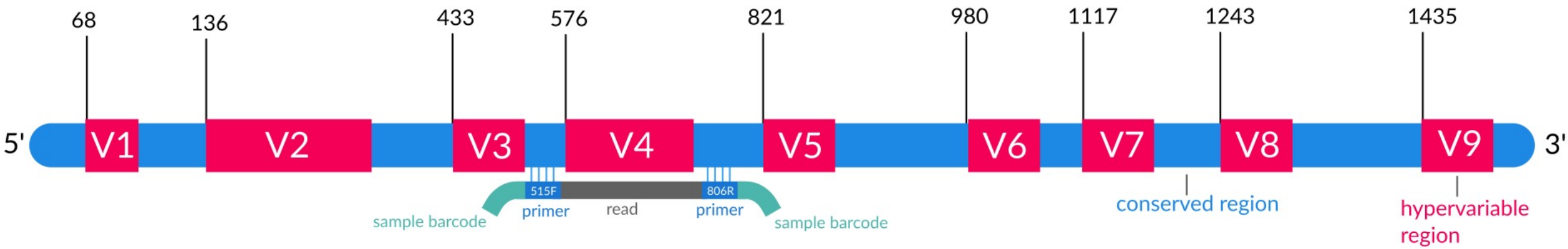
~562 bp for Roche 454 V6-V9

V1-V9 (Full-length)

Pacific Biosciences

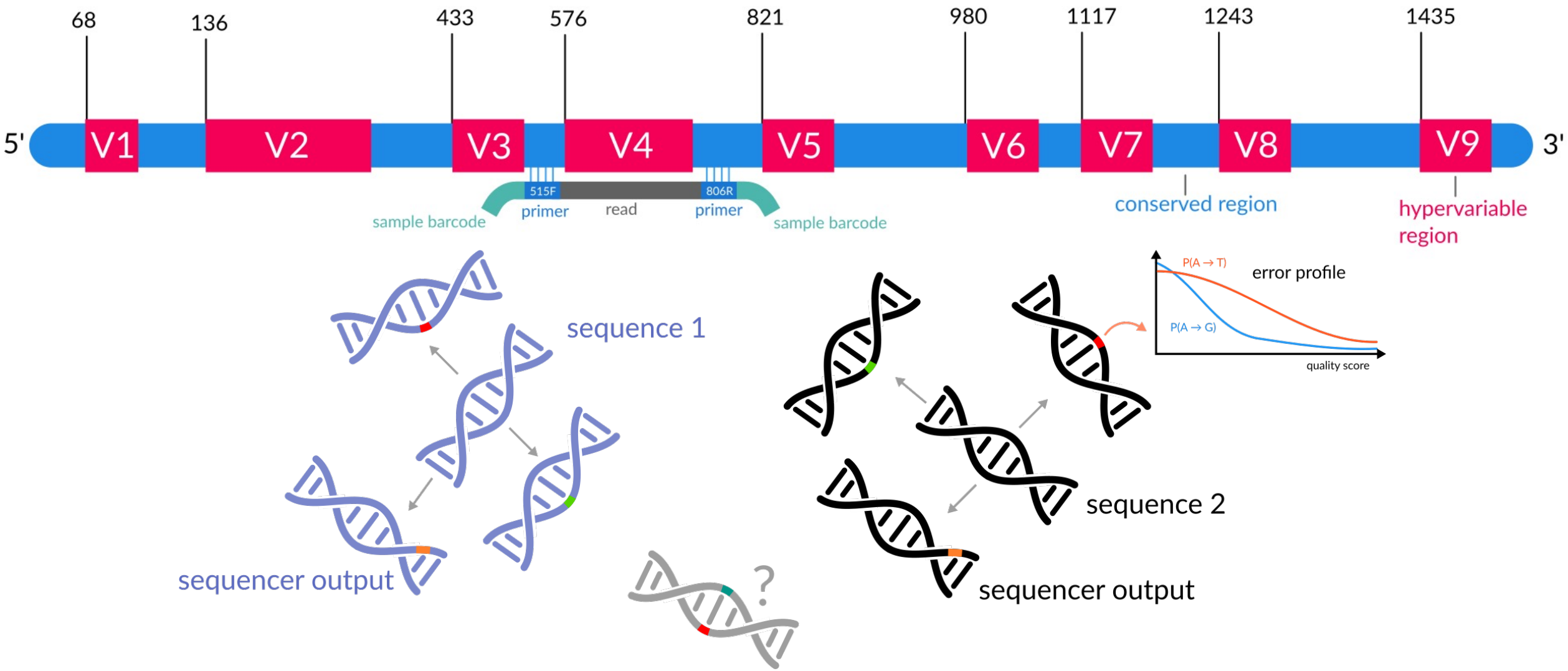
https://help.ezbiocloud.net/16s-rrna-and-16s-rrna-gene/

Metagenome analysis



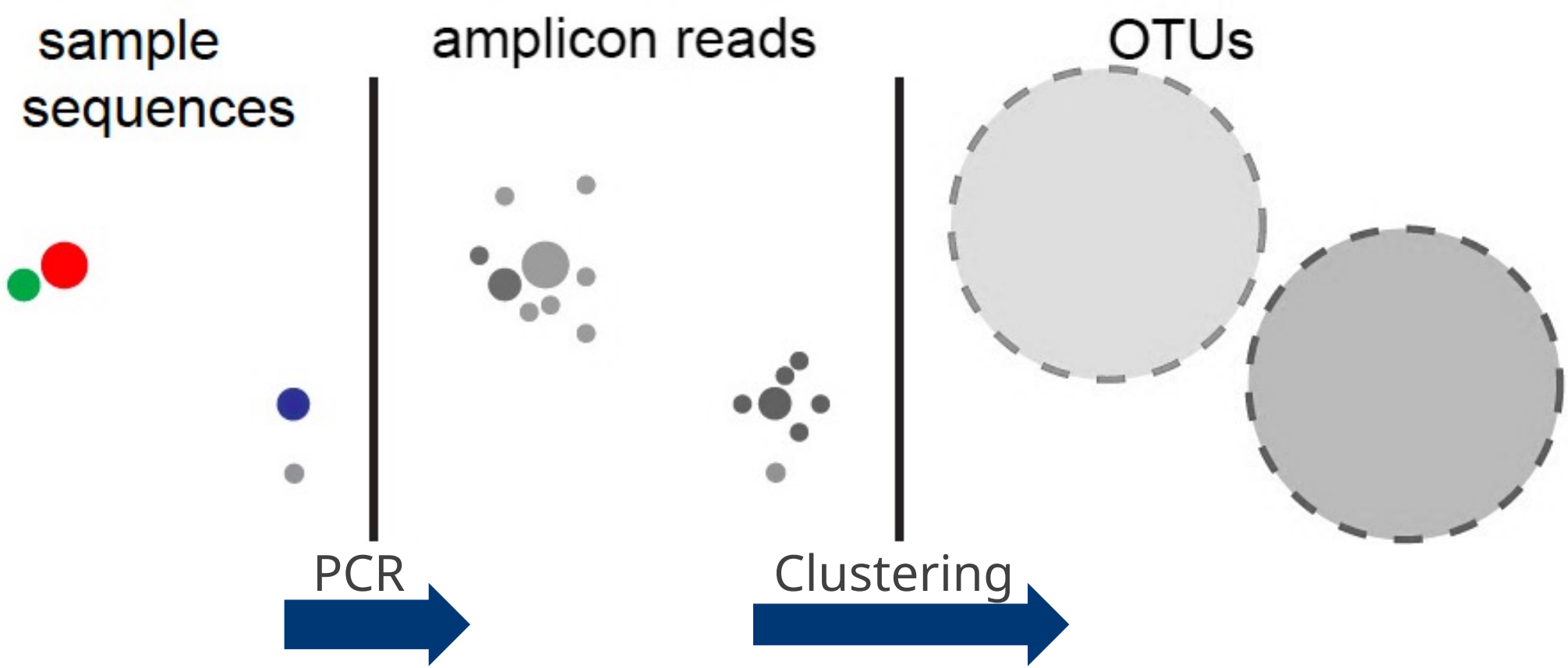
https://github.com/Gibbons-Lab/isb_course_2021/blob/main/docs/16S/assets/16S_gene.png

Limitation



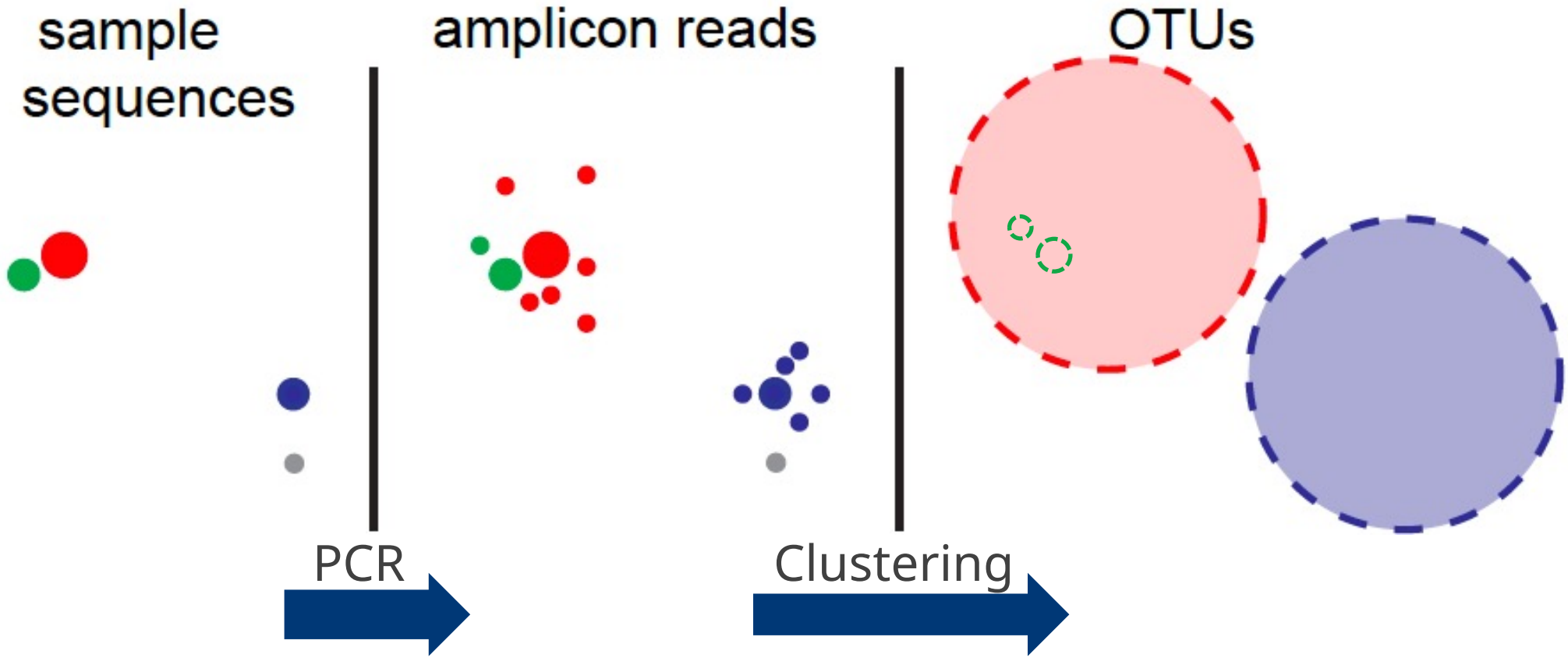
https://github.com/Gibbons-Lab/isb_course_2021/blob/main/docs/16S/assets/dada2.png

Limitation



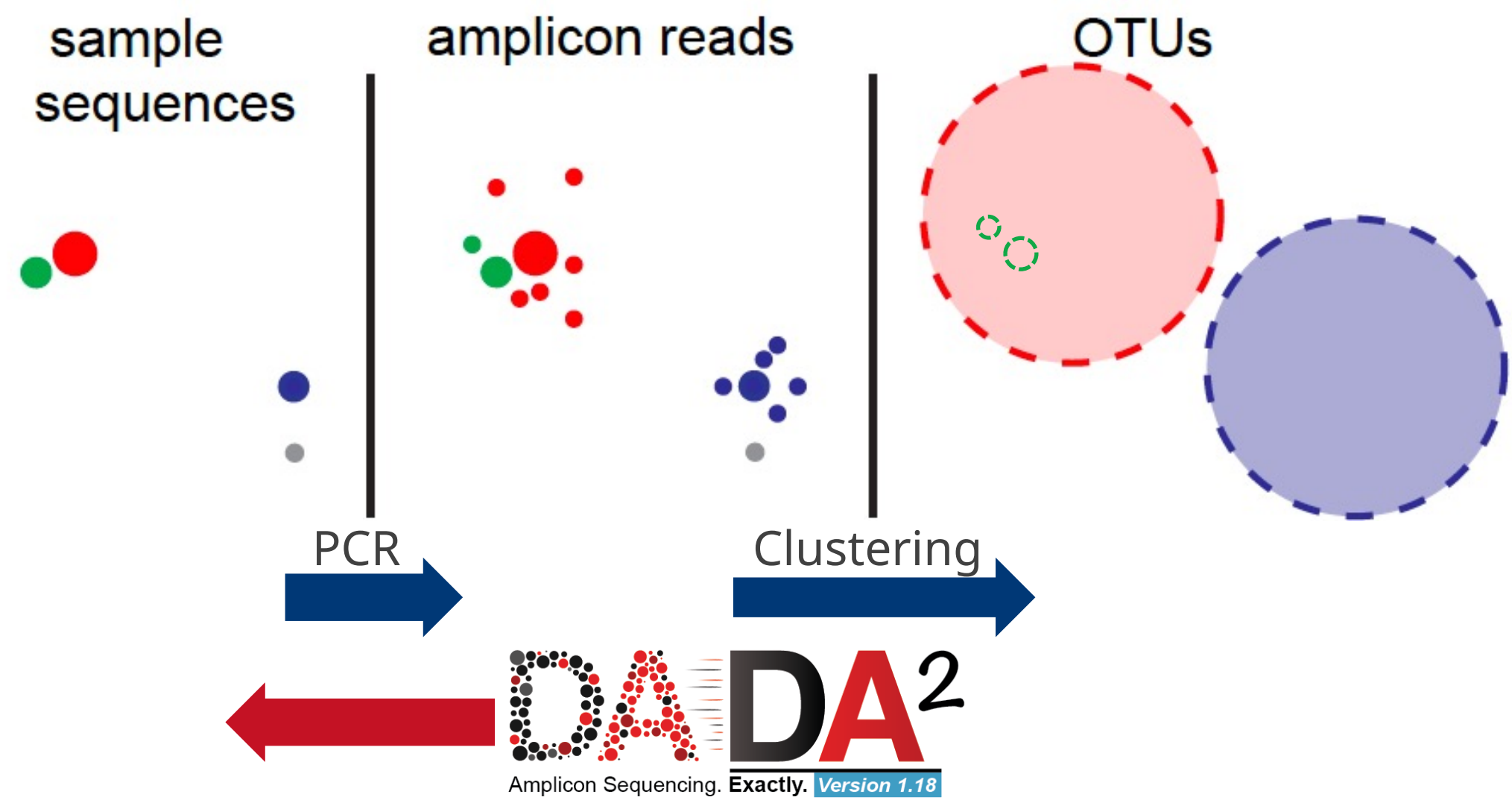
adapted from Supplementary Figure 1. Callahan, Benjamin J et al. “DADA2: High-resolution sample inference from Illumina amplicon data.” Nature methods vol. 13,7 (2016)

Limitation



adapted from Supplementary Figure 1. Callahan, Benjamin J et al. “DADA2: High-resolution sample inference from Illumina amplicon data.” Nature methods vol. 13,7 (2016)

Limitation



Supplementary Figure 1. Callahan, Benjamin J et al. “DADA2: High-resolution sample inference from Illumina amplicon data.” Nature methods vol. 13,7 (2016)

DADA2: Two baseline assumptions

1. DADA2 models errors:
   - Independently within a read
   - Independantly between reads

2. DADA2 models abundance:
   - number of sequence ***i*** that is produced from sequence ***j*** will be Poisson distributed.

adapted from Figure 2. Rosen, M.J. et al. “Denoising PCR-amplified metagenome data.” BMC Bioinformatics 13, 283 (2012)

DADA2: Denoising

DADA2 models errors:
- Independently within a read
- Independantly between reads

error rate

$$\lambda_{ji} = \prod_{l=0}^{L} p(j(l) \rightarrow i(l), qi(l))$$

Probability of *L*th nt in sequence *j* change to *L*th in sequence *i*

Based on the quality

true sequence

j : CCAACGGGAGACAGCAGTGGGGAA

error sequence

i : CCTACGGGAGGCAGCAGTGGGGAA

Fastq File
- Seq_id: @M00763:36:000000000-A8T0A:1:1101:1
- sequence : CCTACGGGAGGCAGCAGTGGGGAATATT
- delim: +
- quality: 88CCCGDBAF)===CEFFGGGG>GGGGGG

DADA2: Denoising

DADA2 models errors:
- Independently within a read
- Independantly between reads

error rate

$$\lambda_{ji} = \prod_{l=0}^{L} p(j(l) \rightarrow i(l), qi(l))$$

Probability of *L*th nt in sequence *j* change to *L*th in sequence *i*

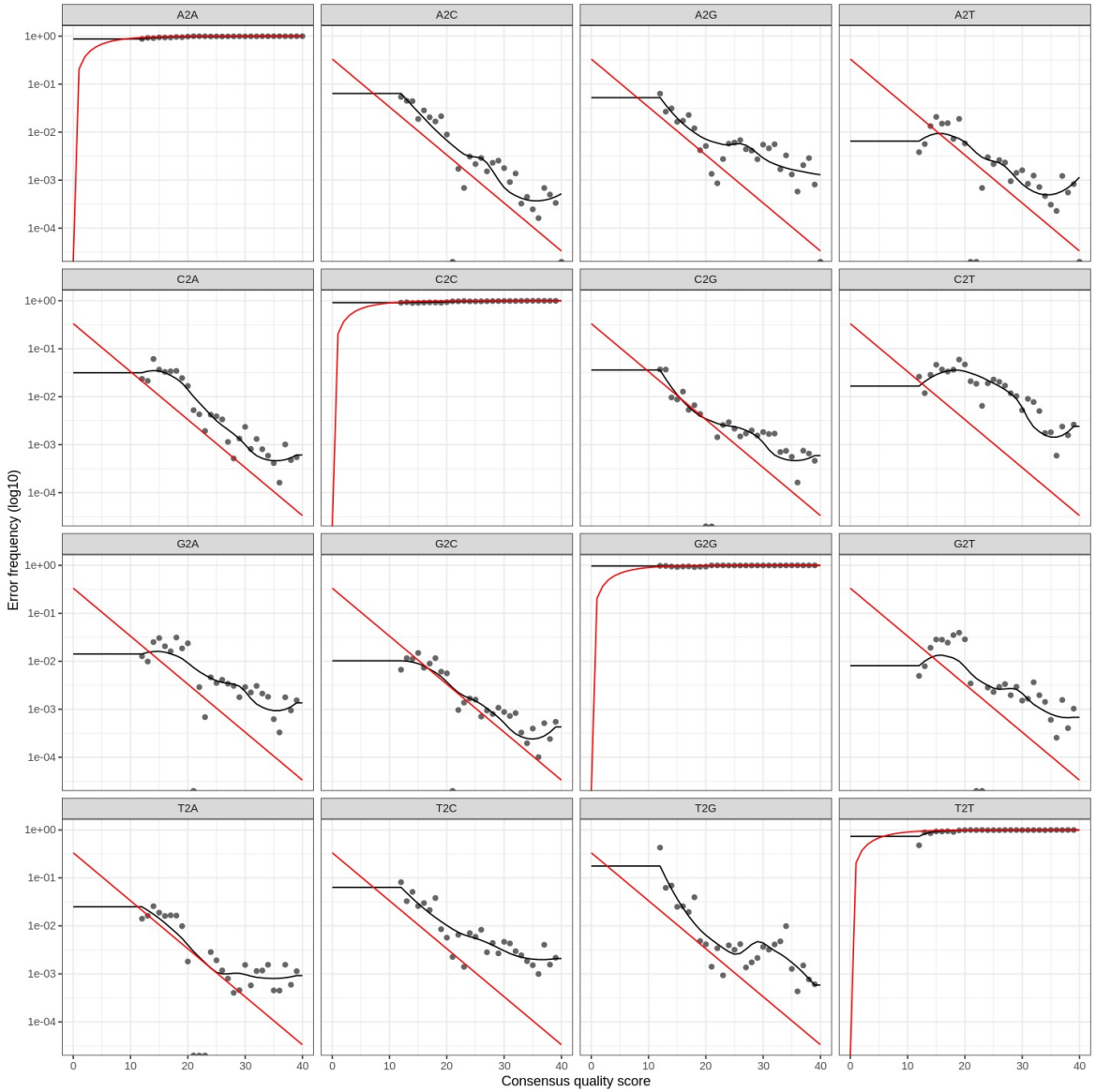
Based on the quality

true sequence

j : CCAACGGGAGACAGCAGTGGGGAA

error sequence

i : CCTACGGGAGGCAGCAGTGGGGAA



DADA2: Denoising

DADA2 models abundance:

- number of sequence ***i*** that is produced from sequence ***j*** will be Poisson distributed

rate of events

$$\mathrm{P}(\underbrace{\mathrm{X} = k}_{\text{observe } k \text{ events}}) = \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!}$$

DADA2 models abundance probability:

error rate

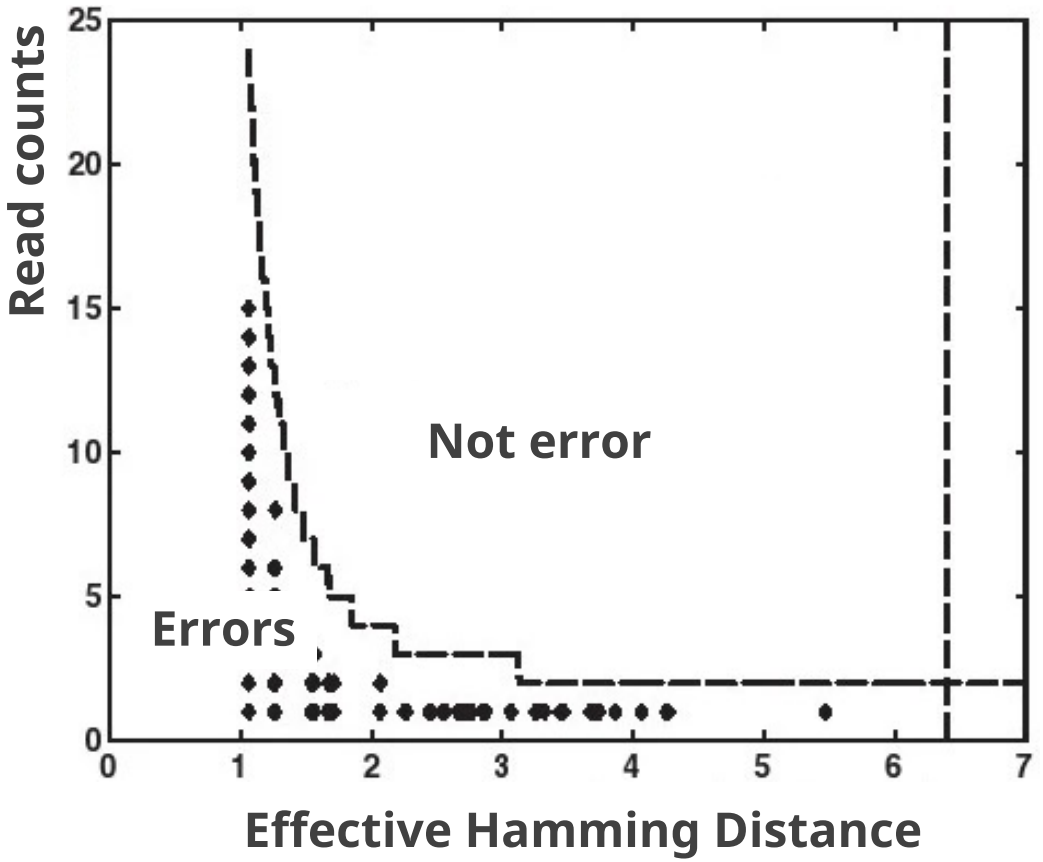
# of sequence *i*

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

# of sequence *j*

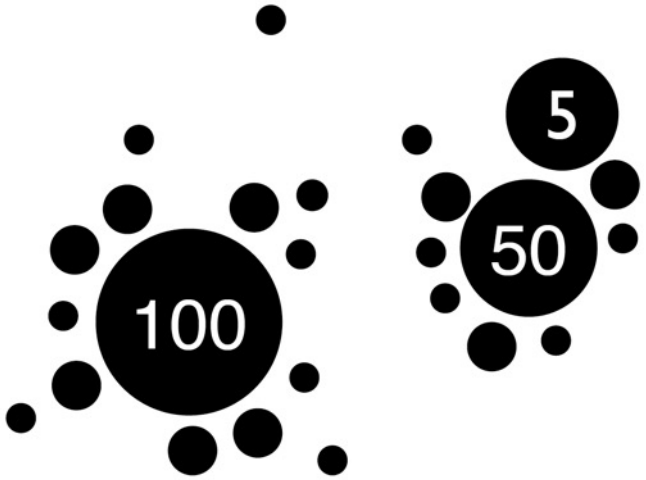
Hamming Distance: 2

j : CCAACGGGAGACAGCAGTGGGGAA
i : CCTACGGGAGGCAGCAGTGGGGAA



adapted from Figure 2. Rosen, M.J. et al. "Denoising PCR-amplified metagenome data." BMC Bioinformatics 13, 283 (2012)

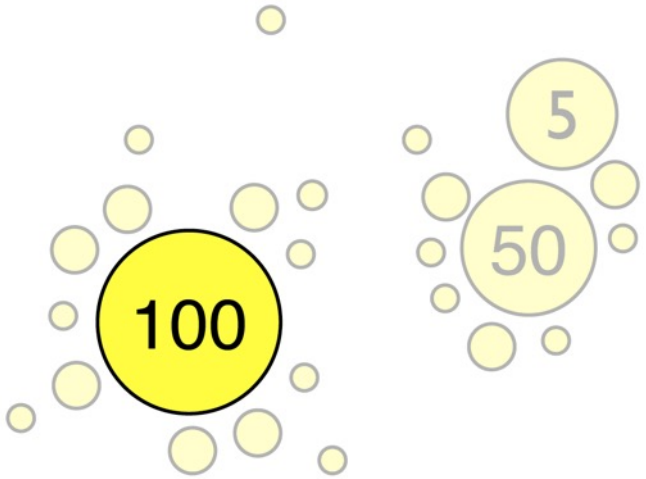
DADA2: Example



1. Assumes : The most abundant sequence as **the only real sequence *j***, and others **as errors *i*.**

https://si.biostat.washington.edu/sites/default/files/modules//2016-SISMID-14-01.pdf

DADA2: Example

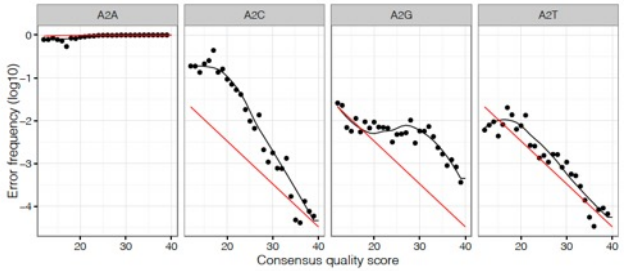


q_{40}	A	C	G	T
A	0.97	0.01	0.01	0.01
C	0.01	0.97	0.01	0.01
G	0.01	0.01	0.97	0.01
T	0.01	0.01	0.01	0.97

* This is a simplified version of 16 * 40 table.
* Note that there are 40 different q values.

1. Assumes : The most abundant sequence as **the only real sequence *j***, and others **as errors *i*.**

2. Learns error model parameter(transition probability).
   Learns error model, and calculate error rate.
   If mismatch > 10%, λ_{ji} is set to 0.



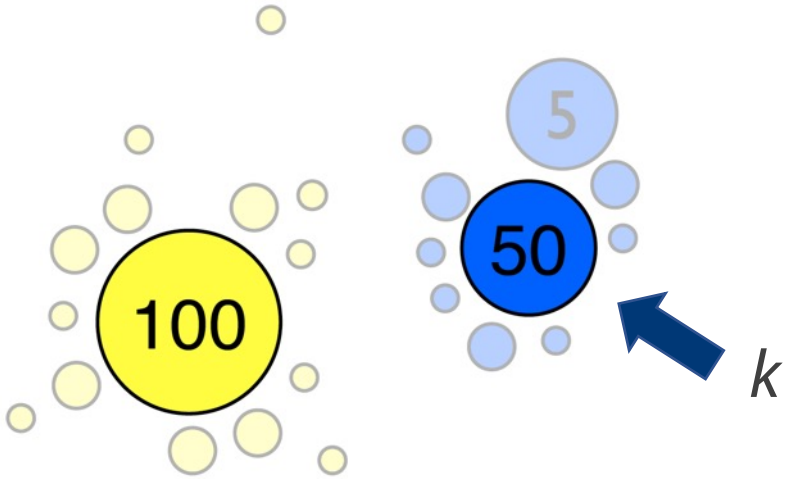
$$\lambda_{ji} = \prod_{l=0}^{L} p(j(l) \rightarrow i(l), qi(l))$$

3. Calculates abundance p value.

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

https://si.biostat.washington.edu/sites/default/files/modules//2016-SISMID-14-01.pdf

DADA2: Example



q_{40}	A	C	G	T
A	0.997	0.001	0.001	0.001
C	0.001	0.997	0.001	0.001
G	0.001	0.001	0.997	0.001
T	0.001	0.001	0.001	0.997

* This is a simplified version of 16 * 40 table.
* Note that there are 40 different q values.

3. Calculates abundance p value.

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

4. Select lowest abundance p value, and set it as another real **sequence *k***.

5. Align errors *i* to where they will have most expected number. $E[\rho_{pois}(n_j; \lambda_{ji})]$ or $E[\rho_{pois}(n_k; \lambda_{ki})]$

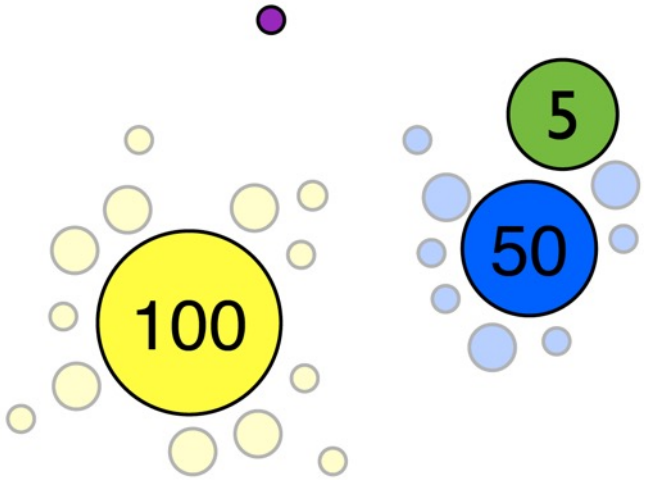
6. Update error models

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

$$\lambda_{ki} = \prod_{l=0}^{L} p(k(l) \to i(l), qi(l))$$

https://si.biostat.washington.edu/sites/default/files/modules//2016-SISMID-14-01.pdf

DADA2: Example



q_{40}	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	$2x10^{-4}$	$6x10^{-3}$	$1x10^{-4}$	0.998

* This is a simplified version of 16 * 40 table.
* Note that there are 40 different q values.

3. Calculates abundance p value.

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

4. Select lowest abundance p value, and set it as another real **sequence *k***.

5. Align errors *i* to where they will have most expected number. $E[\rho_{pois}(n_j; \lambda_{ji})]$ or $E[\rho_{pois}(n_k; \lambda_{ki})]$

6. Update error models

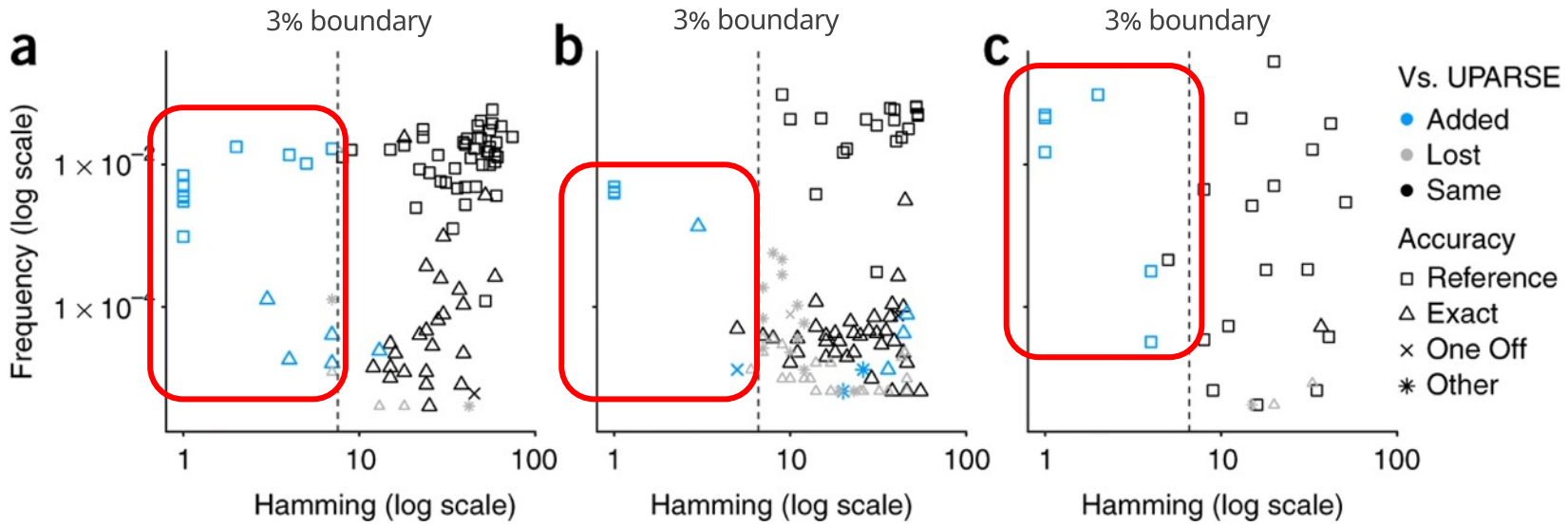
$$\lambda_{ji} = \prod_{l=0}^{L} p(j(l) \to i(l), qi(l))$$

7. Do step 3 to 6, until all abundance p–value are above threshold($1e^{-40}$).

https://si.biostat.washington.edu/sites/default/files/modules//2016-SISMID-14-01.pdf

Mock-up Community

Comparison of sequence variants inferred by DADA2 with OTUs constructed by UPARSE.



Balanced

57 bacteria and archea with nominally equal frequencies

HMP

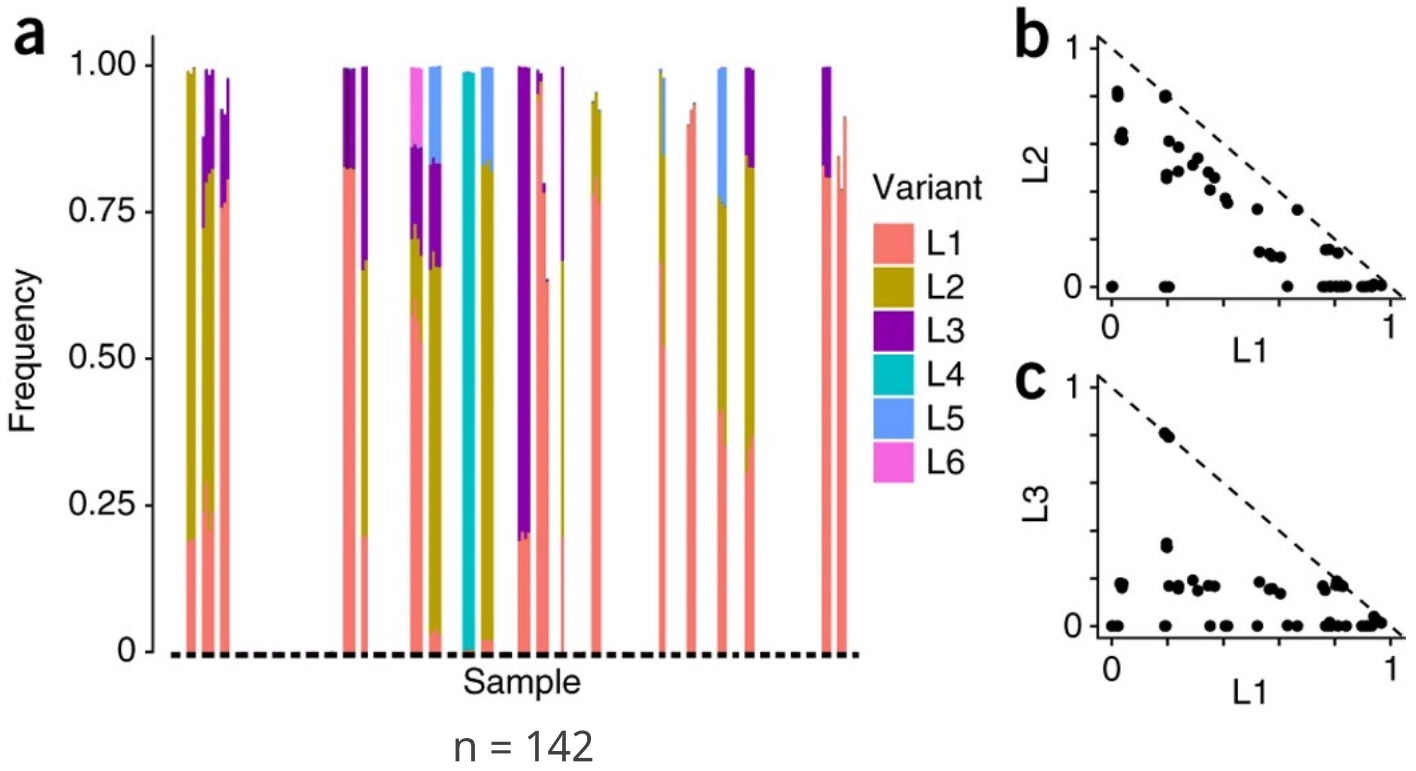
21 bacterial with equal frequencies

Extreme

27 bacterial with different frequencies

adapted from Figure 1. Callahan, Benjamin J et al. “DADA2: High-resolution sample inference from Illumina amplicon data.” Nature methods vol. 13,7 (2016)

Frequency of Lactobacillus *crispatus* (Strain level)



n = 142

adapted from Figure 2. Callahan, Benjamin J et al. “DADA2: High-resolution sample inference from Illumina amplicon data.” Nature methods vol. 13,7 (2016)

Limitation

Strains missed by DADA2	Total Reads	Max Abundance	Hamming to NN	NN Abundance
Prevotella buccalis	5	1	51	9
Clostridium methylpentosum DSM 5476	5	1	25	13
Clostridium phytofermentans ISDg	11	2	15	59564
Parabacteroides sp. D13	3	2	1	28242

Too many # of sequence *j*

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

Singleton removed

Colab

https://colab.research.google.com/drive/1CaLLzB1gEEuMk4P8J_siDFgLUX1d3gor?usp=sharing

Thank you