



# Escuela Regional de Ecología Microbiana de Sistemas Acuáticos Edición 2023

Grupo de Ecología Microbiana de Sistemas Acuáticos,  
Centro Universitario Regional del Este,  
Universidad de la República

# **Bioinformática aplicada al análisis de datos de metabarcoding de ADN ambiental**

Teórico 2

5 de diciembre 2023

Docentes: Paula Huber, Daiana Mir, Luciana Griffero, Cecilia Alonso, Juan Zanetti, Emiliano Pereira

# **Temario**

- 1. Bases de datos de genes marcadores.**
- 1. Algoritmos de clusterización de secuencias.**

# Temario

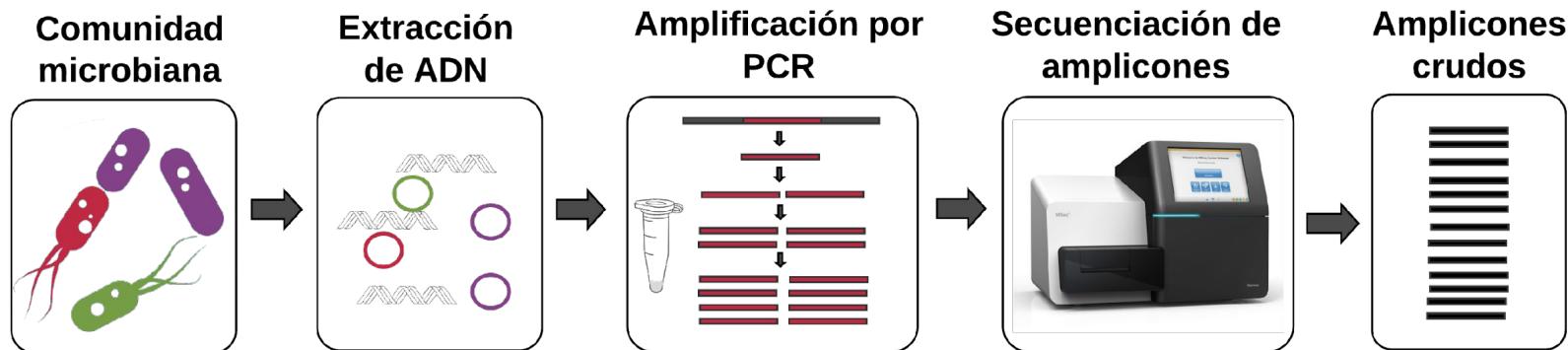
- 1. Bases de datos de genes marcadores.**
- 1. Algoritmos de clusterización de secuencias.**

# Temario

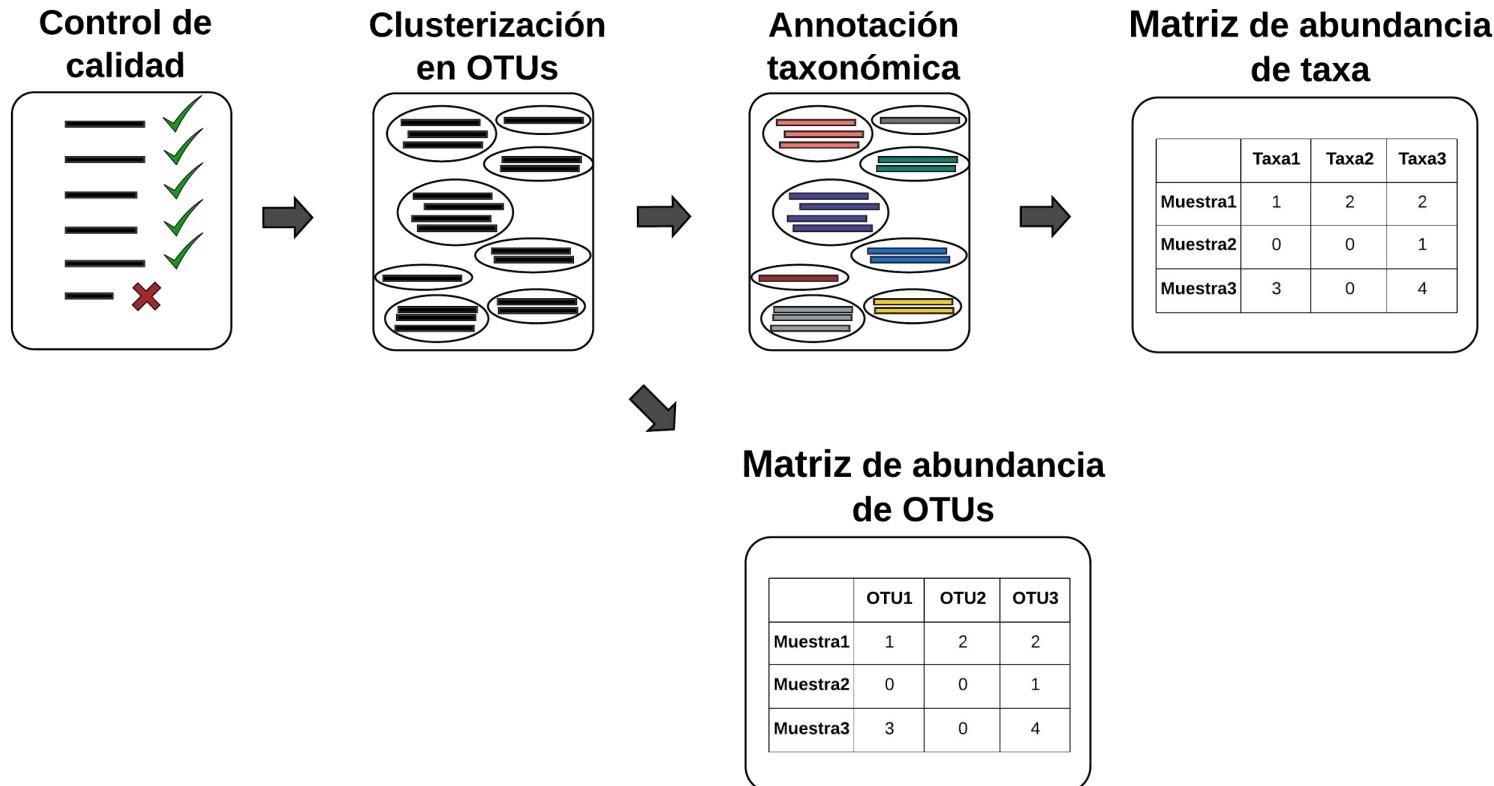
1. Bases de datos de genes marcadores.
1. Algoritmos de clusterización de secuencias.

# **Algoritmos de clusterización de secuencias**

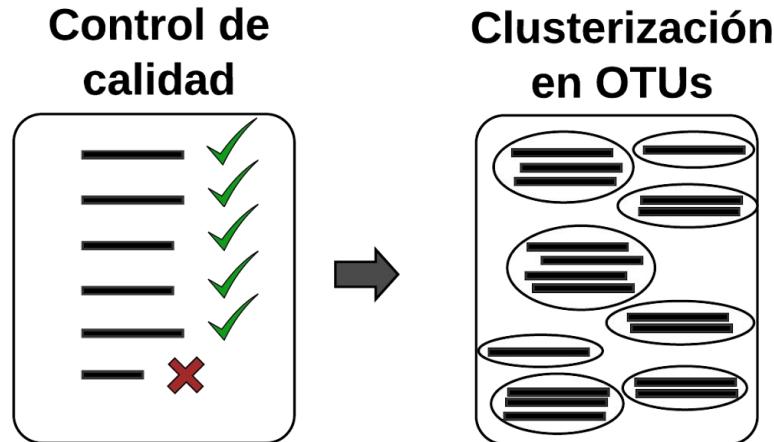
# Punto de partida



# Procesamiento de datos

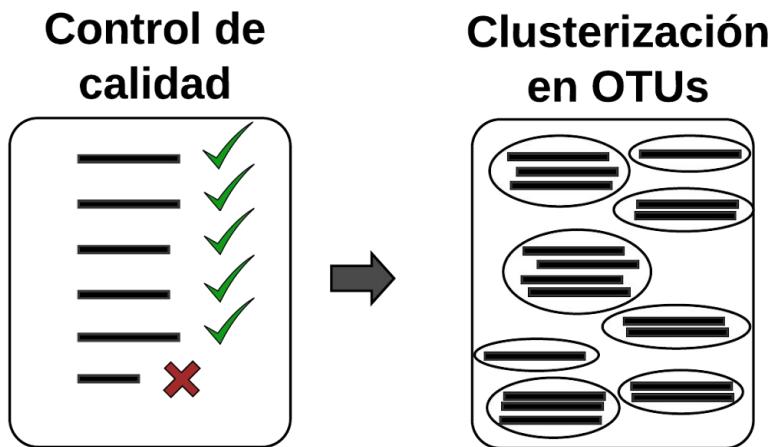


# 1. Procesamiento de datos: clusterización de secuencias



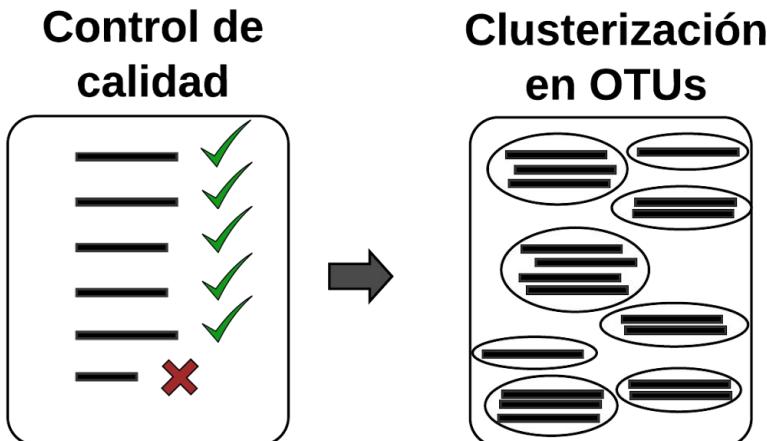
# Procesamiento de datos: clusterización de secuencias

- Agrupar las secuencias en unidades taxonómicas operacionales (OTUs) de acuerdo su identidad.



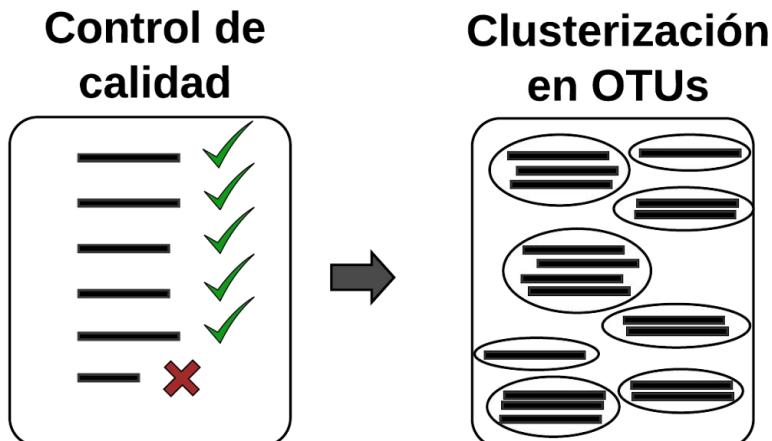
# Procesamiento de datos: clusterización de secuencias

¿Por qué clusterizar los datos de amplicones?



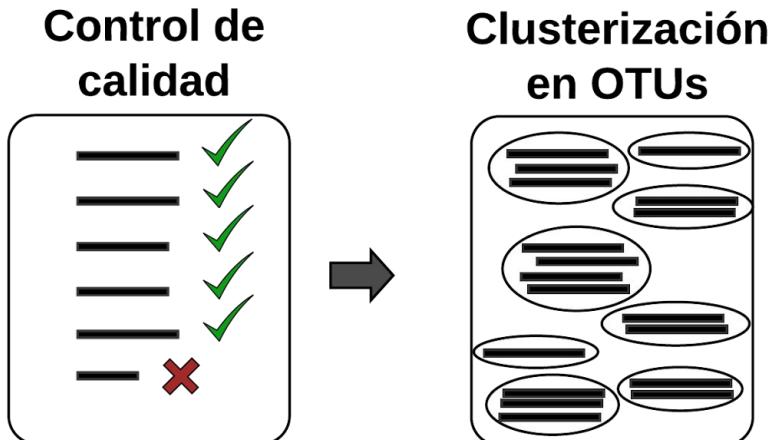
# Procesamiento de datos: clusterización de secuencias

¿Por qué clusterizar los datos de amplicones?



- Darle sentido a grandes volúmenes de datos.
- Estudiar la estructura y diversidad de las comunidades.
- Ignorar (o corregir en el caso de ASVs) la variabilidad artefactual.

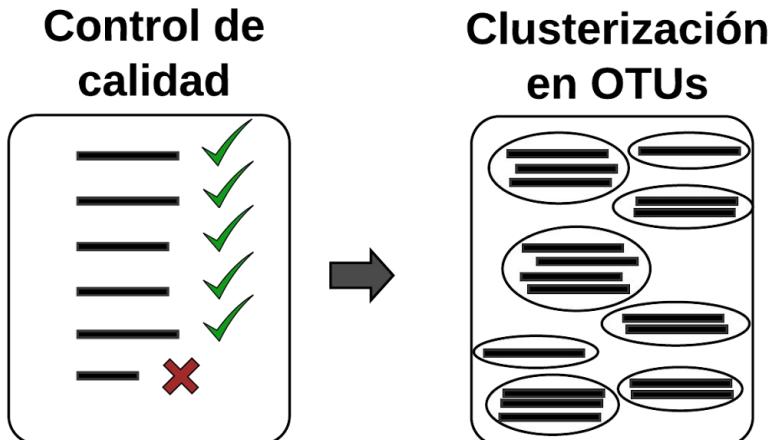
# Procesamiento de datos: clusterización de secuencias



Tres grandes tipos de clusterización:

- De referencia - cerrado
- De referencia - abierto
- De novo

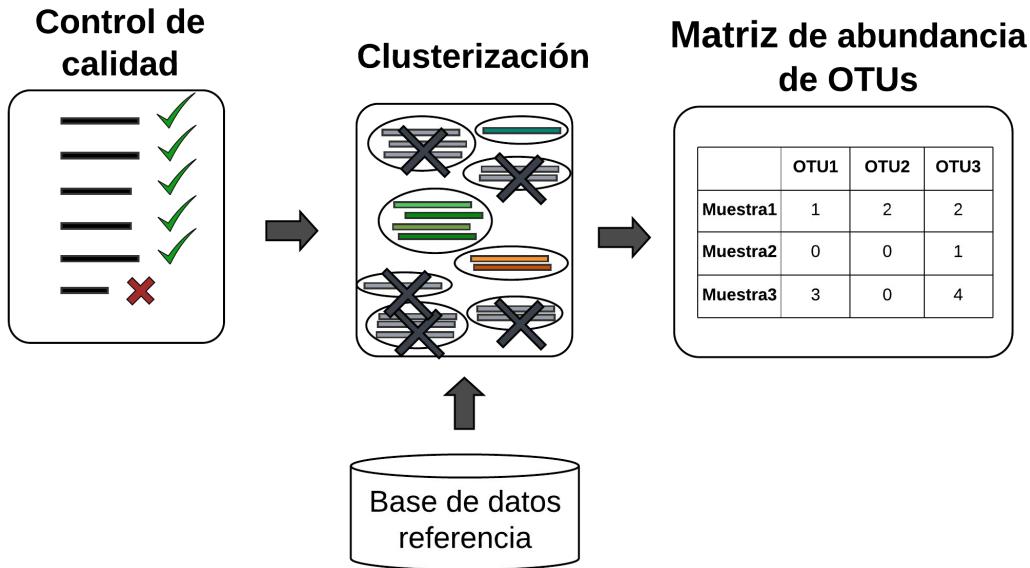
# Procesamiento de datos: clusterización de secuencias



Tres grandes tipos de clusterización:

- De referencia - cerrado
- De referencia - abierto
- **De novo** 

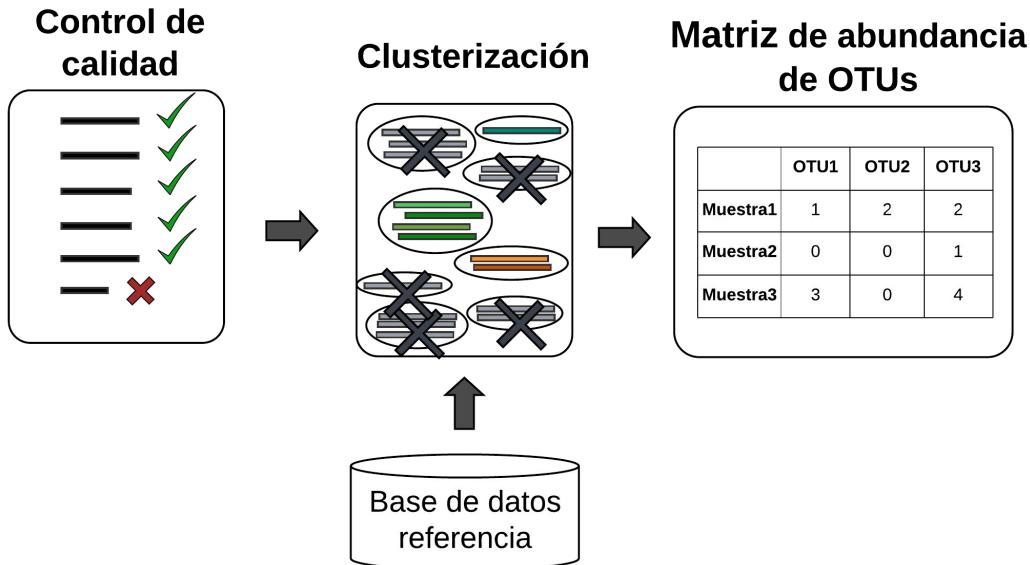
# De referencia - cerrado



## Metodología

- Las secuencias son comparadas contra una base de datos referencia.
- Las secuencias que superan cierto nivel de similitud con una misma secuencia referencia pasan a formar un OTU.

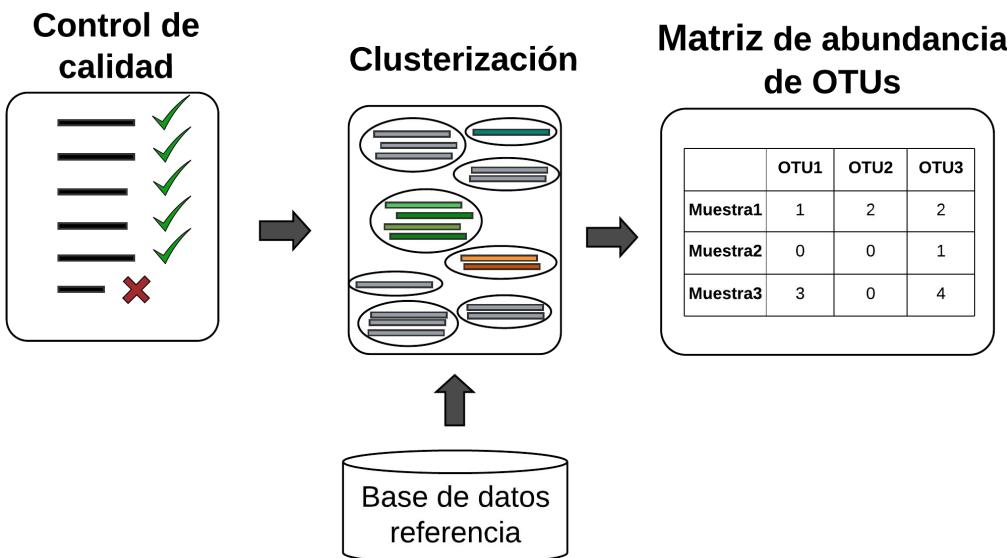
# De referencia - cerrado



## Consideraciones

- Dependiente de la base de datos referencia.
- Secuencias sin representantes cercanos no pueden ser clusterizadas.
- Comúnmente utilizado en la anotación taxonómica.

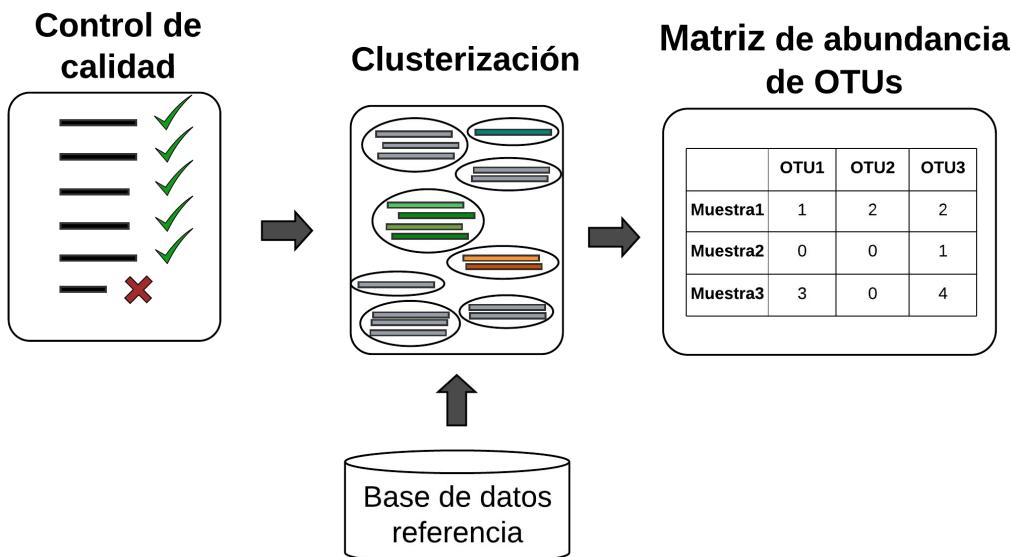
# De referencia - abierto



## Metodología

- Las secuencias son comparadas contra una base de datos referencia.
- Las secuencias que superan cierto nivel de similitud con una misma secuencia referencia pasan a formar un OTU.
- Las secuencias que no tiene referentes cercanos son clusterizadas de novo.

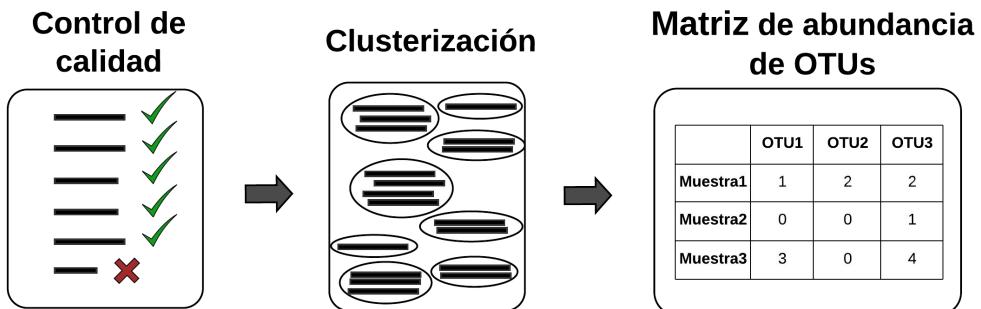
# De referencia - abierto



## Consideraciones

- Combina las fortalezas de novo y de referencia – cerrado.
- Tiene la complicación adicional de que integra OTUs de diferente naturaleza.

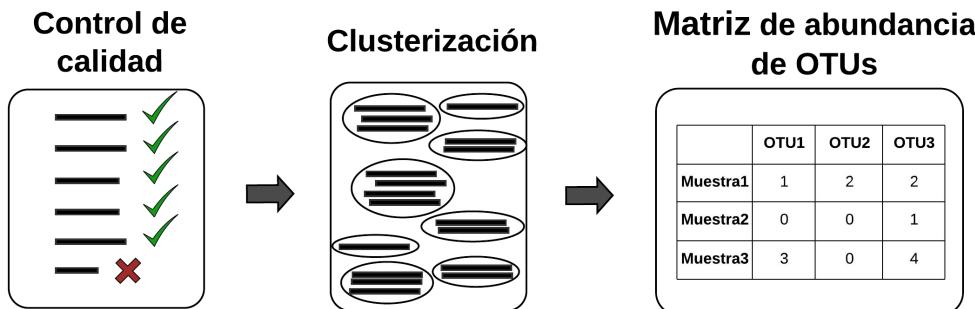
# De novo



## Metodología

- Las secuencias son comparadas entre sí para determinar sus similitudes.
- En base a su similitud, y a la aplicación de distintos tipos de algoritmos, estas son clusterizadas.

## De novo



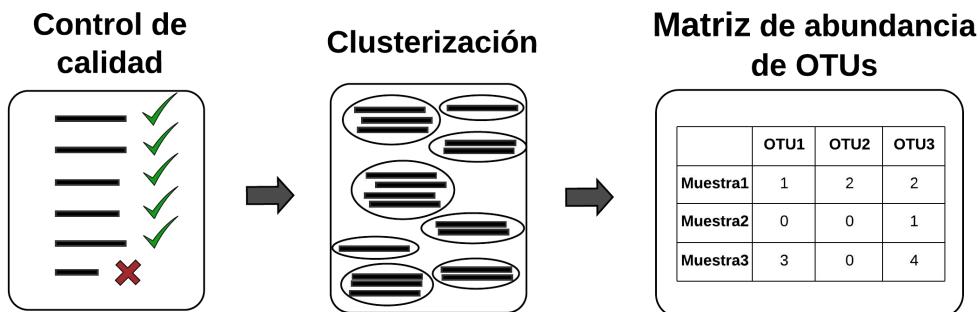
### Consideraciones

- Independencia de bases de datos referencia.
- Mejor performance.
- Ampliamente utilizados.

# De novo



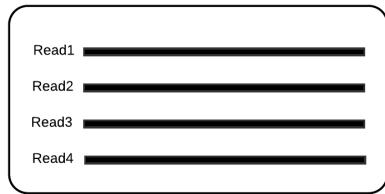
# De novo



- {
- Clusterización jerárquica
  - **Clusterización heurística**

# De novo: clusterización jerárquica

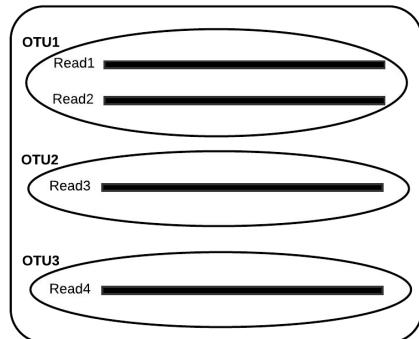
## 1) Amplicones



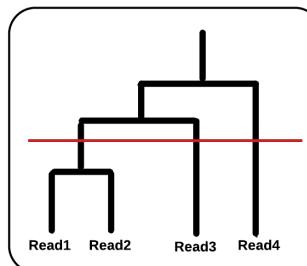
## 2) Matriz de distancias

	Read1	Read2	Read3	Read4
Read1	0	1	4	6
Read2		0	4	5
Read3			0	7
Read4				0

## 4) Clusterización

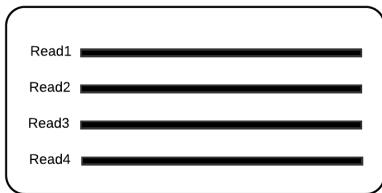


## 3) Árbol jerárquico



# De novo: clusterización jerárquica

## 1) Amplicones

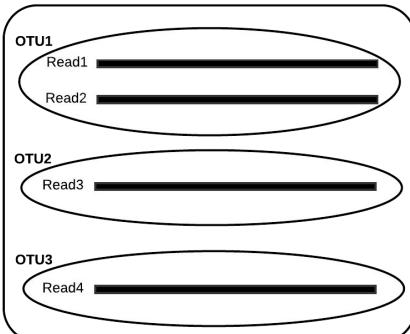


## 2) Matriz de distancias

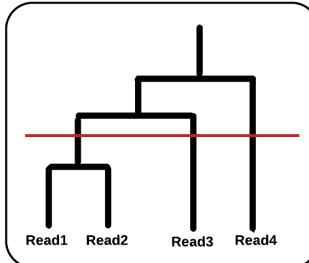
	Read1	Read2	Read3	Read4
Read1	0	1	4	6
Read2		0	4	5
Read3			0	7
Read4				0

- Proceso computacionalmente muy caro.
- Complejidad computacional es  $O(N^2)$ , donde N es el número de secuencias.

## 4) Clusterización

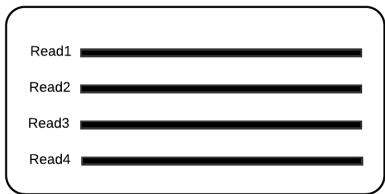


## 3) Árbol jerárquico



# De novo: clusterización jerárquica

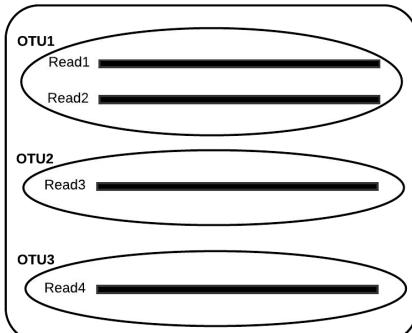
## 1) Amplicones



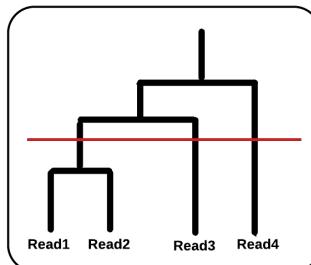
## 2) Matriz de distancias

	Read1	Read2	Read3	Read4
Read1	0	1	4	6
Read2		0	4	5
Read3			0	7
Read4				0

## 4) Clusterización



## 3) Árbol jerárquico



- Al comienzo cada secuencias es un cluster.
- Estas se agrupan de acuerdo a la matriz de distancias.
- Comúnmente aplicando *complete-linkage*, *average-linkage*, o *single-linkage*.

## **De novo: clusterización heurística**

- ¿Qué es una aproximación heurística?

## **De novo: clusterización heurística**

- Heurística: aproximación donde se emplea un método práctico para resolver un problema que no garantiza una solución óptima, pero sí lo “suficientemente buena”.

## De novo: clusterización heurística

- Heurística: aproximación donde se emplea un método práctico para resolver un problema que no garantiza una solución óptima, pero sí lo “suficientemente buena”.



- Las secuencias son procesadas una por una, generalmente, aplicando una estrategia de clusterización incremental codiciosa (*greedy*).

## De novo: clusterización heurística

- Heurística: aproximación donde se emplea un método práctico para resolver un problema que no garantiza una solución óptima, pero sí lo “suficientemente buena”.



- Las secuencias son procesadas una por una, generalmente, aplicando una estrategia de clusterización incremental codiciosa (*greedy*)



- Es evitado el cómputo de distancias de todas las secuencias por pares.

# Procesamiento de datos: clusterización de secuencias

## Clusterización:

De referencia - abierto

De referencia – cerrado

**De novo:** 

Jerárquica

**Heurística:** 

**VSEARCH**

DADA2

# Procesamiento de datos: clusterización de secuencias

Tres herramientas dedicadas a la clusterización de datos de amplicones:

- DADA2 (12,848 citas)
- VSEARCH (2,778 citas)

# VSEARCH: a versatile open source tool for metagenomics

Torbjørn Rognes<sup>1,2</sup>, Tomáš Flouri<sup>3,4</sup>, Ben Nichols<sup>5</sup>, Christopher Quince<sup>5,6</sup> and Frédéric Mahé<sup>7,8</sup>

## VSEARCH

- VSEARCH es una herramienta multiproceso (i.e., *multithreaded*) de código abierto y gratuita que integra varias funciones para procesar y analizar datos de secuencias de nucleótidos (por puesto incluyendo las clusterización de secuencias).
- Alternativa a USEARCH, la cual no es de código abierto y sólo tiene una versión gratuita más limitada.

# VSEARCH

- Algoritmo de clusterización: aproximación heurística.

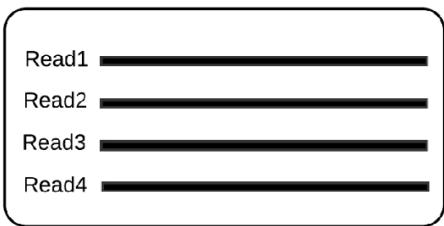
*“Greedy and heuristic centroid-based algorithm”*

# De novo: clusterización heurística

- Estructura del algoritmo

1)

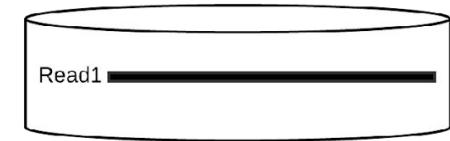
## Amplicones



## OTUs



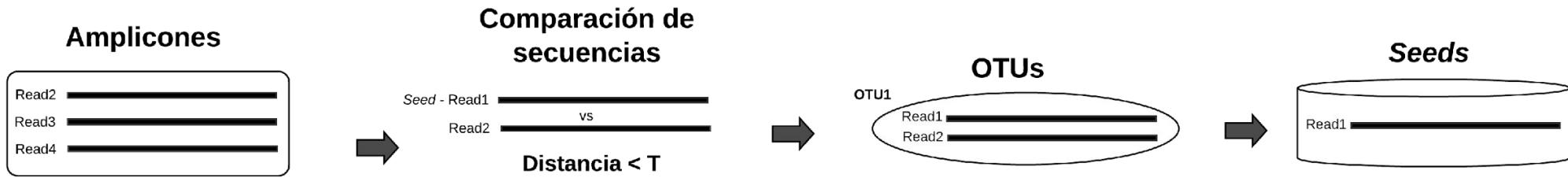
## Seeds



# De novo: clusterización heurística

- Estructura del algoritmo

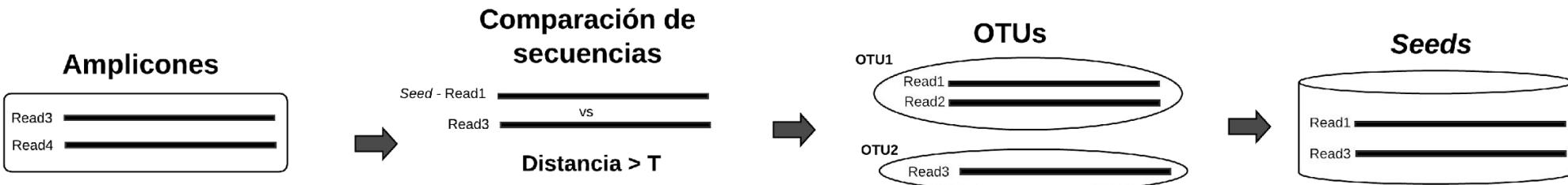
2)



# De novo: clusterización heurística

- Estructura del algoritmo

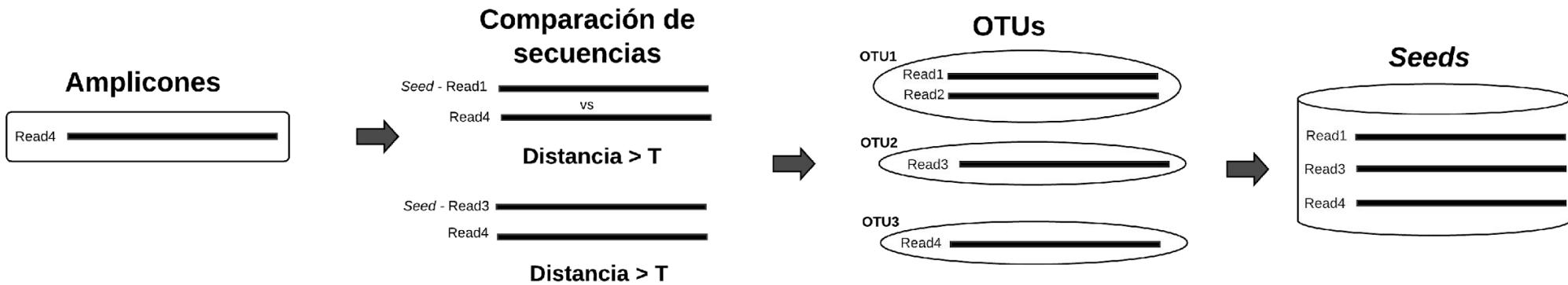
3)



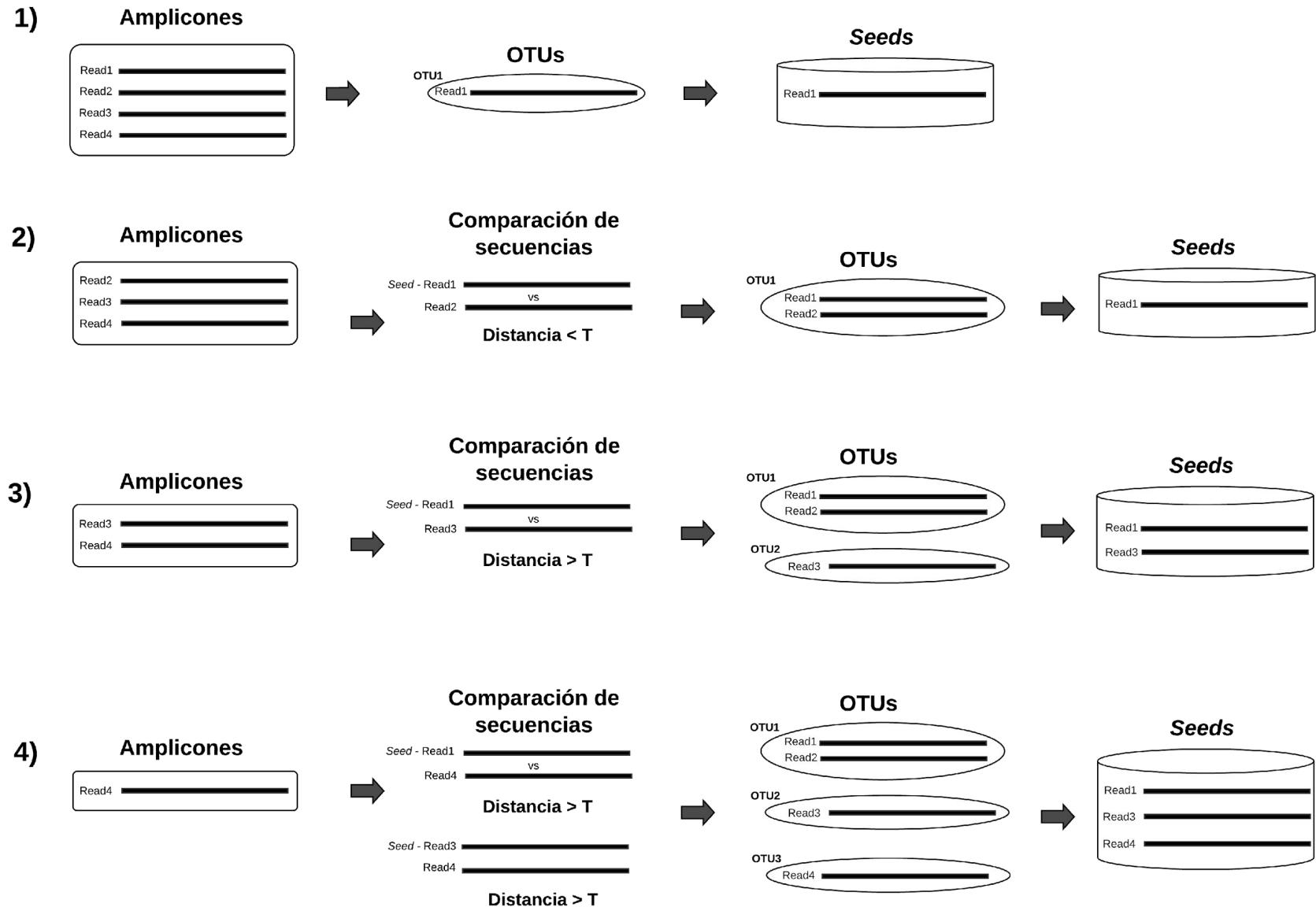
# De novo: clusterización heurística

- Estructura del algoritmo

4)



# De novo: clusterización heurística



# VSEARCH

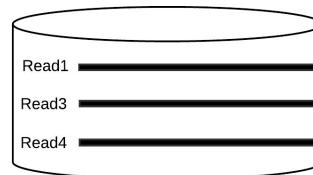
- Particularidades: orden de las secuencias.

1. Dado por el usuario
2. Por longitud decreciente
3. Por abundancia decreciente

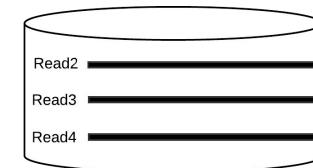


Determina qué secuencias pasan a ser seeds.

Ej.,



Vs.



## VSEARCH

- Particularidades: búsqueda de secuencias.

Dos fases:

1. Selección de candidatos basado en palabras (k-mers) compartidas.
2. Alineamiento de la secuencia problema (*query*) vs. candidatos.

# VSEARCH

- Particularidades: búsqueda de secuencias.

Dos fases:

- 1. Selección de candidatos basado en palabras (k-mers) compartidas.**
2. Alineamiento de la secuencia problema (*query*) vs. candidatos.

# VSEARCH

¿Qué son los k-mers?

Read1

A	G	G	T	A	C	C	A	T	G	T	A	C	C	G	T	T	A	G	G
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

# VSEARCH

¿Qué son los k-mers?

Read1

A	G	G	T	A	C	C	A	T	G	T	A	C	C	G	T	T	A	G	G
A	G	G	T	A	C	C	A												

Primer 8-mer

# VSEARCH

¿Qué son los k-mers?

Read1	A   G   G   T   A   C   C   A   T   G   T   A   C   C   G   T   T   A   G   G
Primer 8-mer	A   G   G   T   A   C   C   A



Nota: por defecto VSEARCH toma k-mers de 8 caracteres (nucleótidos), pero esto puede ser ajustado con el parámetro *wordlength*.

# VSEARCH

¿Qué son los k-mers?

Read1

A	G	G	T	A	C	C	A	T	G	T	A	C	C	G	T	T	A	G	G
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Primer 8-mer

A	G	G	T	A	C	C	A
---	---	---	---	---	---	---	---

Segundo 8-mer

G	G	T	A	C	C	A	T
---	---	---	---	---	---	---	---

# VSEARCH

¿Qué son los k-mers?

Read1

A	G	G	T	A	C	C	A	T	G	T	A	C	C	G	T	T	A	G	G
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Primer 8-mer

A	G	G	T	A	C	C	A
---	---	---	---	---	---	---	---

Segundo 8-mer

G	G	T	A	C	C	A	T
---	---	---	---	---	---	---	---

Tercer 8-mer

G	T	A	C	C	A	T	G
---	---	---	---	---	---	---	---

# VSEARCH

¿Qué son los k-mers?

Read1

A	G	G	T	A	C	C	A	T	G	T	A	C	C	G	T	T	A	G	G
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Primer 8-mer

A	G	G	T	A	C	C	A
---	---	---	---	---	---	---	---

Segundo 8-mer

G	G	T	A	C	C	A	T
---	---	---	---	---	---	---	---

Tercer 8-mer

G	T	A	C	C	A	T	G
---	---	---	---	---	---	---	---

...

T	A	C	C	A	T	G	T
---	---	---	---	---	---	---	---

A	C	C	A	T	G	T	A
---	---	---	---	---	---	---	---

C	C	A	T	G	T	A	C
---	---	---	---	---	---	---	---

C	A	T	G	T	A	C	C
---	---	---	---	---	---	---	---

A	T	G	T	A	C	C	G
---	---	---	---	---	---	---	---

T	G	T	A	C	C	G	T
---	---	---	---	---	---	---	---

G	T	A	C	C	G	T	T
---	---	---	---	---	---	---	---

T	A	C	C	G	T	T	A
---	---	---	---	---	---	---	---

A	C	C	G	T	T	A	G
---	---	---	---	---	---	---	---

$n - k + 1$   
kmers

# VSEARCH

¿Qué son los k-mers?

Read1

A	G	G	T	A	C	C	A	T	G	T	A	C	C	G	T	T	A	G	G
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Primer 8-mer

A	G	G	T	A	C	C	A
---	---	---	---	---	---	---	---

Segundo 8-mer

G	G	T	A	C	C	A	T
---	---	---	---	---	---	---	---

Tercer 8-mer

G	T	A	C	C	A	T	G
---	---	---	---	---	---	---	---

T	A	C	C	A	T	G	T
---	---	---	---	---	---	---	---

A	C	C	A	T	G	T	A
---	---	---	---	---	---	---	---

C	C	A	T	G	T	A	C
---	---	---	---	---	---	---	---

C	A	T	G	T	A	C	C
---	---	---	---	---	---	---	---

A	T	G	T	A	C	C	G
---	---	---	---	---	---	---	---

T	G	T	A	C	C	G	T
---	---	---	---	---	---	---	---

G	T	A	C	C	G	T	T
---	---	---	---	---	---	---	---

T	A	C	C	G	T	T	A
---	---	---	---	---	---	---	---

A	C	C	G	T	T	A	G
---	---	---	---	---	---	---	---

C	C	G	T	T	A	G	G
---	---	---	---	---	---	---	---

$20 - 8 + 1$   
8-mers

# VSEARCH

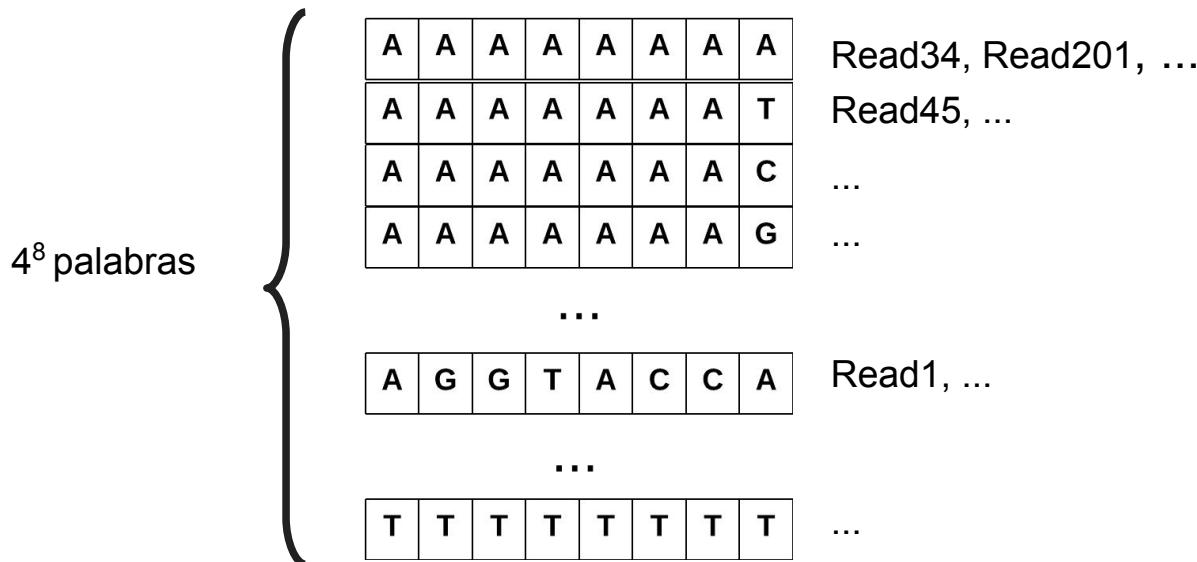
## 1. Selección de candidatos basado en palabras (k-mers) compartidas.

Se crea un índice de todas las  $4^k$  palabras distintas posibles y almacena información sobre en qué secuencias de la base de datos aparecen.

# VSEARCH

## 1. Selección de candidatos basado en palabras (k-mers) compartidas.

Se crea un índice de todas las  $4^k$  palabras distintas posibles y almacena información sobre en qué secuencias de la base de datos aparecen.



# VSEARCH

## 1. Selección de candidatos basado en palabras (k-mers) compartidas.

Se crea un índice de todas las  $4^k$  palabras distintas posibles y almacena información sobre en qué secuencias de la base de datos aparecen.

4<sup>8</sup> palabras



A	A	A	A	A	A	A	A
A	A	A	A	A	A	A	T
A	A	A	A	A	A	A	C
A	A	A	A	A	A	A	G

...

A	G	G	T	A	C	C	A
---	---	---	---	---	---	---	---

...

T	T	T	T	T	T	T	T
---	---	---	---	---	---	---	---

Read34, Read201, ...

Read45, ...

...

...

Read1, ...



Esto permite rápidamente contar el número de k-mers compartidos entre cualquier par de secuencias.

# VSEARCH

1. Selección de candidatos basado en palabras (k-mers) compartidas.



Nota: por defecto VSEARCH selecciona las secuencias con al menos 10 kmers compartidos, pero esto puede ser ajustado con el parámetro *minwordmatches*.

# VSEARCH

1. Selección de candidatos basado en palabras (k-mers) compartidas.



Comparar secuencias basadas en base a k-mers compartidos es un método común para evaluar rápidamente la similitud entre dos secuencias sin alinearlas (lo que sí consume mucho tiempo de cómputo).

## VSEARCH

- Particularidades: búsqueda de secuencias.
- Dos fases:
- 1. Selección de candidatos basado en palabras (k-mers) compartidas.
- 2. Alineamiento de la secuencia problema (*query*) vs. candidatos.

## VSEARCH

- Particularidades: búsqueda de secuencias.
- Dos fases:
- 1. Selección de candidatos basado en palabras (kmers) compartidas.
- **2. Alineamiento de la secuencia problema (*query*) vs. candidatos.**

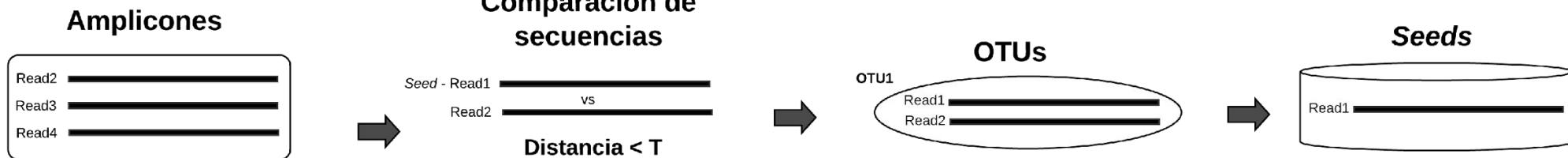
## VSEARCH

- 2. Alineamiento de la secuencia problema (*query*) vs. candidatos.
  - Se seleccionan las secuencias de la base de datos (*seeds*), ordenadas decrecientemente de acuerdo al número de palabras compartidas con la secuencia problema (*query*).
  - Se realiza un alineamiento global (óptimo) entre la secuencia problema (*query*) y las secuencias seleccionadas (*seeds*).

# VSEARCH

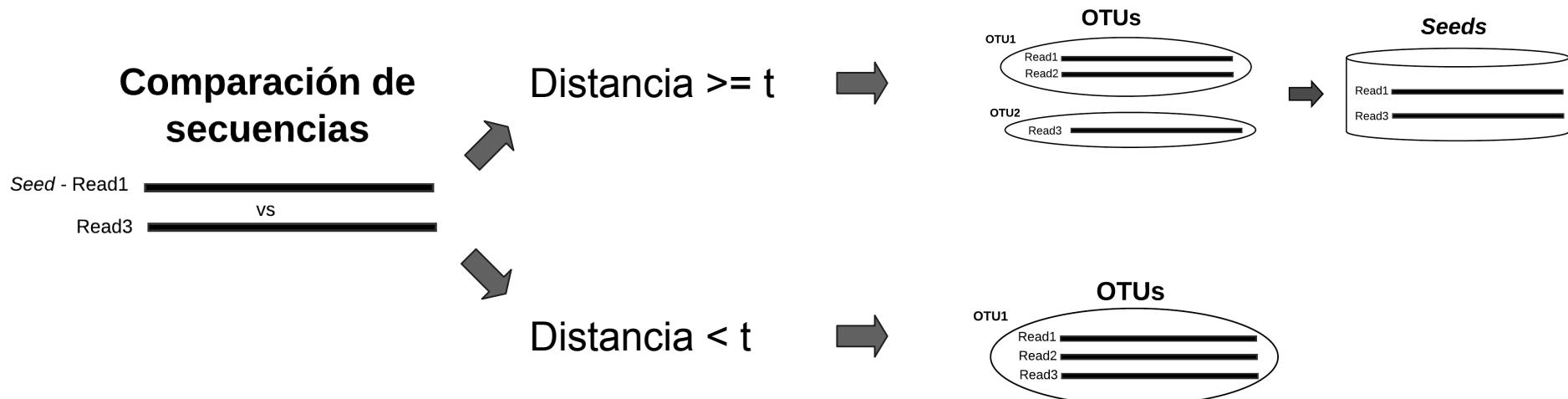
- 2. Alineamiento de la secuencia problema (*query*) vs. candidatos.

2)



# VSEARCH

- 2. Alineamiento de la secuencia problema (*query*) vs. candidatos.



## VSEARCH

- 2. Alineamiento de la secuencia problema (*query*) vs. candidatos.



Nota: Por defecto, la secuencia problema (*query*) se clusteriza con el *seed* que presenta la mayor similitud (*distance-based greedy clustering, DGC*).

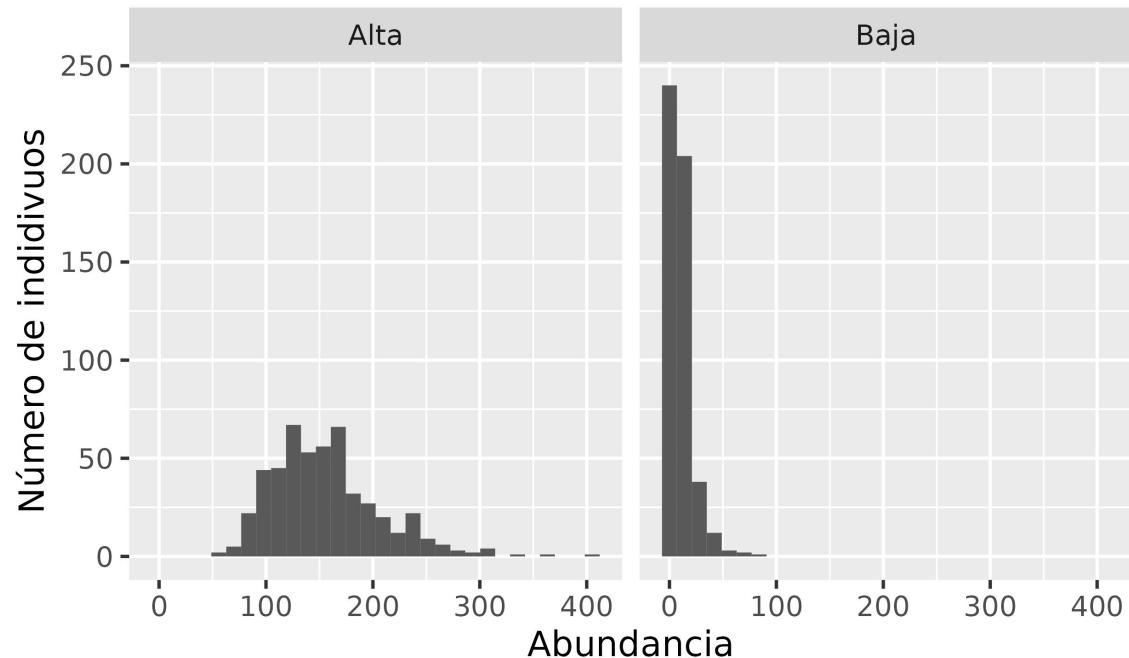
Opcionalmente, se puede utilizar el parámetro *sizeorder* para agrupar la secuencias con el *seed* representando el cluster de mayor abundancia (*abundance-based greedy clustering, AGC*)

# VSEARCH

## Evaluación comparativa: VSEARCH vs USEARCH

Dos set de datos de comunidades referencia artificiales:

- 1) diversidad alta.
- 2) diversidad baja.



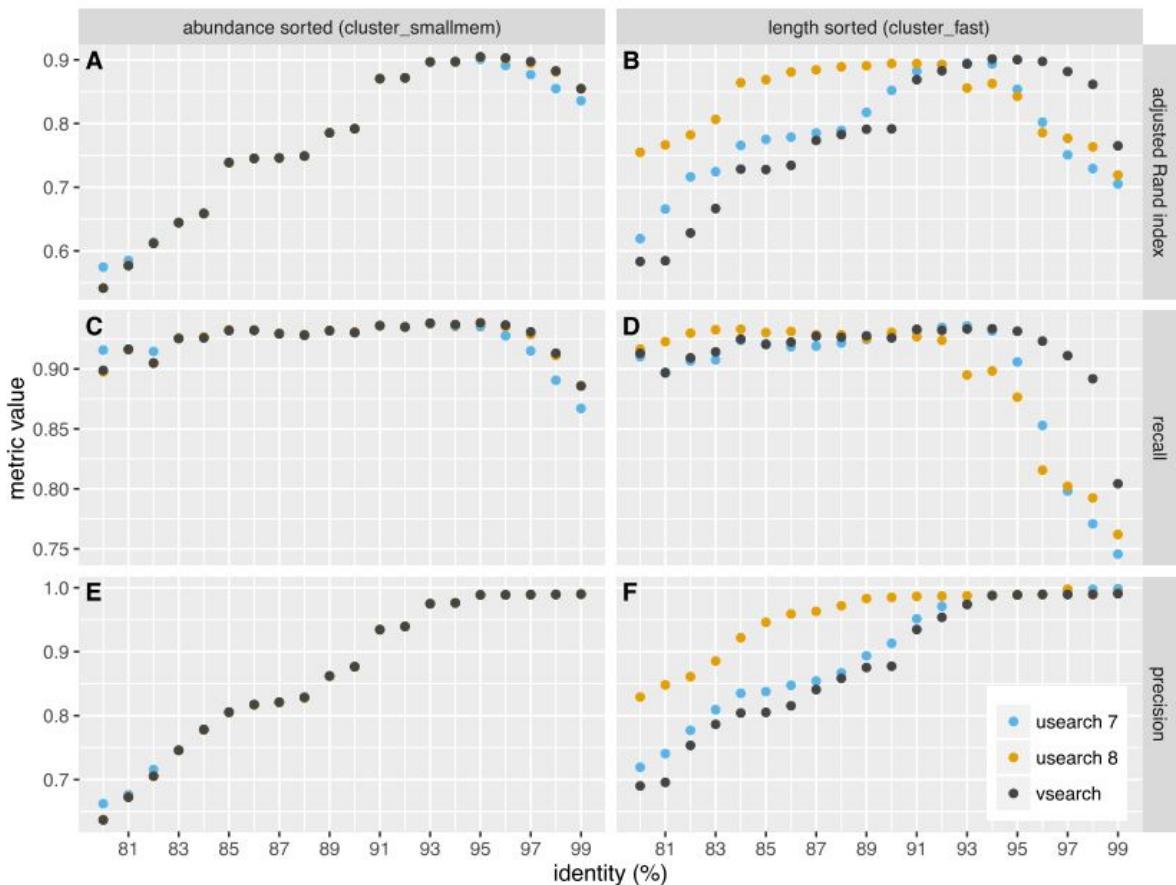
# VSEARCH

## Evaluación comparativa: VSEARCH vs USEARCH

- **Recuperación:** en qué medida los amplicones de la misma especie se agrupan en el mismo OTU (es decir, sin división excesiva).
- **Precisión:** en qué medida los amplicones en una OTU se asignan a la misma especie (es decir, no se agrupan en exceso).
- **Índice Rand:** resumen de Recuperación y Precisión.

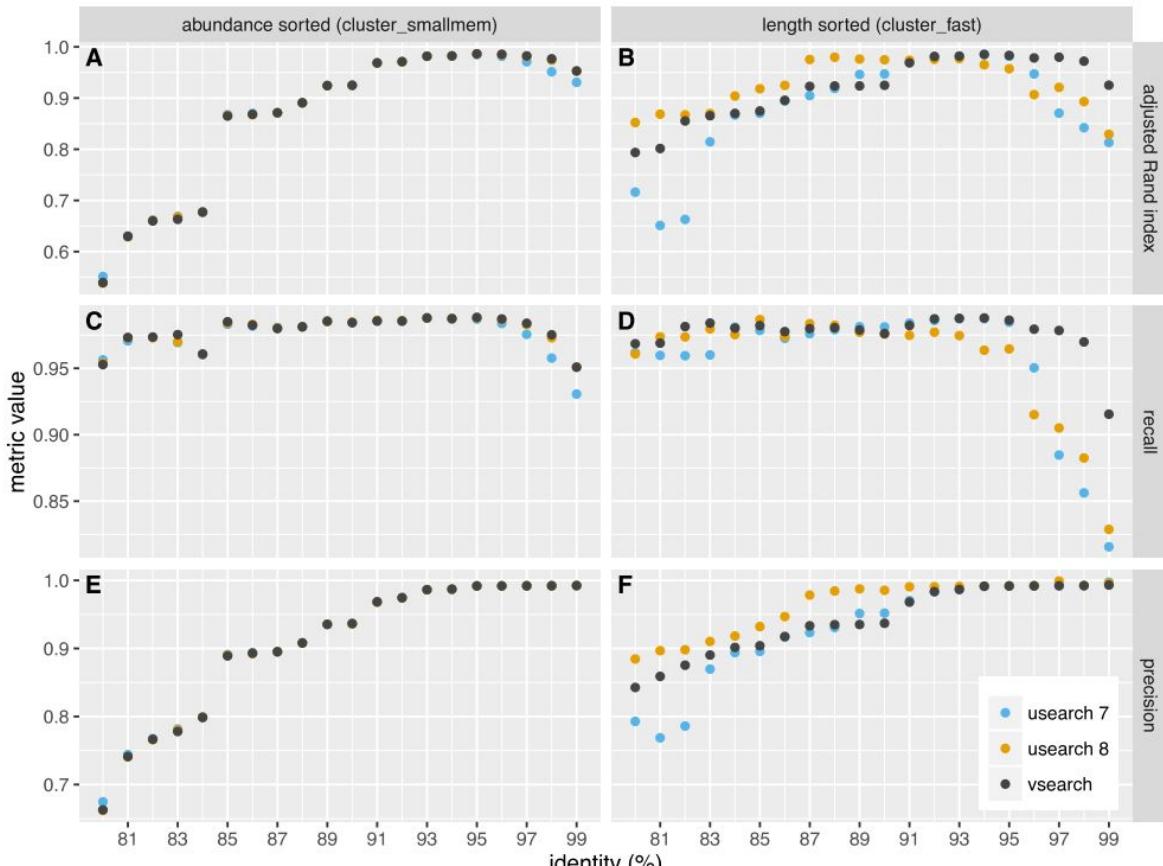
# VSEARCH

**Figura 2 Clusterización de la comunidad de diversidad alta.** USEARCH versión 7 (azul) y 8 (naranja) y VSEARCH (negro) datos ordenados por abundancia (A, C, E) y ordenados por longitud (cluster\_fast) (B, D, F). El rendimiento se indica con las métricas del índice Rand (A, B), recuperación (C, D) y precisión (E, F).



# VSEARCH

**Figura 3 Clusterización de la comunidad de diversidad baja.** USEARCH versión 7 (azul) y 8 (naranja) y VSEARCH (negro) datos ordenados por abundancia (A, C, E) y ordenados por longitud (cluster\_fast) (B, D, F). El rendimiento se indica con las métricas del índice Rand (A, B), recuperación (C, D) y precisión (E, F).



# Procesamiento de datos: clusterización de secuencias

## Clusterización:

De referencia - abierto

De referencia – cerrado

## ✓ De novo:

Jerárquica

## ✓ Heurística:

**VSEARCH**

DADA2

# Procesamiento de datos: clusterización de secuencias

## Clusterización:

De referencia - abierto

De referencia – cerrado

## ✓ De novo:

Jerárquica

## ✓ Heurística:

VSEARCH

**DADA2**

# DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan<sup>1</sup>, Paul J McMurdie<sup>2</sup>,  
Michael J Rosen<sup>3</sup>, Andrew W Han<sup>2</sup>, Amy Jo A Johnson<sup>2</sup> &  
Susan P Holmes<sup>1</sup>

## Schematic of OTU and DADA2 approaches towards amplicon sequencing errors.

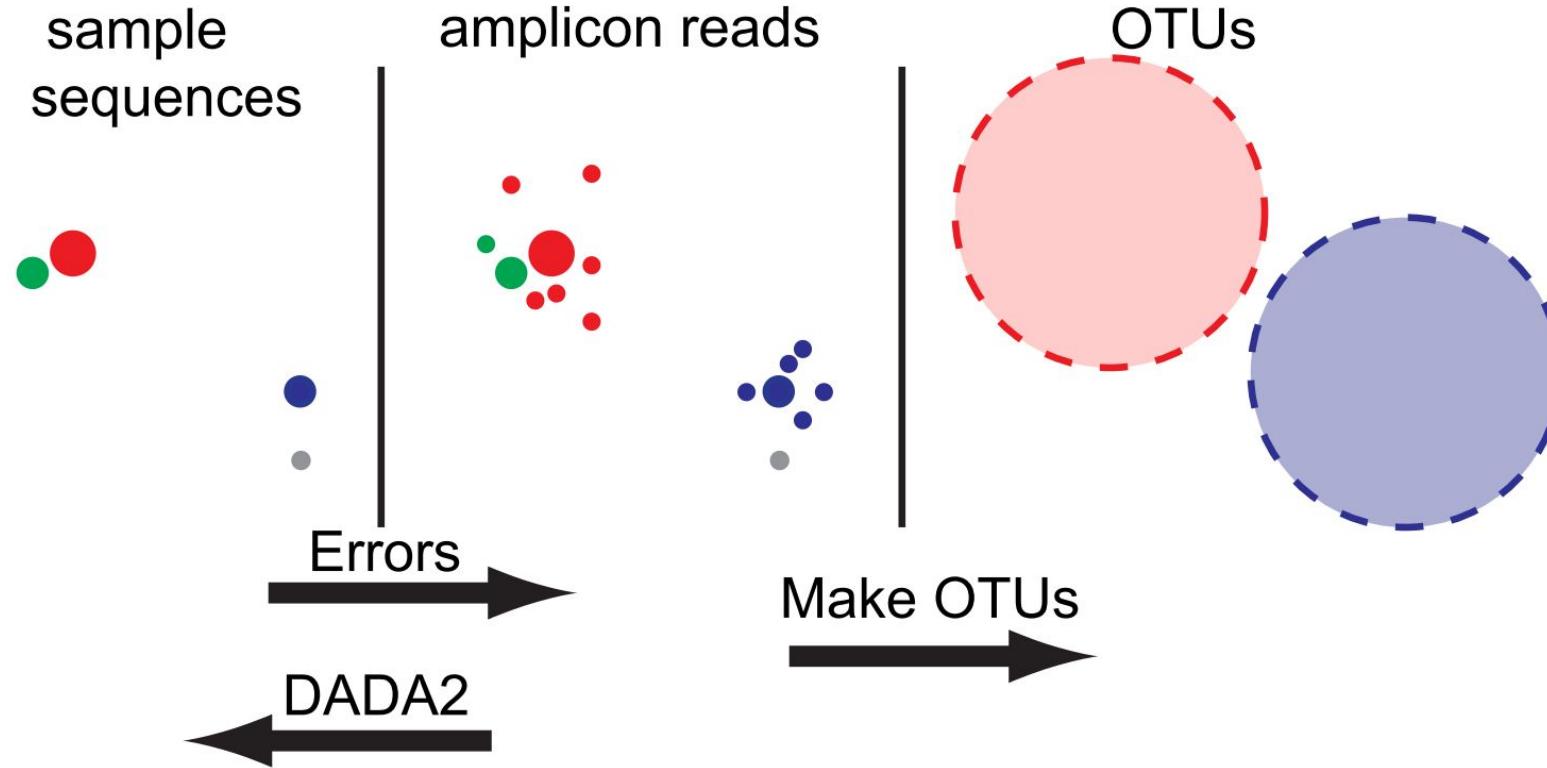


Figura 1. Los círculos representan clusters idénticos de reads con un tamaño escalado por abundancia y color correspondiente a la secuencia verdadera sin errores (hay cuatro secuencias distintas en la muestra: rojo, verde, azul y gris). Los errores se introducen mediante la secuenciación de amplicones desde la parte izquierda hasta la mitad del diagrama. Los OTUs protegen contra inferencias falsas positivas al agrupar secuencias similares. DADA2 utiliza un modelo estadístico de errores de amplicones para inferir directamente las secuencias de muestra subyacentes y, por lo tanto, intenta eliminar el ruido de los datos desde el centro hacia la izquierda.

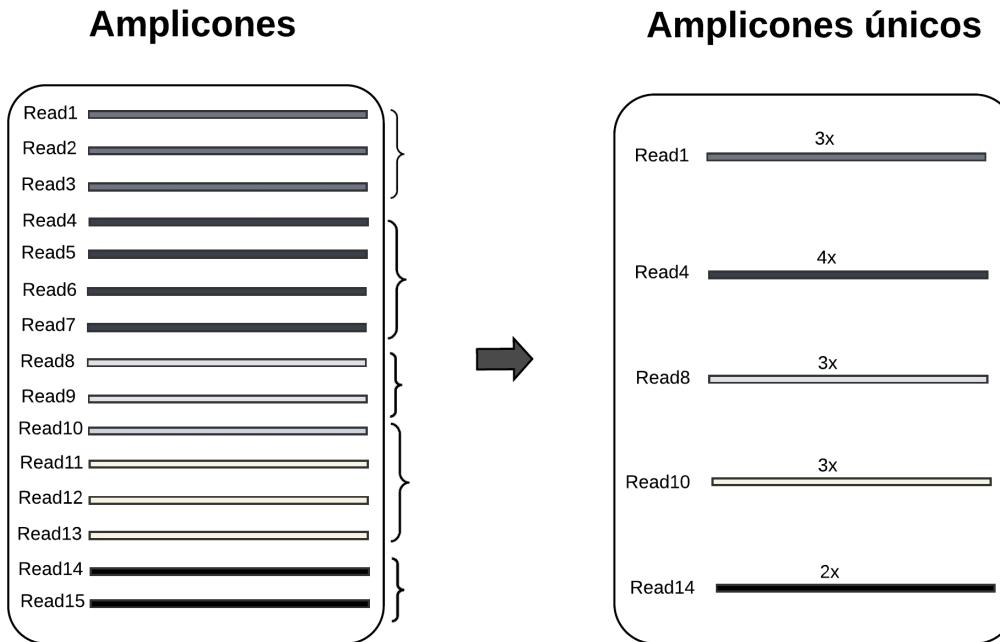
## Divisive Amplicon Denoising Algorithm 2 (DADA2)

- Algoritmo de clusterización: aproximación heurística.

*“Divisive partitioning algorithm”*

# DADA2

- Primer paso: **dereplicado** de **amplicones**.  
Los amplicones con la misma secuencia se agrupan, pasando a tener una abundancia asociada y un perfil de calidad de consenso.

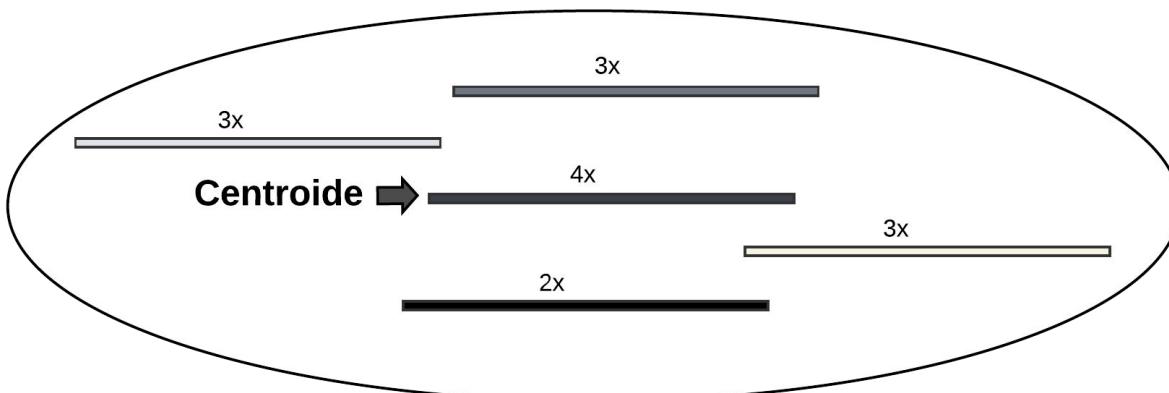


# DADA2

## . Segundo paso: inicialización.

Todas las secuencias pasan a formar un una única partición, quedando la secuencia más abundante como centroide.

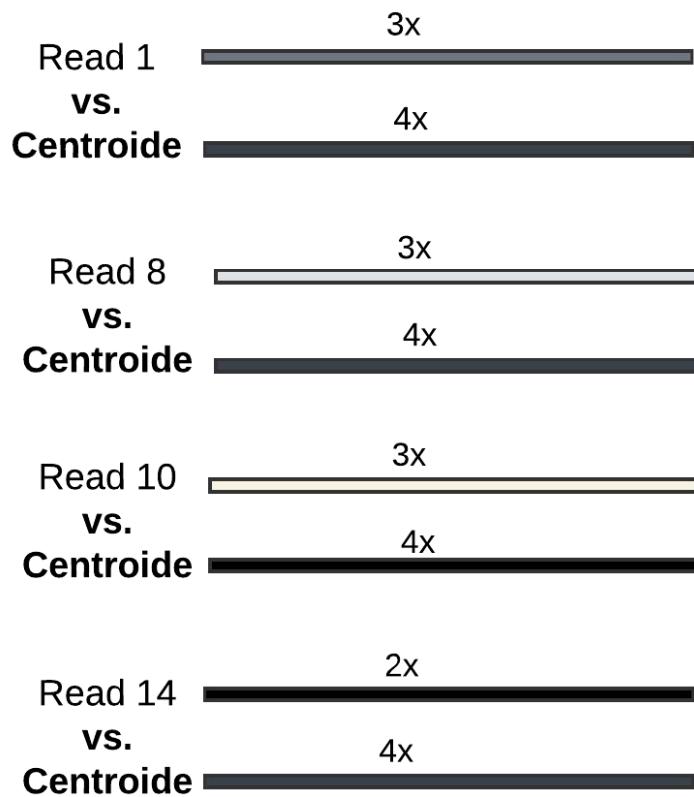
### Amplicones únicos: una única partición



## DADA2

- Tercer paso: comparación de secuencias.**

Todas las secuencias son comparadas con el centroide mediante alineamiento global.



## DADA2

- Cuarto paso: cómputo de tasa de error  $\lambda_{ji}$  y p-valor  $P_A$ .
- $P_A$  cuantifica la noción de que la secuencia “i” podría ser demasiado abundante para explicarla por errores en la secuenciación de amplicones.

Read 1            3x

vs.

Centroide            4x

→      
$$\lambda_{ij} = \prod_{l=0}^L p(j(l) \rightarrow i(l), q_i(l))$$

↓

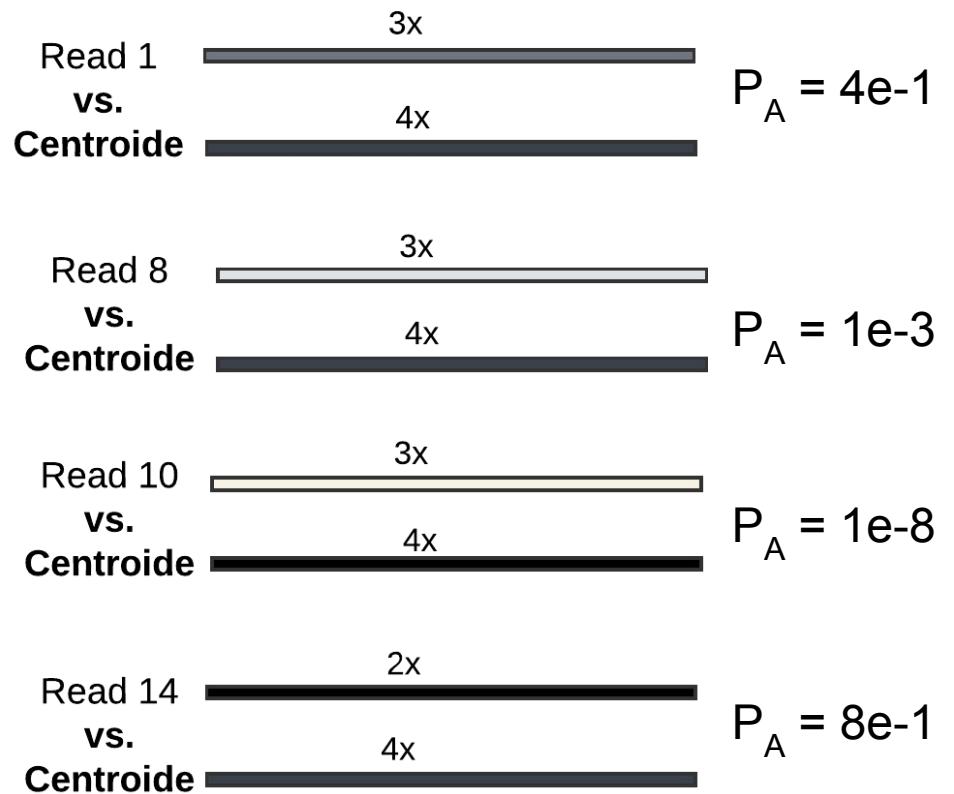
$$P_A(j \rightarrow i) = \frac{\sum_{a=a_i}^{\infty} Pois(n_j \lambda_{ji}, a)}{1 - Pois(n_j \lambda_{ji}, 0)}$$

## DADA2

- **Quinto paso: selección del menor  $P_A$ .**

Seleccionamos la secuencias con el menor

p-valor  $P_A$ .

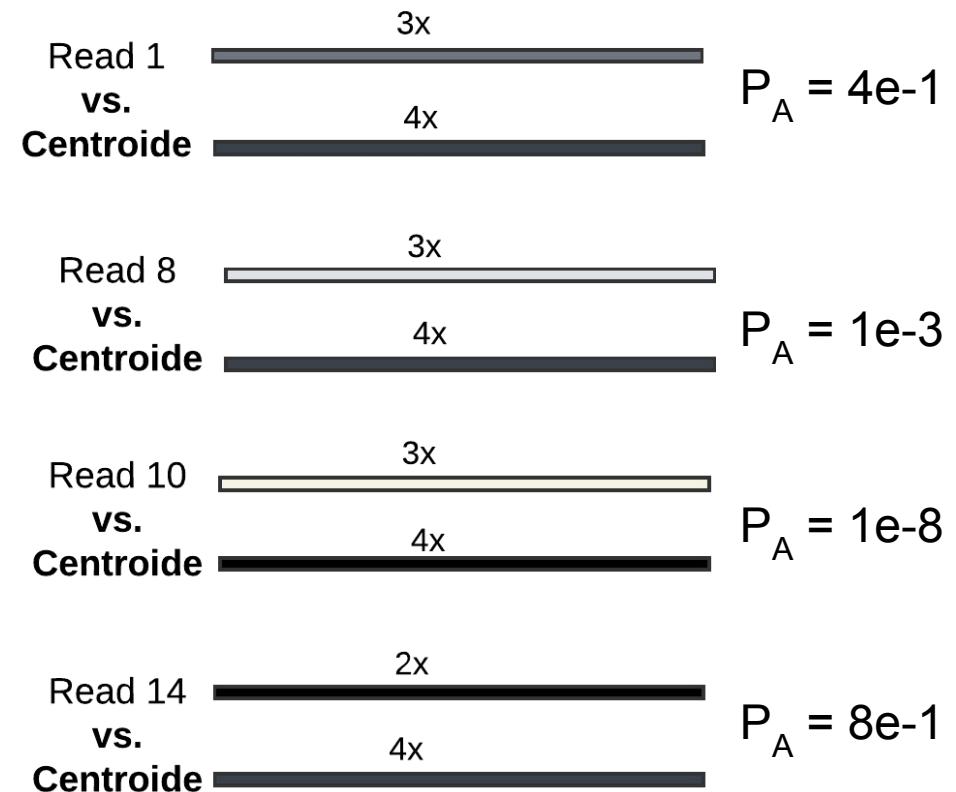


## DADA2

- ### • Quinto paso: selección del menor $P_A$ .

Seleccionamos la secuencias con el menor

p-valor  $P_A$ .



# DADA2

- **Sexto paso: Creación de una nueva partición.**

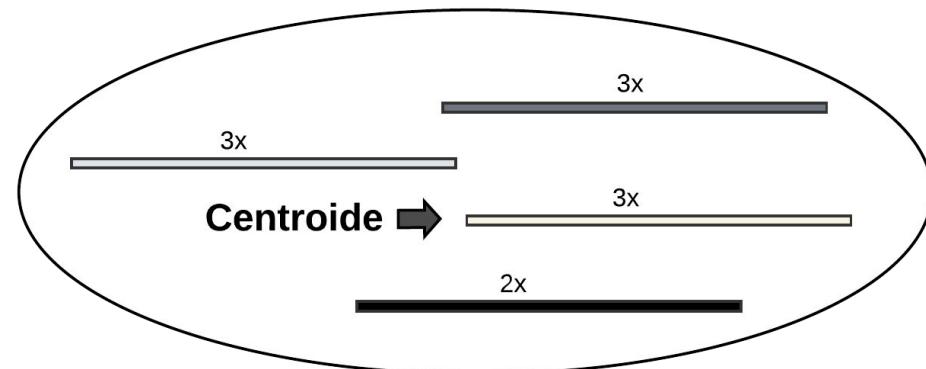
Read 10      3x  
vs.  
Centroide      4x



$$P_A = 1e-8$$



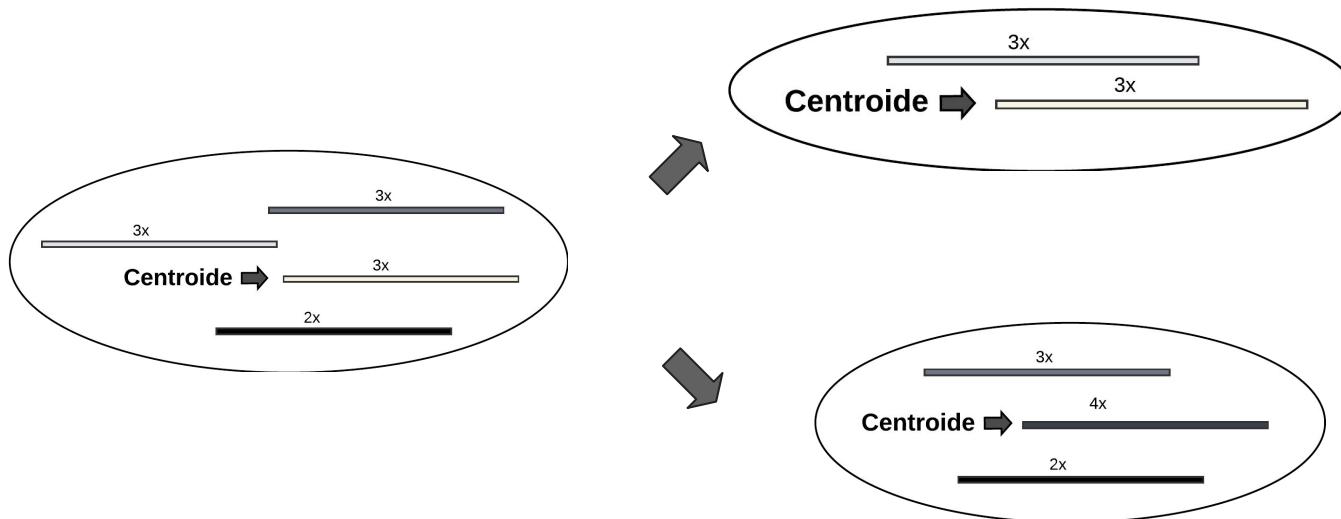
Corrección Bonferroni  
(multiples comparaciones)       $\rightarrow$        $P'_A < t$



## DADA2

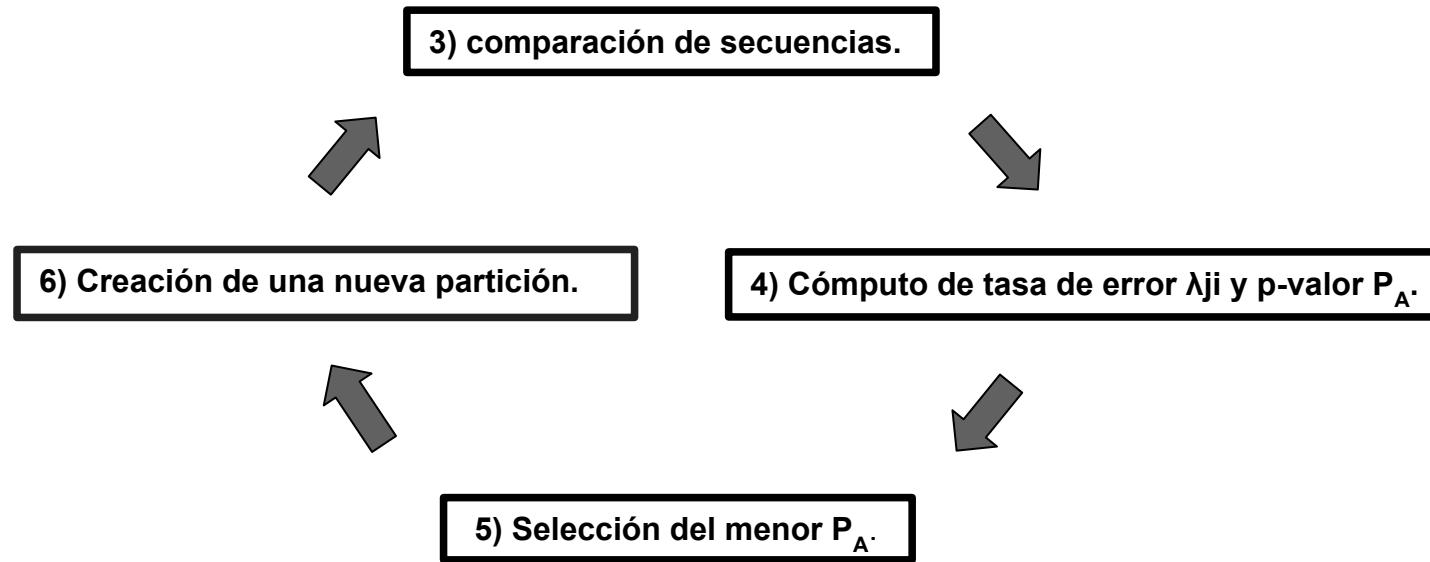
- **Séptimo paso: Creación de una nueva partición.**

Luego de que una partición es formada, todas las secuencias son comparadas con el nuevo centroide y cada secuencia es clusterizada con la partición que tiene más probabilidades de haberlo producido.



# DADA2

## • Iteración



Luego de que una partición es formada, todas las secuencias son comparadas con el nuevo centroide y cada secuencia es clusterizada con la partición que tiene más probabilidades de haberlo producido.

# DADA2

- **Sexto paso: Creación de una nueva partición.**

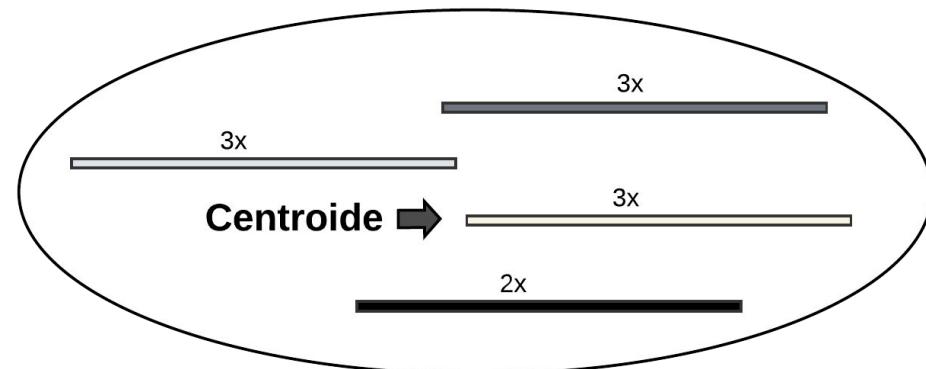
Read 10      3x  
vs.  
Centroide      4x



$$P_A = 1e-8$$

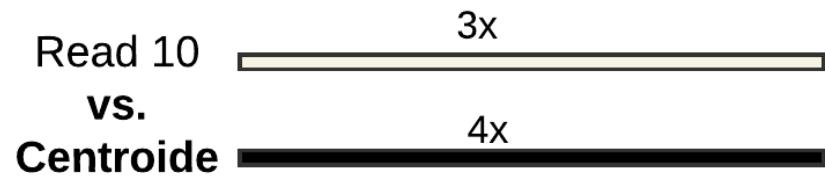


Corrección Bonferroni  
(multiples comparaciones)       $\rightarrow$        $P'_A < t$



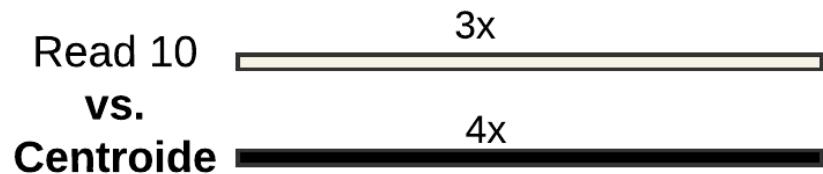
## DADA2

- Heurísticas en la creación de ASVs: comparación de secuencias



## DADA2

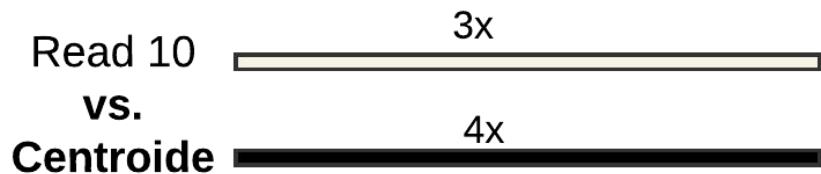
- Heurísticas en la creación de ASVs: comparación de secuencias



El alineamiento entre secuencias es lo más caro computacionalmente en todo el proceso!

## DADA2

- Heurísticas en la creación de ASVs: comparación de secuencias



Primer heurística: los reads a ser comparados son seleccionados previamente en base al números de k-mers compartidos (similarmente a VSEARCH).

## DADA2

## • Heurísticas en la creación de ASVs: comparación de secuencias

A horizontal bar chart comparing two data series. The top series, 'Read 10', is represented by a light gray bar and is labeled '3x' above it. The bottom series, 'Centroide', is represented by a dark gray bar and is labeled '4x' above it. The bars are positioned side-by-side, with 'Read 10' on the left and 'Centroide' on the right.

Segunda heurística: alineamiento en bandas, esto es, se evitan explorar alineamientos que implican un gran número de *gaps*.

	A	G	G	C	T	
A				<b>X</b>	<b>X</b>	<b>X</b>
C					<b>X</b>	<b>X</b>
G	<b>X</b>					<b>X</b>
C	<b>X</b>	<b>X</b>				
T	<b>X</b>	<b>X</b>	<b>X</b>			

## DADA2

- Tasa de error  $\lambda_{ji}$  y p-valor  $P_A$ .



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



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

## DADA2

- Tasa de error  $\lambda_{ji}$ .  
La tasa a la que se produce un amplicón “i” a partir de una secuencia de muestra “j” computado como el producto sobre las probabilidades de transición entre los nucleótidos alineados, integrando la composición nucleotídica y las probabilidades de error.



## DADA2

- Tasa de error  $\lambda_{ji}$ .  
Integra la composición nucleotídica y las probabilidades de error. La probabilidad de transición entre nucleótidos alineados depende del nucleótido original, el de sustitución, y de la puntuación de calidad asociada (es decir,  $16 \times 41$  probabilidades de transición, por ejemplo:  $p(A \rightarrow C, 35)$ ).



## DADA2

- **P-valor**

$P_A$ .

Cuantifica la noción de que la secuencia “i” podría ser demasiado abundante para explicarla por errores en la secuenciación de amplicones.

## DADA2

- **P-valor**

$P_A$ .

Si los errores de secuenciación son independientes entre los *reads*, entonces, el número de lecturas de amplicones con la secuencia “i” que se producirán a partir de la secuencia de muestra “j” tiene una distribución de poisson.

## DADA2

- **P-valor**

$P_A$

Si los errores de secuenciación son independientes entre las reads, entonces, el número de lecturas de amplicones con la secuencia “i” que se producirán a partir de la secuencia de muestra “j” tiene una distribución de poisson.



Esta distribución de poisson tiene una expectativa igual a una tasa de error  $\lambda_{ji}$  multiplicada por las lecturas esperadas de la secuencia de muestra “j”.

## DADA2

- **P-valor  $P_A$ .**
- Si los errores de secuenciación son independientes entre las reads, el número de lecturas de amplicones con la secuencia “i” que se producirán a partir de la secuencia de muestra “j” tiene una distribución de poisson.



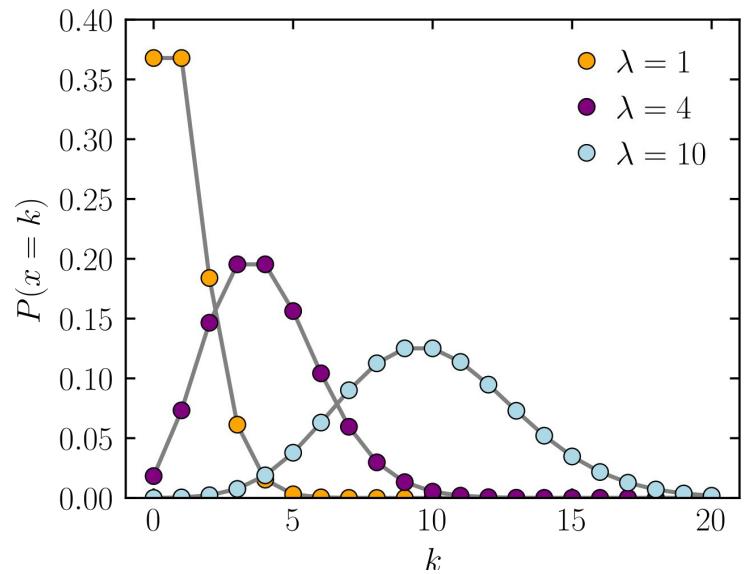
- Esta distribución de poisson tiene una expectativa igual a una tasa de error  $\lambda_{ji}$  multiplicada por las lecturas esperadas de la secuencia de muestra “j”.



$$Pois(n_j \lambda_{ji})$$

# DADA2

- ¿Por qué la distribución de Poisson?



## DADA2

- ¿Por qué Poisson?  
Modelamos la probabilidad de obtener r éxitos en n ensayos.  
En este caso: encontrar una r especies en n reads.

## DADA2

- ¿Por qué Poisson?  
Modelamos la probabilidad de obtener r éxitos en n ensayos.  
En este caso: encontrar una r especies en n reads.



### Distribución binomial

$$P(x=r) = \binom{n}{r} p^r \cdot q^{n-r}$$

## DADA2

- ¿Por qué Poisson?
- Modelamos la probabilidad de obtener r éxitos en n ensayos.
- En este caso: encontrar una r especies en n reads.



### Distribución binomial

$$P(x=r) = \binom{n}{r} p^r \cdot q^{n-r} \quad \rightarrow$$

### Distribución de Poisson

$$P(x=r) = \frac{\lambda^r e^{-\lambda}}{r!}$$

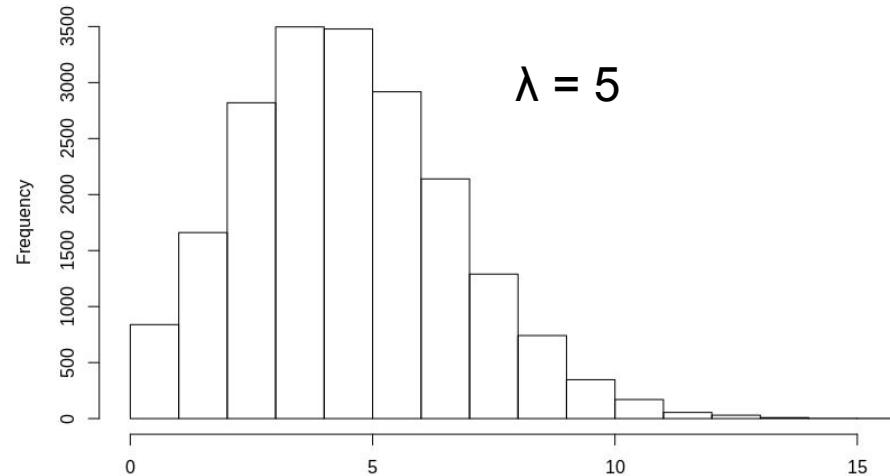
Tende a una distribución de Poisson cuando  $n \rightarrow \infty$ ,  
 $p \rightarrow 0$  y  $np$  permanecen constantes.

## DADA2

- ¿Por qué Poisson?
- Utilizamos Poisson cuando  $n$  es grande y  $p$  es chico.
- La distribución de poisson modela una variable que cuenta el número de eventos que ocurren de forma aleatoria e independiente en una unidad determinada de tiempo/espacio/volumen.

### Distribución de Poisson

$$P(x=r) = \frac{\lambda^r e^{-\lambda}}{r!}$$



## DADA2

- **P-valor**

$P_A$

Si los errores de secuenciación son independientes entre las reads, el número de lecturas de amplicones con la secuencia “i” que se producirán a partir de la secuencia de muestra “j” tiene una distribución de Poisson.



Esta distribución de Poisson tiene una expectativa igual a una tasa de error  $\lambda_{ji}$  multiplicada por las lecturas esperadas de la secuencia de muestra “j”.



$$Pois(n_j \lambda_{ji})$$

## DADA2

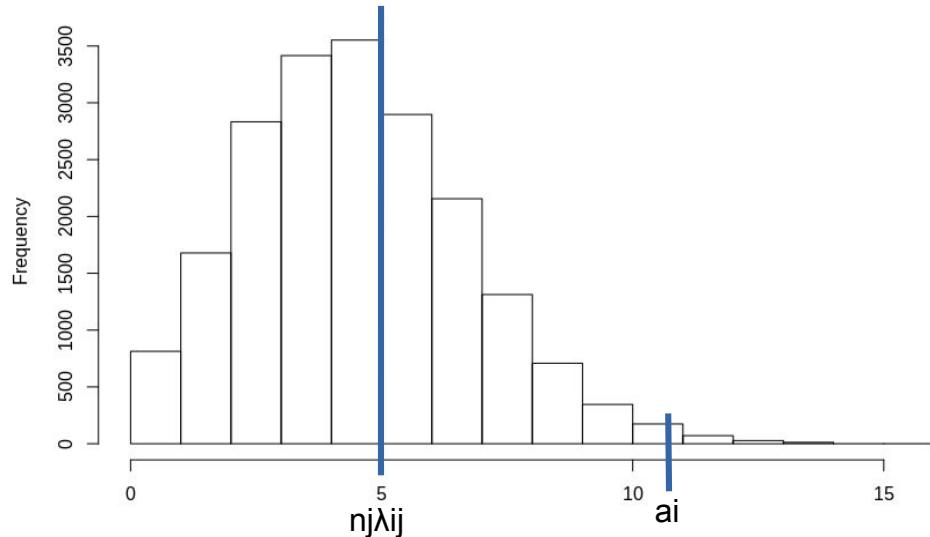
### • P-valor

$P_A$ .

Dada una secuencia “i” con abundancia “ $a_i$ ” en la partición “j” con abundancia “ $n_j$ ”.

Entonces, condicionado a que “i” esté al menos una vez, el valor-p de abundancia es la probabilidad de ver “ $a_i$ ” o más lecturas idénticas.

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



# DADA2

- **P-valor**

**P<sub>A</sub>**.

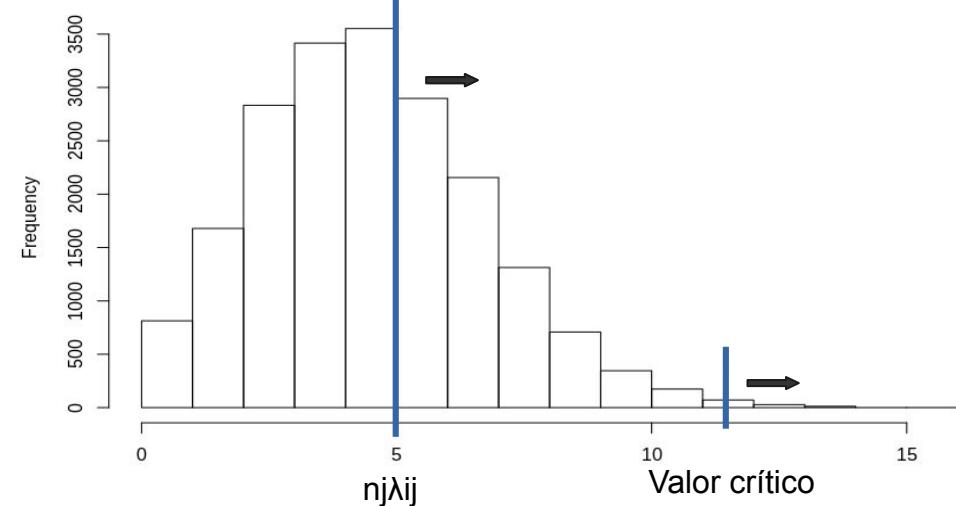
Mayor id% → mayor  $\lambda_{ij}$  →  $n_j \lambda_{ij}$  (i.e., valor esperado de la distribución de Poisson),

→ mayor es el valor crítico asociado al p-valor P<sub>A</sub>.



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

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



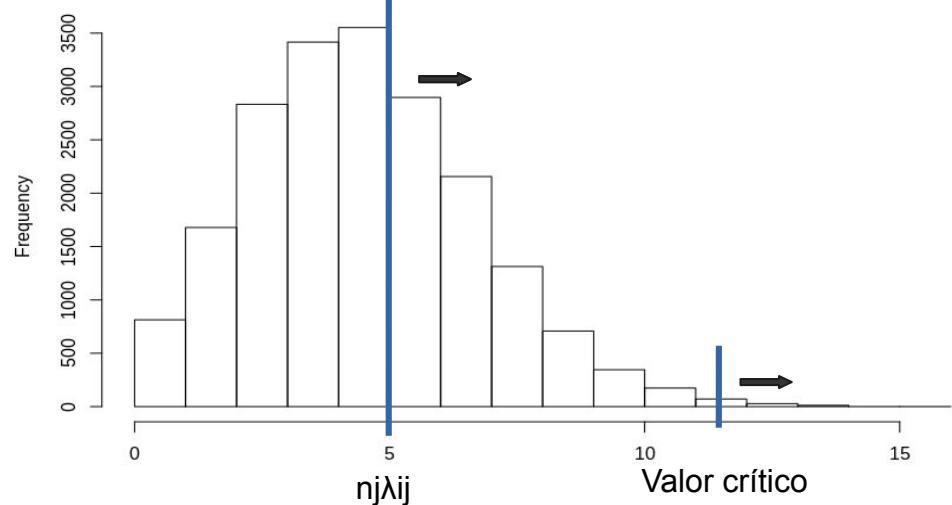
DADA2

- **P-valor  $P_A$** :
  - Una  $P_A$  bajo indica que hay más reads de la secuencia “i” de las que pueden explicarse por errores introducidos durante la amplificación y secuenciación de las “ $n_j$ ” copias de la secuencia de muestra “j”.

Read 1      3x  
vs.  
Centroide      4x

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

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



## DADA2

- Sexto paso: Creación de una nueva partición.

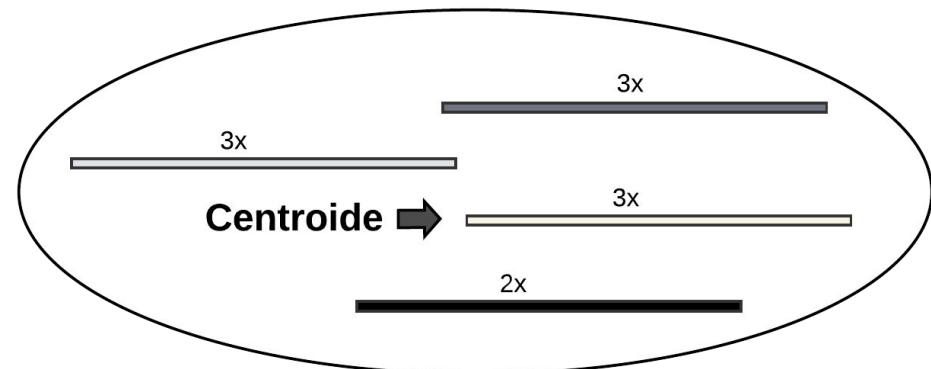
Read 10      3x  
vs.  
Centroide      4x



$$P_A = 1e-8$$



Corrección Bonferroni  
(multiples comparaciones)       $\rightarrow$        $P'_A < t$        $\rightarrow$   
 $P'_A$



# DADA2

- **Parametrización del modelo de error.**

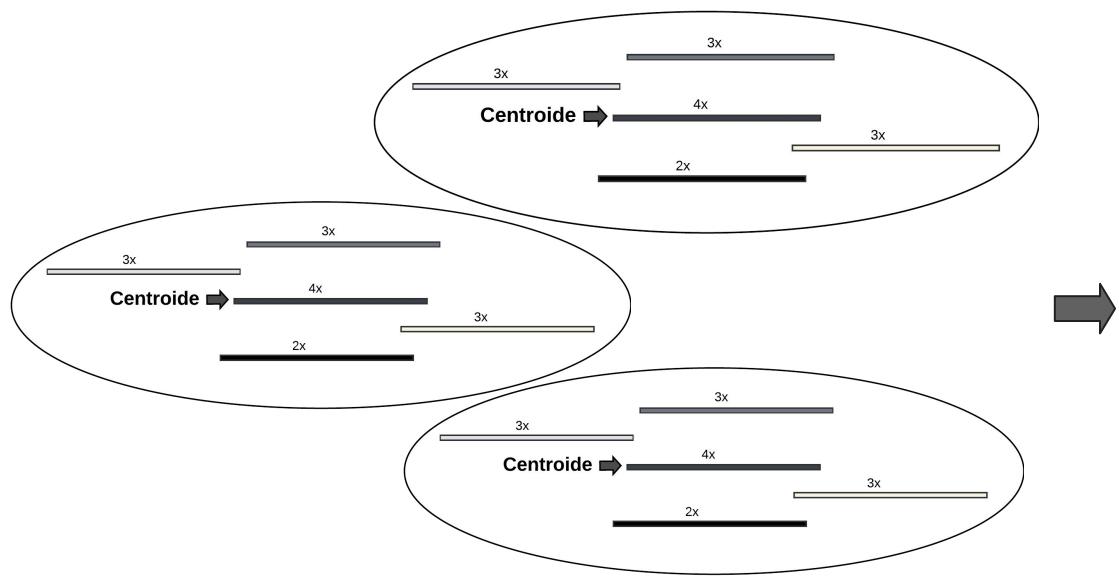
## DADA2

- **Parametrización del modelo de error.**

Es utilizado un modelo paramétrico de los errores introducidos tanto el la amplificación por PCR como la secuenciación. Estos errores generalmente varían entre los distintos protocolos, por lo que los parámetros del modelo son estimados para cada set de datos.

# DADA2

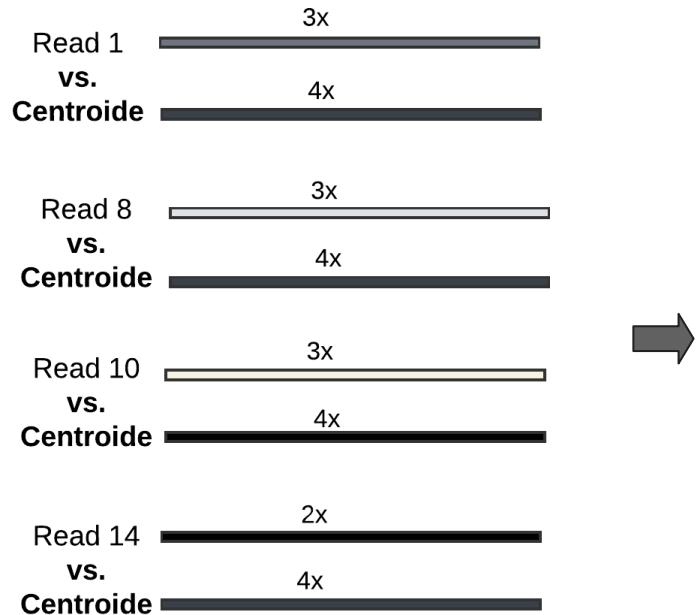
- **Parametrización del modelo de error.**
- Dada una partición las son estimadas las probabilidades de transición de  $16 \times 41$  para cada valor Q, comparando todas las secuencias con el centroide.



Read 1	3x	vs.	3x
	4x	vs.	4x
Centroide	4x	vs.	4x
Read 8	3x	vs.	3x
	4x	vs.	4x
Centroide	4x	vs.	4x
Read 10	3x	vs.	3x
	4x	vs.	4x
Centroide	4x	vs.	4x
Read 14	2x	vs.	2x
	4x	vs.	4x
Centroide	4x	vs.	4x

## DADA2

## • **Parametrización del modelo de error.**



16 x 41

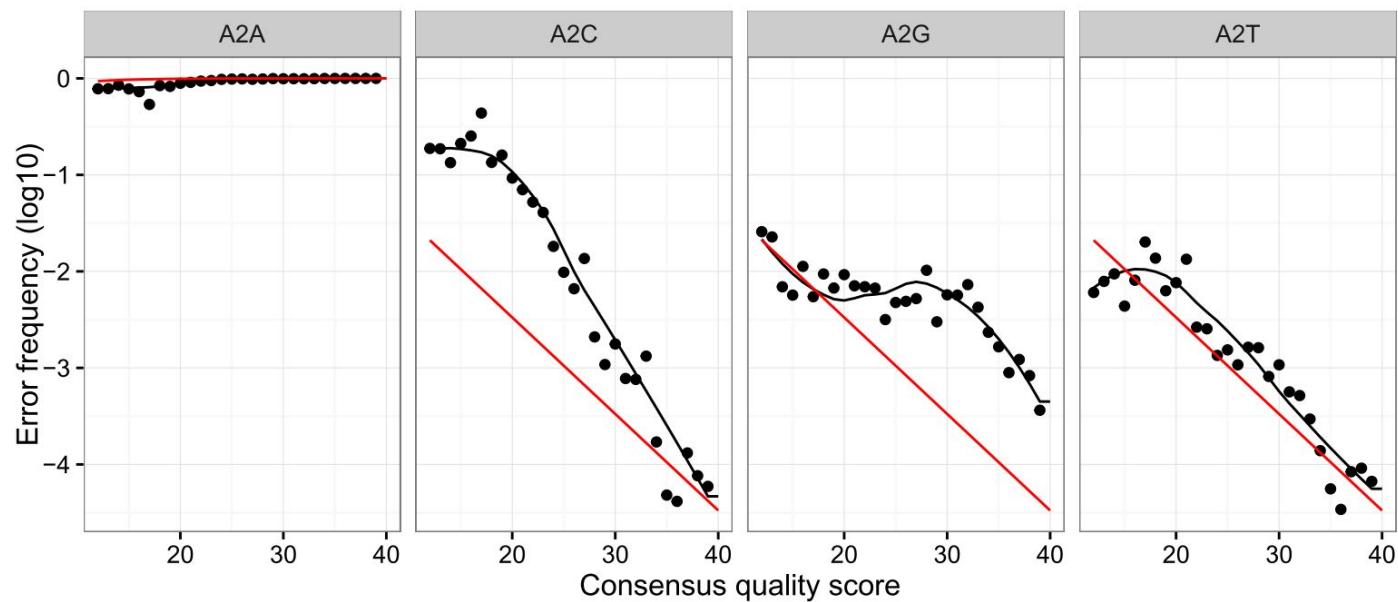
	<b>Q=1</b>	<b>Q=2</b>	<b>Q=3</b>	<b>...</b>	<b>Q=41</b>
$p_{AA}$	$p(A-A,1)$	$p(A-A,2)$	$p(A-A,3)$	$\dots$	$p(A-A,41)$
$p_{CA}$	$p(C-A,1)$	$p(C-A,2)$	$p(C-A,3)$	$\dots$	$p(C-A,41)$
$p_{GA}$	$p(G-A,1)$	$p(G-A,2)$	$p(G-A,3)$	$\dots$	$p(G-A,41)$
$\dots$	$\dots$	$\dots$	$\dots$		
$p_{TT}$	$p(T-T,1)$	$p(T-T,2)$	$p(T-T,3)$	$\dots$	$p(T-T,41)$

# DADA2

## • Parametrización del modelo de error.

Figura 8. Las tasas de error directa observadas para el caso en el que la base correcta es una A. El eje x muestra el puntaje de calidad Q; el eje y la frecuencia de la transición especificada. Los puntos muestran las frecuencias observadas, la línea negra el modelo de error loess y la línea roja las tasas esperadas dado el puntaje de calidad  $Q = -10\log_{10}(\text{perr})$ .

Illumina Miseq error rates as a function of quality.

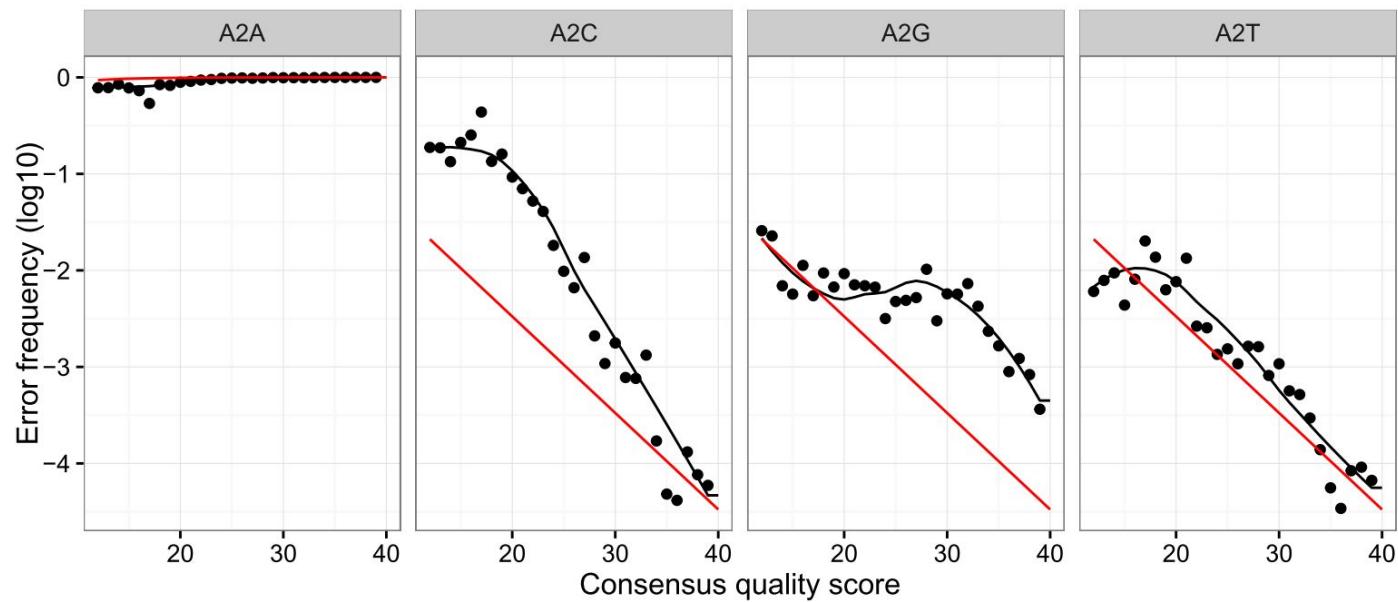


# DADA2

## • Parametrización del modelo de error.

Se ajusta una función de LOESS (*Locally Estimated Scatterplot Smoothing*) ponderado al registro regularizado de las tasas de *mismatches* observadas en función de su calidad.

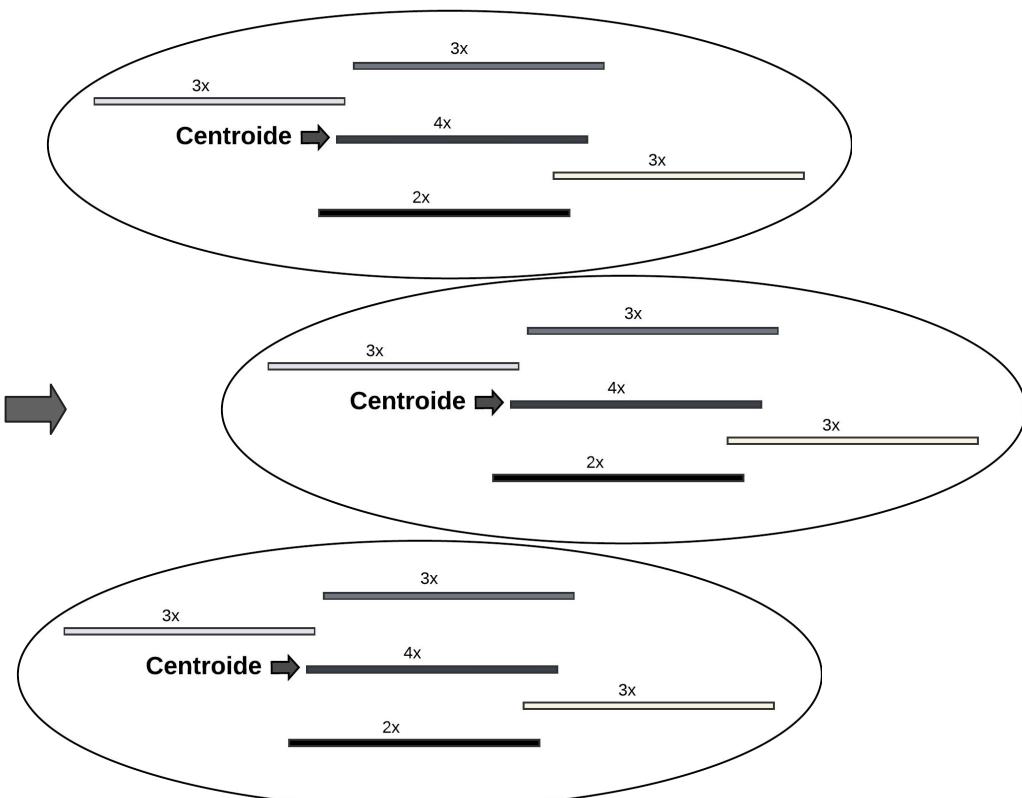
Illumina Miseq error rates as a function of quality.



# DADA2

- Parametrización del modelo de error.

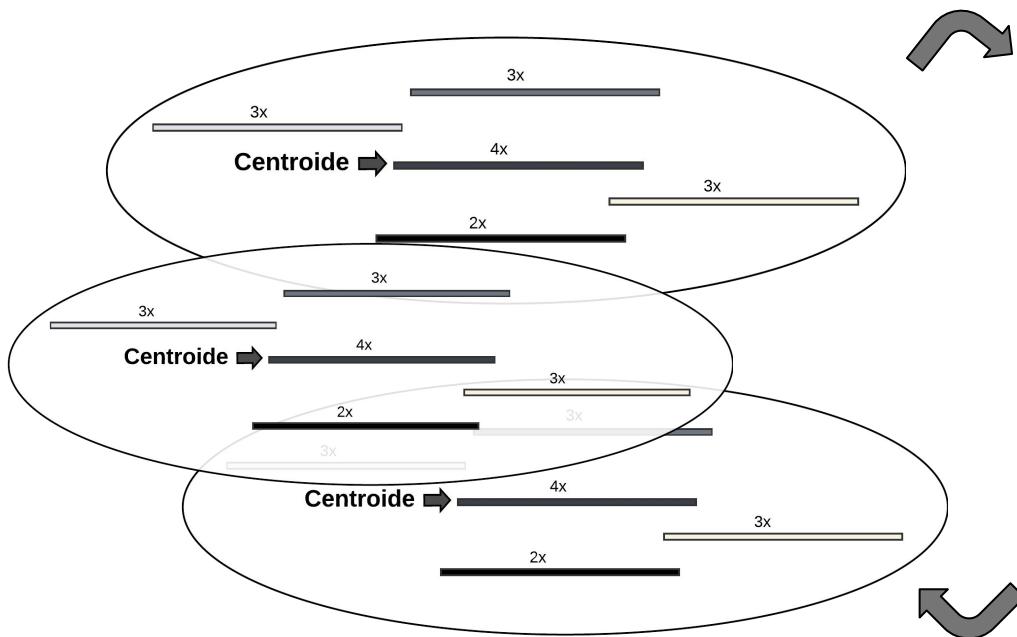
	<b>Q=1</b>	<b>Q=2</b>	<b>Q=3</b>	<b>...</b>	<b>Q=41</b>
pAA	p(A-A,1)	p(A-A,2)	p(A-A,3)	...	p(A-A,41)
pCA	p(C-A,1)	p(C-A,2)	p(C-A,3)	...	p(C-A,41)
pGA	p(G-A,1)	p(G-A,2)	p(G-A,3)	...	p(G-A,41)
...	...	...	...		
pTT	p(T-T,1)	p(T-T,2)	p(T-T,3)	...	p(T-T,41)



# DADA2

## • Parametrización del modelo de error.

Se itera, alternando la inferencia de las particiones (dados los parámetros del modelo de error) con la estimación de parámetros (dadas las particiones) hasta la convergencia.



	Q=1	Q=2	Q=3	...	Q=41
pAA	$p(A-A,1)$	$p(A-A,2)$	$p(A-A,3)$	...	$p(A-A,41)$
pCA	$p(C-A,1)$	$p(C-A,2)$	$p(C-A,3)$	...	$p(C-A,41)$
pGA	$p(G-A,1)$	$p(G-A,2)$	$p(G-A,3)$	...	$p(G-A,41)$
...	...	...	...	...	...
pTT	$p(T-T,1)$	$p(T-T,2)$	$p(T-T,3)$	...	$p(T-T,41)$

## DADA2

- **Pipeline completo para el análisis de amplicones, entre otros, incluye:**
- Dereplicado
- Modelación de errores
- Identificación de ASVs
- Ensamblado de secuencias pareadas
- Identificación de bimeras

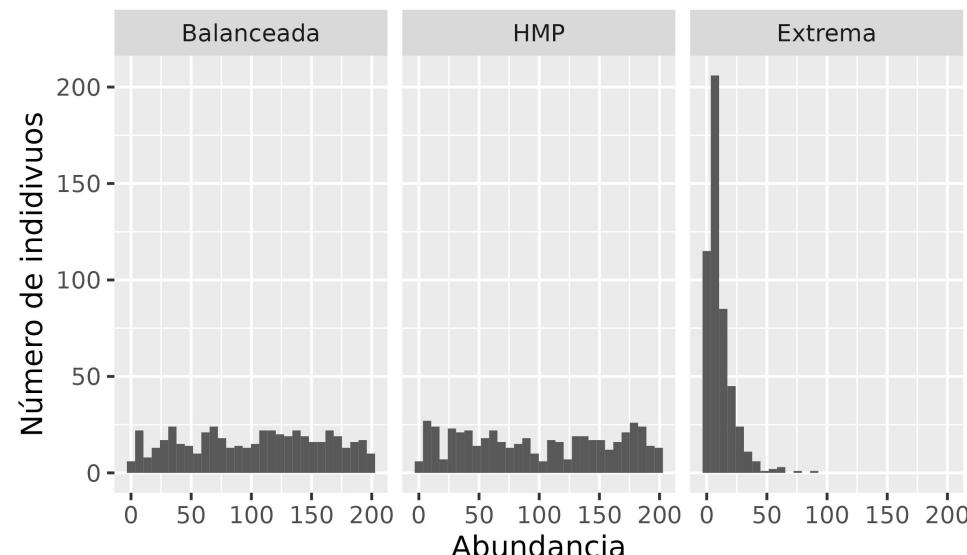
## DADA2

- **Pipeline completo para el análisis de amplicones, entre otros, incluye:**
- Dereplicado
- Modelación de errores
- Identificación de ASVs
- Ensamblado de secuencias pareadas
- Identificación de bimeras

Nota: DADA2 realiza el ensamblado de secuencias pareadas después de la identificación de ASVs. De esta forma, modela de forma independiente los *reads* R1 y R2.

## DADA2

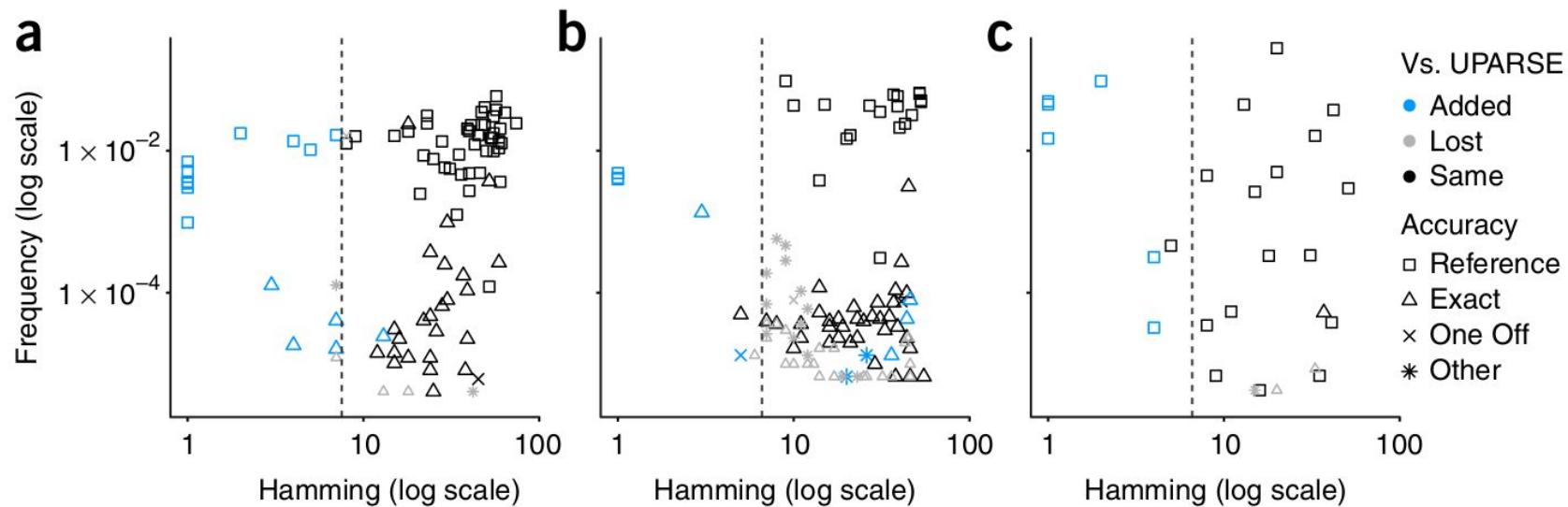
- Evaluación comparativa: DADA2 vs UPARSE (basado en USEARCH, en el cual se basa VSEARCH).
- Tres comunidades referencia (integrando Arqueas y Bacterias) con distinto nivel de diversidad: **Balanceada**, **HMP**, y **Extrema**.



## DADA2

- Evaluación comparativa: DADA2 vs UPARSE (basado en USEARCH).
- Las comunidades fueron secuencias, clusterizadas y comparadas con la referencia conocida, realizando tres tipos de clasificaciones de los resultados.
- ***Exacts***: secuencias con una coincidencia exacta con la referencia u otra base de datos (en el caso de contaminación).
- ***One off***: secuencias con una diferencia con la referencia, u otra base de datos (en el caso de contaminación).
- ***Other***: todo lo demás.

# DADA2



Comparación de ASVs inferidos por DADA2 con OTUs construidos por UPARSE. (a – c) Los ASVs generadas por DADA2 se trazan para tres conjuntos de datos de amplicones de Illumina: (a) Balanceado, (b) HMP y (c) Extremo. La frecuencia se representa en el eje y; La distancia de Hamming a la secuencia más abundante más cercana se traza en el eje x. Cuando una secuencia están bien separadas de otros miembros de la comunidad, los ASVs coinciden en gran medida con las OTU. Sin embargo, DADA2 resuelve variaciones adicionales (azul), especialmente dentro del radio OTU de UPARSE (línea discontinua), al tiempo que genera menos secuencias espurias (One Off y Other).

# Procesamiento de datos: clusterización de secuencias

## **Clusterización:**

De referencia - abierto

De referencia – cerrado

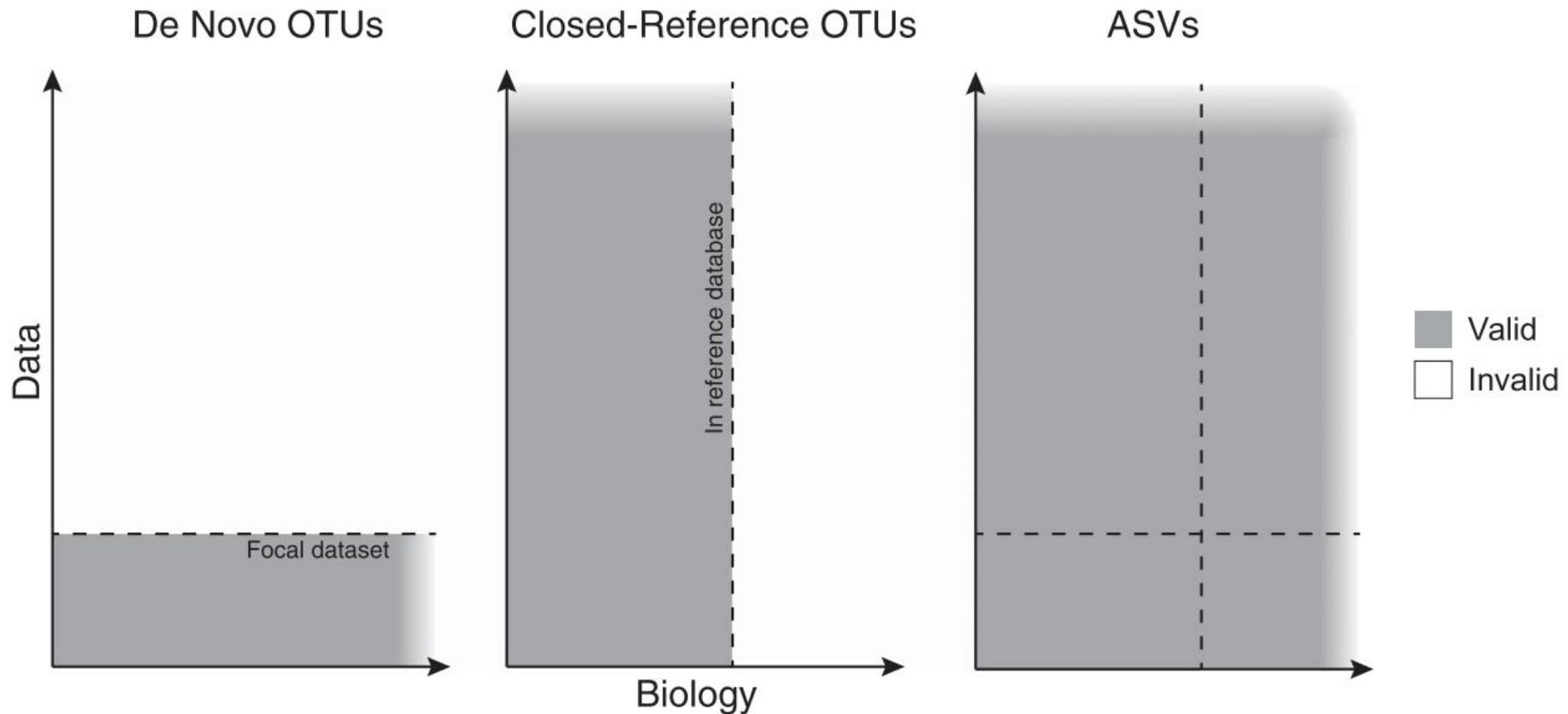
## **De novo:**

Jerárquica

## **Heurística:**

VSEARCH

**DADA2**  **Estado del arte**



# Bibliografía

Wei ZG, Zhang XD, Cao M, Liu F, Qian Y, Zhang SW. Comparison of Methods for Picking the Operational Taxonomic Units From Amplicon Sequences. *Front Microbiol.* 2021 Mar 24;12:644012. doi: 10.3389/fmicb.2021.644012.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ.* 2016 Oct 18;4:e2584. doi: 10.7717/peerj.2584.

Callahan, B., McMurdie, P., Rosen, M. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 13, 581–583 (2016). doi: 10.1038/nmeth.3869

Callahan, B., McMurdie, P. & Holmes, S. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 11, 2639–2643 (2017).  
<https://doi.org/10.1038/ismej.2017.119>