Análisis bioinformático de la diversidad de zooplancton en la zona epipelágica del Golfo de México usando el marcador COI

Grupo Bioinformática

Informe 2020-04-23

## Generalidades del análisis

Los datos provienen de muestras obtenidas en 28 estaciones que cubrió el crucero exploratorio X07. Para cada muestra se generaron amplicones COI. Los amplicones fueron secuenciados desde ambos extremos generando secuencias con una longitud de 300 pares de bases (pb) cada uno, en un secuenciador MiSeq de Illumina (del CICESE-CIGOM). Los siguientes resultados corresponden al análisis bioinformático de las secuencias generadas.

El análisis metagenómico se llevó a cabo en el cluster de cómputo OMICA-CICESE mediante el programa DADA2 versión 1.8. La tubería del análisis fue adaptada de la sugerida por Benjamin Callahan et. al. (2016).

El resumen del análisis se divide en tres etapas principales:

1. Pre-procesamiento y alineamiento de secuencias.
2. Clasificación taxonómica y abundancia relativa.
3. Estimación de los índices de diversidad y presentación gráfica de resultados.

## Indice de figuras

* Figura 1. Promedio del puntaje de calidad Phred a lo largo de las lecturas de las bibliotecas. Bibliotecas Forward (F) y Reverse (R)
* Figura 2. Promedio del puntaje de calidad Phred por posición de nucleótido a lo largo de las lectura. Bibliotecas Forward (F) y Reverse (R)
* Figura 3. Número de secuencias a lo largo del procesamiento
* Figura 4. Panel del tamaño del empalme y amplicon
* Figura 5. Composición de la comunidad a nivel Filo (Barplot).
* Figura 6. Composición de la comunidad a nivel Clase (Barplot).
* Figura 7. Composición de la comunidad a nivel Orden (Heatmap)
* Figura 8. Composición de las comunidades a nivel Familia a lo largo de las estaciones. espectivamente, se indican el nivel de Filos y Clase en las columnas del panel derecho (Heatmap)
* Figura 9. Especies más abundantes a lo largo de todas las estaciones. Se muestran aquellos taxones que representan al menos 10% de la abundancia total por muestra.
* Figura 10. Géneros más abundantes a lo largo de todas las estaciones. Se muestran aquellos taxones que representan al menos 10% de la abundancia total por muestra.
* Figura 11. Curva de rarefacción por muestra. Se presenta el número de ASVs a lo largo de las estaciones en relación al número de secuencias.
* Figura 12. Analisis de diversidad por estación. Se muestran los índices de diversidad calculados para la comunidad de zooplancton cada estación.
* Figura 13. PCoA. Componentes principales (Distancias Jaccard).

## Pre-procesamiento y alineamiento de secuencias.

El pre-procesamiento incluye el escrutinio de lecturas que no cumplen con los parámetros:

Respectivamente, los tamaños 170 y 230 bases de longitud como el valor mínimo para recortar las lecturas de *Forward* (F) y *Reverse* (R). El puntaje de calidad Phred mayor o igual a 15 en promedio. Un umbral de filtrado de calidad para eliminar lecturas con 2 bases erróneamente asignadas en ambas direcciones de las lecturas (F y R). Además, se descartan las lecturas empalmadas que tengan un tamaño menor a 120 bases de longitud y se elimina cualquier lectura que tenga identidad con el genoma del bacteriofago PhiX.

En la Figura 1 y 2 se muestra la distribución y el promedio de la calidad de los nucleótidos, respectivamente, en relación a su posición en la lectura. En este gráfico se observa una buena calidad (≥ 15) por posición a lo largo de toda la lectura. Las lecturas R son significativamente de menor calidad, especialmente al final de la lectura, lo cual es común en secuenciaciones con equipo Illumina.

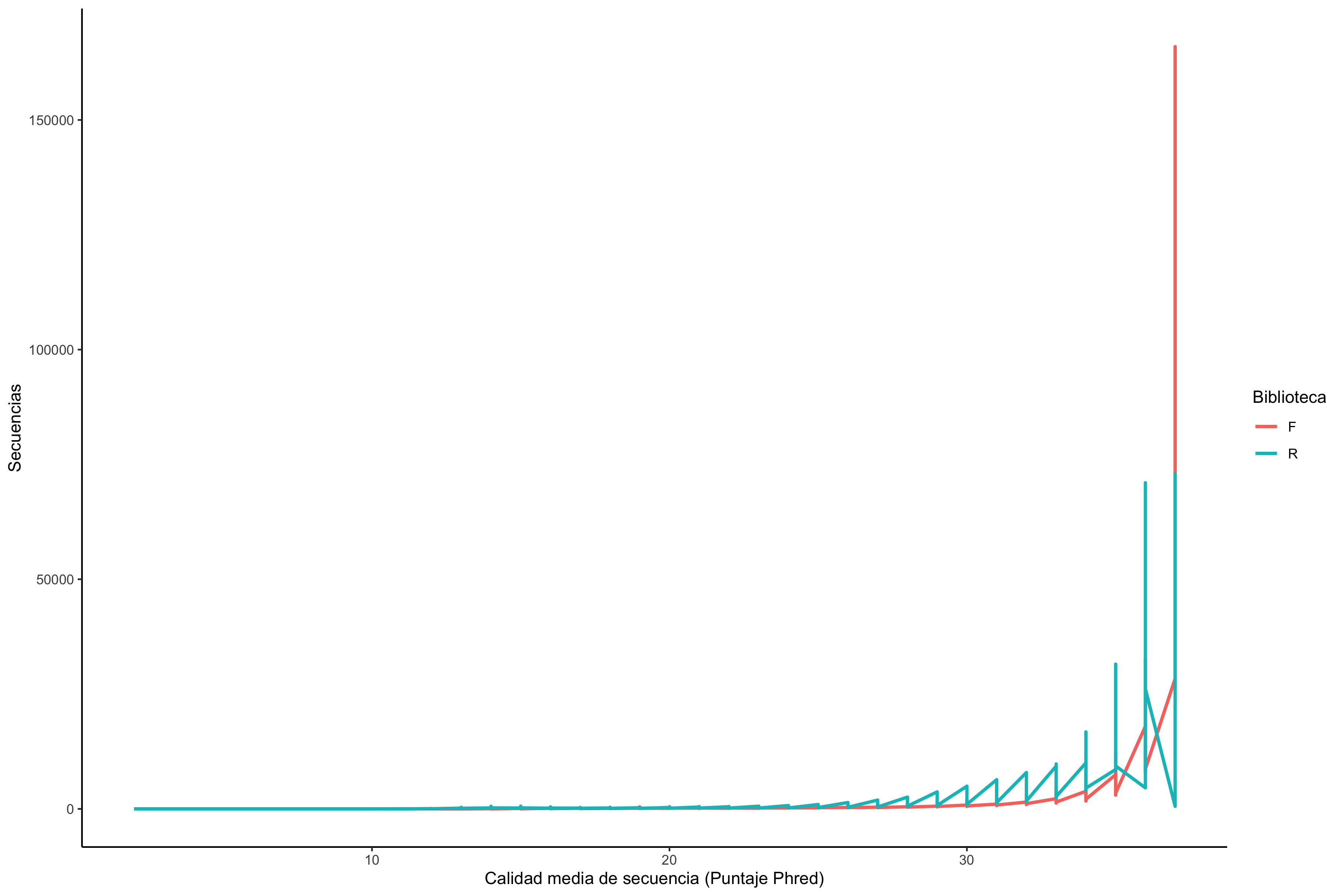


Figura 1. Promedio del puntaje de calidad Phred a lo largo de las lecturas de las bibliotecas. Bibliotecas Forward (F) y Reverse (R).

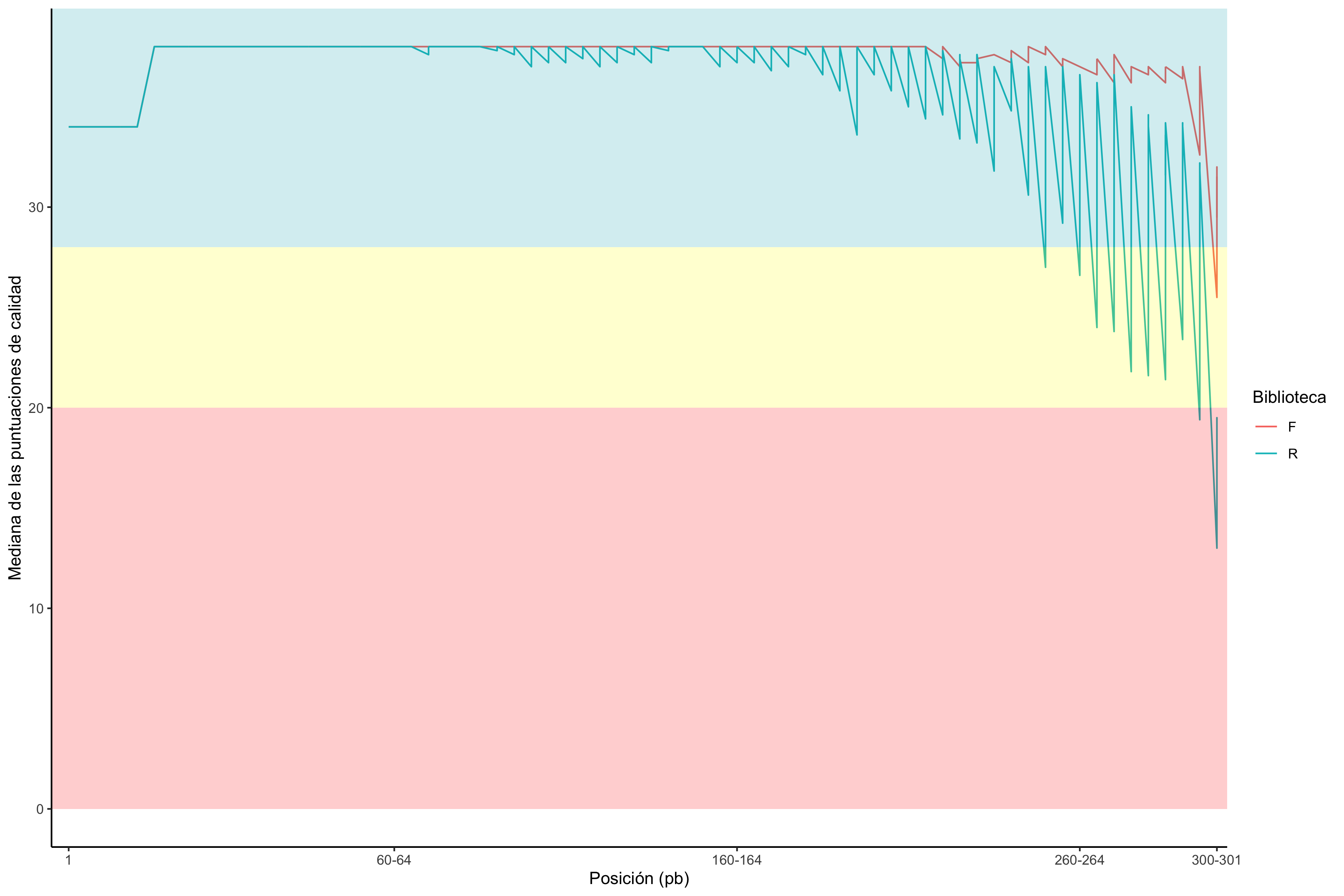


Figura 2. Promedio del puntaje de calidad Phred por posición de nucleótido a lo largo de las lectura. Bibliotecas Forward (F) y Reverse (R).

La Figura 3 muestra un resumen de la cantidad de secuencias por muestra a través de las diferentes etapas del procesamiento de las lecturas. Se denomina ‘amplicon’ a las secuencias resultantes. Cada amplicon representa una secuencia biológica.

El algoritmo de DADA2 reemplaza la selección de unidades operacionales taxonómicas (agrupamiento de similitud de secuencias) por variantes de las secuencias de amplicones (ASV, por su siglas en inglés). La Figura 3 muestra un resumen de las lecturas por muestra a través de las diferentes etapas del flujo de trabajo. Se denomina amplicón ( *ASVs*) a las secuencias empalmadas (no quimericas) que representan una secuencia con variación biológica. Un total de 14165 amplicones fueron obtenidos para el crucero X07. Se consideró el tamaño del empalme como una variable de astringencia durante la generación de amplicones. La figura 4 muestra dos distribuciones, del tamaño del empalme y tamaño del amplicón (ASV) generado.

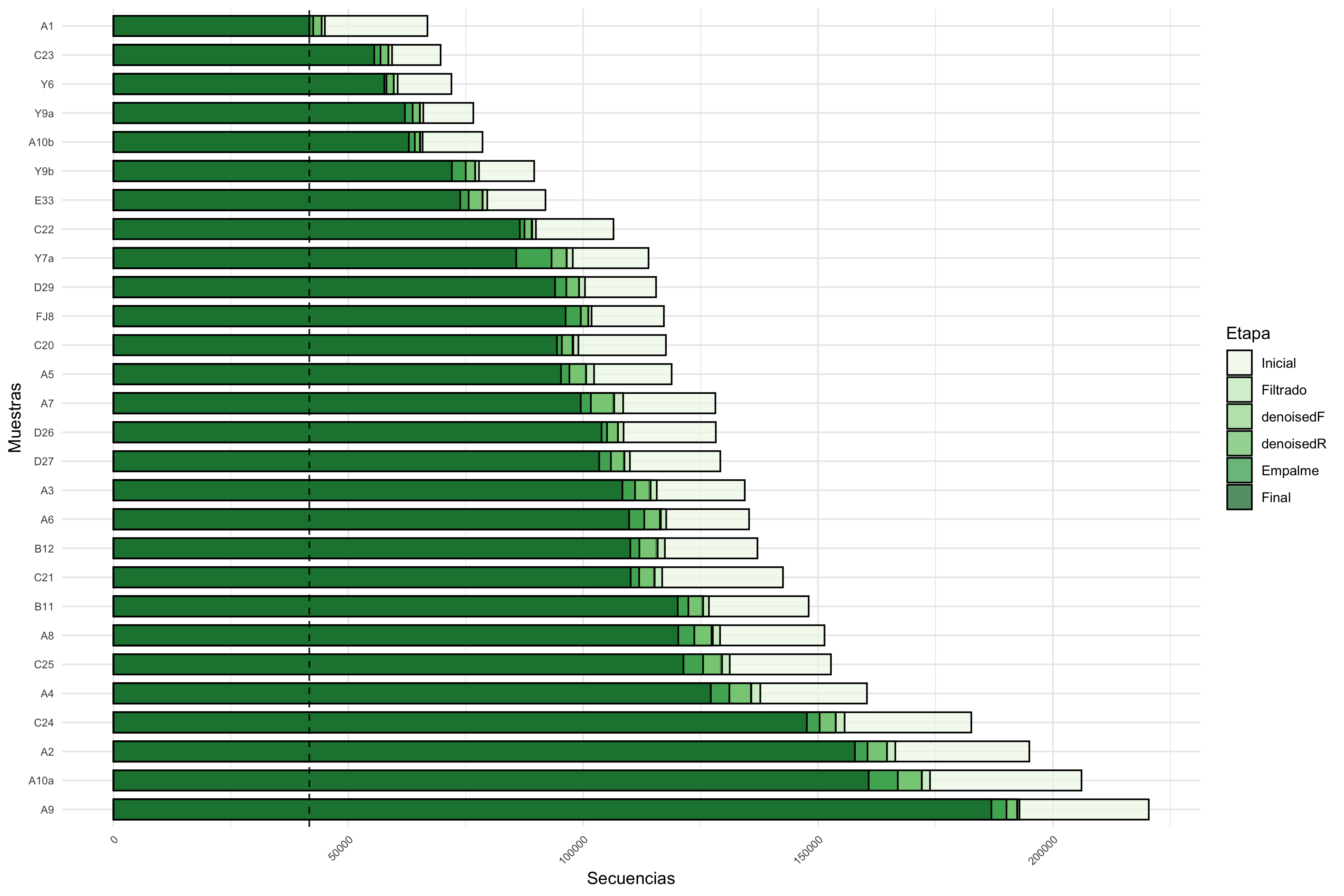
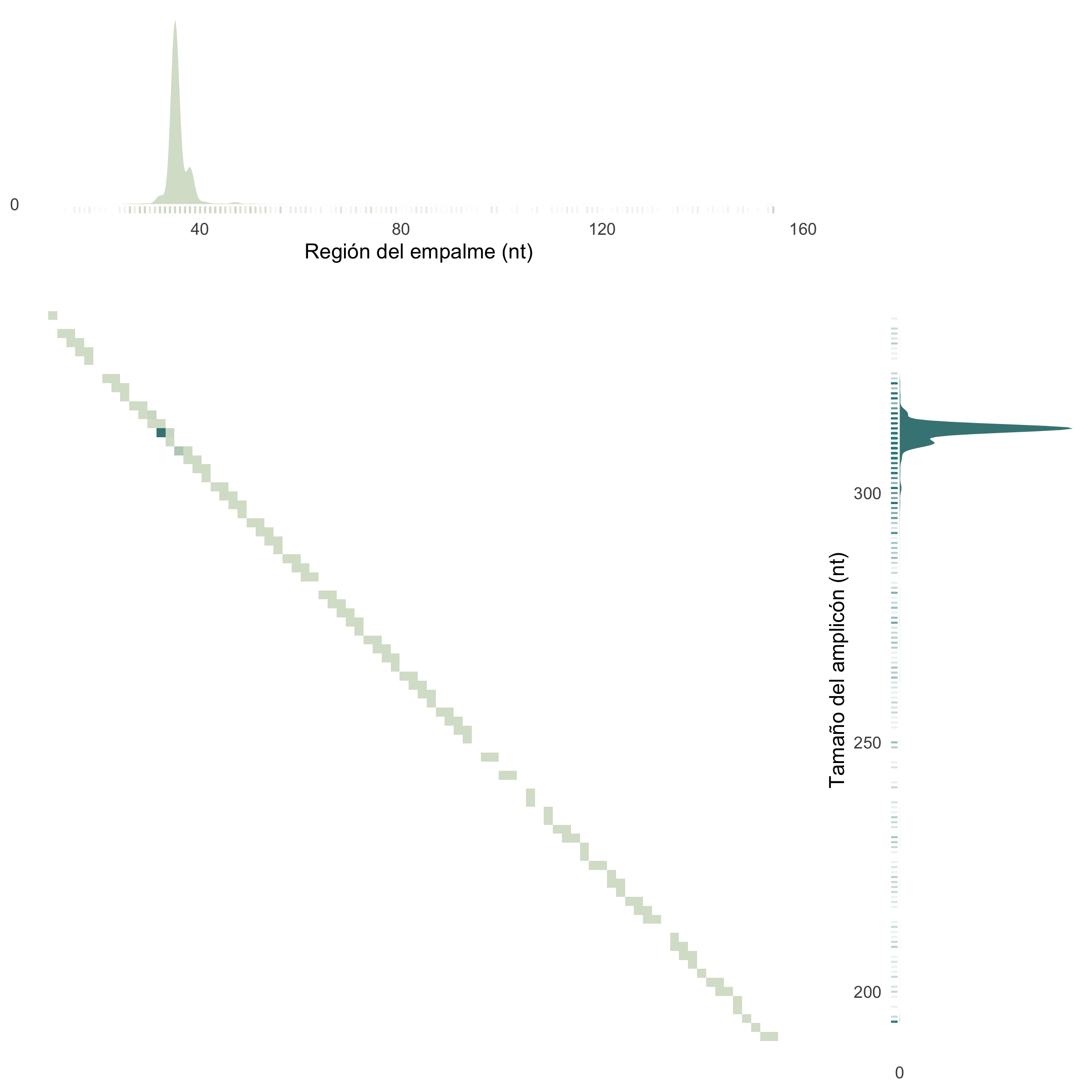
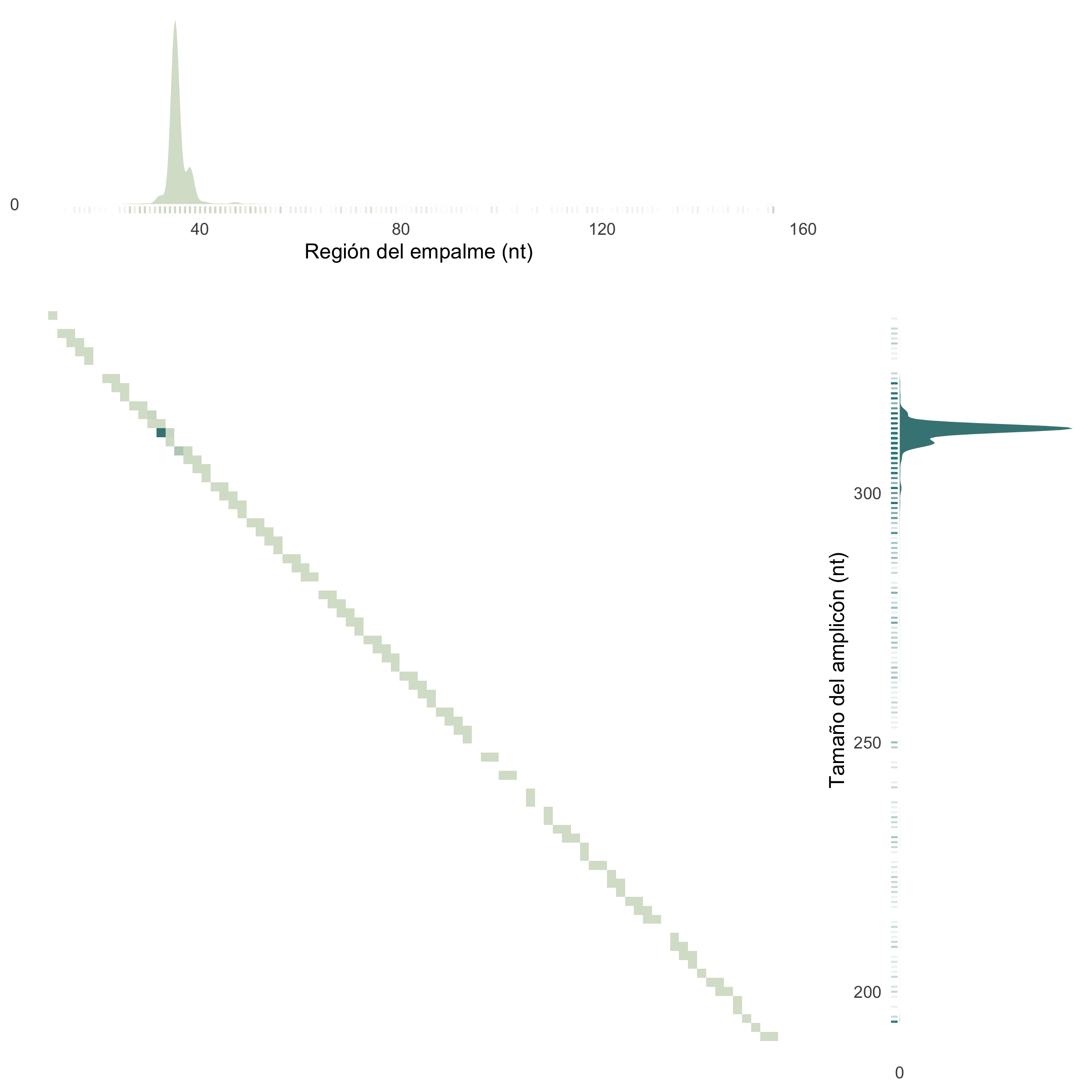


Figura 3. El número de lecturas totales, secuencias filtradas, secuencias empalmadas y secuencias retenidas (quimeras removidas) a lo largo del procesamiento.



Debido a que el número de secuencias para cada muestra individual fue diferente. Con el objetivo de comparar las diferencias entre las muestras, se estandarizó la matriz de distribución de taxones (eg. OTUs, ASVs) en abundancia relativa (RA) basada en la siguiente ecuación:

Donde es la posición de la muestra (1 a ), es la posición del taxon (1 a ) y es el número de secuencias en la muestra i y el taxón de la posición . equivale a la sumatoria de las secuencias en la muestra . Aquí m equivale a 28 muestras y equivale a 9112 taxones.

## Clasificación taxonómica y abundancia relativa

Para la etapa de asignación implementamos la base de datos BOLD. Los amplicones se asignaron utilizando el clasificador bayesiano RDP (wang et. al 2007). El algoritmo de asignación fue configurado con los siguientes parámetros, bootstrap de confidencia = 99, iteraciones = 1000 y tamaño de k-mero = 8 (debido a que cada amplicón representa una variable biológica (amplicones libre de error) se usó un bootstrap de 99 para obtener una asignación astringente. Finalmente, se llevó a cabo la conversión de las asignaciones obtenidas con la base de referencia midori con la taxonomía de worms (world register of marine species) usando un script casero.

La tabla adjunta (X07\_taxones\_identificados.csv) contiene información desglosada de los taxones presentes en este crucero.

Para obtener una perspectiva biológica de la las comunidades se calculó la abundancia de los amplicones colapsados a diferentes niveles taxonómicos. El siguiente grupo de figuras muestra en mapas y graficos de barras de calor la abundancia aglomerada de los taxones por estación a lo largo de los niveles taxonómicos Filo, Clase, Orden y Familia.

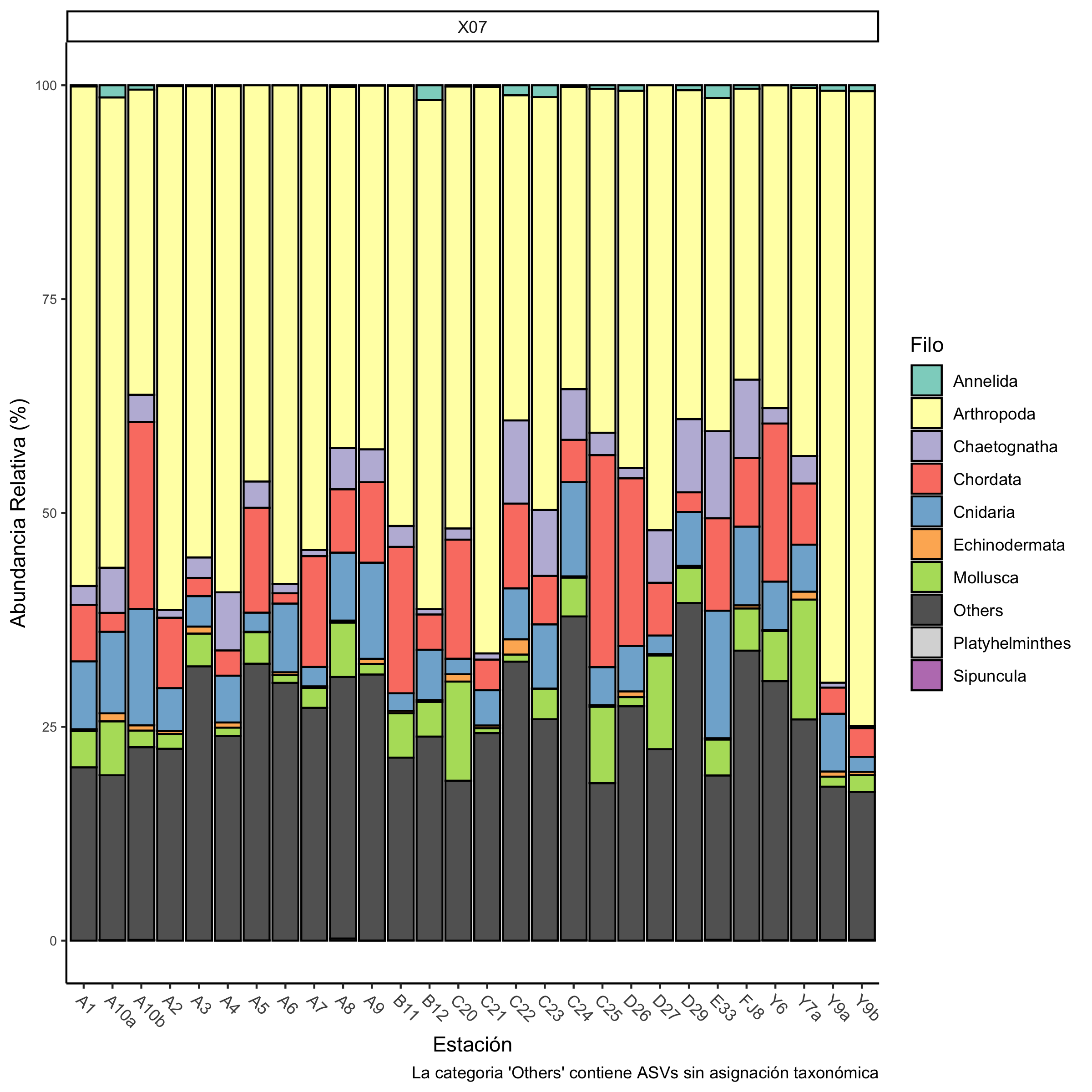


Figura 5. Composición de las comunidades a nivel Filo a lo largo de las estaciones. La categoria ‘Others’ contiene ASVs sin asignacion taxonomica

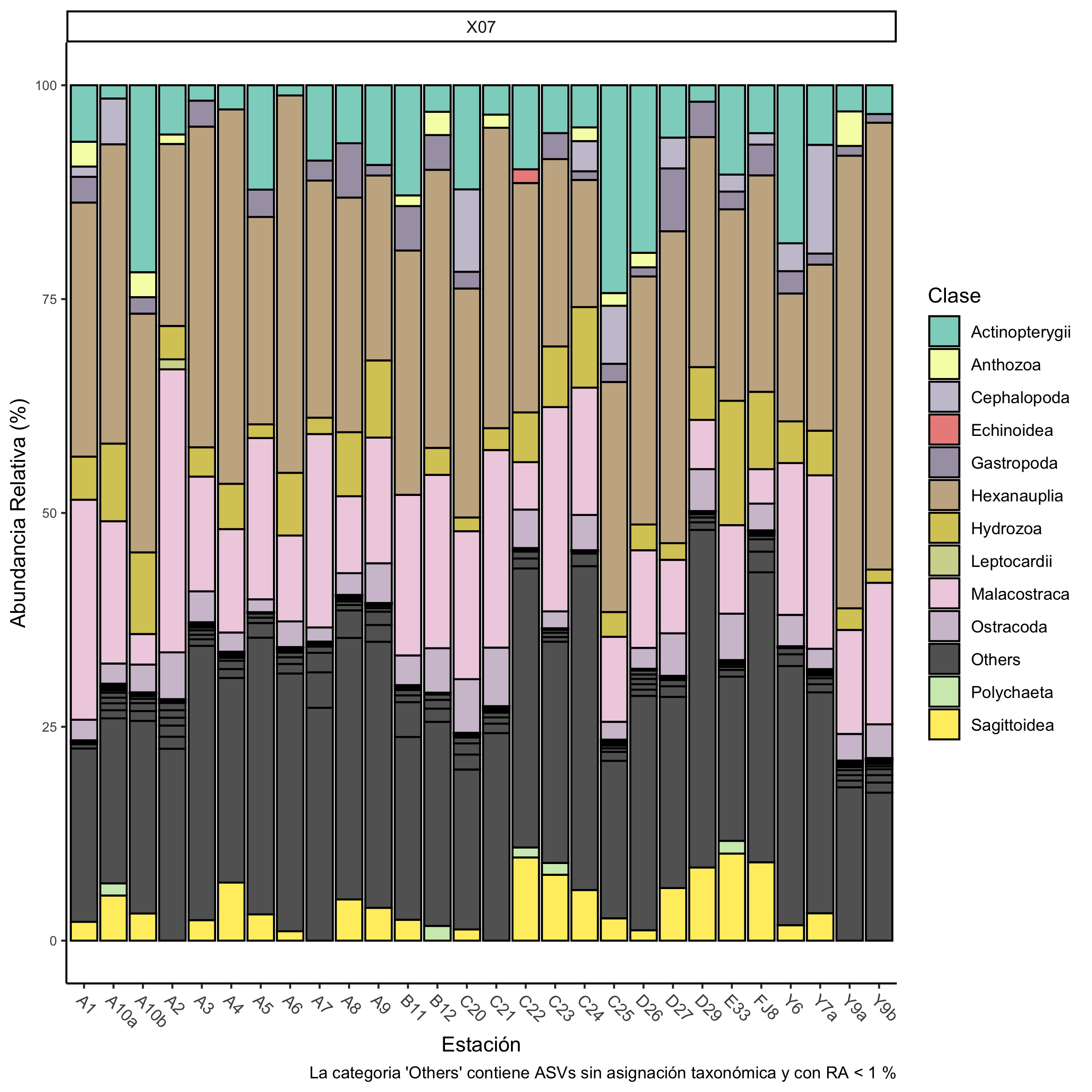
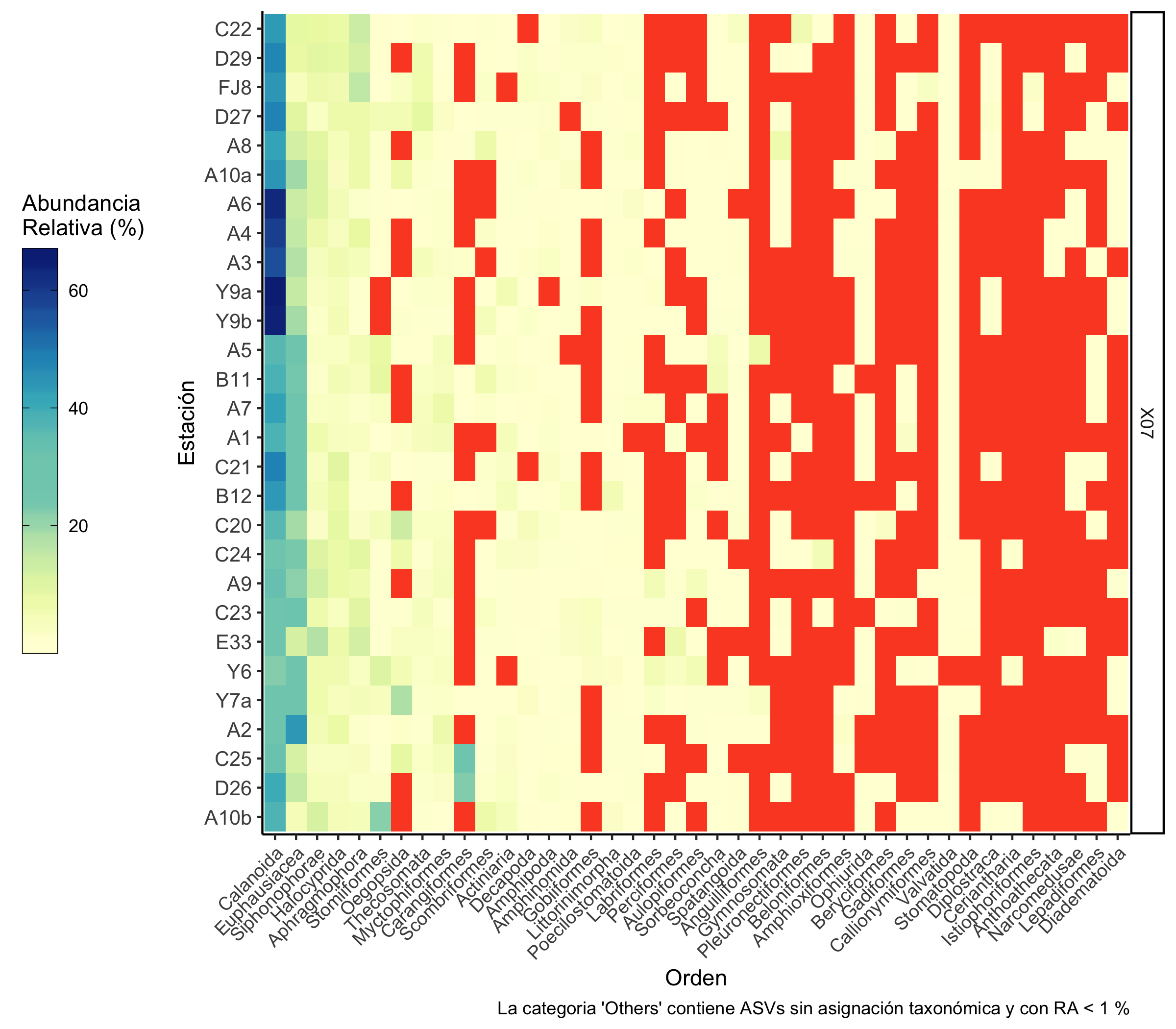
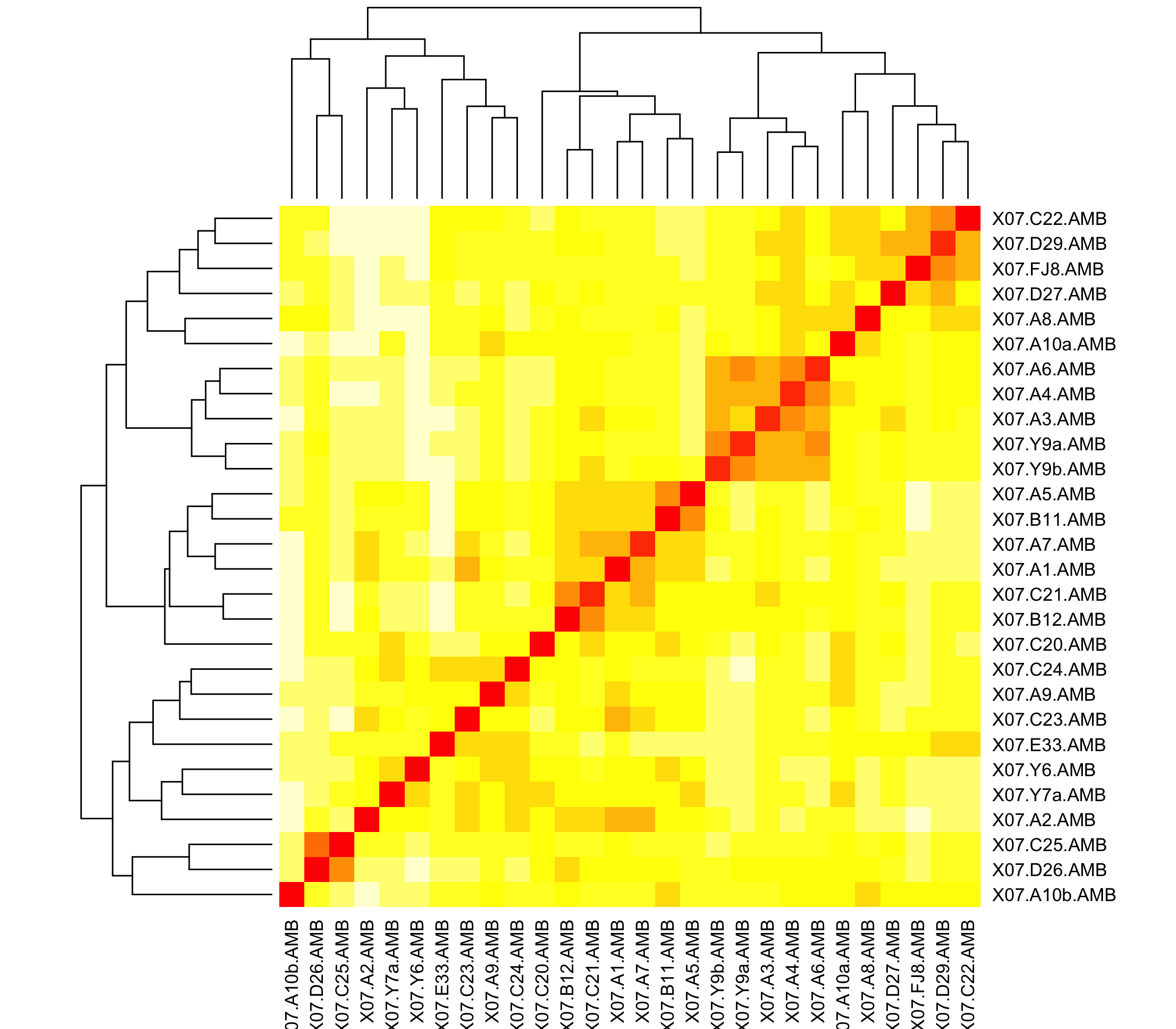


Figura 6. Composición de las comunidades a nivel Clase a lo largo de las estaciones. La categoria ‘Others’ contiene ASVs sin asignacion taxonomica y con RA < 1



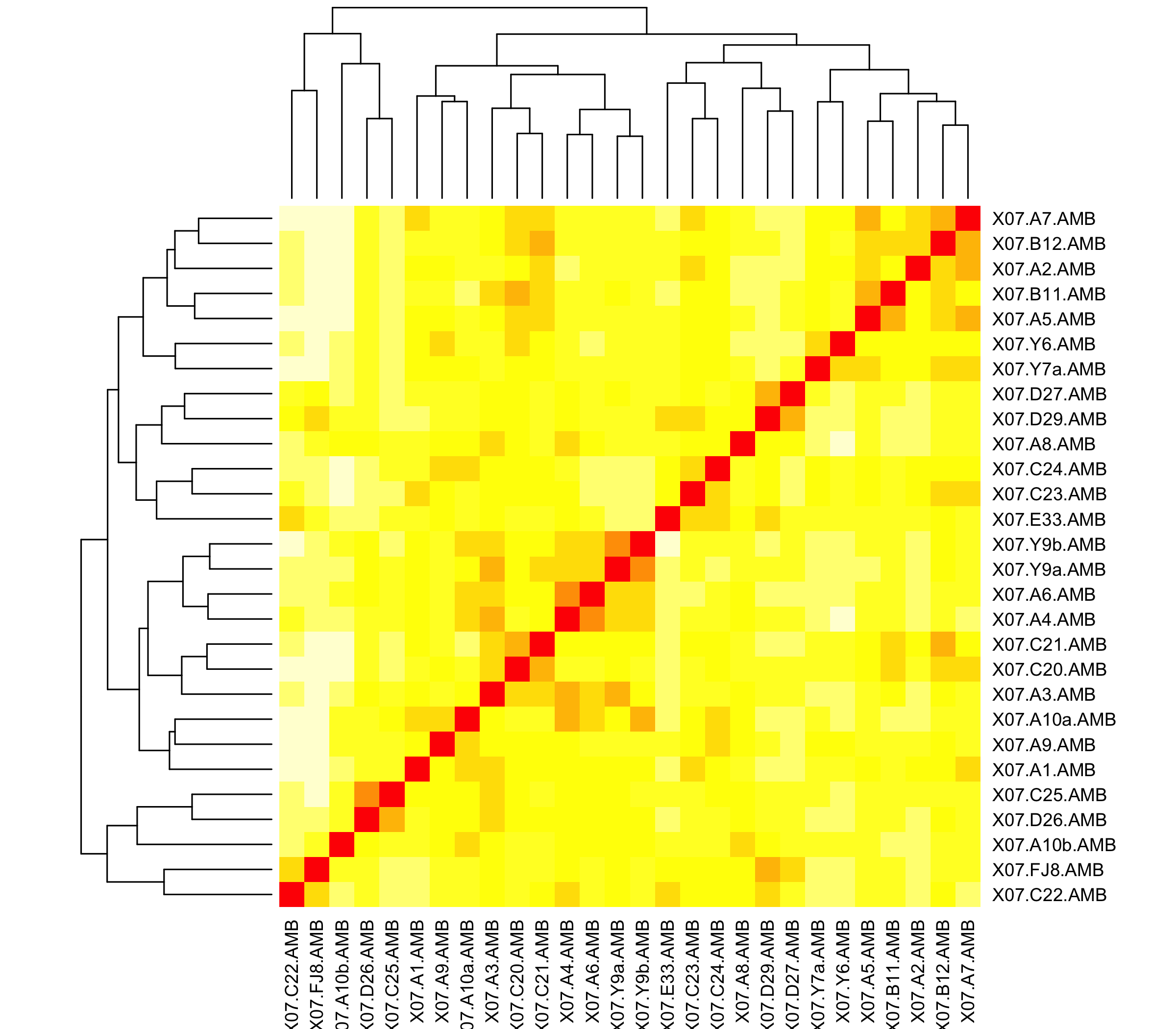


Figura 8. Composición de las comunidades a nivel Familia a lo largo de las estaciones; respectivamente, se indican el nivel de Filos y Clase en las columnas del panel derecho

## quartz\_off\_screen   
## 2

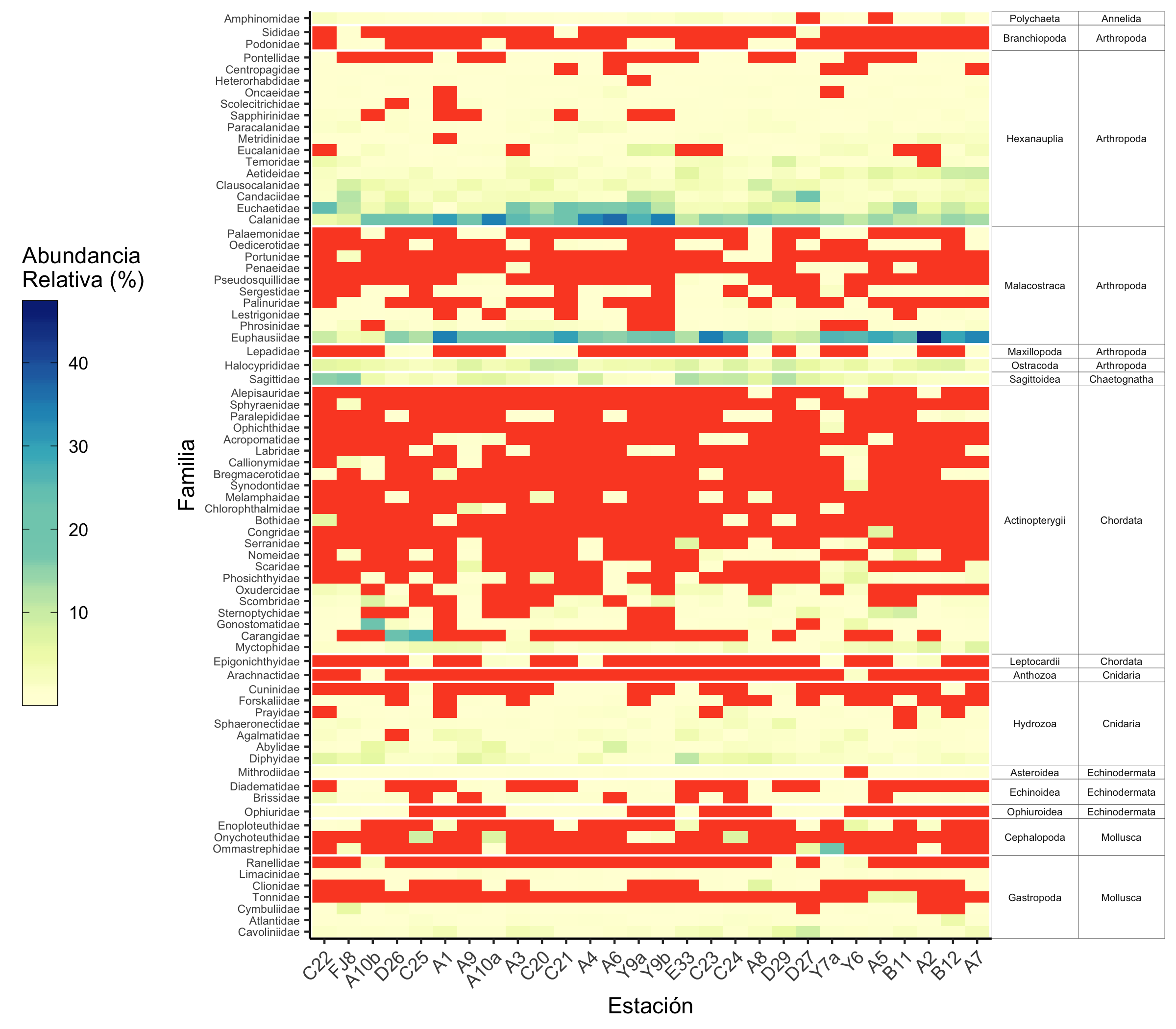


Figura 8. Composición de las comunidades a nivel Familia a lo largo de las estaciones; respectivamente, se indican el nivel de Filos y Clase en las columnas del panel derecho

## quartz\_off\_screen   
## 2

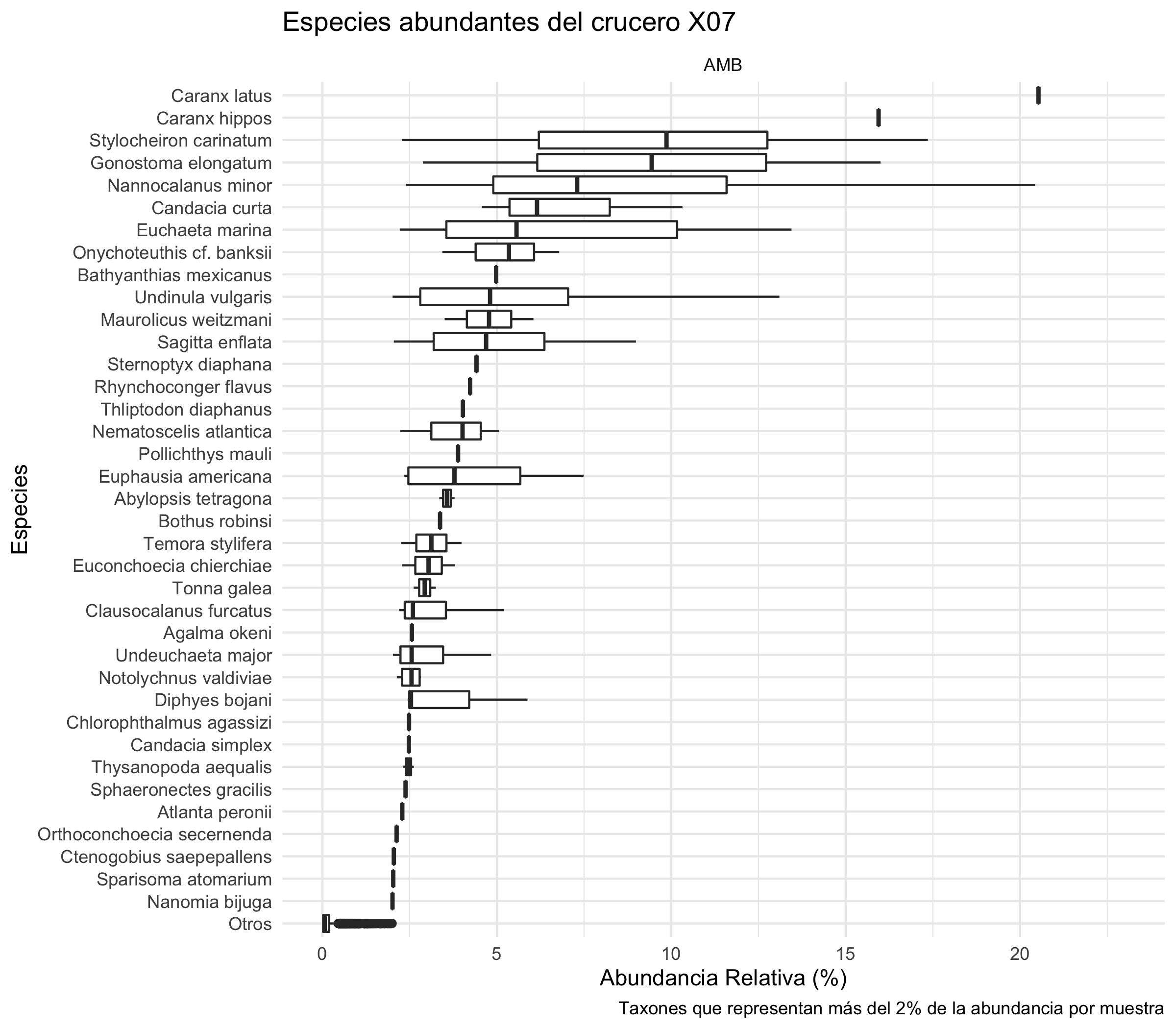
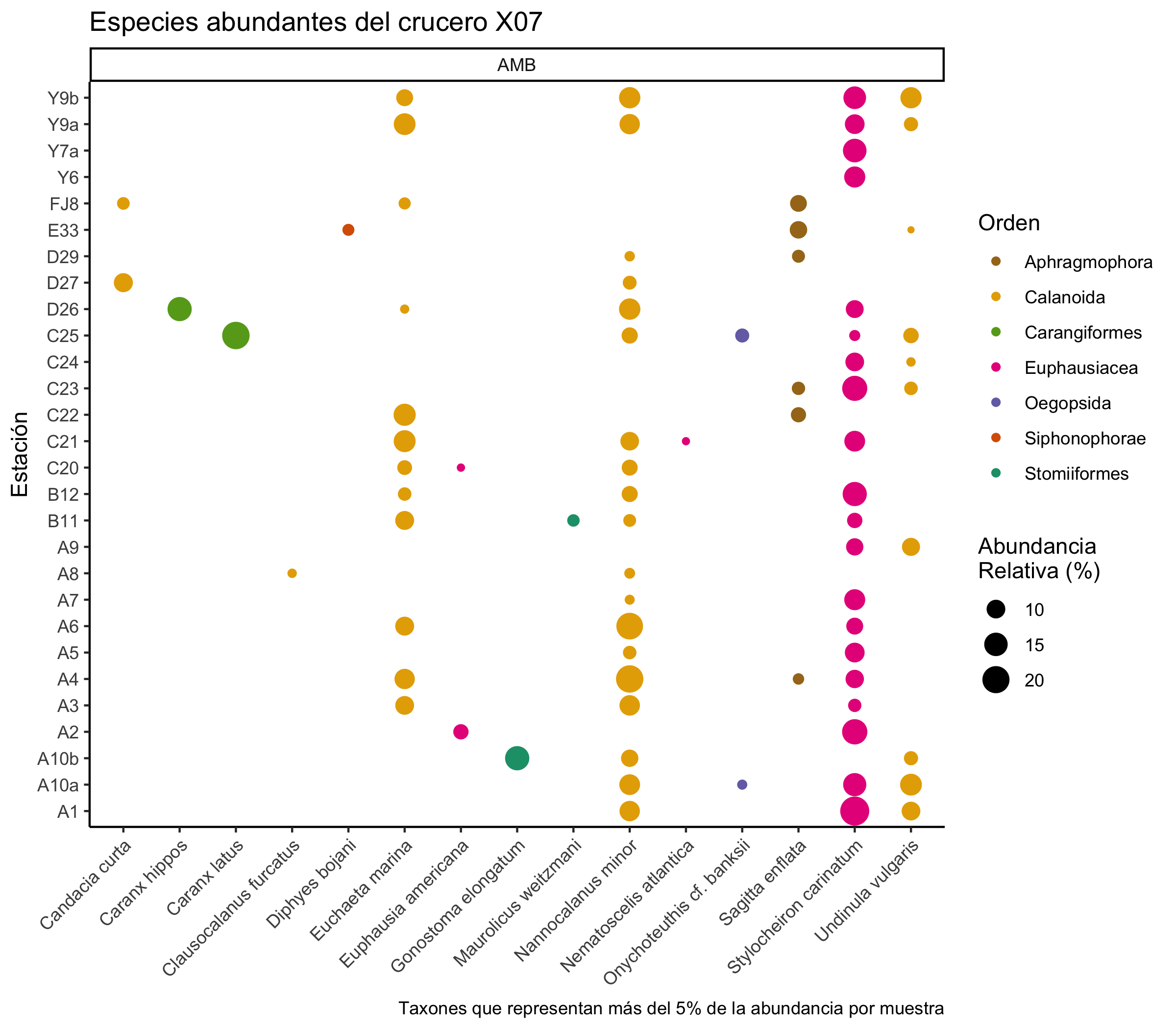


Figura 9. Especies más abundantes a lo largo de todas las estaciones. Se muestran aquellos taxones que representan al menos 10% de la abundancia total por muestra.

## quartz\_off\_screen   
## 2



Especies más abundantes a lo largo de todas las estaciones. Se muestran aquellos taxones que representan al menos 10% de la abundancia total por muestra (Se colorean por orden cada una de las especies).

## Estimación de los índices de diversidad y presentación gráfica de resultados.

La estimación de los índices de diversidad y la generación de representaciones gráficas de los resultados del análisis bioinformático se realizó en R v3.5.0 utilizando diferentes paquetes especializados en análisis de datos ecológicos y gráficos.

Se realizó una curva de rarefacción para evaluar el esfuerzo de secuenciación por estación (Figura 11). Se observó que en la mayoría de las estaciones el número de amplicones puede incrementar aún más si se dedica un mayor esfuerzo de secuenciación.

## quartz\_off\_screen   
## 2

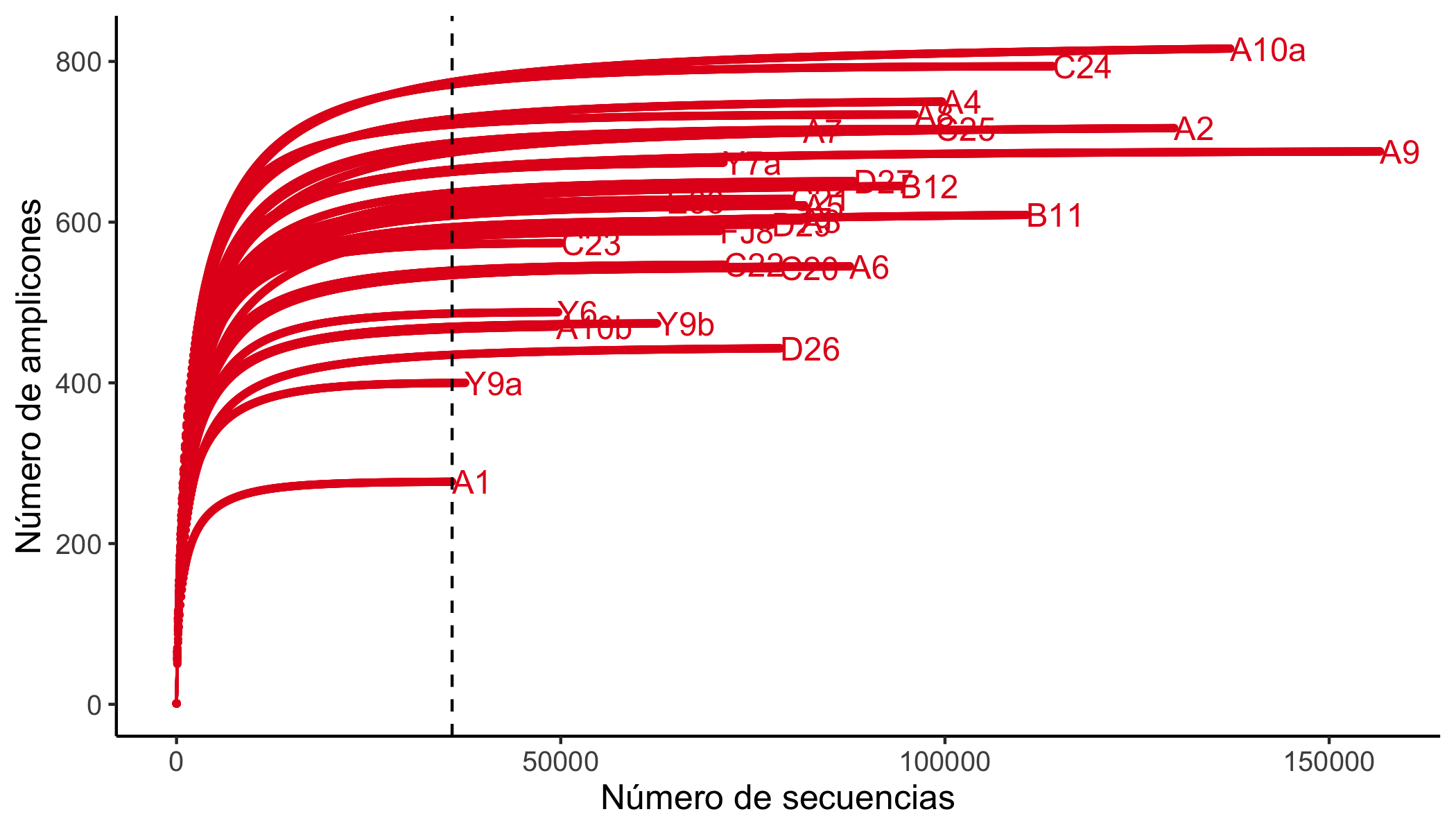


Figura 11. Curva de rarefacción por muestra. Se presenta el número de amplicones a lo largo de las estaciones en relación al número de secuencias.

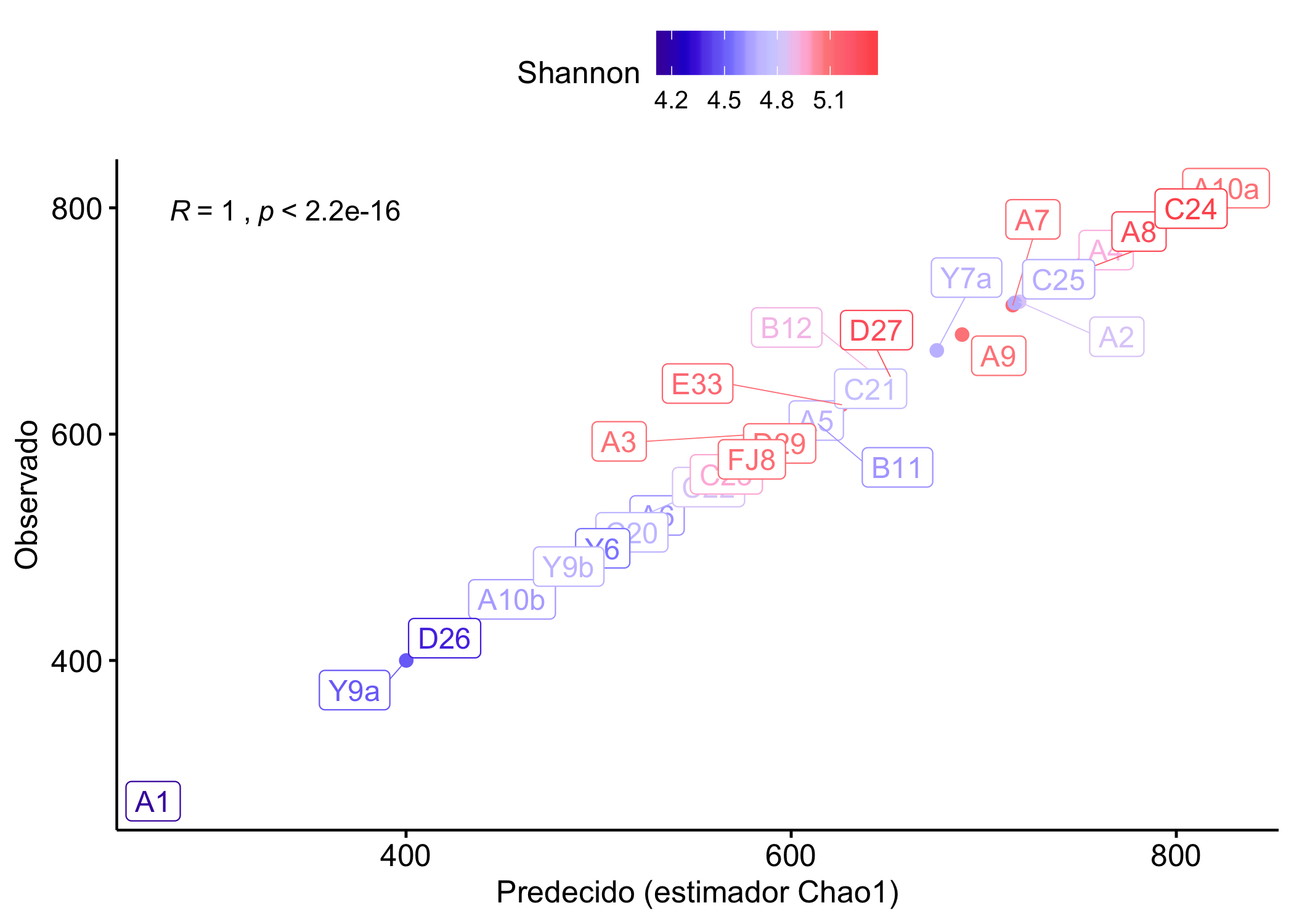


Figura 12. Analisis de diversidad por estación. Se muestran los índices de diversidad calculados para la comunidad de zooplancton cada estación.

La diversidad se obtuvo a través del cálculo de los índices de Shannon, Chao y observado de cada estación (Figura 12). La figura 12 muestra la dispersión de la riqueza de ASVs observados versus la riqueza predicha utilizando los estimadores Chao1 para las estaciones muestreadas. Los colores de los puntos representan el índice de diversidad Shannon para cada estación. La relación 1 indica entre la riqueza observada y predicha. Los valores de índices de diversidad alfa puede encontrarse en la tabla 1.

Tabla 1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Observed | Chao1 | se.chao1 | ACE | se.ACE | Shannon | InvSimpson | Fisher |
| C24 | 794 | 794.08 | 0.32 | 794.37 | 10.78 | 5.34 | 66.19 | 115.08 |
| D27 | 651 | 651.43 | 0.83 | 651.78 | 10.36 | 5.27 | 86.51 | 95.30 |
| A8 | 734 | 734.50 | 1.03 | 734.63 | 9.34 | 5.27 | 69.13 | 108.08 |
| E33 | 626 | 626.20 | 0.53 | 626.55 | 11.08 | 5.12 | 70.54 | 96.35 |
| A7 | 714 | 715.07 | 1.44 | 715.17 | 11.32 | 5.11 | 48.46 | 107.68 |
| FJ8 | 589 | 589.00 | 0.03 | 589.21 | 9.86 | 5.11 | 70.68 | 88.05 |
| A3 | 602 | 602.00 | 0.00 | 602.00 | 9.92 | 5.08 | 54.59 | 88.17 |
| D29 | 597 | 598.25 | 1.73 | 598.10 | 9.41 | 5.08 | 54.52 | 88.04 |
| A10a | 816 | 825.71 | 6.29 | 821.00 | 10.96 | 5.07 | 40.55 | 115.21 |
| A9 | 688 | 688.75 | 1.26 | 689.02 | 7.82 | 5.07 | 42.54 | 92.54 |
| C23 | 574 | 574.27 | 0.66 | 574.57 | 10.21 | 4.95 | 32.42 | 90.92 |
| A4 | 750 | 750.71 | 1.07 | 751.36 | 10.64 | 4.93 | 32.66 | 110.16 |
| B12 | 645 | 649.00 | 3.69 | 646.76 | 10.66 | 4.91 | 33.23 | 93.21 |
| C22 | 547 | 547.30 | 0.70 | 547.63 | 9.59 | 4.84 | 38.64 | 80.61 |
| A2 | 717 | 718.40 | 1.68 | 718.63 | 10.35 | 4.84 | 26.53 | 100.01 |
| C21 | 629 | 629.08 | 0.34 | 629.35 | 10.86 | 4.76 | 28.43 | 93.10 |
| C20 | 543 | 543.30 | 0.70 | 543.52 | 9.46 | 4.71 | 39.91 | 78.62 |
| A5 | 621 | 622.62 | 1.88 | 622.43 | 10.34 | 4.69 | 29.97 | 91.37 |
| Y9b | 474 | 474.33 | 0.76 | 474.61 | 8.57 | 4.69 | 32.15 | 69.70 |
| Y7a | 674 | 675.62 | 1.88 | 675.56 | 11.17 | 4.68 | 24.64 | 103.07 |
| C25 | 716 | 716.07 | 0.31 | 716.39 | 11.00 | 4.67 | 18.62 | 104.48 |
| A10b | 470 | 471.75 | 2.01 | 471.61 | 9.41 | 4.64 | 27.60 | 71.94 |
| B11 | 609 | 614.00 | 4.31 | 611.49 | 10.03 | 4.62 | 34.41 | 84.90 |
| A6 | 545 | 545.30 | 0.70 | 545.64 | 9.00 | 4.62 | 35.66 | 77.51 |
| Y6 | 488 | 488.10 | 0.38 | 488.37 | 9.75 | 4.55 | 24.73 | 75.14 |
| Y9a | 400 | 400.10 | 0.38 | 400.41 | 8.80 | 4.49 | 28.54 | 62.50 |
| D26 | 443 | 443.91 | 1.35 | 444.20 | 8.12 | 4.36 | 24.82 | 61.99 |
| A1 | 277 | 277.20 | 0.62 | 277.37 | 7.14 | 4.14 | 18.01 | 40.87 |

La diferencia en composición de la comunidad entre estaciones se evaluó mediante un análisis de coordenadas principales usando una matriz de distancias Jaccard. En la figura 13 se observa que no hay formación de grupos definidos de estaciones indicando comunidades similares.

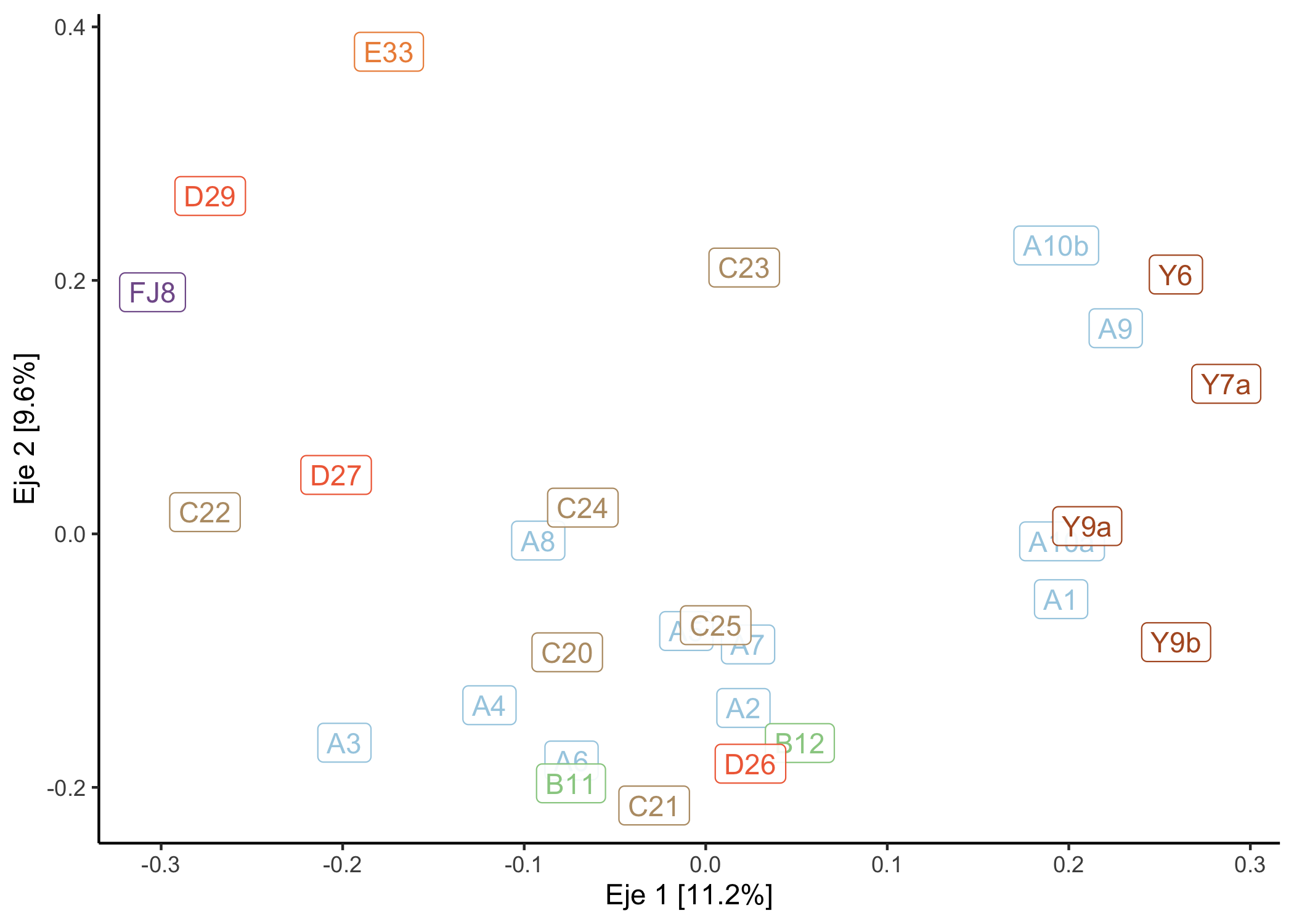


Figura 13. PCoA. Componentes principales (Distancias Jaccard).

## Citas

1. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016). “DADA2: High-resolution sample inference from Illumina amplicon data.” *Nature Methods*, *13*, 581-583.
2. Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The Barcode of Life Data System (<http://www>. barcodinglife. org). Molecular ecology notes, 7(3), 355-364.
3. Wang, Q, G. M. Garrity, J. M. Tiedje, and J. R. Cole. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl Environ Microbiol*. 73(16):5261-7.
4. Callahan BJ, Sankaran K, Fukuyama JA et al. Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses [version 2; referees: 3 approved]. F1000Research 2016, 5:1492
5. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
6. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.