Universidad de la República

Programa de Desarrollo de las Ciencias Básicas (PEDECIBA)

Tesis de Doctorado en Ciencias Biológicas

Utilización de tecnologías de secuenciación de tercera generación para la caracterización de microorganismos

Mag. Cecilia Salazar

Laboratorio de Genómica Microbiana

Institut Pasteur de Montevideo

Montevideo, 2023

Orientador

Dr. Gregorio Iraola

Tribunal

Dra. Claudia Etchebehere (presidente)

Dr. Gustavo Salinas (vocal)

Dr. Paul Cárdenas (vocal)

Agradecimientos

Al Dr. Gregorio Iraola por abrirme las puertas del Institut Pasteur de Montevideo y darme la oportunidad de llevar adelante este trabajo.

A la Dra. Claudia Etchebehere, al Dr. José F. Tort y el Dr. Héctor Romero por haber sido parte de la Comisión de Admisión y Seguimiento de mi doctorado.

Al Dr. Gustavo Salinas, Dr. Paul Cárdenas y la Dra. Claudia Etchebehere por participar como tribunal de esta tesis.

Al Mag. Ignacio Ferrés por sus valiosos aportes en el desarrollo de porefile.

A los compañeros y ex-compañeros del Laboratorio de Genómica Microbiana, de la Unidad de Bioinformática, del Laboratorio de Evolución Experimental de Virus y del Centro de Innovación en Vigilancia Epidemiológica del Institut Pasteur de Montevideo.

A mis padres, a José y a Gea.

Resumen

Uno de los desafíos más urgentes que enfrentan los sistemas de salud a nivel global es la resistencia a los antimicrobianos (RAM). Esta problemática tiene asociadas consecuencias tanto clínicas como económicas para los países, por lo que se han creado guías para la implementación de su vigilancia mediante la identificación de clones de alto riesgo, así como la caracterización de genes que codifican determinantes de resistencia a antimicrobianos y factores de virulencia a través del uso de la genómica en combinación con las metodologías clásicas. El ambiente también juega un rol importante en la diseminación de la RAM. Diferentes especies portadores de genes de resistencia colonizan silenciosamente la microbiota intestinal de poblaciones sanas las cuales alcanzan las aguas residuales urbanas y son sometidas a presiones selectivas por agentes antimicrobianos, metales y desinfectantes allí presentes, convirtiendo el saneamiento en un *hot spot* para la transferencia horizontal de genes de RAM. En este sentido, la implementación de tecnologías de secuenciación portables de tercera generación podrían significar un avance para la caracterización de determinantes de resistencia de forma rápida así como de especies patógenas relevantes para la salud, incluso en el mismo lugar de la toma de la muestra. La metagenómica permite el estudio de la estructura de las comunidades microbianas y la detección de genes de RAM del ambiente así como la caracterización de microorganismos no cultivables y/o desconocidos. A pesar de la disminución de los costos asociados a esta metodología, aún se requieren de procedimientos de relevamiento taxonómico rápido y de bajo costo con capacidad de identificación de alta resolución a nivel de especie. Una de las metodologías más utilizada es la secuenciación del gen 16S del ARN*r* para la detección primaria de microorganismos procariotas a partir de muestras humanas, animales y el ambiente.

El presente trabajo se basa en la implementación de la tecnología de secuenciación Oxford Nanopore Technologies (ONT) para la caracterización genómica de microorganismos procariotas y virales. Específicamente se realizó el análisis genómico retrospectivo de aislamientos recuperados de un brote nosocomial causado por *K. pneumoniae* ocurrido en el año 2017 en una institución hospitalaria local. A partir de los datos genómicos generados exclusivamente con la plataforma portable se obtuvieron secuencias plasmídicas completas donde se identificaron múltiples genes de resistencia, siendo los que codifican para las carbapenemasas KPC-2 y NDM-1 (CR-*Kp*) los más relevantes desde el punto de vista clínico. La generación de genomas de alta resolución con la incorporación de datos de secuenciación

de segunda generación, permitió que se identificara el clon CR-*Kp* ST-11 O2v1:KL64 de distribución global como causante del brote intrahospitalario.

Las tecnologías de secuenciación portables de tercera generación también han sido extensamente utilizadas para la secuenciación de genes marcadores debido a la posibilidad de obtener la secuencia completa del mismo. En este contexto, se desarrolló *porefile* un *pipeline* automatizado para la generación de perfiles taxonómicos a partir de datos de secuenciación de tercera generación del gen 16S del ARNr completo. *Porefile* clasifica las lecturas de secuenciación en base al algoritmo LCA (por *lower common ancestor*) e implementa un paso adicional de pulido a nivel de especie a partir de una base de datos reducida con la finalidad de mejorar la abundancia relativa recuperada a este nivel taxonómico. Utilizando datos simulados, *porefile* mostró resultados similares a los obtenidos con *EMU,* una herramienta basada en el algoritmo EM (por *expectation-maximization)*, en cuanto a la detección a nivel de especie y recuperación de la abundancia relativa esperada de una comunidad de prueba de alta y baja complejidad. Asimismo, mostró niveles similares de recuperación de la taxonomía y abundancia relativa de los géneros más representados cuando se compararon los resultados obtenidos a partir de la región V1-V9 con datos ONT y la región V3-V4 con datos generados con tecnología Illumina para muestras de la microbiota intestinal humana obtenidas de base de datos públicas.

Por otra parte, la genómica ha sido una herramienta fundamental para la vigilancia de variantes de SARS-CoV-2 en el contexto de la reciente pandemia global de COVID-19. La plataforma portable de ONT ha sido utilizada para la caracterización viral en varias regiones del mundo, incluso en regiones de menores ingresos. Si bien aún queda un largo camino por recorrer en cuanto a la implementación de la vigilancia genómica a escalas comparables a las regiones de mayores ingresos, es necesaria la generación de metodologías que permitan realizar la vigilancia de patógenos de forma rápida, confiable y costo-efectiva mientras este camino es recorrido. En este contexto se realizó la caracterización de los primeros genomas de SARS-CoV-2 obtenidos a nivel local en conjunto con otras secuencias uruguayas presentes en la base de datos GISAID correspondiente al mes de marzo del 2020. A partir de estudios filodinámicos, se observaron múltiples introducciones durante el mes de marzo en Uruguay desde distintas regiones que incluyen Sudamérica, Norteamérica y Asia. A su vez, se estima que el virus ya circulaba en el país desde fines de febrero del 2020.

Finalmente, se utilizó tecnología ONT para la generación de secuencias consenso del gen que codifica para la proteína S de SARS-CoV-2. Debido a que gran parte de las mutaciones que definen a una variante de preocupación se encuentran en el gen S, se

implementó un protocolo para la preparación de amplicones solapantes del gen S completo de SARS-CoV-2. Al utilizar esta estrategia, se pudo clasificar gran parte de las muestras en un ensayo piloto, por lo que la metodología podría utilizarse para la clasificación rápida y costo-efectiva de las variantes virales e incluso podría aplicarse a otros marcadores moleculares.

Índice

Introducción general

El monitoreo de microorganismos patógenos emergentes requiere idealmente, de un sistema de vigilancia epidemiológica con la capacidad de detección temprana y mecanismos de intercambio de información rápidos y confiables (1). Un evento de salud pública ocurrido en un lugar puede convertirse rápidamente en una crisis global y en este sentido, la genómica permite generar información con capacidad de ser rápidamente analizada, interpretada y compartida, brindando oportunidad de responder de forma más efectiva frente a potenciales amenazas a la salud (2). Es por esto que la incorporación de estas metodologías en los sistemas de vigilancia epidemiológica podría facilitar el flujo de información a quienes se encuentran en posición de tomar decisiones que conciernen a la salud pública, dentro de un intervalo de tiempo apropiado y buscando minimizar el costo asociado (3).

Los sistemas de vigilancia epidemiológica han evolucionado a lo largo del tiempo y a partir de los mismos se han implementado distintas mejoras, por lo que representan una herramienta cuyo fin último es mantener el bienestar de la población (4,5). La reciente pandemia global de COVID-19 ha dejado en claro la necesidad de contar con estos sistemas de vigilancia activos. Tanto los coronavirus como otros patógenos tienen la capacidad de incorporar cambios en sus genomas que le confieren ventajas para su diseminación, infectividad y rango de hospederos. Contar con metodologías de identificación basadas en genómica de microorganismos emergentes y/o reemergentes, permite la comparación de los datos obtenidos previamente, por lo que su potencial amenaza puede ser detectada de forma temprana. Hasta hace poco tiempo, la mayor parte de los sistemas de vigilancia contaban únicamente con técnicas de aislamiento y tipificación de patógenos (6). Si bien estas metodologías son de uso estándar, es cada vez más frecuente la incorporación de tecnologías de secuenciación masiva en el área de la microbiología. Sin embargo, las mismas se encuentran mayormente a disposición en laboratorios de referencia debido a su alto costo y necesidad de infraestructura y entrenamiento (7,8). A pesar de ello, la aplicación de tecnologías de secuenciación masiva genera un impacto positivo en áreas de vigilancia epidemiológica, investigación de brotes y caracterización de determinantes de resistencia y virulencia de patógenos difíciles de cultivar (8). Asimismo, la generación de datos genómicos a partir de los aislamientos obtenidos de pacientes ingresados en las instituciones hospitalarias puede asistir en la toma de decisiones respecto a los procedimientos de contención una vez identificado y caracterizado un patógeno (9). Una vez instalado un brote

es posible determinar la clonalidad del mismo, rastrear la cadena de transmisión y la distancia genética entre las secuencias obtenidas de las distintas muestras durante la duración del mismo. La reconstrucción de brotes intrahospitalarios permite establecer mejoras en los protocolos de vigilancia epidemiológica así como del muestreo tanto de pacientes y ambiente hospitalario y de los procedimientos de desinfección (10). Por otra parte, un número creciente de de estudios han demostrado la capacidad de predecir la resistencia a antimicrobianos (RAM) de manera precisa a partir de la secuencia genómica, por lo que la implementación de las metodologías de secuenciación masiva en el área clínica permitiría generar perfiles de resistencia de forma rápida y confiable, optimizando los protocolos de uso y administración de antimicrobianos (11).

Una pandemia silenciosa

La RAM se encuentra dentro de las mayores amenazas a la salud pública según la Organización Mundial de la Salud (OMS), afectando tanto humanos como animales y generando un impacto económico significativo (12). La emergencia de microorganismos multirresistentes se da como consecuencia de mutaciones a nivel cromosómico o mediante la adquisición de genes de resistencia de forma heteróloga (13). La exposición a antimicrobianos proporciona una presión selectiva que conlleva el aumento y dispersión de patógenos resistentes. Asimismo, la co-ocurrencia de un consumo aumentado de antibióticos dentro de los centros hospitalarios y el flujo constante de especies patógenas en los mismos proporciona el escenario ideal para la diseminación de la RAM a través de la transferencia horizontal de genes. Diferentes factores juegan un papel importante en esta temática, entre ellos, el manejo de la cantidad de pacientes por unidad e implementación de programas de desinfección y administración de antibióticos (14).

En el 2019 se estimaron 5 millones de muertes a escala global asociadas a la RAM donde 1.27 millones fueron directamente atribuídas a ésta causa. Las infecciones respiratorias fueron las responsables de una gran parte de las muertes asociadas (1.5 millones). A su vez, las infecciones por *Escherichia coli*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumanii* y *Staphylococcus aureus* en asociación con la RAM fueron los causantes del mayor número muertes. Estas especies han sido reconocidas por la OMS como patógenos prioritarios (15). Las regiones de bajos y medianos ingresos o LMICs (por *lower and middle-income countries*) han sido las más impactadas, a lo que se suma la escasez de datos de vigilancia epidemiológica, ya sea por la presencia de sistemas rudimentarios o la recolección de datos incompletos de la cantidad y tipo de antimicrobianos prescritos (16,17).

Más allá de que aún queda un largo camino en la concientización de la importancia del uso racional de antimicrobianos en salud humana y la producción de alimentos, la vigilancia activa es uno de los pilares fundamentales para el manejo de las enfermedades infecciosas multirresistentes. En este sentido, el desarrollo del programa GLASS (por *Global AMR Surveillance System*) de la OMS ha generado recomendaciones para que los países colecten datos del consumo de antimicrobianos, niveles de resistencia y los mecanismos que subyacen estos niveles de resistencia (18). Entre las recomendaciones del programa GLASS se encuentran la implementación de metodologías de secuenciación genómica en combinación con los métodos de caracterización fenotípica para la identificación de clones de alto riesgo y la correlación entre la presencia de factores de virulencia y el desenlace de la infección (19).

Como se mencionó más arriba, la implementación de la secuenciación genómica genera información que permite determinar los mecanismos moleculares de la RAM y su proceso evolutivo, así como la identificación del origen de un foco infeccioso y cadenas de transmisión. En la **Figura 1** se muestra el flujo de trabajo sugerido por la OMS para la incorporación de la secuenciación genómica a los esfuerzos de vigilancia de RAM (19).

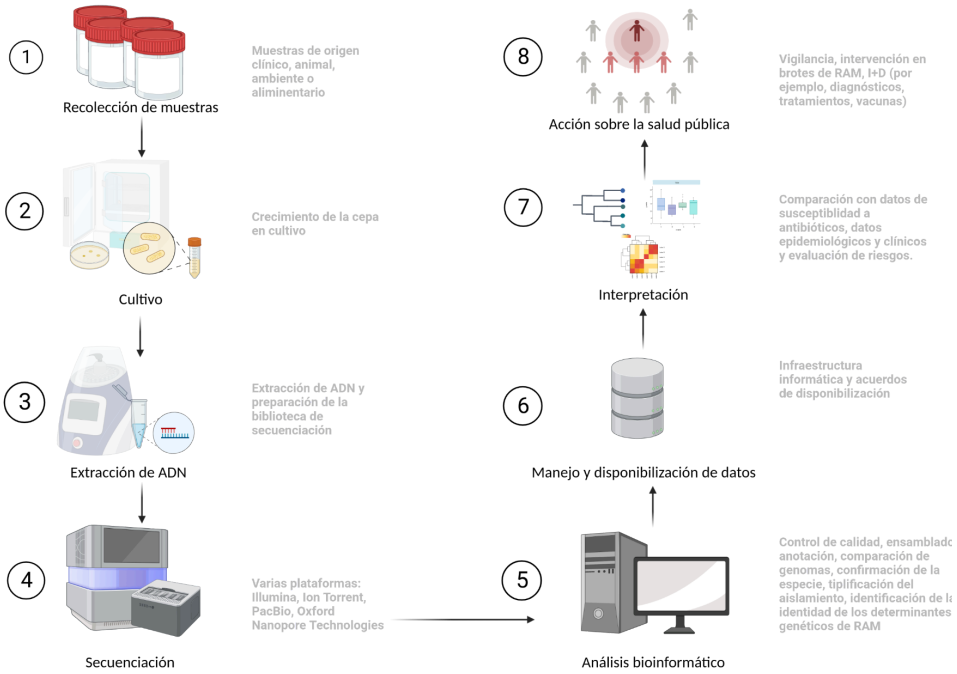

Figura 1: Esquema del flujo de trabajo sugerido para la generación y análisis de los datos de secuenciación genómica recomendada por el programa GLASS (OMS). Adaptado de WHO, 2020 (19). Creado con Biorender.

Los antimicrobianos han sido fundamentales para la medicina moderna, sin embargo, la RAM es un mecanismo natural que no puede ser detenido y esto se debe a la inevitable presión selectiva que se desencadena a través de la administración de antimicrobianos, sin embargo es un mecanismo que puede ser enlentecido. Se estima que el costo asociado a la RAM ascenderá a los 3.000 millones de dólares anuales para el año 2050, estos costos están asociados principalmente, pero no exclusivamente, al tratamiento de infecciones multirresistentes y escalado en uso de los recursos sanitarios, por ejemplo, mayores períodos de hospitalización, estadías más prolongadas en las unidades de cuidados intensivos e infraestructura para generar sitios de aislamiento y prevenir la diseminación de las infecciones (20). Es por este motivo que las acciones frente a la RAM deben ser colectivas y colaborativas entre todas las partes involucradas; estas incluyen agencias gubernamentales y no-gubernamentales, investigadores, responsables políticos, agricultores y pacientes (21).

El uso excesivo de agentes microbianos tanto animales como humanos ha determinado la acumulación de los mismos en el ambiente, impactando a su vez en la selección y diseminación de la RAM. Los genes de RAM se pueden encontrar en aguas residuales urbanas (22), ya que muchos microorganismos portadores de RAM pueden estar presentes en la microbiota de poblaciones sanas. También se encuentran en aguas superficiales y profundas e incluso el agua potable. La acumulación de agentes antimicrobianos, desinfectantes y metales en estos ambientes selecciona los microorganismos resistentes a los mismos, por lo que se ha considerado el sistema de saneamiento como un *hot spot* para la transferencia horizontal de genes de resistencia (23).

El consumo de antibióticos a su vez, perturba en gran medida la estabilidad de la microbiota intestinal, cambiando la composición taxonómica y funcional de la misma y creando oportunidades para la colonización por parte de una gran variedad de microorganismos (24). La perturbación de la microbiota intestinal es conocida como disbiosis, la cual se define como cualquier cambio de los componentes de las comunidades en comparación con las comunidades presentes en individuos sanos. Se caracteriza por la proliferación de patobiontes, pérdida de especies comensales y cambios en la diversidad (25). La utilización de métodos metagenómicos para el estudio de la microbiota intestinal ha expandido el conocimiento respecto a la diversidad microbiana y funcionalidad de la misma, así como la diversidad de genes de RAM y elementos genéticos móviles presentes en la misma (26). Si bien los costos asociados a estos métodos han disminuído en los últimos años, aún se requiere de técnicas de relevamiento taxonómico basados en secuenciación de

marcadores genéticos, como el gen 16S del ARNr, a modo de abordar la composición de las muestras con metodologías más accesibles.

Utilización de marcadores moleculares para la caracterización de comunidades microbianas procariotas

Los estudios sobre la utilización del gen 16S del ARNr (16S) para la determinación de las relaciones filogenéticas entre microorganismos iniciados en los años 60 revolucionaron el área de la microbiología (27). En 1977 Woese y Fox propusieron la utilización del gen 16S como un marcador genético para la identificación taxonómica de microorganismos procariotas y a partir de sus estudios en el área, identificaron el dominio Archaea (28,29). La primera secuencia del gen 16S del ARNr fue obtenida en 1978 y desde entonces este gen ha sido ampliamente utilizado para estudios de caracterización taxonómica de microorganismos. Este gen forma parte del operón ribosomal *rrn* y posee alrededor de 1.550 pares de bases (**Figura 2**). Está compuesto por nueve regiones variables (V1-V9) como resultado de distintas tasas de evolución entre especies y nueve regiones conservadas que sirven como anclajes para cebadores universales de amplificación de las distintas regiones variables. A modo general, la comparación de las secuencias del gen 16S del ARNr permite la diferenciación a nivel de género e incluso a nivel de especies y subespecies (30).

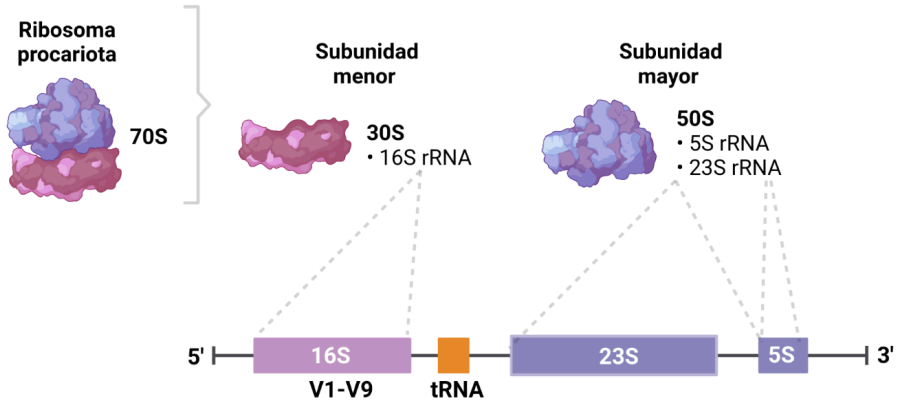

Figura 2: Estructura del ribosoma procariota y de la estructura genética del operón *rrn* del ARN ribosomal. Creado con Biorender

La secuenciación de amplicones de 16S ha sido un procedimiento ampliamente utilizado para la caracterización primaria de comunidades microbianas. El ADN es extraído de la muestra y el gen es amplificado mediante PCR utilizando cebadores cuyo blanco

principal puede ser una o dos regiones hipervariables, el gen 16S completo o incluso el operón *rrn* completo (31). Los amplicones son luego secuenciados y analizados utilizando herramientas bioinformáticas para la clasificación taxonómica y determinación de la abundancia relativa. Los avances en la secuenciación masiva han permitido el desarrollo de protocolos y herramientas que permiten la secuenciación simultánea de múltiples muestras. Es importante destacar que estos procedimientos dependen en gran medida de las bases de datos que alojan la información taxonómica de las secuencias del gen 16S y la calidad de su anotación. En la actualidad, existen más de 23.000 secuencias disponibles en la base de datos NCBI - 16S RefSeq Nucleotide sequence records (32), más de 3.3 millones de secuencias en The Ribosomal Database Project (RDP) (33) y más de medio millón de secuencias completas o casi completas no redundantes correspondientes a la subunidad pequeña del ARNr (SSU Ref NR 99) en la base de datos SILVA (34,35).

Más allá de los sesgos de la información taxonómica alojada en las bases datos, la generación de perfiles taxonómicos tiene asociadas otras limitaciones que justifican la incorporación de controles tanto positivos como negativos para la correcta interpretación de los resultados. Como se mencionó más arriba, la información contenida en las bases de datos limita la información que podemos inferir a partir de las secuencias del gen 16S. Adicionalmente, las bases de datos se encuentran sobrerrepresentadas con bacterias clínicamente relevantes, por lo que los investigadores que trabajan con muestras no-humanas o ambientales encuentran dificultad para asignar taxonomía, particularmente a escalas taxonómicas más profundas (36). Una de las estrategias para abordar esta problemática es la utilización de bases de datos nicho-específicas. En los trabajos de Ritari J. et al., 2015 y Myer P. et al., 2020 se muestra una reducción de secuencias sin asignar al utilizar bases de datos específicas del ambiente, lo que indicaría que un espacio mayor de búsqueda puede ser un factor limitante en la asignación de taxonomía a nivel de especies (37,38).

Las comunidades microbianas en contextos clínicos han sido estudiadas de manera exhaustiva, sin embargo, la secuenciación de ADN de origen ambiental ha revelado una extensa diversidad de microorganismos desconocidos y no cultivables. Adicionalmente, se observa un incremento sostenido de la cantidad de publicaciones en el área del microbioma reflejada en la búsqueda de la palabra clave “microbiome” en PubMed (https://pubmed.ncbi.nlm.nih.gov/) (**Figura 3**). Consecuentemente hubo un incremento en los datos disponibles respecto al repertorio de microorganismos asociados a hospederos y el ambiente (39–43). La generación de perfiles taxonómicos basados en marcadores moleculares como el gen 16S permite estimar la diversidad presente en las muestras y la

cantidad de secuencias sin asignación taxonómica, es decir, la presencia de microorganismos potencialmente desconocidos. Además, permite realizar un relevamiento rápido y de bajo costo de un número elevado de muestras.

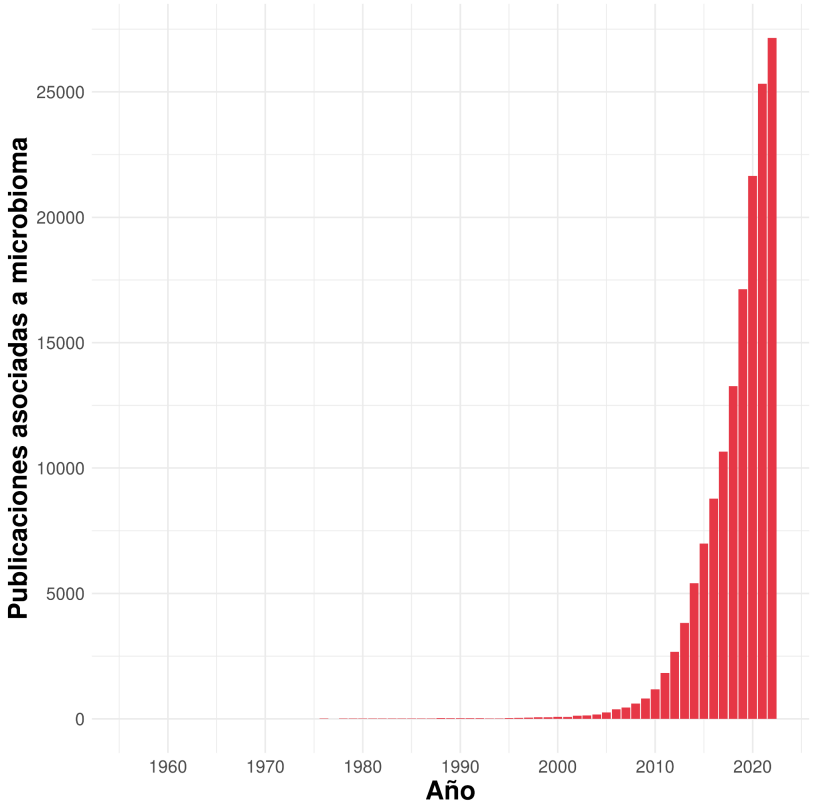

Figura 3: Resultados de la búsqueda de la palabra clave "Microbiome" en PubMed hasta el año 2022.

Como se mencionó anteriormente, desde la aparición de los instrumentos de secuenciación masiva, la generación de perfiles taxonómicos se ha realizado a través de la generación de secuencias parciales del gen 16S del ARNr, por ejemplo las regiones V1-V2, la región V4 o V3-V4 utilizando las plataformas de segunda generación como Illumina. Con el surgimiento de las tecnologías de secuenciación de tercera generación de lecturas largas como Pacific Biosciences (PacBio) y Oxford Nanopore Technologies (ONT), se han reportado una variedad de procedimientos para la preparación de bibliotecas de secuenciación del gen completo 16S (regiones V1-V9) de comunidades microbianas y herramientas de análisis de datos (30,44,45). La utilización de estas plataformas permite superar la limitación impuesta por las tecnologías de segunda generación, donde la secuenciación de una o dos regiones variables introduce sesgos en la detección de diferentes

taxones y a su vez, no permite en general la asignación de taxonomía más allá del nivel de género (46–49). La incorporación de tecnologías de secuenciación de ONT presenta muchas ventajas en este sentido. En primer lugar, se destaca el bajo costo de inversión inicial para la adquisición de la plataforma portátil MinION. No se requiere de infraestructura específica para su instalación en el laboratorio y ONT ha puesto a disposición de sus usuarios varias herramientas de análisis de datos utilizando la interfaz gráfica (https://epi2me.nanoporetech.com/). A pesar de los desarrollos recientes en términos de herramientas para el análisis del gen 16S completo obtenido con tecnologías de tercera generación (44,45), aún se requiere la exploración de distintas estrategias para la correcta manipulación de las lecturas de secuenciación con un porcentaje elevado de error asociado. En un estudio llevado a cabo por Cuscó et al, 2019 utilizando la estrategia de mapeo de los datos de secuenciación contra la base de datos determinó correctamente la taxonomía a nivel de género y especie de una comunidad estándar de prueba y de aislamientos bacterianos (31). Una estrategia similar utilizó Kai et al, 2018, para obtener una correcta clasificación a nivel de especie de los componentes de una comunidad estándar y lo mismo Matsuo et al, 2021 (49,50). Utilizando la estrategia de mapeo contra la base de datos SILVA, Urban et al, 2021 obtuvieron de perfiles taxonómicos generados con el 16S completo a partir de muestras de agua dulce, obteniendo clasificaciones a nivel de especie del 16% de las lecturas de secuenciación, 66% a nivel de género y 77% a nivel de familia (51). Estos estudios indicarían que la estrategia basada en el mapeo de las lecturas de secuenciación utilizando una herramienta como *Minimap2* (52), podría ser útil para interrogar la base de datos conteniendo las secuencias del 16S anotadas, sin embargo, se requieren de nuevas estrategias para la mejora en la clasificación taxonómica de muestras ambientales, así como en la recuperación de la abundancia relativa de sus componentes.

Una pandemia que nos tomó por sorpresa

La utilización de metodologías de secuenciación metagenómica *shotgun* a partir de muestras de lavado bronquioalveolar obtenidas de pacientes hospitalizados cursando una infección respiratoria severa de origen desconocido dió a conocer el agente causal que dio lugar a la pandemia por COVID-19 (*Coronavirus disease* 2019). El nuevo coronavirus del género Betacoronavirus detectado a fines del 2019 (53–55), fue denominado por el *International Committee on Taxonomy of Viruses* (ICTV) como *Severe Acute Respiratory Syndrome Coronavirus 2* (SARS-CoV-2) debido a sus características filogenéticas (56). La infección causada por SARS-CoV-2 se dispersó rápidamente a otras partes del mundo por lo

que la OMS declaró la emergencia sanitaria global el 31 de enero del 2020 y el 11 de marzo de 2020 declaró la pandemia mundial por COVID-19 (57). La respuesta por parte de la comunidad científica para afrontar los desafíos impuestos por un nuevo patógeno humano afectando a la población a escala mundial se ha puesto de manifiesto por la cantidad de literatura asociada a ciencias de la salud, ciencias de la vida, ciencias sociales y ciencias físicas (58). Específicamente en el área de la genómica, fue disponibilizado de forma inmediata un protocolo de secuenciación para SARS-CoV-2 el 20 de enero de 2020 por parte del grupo ARTIC Network (https://artic.network/) en base a sus experiencias previas (59,60). Este protocolo se basa en la amplificación del genoma completo de SARS-CoV-2 utilizando la estrategia Primal Scheme (59). Los esfuerzos de vigilancia genómica a escala global no han tenido precedentes, acumulando a la fecha y luego de 3 años, más de 15 millones de secuencias en la base de datos EpiCov/GISAID (https://www.gisaid.org/) (61). Acompañando la masiva generación de datos genómicos, se establecieron distintos sistemas de nomenclatura para nombrar las variantes del virus que han surgido a lo largo de la pandemia, siendo tal vez las más utilizadas el sistema PANGO y el sistema de letras griegas implementado por la OMS (62,63). Al mismo tiempo, la plataforma Nextstrain (https://nextstrain.org/), respondió rápidamente a la creciente generación de datos genómicos, tomando los datos depositados en las bases de datos públicas, generando análisis filogenéticos a partir de los mismos y permitiendo la visualización en tiempo real de la evolución del virus a escala regional y global (64).

Las regiones de altos ingresos han generado la mayor parte de estas secuencias genómicas depositadas en EpiCoV/GISAID (aproximadamente 13.4 millones). Específicamente, se han depositado más de 7.7 millones de secuencias de origen Europeo, casi 4.5 millones de origen Norteamericano, más de 1.5 millones de origen Asiático y 246 mil de orígen en Oceanía hasta mayo de 2023. Mientras que las secuencias de orígen en Sudamérica y África han contribuido con 413.828 y 162.634, respectivamente (**Figura 4**). Estos datos se relacionan con el número de casos acumulados en estas regiones; según la OMS los países de altos ingresos han acumulado 422 millones de casos de COVID-19, los de medianos a altos ingresos 258 millones, los de bajos a medianos ingresos casi 83 millones y los de bajos ingresos 2 millones de casos confirmados. Sin embargo, estos últimos registran los porcentajes de fallecimientos más altos en proporción a los casos confirmados (https://covid19.who.int/).

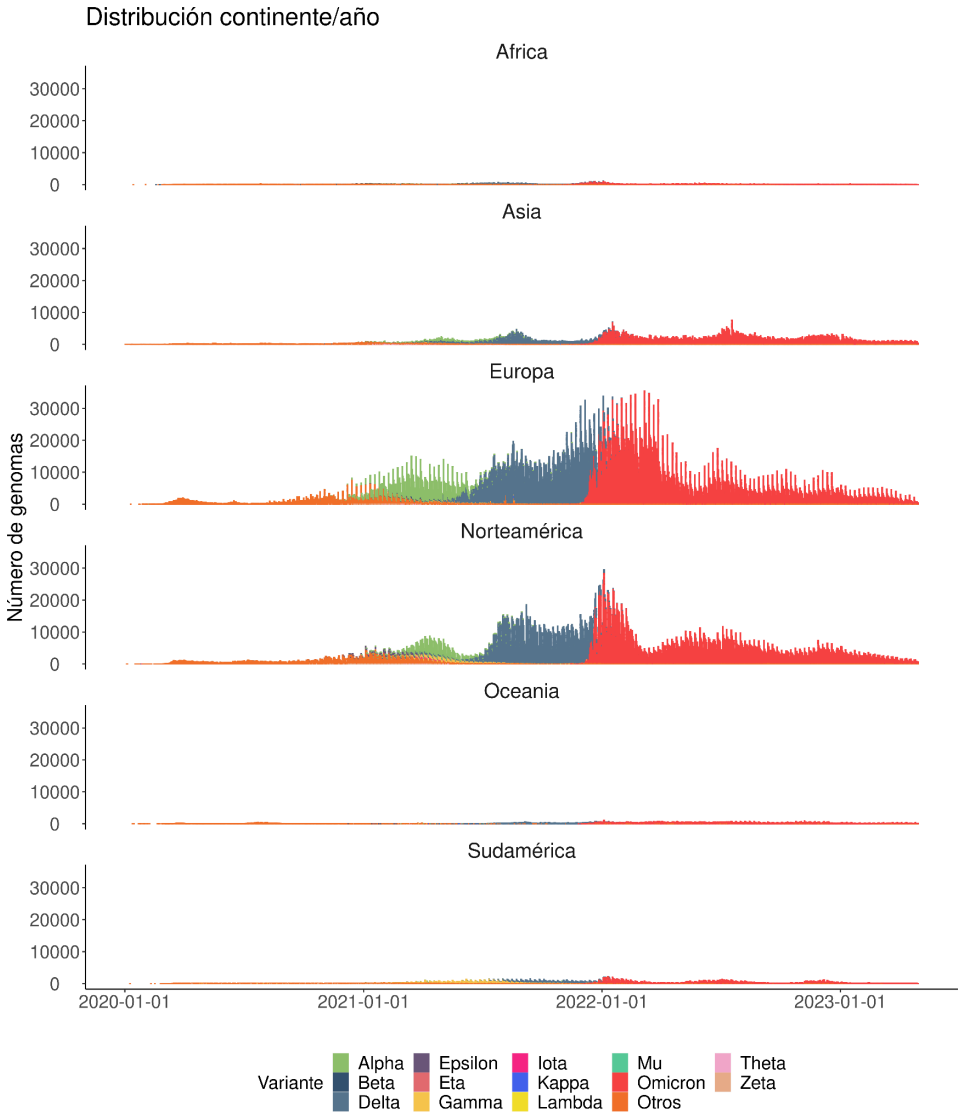

Figura 4: Distribución de las secuencias genómicas de SARS-CoV-2 depositadas en EpiCoV/GISAID a lo largo de la pandemia de COVID-19 en los distintos continentes. Datos obtenidos de https://gisaid.org/.

En un estudio llevado adelante por Brito *et al.* 2023 (65), se demostró que los esfuerzos de vigilancia genómica global se encuentran desbalanceados y esto está ligado con el contexto socioeconómico y las capacidades de los laboratorios previo a la pandemia. Si bien existen distintos aspectos a mejorar respecto a la vigilancia genómica, en particular en los países en vías de desarrollo, la pandemia de COVID-19 dejó en manifiesto la importancia de implementar metodologías metagenómicas para el descubrimiento de nuevos patógenos superando las limitaciones de las estrategias moleculares y fenotípicas utilizadas comúnmente por los sistemas de vigilancia. También ha dejado en claro la necesidad de la implementación de tecnologías costo-efectivas más accesibles, procedimientos estandarizados, acuerdos de

disponibilización de datos y colaboración entre los sectores público y privado con el fin último de maximizar los beneficios a la salud pública global a partir de la vigilancia genómica a escala regional.

Objetivo general

Implementar la utilización de la tecnología de secuenciación de tercera generación (ONT) para la caracterización genómica de patógenos de relevancia para la salud humana y desarrollar tanto protocolos como herramientas para la optimización de recursos y obtención de información epidemiológica acorde a las prestaciones de la tecnología.

Objetivos específicos

1. Realizar el análisis genómico de un brote nosocomial de *K. pneumoniae* a partir de datos de secuenciación obtenidos exclusivamente con la plataforma ONT y realizar un análisis comparativo de los resultados obtenidos con genomas de alta resolución generados con la combinación de datos de secuenciación ONT e Illumina.
2. Desarrollar un flujo de trabajo para el preprocesamiento y asignación taxonómica a partir de datos de secuenciación del gen completo 16S del ARNr obtenidos con la plataforma ONT.
3. Describir la dinámica de SARS-CoV-2 en las primeras semanas luego de su detección en Uruguay a partir de datos genómicos secuenciados a nivel local y obtenidos de la base de datos.
4. Desarrollar un protocolo rápido y costo-efectivo para la asignación de variantes de SARS-CoV-2 a partir de la secuenciación completa del gen S mediante la utilización de la tecnología ONT.

PARTE I:

Caracterización genómica de microorganismos procariotas

Capítulo I

Caracterización genómica de microorganismos de relevancia clínica: *Klebsiella pneumoniae*

Resultados parcialmente incluídos en: Cecilia Salazar, Matias Giménez, Nadia Riera, Andrés Parada, Josefina Puig, Antonio Galiana, Fabio Gril, Mariela Vieytes, Christopher E Mason, Verónica Antelo, Bruno D'Alessandro, Jimena Risso, Gregorio Iraola. Human microbiota drives hospital-associated antimicrobial resistance dissemination in the urban environment and mirrors patient case rates. Microbiome. 2022 Dec 2;10(1):208. doi: https://doi.org/10.1186/s40168-022-01407-8.

1. Introducción

El grupo de patógenos ESKAPE (de la sigla *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae, Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Enterobacter*) es el causante de la mayor parte de las infecciones bacterianas adquiridas en el ambiente hospitalario y una amenaza para la salud debido a las reducidas opciones de tratamientos antimicrobianos disponibles para los mismos (1). Dentro de los patógenos ESKAPE, *K. pneumoniae* es un patógeno gram negativo con fenotipo mucoide capaz de colonizar una gran variedad de ambientes y asociado frecuentemente a infecciones multirresistentes en pacientes ingresados en unidades de cuidados intensivos y/o inmunocomprometidos. La multirresistencia en el ámbito hospitalario viene dada principalmente por *K. pneumoniae* productoras de β-lactamasas de espectro extendido y carbapenemasas (2). Una vez adquirida la infección, *K pneumoniae* coloniza las superficies de las mucosas, incluyendo la nasofaringe y el tracto gastrointestinal. Asimismo, *K. pneumoniae* es capaz de colonizar de forma silenciosa tanto pacientes como al personal hospitalario, dificultando la contención de transmisiones intrahospitalarias (3,4).

K. pneumoniae presenta un genoma de aproximadamente 5-6 megabases, conteniendo aproximadamente 5.000-6.000 genes. Aproximadamente 1.700 genes se encuentran conservados en todos los miembros de la especie (genoma *core*), mientras que los demás genes forman parte del genoma accesorio. Estudios filogenéticos poblacionales muestran que *K. pneumoniae* comprende cientos de linajes, que a su vez se corresponden con grupos clonales definidos por el tipo de MLST del genoma *core* (2). El genoma accesorio de *K. pneumoniae* (compuesto de genes codificados en plásmidos y cromosoma) determina la división de las cepas de *K. pneumoniae* en oportunistas, hipervirulentas y multi-resistentes y a su vez las separa de otras especies estrechamente relacionadas como *K. variicola* y *K. quasipneumoniae* (5). Mientras que *K. pneumoniae* clásica *(*c*Kp)* está asociada con infecciones adquiridas en el ambiente intrahospitalario, *K. pneumoniae* hipervirulenta *(*hv*Kp)* es frecuentemente adquirida de forma comunitaria (2).

Como se mencionó más arriba, la principal problemática de las infecciones causadas por *K. pneumoniae* es la multirresistencia a antimicrobianos. La resistencia se describe en términos fenotípicos (patrones de crecimiento) o genotípicos (presencia y/o expresión de genes) y pueden categorizarse de acuerdo a su origen (intrínseca o adquirida) o según el tipo de resistencia (única, múltiple) (6). La resistencia a antibióticos β-lactámicos es intrínseca en

este patógeno ya que la enzima β-lactamasa está codificada en el genoma *core* de la especie a través del gen cromosómico SHV, pudiendo SHV-1 hidrolizar penicilina y cefalosporinas (7,8). Las β-lactamasas de espectro extendido o BLEE son un mecanismo de resistencia adquirido frecuente en el genoma accesorio de *K. pneumoniae*. Estas enzimas son capaces de hidrolizar cefalosporinas de tercera generación y aztreonam, pero son inhibidas por el ácido clavulánico. Los plásmidos que codifican determinantes de resistencia para antibióticos de este grupo suelen presentar determinantes de resistencia a otros antibióticos y metales pesados (5). Se cree que el grupo clonal 307 (CG307) de *K. pneumoniae* está asociado con infecciones causadas por microorganismos portadores de BLEE. Por otra parte, los antibióticos carbapenémicos han sido la elección para el tratamiento de estas infecciones, sin embargo, las presiones selectivas ejercidas por el uso de estos antibióticos ha determinado el surgimiento de resistencia a los mismos. La resistencia a los antibióticos carbapenémicos está asociada al genoma accesorio y en algunos casos a mutaciones en el genoma *core* a través de la regulación al alza de bombas de flujo, alteraciones en las proteínas de membrana y producción aumentada de BLEE, sin embargo la producción de enzimas hidrolíticas constituyen el mecanismo principal de resistencia (9). En particular, el mecanismo de resistencia asociado a la producción de la carbapenemasa codificada en plásmidos denominada KPC (por *K. pneumoniae carbapenemase*). La transmisión del gen bla_{KPC} está mediada por distintos mecanismos moleculares que incluyen la movilidad de pequeños elementos genéticos como el transposón Tn4401 y la transferencia horizontal a través de la diseminación clonal de plásmidos conteniendo el gen (10). Este perfil de resistencia se encuentra asociado al grupo clonal 258 (CG258). Este grupo incluye el ST258 encontrado en Europa y Norteamérica y el ST11 encontrado principalmente en Asia (5). Otras carbapenemasas han surgido en el genoma accesorio de *K. pneumoniae*, como el gen New Delhi metalloβ-lactamase-1 (NDM-1) y los genes codificados en integrones como el gen de la carbapenemasa Verona-integron encoded metalloβ-lactamase (VIM), Imipenemasas (IMP) y oxacilinasas (OXA) (11–14). La OMS ha catalogado a *K. pneumoniae* productoras de *β*-lactamasas de espectro extendido resistentes al carbapenem (CR-*Kp*) como un grupo crítico debido a las limitadas opciones terapéuticas disponibles para su tratamiento (15). Siete países de la región han reportado CR-*Kp* del grupo clonal 258. Los elementos genéticos móviles asociados a la dispersión de las carbapenemasas más prevalentes son elementos transponibles que incluyen el transposón Tn4401 para el gen bla_{KPC} e ISAba125 para el gen bla_{NDM} localizados en plásmidos de diferentes grupos de incompatibilidad (16).

Otro factor de preocupación es la creciente incidencia de infecciones causadas por *K. pneumoniae* con co-ocurrencia de fenotipo resistente e hipervirulento (CR-*hvKp*). Durante mucho tiempo estos dos fenotipos permanecieron sin solaparse e incluso asociados a grupos clonales distintos, sin embargo en 2018 se registró el primer caso de un brote de *K. pneumoniae* ST11 hipervirulenta y resistente a carbapenem en un hospital de China (17). El estudio genómico del brote mostró que la generación de la cepa CR-*hv*Kp se dió por la adquisición de un plásmido (pLVPK) de aproximadamente 170 kilobases por parte de la CR-*Kp* (18).

Debido a la variedad en los grupos clonales de *K. pneumoniae*, los métodos convencionales de tipificación no tienen la resolución suficiente para distinguir eventos de transmisión relacionados de no relacionados, por lo que el uso de metodologías basadas en genómica para discernir estos eventos es una herramienta con la que podrían beneficiarse los sistemas de vigilancia epidemiológica para la detección temprana de patógenos y la caracterización de brotes en comunidades cerradas (19). Asimismo, las plataformas de secuenciación de tercera generación como la de ONT, permite el análisis de datos genómicos en tiempo real para la rápida identificación de patógenos y los genes de RAM asociados a los mismos.

El objetivo principal de esta sección es determinar las características genómicas que se pueden inferir a partir de los datos de secuenciación de tercera generación obtenidos con ONT. Si bien en la actualidad la tecnología es de para la investigación, debido a la portabilidad y requerimiento mínimos de infraestructura e inversión, esta plataforma tiene el potencial para su implementación en centros hospitalarios para la obtención de información descentralizada, generando datos para la vigilancia genómica de forma rápida y costo-efectiva. Para conocer las limitaciones impuestas para la vigilancia de patógenos, se realizó una comparación con los datos obtenidos con la plataforma Illumina de segunda generación. Si bien las plataformas de segunda generación han sido implementadas como metodología estándar para la vigilancia genómica, se requiere mayor complejidad de infraestructura y una significativa inversión inicial, por lo que las mismas se encuentran principalmente en laboratorios de referencia. La importancia de los laboratorios de referencia a nivel global es indiscutible, sin embargo, la generación de información genómica descentralizada permitirá la puesta en marcha de mecanismos de contención mientras se realiza la validación de los resultados en dichos centros de referencia.

La posterior generación de ambos tipos de datos de secuenciación permite además la investigación más profunda de las cadenas de transmisión. Es por esto, que en este trabajo se

han generado genomas de *K. pneumoniae* de alta resolución para el estudio detallado de los determinantes de resistencia, factores de virulencia, clonalidad del brote intrahospitalario y su relación filogenética con otras secuencias disponibles, así como el contenido plasmídico y mecanismos moleculares asociados a las carbapenemasas detectadas.

2. Objetivo general:

Caracterizar aislamientos de *K. pneumoniae* (CR-*Kp*) obtenidos de un brote nosocomial en cuanto a sus determinantes genéticos de resistencia a antimicrobianos, factores de virulencia, contenido plasmídico y clonalidad utilizando tecnologías de secuenciación de segunda y tercera generación.

2.1 Objetivos específicos:

1. Realizar el análisis comparativo entre la información obtenida a partir de datos de secuenciación ONT, Illumina y la combinación de ambas tecnologías.
2. Analizar el contexto genómico asociado a los genes de resistencia a antibióticos carbapenémicos a modo de determinar el mecanismo más probable de dispersión dentro de la institución hospitalaria.

3. Métodos

3.1 Muestras

Las muestras de ADN genómico procesadas corresponden a 15 aislamientos clínicos de *K. pneumoniae* obtenidos a partir de un brote nosocomial ocurrido entre abril y noviembre del 2017 (**Tabla S1**). Dichas muestras fueron previamente cultivadas en medio LB a 37 ºC y el ADN extraído con el kit de extracción Purelink Genomic DNA kit (Invitrogen). La calidad del ADN se evaluó a través de la medición de la relación de absorción 260/280 nm y 260/230 nm. Su cuantificación se realizó utilizando un método fluorométrico (Qubit™ 1X dsDNA High Sensitivity (HS), Invitrogen™).

3.2 Secuenciación de segunda generación y tercera generación

Aproximadamente 200 ng de cada una de las muestras se enviaron a Sanger Institute (Hinxton, UK) para su secuenciación (2 X 150 pb) en una plataforma Illumina HiSeq 4000.

Una cantidad aproximada de 1.2 ug de ADN de cada una de las muestras fue fragmentada utilizando G-tubes (Covaris) con una media estimada de 8000 pb. La biblioteca de secuenciación para la plataforma ONT se preparó utilizando la estrategia de ligación de índices nativos descrito en el protocolo *Native barcoding genomic DNA by Ligation (SQK-LSK109)* de ONT (https://nanoporetech.com/). Luego de la fragmentación el ADN es reparado y sus extremos preparados con las enzimas NEBNext® FFPE DNA Repair Mix y NEBNext® Ultra™ II End Repair/dA-Tailing (New England Biolabs). La ligación de los índices nativos fue mediada por la T4 ligasa (Blunt/TA Ligase Master Mix, New England Biolabs) y finalmente, los adaptadores de secuenciación fueron añadidos utilizando NEBNext® Quick Ligation Module (New England Biolabs). Aproximadamente 50 fmol de la biblioteca de secuenciación se cargó en una celda R9.4.1 (FLO-MIN106D) y se secuenció en una plataforma ONT MinION Mk1B.

3.3 Pre-procesamiento y análisis de los datos de secuenciación

Para los datos de Illumina se realizó la evaluación de su calidad utilizando la herramienta *FastQC v0.11.9* (20) y se eliminaron adaptadores con *Trimmomatic v0.22* (21). Para el casos de los datos ONT, las muestras fueron demultiplexadas utilizando *Porechop v0.2.4* (22) y filtradas con *NanoFilt v2.6.0* (23). Las lecturas de secuenciación menores a 200 pb y con una calidad menor a 8 no fueron consideradas para el análisis posterior de las secuencias. La clasificación taxonómica primaria para cada muestra se realizó con *Kraken2 v2.1.1* (23) y la tabla de resumen se obtuvo con *Pavian v1.0* (24).

3.4 Ensamblado de genomas de *K. pneumoniae*

A modo de generar los ensamblados completos y de alta resolución de las muestras de *K. pneumoniae* se combinaron los datos de secuenciación Illumina y ONT. Para ellos se utilizó *Unicycler v0.5.0* (24,25) utilizando los parámetros estándar. Los ensamblados se evaluaron con la herramienta *Quast v5.0.2* (26,27) y se visualizaron utilizando *Bandage v0.8.1* (28,29). A modo de comparación, se realizó el ensamblado de las muestras de *K. pneumoniae* solamente con las lecturas cortas de Illumina utilizando *Spades v3.15.5* (30,31) y por otro lado, solamente con las lecturas largas obtenidas con la plataforma ONT utilizando para ello *Flye v2.9.1* (32,33) (**Figura A**).

3.5 Tipificación y caracterización de los determinantes de resistencia a antimicrobianos, factores de virulencia y clasificación de plásmidos de las muestras de *K. pneumoniae*

La tipificación multilocus de secuencias a partir de los genomas ensamblados de *K. pneumoniae* se realizó con la herramienta *MLST v2.23.0* (34). La caracterización de los determinantes de resistencia y contenido plasmídico se realizó con *abricate v1.0.1* (35) utilizando la base de datos CARD (36) y plasmidfinder (37), respectivamente. Para la determinación de los factores de virulencia se utilizó la base de datos VFDB (38). Se utilizó *IntegronFinder 2.0 v1.5.1* (39,40) para la clasificación de integrones y *Kleborate v2.3.0* (41–43) para determinar el tipo de antígeno capsular presente en las muestras. La anotación de los genomas de alta resolución se realizó con *Bakta v1.7.0* (44,45).

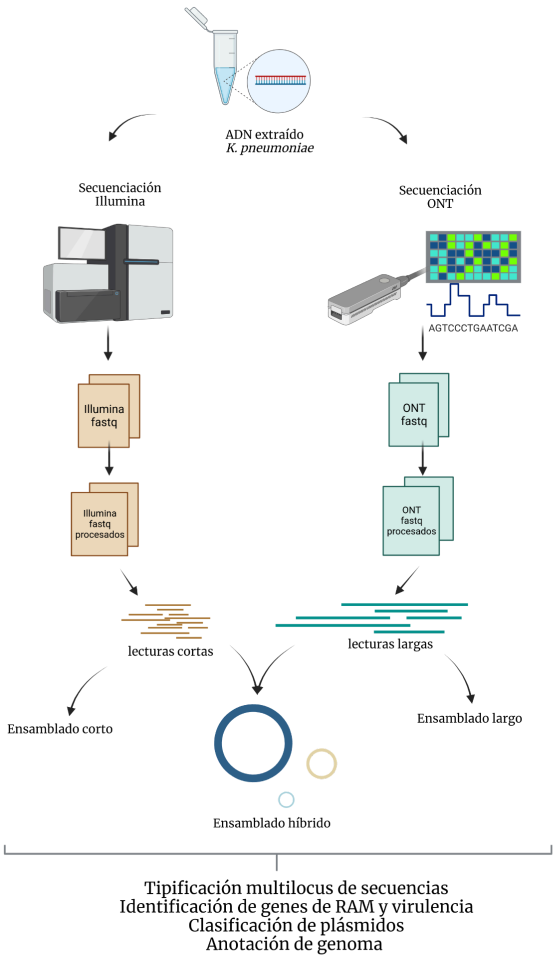

Figura A: Esquema del flujo de trabajo para la obtención de datos de segunda y tercera generación para los aislamientos del brote nosocomial de *K. pneumoniae*. Creado con Biorender.

3.6 Análisis filogenético de las muestras de *K. pneumoniae* ST-11

Se descargaron 290 genomas completos y de buena calidad de *K. pneumoniae* del ST-11 a partir de aislamientos del de la base de datos BV-BRC (https://www.bv-brc.org/). Se utilizó *Parsnp v1.7.4* para el alineamiento del genoma *core*, detección de SNPs (del inglés *single nucleotide polymorphism*) y reconstrucción filogenética (46,47) .

3.7 Visualizaciones

El procesamiento de tablas se realizó con múltiples paquetes de R (48–52) y los gráficos se generaron con el paquete *ggplot2* (53) y *ggtree* (52).

3.8 Disponibilidad de datos y código utilizado

Los datos de secuenciación generados en este estudio fueron depositados en la base de datos de NCBI SRA con el BioProject PRJNA857878. Los comandos para la generación de las figuras se encuentran en https://github.com/Ceci07/klebsiella_genomics

4. Resultados

4.1 Generación de datos de secuenciación ONT e Illumina

Los detalles de los datos de secuenciación obtenidos para las muestras se encuentran en la **Tabla S1**. Para la plataforma ONT se obtuvieron un promedio de 142.440 ± 86.356 lecturas filtradas (calidad > 8 y largo > 200 pb) por muestra con un promedio de calidad 11.1 ± 1.3. Para el caso de la plataforma Illumina, se obtuvieron en promedio 1.660.733,7 ± 93.412,9 de lecturas (2 x 150 pb) (**Figura 1**). Los datos generados se utilizaron para la obtención de ensamblados largos e híbridos de las muestras de *K. pneumoniae*. Se realizó la confirmación inicial de la identidad de las cepas del brote a partir de los resultados de clasificación de las lecturas de secuenciación ONT (**Figura 2**). En promedio, el 89.4% ± 3.3% de las lecturas fueron clasificadas como a *K. pneumoniae*, mientras que 10.6% ± 3.3% de las lecturas fueron clasificadas como otras especies utilizando este método.

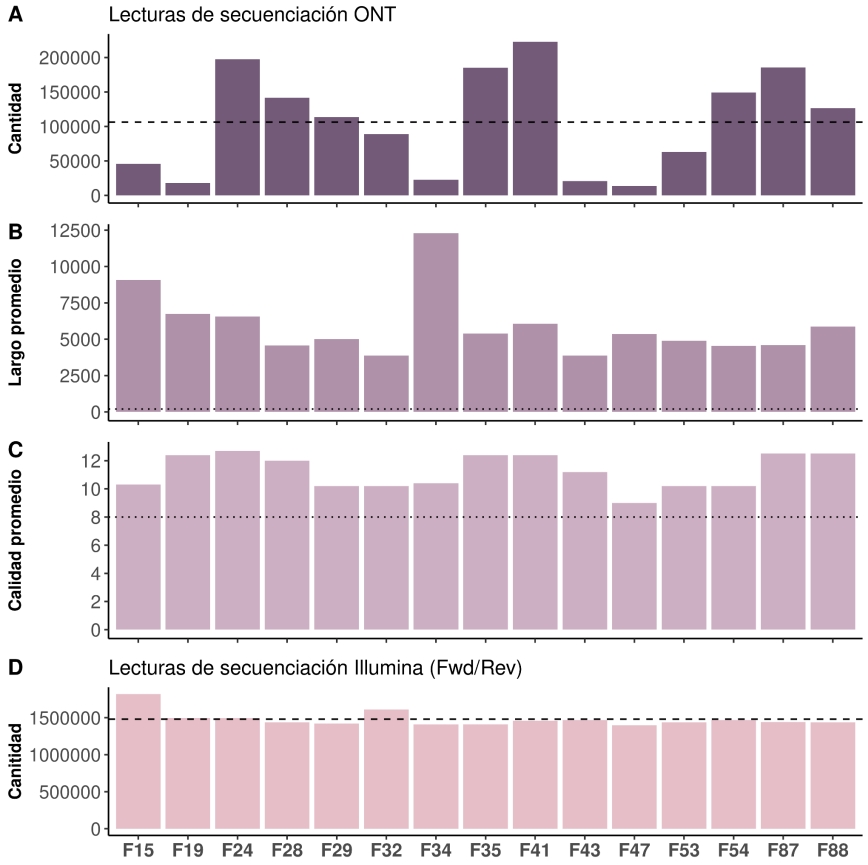

Figura 1: Generación de datos de secuenciación ONT e Illumina. A) Número de lecturas generadas con la plataforma ONT. La línea con punteada indica el promedio de lecturas del total de las muestras (142.440 ± 86.356). B) Largo promedio de las lecturas ONT generadas (5.556,1 ± 2.439,4 pb). La línea punteada indica el límite de largo filtrado (200 pb). C) Calidad promedio de las muestras (11.1 ± 1.3). La línea punteada indica el límite de calidad filtrado (*quality score* ≥ 8). D) Cantidad de lecturas de Illumina (2 x 150 pb). La línea punteada indica el promedio de lecturas del total de las muestras (1.660.733,7 ± 93.412,9) .

4.2 Generación de ensamblados genómicos *de novo* a partir de datos de secuenciación Illumina, ONT y ONT+Illumina

A partir de los ensamblados con lecturas de secuenciación cortas, largas y ambas (ensamblado híbrido) obtenidos de las muestras de *K. pneumoniae* secuenciadas con distintas plataformas, se observan marcadas diferencias en cuanto número de contigs y el tamaño del contig mayor entre los ensamblados con lecturas cortas de *Spades* y a partir de las lecturas largas (*Flye*) y combinadas (*Unicycler*), respectivamente (**Figura S1**). Estas diferencias no son tan marcadas entre los ensamblados obtenidos con *Flye* y *Unicycler* (**Figura 3**). Más detalles de los parámetros medidos de los ensamblados largos e híbridos se encuentran en la **Tabla S2**.

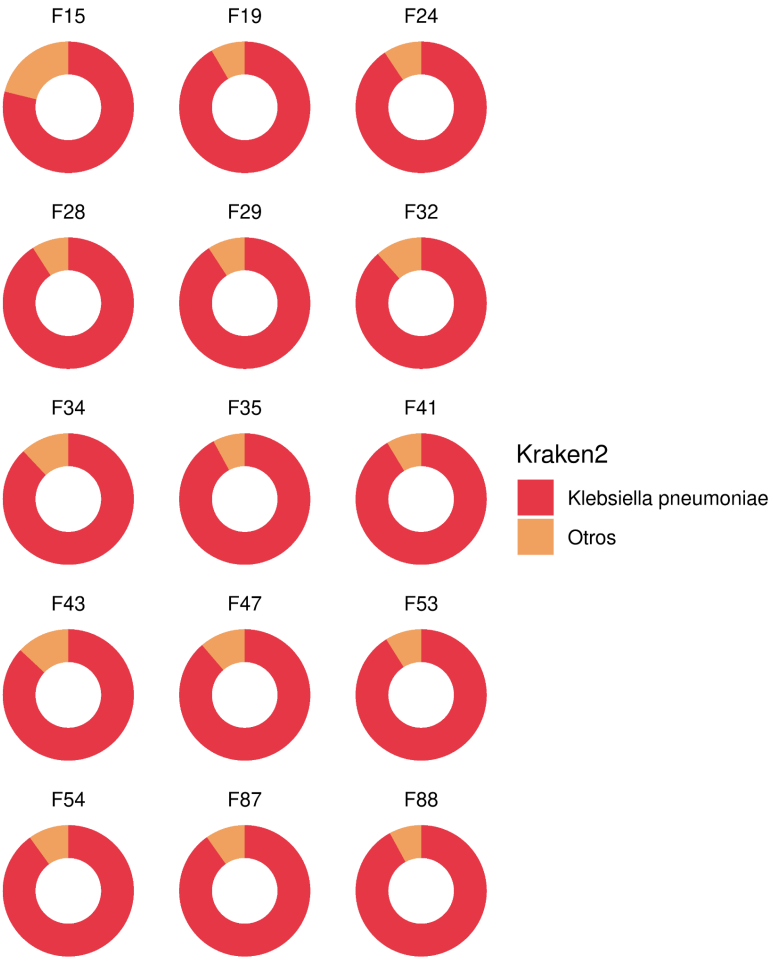

Figura 2: Clasificación de lecturas ONT con *Kraken2* para la confirmación inicial de la identidad de las cepas del brote nosocomial. Casi el 90% de las lecturas generadas con la plataforma ONT fueron clasificadas a nivel de especie como *K. pneumoniae*.

4.3 Tipificación multilocus de secuencias (MLST) basado en los genomas de *K. pneumoniae* ensamblados con el método híbrido

Con la finalidad de determinar la posibilidad de asignar el MLST a partir de los datos ensamblados largos, se utilizó la herramienta MLST y se comparó con el resultado obtenido para el ensamblado híbrido de cada una de las muestras. No se pudo determinar el MLST a partir de los ensamblados largos, sin embargo a partir de los ensamblados cortos e híbridos se pudo determinar que el clon causante del brote nosocomial corresponde al ST-11 de *K. pneumoniae* (**Tabla S2**).

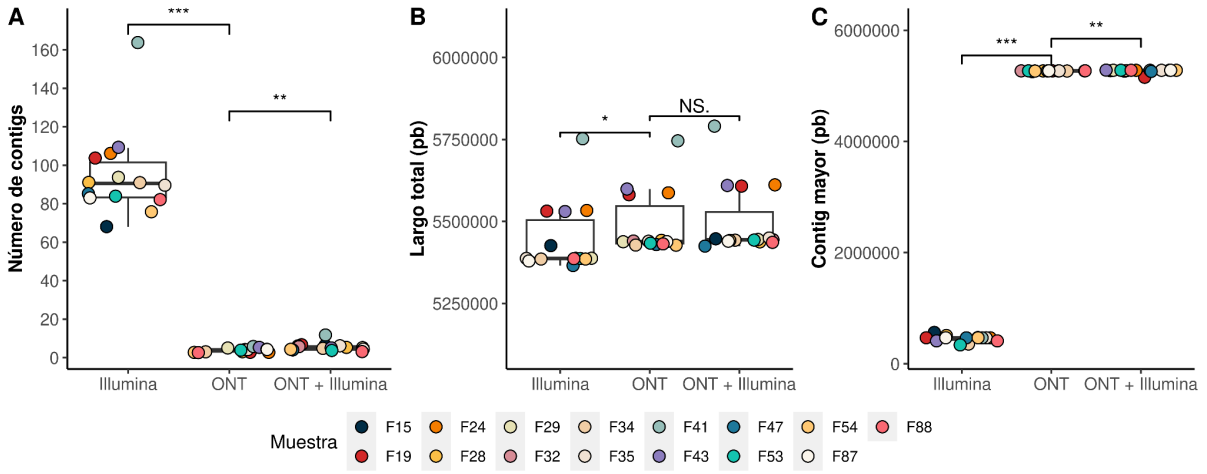

Figura 3: Ensamblados genómicos *de novo* utilizando lecturas de secuenciación Illumina (*Spades*), ONT (*Flye*) y ambas (*Unicycler*). A) Número de contigs generados por cada una de las estrategias de ensamblado. Se obtuvieron un promedio 94.8 ± 22.9 contigs con el ensamblado obtenido con *Spades*, 3.9 ± 0.9 con *Flye* y 5.5 ± 2.1 con *Unicycler*. B) Comparación del tamaño total del ensamblado. Se obtuvo un tamaño total promedio de 5.4 Mb ± 0.11 Mb con datos *Spades*, 5.5 ± 0.10 Mb con *Flye* y 5.5 Mb ± 0.11 Mb con *Unicycler*. C) Comparación del contig de mayor tamaño. Los contigs más largos fueron obtenidos con los ensamblados de *Unicycler* 5.3 Mb ± 0.003 Mb, seguido de 5.3 Mb ± 0.0004 Mb con *Flye* y 0.45 Mb ± 0.006 Mb con *Spades*. La comparación se realizó utilizando el test de Wilcoxon.

4.4 Detección de genes que confieren resistencia a antimicrobianos, contenido plasmídico y factores de virulencia

El perfil de genes de resistencia obtenidos con la herramienta *abricate* con ensamblados largos e híbridos se muestran en la **Figura 4A**. Se observa que en general existe una concordancia entre la anotación obtenida para ambos ensamblados, salvo para la muestra F15, donde no se pudo detectar la presencia del gen *QnrB20* a partir del ensamblado obtenido con *Flye*. Adicionalmente, se detectaron tres genes (*msbA*, *CMY-59* y *bacA*) que no se encuentran presentes en el ensamblado de alta resolución de dicha muestra. Para la muestra F47 se detectó una copia del gen $bla_{SHV\text{-}158}$, mientras que en los ensamblados de alta resolución se detectaron en todos los casos el gen $bla_{SHV\text{-}182}$. A pesar de estas discrepancias, en general se pudo detectar el tipo de genes de resistencia que circularon durante el período del brote intrahospitalario utilizando datos de secuenciación ONT. Los genes adquiridos corresponden a las carbapenemasas bla_{KPC} y $bla_{NDM\text{-}1}$ y $bla_{OXA\text{-}1}$ y la β-lactamasa $bla_{CTX\text{-}M\text{-}15}$. Éstos a su vez fueron detectados utilizando tanto datos de secuenciación ONT, como ONT+Illumina.

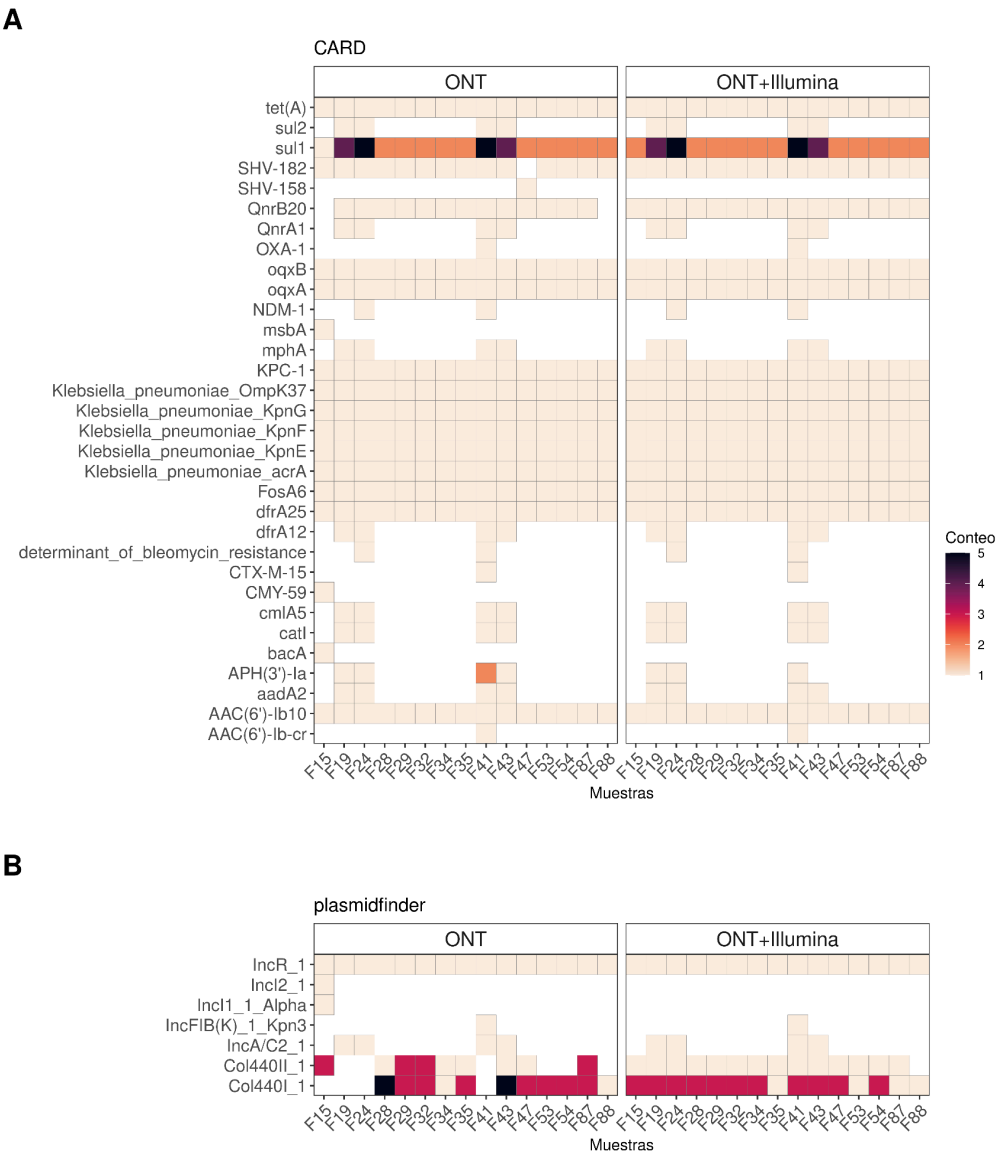

Figura 4: Comparación entre los ensamblados obtenidos con datos de secuenciación de tercera generación (ONT) y datos de tercera y segunda generación (ONT+Illumina) en términos de genes de resistencia detectados y contigs plasmídicos. A) Identificación de los genes RAM utilizando la base de datos CARD. B) Identificación de los contigs plasmídicos en base al criterio de *plasmidfinder*. En ambos casos se utilizó *abricate* para la búsqueda de los genes de RAM y plásmidos. Se muestran solamente los datos con una cobertura mayor al 95% y una identidad mayor al 90% para el caso de los genes RAM y 98% y 90%, respectivamente para el caso de plásmidos.

Respecto a la detección del contenido plasmídico de las cepas de estudio, el plásmido perteneciente al grupo de incompatibilidad IncR_1 fue encontrado en todas las muestras analizadas (**Figura 4B**). También fueron encontrados los plásmidos pertenecientes a los grupos de incompatibilidad IncA/C2_1 (en muestras F19, F24, F41 y F43) y IncFIB(K)_1_Kpn3 (en muestra F41). Estos resultados fueron obtenidos tanto para los ensamblados ONT como para ONT+Illumina. Sin embargo, para la muestra F15, se obtuvieron clasificaciones de plásmidos de distintos contigs que no se corroboran con el

ensamblado de alta resolución de la misma muestra. Adicionalmente, solamente a partir de los datos ONT+Illumina, se obtuvo de manera consistente la clasificación de plásmidos pequeños Col440I_1 y Col440II_1. Los mismos se encuentran presentes en todas las muestras del estudio, a excepción de F88 en el cual Col440II_1 no fue detectado.

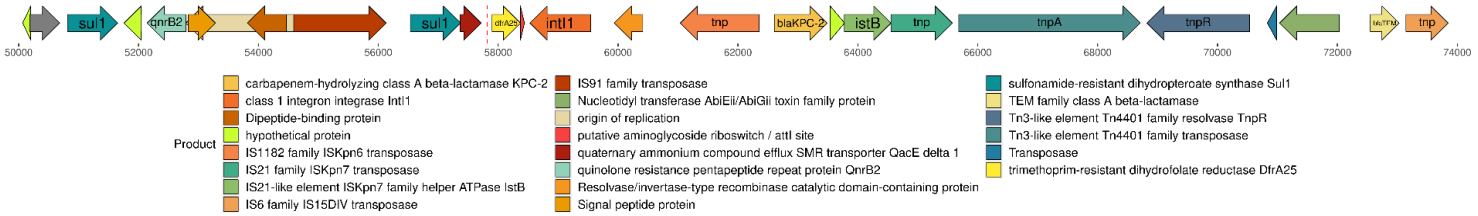

Figura 5: Estructura general del contexto genómico del gen *bla*$_{KPC-2}$ del plásmido IncR en las muestras de *K. pneumoniae* del brote nosocomial. El gen *bla*$_{KPC-2}$ se encuentra adyacente a un integrón clase I o "clínico". Se muestra la anotación obtenida con *Bakta* de la secuencia del plásmido IncR entre las posiciones 50.000 y 75.000 de una muestra representativa (F19). La línea punteada indica el sitio attC detectado por *IntegronFinder* 2.0.

A partir de los genomas de alta resolución, se realizó la búsqueda de integrones y la anotación de los genes del plásmido IncR a modo de visualizar el contexto genómico de la carbapenemasa. La anotación del plásmido con *Bakta* y la búsqueda de determinantes de RAM con *Kleborate* determinó que la carbapenemasa fuera clasificada como *bla*$_{KPC-2}$ y no como *bla*$_{KPC-1}$ obtenido con la herramienta de clasificación *abricate*. En la **Figura 5** se muestra el contexto genómico general observado en el segmento que incluye el gen *bla*$_{KPC-2}$ del plásmido IncR de la mayor parte de las muestras. En el mismo se observa asociado a distintos elementos de la familia de transposas Tn4401, así como las transposas ISKpn6. También se detectó el gen *bla*$_{TEM}$ asociado a la resistencia a antibióticos β-lactámicos. El detalle del mismo segmento para los plásmidos IncR de los demás genomas se encuentra en la **Figura S2**. En general los genes RAM detectados en el plásmido IncR confieren resistencia a antibióticos β-lactámicos y carbapenémicos (carbapenem, cefalosporinas, monobactam y penam). La predicción obtenida con *integron-finder* para todas los plásmidos IncR de los aislamientos clínicos confirmó la presencia de la integrasa *intI1* y del sitio attC, clasificando las secuencias encontradas como integrones de clase I (**Tabla S3**).

Los genes de resistencia asociados a IncA/C2_1 corresponden a *sul1*, *sul2*, *catI*, *dfrA12*, *aadA2*, *QnrA1*, *cmlA5*, *mphA*, *APH(3')-Ia, determinant_of_bleomycin_resistance* y *bla*$_{NDM-1}$. Estos genes están asociados a la resistencia a sulfonamidas, fenicol,

diaminopirimidinas, aminoglucósidos, fluoroquinolonas, macrólidos, glicopéptidos y carbapenem. Las muestras F24 y F41 portan el gen de la carbapenemasa *bla*$_{NDM-1}$, asociados éstos a resistencia a carbapenem, cefalosporinas, cefamicinas y penam, respectivamente (**Figura 6**). Por su parte el plásmido IncFIB(K)_1_Kpn3 contiene los genes de resistencia *APH(3')-Ia*, *bla*$_{OXA-1}$, *AAC(6')-Ib-cr* y *bla*$_{CTX-M-15}$ asociados con la resistencia a aminoglucósidos, cefalosporinas, penan y fluoroquinolonas.

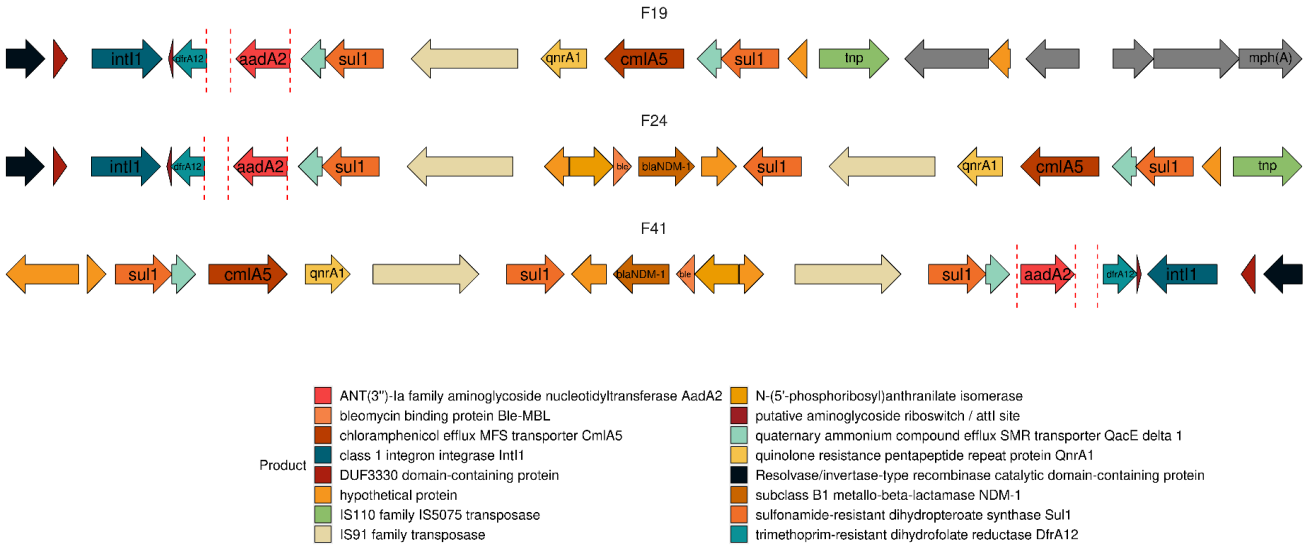

Figura 6: Estructura general del contexto genómico del gen *bla*$_{NDM-1}$ del plásmido IncA/C2 en las muestras de *K. pneumoniae* del brote nosocomial. La muestra F19 presenta el plásmido IncA/C2, pero no presenta el gen *bla*$_{NDM-1}$. El gen *bla*$_{NDM-1}$ en las muestras positivas se encuentra asociado a un transposón de la familia IS91. La anotación fue obtenida con *Bakta* y las líneas punteadas indican los sitios *attC* detectados por *IntegronFinder* v2.0.

El análisis del perfil de resistencia indica que las muestras de aislamientos de *K. pneumoniae* nosocomial obtenidas entre marzo y noviembre del 2017 presentan multirresistencia a diferentes familias de antimicrobianos, confirmándose lo observado por otras metodologías de relevamiento clínico realizadas por la institución hospitalaria (no se muestra). Se determinó la presencia del gen *bla*$_{KPC-2}$ en un plásmido clasificado como IncR_1 adyacente a un integrón de clase I en todas las muestras y adicionalmente, el gen NDM-1 en un plásmido clasificado como IncA/C2_1 en dos muestras y *bla*$_{OXA-1}$ en un plásmido clasificado como IncFIB(K)_1_Kpn3. La muestra F41 presenta simultáneamente éstos tres genes de relevancia clínica. De acuerdo a la tipificación de secuencia multilocus (MLST) a partir de los genomas de alta resolución todas las muestras corresponden al ST11 del CG258 portando el gen de virulencia *yersiniabactin*, el tipo capsular KL64 y el locus O correspondiente a O2v1 (**Tabla S4**). Adicionalmente, se detectó una mutación en el gen *mgrB*

el cual determina un fenotipo resistente a la colistina y mutaciones en los genes *gryA* y *parC* que confieren resistencia a las fluoroquinolonas.

Por otra parte, en la **Figura** 7 se muestra el análisis filogenético basado en SNPs del genoma *core* obtenido a partir de los genomas completos y de alta calidad de los aislamientos de *K. pneumoniae* ST-11. En la misma se muestra que las secuencias de Uruguay forman un cluster homogéneo, presentando similaridad con secuencias asiáticas y europeas (**Figura S3**).

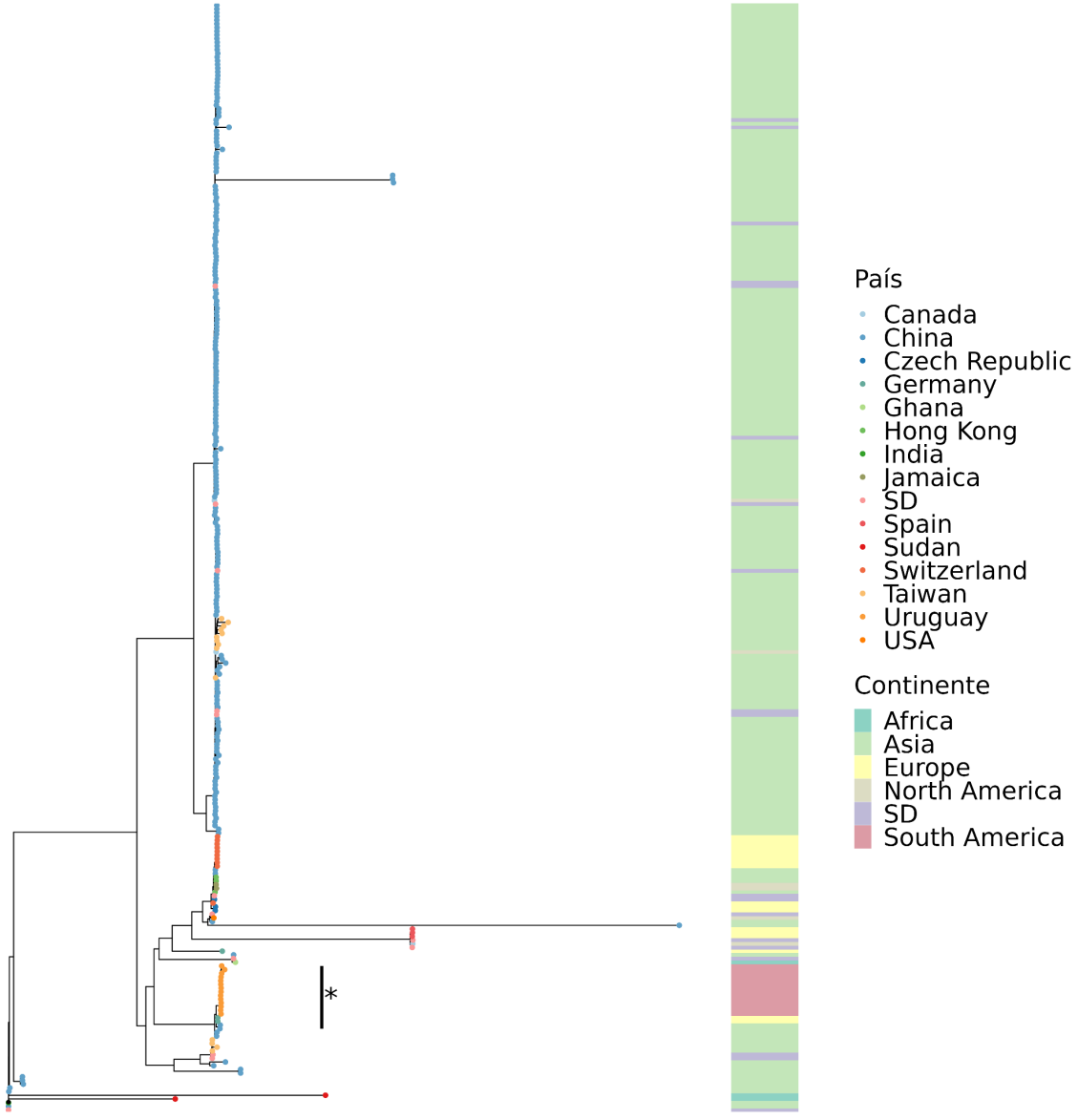

Figura 7: Análisis filogenético de SNPs del genoma *core* de los genomas de *K. pneumoniae* ST-11 de Uruguay analizados y los obtenidos de la base de datos. El asterisco indica el nodo que contiene las secuencias generadas en este estudio.

5. Discusión

El aumento en la incidencia de RAM es una temática prioritaria a ser abordada, particularmente en centros hospitalarios y más aún en las unidades de cuidados intensivos. Rutinariamente se realiza la caracterización de la resistencia a antimicrobianos a partir del aislamiento de microorganismos y pruebas fenotípicas como las basadas en la concentración mínima inhibitoria (MIC, por *minimum inhibitory concentration*). También se han implementado otros métodos basados en el genotipado o métodos moleculares que eliminan la necesidad de el cultivo extenso de los microorganismos. Estos incluyen la PCR, los microarrays y LAMP (54). Una de las herramientas con potencial aplicación para la caracterización de los determinantes de RAM de forma rápida y de alta resolución es la secuenciación masiva. Sin embargo, deben superarse las limitaciones para su implementación en el ámbito clínico, tales como la reducción en los tiempos requeridos y costos, estandarización de los procedimientos, generación de herramientas de análisis de datos de uso universal y mejoras en las bases de datos disponibles. El uso de la genómica en la práctica clínica ha aumentado en la última década, particularmente para la tipificación de patógenos e investigación de brotes. Recientemente se reportó un flujo de trabajo certificado con la norma ISO para la detección y validación de los resultados obtenidos de RAM utilizando datos de secuenciación masiva (55).

En el presente estudio se realizó la caracterización genómica retrospectiva de un brote clonal de CR-*Kp* en una institución pública de la ciudad de Montevideo durante el año 2017 utilizando tecnología de secuenciación portable ONT. Previamente, ya habían ocurrido brotes intrahospitalarios de CR-*Kp* en Uruguay. Específicamente, en el año 2011 se identificó un brote de *K. pneumoniae* en el Hospital Central de las Fuerzas Armadas de la ciudad de Montevideo con una duración de 30 meses. Mediante metodologías de PCR, se determinó que el brote fue causado por una CR-*Kp* portando el gen bla_{KPC-2} (56). En el año 2014, se describen dos casos clínicos de CR-*Kp* bla_{KPC-2} ocurridos en el año 2011 en un hospital privado de la ciudad de Maldonado portando un fenotipo multirresistente que incluye resistencia a colistina. Mediante técnicas de electroforesis en gel de campo pulsado (PFGE) se determinó que se trataba de un único clon perteneciente al ST258 (57). Si bien los primeros casos de CR-Kp bla_{KPC-2} fueron detectados en el año 2011, un estudio llevado a cabo con aislamientos del período de julio del 2010 y agosto del 2011 se detectaron pacientes colonizados con *K. pneumoniae* no productora de carbapenemasas tipificados con el ST258. En otro estudio se relevaron 8364 enterobacterias aisladas entre el año 2012 y 2016 en el

Hospital de Clínicas "Dr Manuel Quintela" donde el 30% correspondían a *K. pneumoniae*. El análisis de tipificación multilocus de estas secuencias determinó la presencia del ST11, ST14 y ST661 (58). En el año 2018 se reporta el genoma del primer aislamiento de CR-Kp bla_{KPC-2} en Uruguay y un análisis filogenético determinó que se trataba del ST258. Se identificaron además los plásmidos del grupo de incompatibilidad IncFIB(K) y IncFII(K) y una inserción en el gen cromosómico *mgrB*, involucrado en la generación de resistencia a colistina (59). En el mismo año de recolección de las muestras de este estudio, se registraron un total de 10 brotes de microorganismos productores de carbapenemasas (MPC) en Uruguay (60).

A partir de los datos de secuenciación ONT obtenidos para los aislamientos de *K. pneumoniae*, se pudo confirmar la identidad del patógeno utilizando metodologías rápidas, obteniéndose casi 90% de clasificación a nivel de especie utilizando las lecturas de secuenciación. También se pudieron caracterizar los determinantes de resistencia y virulencia y los grupos de incompatibilidad de los plásmidos asociados al clon. Entre ellos los genes que confieren resistencia a antibióticos carbapenémicos como el gen bla_{KPC-2} y el gen bla_{NDM-1}. Debido a que se requiere coincidencias perfectas en los siete locus para determinación del MLST, no se obtuvo este dato utilizando los genomas ensamblados únicamente con las lecturas ONT, ya que un error común asociado a estos ensamblados son las inserciones y deleciones (indels) que acortan artificialmente las proteínas al introducir codones *stop* prematuros o cambios en el marco de lectura (61). Recientemente, se publicó un estudio comparativo para la obtención del MLST, tipificación del locus K/O y determinación de genes de RAM y virulencia con datos ONT y ONT+Illumina. Utilizando un modelo reciente para la asignación de bases de muy alta precisión de ONT (Guppy v4.0.14, r9.4.1_450bps_sup) y generando ensamblados con o sin un paso adicional de pulido (Medaka) un total de 87.3% de las muestras fueron correctamente asignadas al MLST (62). Asimismo, los avances recientes en la química de secuenciación de ONT (R10.4.1) así como en los modelos de asignación de bases harían posible próximamente la determinación del MLST solamente con datos ONT. Para el caso de las muestras del brote nosocomial la tipificación de secuencia multilocus corresponde al ST11 del CG258 obtenido a partir de los ensamblados de alta resolución (ONT+Illumina). Asimismo, presentan el locus *ybt* 9 en un elemento integrativo conjugativo ICEKp3. *ybt* codifica la biosíntesis del sideróforo *yersiniabactin* y su receptor, considerados factores de virulencia claves ya que proporcionan un mecanismo de captura de hierro del hospedero, incrementando su habilidad de supervivencia y replicación (63). Tanto el lipopolisacárido (o antígeno O) como el polisacárido capsular (antígeno K) actúan como factores de virulencia y antígenos de

superficie de *K. pneumoniae*. Los aislamientos del brote nosocomial corresponden a KL64, las *K. pneumoniae* ST-11-KL64 son particularmente comunes en Asia y Sudamérica (5,64). El locus O corresponde a O2v1. Recientemente se reportó el reemplazo gradual del subclon OL101:KL47 del ST-11 de CP-*Kp* por el subclon O2v1:KL64 en China. El mismo contiene una mutación puntual en el gen *recC*, la cual promueve significativamente eventos de recombinación y adicionalmente se observan mayor cantidad de elementos genéticos móviles (65). La sustitución en el gen *recC* reportada en ese estudio (His935Arg) también fue detectada en las muestras del brote nosocomial del presente estudio. Adicionalmente, se detectó una mutación en el gen cromosómico *mgrB,* el cual determina un fenotipo resistente a la colistina y mutaciones en los genes cromosómicos *gryA* y *parC*. Estos genes se encuentran dentro de las regiones determinantes de resistencia a quinolonas (QRDR). Las mutaciones en el gen *gryA* se encuentran principalmente en bacterias Gram negativas y se halla vinculada a cambios estructurales en la ADN girasa y/o topoisomerasas reduciendo la afinidad de las mismas a las fluoroquinolonas, impidiendo la replicación y transcripción del microorganismo. Asimismo, se detectó en todas las muestras del estudio el gen plasmídico *qnrB2* que también confiere resistencia a fluoroquinolonas (66).

La multirresistencia de las muestras recolectadas se atribuyó a distintas familias de antimicrobianos incluyendo β-lactámicos y carbapenémicos. Los genes de las carbapenemasas bla_{KPC-2} y bla_{NDM-1} se encuentran en plásmidos de los grupos de incompatibilidad IncR y IncA/C2, respectivamente. El gen bla_{KPC-2} se encuentra asociado al transposón Tn4401. El mismo está compuesto por los genes transposasa (tnpA) y resolvasa (tnpR) y dos secuencias de inserción ISKpn6 y ISKpn7 además del gen bla_{KPC-2} (67). El mismo se encontró adyacente a un segmento conteniendo el gen *IntI* y un sitio *attC* además de los genes *dfrA25* (asociado a la resistencia al trimetoprim), *qacEdelta1* (asociado a la resistencia al amonio cuaternario) y *sul1* (asociado a la resistencia a la sulfonamida), característicos de la estructura de un integrón clase I. Los integrones son elementos transferidos horizontalmente y juegan un rol importante en la diseminación de genes de RAM. Se cree que han sido de los primeros elementos de resistoma ambiental a moverse al ecosistema humano y se han convertido en uno de los vectores de resistencia antimicrobiana más relevante del punto de vista clínico (68). Por su parte, el gen bla_{NDM-1} fue detectado en dos de las tres muestras donde el plásmido IncA/C2 fue detectado. El mecanismo de inserción a la secuencia plasmídica parece estar asociado a un transposón de la familia IS91. Una de las características llamativas es la presencia de dos regiones idénticas de esta transposasa flanqueando la región que contiene el gen $bla_{NDM-1.}$ Estos elementos carecen de

los repetidos invertidos convencionales en sus terminaciones y pueden movilizar secuencias de ADN en un proceso llamado replicación del círculo rodante (69). También se ha reportado este elemento de transposición asociado a la presencia de múltiples copias en tándem del gen bla_{NDM-1} (70).

En resumen, en el presente estudio se determinó que el clon causante del brote nosocomial corresponde al ST-11 de *K. pneumoniae*, portando multirresistencia a antibióticos β-lactámicos y carbapenémicos. El ST-11, junto con el ST-258 forman parte del CG258 de *K. pneumoniae*. El ST-258 ha contribuido significativamente a la dispersión de resistencia a antibióticos carbapenémicos en Estados Unidos y Europa. Sin embargo, el ST-11, predominante en China y con distribución global, corresponde a una variante de un solo locus (*tonB*) del ST258, lo que indica una estrecha relación entre ellos (64). Dada la detección de la circulación de un clon de CR-*Kp* ST-11 de distribución mundial y el aumento en la prevalencia de CR-*Kp* del fenotipo híbrido de *K. pneumoniae* clásica multirresistente e hipervirulenta, es pertinente la implementación de sistemas de vigilancia genómica descentralizados que apunten a la caracterización temprana de estos patógenos tanto en el contexto clínico como en el ambiente. Debido a los indicios de que el *K. pneumonie* ST-11 presenta una mayor supervivencia en superficies abióticas dentro del contexto hospitalario, este clon podría diseminarse al ambiente de forma más exitosa y contribuir a la dispersión de la RAM al ambiente (71). Posteriormente, el clon causante del brote nosocomial de CR-*Kp* ST11 fue detectado en aguas residuales del ambiente urbano de la ciudad de Montevideo, poniendo en evidencia el rol del microbioma intestinal y de los sistemas de saneamiento en la dispersión de patógenos así como de la necesidad de implementación de sistemas de tratamiento de las aguas residuales hospitalarias (72). En este sentido, la utilización de tecnologías de secuenciación de tercera generación portátil podrían constituir en una estrategia costo-efectiva para el monitoreo de la presencia de genes de RAM y factores de virulencia en el ambiente. Adicionalmente, las constantes mejoras en cuanto a la química de secuenciación y las herramientas de análisis permitirá en un futuro llevar a cabo la anotación de genomas y análisis filogenéticos prescindiendo de datos de secuenciación de segunda generación.

6. Material suplementario

Tabla S1: Tabla de datos de secuenciación filtrados obtenidos de las plataformas ONT y Illumina para las muestras de ADN de *K. pneumoniae* obtenidas del brote nosocomial.

	ONT				Illumina (2 X 150 pb)
	Número de lecturas	**Promedio de largo de lecturas**	**Promedio de calidad de lecturas**	**N50 de lecturas**	**Número de lecturas (Fwd/Rev)**
F15	45.733	9.073,8	10.3	14.298	1.821.192
F19	178.560	6.751,0	12.4	7.486	1,497,448
F24	197.526	6.570,5	12.7	7.508	1,494,135
F28	141.525	4.562,3	12.0	5.198	1,437,097
F29	113.586	5.010,6	10.2	6.015	1,422,106
F32	88.927	3.878,1	10.2	4.511	1.612.242
F34	22.785	12.298,1	10.4	15.320	1.409.196
F35	185.235	5.394,5	12.4	6.470	1,411,007
F41	222.921	6.065,6	12.4	7.365	1,458,385
F43	208.720	3.885,5	11.2	4.440	1,469,955
F47	13.708	5.355,3	9.0	7.099	1,399,320
F53	63.071	4.896,3	10.2	5.877	1,436,953
F54	149.083	4.544,3	10.2	5.641	1,470,372
F87	185.327	4.587,7	12.5	6.458	1,442,849
F88	126.724	5.875,3	12.5	7.441	1,437,070
Promedio	**142.440**	**5.556,1**	**11,1**	**7.388,5**	**1.660.733,7**
Desvío Estándar	**86.356**	**2.439,4**	**1,3**	**3.212,0**	**93.412,9**

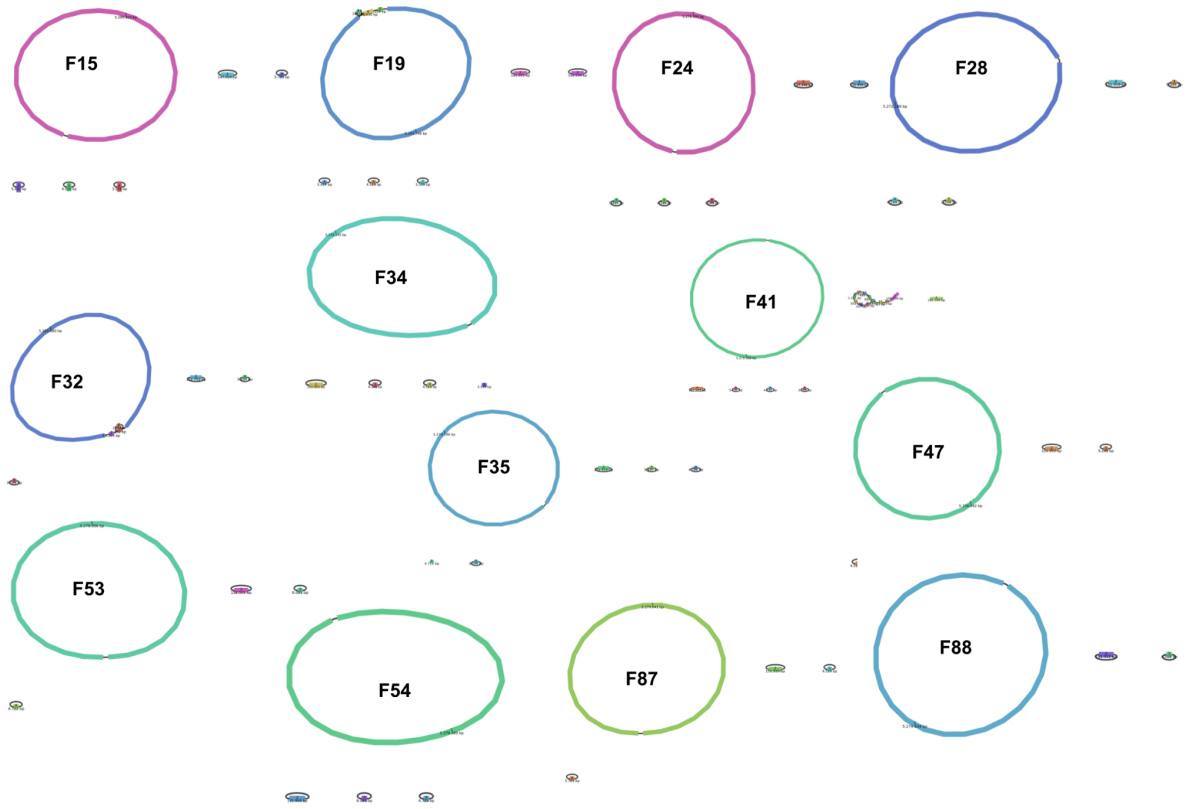

Figura S1: Visualización de grafos de los genomas completos o casi completos de alta resolución generados con datos de secuenciación de segunda y tercera generación utilizando *Unicycler*.

Tabla S2: Resultados de *Quast* a partir de los ensamblados obtenidos con datos ONT y datos ONT+Illumina de las muestras del brote *K. pneumoniae* nosocomial

Muestra	Largo total (pb)	Contigs	Contig más largo (pb)	ST	Plataforma
F15	6.041.036	21	5.267.610	SD	ONT
F15	5.384.293	68	563.398	11	Illumina
F15	5.447.074	6	5.280.823	11	ONT + Illumina
F19	5.581.492	3	5.268.543	SD	ONT
F19	5.522.525	104	467.234	11	Illumina
F19	5.607.794	7	5.153.748	11	ONT + Illumina
F24	5.587.311	3	5.268.256	SD	ONT
F24	5.523.873	106	466.934	11	Illumina
F24	5.611.603	6	5.278.990	11	ONT + Illumina

F28	5.442.388	3	5.268.278	SD	ONT
F28	5.379.552	91	507.743	11	Illumina
F28	5.437.714	5	5.272.289	11	ONT + Illumina
F29	5.438.101	5	5.266.683	SD	ONT
F29	5.380.514	94	470.985	11	Illumina
F29	5.445.043	5	5.279.618	11	ONT + Illumina
F32	5.440.252	4	5.268.921	SD	ONT
F32	5.444.342	6	5.263.880	11	ONT + Illumina
F34	5.427.989	3	5.266.581	SD	ONT
F34	5.379.046	91	354.511	11	Illumina
F34	5.442.900	5	5.274.145	11	ONT + Illumina
F35	5.439.556	4	5.269.067	SD	ONT
F35	5.379.474	90	467.153	11	Illumina
F35	5.449.543	6	5.279.196	11	ONT + Illumina
F41	5.745.832	6	5.268.051	SD	ONT
F41	5.680.869	164	466.646	11	Illumina
F41	5.790.557	12	5.279.088	11	ONT + Illumina
F43	5.598.963	5	5.268.896	SD	ONT
F43	5.609.704	5	5.282.442	11	ONT + Illumina
F47	5.429.538	4	5.254.936	SD	ONT
F47	5.680.869	164	466.646	11	Illumina
F47	5.424.753	4	5.259.042	11	ONT + Illumina
F53	5.433.664	4	5.269.725	SD	ONT
F53	5.378.887	84	338.928	11	Illumina
F53	5.443.230	4	5.279.956	11	ONT + Illumina
F54	5.427.449	3	5.265.845	SD	ONT
F54	5.379.376	76	467234	11	Illumina

F54	5.442.117	4	5.279.389	11	ONT + Illumina
F87	5.438.415	4	5.269.710	SD	ONT
F87	5.372.170	83	466.934	11	Illumina
F87	5.439.792	4	5.279.641	11	ONT + Illumina
F88	5.431.613	3	5.270.272	SD	ONT
F88	5.376.559	82	411.276	11	Illumina
F88	5.436.256	3	5.279.834	11	ONT + Illumina

SD: sin determinar

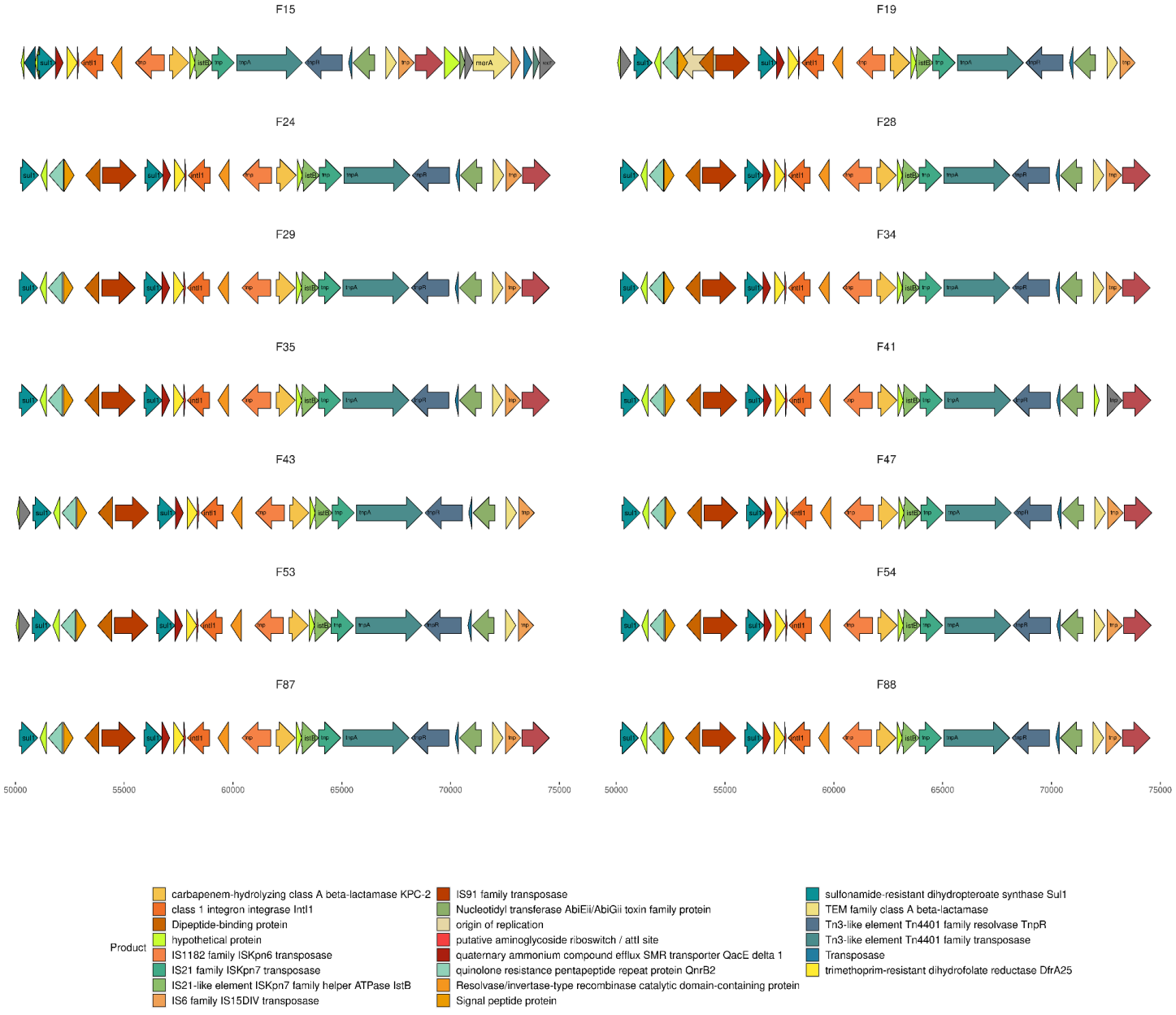

Figura S2: Contexto genómico de las muestras de *K. pneumoniae* ST-11 del brote nosocomial. No se muestra la muestra F32 en la figura.

Tabla S3: Detección de integrones con *IntegronFinder 2.0* completos en las muestras CR-*Kp* del brote intrahospitalario.

	Contig	Tamaño (pb)	Sitios attC	Plásmido
F15	2	147.024	1	IncR_1
F19	2	161.001	3	IncA/C2_1
F19	3	152.544	1	IncR_1
F24	2	167.148	3	IncA/C2_1
F24	3	151.959	1	IncR_1
F29	2	151.919	1	IncR_1
F32	2	151.919	1	IncR_1
F34	2	151.921	1	IncR_1
F35	2	151.921	1	IncR_1
F41	3	159.506	3	IncA/C2_1
F41	4	151.949	1	IncR_1
F43	2	161.001	3	IncA/C2_1
F43	3	152.543	1	IncR_1
F47	2	151.993	1	IncR_1
F53	2	152.501	1	IncR_1
F54	2	151.955	1	IncR_1
F87	2	151.920	1	IncR_1
F88	2	151.912	1	IncR_1

Tabla S4: Resultados parciales obtenidos con *Kleborate*

strain	species	ST	Yersiniabactin	YbST	wzi	K_locus	O_locus	Sul_acquired	Bla_Carb_acquired	Bla_chr	Col_mutations
F15	Klebsiella pneumoniae	ST11	ybt 9; ICEKp3	183	wzi64	KL64	O2v1	sul1;sul1	KPC-2	SHV-11.v1	MgrB-62%
F19	Klebsiella pneumoniae	ST11	ybt 9; ICEKp3	183	wzi64	KL64	O2v1	sul1;sul1;sul1;sul1;sul2	KPC-2	SHV-11.v1	MgrB-62%
F24	Klebsiella pneumoniae	ST11	ybt 9; ICEKp3	183	wzi64	KL64	O2v1	sul1;sul1;sul1;sul1;sul1;sul2	KPC-2;NDM-1	SHV-11.v1	MgrB-62%
f28	Klebsiella pneumoniae	ST11	ybt 9; ICEKp3	183	wzi64	KL64	O2v1	sul1;sul1	KPC-2	SHV-11.v1	MgrB-62%
F29	Klebsiella pneumoniae	ST11	ybt 9; ICEKp3	183	wzi64	KL64	O2v1	sul1;sul1	KPC-2	SHV-11.v1	MgrB-62%
F32	Klebsiella pneumoniae	ST11	ybt 9; ICEKp3	183	wzi64	KL64	O2v1	sul1;sul1	KPC-2	SHV-11.v1	MgrB-62%
F34	Klebsiella pneumoniae	ST11	ybt 9; ICEKp3	183-1LV	wzi64	KL64	O2v1	sul1;sul1	KPC-2	SHV-11.v1	MgrB-62%
F35	Klebsiella pneumoniae	ST11	ybt 9; ICEKp3	183	wzi64	KL64	O2v1	sul1;sul1	KPC-2	SHV-11.v1	MgrB-62%
F41	Klebsiella pneumoniae	ST11	ybt 9; ICEKp3	183	wzi64	KL64	O2v1	sul1;sul1;sul1;sul1;sul1*;sul2	KPC-2;NDM-1	SHV-11.v1	MgrB-62%
F43	Klebsiella pneumoniae	ST11	ybt 9; ICEKp3	183	wzi64	KL64	O2v1	sul1;sul1;sul1;sul1;sul2	KPC-2	SHV-11.v1	MgrB-62%

F47	Klebsiella pneumoniae	ST11	ybt 9; ICEKp3	183	wzi64	KL64	O2v1	sul1;sul1	KPC-2	SHV-11.v1	MgrB-62%
F53	Klebsiella pneumoniae	ST11	ybt 9; ICEKp3	183	wzi64	KL64	O2v1	sul1;sul1	KPC-2	SHV-11.v1	MgrB-62%
F54	Klebsiella pneumoniae	ST11	ybt 9; ICEKp3	183	wzi64	KL64	O2v1	sul1;sul1	KPC-2	SHV-11.v1	MgrB-62%
F87	Klebsiella pneumoniae	ST11	ybt 9; ICEKp3	183	wzi64	KL64	O2v1	sul1;sul1	KPC-2	SHV-11.v1	MgrB-62%
F88	Klebsiella pneumoniae	ST11	ybt 9; ICEKp3	183	wzi64	KL64	O2v1	sul1;sul1	KPC-2	SHV-11.v1	MgrB-62%

Figura S3: Árbol filogenético basado en SNP del genoma *core* de *K. pneumoniae*. Las secuencias analizadas de Uruguay son similares entre sí y se encuentran formando un cluster con secuencias obtenidas de Europa y Asia.

7. Referencias

1. De Oliveira DMP, Forde BM, Kidd TJ, Harris PNA, Schembri MA, Beatson SA, et al. Antimicrobial Resistance in ESKAPE Pathogens. Clin Microbiol Rev. 17 de junio de 2020;33(3):e00181-19.
2. Wyres KL, Lam MMC, Holt KE. Population genomics of Klebsiella pneumoniae. Nat Rev Microbiol. junio de 2020;18(6):344-59.
3. Selden R. Nosocomial Klebsiella Infections: Intestinal Colonization as a Reservoir. Ann Intern Med. 1 de mayo de 1971;74(5):657.
4. Jarvis WR, Munn VP, Highsmith AK, Culver DH, Hughes JM. The Epidemiology of Nosocomial Infections Caused by *Klebsiella pneumoniae*. Infect Control. febrero de 1985;6(2):68-74.
5. Martin RM, Bachman MA. Colonization, Infection, and the Accessory Genome of *Klebsiella pneumoniae*. Front Cell Infect Microbiol. 22 de enero de 2018;8:4.
6. Davison HC, Woolhouse MEJ, Low JC. What is antibiotic resistance and how can we measure it? Trends Microbiol. diciembre de 2000;8(12):554-9.
7. Babini GS, Livermore DM. Are SHV β-Lactamases Universal in *Klebsiella pneumoniae*? Antimicrob Agents Chemother. agosto de 2000;44(8):2230-2230.
8. Bourouis A, Ben Moussa M, Belhadj O. Multidrug-resistant phenotype and isolation of a Novel SHV- beta-Lactamase variant in a clinical isolate of *Enterobacter cloacae*. J Biomed Sci. diciembre de 2015;22(1):27.
9. Suay-García, Pérez-Gracia. Present and Future of Carbapenem-resistant Enterobacteriaceae (CRE) Infections. Antibiotics. 19 de agosto de 2019;8(3):122.
10. Chen L, Mathema B, Chavda KD, DeLeo FR, Bonomo RA, Kreiswirth BN. Carbapenemase-producing *Klebsiella pneumoniae*: molecular and genetic decoding. Trends Microbiol. diciembre de 2014;22(12):686-96.
11. Lauretti L, Riccio ML, Mazzariol A, Cornaglia G, Amicosante G, Fontana R, et al. Cloning and Characterization of *bla* $_{VIM}$, a New Integron-Borne Metallo-β-Lactamase Gene from a *Pseudomonas aeruginosa* Clinical Isolate. Antimicrob Agents Chemother. julio de 1999;43(7):1584-90.
12. Yong D, Toleman MA, Giske CG, Cho HS, Sundman K, Lee K, et al. Characterization of a New Metallo-β-Lactamase Gene, *bla* $_{NDM-1}$, and a Novel Erythromycin Esterase Gene Carried on a Unique Genetic Structure in *Klebsiella pneumoniae* Sequence Type 14 from India. Antimicrob Agents Chemother. diciembre de 2009;53(12):5046-54.
13. Watanabe M, Iyobe S, Inoue M, Mitsuhashi S. Transferable imipenem resistance in *Pseudomonas aeruginosa*. Antimicrob Agents Chemother. enero de 1991;35(1):147-51.
14. Poirel L, Héritier C, Tolün V, Nordmann P. Emergence of Oxacillinase-Mediated Resistance to Imipenem in *Klebsiella pneumoniae*. Antimicrob Agents Chemother. enero de 2004;48(1):15-22.

15. WHO. WHO publishes list of bacteria for which new antibiotics are urgently needed [Internet]. 2017. Disponible en: https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed
16. Reyes JA, Melano R, Cárdenas PA, Trueba G. Mobile genetic elements associated with carbapenemase genes in South American Enterobacterales. Braz J Infect Dis. mayo de 2020;24(3):231-8.
17. Gu D, Dong N, Zheng Z, Lin D, Huang M, Wang L, et al. A fatal outbreak of ST11 carbapenem-resistant hypervirulent *Klebsiella pneumoniae* in a Chinese hospital: a molecular epidemiological study. Lancet Infect Dis. enero de 2018;18(1):37-46.
18. Ernst CM, Braxton JR, Rodriguez-Osorio CA, Zagieboylo AP, Li L, Pironti A, et al. Adaptive evolution of virulence and persistence in carbapenem-resistant *Klebsiella pneumoniae*. Nat Med. 1 de mayo de 2020;26(5):705-11.
19. Ruppé E, Olearo F, Pires D, Baud D, Renzi G, Cherkaoui A, et al. Clonal or not clonal? Investigating hospital outbreaks of KPC-producing Klebsiella pneumoniae with whole-genome sequencing. Clin Microbiol Infect. julio de 2017;23(7):470-5.
20. Simon Andrews. FastQC [Internet]. Disponible en: https://github.com/s-andrews/FastQC
21. Trimmomatic [Internet]. Disponible en: https://github.com/usadellab/Trimmomatic
22. Ryan Wick. porechop [Internet]. Disponible en: https://github.com/rrwick/Porechop
23. Wouter De Coster. Nanofilt [Internet]. Disponible en: https://github.com/wdecoster/nanofilt
24. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. Phillippy AM, editor. PLOS Comput Biol. 8 de junio de 2017;13(6):e1005595.
25. Ryan Wick. Unicycler [Internet]. Disponible en: https://github.com/rrwick/Unicycler
26. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 15 de abril de 2013;29(8):1072-5.
27. Center for Algorithmic Biotechnology. Quast [Internet]. Disponible en: https://github.com/ablab/quast
28. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of *de novo* genome assemblies. Bioinformatics. 15 de octubre de 2015;31(20):3350-2.
29. Bandage [Internet]. Disponible en: https://github.com/rrwick/Bandage
30. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. mayo de 2012;19(5):455-77.
31. Center for Algorithmic Biotechnology. Spades [Internet]. Disponible en: https://github.com/rrwick/Bandage
32. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat

graphs. Nat Biotechnol. mayo de 2019;37(5):540-6.
33. Flye [Internet]. Disponible en: https://github.com/fenderglass/Flye
34. Torsten Seemann. MLST [Internet]. Disponible en: https://github.com/tseemann/mlst
35. Torsten Seemann. Abricate.
36. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 29 de octubre de 2019;gkz935.
37. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, et al. *In Silico* Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. Antimicrob Agents Chemother. julio de 2014;58(7):3895-903.
38. Chen L. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 17 de diciembre de 2004;33(Database issue):D325-8.
39. Néron B, Littner E, Haudiquet M, Perrin A, Cury J, Rocha E. IntegronFinder 2.0: Identification and Analysis of Integrons across Bacteria, with a Focus on Antibiotic Resistance in Klebsiella. Microorganisms. 24 de marzo de 2022;10(4):700.
40. IntegronFinder V2.0 [Internet]. Disponible en: https://github.com/gem-pasteur/Integron_Finder
41. Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, Holt KE. A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex. Nat Commun. 7 de julio de 2021;12(1):4188.
42. Wyres KL, Wick RR, Gorrie C, Jenney A, Follador R, Thomson NR, et al. Identification of Klebsiella capsule synthesis loci from whole genome data. Microb Genomics [Internet]. 1 de diciembre de 2016 [citado 6 de abril de 2023];2(12). Disponible en: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000102
43. Kleborate [Internet]. Disponible en: https://github.com/klebgenomics/Kleborate
44. Bakta [Internet]. Disponible en: https://github.com/oschwengers/bakta
45. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb Genomics [Internet]. 5 de noviembre de 2021 [citado 18 de mayo de 2023];7(11). Disponible en: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000685
46. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. noviembre de 2014;15(11):524.
47. Maryland Bioinformatics Labs. parsnp [Internet]. Disponible en: https://github.com/marbl/parsnp
48. David Wilkins. gggenes: Draw Gene Arrow Maps in «ggplot2» [Internet]. 2023. Disponible en: https://CRAN.R-project.org/package=gggenes
49. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 21 de noviembre de 2019;4(43):1686.

50. Hadley Wickham, RStudio. stringr [Internet]. 2022. Disponible en: https://mirror.linux.duke.edu/cran/web/packages/stringr/stringr.pdf
51. Ahlmann-Eltze C, Patil I. ggsignif: R Package for Displaying Significance Brackets for «ggplot2» [Internet]. PsyArXiv; 2021 mar [citado 1 de marzo de 2023]. Disponible en: https://osf.io/7awm6
52. Yu G. Using ggtree to Visualize Data on Tree‐Like Structures. Curr Protoc Bioinforma [Internet]. marzo de 2020 [citado 6 de mayo de 2023];69(1). Disponible en: https://onlinelibrary.wiley.com/doi/10.1002/cpbi.96
53. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd ed. 2016. Cham: Springer International Publishing : Imprint: Springer; 2016. 1 p. (Use R!).
54. Khan ZA, Siddiqui MF, Park S. Current and Emerging Methods of Antibiotic Susceptibility Testing. Diagnostics. 3 de mayo de 2019;9(2):49.
55. Sherry NL, Horan KA, Ballard SA, Gonçalves Da Silva A, Gorrie CL, Schultz MB, et al. An ISO-certified genomics workflow for identification and surveillance of antimicrobial resistance. Nat Commun. 4 de enero de 2023;14(1):60.
56. Borthagary G, Nabón A. Primer brote de *Klebsiella pneumoniae* productora de Carbapenemasa tipo KPC en un hospital de tercer nivel. SALUD Mil [Internet]. 1 de junio de 2018 [citado 24 de mayo de 2023];37(2). Disponible en: http://revistasaludmilitar.uy/ojs/index.php/Rsm/article/view/3
57. Marquez C, Ingold A, Echeverría N, Acevedo A, Vignoli R, García-Fulgueiras V, et al. Emergence of KPC-producing *Klebsiella pneumoniae* in Uruguay: infection control and molecular characterization. New Microbes New Infect. mayo de 2014;2(3):58-63.
58. Papa-Ezdra R, Caiata L, Palacio R, Outeda M, Cabezas L, Bálsamo A, et al. Prevalence and molecular characterisation of carbapenemase-producing Enterobacterales in an outbreak-free setting in a single hospital in Uruguay. J Glob Antimicrob Resist. marzo de 2021;24:58-62.
59. Álvarez VE, Campos J, Galiana A, Borthagaray G, Centrón D, Márquez Villalba C. Genomic analysis of the first isolate of KPC-2-producing *Klebsiella pneumoniae* from Uruguay. J Glob Antimicrob Resist. diciembre de 2018;15:109-10.
60. Lucía Alonso, Gustavo Gagliano, Fabio Grill, Adriana Nabón, Verónica Seija. Plan Nacional de Acción contra la Resistencia Antimicrobiana. Abordaje desde la Salud Pública [Internet]. 2018. Disponible en: https://www.gub.uy/ministerio-salud-publica/sites/ministerio-salud-publica/files/documentos/noticias/MSP%20PLAN%20NACIONAL%20ACCION%20CONTRA%20RESISTENCIA%20ANTIMICROBIANA.pdf
61. De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND, Swann J, et al. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microb Genomics [Internet]. 1 de septiembre de 2019 [citado 18 de junio de 2023];5(9). Disponible en: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000294

62. Foster-Nyarko E, Cottingham H, Wick RR, Judd LM, Lam MMC, Wyres KL, et al. Nanopore-only assemblies for genomic surveillance of the global priority drug-resistant pathogen, Klebsiella pneumoniae. Microb Genomics [Internet]. 8 de febrero de 2023 [citado 9 de abril de 2023];9(2). Disponible en: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000936
63. Lam MMC, Wick RR, Wyres KL, Gorrie CL, Judd LM, Jenney AWJ, et al. Genetic diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in *Klebsiella pneumoniae* populations. Microb Genomics [Internet]. 1 de septiembre de 2018 [citado 8 de junio de 2023];4(9). Disponible en: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000196
64. Wang J, Feng Y, Zong Z. The Origins of ST11 KL64 *Klebsiella pneumoniae:* a Genome-Based Study. Wang H, editor. Microbiol Spectr. 13 de abril de 2023;11(2):e04165-22.
65. Zhou K, Xue CX, Xu T, Shen P, Wei S, Wyres KL, et al. A point mutation in recC associated with subclonal replacement of carbapenem-resistant Klebsiella pneumoniae ST11 in China. Nat Commun. 28 de abril de 2023;14(1):2464.
66. Kareem SM, Al-kadmy IM, Kazaal SS, Mohammed Ali AN, Aziz SN, Makharita RR, et al. Detection of gyrA and parC Mutations and Prevalence of Plasmid-Mediated Quinolone Resistance Genes in Klebsiella pneumoniae. Infect Drug Resist. febrero de 2021;Volume 14:555-63.
67. Cuzon G, Naas T, Nordmann P. Functional Characterization of Tn *4401* , a Tn *3* -Based Transposon Involved in *bla* $_{KPC}$ Gene Mobilization. Antimicrob Agents Chemother. noviembre de 2011;55(11):5370-3.
68. Gillings MR. Class 1 integrons as invasive species. Curr Opin Microbiol. agosto de 2017;38:10-5.
69. Garcillan-Barcia MP, Cruz F. Distribution of IS91 family insertion sequences in bacterial genomes: evolutionary implications. FEMS Microbiol Ecol. noviembre de 2002;42(2):303-13.
70. Hu Y, Zhang W, Shen X, Qu Q, Li X, Chen R, et al. Tandem Repeat of blaNDM-1 and Clonal Dissemination of a fosA3 and blaKPC-2 Co-Carrying IncR-F33: A–: B– Plasmid in *Klebsiella pneumoniae*Isolates Collected in a Southwest Hospital in China, 2010–2013. Infect Drug Resist. diciembre de 2022;Volume 15:7431-47.
71. Liu Y, Zhang X, Cai L, Zong Z. Enhanced survival of ST-11 carbapenem-resistant *Klebsiella pneumoniae* in the intensive care unit. Infect Control Hosp Epidemiol. junio de 2020;41(6):740-2.
72. Salazar C, Giménez M, Riera N, Parada A, Puig J, Galiana A, et al. Human microbiota drives hospital-associated antimicrobial resistance dissemination in the urban environment and mirrors patient case rates. Microbiome. 2 de diciembre de 2022;10(1):208.

Capítulo II

Generación de perfiles taxonómicos basados en la secuenciación del gen 16S completo del ARN*r*

1. Introducción

El gen 16S del ARN*r* (16S) codifica para el ARN*r* 16S que compone la subunidad ribosomal pequeña procariota (SSU). El mismo es comúnmente utilizado como un marcador taxonómico para la identificación de microorganismos procariotas y la determinación de sus relaciones filogenéticas. Este gen posee alrededor de 1.550 pares de bases y está compuesto por nueve regiones con alto grado de variabilidad (V1-V9) y nueve regiones conservadas que sirven como anclaje para la amplificación con cebadores universales. Debido a éstas características el gen 16S ha sido utilizado como marcador de la diversidad presente en el microbioma de ecosistemas complejos, como el suelo, la microbiota intestinal humana, entre otros (1–3). La posibilidad de generación de clasificaciones a nivel de especie e incluso a niveles más profundos permite la obtención de resultados relevantes, ya que en el caso de la microbiota humana, las comunidades procariotas presentes en la misma incluyen tanto especies comensales como patogénicas del mismo género (4). Si bien, las metodologías de secuenciación metagenómicas se han vuelto más accesibles, la secuenciación del gen 16S sigue siendo aún una metodología costo-efectiva para estudios de relevamiento de comunidades presentes en una gran variedad de ambientes. En términos generales, la mayor parte de los estudios están basados en la secuenciación parcial del gen 16S utilizando plataformas de segunda generación. La estrategia de análisis incluye el agrupamiento de lecturas de secuenciación (OTUs, por *Operational Taxonomic Units*) según un porcentaje de identidad fijo, por lo general del ≥ 97%. El agrupamiento y posterior generación de consensos reduce la complejidad en la comparación con la base de datos, sin embargo, limita la clasificación a niveles taxonómicos más allá del género. La herramienta QIIME se basa en esta estrategia y es uno de los softwares más utilizados para la caracterización de comunidades a partir del gen 16S (5). Otra de las estrategias ampliamente utilizadas, corresponde a la generación de un modelo de error basado en la calidad de la secuenciación. Este modelo luego realiza la predicción de errores verdaderos como consecuencia de la variación biológica y distingue los errores que se generan debido a las características de la plataforma de secuenciación. Las secuencias marcadas como "verdaderas" que difieren a partir de un único nucleótido se definen como ASVs (por *amplicon sequence variants*) independientes (6). El paquete DADA2, es la herramienta más utilizada basada en esta estrategia (7). A pesar de su extensa utilización por distintos grupos de investigación, ambas estrategias se encuentran limitadas por las regiones variables del gen 16S seleccionadas, ya que su utilización se restringe a datos de secuenciación generados

por plataformas de segunda generación que permiten la obtención de datos de una o dos regiones variables del gen 16S. Para sacar el mayor provecho de estas estrategias, resulta crítico determinar qué región presenta un mayor poder discriminatorio a nivel de especies en determinado ecosistema (4). Una de las alternativas para la generación de perfiles taxonómicos con resolución a nivel de especie es la utilización de la secuenciación del gen completo 16S con tecnologías de tercera generación. Estas permiten que todas las regiones hipervariables sean contrastadas contra una base de datos, generando mayores posibilidades de clasificación taxonómica a nivel de especie. Diferentes estudios han mostrado el potencial de las plataformas Pacific Biosciences (PacBio) y Oxford Nanopore Technologies (ONT) para la generación de perfiles taxonómicos de comunidades procariotas de baja y alta complejidad en base a lecturas de secuenciación del gen completo 16S (8–15). Debido a las diferencias entre las lecturas de secuenciación generadas con plataformas de segunda respecto a las de tercera generación, específicamente resultado de mayores proporciones de inserciones y deleciones en las mismas, las herramientas estándar utilizadas para el análisis de los datos de Illumina no resultan apropiadas para el análisis de datos generados con PacBio u ONT (16). Para el caso de ONT, se ha puesto a disposición una herramienta de código abierto (https://github.com/epi2me-labs/wf-metagenomics) para el análisis de secuencias completas del 16S generadas con la plataforma. La misma se basa en el algoritmo de *Kraken2* (17) en combinación con *Braken* (18) o *Minimap2* (19) utilizando la base de datos de NCBI RefSeq (20). Más recientemente, EMU, una herramienta diseñada específicamente para secuencias generadas con plataformas de tercera generación fue desarrollada utilizando el algoritmo *expectation-maximization* (EM) (21). La misma demostró mejores resultados en la asignación taxonómica y recuperación de la abundancia relativa de comunidades simuladas respecto de *Kraken2*, *Braken*, *NanoCLUST* (22), *Centrifuge* y *QIIME* (5).

En este trabajo se presenta el desarrollo y evaluación de *porefile*, una herramienta con diferentes módulos de análisis para la generación de perfiles taxonómicos basados en el gen 16S. El primer módulo realiza el preprocesamiento de las lecturas de secuenciación generadas con las plataformas de tercera generación, que consiste en el demultiplexado y el filtrado por tamaño y calidad de las lecturas de secuenciación de cada una de las muestras. En el siguiente módulo se interroga la última versión de la base de datos SILVA (23,24) mediante una estrategia de mapeo de las lecturas de secuenciación contra la base de datos utilizando *Minimap2* (19) y las clasificaciones resultantes se agrupan en los niveles taxonómicos principales en base al algoritmo del ancestro común más reciente o LCA (por

lower common ancestor) implementada en MEGAN6 (25). A partir de los datos de clasificación obtenidos, un paso opcional de pulido de los resultados a nivel de especie se encuentra disponible. En el mismo, se reduce la base de datos a las especies clasificadas en el primer paso de interrogación o *match* generando una base de datos personalizada. Luego las lecturas de secuenciación no asignadas o con un valor de abundancia relativa por debajo de determinado umbral son reclasificadas en base a esta base de datos reducida utilizando criterios más estrictos de mapeo o *polish*. La distribución de la abundancia relativa es calculada y los resultados se muestran en el formato de tablas (tabla de conteo y clasificación taxonómica). También se generan archivos .rma compatibles con la versión GUI de MEGAN6 para la visualización de datos y análisis.

2. Objetivo general:

Desarrollar una herramienta de clasificación de lecturas de secuenciación del gen completo 16S obtenidas con plataformas de secuenciación de tercera generación, maximizando la recuperación de clasificaciones a nivel de especie y la distribución de la abundancia relativa.

2.1 Objetivos específicos:

1. Implementar un flujo de trabajo que incluya el preprocesamiento de las lecturas de secuenciación de tercera generación, con énfasis en ONT.
2. Implementar el algoritmo LCA en base al *score* de alineamiento de las lecturas de secuenciación del gen 16S contra la base de datos para el posterior agrupamiento de las clasificaciones y generación de tablas de conteo y taxonomía.
3. Generar clasificaciones de alta resolución a nivel de especie para comunidades de alta y baja complejidad a partir de datos de secuenciación ONT.
4. Comparar los resultados obtenidos con *porefile* con otras herramientas recientes para la obtención de perfiles taxonómicos con tecnologías de tercera generación.
5. Evaluar la clasificación taxonómica obtenida de *porefile* a partir de un conjunto de datos reales de secuenciación del gen 16S con ONT y comparar los resultados obtenidos con metodologías de segunda generación para las mismas muestras.

3. Materiales y métodos

3.1 Datos simulados

Se simularon datos de secuenciación ONT para comunidades microbianas de baja (CBC) y alta complejidad (CAC). La CBC se simuló en base a una comunidad estándar de prueba comercial (ZymoBiomics Microbial Community Standard, Zymo Research), la cual contiene ocho componentes bacterianos que incluyen: *Pseudomonas aeruginosa*, *Escherichia coli, Salmonella enterica*, *Lactobacillus fermentum* (o *Limosilactobacillus fermentum*), *Enterococcus faecalis*, *Staphylococcus aureus*, *Listeria monocytogenes* y *Bacillus subtilis*. Las secuencias del gen 16S del ARN*r* de referencia para la simulación fueron obtenidas del fabricante de la comunidad de prueba comercial (https://files.zymoresearch.com/protocols/_d6300_zymobiomics_microbial_community_standard.pdf). La simulación de las lecturas de secuenciación se realizó con *Badread v0.2.0* (26). Un total de 40 Mb de lecturas de 1500 pb ± 200 pb con una distribución gama fueron simuladas a partir de las secuencias de referencia. Se utilizó el modelo de error "nanopore2020" y se generaron cuatro conjuntos de datos con un porcentaje de identidad contra la referencia entre 85% y 95% en base a una distribución beta con una desviación estándar del 5% y un máximo de 99%. Adicionalmente, se seleccionaron 112 genomas completos de alta calidad de la base de datos BV-BRC (https://www.bv-brc.org/view/Taxonomy/234#view_tab=overview) de bacterias de origen humano y ambiental (**Tabla S1**). Para la CAC, las secuencias del gen 16S del ARN*r* fueron extraídas con *barrnap* v0.9 (https://github.com/tseemann/barrnap) y se simularon lecturas de secuenciación para los genes 16S del ARN*r* extraídos como se mencionó anteriormente. Para las lecturas tanto de alta como de baja complejidad, se utilizó *Nanofilt v2.8.0* (27) para retener solamente aquellas con un largo entre 1000 y 1700 pb y una calidad mayor a 10. La CBC se generó combinando las proporciones teóricas informadas por el fabricante de ZymoBiomics microbial community standard y para la CAC se combinaron cantidades iguales de lecturas de todas las referencias simuladas.

3.2 Datos reales

Se re-analizaron datos de secuenciación del gen 16S del ARN*r* completo (regiones V1-V9) obtenidos con la plataforma ONT y a su vez, los datos de secuenciación de la

región V3-V4 obtenidos con la plataforma Illumina a partir de un conjunto de datos correspondiente a muestras obtenidas de la microbiota fecal humana publicado por Matsuo et al, 2021 (13). En este trabajo además se incluye un estándar como control positivo de ADN genómico conteniendo diez cepas correspondientes a *Bacillus cereus*, *Bifidobacterium adolescentis*, *Clostridium beijerinckii*, *Deinococcus radiodurans*, *Enterococcus faecalis*, *Escherichia coli*, *Lactobacillus gasseri*, *Cereibacter sphaeroides*, *Staphylococcus epidermidis*, *Streptococcus mutans* (MSA-1000, ATCC). Los datos de secuenciación fueron obtenidos de la base de datos NCBI SRA bajo el BioProject PRJDB9744 (**Tabla S2**). Para los datos de secuenciación ONT se utilizó *Nanofilt v2.8.0* (27) para retener las lecturas entre 1000 y 1700 pb y con una calidad mayor a 10. Para los datos de Illumina obtenidos en la plataforma MiSeq (2 X 250 pb), las mismas fueron procesadas con *Trimmomatic v0.39* (28) para la remoción de adaptadores de secuenciación y la calidad de las lecturas filtradas fueron evaluadas con *FastQC v0.11.9* (29).

3.3 Generación de perfiles taxonómicos

La clasificación taxonómica a nivel de especie a partir de las lecturas tanto simuladas como reales se realizó con *porefile*, *EMU* y *wf-metagenomics* (https://github.com/epi2me-labs/wf-metagenomics). La clasificación taxonómica a nivel de género para los datos obtenidos con la plataforma Illumina se realizó con el paquete de *DADA2* (7).

3.4 Comparación de resultados de clasificación taxonómica

Los resultados de *porefile* fueron procesados con *phyloseq v1.34.0* (30) y las tablas de abundancia de cada una de las herramientas se procesaron con múltiples paquetes de R (31–33). La clasificación taxonómica obtenida a partir de los datos simulados para cada herramienta fue evaluada con distintos parámetros. La precisión fue calculada como $\frac{vp}{vp+fp}$ y la sensibilidad como $\frac{vp}{vp+fn}$, donde *vp* corresponde a los verdaderos positivos, es decir, el número de lecturas clasificadas correctamente a nivel de especie; *fp* corresponde a los falsos positivos, es decir, el número de clasificaciones incorrectas a nivel de especie y *fn* corresponden a los falsos negativos, es decir, el número de lecturas que no fueron clasificadas a nivel de especie. Adicionalmente, se utilizó el paquete *Metrics* (34) para determinar la raíz del error cuadrático medio (RECM) entre los valores de abundancia relativa esperada versus la observada para cada método y a su vez para cada una de las

especies. El límite de detección se determinó en base a la clasificación de una comunidad estándar de prueba. Se determinó el límite de detección como la abundancia relativa más baja a la que se detecta alguno de los componentes de la comunidad de prueba.

3.5 Disponibilidad de código utilizado

Los comandos para la generación de las figuras se encuentran disponibles en https://github.com/Ceci07/porefile_figures

4. Resultados

4.1 Implementación de *porefile*

Se desarrolló un flujo de trabajo para el preprocesamiento y generación de perfiles taxonómicos a partir de datos de secuenciación del gen 16S del ARN*r* generados con tecnología de tercera generación. Los módulos de *porefile* fueron montados en el sistema de manejo de software Nextflow (35) y se encuentran disponibles en https://github.com/microgenlab/porefile. *Porefile* es una herramienta de clasificación de lecturas de secuenciación generadas a partir del gen 16S. La clasificación se basa en el mapeo de las lecturas contra de la última versión de la base de datos SILVA (*match*) (**Figura 1**).

Las clasificaciones son agrupadas en base al algoritmo del ancestro común más reciente o LCA (del *lower common ancestor*) implementado en MEGAN6. Este algoritmo es ampliamente utilizado para el agrupamiento de lecturas de secuenciación cortas en los nodos de un árbol taxonómico (como el de NCBI) en base al *score* de alineamiento de las lecturas contra una base de datos. Utilizando el algoritmo LCA *naive* implementado en MEGAN6 las lecturas de secuenciación del gen 16S son asignadas al nodo común a todos los *hits* significativos contra la base de datos (25,36). Para *porefile*, estos *hits* significativos corresponden al mejor *score* de alineamiento y todos aquellos que se encuentran dentro del mejor 10% de ese valor. A su vez, *porefile* permite la creación de la nomenclatura sinónima entre la versión más reciente de la base de datos SILVA y la de NCBI.

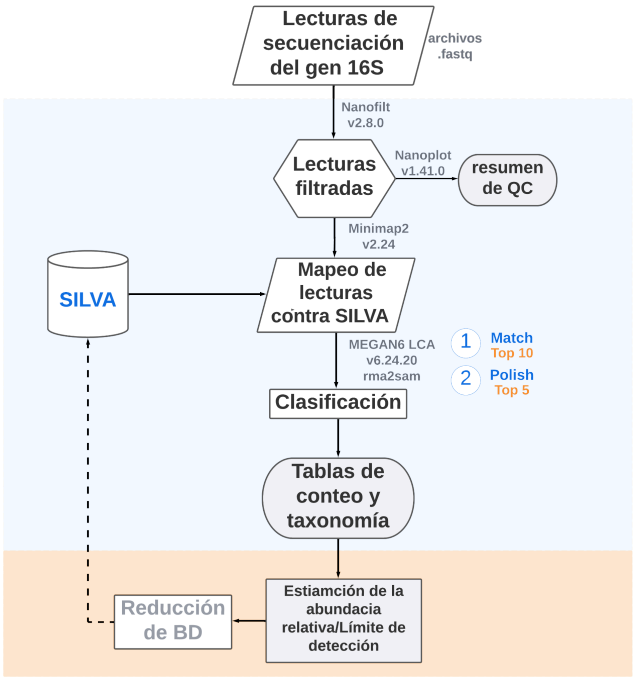

Figura 1: Esquema del flujo de trabajo de *porefile*. Las lecturas de secuenciación obtenidas en formato *fastq* son filtradas con *Nanofilt*. Las lecturas con largos entre 1000 y 1500 pb y un q-score mayor a 10 son retenidas y mapeadas contra la base de datos SILVA utilizando *Minimap2*. La clasificación de las lecturas se realiza utilizando el algoritmo LCA implementado en MEGAN6. En la etapa de *match* contra la base de datos se agrupan las lecturas dentro 10% del mejor score de alineamiento contra la base de datos. Los hits son colocados en un árbol taxonómico con la nomenclatura NCBI y a partir del mismo se generan tablas de conteo para cada uno de los nodos del árbol y su respectiva clasificación taxonómica. En el paso de *polish*, la base de datos se reduce a los taxones clasificados en el paso de *match*. Las lecturas son mapeadas contra la base de datos reducida y las lecturas se agrupan dentro del 5% del mejor *score* de alineamiento contra la base de datos.

A modo de ejemplificar los pasos del algoritmo, en la **Figura 2** se muestra que una lectura de secuenciación es clasificada como *Campylobacter lari* RM2100, *Helicobacter hepaticus* ATCC 51449 y *Wolinella succinogenes* utilizando la estrategia de alineamiento contra la base de datos. Debido a que no hay una diferencia significativa entre el *score* de alineamiento entre ellas mayor al 10% del mejor *hit*, el algoritmo asigna esa lectura al orden *Campylobacterales,* el cual es el ancestro común a todos los *hits* significativos (25). Una vez aplicado el algoritmo LCA, se realiza el conteo de las lecturas asignadas a cada nodo del árbol taxonómico NCBI, generando una tabla de conteo y una tabla con la taxonomía conteniendo la anotación completa de la clasificación (Dominio, Reino, Filo, Clase, Orden, Familia, Género y Especie). A partir de la tabla de asignación taxonómica y de conteo se puede inferir abundancia relativa de los taxones y generar resultados estándar de estudios de perfilado taxonómico (curvas de rarefacción, determinación de la diversidad alfa y beta, etc). Adicionalmente, *Porefile* incluye un paso de refinamiento de la clasificación a nivel de especie (*polish*). Para este paso, un módulo de *porefile* extrae de la base de datos SILVA las

secuencias de referencia que corresponden a las clasificaciones obtenidas a nivel de especie en las muestras (base de datos reducida) y mapea las lecturas de secuenciación que no fueron asignadas o se encuentran por debajo de un umbral arbitrario contra la base de datos reducida. Los archivos de mapeo son analizados por el algoritmo LCA, con los parámetros de *score* de alineamiento más estrictos, específicamente se asigna una especie a la lectura *r* si *r* tiene un *score* de alineamiento contra la especie *s* dentro del 5% por ciento superior al mejor *hit* obtenido para la especie *s*. La finalidad de este paso es la generación de una nueva base de datos personalizada para el tipo de muestra analizada, disminuir la cantidad de alineamientos contra la base de datos y mejorar la recuperación de la abundancia relativa de los componentes de la muestra en cuestión.

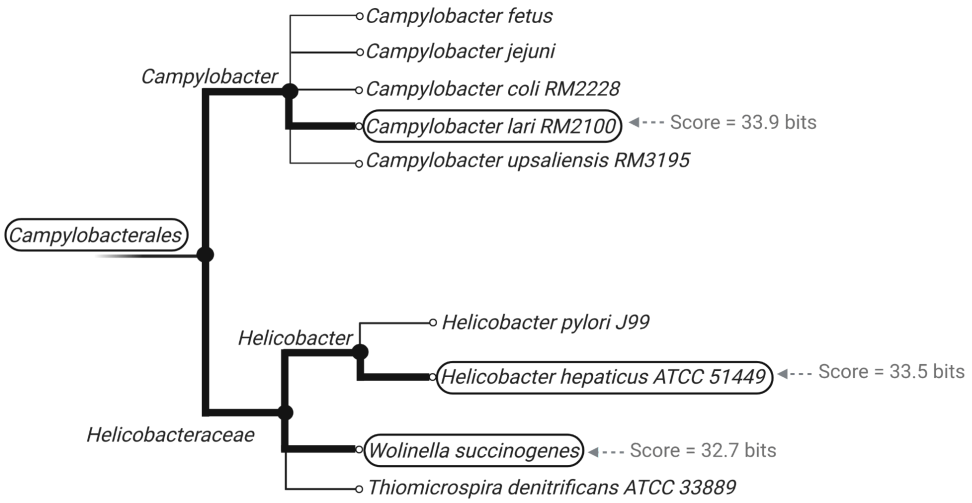

Figura 2: Algoritmo LCA utilizado por *porefile*. A la derecha se muestra en score de alineamiento obtenido para una lectura de secuenciación *r* contra *Campylobacter lari*, *Helicobacter hepaticus* y *Wolinella succinogenes*. El algoritmo LCA asigna la lectura *r* al taxón *Campylobacterales*, ya que es el ancestro común más reciente para las tres especies con la que *r* encontró coincidencias en la base de datos. Adaptado de Hudson *et al.*, 2007.

4.2 Porefile recupera la mayor parte de los componentes de una comunidad de prueba simulada

Con la finalidad de remover los sesgos introducidos durante el procesamiento de muestras y amplificación del gen S, se realizó la simulación de lecturas de secuenciación ONT en base a secuencias conocidas de distintas especies. A modo de determinar el mejor rango para el parámetro de identidad de las simulaciones generadas con *Badread*, se generaron comunidades de prueba con un 85% a 95% de identidad contra las respectivas referencias (*Pseudomonas aeruginosa*, *Escherichia coli, Salmonella enterica*, *Lactobacillus fermentum*, *Enterococcus faecalis*, *Staphylococcus aureus*, *Listeria monocytogenes* y

Bacillus subtilis). Asimismo, se utilizó el parámetro de identidad por defecto asignada por el simulador, correspondiente a 87.5% ± 5.5% con un máximo de 97.5%, generando diez réplicas de comunidades de pruebas simuladas de baja complejidad. Se realizó la asignación taxonómica con *porefile* utilizando la opción de pulido a nivel de especie y adicionalmente se realizó la clasificación con *EMU* y *wf-metagenomics* (*Minimap2*).

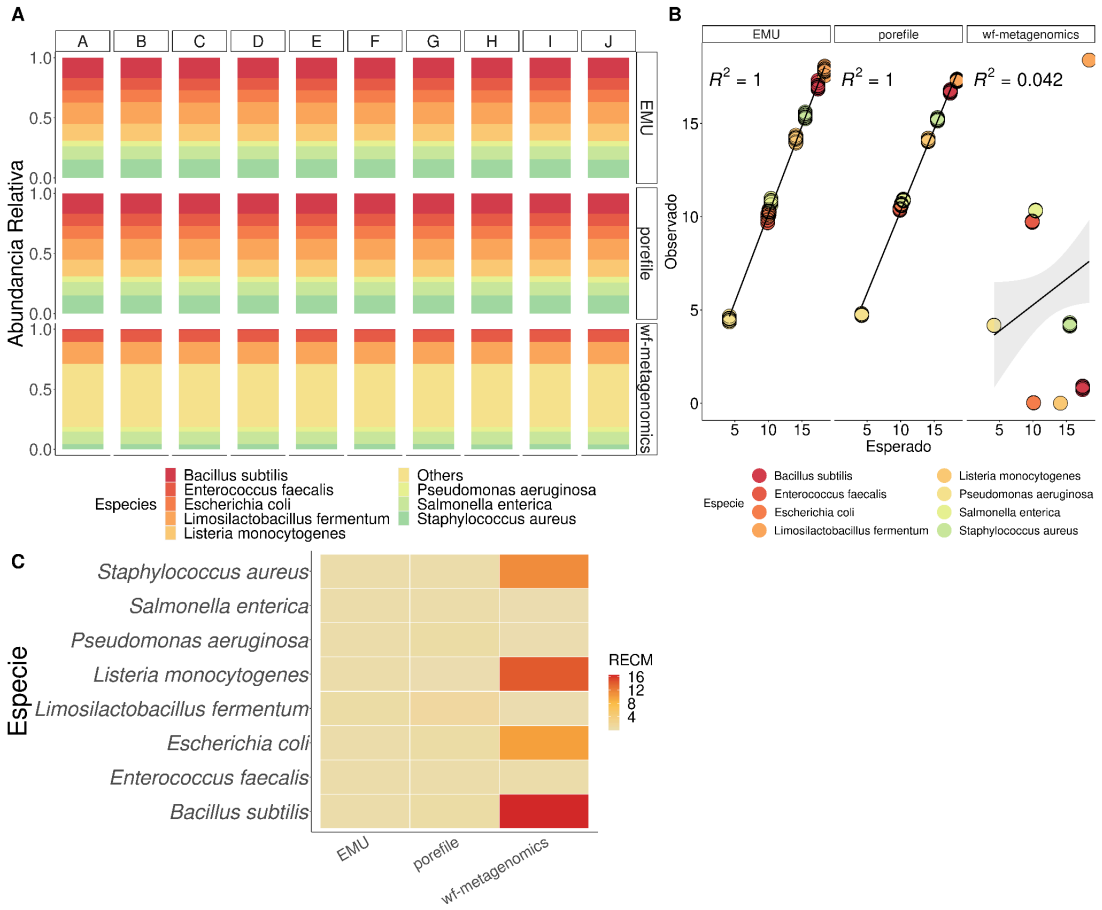

Figura 3: Clasificación taxonómica de una comunidad simulada de baja complejidad. A) Clasificación de las réplicas de las muestras simuladas con *porefile*, *EMU* y *wf-metagenomics*. B) Regresión lineal entre las abundancias relativas esperadas y las observadas para cada uno de los componentes de la comunidad de prueba. C) RECM de cada uno de los componentes en las distintas réplicas de la comunidad de prueba. La clasificación *Bacillus spizizenii* fue reemplazada por *Bacillus subtilis*, ya que *Bacillus spizizenii* se considera una subespecie de *Bacillus subtilis* (*Bacillus subtilis sub. spizizenii*).

Durante el primer paso de *porefile,* el límite de detección (LD) de la abundancia relativa de los componentes de la comunidad fue 0.0258 ± 0.0005. Se tomaron todas las secuencias de la base de datos por encima de este límite en la base de datos SILVA (utilizando la opción --lowAbundanceThreshold) y se creó una nueva base de datos reducida. En la etapa de pulido a nivel de especie (*polish*) las lecturas sin asignar o por debajo del LD de secuenciación son contrastadas nuevamente contra esta versión de la base de datos y la clasificación se realiza utilizando el algoritmo LCA. En la **Figura 3A** se

observan los resultados de la clasificación taxonómica obtenida con *porefile*, *EMU* y *wf-metagenomics*. Tanto con *porefile* como con *EMU* y *wf-metagenomics* se detectaron todos los componentes de la comunidad de prueba simulada (**Figura 3A**), sin embargo, solamente con *porefile* y *EMU* se obtuvieron una relación lineal óptima entre la abundancia relativa esperada versus la observada (**Figura 3B**).

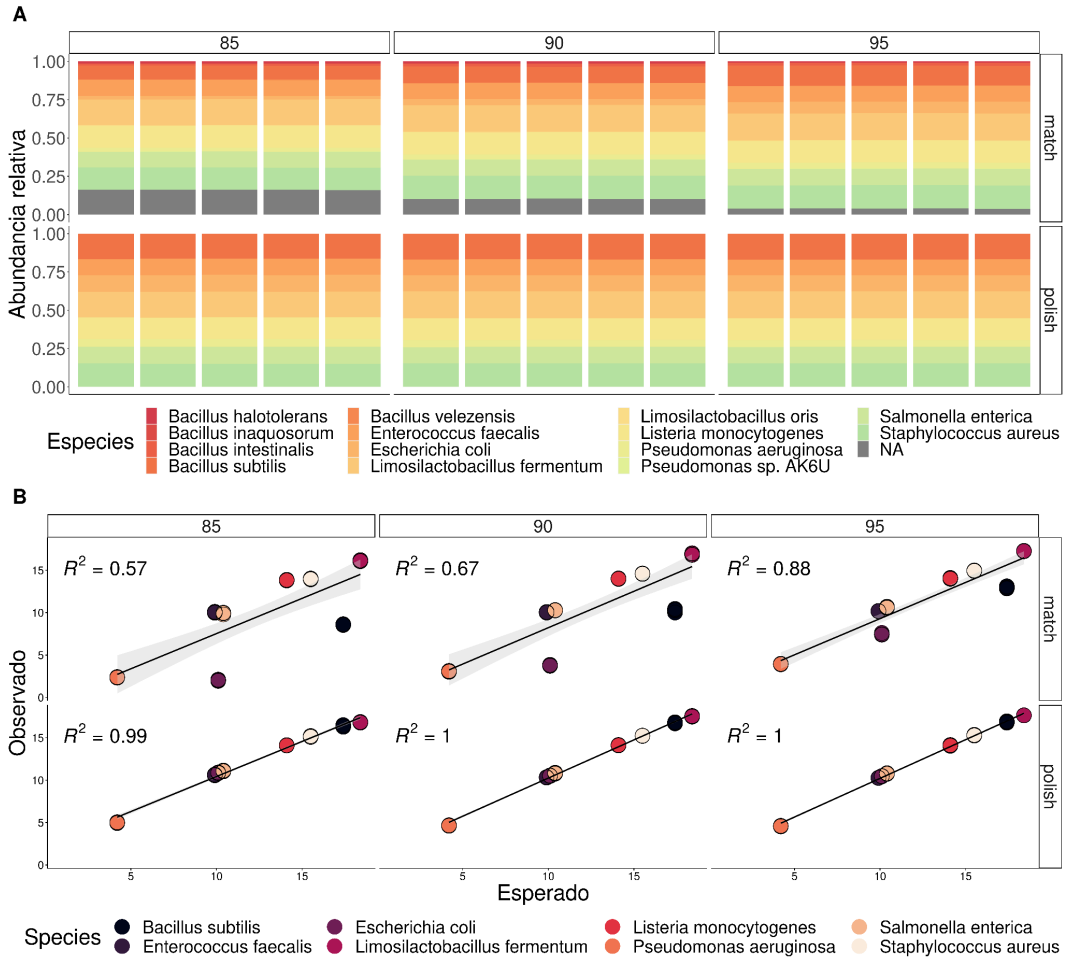

Figura 4: Clasificación taxonómica y recuperación de la abundancia relativa de una comunidad de prueba simulada de baja complejidad variando el parámetro de identidad contra las referencias utilizadas para la simulación. A) Clasificación taxonómica obtenida en el primer y segundo paso de interrogación de la base de datos SILVA. B) Ajuste lineal entre las abundancias relativas esperadas y observadas.

A modo de determinar qué especies contribuyen en la desviación de las abundancias relativas esperadas para cada herramienta, se determinó el RECM para cada una de las especies y cada uno de los resultados obtenidos con *porefile, EMU* y *wf-metagenomics*. En general se obtuvo un valor cercano a cero de RECM con *porefile* y *EMU* (0.53 ± 0.29), sin embargo con wf-metagenomics se obtuvo un valor de 6.53 ± 7.17, siendo *Bacillus subtilis*, *Listeria monocytogenes*, *Staphylococcus aureus* y *Escherichia coli* los mayores

contribuyentes en la desviación de la abundancia relativa esperada para la comunidad de prueba simulada (**Figura 3C**). En la **Figura S1** se pueden observar específicamente la abundancia relativa recuperada de los componentes de la comunidad de prueba para cada una de las herramientas. Adicionalmente, las lecturas simuladas obtenidas con otros porcentajes de identidad respecto a las referencias de la comunidad de prueba fueron evaluadas y los resultados se muestran en la **Figura 4.** Se observa que para el rango de identidades entre 85% y 95% los componentes de la comunidad de prueba son detectados en los dos pasos (*match* y *polish*) de interrogación de la base de datos. En segundo lugar, hay una disminución en la cantidad de lecturas sin asignar o mal asignadas luego del segundo paso de interrogación (*polish*) contra la base de datos reducida para todas las identidades ensayadas. Esto tiene como resultado un aumento tanto de la precisión como de la sensibilidad cuando se compara el primer paso (precisión 86% ± 1.1%, sensibilidad 72.9% ± 8.5%), respecto al segundo paso (precisión 93.3% ± 2.7% y sensibilidad 83.2% ± 4.4%). Asimismo, se observa una mejora en la recuperación de la abundancia relativa esperada en todos los rangos de identidad ensayados.

En la **Figura 5A** se muestra un muestreo al azar de 20 componentes de las comunidades de pruebas de alta complejidad. Luego del paso de pulido a nivel de especie llevado a cabo por *porefile*, se obtiene para la CAC una mejor recuperación de la abundancia relativa esperada, respecto a la obtenida con *wf-metagenomics* y *EMU* **(Figura 5B)**. Asimismo, se detectó el 91.9% de los componentes de CAC con *porefile*, mientras que con *EMU* se detectaron el 86.6% y *wf-metagenomics* pudo detectar el 89.3%. Para el caso de las muestras humanas, *porefile* detectó el 91% de los componentes, mientras que *EMU* 89% y *wf-metagenomics* 89%. Por otra parte, *porefile* pudo detectar el 100% de los componentes de la comunidad de prueba de origen ambiental, mientras que *EMU* pudo detectar el 66.7% y *wf-metagenomics* el 83.3% (**Figura 5C**). Las especies que no fueron detectadas por los tres métodos fueron *Acinetobacter gyllenbergii*, *Erysipelotrichaceae bacterium 66202529* y *Anaerococcus mediterraneensis*. En ningún caso se observa una regresión lineal óptima entre lo esperado y lo observado. Es decir, la abundancia relativa no es recuperada de forma homogénea entre los componentes de la CAC. Sin embargo, se obtiene un mejor coeficiente R^2 para los resultados generados con *porefile*, respecto de *EMU* y *wf-metagenomics*. En resumen, *porefile* pudo recuperar más componentes del conjunto de comunidades de prueba simuladas con componentes de orígen humano y ambiental, respecto a los demás métodos y a su vez, si bien la recuperación de abundancia relativa no fue óptima, *porefile* mostró un coeficiente de ajuste lineal superior respecto a lo

observado con *EMU* y *wf-metagenomics*. Por lo tanto, *porefile* mejora la recuperación de la abundancia relativa de comunidades de prueba simuladas de alta complejidad y detecta una mayor cantidad de componentes de la CAC, particularmente, las de origen ambiental, en contraposición de *wf-metagenomics* y *EMU*.

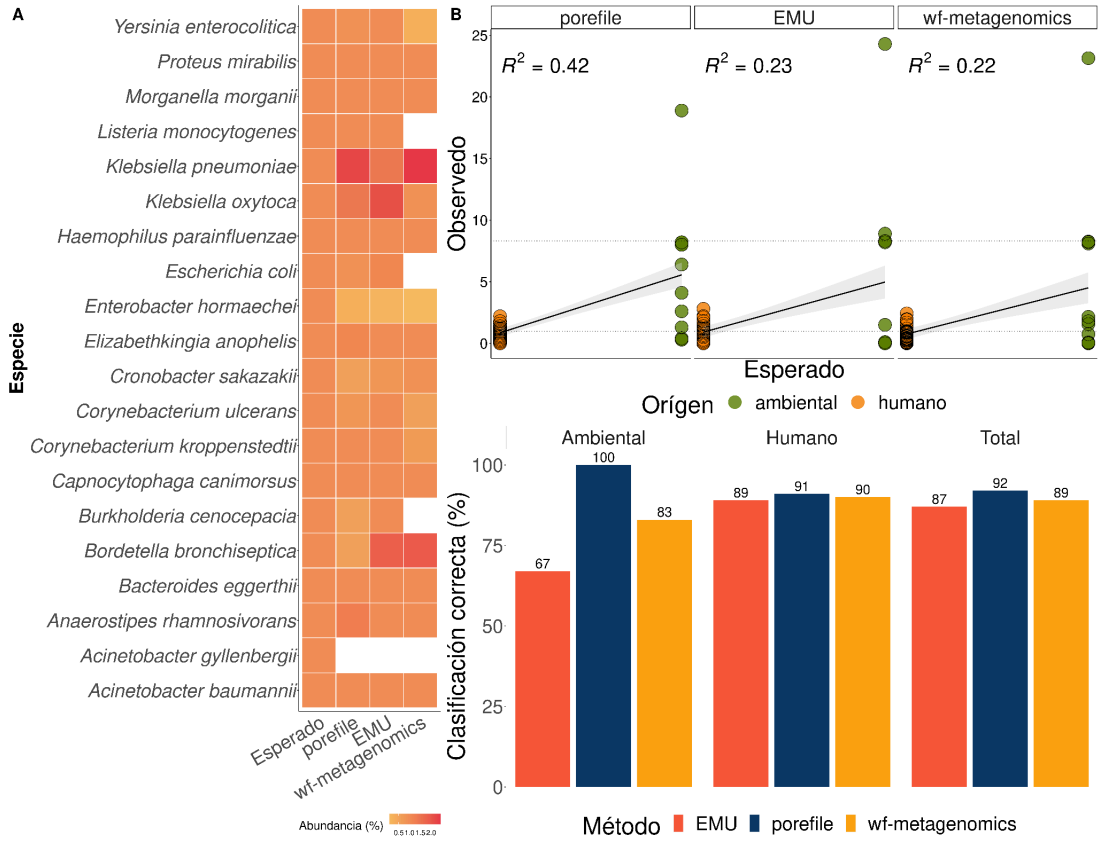

Figura 5: Clasificación taxonómica de la comunidad de prueba simulada de alta complejidad. A) Muestreo al azar de veinte componentes de las comunidades de prueba de alta complejidad simuladas (origen humano y ambiental) clasificadas con *porefile*, *EMU* y *wf-metagenomics*. B) Regresión lineal entre la abundancia relativa esperada y la observada para las comunidades de prueba simuladas de origen humano y ambiental. C) Porcentaje de clasificaciones correctas obtenidas con cada uno de los métodos de clasificación tanto para las comunidades de origen humano, ambiental y la totalidad.

4.3 Porefile genera perfiles taxonómicos a nivel de especie a partir del microbioma intestinal humano y recupera en gran parte la taxonomía del perfil generado con tecnologías de segunda generación a nivel de género.

A modo de evaluar las capacidades de *porefile* de generar perfiles taxonómicos a partir de datos de secuenciación reales del gen completo del 16S del ARN*r*, se utilizó un conjunto de datos previamente generados por Matsuo et al., 2021 (13). En este trabajo se incluyen datos de la amplificación de las regiones V1-V9 del gen 16S secuenciados con la

plataforma ONT. A su vez, se generaron amplicones de las regiones V3-V4 y dichos amplicones fueron secuenciados con la plataforma Illumina MiSeq (2 x 250 pb). Este conjunto de datos incluye además una comunidad estándar de prueba de ADN (MSA-1000, ATCC). Los datos ONT fueron evaluados con *porefile* y *EMU*, mientras que los datos de Illumina fueron evaluados con la estrategia de generación de ASVs con el paquete *DADA2*. Para el caso de *porefile* se muestra la comparación en la recuperación de la abundancia relativa en el paso de *match* y *polish* en la **Figura S2**.

La inclusión de una comunidad de prueba estándar de composición teórica conocida, permitió que se obtuviera información respecto a la capacidad de detección de los componentes, la recuperación de la abundancia relativa y el LD de los componentes de la misma. El LD de los componentes de la comunidad estándar luego del primer paso de interrogación (*match*) contra la base de datos fue de 0.004.

La abundancia relativa recuperada para la comunidad de prueba refleja potenciales sesgos. En este caso, la comunidad de prueba se añadió a partir del paso de amplificación, por lo que la desviación en la abundancia relativa observada respecto a la esperada es atribuíble a los cebadores seleccionados, a las condiciones de la PCR y la estrategia de preparación de la biblioteca de secuenciación. Se observa que todos los componentes de la comunidad de prueba son detectados tanto con *porefile* como con *EMU* (**Figura 6**). Se observa que para el caso de *porefile* las mayores desviaciones respecto a la composición teórica vienen dadas por la subestimación de las bacterias gram negativas *Bifidobacterium adolescentis, Escherichia coli* y la bacteria gram positiva *Bacillus cereus*. Asimismo, *Rhodobacter sphaeroides*, una bacteria gram negativa se encuentra sobreestimada. Para el caso de *EMU*, las desviaciones vienen dadas principalmente por la sobre estimación de la bacteria gram positivas *Clostridium beijerinckii* y subestimación de *Bifidobacterium adolescenti*s y *Bacillus cereus*. La capacidad de recuperación de la abundancia relativa teórica fue medida con la raíz del error cuadrático medio (RECM). Los valores de RECM son cercanos a cero tanto con *porefile* como *EMU*. Específicamente, *porefile* presenta un RECM (0.049) levemente mayor que a *EMU* (0.037). Sin embargo, el porcentaje de clasificaciones por debajo del LD, es mayor con *porefile* (5.5%), respecto a EMU (0.5%) a nivel de especies.

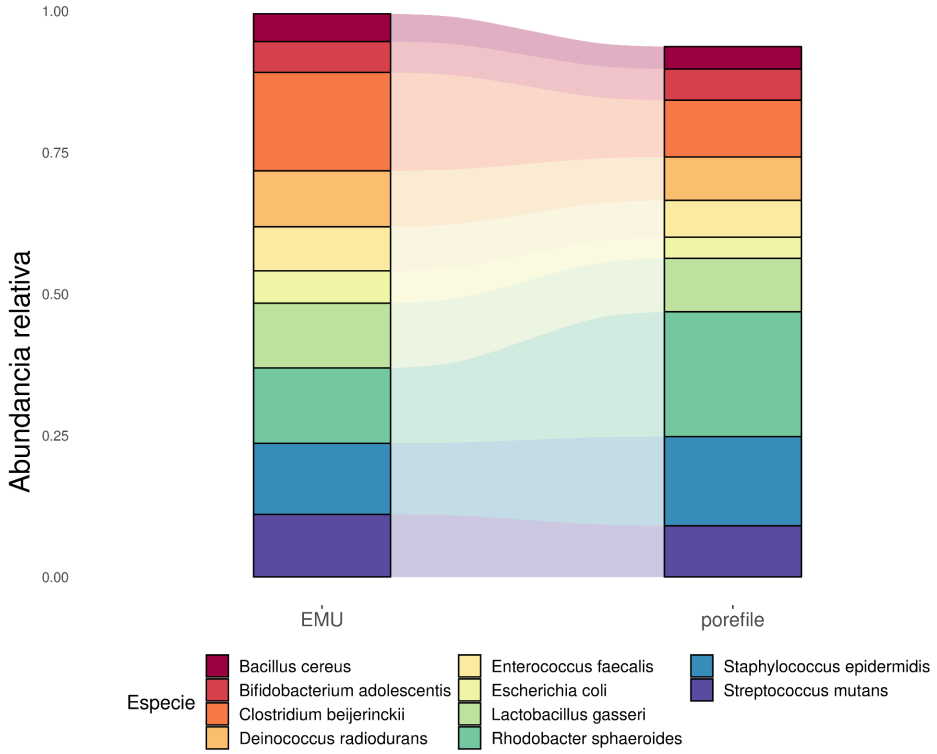

Figura 6: Clasificación taxonómica obtenida con *EMU* y *porefile* de la comunidad estándar de prueba MSA-1000 (ATCC) correspondiente a la secuenciación con la plataforma ONT del conjunto de datos obtenidos de Matsuo et al, 2021. La comunidad estándar está compuesta por cantidades iguales de ADN genómico de *Bacillus cereus* (ATCC 10987), *Bifidobacterium adolescentis* (ATCC 15703), *Clostridium beijerinckii* (ATCC 35702), *Deinococcus radiodurans* (ATCC BAA816), *Enterococcus faecalis* (ATCC 47077), *Escherichia coli* (ATCC 700926), *Lactobacillus gasseri* (ATCC 33323), *Rhodobacter sphaeroides* (ATCC 17029), *Staphylococcus epidermidis* (ATCC 12228) y *Streptococcus mutans* (ATCC 700610). Se muestran las clasificaciones por encima del límite de detección de *porefil*e (0.037). El RECM para *porefile* fue de 0.049 y 0.037 para EMU. Adicionalmente, la cantidad de clasificaciones incorrectas fue del 5.5% para *porefile* y 0.5% para EMU.

4.4 *Porefile* aumenta la resolución en la clasificación al utilizar una base de datos específica del ambiente

Se evaluaron los datos de secuenciación del gen 16S obtenidos a partir de muestras de microbiota intestinal humana del trabajo de Matsuo et al, 2021. A partir del primer paso de interrogación contra la base de datos (*match*) se obtuvieron los taxones para la creación de la base de datos reducida.

Al realizar la reclasificación de las secuencias (*polish*) utilizando la base de datos reducida y un criterio de agrupamiento en los nodos del árbol taxonómico más estrictos (top 5%), se clasificaron la totalidad de las secuencias de las muestras, aumentando la abundancia relativa de muchas especies que habían sido recuperadas en menor proporción en el paso de *match* (**Figura 7**).

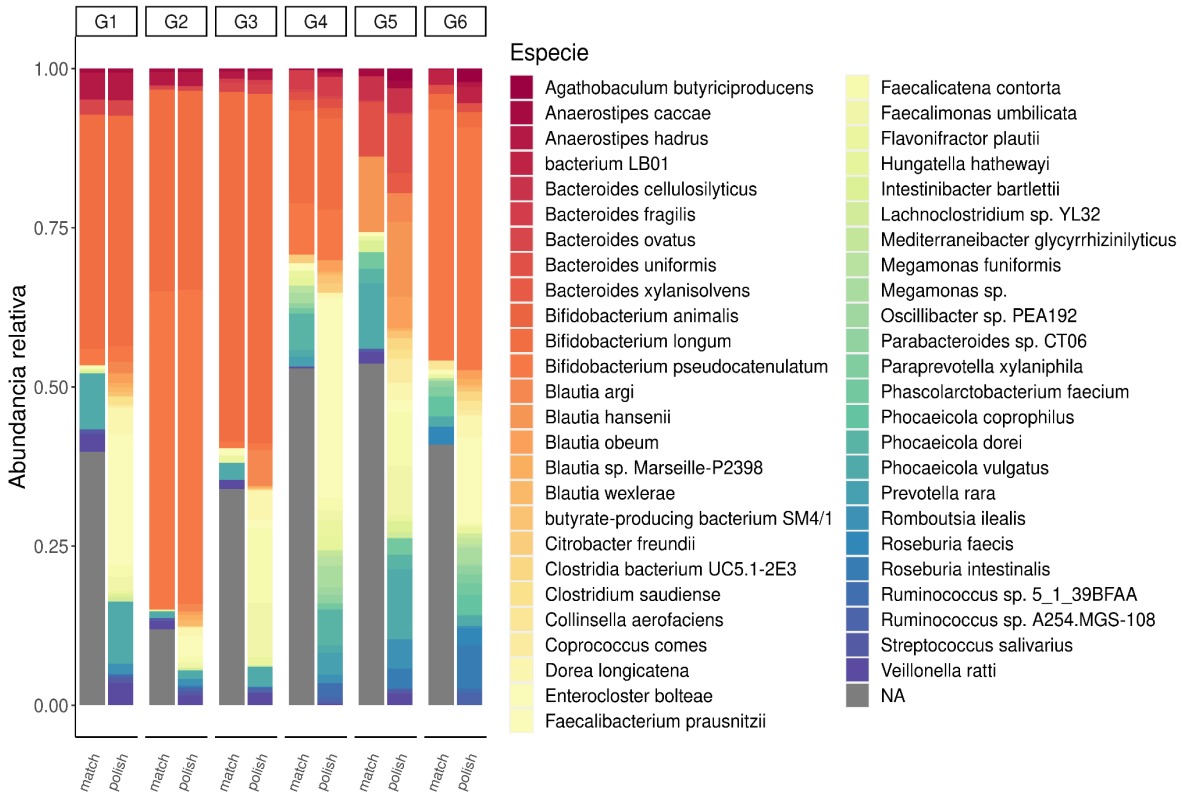

Figura 7: Análisis de datos de secuenciación del gen 16S completo con la plataforma ONT para el dataset de muestras de microbiota intestinal humana del trabajo de Matsuo et al, 2021. Se muestra la clasificación obtenida en el paso de *match* y *polish* con *porefile* de cada una de las muestras (G1-G6).

4.5 *Porefile* recupera en gran medida la clasificación a nivel de Género respecto a la clasificación obtenidas con la estrategia de ASVs a partir de datos de Illumina.

Se realizó la comparación de los datos de secuenciación del gen 16S obtenidos de la región V1-V9 en la plataforma ONT y las regiones V3-V4 obtenidos con la plataforma Illumina del conjunto de datos publicado por Matsuo et al, 2021 de las muestras de microbiota intestinal humana. A modo de determinar si los resultados obtenidos con *porefile* son consistentes con los datos obtenidos con la metodología estándar de secuenciación de lecturas cortas. Las mismas fueron analizadas con la estrategia de obtención de ASVs del paquete DADA2. Asimismo, se incluyeron los datos de clasificación obtenidos para las regiones V1-V9 con EMU. En la **Figura 8** se muestra una comparación de 56 géneros con una abundancia relativa mayor a 0.005. La totalidad de las clasificaciones se muestra en la

Figura S3. Del total de comparaciones se obtuvo que *porefile* recupera el 82% de los géneros clasificados con *DADA2*, mientras que *EMU* recupera el 85%.

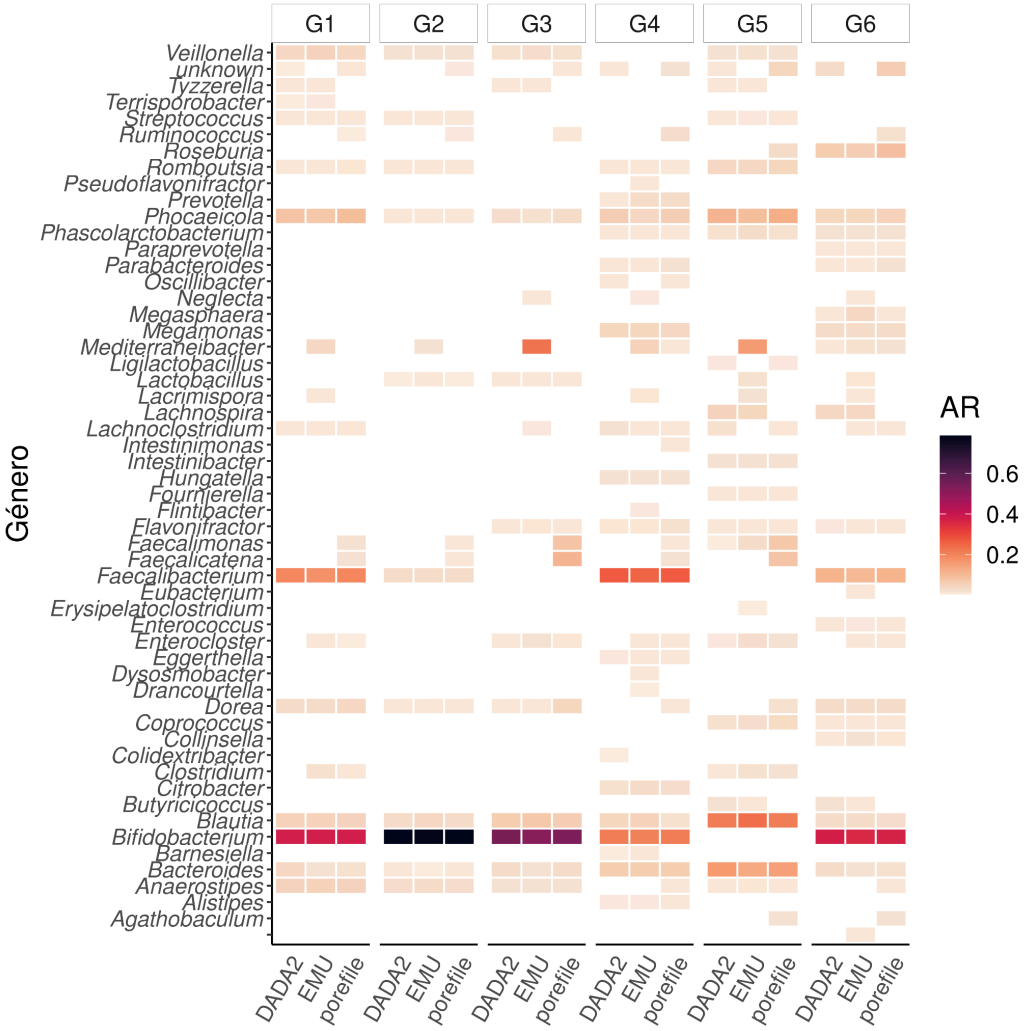

Figura 8: Comparación entre las clasificaciones a nivel de género obtenidas con datos de secuenciación ONT e Illumina. Las secuencias ONT fueron analizadas con *porefile* y *EMU* y las secuencias Illumina fueron analizadas con *DADA2*. Se muestran todas las clasificaciones con una abundancia relativa mayor a 0.005.

El género predominante detectado con la secuenciación ONT tanto por *porefile* como con EMU corresponde al género *Bifidobacterium*. Este resultado es confirmado con todas las estrategias de análisis seleccionadas. El detalle de las clasificaciones detectadas a nivel de género y especie se encuentran en las **Figuras S4 y S5.** A través del análisis de los datos de secuenciación del gen 16S completo se detectó que las especies predominantes en el conjunto de datos corresponde a *Bifidobacterium longum* y *Bifidobacterium pseudocatenulatum.*

5. Discusión

La utilización de marcadores moleculares como el gen 16S han sido utilizados extensamente desde la aparición de las tecnologías de secuenciación masiva. Esto determinó el desarrollo de distintas herramientas bioinformáticas para el análisis de los datos generados, muchas de las cuales se han integrado a procedimientos estándar para el estudio de comunidades microbianas. El uso más frecuente de este tipo de abordaje es la generación de perfiles taxonómicos y de diversidad entre distintas condiciones a partir de muestras de distintos orígenes de una forma costo-efectiva. En este sentido, un estudio reciente ha mostrado que la información obtenida con la secuenciación de las regiones hipervariables V3-V4 del gen 16S es comparable con la información taxonómica obtenida a través de la secuenciación *shotgun* de baja cobertura (37). Si bien éstas técnicas son ampliamente utilizadas, es necesario conocer los sesgos introducidos en las distintas etapas del procesamiento; desde la extracción de ADN de las muestras y amplificación hasta el análisis de datos. Para ello es indispensable el uso de controles tanto positivos como negativos para la correcta interpretación de los resultados. La utilización de controles positivos, como las comunidades microbianas de prueba, permite determinar si existen sesgos hacia algún tipo de microorganismo, el límite de detección confiable y el impacto de potenciales contaminaciones cruzadas. Es sabido que los procedimientos de extracción de ADN de las muestras constituyen uno de los puntos más susceptibles a la introducción de sesgos, éstos provienen principalmente de las diferencia en las eficiencias de extracción debido a la composición de la pared celular de los distintos microorganismos, por lo que la inclusión de comunidades microbianas de prueba que incluyan distintos tipos de cepas (por ejemplo, Gram positivas y negativas) en esta etapa permite conocer si la metodología utilizada para la extracción de ADN impacta en la detección de la abundancia relativa de los componentes de la muestra y en qué extensión (38,39). Incluso se ha propuesto un modelo matemático que apunta a corregir estos sesgos en base a la medición de la eficiencia obtenida en cada paso del flujo de trabajo (40). Otro de los puntos de introducción de sesgos, es la etapa de amplificación del gen 16S a través del uso de cebadores “universales”. Sin embargo, es sabido que dichos cebadores universales no amplifican con la misma eficiencia el gen 16S de todos las especies presentes en una determinada muestra. Otros de los factores que introducen sesgos es el tipo de polimerasa, las *Taq* polimerasas tienen un porcentaje mayor de errores, respecto a las polimerasas de alta fidelidad (41). A partir de esto, se han propuesto estrategias alternativas para disminuir los sesgos asociados a la amplificación del gen 16S, como la

utilización de cebadores capaces de amplificar el gen completo del 16S o incluso el operón *rrn* completo a modo de recuperar una mayor variabilidad en esa región y aumentar el poder discriminatorio entre las especies (15,42–44).

Con el surgimiento de las nuevas tecnologías de secuenciación de tercera generación, se han creado oportunidades para explorar la diversidad presente en muestras de distintos orígenes utilizando el gen completo 16S a partir de datos de secuenciación masiva. Esto ha determinado la necesidad de desarrollo de herramientas bioinformáticas para el análisis de este tipo de datos, teniendo en cuenta el porcentaje de error asociado a la plataforma de secuenciación. En este contexto se desarrolló *porefile*, una herramienta de clasificación taxonómica, cuyos módulos fueron integrados dentro del sistema Nextflow. *Porefile* utiliza la estrategia de mapeo contra la última versión de la base de datos SILVA para determinar los taxones presentes en las muestras analizadas. El agrupamiento de las clasificaciones se realiza utilizando el algoritmo LCA, luego implementa una etapa de pulido a nivel de especie y para ello recluta de la base de datos las secuencias detectadas en la primera etapa de comparación (*match*) para crear una base de datos reducida. Luego se contrastan nuevamente las lecturas de secuenciación no asignadas o en baja abundancia contra esta base de datos reducida (*polish*) y se reporta los resultados a nivel de especie para las comunidades presentes en las muestras. A partir de datos de secuenciación ONT simulados, es decir, sin los sesgos clásicos de los estudios del microbioma mencionados anteriormente, se determinó que *porefile* es una herramienta capaz de detectar los componentes de comunidades microbianas de baja y alta complejidad. Incluso, se mostró que presenta mejor capacidad de detección y recuperación de la abundancia relativa de muestras ambientales, respecto a otras herramientas utilizadas. Otros autores han propuesto la utilización de base de datos nicho-específicas para mejorar la clasificación de las secuencias del gen 16S (4,45,46). *Porefile* incorpora estos conceptos mediante la obtención de la información contenida en las muestras para la generación de una base de datos reducida a las especies detectados en el *dataset*, creando una base de datos específica para el ambiente, permitiendo una mejora en la recuperación de la abundancia relativa de los componentes además de su detección.

Cuando se evaluó *porefile* con un *dataset* de secuenciación del gen 16S previamente publicado (13) se obtuvieron resultados similares tanto con *porefile* como con EMU para la comunidad de prueba comercial presente en el dataset en términos de detección y RCEM de la abundancia relativa. Se observaron desviaciones puntuales de la abundancia relativa para algunos de sus componentes, pero estas desviaciones no parecen tener un impacto mayor cuando son evaluados en su conjunto. Cabe destacar que tanto *porefile* como EMU

recuperaron todos los componentes de la comunidad estándar de prueba a nivel de especie utilizando la base de datos SILVA, sin embargo, los autores de la publicación no pudieron detectar *Escherichia coli* en la misma utilizando una estrategia de mapeo con *minimap2* contra una base de datos personalizada (GenomeSync). Al evaluar los datos generados en este trabajo para las muestras de microbiota intestinal humana, tanto a partir de las regiones V1-V9 del 16S con la plataforma ONT, como para la región V3-V4 con la plataforma Illumina, se observó que *porefile* recupera en similares proporciones los componentes de la microbiota fecal a nivel de género, respecto de EMU. Adicionalmente, se detectó que el género *Bifidobacterium* es predominante en cinco de seis muestras, siendo *Bifidobacterium longum* y *Bifidobacterium pseudocatenulatum* las especies más representadas. Las mismas han sido previamente reportadas como parte de la microbiota intestinal (47). Específicamente, *B. longum* estaría involucrada en el desarrollo de la respuesta inmune temprana. Por su parte, *B. pseudocatenulatum* se encuentra dentro del mismo cluster filogenético que *B. adolescentis* y su funciones están asociadas a la fermentación de una gran variedad de carbohidratos. Algunas cepas de esta especie han sido reportadas con potencial probiótico (48). Los autores del trabajo no reportaron *B. pseudocatenulatum* en las muestras, pero si *B. adolescentis* y otras especies de *Bifidobacterium* aisladas previamente de la microbiota intestinal y otras especies de origen no humano. El género *Bifidobacterium* tiene un rol central en la homeostasis de la microbiota intestinal, por lo que contar con herramientas capaces de obtener información de alta resolución y de bajo costo de este género generaría impacto positivo en esta área de estudio.

Si bien se requiere que un número mayor de *datasets* sean evaluados, especialmente, a partir de muestras de distintos orígenes, estos resultados indican que *porefile* es una herramienta capaz de clasificar lecturas de secuenciación ONT de forma automática y reproducible, ya que sus módulos son manejados en un sistema que permite el rastreo de todos los procesos y resultados generados (35). Asimismo, genera archivos de salida compatibles con herramientas de visualización gráfica de los resultados y archivos de conteo y taxonomía compatibles con herramientas de uso común en el estudio de los perfiles taxonómicos como *phyloseq* (30).

Finalmente, la metodología implementada en *porefile* tiene el potencial de ser extendido a otros marcadores moleculares utilizados en el área de la microbiología, como el gen 18S para el relevamiento de las comunidades eucariotas, ya que es posible hacer el ajuste de diversos parámetros como el tamaño del amplicón. Además utiliza la base de datos SILVA que también incluye una extensa colección de secuencias del gen 18S del ARN*r*.

6. Material suplementario

Tabla S1: Genomas completos obtenidos de BV-BRC para la simulación de lecturas de secuenciación ONT para la CAC

ID BV-BRC	Especie	Genoma	Fecha de la muestra	Grupo geográdico	Origen
1031746.3	*Campylobacter hyointestinalis*	Completo	1986	Europa	Humano
1032071.3	*Aliarcobacter cryaerophilus*	Completo		Norteamérica	Humano
103.855.148	*Bordetella hinzii*	Completo		Norteamérica	Humano
1051974.3	*Granulibacter bethesdensis*	Completo			Humano
106.654.153	*Acinetobacter nosocomialis*	Completo	2015-03-18	Asia	Humano
1095685.4	*Neisseria meningitidis*	Completo	2000-03	Asia	Humano
1.117.645.837	*Elizabethkingia anophelis*	Completo	2012-12-08	Asia	Humano
1.134.687.569	*Klebsiella michiganensis*	Completo	2016-12-12	Asia	Humano
1215914.3	*Lacticaseibacillus casei*	Completo			Humano
1229621.4	*Anaerostipes rhamnosivorans*	Completo	2011	Europa	Humano
1262462.5	*Yersinia enterocolitica*	Completo			Humano
128.038.018	*Staphylococcus aureus*	Completo	2019	Africa	Humano
128944.3	*Aerococcus urinaehominis*	Completo	2001	Europa	Humano
131.343.095	*Streptococcus pneumoniae*	Completo	2009	Europa	Humano
13.142.833	*Streptococcus pyogenes*	Completo	2009	Asia	Humano
1325724.3	*Brevundimonas vancanneytii*	Completo	1800-2017		Humano
1335.4	*Streptococcus equinus*	Completo	2014-10-27	Norteamérica	Humano
134534.15	*Acinetobacter gyllenbergii*	Completo	2013-01-01	Asia	Humano
13.516.515	*Enterococcus faecalis*	Completo	2021		Humano
135.214.893	*Enterococcus faecium*	Completo	2017-04-12	Asia	Humano
1385930.3	*Burkholderia dolosa*	Completo			Humano
1385939.3	*Bifidobacterium breve*	Completo			Humano
1.408.387	*Bacillus pumilus*	Completo	2005-01-01	Norteamérica	Humano
1416803.3	*Bordetella genomosp. 9*	Completo	2008		Humano
1.463.165.355	*Klebsiella quasipneumoniae*	Completo		Norteamérica	Humano
146827.6	*Corynebacterium simulans*	Completo	2000	Europa	Humano
147206.17	*Collinsella stercoris*	Completo	not applicable		Humano
149.616.373	*Clostridioides difficile*	Completo	2020-11-30	Asia	Humano
1530123.47	*Acinetobacter seifertii*	Completo	2010-2017	Asia	Humano
154.046.362	*Hungatella hathewayi*	Completo	not applicable		Humano
1585976.7	*Bergeyella cardium*	Completo	2016-05-31	Asia	Humano
1.588.362.081	*Enterobacter hormaechei*	Completo	2020	Asia	Humano
161879.63	*Corynebacterium kroppenstedtii*	Completo	2021-10-05	Asia	Humano
1.639.116	*Listeria monocytogenes*	Completo	2008-07-01	Europa	Humano
1680.5	*Bifidobacterium adolescentis*	Completo	2006-10-30	Asia	Humano

1689.18	*Bifidobacterium dentium*	Completo	1900-1974		Humano
1.717.786	*Corynebacterium diphtheriae*	Completo	2017	Asia	Humano
17.471.081	*Cutibacterium acnes*	Completo	2016-01-05	Asia	Humano
177.323.679	*Mycobacterium tuberculosis*	Completo		Asia	Humano
1.774.124	*Mycobacteroides chelonae*	Completo	2019-10	Europa	Humano
1870984.37	*Anaerococcus mediterraneensis*	Completo	2019-05-10	Europa	Humano
187.327.398	*Acidaminococcus intestini*	Completo	not applicable		Humano
1.953.312	*Campylobacter coli*	Completo	2018	Asia	Humano
19.723.707	*Campylobacter jejuni*	Completo	2021	Europa	Humano
1.992.267	*Campylobacter concisus*	Completo	2009-01-02	Oceania	Humano
207340.31	*Roseomonas mucosa*	Completo	01-nov-2013	Norteamérica	Humano
208.224.339	*Enterobacter kobei*	Completo	2021	Asia	Humano
2098.77	*Metamycoplasma hominis*	Completo			Humano
21.010.419	*Helicobacter pylori*	Completo	2015	Asia	Humano
2.104.439	*Mycoplasmoides pneumoniae*	Completo	2016	Asia	Humano
2.168.162.621	*Bifidobacterium longum*	Completo	2020-03-01	Asia	Humano
239.935.255	*Akkermansia muciniphila*	Completo	2017-11-06	Asia	Humano
244.366.383	*Klebsiella variicola*	Completo	2002-10-23	Oceania	Humano
253.62	*Chryseobacterium indologenes*	Completo	2010-08	Norteamérica	Humano
262728.6	*Haemophilus influenzae*	Completo			Humano
2676062.3	*Erysipelotrichaceae bacterium 66202529*	Completo	2019-10-18	Europa	Humano
272621.13	*Lactobacillus acidophilus*	Completo	1970		Humano
2763670.3	*Wujia chipingensis*	Completo	2018-06-01	Asia	Humano
28.025.133	*Bifidobacterium animalis*	Completo	2010-10	Asia	Humano
28.111.543	*Bacteroides eggerthii*	Completo	not applicable		Humano
281.161.841	*Bacteroides ovatus*	Completo	2002-04-16	Norteamérica	Humano
28.141.828	*Cronobacter sakazakii*	Completo	2017-07-03	Europa	Humano
28188.25	*Capnocytophaga canimorsus*	Completo	2018-03-21	Oceania	Humano
2.872.558	*Pseudomonas aeruginosa*	Completo	2012		Humano
2.890.118.084	*Salmonella enterica*	Completo	2016-07-15	Asia	Humano
29.459.447	*Brucella melitensis*	Completo	2015-03-01	Asia	Humano
29.518.255	*Borreliella afzelii*	Completo	2001	Europa	Humano
29.519.483	*Borreliella garinii*	Completo	1996	Europa	Humano
32020.98	*Campylobacter fetus*	Completo	2012-05	Norteamérica	Humano
33.035.113	*Blautia producta*	Completo			Humano
35814.91	*Bordetella holmesii*	Completo	1984	Norteamérica	Humano
36.809.384	*Mycobacteroides abscessus*	Completo	2015		Humano
3.716.011.141	*Bacteroides xylanisolvens*	Completo	2007-07-11	Norteamérica	Humano
40.215.152	*Acinetobacter junii*	Completo	2018-04-17	Asia	Humano
40.324.208	*Stenotrophomonas maltophilia*	Completo	2019-12-04	Asia	Humano

405532.5	*Bacillus cereus*	Completo	1969		Humano
463025.4	*Bordetella bronchialis*	Completo	2009		Humano
47.018.285	*Acinetobacter baumannii*	Completo			Humano
48.518.128	*Neisseria gonorrhoeae*	Completo	2014	Europa	Humano
504.107	*Kingella kingae*	Completo	2018-08-13	Oceania	Humano
518.68	*Bordetella bronchiseptica*	Completo	1995	Norteamérica	Humano
519.5	*Bordetella parapertussis*	Completo	2012	Norteamérica	Humano
5.201.353	*Bordetella pertussis*	Completo	1983	Oceania	Humano
545.196	*Citrobacter koseri*	Completo	2019-10-10	Asia	Humano
5.461.553	*Citrobacter freundii*	Completo	2020-11-23	Asia	Humano
548.155	*Klebsiella aerogenes*	Completo	1997	Europa	Humano
548.659	*Klebsiella aerogenes*	Completo	2013	Asia	Humano
5.503.005	*Enterobacter cloacae*	Completo	2014	Europa	Humano
562.110.757	*Escherichia coli*	Completo	2019-05-22	Norteamérica	Humano
571.73	*Klebsiella oxytoca*	Completo	2018	Europa	Humano
5.733.984	*Klebsiella pneumoniae*	Completo	2016	Europa	Humano
57.706.158	*Citrobacter braakii*	Completo	2019-05-23	Asia	Humano
582.135	*Morganella morganii*	Completo	2015-07-31	Asia	Humano
584.296	*Proteus mirabilis*	Completo	2014-05-14	Asia	Humano
587.328	*Providencia rettgeri*	Completo	2013-10	Sudamérica	Humano
6.152.494	*Serratia marcescens*	Completo	2016	Europa	Humano
61.645.623	*Enterobacter asburiae*	Completo	2020	Africa	Humano
65.058.243	*Corynebacterium ulcerans*	Completo	2016	Asia	Humano
654.725	*Aeromonas veronii*	Completo	2019	Asia	Humano
68.892.148	*Streptococcus infantis*	Completo	2018-03-11	Asia	Humano
729.339	*Haemophilus parainfluenzae*	Completo			Humano
730.14	*[Haemophilus] ducreyi*	Completo	1982		Humano
730.31	*[Haemophilus] ducreyi*	Completo	2014	Africa	Humano
76859.5	*Fusobacterium nucleatum*	Completo	2018	Asia	Humano
777.319	*Coxiella burnetii*	Completo	2014	Europa	Humano
783.32	*Rickettsia rickettsii*	Completo	1944	Norteamérica	Humano
803.16	*Bartonella quintana*	Completo	2020-05-01	Norteamérica	Humano
813.33	*Chlamydia trachomatis*	Completo	1994-09-29	Norteamérica	Humano
8.172.206	*Bacteroides fragilis*	Completo	1985-09-25	Norteamérica	Humano
824.5	*Campylobacter gracilis*	Completo	1977		Humano
83558.16	*Chlamydia pneumoniae*	Completo		Europa	Humano
84.112.435	*Eggerthella lenta*	Completo	2020-12-08	Europa	Humano
857417.11	*Janibacter indicus*	Completo	2017-09-30	Asia	Humano
87.883.732	*Burkholderia multivorans*	Completo	2010	Europa	Humano
95.486.299	*Burkholderia cenocepacia*	Completo	1996-09-10	Norteamérica	Humano

1196182.3	*Listeria monocytogenes*	Completo	2008	Norteamérica	Ambiental
12.804.443	*Staphylococcus aureus*	Completo	2019-06-01	Europa	Ambiental
13.517.158	*Enterococcus faecalis*	Completo	2018-10	Asia	Ambiental
155.322.309	*Bacillus toyonensis*	Completo	2019	Norteamérica	Ambiental
1667327.38	*Klebsiella quasipneumoniae*	Completo			Ambiental
1.812.935.393	*Enterobacter roggenkampii*	Completo	2021-07	Asia	Ambiental
2489014.3	*Klebsiella variicola*	Completo			Ambiental
388357.3	*Kocuria turfanensis*	Completo	2006-08-20	Asia	Ambiental
41170.17	*Exiguobacterium acetylicum*	Completo	2019	Norteamérica	Ambiental
41276.2	*Brevundimonas vesicularis*	Completo	2012-09	Asia	Ambiental
56.277.543	*Escherichia coli*	Completo	2001	Oceania	Ambiental
57.341.322	*Klebsiella pneumoniae*	Completo	2000		Ambiental

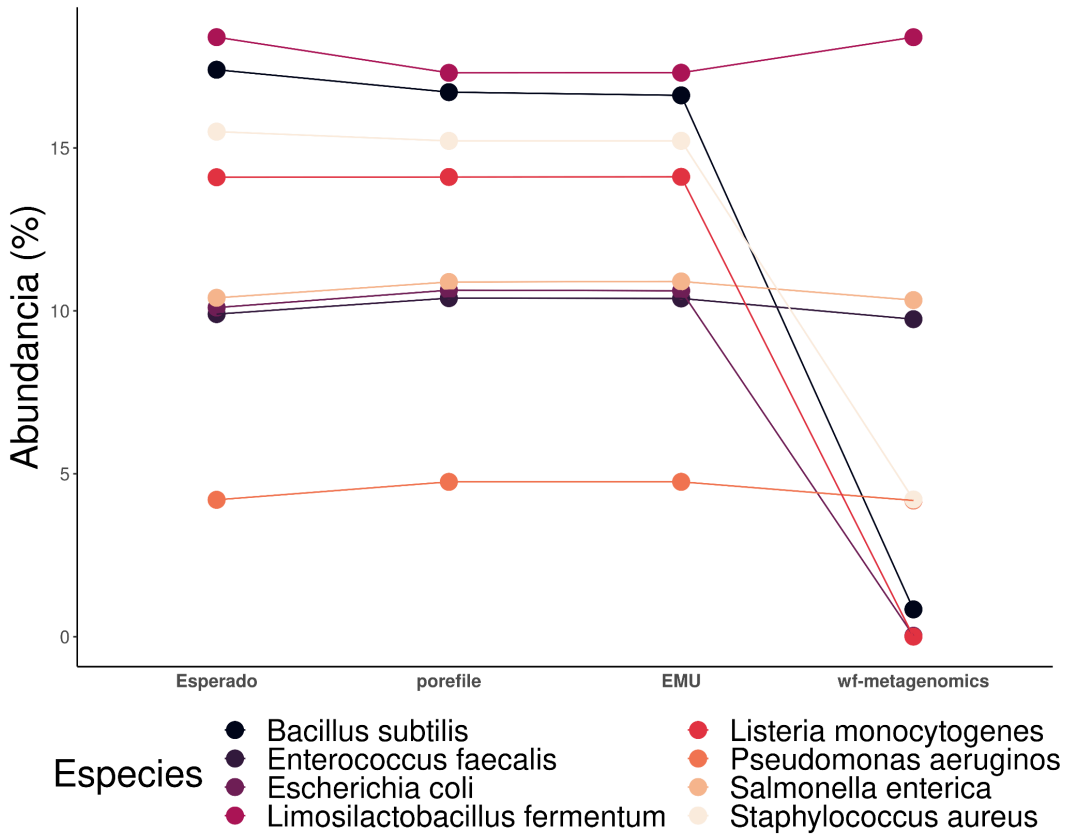

Figura S1: Detección de componentes de CBC y abundancia relativa obtenida para cada una de las herramientas de clasificación

Tabla S2: Datos de secuenciación ONT e Illumina de Matsuo et al., 2021

SRA	Muestra	Tipo	ID	Datos
DRR225048	gut microbiome 1	Sample	G1	ONT
DRR225051	gut microbiome 2	Sample	G2	ONT
DRR225054	gut microbiome 3	Sample	G3	ONT
DRR225057	gut microbiome 4	Sample	G4	ONT
DRR225060	gut microbiome 5	Sample	G5	ONT
DRR225063	gut microbiome 6	Sample	G6	ONT
DRR225046	10 Strain Even Mix Genomic Material	Mock	M2	ONT
DRR225050	gut microbiome 1	Sample	G1	Illumina
DRR225053	gut microbiome 2	Sample	G2	Illumina
DRR225056	gut microbiome 3	Sample	G3	Illumina
DRR225059	gut microbiome 4	Sample	G4	Illumina
DRR225062	gut microbiome 5	Sample	G5	Illumina
DRR225065	gut microbiome 6	Sample	G6	Illumina

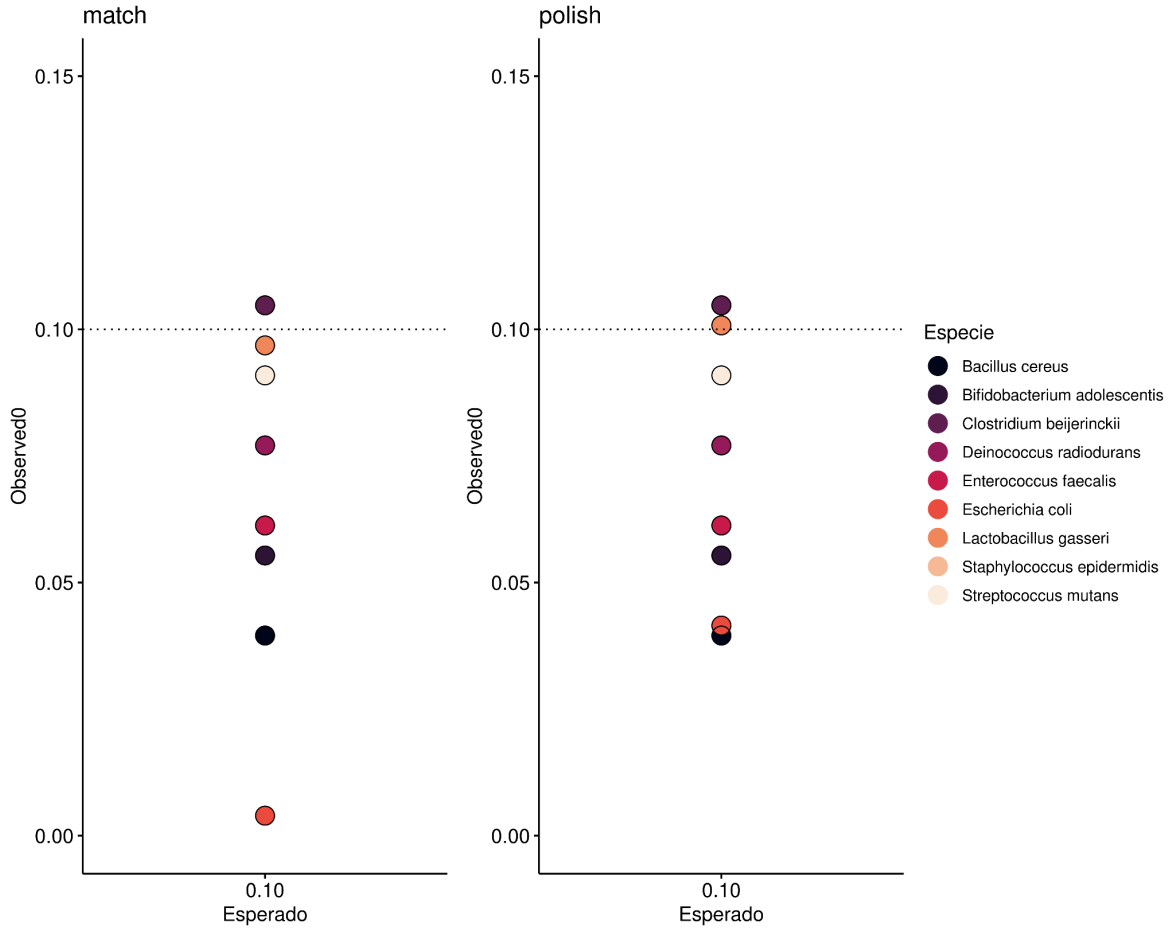

Figura S2: Comparación de la recuperación relativa de la comunidad estándar de prueba MSA-1000 (ATCC) del dataset de Matsuo et al, 2021.

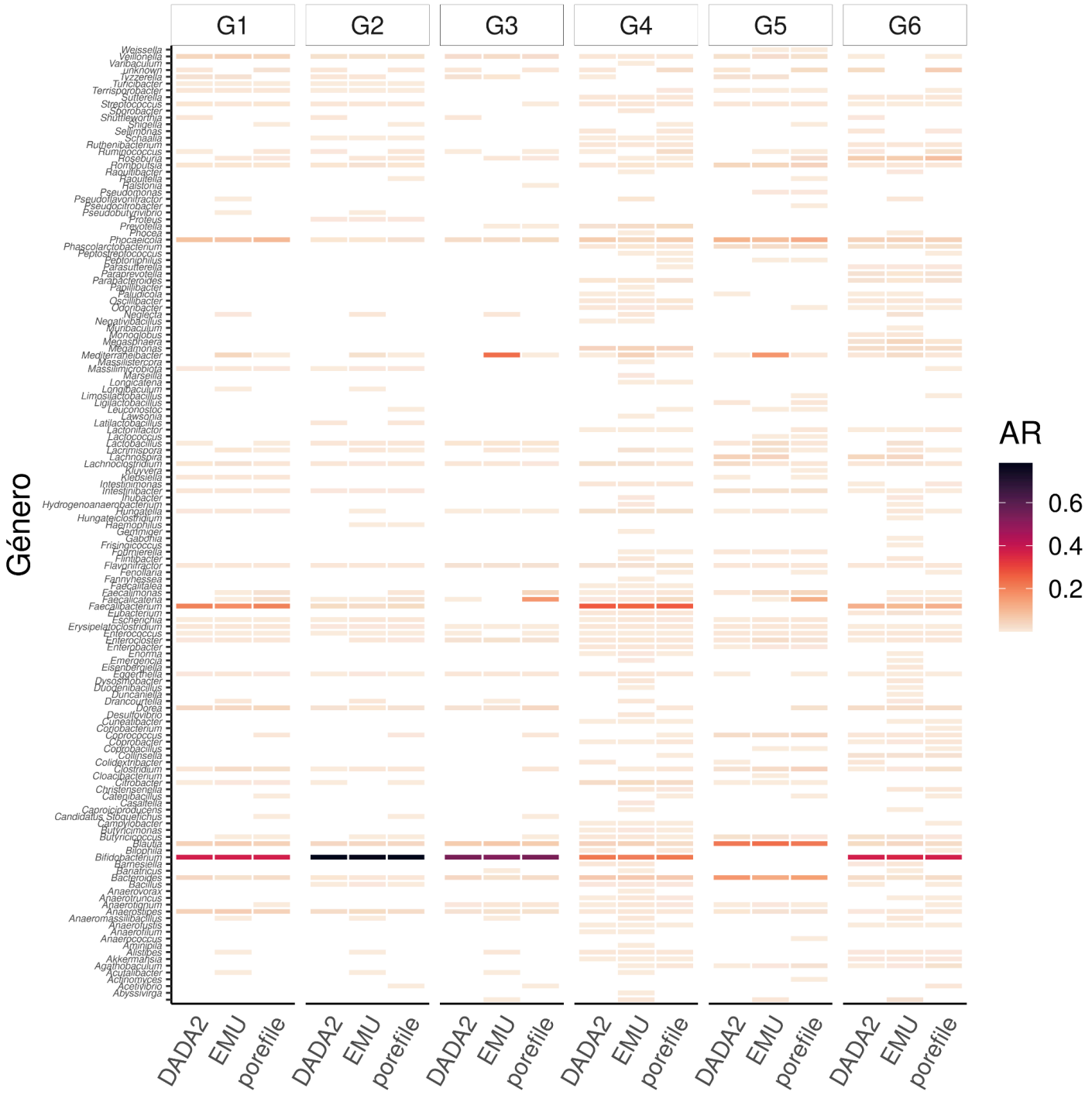

Figura S3: Comparación de las clasificaciones obtenidas a nivel de género para las muestras de la microbiota fecal humana (G1-G6) con datos de secuenciación del gen 16S obtenidos con la plataforma ONT (V1-V9) y la plataforma Illumina (V3-V4). Se utilizaron las herramientas *porefile* y *EMU* para la clasificación de los datos ONT y *DADA2* para los datos Illumina.

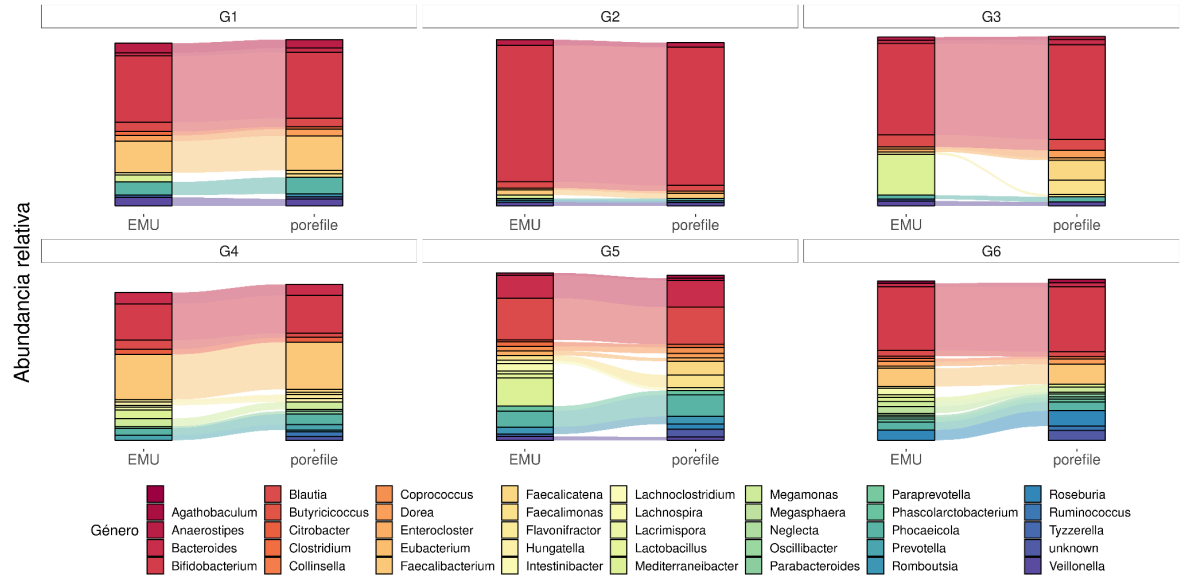

Figura S4: Clasificación a nivel de género con *porefile* y *EMU* de las muestras de microbioma intestinal humano obtenidos de Matsuó et al. 2021. Se muestran los géneros con una abundancia relativa mayor a 0.01.

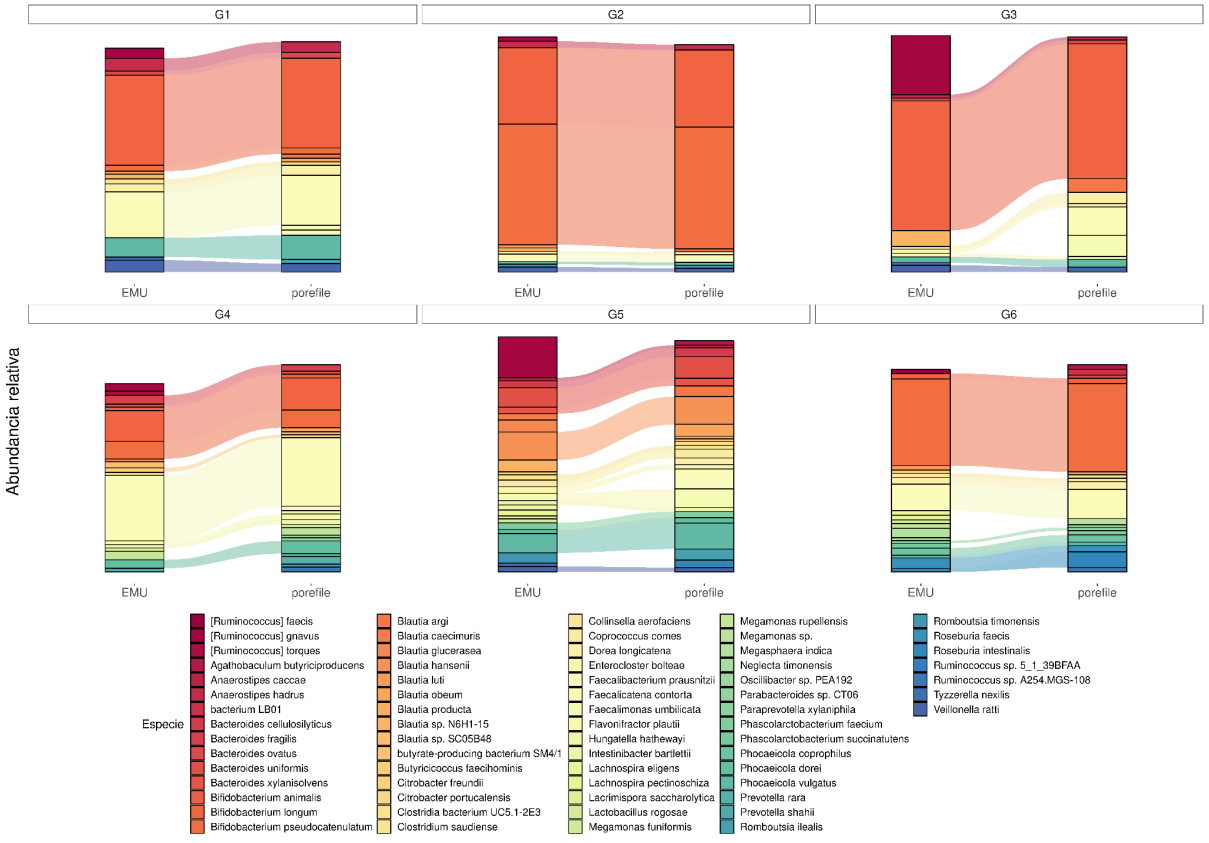

Figura S5: Clasificación a nivel de especies con *porefile* y *EMU* de las muestras de microbioma intestinal humano obtenidos de Matsuo et al. 2021. Se muestran las especies con una abundancia relativa mayor a 0.01.

7. Referencias

1. Ramirez KS, Leff JW, Barberán A, Bates ST, Betley J, Crowther TW, et al. Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. Proc R Soc B Biol Sci. 22 de noviembre de 2014;281(1795):20141988.
2. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 23 de noviembre de 2017;551(7681):457-63.
3. Segota I, Watrous JD, Kantz ED, Nallamshetty S, Tiwari S, Cheng S, et al. Reconstructing the landscape of gut microbial species across 29,000 diverse individuals. Nucleic Acids Res. 22 de mayo de 2023;51(9):4178-90.
4. F. Escapa I, Huang Y, Chen T, Lin M, Kokaras A, Dewhirst FE, et al. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. Microbiome. diciembre de 2020;8(1):65.
5. Kuczynski J, Stombaugh J, Walters WA, González A, Caporaso JG, Knight R. Using QIIME to Analyze 16S rRNA Gene Sequences from Microbial Communities. Curr Protoc Microbiol [Internet]. noviembre de 2012 [citado 7 de junio de 2023];27(1). Disponible en: https://onlinelibrary.wiley.com/doi/10.1002/9780471729259.mc01e05s27
6. Chiarello M, McCauley M, Villéger S, Jackson CR. Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold. Moreno-Hagelsieb G, editor. PLOS ONE. 24 de febrero de 2022;17(2):e0264443.
7. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. julio de 2016;13(7):581-3.
8. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 6 de noviembre de 2019;10(1):5029.
9. Benítez-Páez A, Portune KJ, Sanz Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. GigaScience. diciembre de 2016;5(1):4.
10. Nygaard AB, Tunsjø HS, Meisal R, Charnock C. A preliminary study on the potential of Nanopore MinION and Illumina MiSeq 16S rRNA gene sequencing to characterize building-dust microbiomes. Sci Rep. 21 de febrero de 2020;10(1):3209.
11. Kai S, Matsuo Y, Nakagawa S, Kryukov K, Matsukawa S, Tanaka H, et al. Rapid bacterial identification by direct PCR amplification of 16S rRNA genes using the MinION™ nanopore sequencer. FEBS Open Bio. marzo de 2019;9(3):548-57.

12. Rozas M, Brillet F, Callewaert C, Paetzold B. MinION™ Nanopore Sequencing of Skin Microbiome 16S and 16S-23S rRNA Gene Amplicons. Front Cell Infect Microbiol. 5 de enero de 2022;11:806476.
13. Matsuo Y, Komiya S, Yasumizu Y, Yasuoka Y, Mizushima K, Takagi T, et al. Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION™ nanopore sequencing confers species-level resolution. BMC Microbiol. diciembre de 2021;21(1):35.
14. Urban L, Holzer A, Baronas JJ, Hall MB, Braeuninger-Weimer P, Scherm MJ, et al. Freshwater monitoring by nanopore sequencing. eLife. 19 de enero de 2021;10:e61504.
15. Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the rrn operon. F1000Research. 1 de agosto de 2019;7:1755.
16. Santos A, Van Aerle R, Barrientos L, Martinez-Urtaza J. Computational methods for 16S metabarcoding studies using Nanopore sequencing data. Comput Struct Biotechnol J. 2020;18:296-305.
17. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 28 de noviembre de 2019;20(1):257.
18. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci. 2 de enero de 2017;3:e104.
19. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics. 15 de septiembre de 2018;34(18):3094-100.
20. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 1 de enero de 2012;40(D1):D136-43.
21. Curry KD, Wang Q, Nute MG, Tyshaieva A, Reeves E, Soriano S, et al. Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. Nat Methods. julio de 2022;19(7):845-53.
22. Rodríguez-Pérez H, Ciuffreda L, Flores C. NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data. Inanc B, editor. Bioinformatics. 12 de julio de 2021;37(11):1600-1.
23. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 14 de noviembre de 2007;35(21):7188-96.
24. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 27 de noviembre de 2012;41(D1):D590-6.
25. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. marzo de 2007;17(3):377-86.
26. Wick R. Badread: simulation of error-prone long reads. J Open Source Softw. 4 de abril de 2019;4(36):1316.

27. Wouter De Coster. Nanofilt [Internet]. Disponible en: https://github.com/wdecoster/nanofilt
28. Trimmomatic [Internet]. Disponible en: https://github.com/usadellab/Trimmomatic
29. Simon Andrews. FastQC [Internet]. Disponible en: https://github.com/s-andrews/FastQC
30. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. Watson M, editor. PLoS ONE. 22 de abril de 2013;8(4):e61217.
31. Hadley Wickham, RStudio. stringr [Internet]. 2022. Disponible en: https://mirror.linux.duke.edu/cran/web/packages/stringr/stringr.pdf
32. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 21 de noviembre de 2019;4(43):1686.
33. Ahlmann-Eltze C, Patil I. ggsignif: R Package for Displaying Significance Brackets for «ggplot2» [Internet]. PsyArXiv; 2021 mar [citado 1 de marzo de 2023]. Disponible en: https://osf.io/7awm6
34. Hamner B, Michael Frasco. Metrics: Evaluation Metrics for Machine Learning. R package version 0.1.4. [Internet]. 2018. Disponible en: https://CRAN.R-project.org/package=Metrics
35. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. abril de 2017;35(4):316-9.
36. Huson DH, Albrecht B, Bağcı C, Bessarab I, Górska A, Jolic D, et al. MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. Biol Direct. enero de 2018;13(1):6.
37. Usyk M, Peters BA, Karthikeyan S, McDonald D, Sollecito CC, Vazquez-Baeza Y, et al. Comprehensive evaluation of shotgun metagenomics, amplicon sequencing, and harmonization of these platforms for epidemiological studies. Cell Rep Methods. enero de 2023;3(1):100391.
38. Vaginal Microbiome Consortium (additional members), Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol. diciembre de 2015;15(1):66.
39. Nearing JT, Comeau AM, Langille MGI. Identifying biases and their potential solutions in human microbiome studies. Microbiome. 18 de mayo de 2021;9(1):113.
40. McLaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic sequencing experiments. eLife. 10 de septiembre de 2019;8:e46923.
41. McInerney P, Adams P, Hadi MZ. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. Mol Biol Int. 17 de agosto de 2014;2014:1-8.
42. Seol D, Lim JS, Sung S, Lee YH, Jeong M, Cho S, et al. Microbial Identification Using rRNA Operon Region: Database and Tool for Metataxonomics with Long-Read Sequence. Xu ZZ, editor. Microbiol Spectr. 27 de abril de 2022;10(2):e02017-21.
43. de Oliveira Martins L, Page AJ, Mather AE, Charles IG. Taxonomic resolution of the ribosomal RNA operon in bacteria: implications for its use with long-read sequencing. NAR Genomics Bioinforma. 1 de marzo de 2020;2(1):lqz016.

44. Kinoshita Y, Niwa H, Uchida-Fujii E, Nukada T. Establishment and assessment of an amplicon sequencing method targeting the 16S-ITS-23S rRNA operon for analysis of the equine gut microbiome. Sci Rep. 4 de junio de 2021;11(1):11884.
45. Ritari J, Salojärvi J, Lahti L, de Vos WM. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. BMC Genomics. diciembre de 2015;16(1):1056.
46. Myer PR, McDaneld TG, Kuehn LA, Dedonder KD, Apley MD, Capik SF, et al. Classification of 16S rRNA reads is improved using a niche-specific database constructed by near-full length sequencing. Schierwater B, editor. PLOS ONE. 13 de julio de 2020;15(7):e0235498.
47. Yin P, Zhang C, Du T, Yi S, Yu L, Tian F, et al. Meta-analysis reveals different functional characteristics of human gut Bifidobacteria associated with habitual diet. Food Res Int. agosto de 2023;170:112981.
48. Drey E, Kok CR, Hutkins R. Role of Bifidobacterium pseudocatenulatum in Degradation and Consumption of Xylan-Derived Carbohydrates. Nikel PI, editor. Appl Environ Microbiol. 26 de octubre de 2022;88(20):e01299-22.

PARTE II:

Caracterización de variantes de SARS-CoV-2 y generación de herramientas alternativas para la vigilancia epidemiológica

Capítulo III

Introducción de SARS-CoV-2 en Uruguay

1. Introducción

Una infección respiratoria severa de origen desconocido fue detectada a finales del 2019 en la provincia de Hubei, China. Los métodos de diagnóstico apuntaban a una infección viral causada por un coronavirus. Tras el análisis metagenómico de muestras obtenidas de pacientes hospitalizados se determinó que el agente etiológico correspondía a nuevo Betacoronavirus (1–3) denominado posteriormente *Severe Acute Respiratory Syndrome Coronavirus 2* (SARS-CoV-2) por el ICTV (*International Committee on Taxonomy of Viruses*) en base a sus características filogenéticas (4). La enfermedad causada por SARS-CoV-2 se denominó COVID-19 (por *Coronavirus disease 2019*) por parte de la Organización Mundial de la Salud (OMS) (5). La misma se dispersó rápidamente a distintas regiones del mundo como consecuencia de la conectividad y movilidad internacional y el 11 de marzo de 2020 es declarada la pandemia global de COVID-19 por parte de la OMS (6). Los esfuerzos de vigilancia genómica llevados para SARS-CoV-2 a nivel global y la rápida disponibilización de la información ha permitido el análisis y rastreo epidemiológico, así como el monitoreo de la evolución viral a escala global prácticamente en tiempo real como ningún otro patógeno (7).

La OMS y el ECDC (por *European Centre for Disease Prevention and Control*) han generado guías técnicas para el muestreo y secuenciación de SARS-CoV-2 (8,9). La obtención de secuencias genómicas del virus permite llevar a cabo estudios filodinámicos y filogenéticos de SARS-CoV-2, que permiten comprender la transmisión, epidemiología y dispersión espacial de SARS-CoV-2, así como el monitoreo de nuevas mutaciones que podrían conferir ventajas biológicas al virus tales como mayor transmisibilidad, capacidad de evasión de la respuesta inmune natural o inducida o posibles escapes a las terapias disponibles (7). Para la generación de las secuencias genómicas de SARS-CoV-2, se han utilizado al menos dos metodologías; la primera consistió en un abordaje metagenómico a partir de muestras clínicas obtenidas de pacientes enfermos. Estas metodologías son útiles cuando no se tiene información previa del agente causal y resulta extremadamente útil para la detección de nuevos patógenos, tal como fue el caso de SARS-CoV-2. La segunda estrategia y tal vez la más utilizada, es la secuenciación direccionada basada en la amplificación de segmentos solapantes del genoma y reconstrucción del mismo guiado por una referencia. Esta estrategia ha sido implementada para múltiples plataformas de secuenciación (10–12). La rápida disponibilización de las primeras secuencias de SARS-CoV-2 en etapas tempranas de la emergencia sanitaria permitió el desarrollo de métodos diagnósticos específicos y además

permitió la implementación de protocolos de secuenciación, en base a protocolos previos (13,14). En este sentido, en enero del 2020 ARTIC Network puso a disposición un esquema de cebadores para la amplificación de SARS-CoV-2 en conjunto con protocolos de secuenciación de fragmentos solapantes del genoma y análisis para la plataforma ONT (https://artic.network/ncov-2019). Con este esquema el genoma de SARS-CoV-2 es amplificado a partir del producto de retrotranscripción en fragmentos de aproximadamente 400 pares de bases utilizando el método Primal Scheme (14). A partir de esta estrategia, otros protocolos han surgido para la generación de amplicones solapantes del genoma viral y preparación de bibliotecas de secuenciación en un tiempo más corto (15). Se estima que la mayor parte de las secuencias genómicas de SARS-CoV-2 generadas se han obtenido con la plataforma Illumina y aproximadamente el 25% con la plataforma ONT (16).

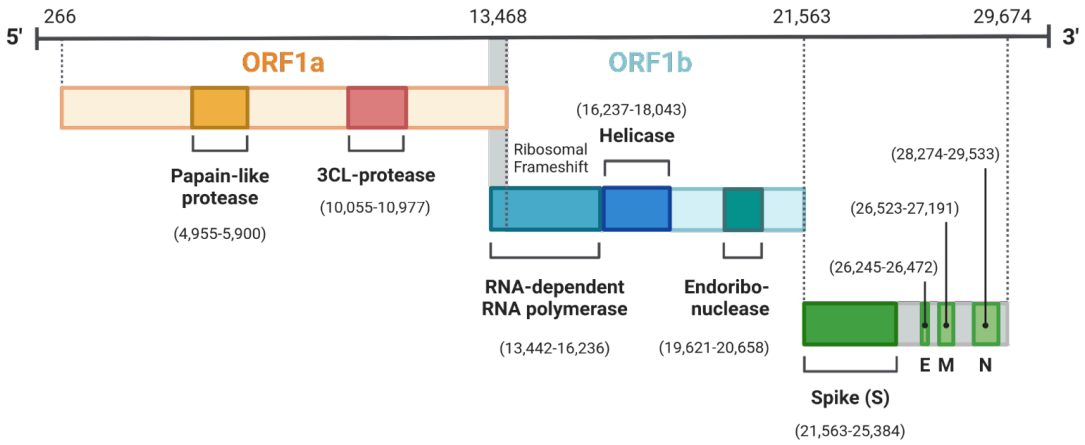

Figura 1: Organización genómica del SARS-CoV-2. Obtenido de Biorender.

A partir de las primeras secuencias obtenidas del genoma de SARS-CoV-2 se determinó que el mismo pertenece al género Betacoronavirus, los cuales representan virus envueltos con un genoma ARN de cadena simple y sentido positivo, con un tamaño de aproximadamente 30 kilobases. El mismo codifica 16 proteínas no estructurales (nsp), 4 estructurales y 6 accesorias. Las 16 nsps involucradas en la transcripción viral, replicación y evasión del sistema inmune del hospedero son productos del clivado de una poliproteína codificada por el ORf1a y Orf1b que en conjunto ocupan aproximadamente el 70% del genoma viral (17). El 30% restante del genoma codifica para proteínas accesorias y proteínas estructurales correspondientes a la envoltura (E), membrana (M), nucleocápside (N) y espícula o *Spike* (S) (**Figura 1**). Esta última corresponde a una glicoproteína que interactúa

con el receptor ACE2 (por *angiotensin convertase enzyme* 2) mediando la fusión del virus con las células del hospedero. Asimismo, en la proteína S se encuentra el sitio RBD (por *receptor binding domain*) el cual es el principal blanco de los anticuerpos neutralizantes generados luego de la infección natural o tras la vacunación (18).

Los primeros casos de COVID-19 detectados en Sudamérica fueron reportados en Brasil, mientras que Venezuela y Uruguay fueron los últimos países en reportar los primeros casos confirmados (19). A partir de esto, Uruguay procede a cerrar parcialmente las fronteras permitiendo el retorno de ciudadanos uruguayos y a su vez declarando cuarentena voluntaria (20) (**Figura 2**). En este contexto, se realizó la caracterización genómica a nivel local de diez muestras de individuos con diagnóstico positivo para COVID-19 entre el 16 y 19 de marzo de 2020 con protocolos de secuenciación generados en las etapas tempranas de la pandemia con la finalidad de conocer los genotipos, así como estimar la fecha y orígen geográfico más probable del ingreso de SARS-CoV-2 en Uruguay.

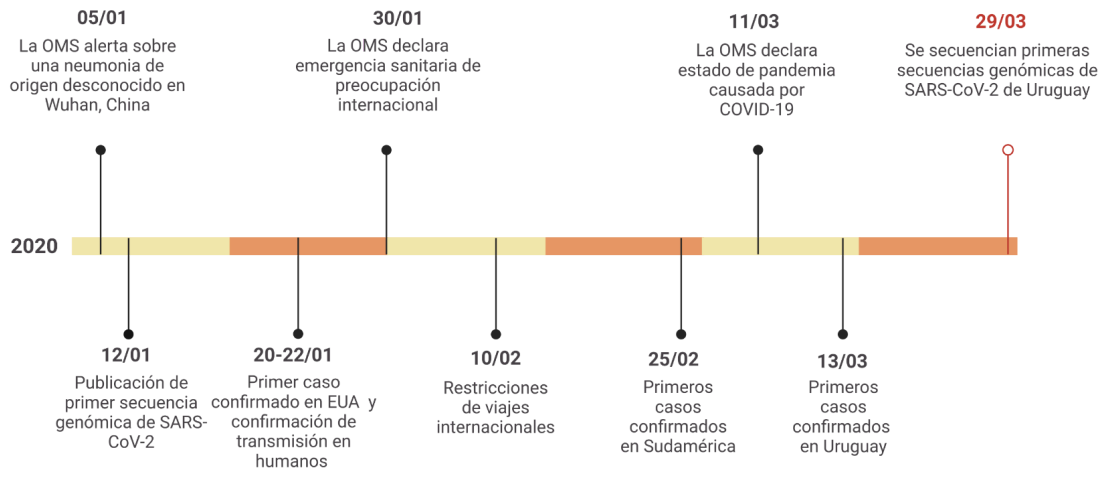

Figura 2: Línea de tiempo de SARS-CoV-2 del comienzo de la emergencia sanitaria por COVID-19 en Uruguay. Creado en Biorender.

2 . Objetivo general

Describir los linajes de SARS-CoV-2 que circularon en marzo del 2020 en Uruguay luego de la confirmación de los primeros casos de COVID-19 en el país.

2.1 Objetivos específicos

1. Determinar la fecha y orígen más probable de la introducción de SARS-CoV-2 en Uruguay mediante métodos filodinámicos.
2. Determinar las características principales de los linajes de SARS-CoV-2 detectados en marzo del 2020 de los genomas depositados en la base de datos EpiCoV/GISAID.

3. Métodos

3.1 Estudios filodinámicos de SARS-CoV-2 en Uruguay en marzo del 2020.

A modo de estimar la fecha, el origen geográfico y la variación genética de las primeras secuencias generadas de SARS-CoV-2 en Uruguay, se utilizó un conjunto de secuencias de la base de datos EpiCoV/GISAID correspondiente al *Nextstrain build South America*, 2021-03-03 obtenido de Sant'Anna et al, 2021 (21) y conteniendo un total de 3642 secuencias muestreadas entre 2019-12-26 y 2021-05-13. Las secuencias fueron divididas en nueve regiones que incluyen Uruguay (n = 57), países limítrofes a Uruguay (Argentina n = 127 y Brasil n = 988), del resto de Sudamérica (n = 899), Norteamérica (n = 167), Europa (n = 712), Asia (n = 413), África (n = 232) y Oceanía (n = 48). Del total de secuencias de Uruguay (EPI_SET_230830yk), 38 fueron añadidas posteriormente al conjunto de datos, donde 10 fueron generadas por el grupo de trabajo del Institut Pasteur de Montevideo (22). Las secuencias uruguayas corresponden a muestras obtenidas de EpiCoV/GISAID colectadas hasta el 2020-03-30 y forman parte del foco de estudio de la dinámica de SARS-CoV-2 en ese período (**Tabla S1**). Para todas las muestras el clado Nextrain fue asignado utilizando *Nextclade CLI v1.10.3* (23). Para estimar la fecha de ingreso y orígen geográfico de las primeras secuencias de SARS-CoV-2 en Uruguay se utilizó un modelo de reloj molecular (24). Es decir, se asume que los cambios observados a lo largo del tiempo se dieron a una tasa evolutiva relativamente constante. El conjunto de datos completo de secuencias fue alineado con *Nextalign CLI v1.4.5* (23) utilizando como referencia la secuencia WIV04 (EPI_ISL_402124). Múltiples iteraciones de generación de árboles filogenéticos de máxima verosimilitud fueron inferidos con *IQ-TREE v2.1.4* utilizando el modelo de sustitución GTR+G (25). A modo de evaluar incongruencias entre la divergencia genética y fecha de muestreo que afectan la señal temporal se utilizaron los métodos de evaluación *root-to-tip* implementados en *TreeTime v0.9.0* (26) y *TempEst v1.5.3* (27). Para la obtención de un mejor

ajuste la topología inicial del árbol filogenético para las estimaciones espacio-temporales, la filogenia final fue inferida utilizando ModelFinder implementada en *IQ-TREE* (28). La estimación de la filogenia a escala temporal se realizó con *TreeTime* utilizando el método de enraizado a partir de la secuencia más antigua (*–reroot oldest*). Se estimó la filogenia a escala temporal utilizando la configuración con máxima verosimilitud (*joint)*. Las transiciones geográficas entre las distintas regiones fueron estimadas con el modelo de transiciones de estados discretos o *mugration* de *TreeTime*, donde las posibles regiones ancestrales fueron especificadas como Uruguay, Argentina, Brasil, Sudamérica, Norteamérica, Europa, Asia, África y Oceanía.

3.2 Análisis de datos y disponibilidad del código

La manipulación de tablas se realizó con múltiples paquetes de R (29–32). Los alineamientos fueron visualizados con *Aliview v1.28* (33) y las filogenias temporales con *FigTree v1.4.4* (https://github.com/rambaut/figtree/). Los comandos ejecutados y el código para la generación de las figuras se encuentran disponibles en https://github.com/Ceci07/sars-cov-2_genomics.

4. Resultados

4.1 Linajes PANGO y clado Nextrain circulantes en Uruguay en marzo del 2020

En la **Figura 3A** se muestra la distribución de linajes y clados correspondientes a las secuencias genómicas de Uruguay obtenidas en marzo del 2020 disponibles en la base de datos EpiCoV/GISAID. La mayor parte de las secuencias corresponde al clado 19B (n = 25), seguido de 19A (n = 3), 20A (n = 4), 20D (n = 3) y 20B (n = 1). En la **Figura 3B** se muestra la distribución de las sustituciones aminoacídicas detectadas en estos genomas a lo largo del mes de marzo. La primera sustitución detectada a nivel del gen S, corresponde a la sustitución D614G. La secuencia más antigua conteniendo la sustitución D614G en la proteína S data del 17 de marzo del 2020 (en negro) (**Tabla S1**).

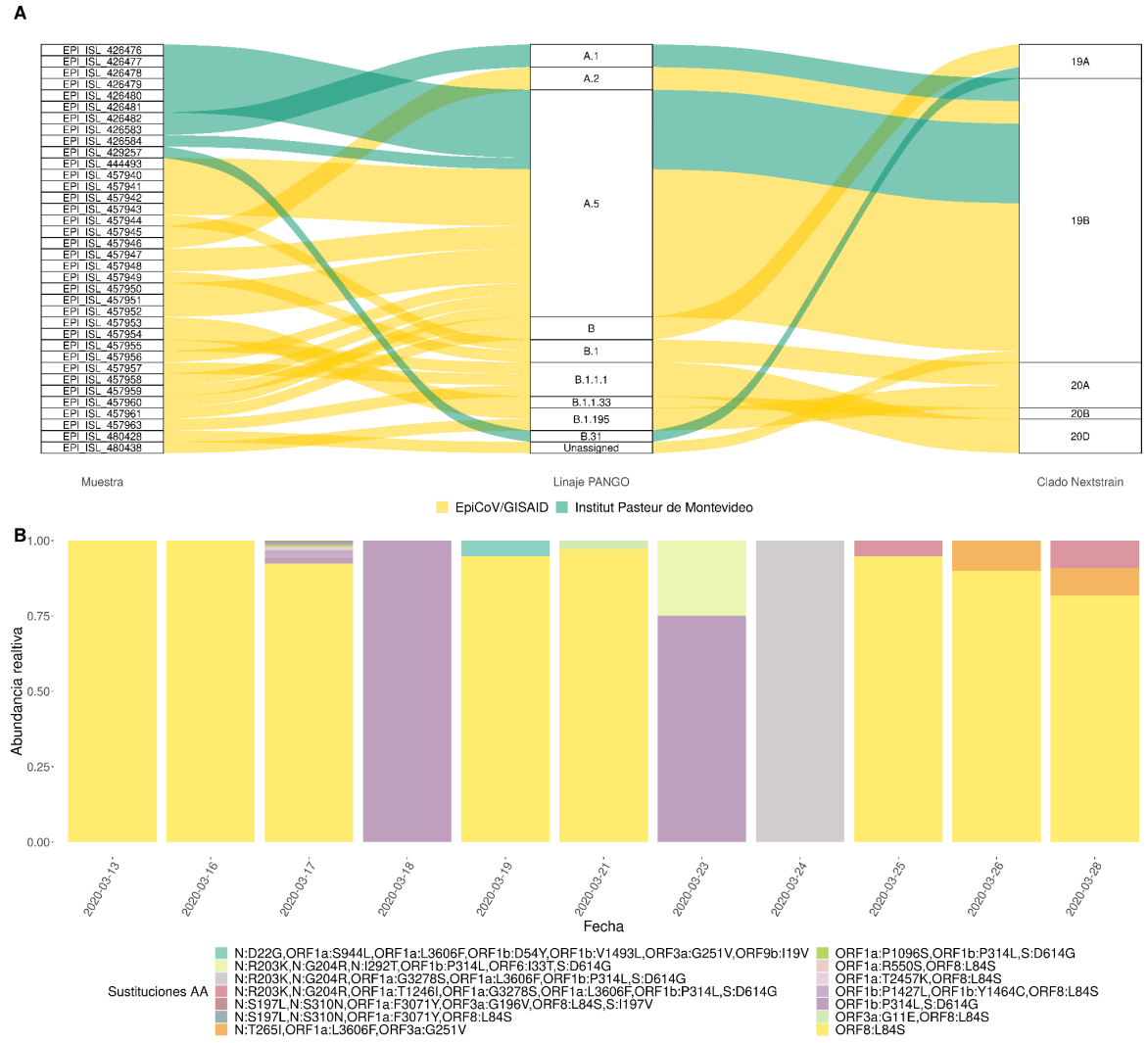

Figura 3: Distribución de las secuencias obtenidas de Uruguay en la base de datos EpiCoV/GISAID de marzo del 2020. A) Distribución de linajes PANGO y clado Nextstrain. B) Detección de sustituciones aminoacídicas en los genomas de la base de datos.

4.2 Generación de la filogenia de máxima verosimilitud del conjunto de datos genómicos.

A modo de estimar la fecha de ingreso y el/los orígen/es geográfico/s más probable del ingreso de SARS-CoV-2 en Uruguay se se utilizó un conjunto de datos previamente muestreado de la base de datos añadiendo todas las secuencias obtenidas de Uruguay (37 secuencias) de marzo del 2020. El conjunto de datos inicial fue sometido a varias iteraciones de inferencia filogenética de máxima verosimilitud y evaluación de la señal temporal con el método *root-to-tip* de *TreeTime* y *TempEst* (27). El TMRCA (por *time to the most recent common ancestor*) de la topología inicial del árbol filogenético se situó a fines de octubre de

2019 (2019.8 o 2019-10-20). Este conjunto final de datos incluye 3365 secuencias con una señal temporal aceptable para la inferencia de una filogenia a escala temporal (**Figura S1**). La tasa de mutación estimada para el conjunto de datos analizados fue de 6.878 x 10^{-04} sustituciones/sitio/año en base a la filogenia en escala temporal obtenida con el método *joint*.

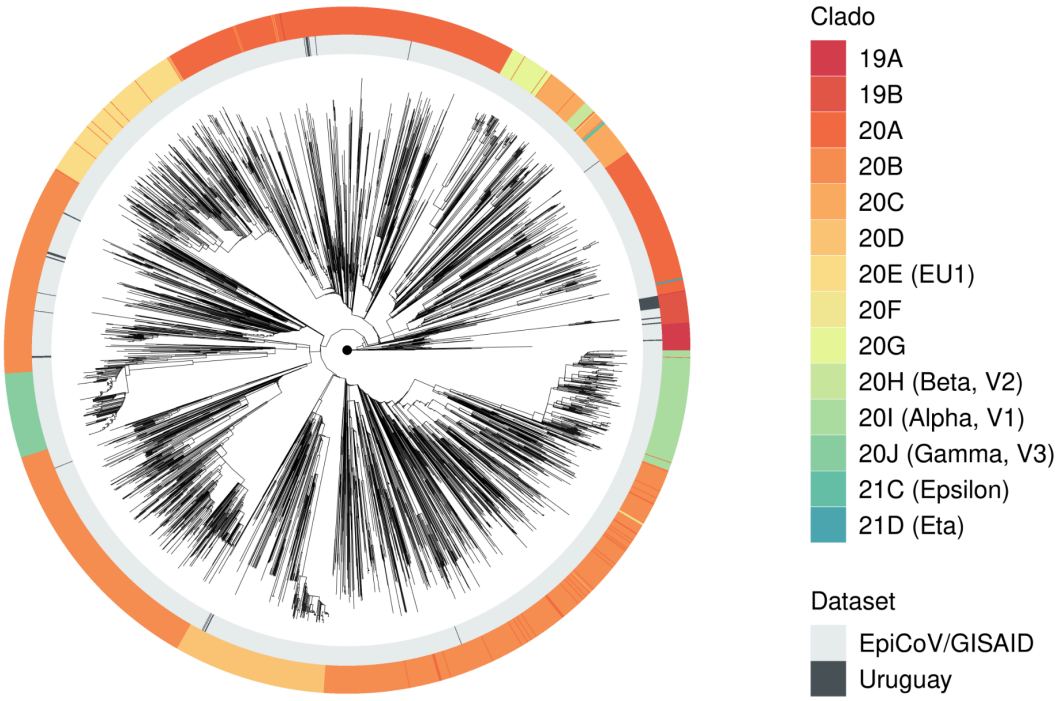

Figura 4: Árbol filogenético a escala temporal del *dataset* de estudio. El árbol filogenético contiene 3365 secuencias cuyo TMRCA observado (2019.9, $R^2 = 0.58$) y la tasa de mutación se corresponde a 6.878 x 10^{-04} sustituciones/sitio/año.

4.3 Estimación temporal y geográfica de la introducción de SARS-CoV-2 en Uruguay

La relación entre la divergencia genética y la fecha de muestreo indican un comportamiento de reloj molecular por parte del conjunto de datos **(Figura S1**). A partir de la calibración de la filogenia a escala temporal se obtuvo que TMRCA del árbol se ubica a mediados de noviembre del 2019 (2019-11-12). Se detectaron al menos nueve introducciones independientes durante el mes de marzo del 2020 (**Tabla S2 y Figura S2**). El cluster 4 que dió lugar al mayor cluster de transmisión detectado en marzo (n = 21). Se estima que esta introducción tuvo tuvo lugar hacia fines de febrero con un orígen probable en Sudamérica (excepto Brasil y Argentina). Este incluye la secuencia de SARS-CoV-2 de Uruguay con fecha de recolección más antigua (2020-03-13) (**Figura 5**).

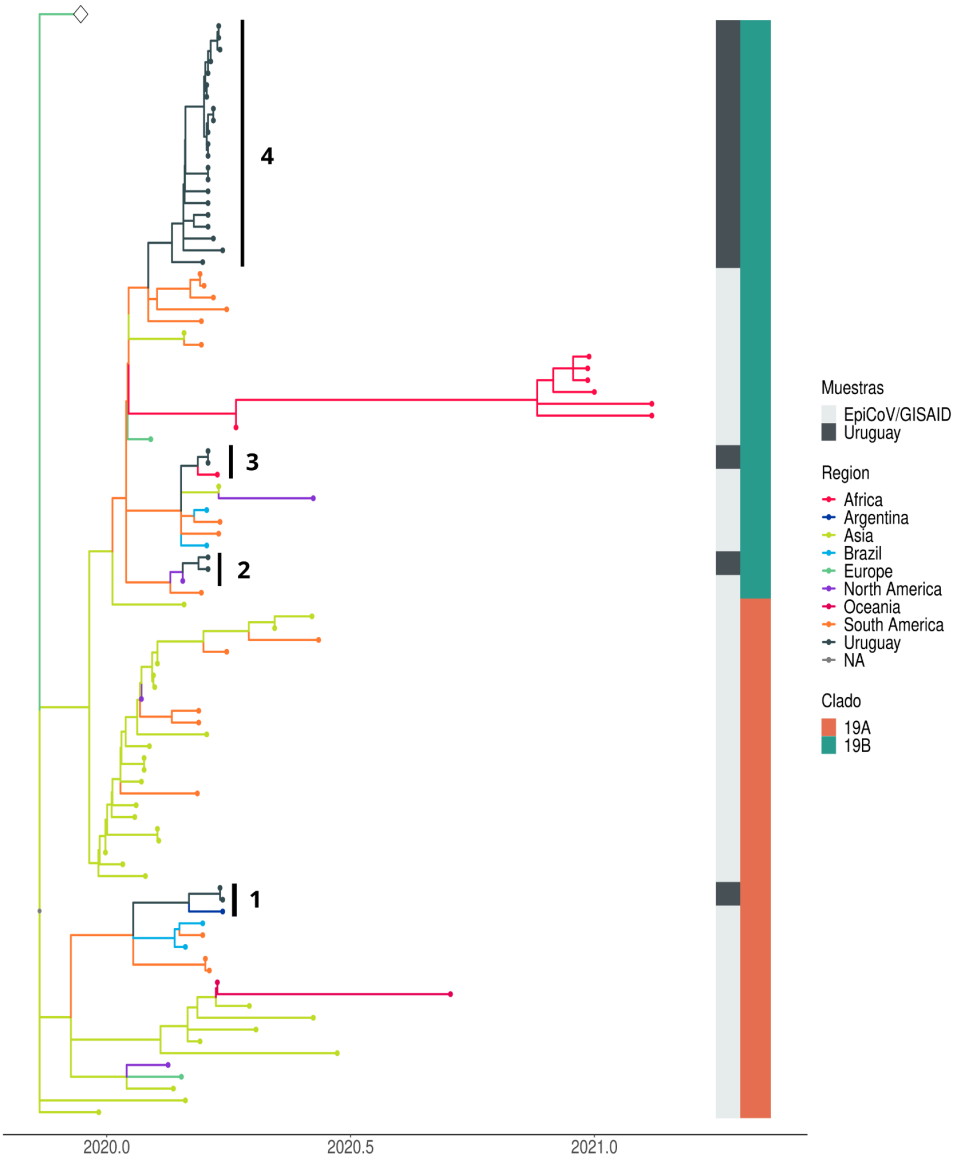

Figura 5: Árbol filogenético a escala temporal del conjunto de secuencias correspondientes a los clados que circularon en marzo del 2020 en Uruguay. El orígen geográfico más probable de las primeras introducciones de SARS-CoV-2 a Uruguay se estimó con el modelo de transiciones entre estados discretos de *TreeTime*. El color de las ramas corresponde al origen geográfico más probable. Se muestran cuatro de las nueve introducciones detectadas en marzo del 2020. La figura en forma de diamante contiene todas las ramas derivadas de ese nodo. Los números representan algunos de los distintos clusters de secuencias uruguayas del mes de marzo.

Se asume que las regiones geográficas están parametrizadas en función del tiempo de forma análoga a modelos que describen la evolución de las secuencias de los genomas con el modelo de reloj molecular. En este caso, las regiones geográficas fueron definidas como Sudamérica, Norteamérica, Europa, Asia, África, Oceanía, Argentina, Brasil y Uruguay. La mayor parte de las introducciones tienen un origen probable en Sudamérica. Asimismo, se registraron introducciones desde Norteamérica, Asia, Europa, Argentina y Brasil (**Tabla S2**).

5. Discusión

Los coronavirus son un grupo de virus de ARN que han causado dos epidemias de gran magnitud en las últimas décadas, específicamente debido a SARS-CoV y MERS-CoV (por *Middle East respiratory syndrome coronavirus*) (34–36). Tras la emergencia del nuevo coronavirus SARS-CoV-2, la declaración de la pandemia mundial y la confirmación de los primeros casos de COVID-19 en Uruguay, diferentes grupos de investigadores del sector público y privado pusieron a disposición sus recursos para conocer las distintas variantes que han circulado desde el inicio de la pandemia a nivel local (21,33–42). En base a las secuencias depositadas en la base de datos se observa que hubo circulación de los clados 19A, 19B, 20A, 20B y 20D durante marzo del 2020 en Uruguay. A partir de estas secuencias se pudo determinar que ya en los inicios de la pandemia en Uruguay circulaban linajes conteniendo una sustitución de ácido aspártico a glicina (D614G) en la proteína S del virus. Estos linajes identificados como portadores de la sustitución D614G circulando en Uruguay corresponden a B.1, B.1.195, B.1.1.33 y B.1.1.1. Esta sustitución luego estaría ampliamente distribuída en los distintos linajes posteriores a las cepas originales de SARS-CoV-2 y a su vez se encuentra presente en todos las variantes de preocupación reportadas hasta el momento. Algunos estudios han sugerido que esta sustitución mejora la eficiencia del ingreso de las células que expresan el receptor de la convertasa de angiotensina 2 (ACE2), respecto a la cepa original (46). Esta mayor eficiencia estaría asociada a la incorporación de un sitio de clivaje, estabilización de la interacción proteína S-ACE2 y mayor empaquetamiento de la proteína en la partícula viral (47). La frecuencia de la sustitución D614G tuvo un marcado aumento en abril del 2020 a nivel global y se cree que aún en los linajes más recientes de SARS-CoV-2, como la variante Omicron y sus subvariantes, esta sustitución podría estar implicada en la evasión de la respuesta inmune, transmisibilidad y susceptibilidad a la reinfección (48,49). En este sentido, durante el período julio-agosto del 2020 se registró la primera ola de contagios en el sur de Brasil con predominancia de los linajes B.1.1.33 y B.1.1.28, ambas conteniendo la mutación D614G (50). Estudios de la dinámica de SARS-CoV-2 en la frontera seca con Brasil, mostraron múltiples introducciones desde Brasil tanto de B.1.1.33 como de B.1.1.28 a través de las localidades fronterizas con el estado de Río Grande do Sul en el período mayo-julio (42). La cantidad de casos y fallecimientos por COVID-19 se mantuvieron en números relativamente bajos dado el contexto regional hasta diciembre del 2020, momento en el cual se inició la primera ola de contagios masiva en el país. Asimismo se identificó un sublinaje de B.1.1.28, denominado luego P.6, con dos mutaciones adicionales en el gen que

codifica la proteína S (Q675H + Q677H), cuyo origen más probable se encuentra situado en Uruguay. Se cree que estos cambios aminoacídicos cerca del sitio de clivaje S1/S2 pudieron haber estado involucrados en un aumento de la transmisibilidad de este linaje respecto a otros linajes co-circulantes en el país (39). También se reportó y caracterizó el ingreso de la variante de preocupación Gamma (alias P.1) al país (38).

Para SARS-CoV-2, y muchos virus de ARN, los cambios evolutivos son medibles entre distintos tiempos de muestreo, permitiendo el uso del modelo de reloj molecular para su reconstrucción filogenética a escala temporal. A partir de esto, el conjunto de datos seleccionado para evaluar el análisis de la introducción de SARS-CoV-2 en Uruguay mostró adecuarse a este modelo de reloj molecular evaluado a través del método *root-to-tip* (27). Por otro lado, la estimación de tasa de mutación de SARS-CoV-2 observada es similar a la media estimada anteriormente por Duchêne y colaboradores, 2020 en los inicios de la pandemia de COVID-19 (24). La edad de la raíz del árbol filogenético a escala temporal fue estimada a mediados de noviembre del 2019, la cual corresponde con otros reportes del origen de SARS-CoV-2 (51–53). La fecha estimada de introducción de SARS-CoV-2 en Uruguay se estimó a fines de febrero del 2020 y con una mayor cantidad de introducciones independientes desde Sudamérica respecto a otros orígenes. Además de Sudamérica, se detectaron introducciones desde Asia, Norteamérica y Brasil. Estimaciones similares fueron reportadas por Elizondo y colaboradores, 2021, donde el TMRCA del cluster principal conteniendo la secuencia de SARS-CoV-2 más antigua fue estimado entre el 2 y 5 de marzo del 2020 utilizando métodos bayesianos de inferencia espacio-temporal (54).

La rapidez sin precedentes de la disponibilización de los datos genómicos de SARS-CoV-2 ha permitido la implementación de forma consistente de la vigilancia genómica a escala global para el seguimiento en tiempo real de la evolución viral mediante el genotipado, serotipado y asignación de linajes. También ha asistido al rastreo de contactos e identificación de cadenas de transmisiones así como la identificación de variantes de preocupación antes de su dispersión internacional. Adicionalmente, ha contribuido a la generación del paisaje mutacional que ha permitido evaluar y desarrollar tratamientos y vacunas. Si bien se han realizado avances metodológicos significativos para afrontar la masiva generación de datos, las variaciones regionales y temporales en la generación de los mismos han estado fuertemente relacionados con las disparidades socioeconómicas a nivel mundial (55,56).

6. Material suplementario

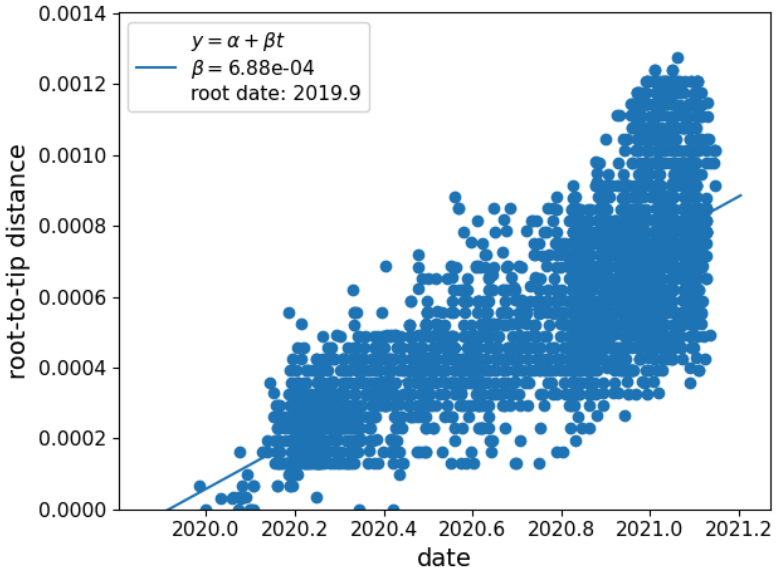

Figura S1: Correlación entre la divergencia y las fechas de muestreo de los componentes del *dataset* de estudio. La misma fue evaluada con el método *root-to-tip* de *TreeTime*.

Tabla S1: Secuencias genómicas de Marzo 2020 de Uruguay en EpiCoV/GISAID ordenadas según la fecha de muestreo. En negrita las secuencias conteniendo la sustitución D614G en la proteína S

Muestra	GISAID_ID	Región	Fecha	Linaje PANGO	Clado	Sustituciones en AA
Mdeo-1	EPI_ISL_444493	Uruguay	2020-03-13	A.5	19B	ORF8:L84S
UY-1	EPI_ISL_426476	Uruguay	2020-03-16	A.5	19B	ORF8:L84S
UY-2	EPI_ISL_426477	Uruguay	2020-03-16	A.5	19B	ORF8:L84S
UY-3	EPI_ISL_426478	Uruguay	2020-03-17	A.5	19B	ORF8:L84S
UY-4	EPI_ISL_426479	Uruguay	2020-03-17	A.5	19B	ORF8:L84S
UY-5	EPI_ISL_426480	Uruguay	2020-03-17	A.5	19B	ORF8:L84S
UY-6	EPI_ISL_426481	Uruguay	2020-03-17	A.5	19B	ORF1a:T2457K,ORF8:L84S
UY-7	EPI_ISL_426482	Uruguay	2020-03-17	A.1	19B	ORF1b:P1427L,ORF1b:Y1464C,ORF8:L84S
UY-8	EPI_ISL_426583	Uruguay	2020-03-17	A.1	19B	ORF1b:P1427L,ORF1b:Y1464C,ORF8:L84S
UY-NYUMC844	EPI_ISL_457940	Uruguay	2020-03-17	A.5	19B	ORF8:L84S
UY-NYUMC845	EPI_ISL_457941	Uruguay	2020-03-17	A.5	19B	ORF1a:R550S,ORF8:L84S
UY-NYUMC846	EPI_ISL_457942	Uruguay	2020-03-17	A.5	19B	ORF8:L84S

UY-NYUMC847	EPI_ISL_457943	Uruguay	2020-03-17	A.5	19B	ORF8:L84S
UY-NYUMC848	**EPI_ISL_457944**	**Uruguay**	**2020-03-17**	**B.1**	**20A**	**ORF1b:P314L,S:D614G**
UY-NYUMC849	EPI_ISL_457945	Uruguay	2020-03-17	A.2	19B	N:S197L,N:S310N,ORF1a:F3071Y,ORF8:L84S
UY-NYUMC850	EPI_ISL_457946	Uruguay	2020-03-17	A.2	19B	N:S197L,N:S310N,ORF1a:F3071Y,ORF3a:G196V,ORF8:L84S,S:I197V
UY-NYUMC851	EPI_ISL_457947	Uruguay	2020-03-17	A.5	19B	ORF8:L84S
UY-NYUMC852	EPI_ISL_457948	Uruguay	2020-03-17	A.5	19B	ORF8:L84S
UY-NYUMC853	**EPI_ISL_457949**	**Uruguay**	**2020-03-17**	**B.1**	**20A**	**ORF1a:P1096S,ORF1b:P314L,S:D614G**
UY-NYUMC940	**EPI_ISL_480438**	**Uruguay**	**2020-03-18**	**B.1.195**	**20A**	**ORF1b:P314L,S:D614G**
UY-10	EPI_ISL_429257	Uruguay	2020-03-19	B.31	19A	N:D22G,ORF1a:S944L,ORF1a:L3606F,ORF1b:D54Y,ORF1b:V1493L,ORF3a:G251V,ORF9b:I19V
UY-9	EPI_ISL_426584	Uruguay	2020-03-19	A.5	19B	ORF8:L84S
UY-NYUMC854	EPI_ISL_457950	Uruguay	2020-03-21	A.5	19B	ORF3a:G11E,ORF8:L84S
UY-NYUMC855	EPI_ISL_457951	Uruguay	2020-03-21	A.5	19B	ORF8:L84S
UY-NYUMC856	EPI_ISL_457952	Uruguay	2020-03-21	A.5	19B	ORF8:L84S
UY-NYUMC857	**EPI_ISL_457953**	**Uruguay**	**2020-03-23**	**B.1.1.33**	**20B**	**N:R203K,N:G204R,N:I292T,ORF1b:P314L,ORF6:I33T,S:D614G**
UY-NYUMC858	**EPI_ISL_457954**	**Uruguay**	**2020-03-23**	**B.1.195**	**20A**	**ORF1b:P314L,S:D614G**
UY-NYUMC859	**EPI_ISL_457955**	**Uruguay**	**2020-03-24**	**B.1.1.1**	**20D**	**N:R203K,N:G204R,ORF1a:G3278S,ORF1a:L3606F,ORF1b:P314L,S:D614G**
UY-NYUMC860	EPI_ISL_457956	Uruguay	2020-03-25	A.5	19B	ORF8:L84S
UY-NYUMC861	**EPI_ISL_457957**	**Uruguay**	**2020-03-25**	**B.1.1.1**	**20D**	**N:R203K,N:G204R,ORF1a:T1246I,ORF1a:G3278S,ORF1a:L3606F,ORF1b:P314L,S:D614G**
UY-NYUMC862	EPI_ISL_457958	Uruguay	2020-03-25	A.5	19B	ORF8:L84S
UY-NYUMC863	EPI_ISL_457959	Uruguay	2020-03-26	B	19A	N:T265I,ORF1a:L3606F,ORF3a:G251V
UY-NYUMC930	EPI_ISL_480428	Uruguay	2020-03-26	Unassigned	19B	ORF8:L84S
UY-NYUMC864	EPI_ISL_457960	Uruguay	2020-03-28	A.5	19B	ORF8:L84S
UY-NYUMC865	EPI_ISL_457961	Uruguay	2020-03-28	B	19A	N:T265I,ORF1a:L3606F,ORF3a:G251V
UY-NYUMC867	**EPI_ISL_457963**	**Uruguay**	**2020-03-28**	**B.1.1.1**	**20D**	**N:R203K,N:G204R,ORF1a:T1246I,ORF1a:G3278S,ORF1a:L3606F,ORF1b:P314L,S:D614G**

Tabla S2: Introducciones de SARS-CoV-2 en Uruguay en marzo del 2020. Los nodos corresponden al árbol timetree

	nodo mugration	nodo time tree	fecha del nodo	fecha numérica	número de secuencias	número de secuencias de Uruguay	orígen probable
Cluster 1	NODE_0000013	NODE_0001265	2020-03-03	2020.1699574	3	2	Sudamérica
Cluster 2	NODE_0000042	NODE_0001299	2020-03-10	2020.189436	2	2	Norteamérica
Cluster 3	NODE_0000043	NODE_0001309	2020-03-09	2020.188186	3	2	Sudamérica
Cluster 4	NODE_0000060	NODE_0001320	2020-02-19	2020.135013	21	21	Sudamérica
Cluster 5	NODE_0000745	NODE_0000929	2020-02-22	2020.143515	8	1	Europa
Cluster 6	NODE_0000762	NODE_0000948	2020-03-17	2020.209016	9	1	Sudamérica
Cluster 7	NODE_0001342	NODE_0001811	2020-03-23	2020.225410	1	1	Brasil
Cluster 8	NODE_0000782	NODE_0002835	2020-03-23	2020.225024	7	4	Argentina
Cluster 9	NODE_0000782	NODE_0001053	2020-03-17	2020.209055	11	6	Sudamérica

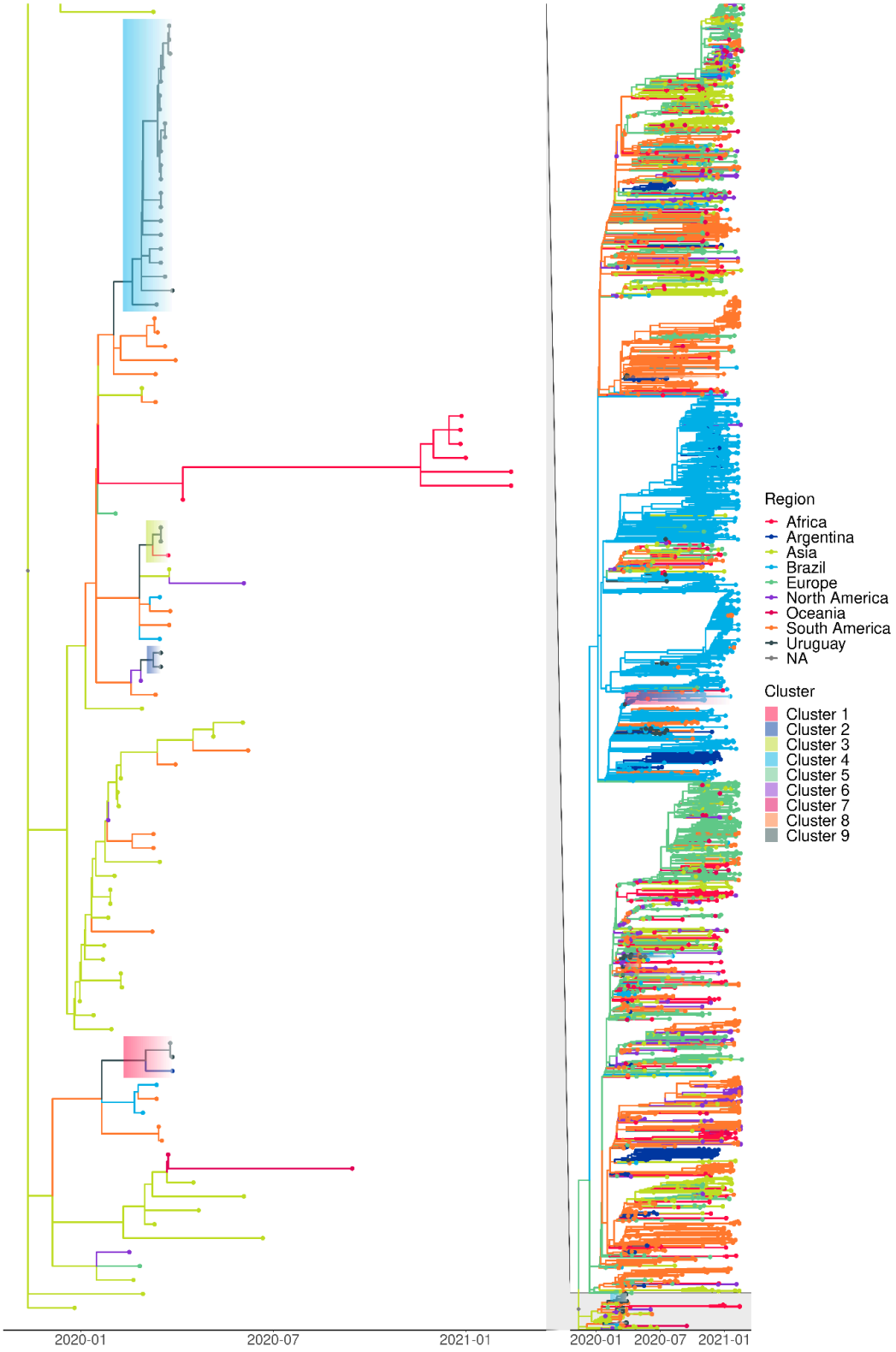

Figura S2: En el panel de la derecha se muestra la reconstrucción filogenética a escala espacio-temporal del conjunto de datos de estudio inferido con el modelo *joint* de *TreeTime*. Los colores de las ramas reflejan las transiciones geográficas entre las regiones ancestrales (Africa, Asia, Europa, Norteamérica, Sudamérica, Oceanía, Argentina y Brasil). En el panel de la izquierda se muestran sombreadas 4 de las 9 introducciones de SARS-CoV-2 detectadas en Uruguay en marzo del 2020.

7. Referencias

1. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 12 de marzo de 2020;579(7798):270-3.
2. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 12 de marzo de 2020;579(7798):265-9.
3. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet. febrero de 2020;395(10224):565-74.
4. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, Gorbalenya AE, Baker SC, Baric RS, De Groot RJ, Drosten C, et al. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2 de marzo de 2020;5(4):536-44.
5. WHO. Naming the coronavirus disease (COVID-19) and the virus that causes it [Internet]. 2020. Disponible en: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it#:~:text=For%20that%20reason%20and%20others,as%20agreed%20by%20the%20ICTV.
6. WHO. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 [Internet]. 2020. Disponible en: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020
7. Attwood SW, Hill SC, Aanensen DM, Connor TR, Pybus OG. Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. Nat Rev Genet. septiembre de 2022;23(9):547-62.
8. WHO. Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health [Internet]. 2021. Disponible en: https://www.who.int/publications/i/item/9789240018440
9. ECDC. Methods for the detection and characterisation of SARS-CoV-2 variants – second update [Internet]. 2022. Disponible en: https://www.ecdc.europa.eu/sites/default/files/documents/Methods-for-the-detection-char-SARS-CoV-2-variants_2nd%20update_final.pdf
10. Baker DJ, Aydin A, Le-Viet T, Kay GL, Rudder S, De Oliveira Martins L, et al.

CoronaHiT: high-throughput sequencing of SARS-CoV-2 genomes. Genome Med. diciembre de 2021;13(1):21.

11. Bull RA, Adikari TN, Ferguson JM, Hammond JM, Stevanovski I, Beukers AG, et al. Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. Nat Commun. 9 de diciembre de 2020;11(1):6272.
12. Carbo EC, Mourik K, Boers SA, Munnink BO, Nieuwenhuijse D, Jonges M, et al. A comparison of five Illumina, Ion Torrent, and nanopore sequencing technology-based approaches for whole genome sequencing of SARS-CoV-2. Eur J Clin Microbiol Infect Dis. junio de 2023;42(6):701-13.
13. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. febrero de 2016;530(7589):228-32.
14. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc. junio de 2017;12(6):1261-76.
15. Freed NE, Vlková M, Faisal MB, Silander OK. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. Biol Methods Protoc. 1 de enero de 2020;5(1):bpaa014.
16. Berno G, Fabeni L, Matusali G, Gruber CEM, Rueca M, Giombini E, et al. SARS-CoV-2 Variants Identification: Overview of Molecular Existing Methods. Pathogens. 17 de septiembre de 2022;11(9):1058.
17. Brant AC, Tian W, Majerciak V, Yang W, Zheng ZM. SARS-CoV-2: from its discovery to genome structure, transcription, and replication. Cell Biosci. diciembre de 2021;11(1):136.
18. Jackson CB, Farzan M, Chen B, Choe H. Mechanisms of SARS-CoV-2 entry into cells. Nat Rev Mol Cell Biol. enero de 2022;23(1):3-20.
19. Poterico JA, Mestanza O. Genetic variants and source of introduction of SARS‐CoV‐2 in South America. J Med Virol. octubre de 2020;92(10):2139-45.
20. Uruguay Presidencia. Medidas del Gobierno para atender la emergencia sanitaria por coronavirus (COVID-19) en materia de Relaciones Exteriores [Internet]. 2021. Disponible en: https://www.gub.uy/presidencia/politicas-y-gestion/medidas-del-gobierno-para-atender-emergencia-sanitaria-coronavirus-covid-19-4
21. Sant'Anna FH, Muterle Varela AP, Prichula J, Comerlato J, Comerlato CB, Roglio VS,

et al. Emergence of the novel SARS-CoV-2 lineage VUI-NP13L and massive spread of P.2 in South Brazil. Emerg Microbes Infect. 1 de enero de 2021;10(1):1431-40.

22. Salazar C, Díaz-Viraqué F, Pereira-Gómez M, Ferrés I, Moreno P, Moratorio G, et al. Multiple introductions, regional spread and local differentiation during the first week of COVID-19 epidemic in Montevideo, Uruguay [Internet]. Microbiology; 2020 may [citado 27 de mayo de 2023]. Disponible en: http://biorxiv.org/lookup/doi/10.1101/2020.05.09.086223
23. Aksamentov I, Roemer C, Hodcroft E, Neher R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw. 30 de noviembre de 2021;6(67):3773.
24. Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G. Temporal signal and the phylodynamic threshold of SARS-CoV-2. Virus Evol. 1 de julio de 2020;6(2):veaa061.
25. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol. enero de 2015;32(1):268-74.
26. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol [Internet]. 1 de enero de 2018 [citado 6 de mayo de 2023];4(1). Disponible en: http://academic.oup.com/ve/article/doi/10.1093/vex042/4794731
27. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. enero de 2016;2(1):vew007.
28. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. junio de 2017;14(6):587-9.
29. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 21 de noviembre de 2019;4(43):1686.
30. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd ed. 2016. Cham: Springer International Publishing : Imprint: Springer; 2016. 1 p. (Use R!).
31. Hadley Wickham, RStudio. stringr [Internet]. 2022. Disponible en: https://mirror.linux.duke.edu/cran/web/packages/stringr/stringr.pdf
32. Yu G. Using ggtree to Visualize Data on Tree‐Like Structures. Curr Protoc Bioinforma [Internet]. marzo de 2020 [citado 6 de mayo de 2023];69(1). Disponible en: https://onlinelibrary.wiley.com/doi/10.1002/cpbi.96

33. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 15 de noviembre de 2014;30(22):3276-8.
34. Lee N, Hui D, Wu A, Chan P, Cameron P, Joynt GM, et al. A Major Outbreak of Severe Acute Respiratory Syndrome in Hong Kong. N Engl J Med. 15 de mayo de 2003;348(20):1986-94.
35. Yu ITS, Li Y, Wong TW, Tam W, Chan AT, Lee JHW, et al. Evidence of Airborne Transmission of the Severe Acute Respiratory Syndrome Virus. N Engl J Med. 22 de abril de 2004;350(17):1731-9.
36. Bermingham A, Chand MA, Brown CS, Aarons E, Tong C, Langrish C, et al. Severe respiratory illness caused by a novel coronavirus, in a patient transferred to the United Kingdom from the Middle East, September 2012. Eurosurveillance [Internet]. 4 de octubre de 2012 [citado 18 de agosto de 2023];17(40). Disponible en: https://www.eurosurveillance.org/content/10.2807/ese.17.40.20290-en
37. Panzera Y, Ramos N, Frabasile S, Calleros L, Marandino A, Tomás G, et al. A deletion in SARS‐CoV‐2 ORF7 identified in COVID‐19 outbreak in Uruguay. Transbound Emerg Dis. noviembre de 2021;68(6):3075-82.
38. Rego N, Costábile A, Paz M, Salazar C, Perbolianachis P, Spangenberg L, et al. Real-Time Genomic Surveillance for SARS-CoV-2 Variants of Concern, Uruguay. Emerg Infect Dis. noviembre de 2021;27(11):2957-60.
39. Rego N, Salazar C, Paz M, Costábile A, Fajardo A, Ferrés I, et al. Emergence and Spread of a B.1.1.28-Derived P.6 Lineage with Q675H and Q677H Spike Mutations in Uruguay. Viruses. 10 de septiembre de 2021;13(9):1801.
40. Panzera Y, Goñi N, Calleros L, Ramos N, Frabasile S, Marandino A, et al. Genome Sequences of SARS-CoV-2 P.1 (Variant of Concern) and P.2 (Variant of Interest) Identified in Uruguay. Roux S, editor. Microbiol Resour Announc. 27 de mayo de 2021;10(21):e00410-21.
41. Panzera Y, Ramos N, Calleros L, Marandino A, Tomás G, Techera C, et al. Transmission cluster of COVID-19 cases from Uruguay: emergence and spreading of a novel SARS-CoV-2 ORF6 deletion. Mem Inst Oswaldo Cruz. 2021;116:e210275.
42. Mir D, Rego N, Resende PC, Tort F, Fernández-Calero T, Noya V, et al. Recurrent Dissemination of SARS-CoV-2 Through the Uruguayan–Brazilian Border. Front Microbiol. 28 de mayo de 2021;12:653986.
43. Panzera Y, Calleros L, Goñi N, Marandino A, Techera C, Grecco S, et al. Consecutive deletions in a unique Uruguayan SARS-CoV-2 lineage evidence the genetic variability

potential of accessory genes. Tse H, editor. PLOS ONE. 17 de febrero de 2022;17(2):e0263563.

44. Cancela F, Ramos N, Smyth DS, Etchebehere C, Berois M, Rodríguez J, et al. Wastewater surveillance of SARS-CoV-2 genomic populations on a country-wide scale through targeted sequencing. Raju N, editor. PLOS ONE. 21 de abril de 2023;18(4):e0284483.
45. Salazar C, Costabile A, Ferrés I, Perbolianachis P, Pereira-Gómez M, Simón D, et al. Case Report: Early Transcontinental Import of SARS-CoV-2 Variant of Concern 202012/01 (B.1.1.7) From Europe to Uruguay. Front Virol. 28 de mayo de 2021;1:685618.
46. Zhang L, Jackson CB, Mou H, Ojha A, Peng H, Quinlan BD, et al. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. Nat Commun. 26 de noviembre de 2020;11(1):6013.
47. Jackson CB, Zhang L, Farzan M, Choe H. Functional importance of the D614G mutation in the SARS-CoV-2 spike protein. Biochem Biophys Res Commun. enero de 2021;538:108-15.
48. Chakraborty C, Saha A, Bhattacharya M, Dhama K, Agoramoorthy G. Natural selection of the D614G mutation in SARS-CoV-2 Omicron (B.1.1.529) variant and its subvariants. Mol Ther - Nucleic Acids. marzo de 2023;31:437-9.
49. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. agosto de 2020;182(4):812-827.e19.
50. Varela APM, Prichula J, Mayer FQ, Salvato RS, Sant'Anna FH, Gregianini TS, et al. SARS-CoV-2 introduction and lineage dynamics across three epidemic peaks in Southern Brazil: massive spread of P.1. Infect Genet Evol. diciembre de 2021;96:105144.
51. Nie Q, Li X, Chen W, Liu D, Chen Y, Li H, et al. Phylogenetic and phylodynamic analyses of SARS-CoV-2. Virus Res. octubre de 2020;287:198098.
52. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. abril de 2020;26(4):450-2.
53. Singh D, Yi SV. On the origin and evolution of SARS-CoV-2. Exp Mol Med. abril de 2021;53(4):537-47.
54. Elizondo V, Harkins GW, Mabvakure B, Smidt S, Zappile P, Marier C, et al. SARS-CoV-2 genomic characterization and clinical manifestation of the COVID-19 outbreak in Uruguay. Emerg Microbes Infect. 1 de enero de 2021;10(1):51-65.

55. Tosta S, Moreno K, Schuab G, Fonseca V, Segovia FMC, Kashima S, et al. Global SARS-CoV-2 genomic surveillance: What we have learned (so far). Infect Genet Evol. marzo de 2023;108:105405.
56. Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, Kraemer MUG, et al. Global disparities in SARS-CoV-2 genomic surveillance. Nat Commun. 16 de noviembre de 2022;13(1):700

Capítulo IV

Generación de metodologías para el relevamiento rápido y costo-efectivo de SARS-CoV-2

Resultados incluídos en la publicación: Cecilia Salazar, Ignacio Ferrés, Mercedes Paz, Alicia Costábile, Gonzalo Moratorio, Pilar Moreno, Gregorio Iraola. Fast and cost-effective SARS-CoV-2 variant detection using Oxford Nanopore full-length spike gene sequencing. Microb Genom. 2023; 9(5): doi: https://doi.org/10.1099/mgen.0.001013.

1. Introducción

El constante desarrollo de metodologías costo-efectivas que permitan el monitoreo de agentes infecciosos constituyen una parte fundamental de los sistemas de vigilancia. Esto se ha puesto de manifiesto durante la reciente pandemia de COVID-19. Luego de su primera identificación en la provincia de Hubei, China, la infección se diseminó rápidamente a distintos países siendo declarada pandemia mundial por la OMS el 11 de marzo de 2020 (1). El nuevo Betacoronavirus SARS-CoV-2, el agente causal de la COVID-19; se trata de un virus envuelto con un genoma de ARN cadena simple de polaridad positiva (2–4). Como es el caso de los coronavirus, SARS-CoV-2 se une al receptor de la enzima de la convertasa de angiotensina 2 (ACE2, por *angiotensin convertase enzyme 2*) a través de la glicoproteína espícula o *Spike* (S), mediando la entrada del virus a la célula a través de la activación proteolítica de la misma por parte de la proteasa furina y otras (5). Desde la disponibilización del primer genoma de SARS-CoV-2 a inicios del 2020, millones de secuencias han sido compartidas a través de la base de datos EpiCoV/GISAID (https://gisaid.org/) (6). La disponibilización temprana de los datos de secuenciación de manera pública ha facilitado el desarrollo de metodologías de diagnóstico molecular específicas y ha permitido a su vez, conocer las cadenas de transmisión, así como las variantes genéticas del virus y su evolución (7,8).

A lo largo de la pandemia han surgido distintos sistemas de nomenclatura para identificar a SARS-CoV-2 (9,10). Probablemente la más utilizada corresponde a la nomenclatura PANGO. Los linajes PANGO son designados solamente si cierto linaje contiene al menos cinco secuencias con alto porcentaje de cobertura del genoma (≥ 95%), mientras que las secuencias que no están incluídas en el conjunto de designación son asignadas a una estimación de linaje (11,12). Más recientemente, se estableció una nueva nomenclatura para clasificar variantes de SARS-CoV-2 tomando en cuenta características de transmisibilidad, severidad de la enfermedad, incidencia de re-infecciones (por escape a la inmunidad natural) y efectividad de las vacunas (debido al escape de la inmunidad inducida por vacunas) (13). Estas variantes están categorizadas en variantes de interés (VOI, por *variants of interest*), variantes de preocupación (VOC, por *variants of concern*) y variantes de alta consecuencia (VHC, por *variants of high consequence*). Si bien, no se identificó ninguna VHC hasta el momento, se han designado cinco VOCs basados en gran medida por las mutaciones presentes en el gen S; estos son B.1.1.7, B.1.351, P.1 ,B.1.617.2 y B.1.1.529. Estas a su vez se corresponden con el sistema de nomenclatura que engloba colectivamente a

estos linajes y sublinajaes. Las mismas se denominan Alfa, Beta, Gamma, Delta y Omicron. En la actualidad, la mayor parte de las VOCs circulantes comprenden los linajes asociados a Omicron y sus descendientes.

Como sucede con la mayor parte de los virus ARN, las mutaciones en el genoma de SARS-CoV-2 se espera que sean deletéreas o neutrales y por lo tanto, solamente una pequeña proporción de éstas mutaciones afectaría las funciones del virus en cuanto a su infectividad, severidad y antigenicidad. Las sustituciones que afectan a la proteína S podrían generar cambios en la interacción con el receptor ACE2 pudiendo facilitar el ingreso a la célula hospedera. Además se podrían afectar los determinantes antigénicos de la proteína y esto resulta de interés ya que la proteína S es el principal blanco de los anticuerpos neutralizantes generados en la infección y/o vacunación (14). La selección positiva de las mutaciones en el gen S ha tenido como consecuencia la emergencia de variantes de SARS-CoV-2 y posiblemente una mejor adecuación biológica. Un ejemplo de esto es la introducción de la sustitución D614G presente en las cinco VOCs (**Figura 1**) (15,16).

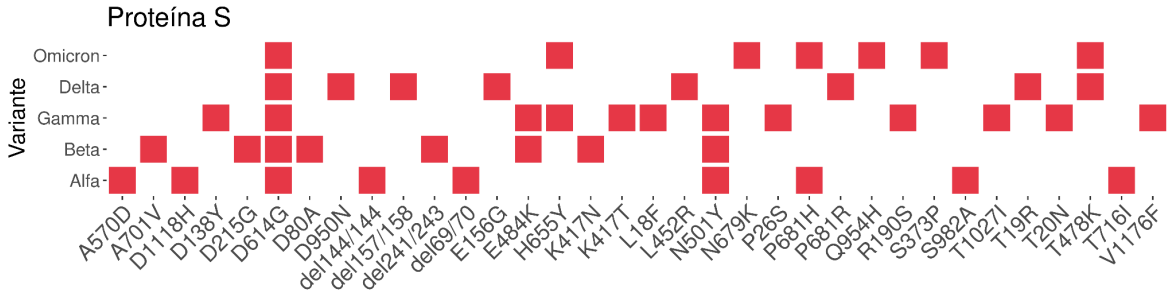

Figura 1: Perfil de sustituciones aminoacídicas en la secuencia de la proteína S de SARS-CoV-2 (outbreak.info). La sustitución D614G emergió al inicio de la pandemia en el año 2020 y se encuentra presente en todas las VOCs.

Las metodologías estándar para la caracterización de variantes de SARS-CoV-2 es la secuenciación del genoma viral completo o WGS (por *whole genome sequencing*). Sin embargo, desde los inicios de declarada la emergencia sanitaria, se han hecho esfuerzos para la determinación de variantes de SARS-CoV-2 basadas en RT-PCR cuantitativas, reduciendo los costos y el tiempo de reporte en comparación con métodos de WGS (17). La gran limitación de esta estrategia es que provee de información estática, no permitiendo la identificación de cambios potencialmente relevantes presentes en la secuencia del gen blanco como es el caso del gen que codifica para la proteína S. A modo de superar esta limitación, se han implementado estrategias alternativas utilizando la secuenciación de alto rendimiento del gen S utilizando la plataforma Illumina (18). Estas estrategias apuntan al relevamiento de un

gran volumen de muestras y/o se requiere filtrar la cantidad de muestras que pasan a la etapa de secuenciación genómica.

En esta sección se presentan dos estrategias para la determinación de variantes de SARS-CoV-2 basado únicamente en la secuencia del gen S, llamadas metodologías "estándar" y "rápida", respectivamente. En las mismas se utiliza la plataforma de secuenciación de Oxford Nanopore Technologies (ONT) y se basan en la generación de amplicones solapantes de la región del gen S utilizando cebadores previamente reportados por ARTIC Network (https://artic.network/) para la secuenciación del genoma completo de SARS-CoV-2 (19). Asimismo, se incorpora al análisis de los datos una herramienta recientemente desarrollada para la asignación de linajes PANGO basada exclusivamente en la secuencia del gen S. Esto permite la clasificación de variantes de SARS-CoV-2 bajo el concepto de "set de linajes", donde se tiene en cuenta que distintos linajes comparten la misma secuencia del gen S (20). Debido a que la gran parte de las mutaciones que definen una VOC han sido identificadas en el gen S, se espera que nuevas mutaciones que implican ventajas evolutivas para el virus sean fijadas a lo largo del tiempo. Por este motivo, el desarrollo de estrategias que apuntan a generar información rápida y con costos más accesibles respecto a los métodos de WGS resulta relevante para la identificación temprana de cambios en el principal blanco de las vacunas actuales contra SARS-CoV-2. Adicionalmente, éstas estrategias pueden extenderse a otros genes y otros patógenos de interés para la salud, ya sea con fines de caracterización genética primaria o de relevamiento epidemiológico. También permite generar criterios para la toma de decisiones sobre qué muestras deben pasar a la etapa de WGS disminuyendo la redundancia de la información genómica generada en un mismo contexto.

2. Objetivo general

Implementar una metodología de clasificación rápida de SARS-CoV-2 utilizando la secuenciación de amplicones solapantes de la región que codifica para la proteína S en un ensayo piloto.

2.1 Objetivos específicos

1. Generar secuencias consenso del gen S de SARS-CoV-2 a partir de datos de secuenciación ONT.

2. Determinar el promedio de profundidad de secuenciación y completitud del gen S necesaria para la asignación del linaje y/o variante.
3. Reducir el tiempo y el costo en la preparación de la biblioteca de secuenciación.

3. Materiales y métodos

3.1 Secuenciación genómica de SARS-CoV-2

3.1.1 Muestras clínicas de SARS-CoV-2

Un total de 44 muestras de ARN extraído a partir del exudado nasal de pacientes SARS-CoV-2 positivos fueron referidos al Centro de Innovación en Vigilancia Epidemiológica (CIVE) para su secuenciación genómica. Los detalles disponibles de las muestras se pueden encontrar en la **Tabla S1**.

3.1.2 Estrategia de amplificación del genoma viral y secuenciación ONT de SARS-CoV-2

La amplificación del genoma y preparación de la biblioteca de secuenciación se realizó principalmente en base al protocolo descrito por Freed et al. 2020 (21). Detalles adicionales del procedimiento se describen en Salazar et al. 2023 (22).

3.1.3 Asignación de bases, demultiplexado y generación de las secuencias consenso

La asignación de bases de alta precisión y el demultiplexado de las muestras se realizó con un Guppy v4.0 o superior (https://nanoporetech.com/). La generación de los consensos, determinación de la completitud del genoma, los linajes PANGO y clado Nextrain se llevó a cabo con el pipeline *poreCov* (23).

3.2 Secuenciación del gen S de SARS-CoV-2

Al igual que para el genoma viral, el ARN extraído de las muestras positivas para SARS-CoV-2, fue sometido a transcripción reversa utilizando LunaScript® RT SuperMix Kit (New England Biolabs, MA) de acuerdo a las instrucciones del fabricante en un volumen final de 10 µL. Los cebadores utilizados corresponden a los descritos en el protocolo ONT *Spike Seq RT PCR Expansion (SQK-RBK110.96 and EXP-SRT001)*. Estos cebadores pertenecen al esquema ARTIC V3 (https://github.com/artic-network/primer-schemes/blob/master/nCoV-2019/V3/nCoV-2019.tsv) para la amplificación del genoma y generan un producto de amplificación de

aproximadamente 1000 pares de bases, los cuales cubren el rango de posiciones entre 21.076 y 26.315 de la secuencia de referencia EPI_ISL_402124 (WIV04). Adicionalmente, los cebadores fueron sintetizados con etiquetas específicas en sus extremos 5' (Fw-TTTCTGTTGGTGCTGATATTGC y Rv-ACTTGCCTGTCGCTCTATCTTC) para la generación de muestras indexadas con la metodología de PCR utilizando el protocolo *Ligation sequencing amplicons - PCR barcoding (SQK-LSK109 with EXP-PBC096)* de ONT (**Tabla S2**).

3.2.1 Metodología estándar

La amplificación del gen S se realizó utilizando dos reacciones por muestra en el formato pool A con los cebadores pares y pool B con los cebadores impares. Cada reacción consiste en un volumen final de 12.5 µL conteniendo 10.5 µL de Q5® Hot Start High-Fidelity 2X Master Mix (New England Biolabs, MA) y una concentración final de 0.6 µM del pool de cebadores A y B, respectivamente al cual se añadió 2 µL del producto de retrotranscripción. Se incubó en el termociclador durante 30 s a 98 ºC, seguido de 20 ciclos de 15 s a 98 ºC y 3 m a 63 ºC. Luego de la amplificación los productos de PCR del pool A y B fueron combinados y diluídos 1:10 en H_2O libre de nucleasas. Los índices ONT fueron añadidos en un segundo paso de PCR. El mix de PCR de indexado consiste en 6 µL de Q5® Hot Start High-Fidelity 2X Master Mix, 1 µL de cebadores de indexado del kit EXP-PBC096 (ONT) y 5 µL del producto de amplificación diluido. Se incubó en el termociclador durante 30 s a 98 ºC, seguido de 15 ciclos de 7 s a 98 ºC, 15 s a 62 ºC y 30 s a 72 ºC, con un paso final de 2 m a 72 ºC. Las reacciones de indexado fueron combinadas y el pool de muestras se incubó con un volumen de 0.5X de Agencourt AMPure XP (Beckman Coulter™) a temperatura ambiente y en agitación. Luego se incubó el pool de muestras en un rack magnético durante 5 m, se descartó el sobrenadante y se realizaron dos lavados con alcohol 70% sin retirar el tubo del rack magnético. Se dejó secar durante ~30 s, se añadió 50 µL de agua libre de nucleasas y se incubó durante 2 m a temperatura ambiente. Luego, el pool de muestras se cuantificó utilizando un método fluorométrico (Qubit dsDNA HS Assay Kit, Thermo Fisher Scientific). Se preparó 1 µg del pool de muestras purificadas en 48 µL y se incubó con 3.5 µL de NEBNext FFPE DNA Repair Buffer (New England Biolabs, MA), 2 µL de NEBNext FFPE DNA Repair Mix (New England Biolabs, MA), 3.5 µL Ultra II End-prep reaction buffer y 3 µL Ultra II End-prep enzyme mix. Se incubó durante 5 m a 20 ºC y 5 m a 65 ºC. Luego de la preparación de extremos y reparación del ADN se procedió a una

purificación adicional con un volumen 1X de Agencourt AMPure. El pool de muestras reparado fue eluído en 61 µL de agua libre de nucleasas. La ligación del adaptador de secuenciación se realizó con 60 µL del pool de muestras, 25 µL del buffer de ligación (LNB) del kit SQK-LSK109 (ONT), 10 µL de NEBNext Quick T4 DNA Ligase (New England Biolabs, MA) y 5 µL de AMX adapter del kit EXP-PBC096 (ONT). La reacción se incubó aproximadamente 10 min a temperatura ambiente. La biblioteca de secuenciación final fue purificada incubando con un volumen 0.4X de Agencourt AMPure XP. Se incubó luego 5 m en un rack magnético y se descartó el sobrenadante. Se realizaron dos lavados del pellet con el buffer SFB del kit SQK-LSK109 (ONT) y se eluyó la biblioteca en 15 µL del buffer EB del mismo kit durante aproximadamente 10 m a 37 ºC. Se cuantificó la biblioteca de secuenciación final y se preparó la misma para la secuenciación. Se utilizaron aproximadamente 300 ng de la biblioteca en un volumen final de 12 µL y se mezclaron con 37.5 µL de SQB y 25.5 µL de LB en un volumen final de 75 µL y se secuenció durante aproximadamente 18 hs.

3.2.2 Metodología rápida

A diferencia del método estándar, con la metodología rápida la amplificación del gen S y el indexado de la muestra se realizan en el mismo tubo. En primer lugar se inmovilizan los cebadores de indexado en la tapa del tubo basado en un método descrito anteriormente (24). Brevemente, 1 µL de cebadores de indexado son colocados en el centro de la tapa de tubos *strips*. A modo de visualizar la gota, se agregó previamente 1 µL de una dilución 1:10 de un colorante específico de ADN como el TriTrack DNA Loading Dye (6X) (Thermo Scientific). Se incubó a temperatura ambiente hasta lograr el secado total del líquido en la tapa. A su vez, se prepararon dos reacciones de PCR (pool A y B) en un volumen final 12 µL utilizando una LongAmp® Taq 2X Master Mix (New England Biolabs, MA) e igual concentración de cebadores que para el método estándar (**Tabla 2**). Se realizó el primer paso de amplificación incubando en el termociclador durante 30 s a 94 ºC con 20 ciclos a 15 s a 94 ºC y 3 m a 63 ºC. El segundo paso de indexado se realiza retirando los tubos del termociclador y homogeneizando los cebadores de indexado, previamente inmovilizados en la tapa, en el mix de PCR sin necesidad de abrir los tubos. Se realiza una breve centrifugación de los tubos y se incuban nuevamente en el termociclador 3 m a 95 ºC, seguido de 15 ciclos de 15 s a 95 ºC, 15 s a 62 ºC, 50 s a 65 ºC, con una extensión final a 65 ºC durante 10 m. Una vez finalizada la amplificación e indexado se combinaron todos los productos de PCR y se

realizó una limpieza con un volumen 0.5X de Agencourt AMPure XP y se procedió a realizar la preparación de extremos, reparación y unión de adaptadores de secuenciación al igual que para el método estándar. Se cuantificó la biblioteca final con un método fluorométrico y se cargó aproximadamente 800 ng en una celda de secuenciación FLO-MIN106D y se realizó el seguimiento de la secuenciación durante aproximadamente 12 h.

3.2.3 Generación de secuencias consenso para el gen S y estimación del linaje PANGO

Se obtuvieron cinco réplicas de un muestreo con reposición de los datos de secuenciación de cada una de las muestras con *fastq-tools v0.8.3* (https://github.com/dcjones/fastq-tools). Las secuencias consenso para el gen S fueron obtenidas con el *pipeline* de Nextflow epi2me-lab/wf-artic (25) utilizando el esquema spike-seq, V1. A modo de obtener una estimación del linaje PANGO se utilizó *hedgehog v1.0.19* (20,26). La completitud del gen S se estimó con *president v0.6.3* (27) utilizando la región correspondiente al gen S de la secuencia de referencia WIV04. La profundidad de secuenciación promedio se obtuvo con *samtools v1.14* (28). Las diferencias entre la profundidad de secuenciación y la completitud del gen fueron evaluadas con el test de Wilcoxon implementado en el paquete estadístico *ggsignif* (29). Debido a que las variantes Gamma, Delta y Omicron comprenden un set diverso de linajes y sublinajes, se utilizó una versión reducida de la nomenclatura a modo de realizar las comparaciones. Se incluyó un asterisco para englobar todas los subtipos de determinado linaje o sublinaje (por ejemplo, BA.1.1 está representada como BA.1.*).

3.3 Evaluación de los sublinajes de Omicron utilizando el gen S

A modo de evaluar la posibilidad de utilizar el gen S como un potencial marcador genético para clasificar linaje y sublinajes, se utilizaron 5.000 secuencias obtenidas aleatoriamente de la base de datos EpiCoV/GISAID (EPI_SET_220929wd). Las mismas fueron alineadas con *Nextalign CLI v1.4.5* (30) utilizando la secuencia de referencia WIV04. Las secuencias del alineamiento entre las posiciones 21.076 y 26.6315 fueron extraídas manualmente utilizando *Aliview* (31). El set de linaje PANGO fue estimado con *hedgehog v1.0.19* como se describió anteriormente. La asignación de linaje obtenido con el gen S y el

reportado para el genoma de donde proviene S fueron comparados y visualizados con *ggplot2* (32).

4. Resultados

4.1 Detección de linajes de SARS-CoV-2 basada en la secuencia del gen S obtenida a través de la metodología estándar de secuenciación de amplicones

A modo de determinar si la estrategia de generación de bibliotecas de secuenciación basadas en el protocolo de ligación e indexado por PCR de ONT era adecuada para la obtención de secuencias consenso del gen S y estimación del linaje PANGO y/o variantes, los cebadores reportados para la amplificación del gen S en un protocolo basado en la química de trasposición de *ONT Spike Seq RT PCR Expansion (SQK-RBK110.96 and EXP-SRT001)* fueron adaptados para ser utilizados con la estrategia de *PCR barcoding* de ONT. Como método alternativo, se utilizaron cebadores conteniendo etiquetas compatibles para el protocolo de indexado de amplicones a través de la estrategia de PCR (protocolo *Ligation - PCR barcoding (SQK-LSK109 + EXP-PBC096)*). Con esta metodología, aproximadamente el 70% de los linajes fueron asignados correctamente (**Tabla S3**). Se observa que las muestras sin asignar o las clasificaciones incorrectas están asociadas a una completitud menor al 50% (**Figura 2A**). En este sentido, se observaron diferencias significativas ($P \leq 0.001$) en cuanto a la completitud del gen y la profundidad de secuenciación entre las asignaciones correctas e incorrectas del set de linaje PANGO. Específicamente, las asignaciones incorrectas están asociadas a una menor completitud del gen S, la media de la completitud fue de 56% ± 36% acompañados de una profundidad de secuenciación de 219X ± 210X. Por su parte, las asignaciones correctas presentaron una completitud del gen de 95% ± 7% y una profundidad de secuenciación de 463X ± 99X (**Figura 2B**). Esto podría servir de guía para la determinación del esfuerzo de secuenciación necesario para la obtención del linaje y/o variante con alto grado de confianza.

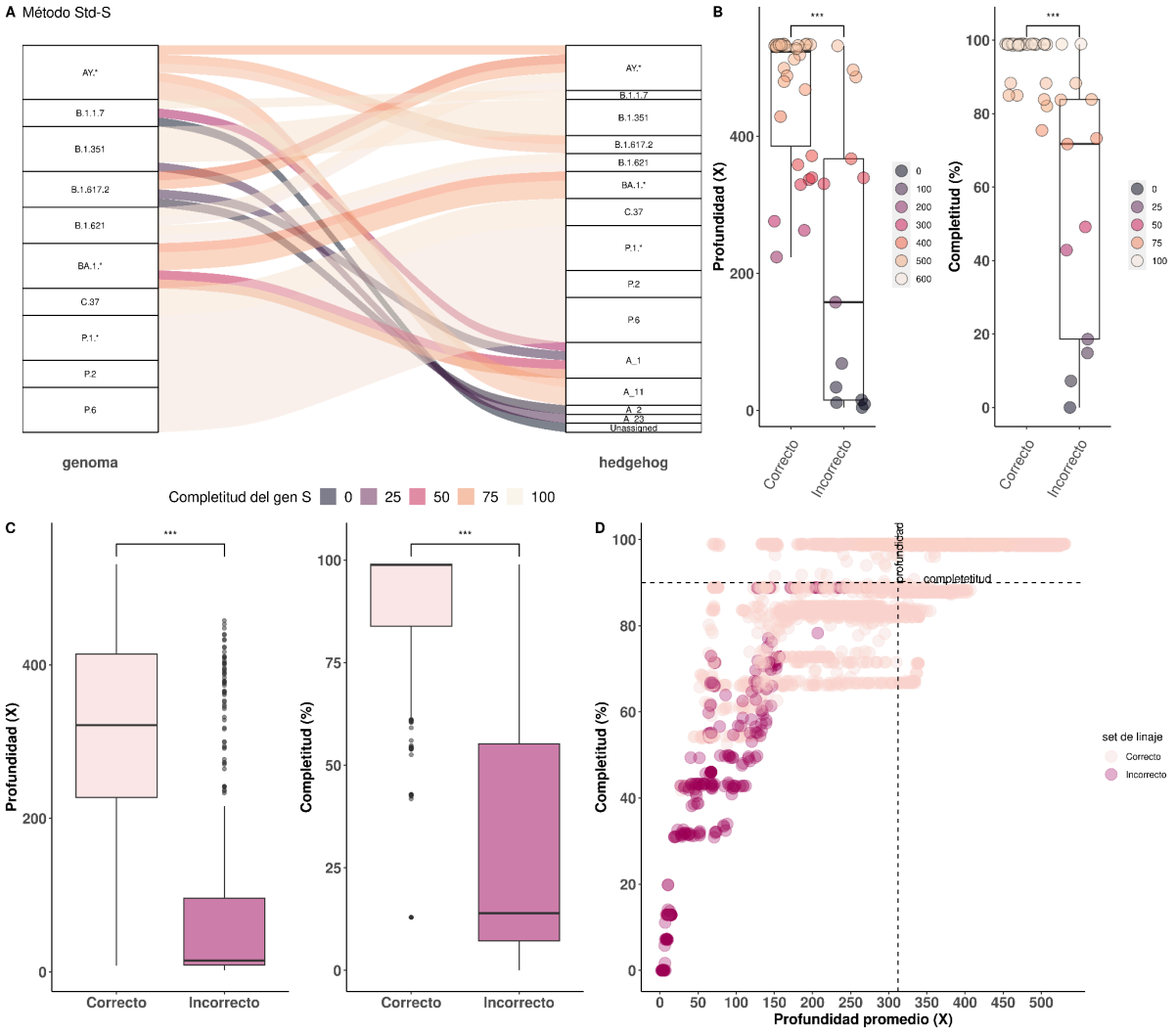

Figura 2: Determinación del set de linaje de muestras VOC, VOI y no VOC/VOI (P.2 y P.6) obtenidas con el protocolo estándar de secuenciación del gen S. A) Comparación del linaje obtenido con la secuenciación del genoma completo y el gen S. Las conexiones entre ambos resultados reflejan la completitud obtenida para el gen S. B) Comparación entre las asignaciones correctas e incorrectas en términos de profundidad de secuenciación (X) y completitud del gen S (%). Se observaron diferencias significativas entre las asignaciones correctas e incorrectas ($P \leq 0.001$). C) Asignación de linajes correctos e incorrectos obtenidos a partir de las secuencias consenso generadas tras el muestreo de los datos de secuenciación (500 a 10.000 lecturas) para cada una de las muestras. Al igual que para las muestras secuenciadas, en los consensos simulados se observan diferencias significativas entre la profundidad de secuenciación y la completitud del gen. D) Promedio de profundidad de secuenciación y completitud del gen. Las asignaciones correctas se maximizan en valores por encima del promedio tanto para la profundidad de secuenciación como la completitud del gen (cuadrante superior derecho).

Con la finalidad de entender el efecto que tiene la profundidad de secuenciación y la completitud del gen en la obtención de asignaciones de variantes VOC, VOI y no VOC/VOI, se muestrearon al azar lecturas de secuenciación para cada una de las muestras obteniendo desde 500 hasta 10.000 lecturas (**Figura S1 y S2**). Luego se generó el consenso del gen S para cada punto de muestreo y se determinó el set de linaje PANGO, la profundidad promedio de secuenciación alcanzada y la completitud del gen S. A partir de los consensos generados con los datos muestreados se observó una diferencia significativa entre la

completitud del gen S y la profundidad de secuenciación, tal como se había observado con los consensos generados para cada una de las muestras. Particularmente, se observó que las clasificaciones correctas presentan en promedio 90% ± 15% de completitud del gen y una profundidad de secuenciación de 312X ± 127X (**Figura 2C**). Si tomamos en cuenta éstos valores promedio, el 97% de los puntos de muestreo que se encuentran por encima de este umbral fueron asignados correctamente. Esto sugiere que las muestras que alcanzan y exceden estos valores umbrales son asignadas con un alto nivel de confianza (**Figura 2D**).

En resumen, los resultados muestran que las secuencias consenso generadas con el protocolo estándar de secuenciación de amplicones resulta en una herramienta útil para la determinación del set de linajes PANGO a partir de datos de secuenciación del gen S. Asimismo, se estimó un umbral de profundidad de secuenciación y completitud del gen. Si el gen presenta una completitud mayor al 90% y una profundidad de secuenciación mayor a 300X se maximizan las asignaciones correctas.

4.2 La utilización del protocolo rápido permite la detección de VOCs y VOIs a un costo reducido.

Luego de corroborar la posibilidad de utilizar secuenciación ONT a partir de la amplificación del gen S en fragmentos solapantes utilizando el protocolo estándar, el siguiente paso consistió en generar un protocolo donde se pudiera reducir los costos y el tiempo de manipulación y por lo tanto del reporte del resultado. A partir de esto, se desarrolló el protocolo rápido cuya principal diferencia respecto al protocolo estándar consiste en que la amplificación y el indexado de la muestra se llevan a cabo en el mismo tubo utilizando el mismo mix de PCR para ambos pasos. Mientras que la PCR de amplificación (round 1) que genera amplicones con etiquetas en los extremos se lleva a cabo en el fondo del tubo, la PCR de indexado (round 2) ocurre una vez finalizado el round 1. En este momento los cebadores de indexado previamente inmovilizados en la tapa del tubo se disuelven en la reacción de PCR mediante inversión y centrifugación. Luego los tubos se colocan en el termociclador para continuar con el ciclo de indexado. Estos cebadores añaden el índice (o *barcode*) a partir de las etiquetas añadidas en el primer paso de amplificación.

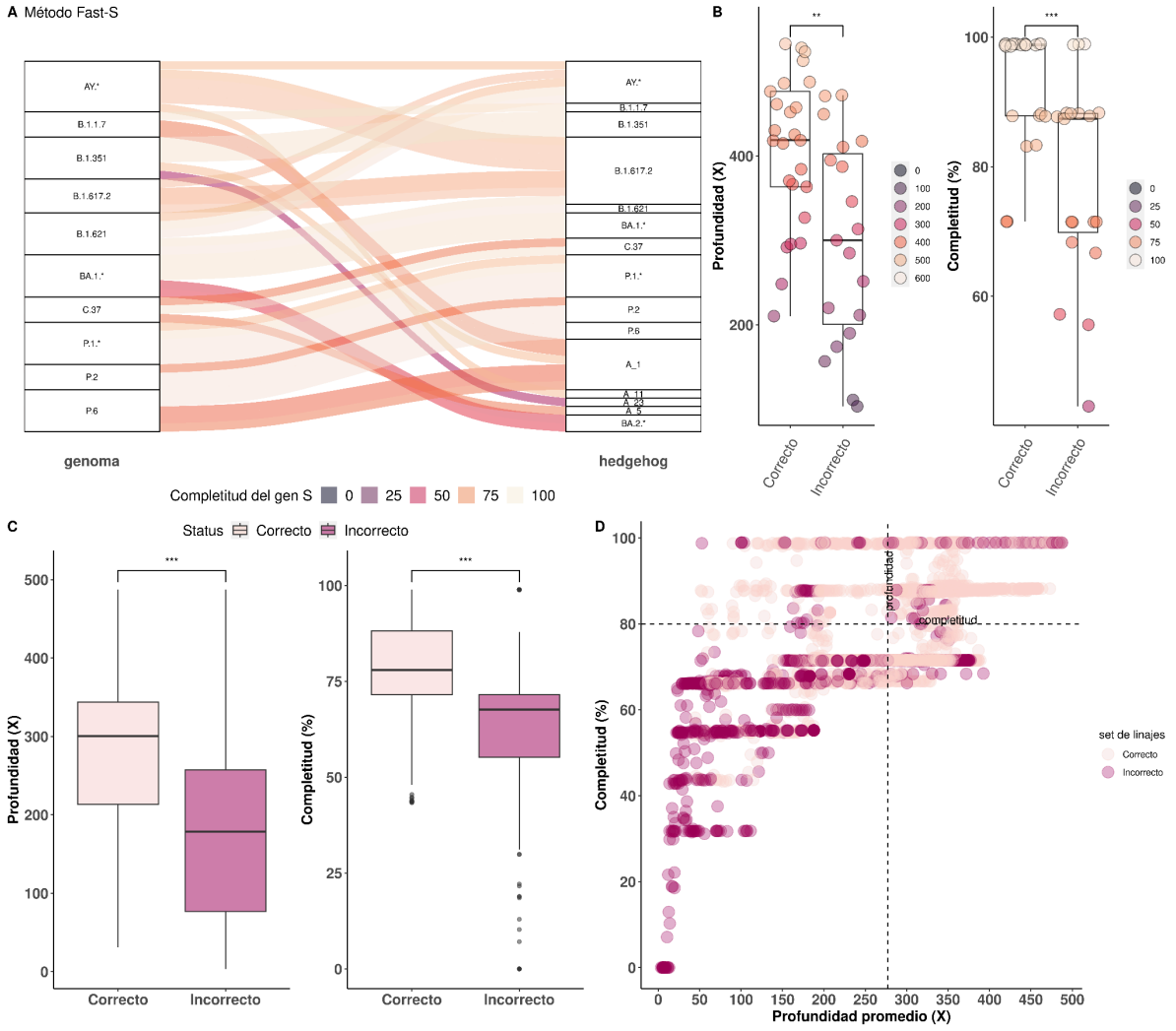

Figura 3: Determinación del set de linaje de muestras VOC, VOI y no VOC/VOI (P.2 y P.6) obtenidas con el protocolo rápido de secuenciación del gen S. A) Comparación del linaje obtenido con la secuenciación del genoma completo y el gen S. Las conexiones entre ambos resultados reflejan la completitud obtenida para el gen S. B) Comparación entre las asignaciones correctas e incorrectas en términos de profundidad de secuenciación (X) y completitud del gen S (%). Se observaron diferencias significativas entre las asignaciones correctas e incorrectas ($P \leq 0.01$ y $P \leq 0.001$). C) Asignación de linajes correctos e incorrectos obtenidos a partir de las secuencias consenso generadas tras el muestreo de los datos de secuenciación (500 a 10.000 lecturas) para cada una de las muestras. Al igual que para las muestras secuenciadas, en los consensos simulados se observan diferencias significativas ($P \leq 0.001$) entre la profundidad de secuenciación y la completitud del gen. D) Promedio de profundidad de secuenciación y completitud del gen. Las asignaciones correctas se maximizan en valores por encima del promedio tanto para la profundidad de secuenciación como la completitud del gen (cuadrante superior derecho).

Como con el protocolo estándar, las muestras asignadas incorrectamente presentaron una menor completitud del gen S (73.6% ± 14.3%), a excepción de la las muestras correspondientes al linaje PANGO B.1.621 que presentaron alto porcentaje de completitud (96.1% ± 5.5%) (**Figura 3A**). Se observaron diferencias significativas en términos de profundidad de secuenciación ($P \leq 0.01$) y completitud del gen S ($P \leq 0.001$). Las muestras

clasificadas correctamente presentaron una completitud promedio de 94% ± 8% y una profundidad de secuenciación de 407X (± 90X).

Las asignaciones incorrectas presentaron una completitud del 78% ± 15% y un promedio de profundidad de secuenciación del 298X ± 122X (**Figura 3B**). Luego del muestreo las lecturas de secuenciación y reasignación de los consensos generados, se observan diferencias significativas ($P \leq 0.001$) tanto para la profundidad de secuenciación como para la completitud del gen S (**Figura 3C**). En este caso, las asignaciones incorrectas estuvieron asociadas a una profundidad de secuenciación y completitud del gen menores a 178X y 64%, respectivamente. Se requiere la combinación de una profundidad de secuenciación mayor a 277X ± 92X y completitud del gen S de 80% ± 13% para maximizar la correcta asignación del set de linajes. Específicamente, el 82% de los consensos en el cuadrante superior derecho fueron asignados de forma correcta (**Figura 3D**).

4.3 La robustez de la clasificación se mantiene en Omicron y sus sublinajes

A modo de evaluar si la herramienta de asignación del set de linajes mantiene su robustez con los distintos sublinajes de la variante Omicron, se utilizaron 5000 secuencias genómicas obtenidas al azar de la base de datos EpiCoV/GISAID (**Figura S5**). El set de linajes obtenidos con *hedgehog* fue comparado con el linaje reportado para la secuencia genómica de las muestras. Se obtuvo ~ 98% de asignaciones correctas utilizando el nivel de asignación BA.1*, BA.2*, etc (**Figura 4A**). Específicamente, 2.587 de 2634 (98.2%) secuencias BA.1* fueron clasificadas como tales utilizando el gen S. 2.048 de 2.096 (97.7%) secuencias de BA.2* fueron clasificadas correctamente, 55 de 57 (96.5%) de secuencias BA.4* y 158 de 161 (98.1%) de las secuencias BA.5* (**Figura 4B**). El detalle de las clasificaciones exactas basadas en el gen S se encuentran en la **Figura S6** y en la Tabla S5. Adicionalmente, se observan diferencias significativas entre las asignaciones correctas e incorrectas (**Figura S7**). Las asignaciones correctas tienen una completitud del gen del 97.2% ± 8.6% y las asignaciones incorrectas 94.3% ± 5.1%. Esto sugiere que se requiere altos niveles de completitud para alcanzar el nivel de asignación de linaje y sublinajes de Omicron.

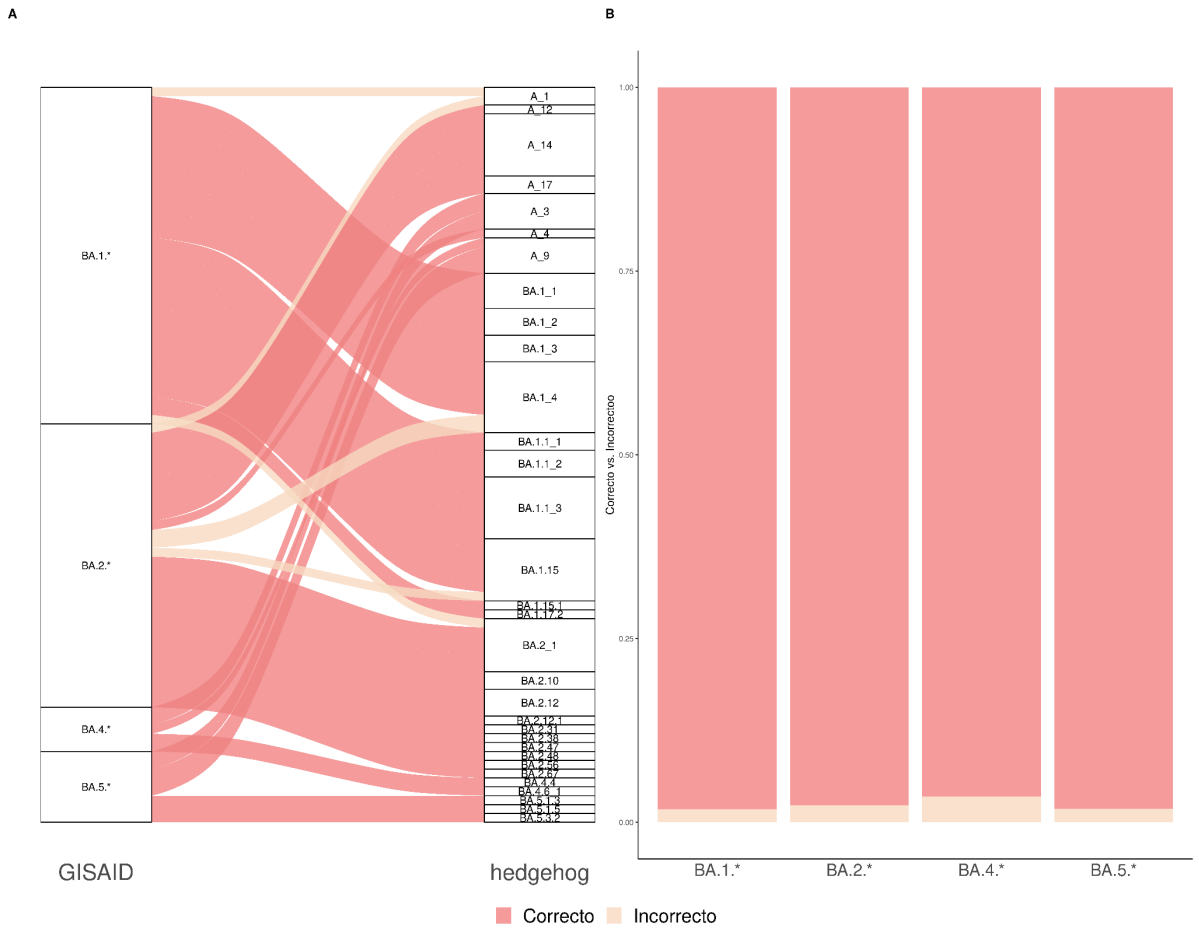

Figura 4: Asignación de linajes PANGO basados en genomas en comparación con el set de linaje basado en el gen S obtenido con la herramienta hedgehog. A) comparación de linajes obtenidos con el genoma completo y el gen S de un subset de 300 secuencias de Omicron. B) Clasificación correcta e incorrectas del dataset de Omicron: una correcta asignación fue obtenida para el 98.2% de las secuencias BA.1* (n = 2643), 97.7% para BA.2* (n = 2096), 96.5% para BA.4* (n = 57) y 98.1% para BA.5* (n = 161).

4.4 Propuesta de flujo de trabajo para la vigilancia genómica de SARS-CoV-2

Desde el inicio de la pandemia global de COVID-19, las secuencias genómicas de SARS-CoV-2 han sido compartidas en bases de datos públicas como EpiCoV/GISAID. Los genomas representados corresponden a muestras clínicas, ambientales y hospederos no-humanos. Muchos países como Estados Unidos y Reino Unido han llevado adelante distintos programas de vigilancia genómica procesando miles de secuencias al mes a modo de caracterizar las variantes circulantes de SARS-CoV-2 (33,34). A pesar de que la OMS estableció la vigilancia genómica de SARS-CoV-2 como una prioridad (35), la implementación de los sistemas de vigilancia de variantes en regiones de menores recursos ha sido limitada, debido en parte a los requerimientos de infraestructura para algunas de las

plataformas, altos costos de los reactivos y compromiso de las partes involucradas. Este aspecto se ve reflejado en gran medida en las disparidades de la representación de genomas entre los distintos continentes (**Figura 5**).

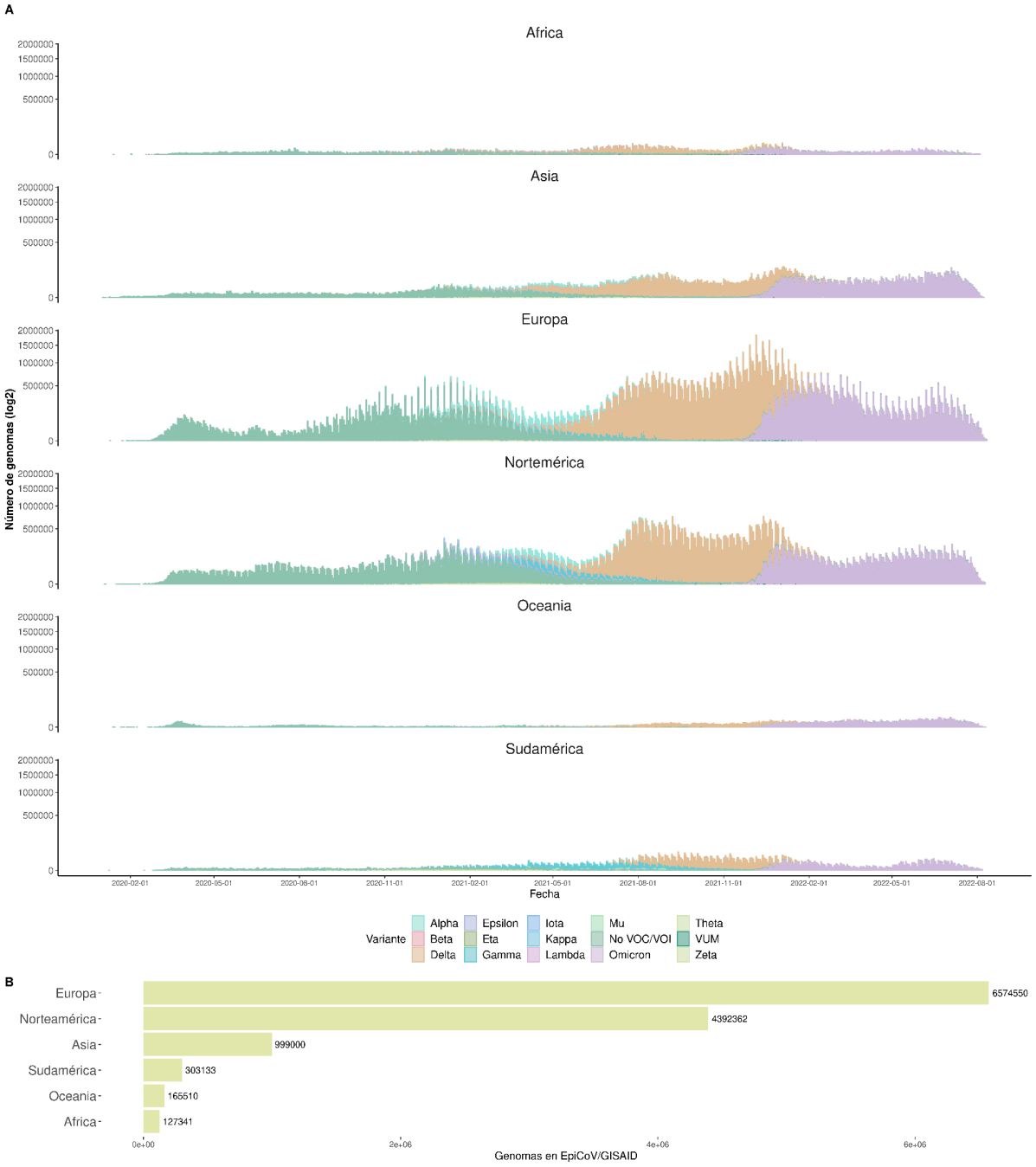

Figura 5: Representación continental de las secuencias genómicas de SARS-CoV-2 en EpiCoV/GISAID desde enero del 2020 hasta agosto del 2022. A) Distribución de variantes por continente. B) Número absoluto de genomas depositados por continente.

En el presente trabajo se propone un flujo de trabajo que asista en la toma de decisiones respecto a la secuenciación de SARS-CoV-2 en el contexto de la vigilancia

genómica, el cual representa un punto intermedio entre las metodologías de *q*PCR y WGS. En este flujo de trabajo el ARN residual obtenido de las muestras positivas para COVID-19 son recibidas en el laboratorio de secuenciación. Las mismas pueden provenir de comunidades cerradas, infecciones irruptivas o vigilancia centinela. Las mismas son relevadas de forma masiva a través de la secuenciación del gen S para la determinación de la variante correspondiente. Solamente las muestras que alcanzan un umbral mínimo de completitud y profundidad de secuenciación son tomadas en cuenta para el reporte de resultados. Si las mismas no son asignadas o presentan cambios que indiquen la presencia de una nueva variante del gen, las mismas son remitidas para la secuenciación de su genoma completo para su caracterización (**Figura 6**).

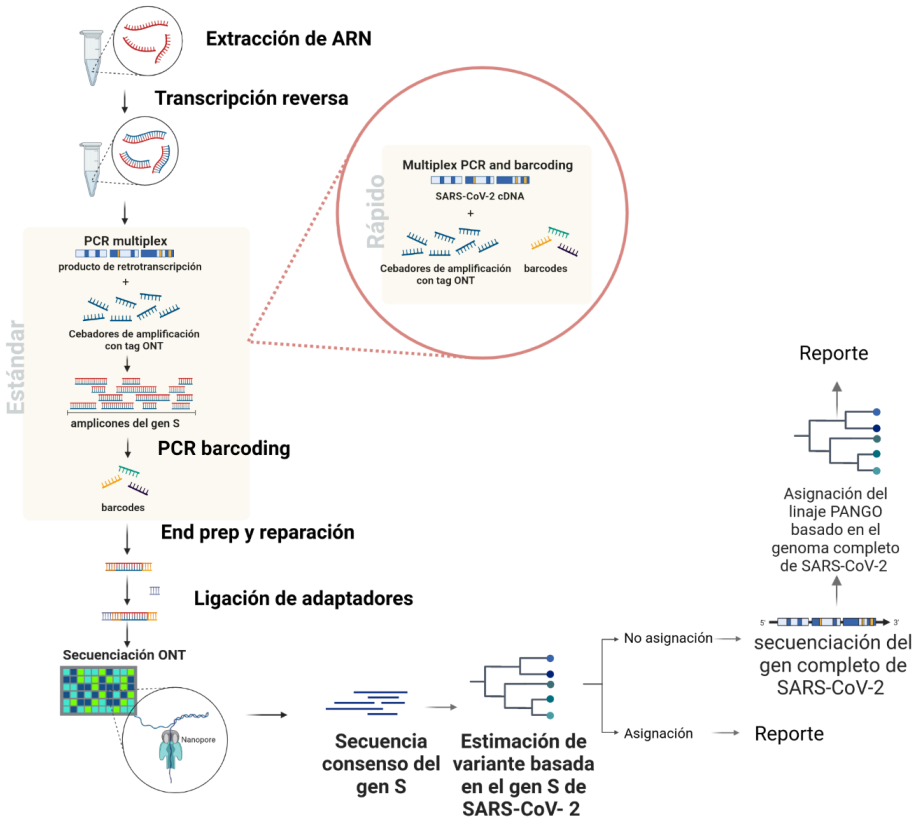

Figura 6: Propuesta de esquema de vigilancia de SARS-CoV-2 basada en la secuenciación gen S. Se muestra el esquema de amplificación e indexado de muestras con el método estándar y con el método rápido.

Este procedimiento tiene el objetivo de optimizar recursos para la vigilancia molecular de SARS-CoV-2. El costo estimado por muestra utilizando el protocolo rápido se estimó en el entorno de los USD 6, en contraste con los aproximadamente USD 15 estimados para el genoma completo. El tiempo estimado de ejecución del protocolo es menor a 20 hs

para el caso de la secuenciación del gen S para 96 muestras, mientras que se estima en el entorno a las 48 hs para la misma cantidad de genomas completos. Las principales limitaciones de la utilización del gen S para estudios de relevamiento de variantes, es que se obtiene un grado menor de resolución filogenética en comparación con el genoma completo y la imposibilidad de identificar variantes recombinantes.

5. Discusión

El gen que codifica para la proteína S de SARS-CoV-2 presenta un rol protagónico en la interacción con la célula del hospedero. Asimismo, se ha acumulado evidencia que las distintas variantes virales son capaces de evadir la respuesta inducida por las vacunas a través de la aparición de mutaciones en el gen S asociados a la región RBD (36). La OMS declaró el fin de la emergencia sanitaria por COVID-19 a principios del 2023, sin embargo aún se registran muertes como consecuencia de la infección y se espera que el virus continúe adquiriendo nuevas mutaciones que podrían afectar la efectividad a largo plazo de las actuales vacunas. Es por esto que resulta conveniente mantener los sistemas de vigilancia genómica centinela a modo de poder detectar de forma temprana el surgimiento de una variante que potencialmente escape al sistema inmune de poblaciones inmunizadas. En este sentido, el desarrollo de metodologías costo-efectivas, particularmente en aquellas regiones con menores recursos, puede asistir en respecto al relevamiento inicial de muestras de SARS-CoV-2, previo a la caracterización del genoma completo. Con esta finalidad, en esta sección se ha propuesto la utilización de la secuenciación del gen S para la determinación de variantes y como guía para determinar si una muestra pasa a la etapa de caracterización del genoma completo en el contexto de los sistemas de vigilancia genómica. Para ello se implementó un protocolo para el relevamiento basado en la secuenciación del gen S utilizando la plataforma portable y de bajo costo de ONT. La ventaja frente a otros protocolos radica que con la versión rápida del mismo, el paso de amplificación e indexado de las muestras se realiza en un único tubo, reduciendo los costos y disminuyendo la posibilidad de contaminaciones cruzadas, así como los tiempos de manipulación de las muestras. Varias estrategias de caracterización basadas en el gen S de SARS-CoV-2 han surgido a lo largo de los últimos años. Una de éstas estrategias consistió en la generación de amplicones solapantes del ectodominio del gen S con la plataforma Illumina. La misma se utilizó para tipificar entre el 10-50% de los casos de COVID-19 de Austria desde enero hasta junio del 2021 (18). La estrategia HiSpike fue desarrollada utilizando un protocolo de tres pasos con un tiempo de ejecución de aproximadamente 30 hs. La misma consiste en una RT-PCR1 y PCR2, seguida

de la generación de la biblioteca de secuenciación para la plataforma MiSeq (37). Distintos métodos fueron desarrollados en base a la secuenciación de Sanger (38–44) y también en base a la plataforma ONT (45). Sin embargo, los pasos implementados en los protocolos de este trabajo presentan el potencial para llevar adelante la vigilancia mutacional del gen S con requerimientos mínimos de infraestructura tanto de un número reducido de muestras (12 a 96) hasta cientos de muestras en combinación con estrategias de doble indexado (46) o indexado personalizados (45). Debido a que los cebadores utilizados amplifican diferentes regiones del gen S, un aspecto a tener en cuenta para su potencial aplicación, es el monitoreo de la amplificación de cada uno de los amplicones a modo de detectar el posible fallo de los mismos y por lo tanto fallos en la clasificación de determinada variante. Otro de los aspectos a tener en cuenta es la obtención de secuencias consenso con una completitud y profundidad de secuenciación adecuada para la asignación. Es recomendable la inclusión de controles previamente caracterizados mediante WGS para el monitoreo de la profundidad requerida de secuenciación para obtener resultados con alto grado de confianza. En este sentido, la generación de información confiable, de bajo costo y a gran escala es un aspecto clave para continuar con la vigilancia de variantes de SARS-CoV-2. El desarrollo previo de herramientas de análisis específicas que consideren la información contenida solamente en el gen S ha sido un factor importante para el desarrollo de este trabajo, ya que de este modo se reduce la generación de resultados ambiguos. Esto es porque las herramientas utilizadas para la asignación de linajes PANGO a partir de genomas de SARS-CoV-2 están entrenados con un conjunto de datos de genomas que han sido designados a un linaje PANGO en base a la información contenida a lo largo de todo el genoma. La herramienta *hedgehog* introduce el concepto de “set de linajes” para identificar aquellos linajes que comparten la misma secuencia del gen S (20). Por este motivo el desarrollo de protocolos de secuenciación del gen S como los descritos en este trabajo, en combinación con herramientas dedicadas para el análisis de los mismos presentan una potencial alternativa de bajo costo para el seguimiento de variantes de SARS-CoV-2.

En los últimos años muchas de las decisiones sobre temas de salud pública fueron determinadas en base a el monitoreo en tiempo real de la diversidad genética de SARS-CoV-2. Como consecuencia, una ola masiva de datos genómicos fue generada en respuesta a la vigilancia genómica a escala global. Las regiones de altos ingresos como Europa y Norteamérica han liderado en términos de la cantidad de secuencias depositadas en la base de datos EpiCoV/GISAID, contribuyendo con 6.6 millones y 4.4 millones de secuencias hasta agosto del 2022, respectivamente. Las demás regiones (Sudamérica, Asia,

África y Oceanía) han contribuido colectivamente con 1.6 millones de secuencias, evidenciando la disparidad en las capacidades de secuenciación a nivel mundial. A pesar de que los costos de secuenciación han disminuído de manera considerable y las plataformas portátiles han aliviado los requerimientos de infraestructura, la secuenciación genómica es aún un desafío en países en desarrollo debido a los costos de reactivos, la logística para la obtención de los mismos y la presencia de personal especializado (47). En países de bajos y medianos ingresos, la mayor parte de la secuenciación se realiza en instituciones académicas, las cuales han realizado esfuerzos de secuenciación genómica de manera intermitente con tiempos de disponibilización de datos en muchos casos más lentos. A pesar de que la importancia de disponibilizar la información es reconocida por estas instituciones y la información es compartida a las autoridades encargadas de la salud pública, se ha establecido la discusión respecto a que los frutos de estos esfuerzos no serían volcados a estos países, ya que la menor capacidad de análisis de los datos obstaculiza la generación de publicaciones de alto impacto, así como de patentes y acceso subsidios internacionales (48).

6. Material suplementario

Tabla S1: Descripción de las muestras de ARN analizadas.

ID de la muestra	Clado Nextstrain	Linaje PANGO	Ct diagnóstico	Departamento	Fecha de diagnóstico	Fecha de recolección
CUY1-000006	20B	P.2	18	Artigas	2021/03/04	2021/03/04
CUY12-002713	20J (Gamma, V3)	P.1	15.3	Montevideo	2021/06/07	2021/06/06
CUY12-002716	20J (Gamma, V3)	P.1	16.8	Montevideo	2021/06/07	2021/06/06
CUY12-002721	20J (Gamma, V3)	P.1	19.9	Montevideo	2021/06/07	2021/06/05
CUY12-002725	20J (Gamma, V3)	P.1	15.9	Montevideo	2021/06/07	2021/06/05
CUY16-003495	20J (Gamma, V3)	P.1	14.8	S/D	2021/02/19	S/D
CUY16-003498	20B	P.6	15.8	Montevideo	2021/02/18	S/D
CUY16-003501	20B	P.6	16.5	Montevideo	2021/02/17	S/D
CUY16-003536	20B	P.6	19.6	Montevideo	2021/01/29	S/D
CUY16-003538	20B	P.6	17.7	Montevideo	2021/01/28	S/D
CUY16-003792	21A (Delta)	B.1.617.2	15.3	Montevideo	2021/07/07	2021/07/05
CUY16-003792	21A (Delta)	B.1.617.2	15.3	Montevideo	2021/07/07	2021/07/05
CUY16-003792	21A (Delta)	B.1.617.2	15.3	Montevideo	2021/07/07	2021/07/05
CUY16-003792	21A (Delta)	B.1.617.2	15.3	Montevideo	2021/07/07	2021/07/05
CUY16-003800	20H (Beta, V2)	B.1.351	16.7	Montevideo	2021/07/07	2021/07/06
CUY16-003801	20H (Beta, V2)	B.1.351	17.7	Montevideo	2021/07/07	2021/07/06
CUY16-003802	20H (Beta, V2)	B.1.351	13.5	Montevideo	2021/07/07	2021/07/06
CUY16-003803	20H (Beta, V2)	B.1.351	15.2	Montevideo	2021/07/07	2021/07/06
CUY16-003804	20H (Beta, V2)	B.1.351	14.5	Montevideo	2021/07/07	2021/07/06
CUY17-003849	20I (Alpha, V1)	B.1.1.7	21	Canelones	2021/07/07	2021/07/07
CUY17-003866	20I (Alpha, V1)	B.1.1.7	25.8	S/D	2021/07/12	S/D
CUY17-003867	20I (Alpha, V1)	B.1.1.7	27.9	S/D	2021/07/12	S/D
CUY17-003892	21I (Delta)	AY.26	15.7	Montevideo	2021/07/14	S/D
CUY17-003892	21I (Delta)	AY.26	15.7	Montevideo	2021/07/14	S/D
CUY17-003892	21I (Delta)	AY.26	15.7	Montevideo	2021/07/14	S/D
CUY17-003892	21I (Delta)	AY.26	15.7	Montevideo	2021/07/14	S/D
CUY17-003893	21A (Delta)	B.1.617.2	15.7	Montevideo	2021/07/14	S/D
CUY17-003893	21A (Delta)	B.1.617.2	15.7	Montevideo	2021/07/14	S/D
CUY17-003893	21A (Delta)	B.1.617.2	15.7	Montevideo	2021/07/14	S/D
CUY17-003893	21A (Delta)	B.1.617.2	15.7	Montevideo	2021/07/14	S/D
CUY17-003901	21J (Delta)	AY.122	10	Montevideo	2021/07/09	2021/07/09
CUY17-003909	21A (Delta)	B.1.617.2	18.4	Montevideo	2021/07/14	2021/07/12
CUY17-003910	21J (Delta)	AY.122	18.6	Montevideo	2021/07/14	2021/07/14
CUY17-003939	20J (Gamma, V3)	P.1	14.8	Montevideo	2021/07/14	2021/07/13
CUY18-004019	21H (Mu)	B.1.621	22.2	S/D	S/D	S/D

CUY18-004022	21H (Mu)	B.1.621	20.2	S/D	S/D	S/D
CUY18-004023	21H (Mu)	B.1.621	20.8	S/D	S/D	S/D
CUY18-004026	21H (Mu)	B.1.621.1	22.3	S/D	2021/07/16	
CUY19-004183	21G (Lambda)	C.37	20.2	S/D	S/D	S/D
CUY23-004816	21J (Delta)	AY.20	19	Montevideo	2021/08/22	2021/08/22
CUY23-004816	21J (Delta)	AY.20	19	Montevideo	2021/08/22	2021/08/22
CUY23-004816	21J (Delta)	AY.43	19	Montevideo	2021/08/22	2021/08/22
CUY23-004816	21J (Delta)	AY.43	19	Montevideo	2021/08/22	2021/08/22
CUY24-004876	21J (Delta)	AY.25.1	15	Montevideo	2021/08/23	S/D
CUY24-004954	21J (Delta)	AY.99.2	15.3	Montevideo	S/D	S/D
CUY24-004954	21J (Delta)	AY.99.2	15.3	Montevideo	S/D	S/D
CUY24-004954	21J (Delta)	AY.99.2	15.3	Montevideo	S/D	S/D
CUY24-004954	21J (Delta)	AY.99.2	15.3	Montevideo	S/D	S/D
CUY4-000486	20B	P.2	1	Montevideo	S/D	2021/04/03
CUY42-007286	21K (Omicron)	BA.1	19	Canelones	2022/01/04	2022/01/04
CUY42-007288	21K (Omicron)	BA.1	19	Canelones	2022/01/04	2022/01/04
CUY42-007289	21K (Omicron)	BA.1	20	Montevideo	2022/01/04	2022/01/04
CUY42-007291	21K (Omicron)	BA.1	22	San José	2022/01/04	2022/01/04
CUY42-007294	21K (Omicron)	BA.1.1	18	San José	2022/01/04	2022/01/04
CUY5-000547	20B	P.6	17	Montevideo	S/D	2021/04/10
CUY6-001233	21G (Lambda)	C.37.1	23.2	Canelones	2021/04/19	2021/04/18
CUY8-001646	21G (Lambda)	C.37.1	19.4	Canelones	2021/04/17	2021/04/17

S/D sin datos

Tabla S2: Cebadores *forward* y *reverse* para la amplificación del gen S de SARS-CoV-2 el formato pool A (pares) y B (impares)

ONT_Sseq_1_LEFT	**TTTCTGTTGGTGCTGATATTGC ACAAAAGAAAATGACTCTAAAGAGGGTTT**
ONT_Sseq_1_RIGHT	**ACTTGCCTGTCGCTCTATCTTC ACTCTGAACTCACTTTCCATCCAAC**
ONT_Sseq_3_LEFT	**TTTCTGTTGGTGCTGATATTGC AGAGTCCAACCAACAGAATCTATTGT**
ONT_Sseq_3_RIGHT	**ACTTGCCTGTCGCTCTATCTTC ACCTGTGCCTGTTAAACCATTGA**
ONT_Sseq_5_LEFT	**TTTCTGTTGGTGCTGATATTGC CAACTTACTCCTACTTGGCGTGT**
ONT_Sseq_5_RIGHT	**ACTTGCCTGTCGCTCTATCTTC TGGAGCTAAGTTGTTTAACAAGCG**
ONT_Sseq_7_LEFT	**TTTCTGTTGGTGCTGATATTGC GGGCTATCATCTTATGTCCTTCCCT**
ONT_Sseq_7_RIGHT	**ACTTGCCTGTCGCTCTATCTTC AGGTGTGAGTAAACTGTTACAAACAAC**
ONT_Sseq_2_LEFT	**TTT TGTTGGTGCTGATATTGC ACACGTGGTGTTTATTACCC TGAC**
ONT_Sseq_2_RIGHT	**ACTTGCCTGTCGCTCTATCTTC GCAACACAGTTGCTGATTCTCTTC**
ONT_Sseq_4_LEFT	**TTTCTGTTGGTGCTGATATTGC CCAGCAACTGTTTGTGGACCTA**
ONT_Sseq_4_RIGHT	**ACTTGCCTGTCGCTCTATCTTC TGTGTACAAAAACTGCCATATTGCA**
ONT_Sseq_6_LEFT	**TTTCTGTTGGTGCTGATATTGC TTGCCTTGGTGATATTGCTGCT**
ONT_Sseq_6_RIGHT	**ACTTGCCTGTCGCTCTATCTTC TGCCAGAGATGTCACCTAAATCAA**

ONT_Sseq_8_LEFT	**TTTCTGTTGGTGCTGATATTGC TGCTGTAGTTGTCTCAAGGGCT**
ONT_Sseq_8_RIGHT	**ACTTGCCTGTCGCTCTATCTTC ACGAAAGCAAGAAAAAGAAGTACGC**

Tabla S3: Resultados de la secuenciación del gen S a partir de muestras de ARN de SARS-CoV-2

		Genome			Standard-S						Fast-S					
Sample	**biosample**	**WGS reads**	**Nextclade**	**PANGO lineage**	**Barcode**	**Std-S reads**	**Depth**	**Completeness**	**hedgehog**	**set_description**	**Barcode**	**Fast-S reads**	**Depth**	**Completeness**	**hedgehog**	**set_description**
CUY1-000006	SAMN31488805	79.526	20B	P.2	barcode83	318.857	534,8	98,97	P.2	P.2	barcode81	30.004	384,3	98,97	P.2	P.2
CUY12-002713	SAMN31488806	55.916	20J (Gamma, V3)	P.1	barcode45	96.014	534,5	98,97	P.1	P.1_1	barcode43	92.921	487,6	98,95	P.1	P.1_1
CUY12-002716	SAMN31488807	42.759	20J (Gamma, V3)	P.1	barcode57	86.224	533,9	98,97	P.1	P.1_1	barcode55	101.068	485,9	98,95	P.1	P.1_1
CUY12-002721	SAMN31488808	51.674	20J (Gamma, V3)	P.1	barcode69	5.405	428,8	98,93	P.1	P.1_1	barcode67	6.631	292,0	87,88	P.1	P.1_1
CUY12-002725	SAMN31488809	50.346	20J (Gamma, V3)	P.1	barcode81	77.297	533,8	98,95	P.1	P.1_1	barcode79	142.683	532,8	98,95	P.1	P.1_1
CUY16-003496	SAMN31488810	10.523	20B	P.2	barcode71	62.622	531,0	98,97	P.2	P.2	barcode69	7.533	327,0	71,53	P.2	P.2
CUY16-003498	SAMN31488811	14.866	20B	P.6	barcode36	24.253	499,3	98,97	P.6	P.6	barcode34	2.132	156,6	66,68	A	A_1
CUY16-003501	SAMN31488812	17.526	20B	P.6	barcode48	48.057	532,1	98,97	P.6	P.6	barcode46	33.906	366,2	98,97	P.6	P.6
CUY16-003536	SAMN31488813	17.579	20B	P.6	barcode60	25.024	530,5	98,97	P.6	P.6	barcode58	3.970	219,8	68,32	A	A_1
CUY16-003538	SAMN31488814	23.725	20B	P.6	barcode72	5.522	336,8	98,97	P.6	P.6	barcode70	6.314	211,4	71,41	A	A_1
CUY16-003792	SAMN31488815	12.612	21A (Delta)	B.1.617.2	barcode71	88.338	339,4	71,72	AY.30	AY.30	barcode62	5.074	457,3	88,19	B.1.617.2	B.1.617.2_7
CUY16-003800	SAMN31488816	14.559	20H (Beta, V2)	B.1.351	barcode68	146	15,3	18,59	A	A_23	barcode66	170.148	512,7	98,80	B.1.351	B.1.351
CUY16-003801	SAMN31488817	13.649	20H (Beta, V2)	B.1.351	barcode92	37.334	519,0	98,80	B.1.351	B.1.351	barcode90	33.130	284,9	87,71	A	A_1
CUY16-003802	SAMN31488818	22.218	20H (Beta, V2)	B.1.351	barcode80	348.485	533,7	98,80	B.1.351	B.1.351	barcode78	185.166	527,8	98,78	B.1.351	B.1.351
CUY16-003803	SAMN31488819	12.872	20H (Beta, V2)	B.1.351	barcode21	132.790	533,5	98,80	B.1.351	B.1.351	barcode19	50.019	430,0	98,80	B.1.351	B.1.351
CUY16-003804	SAMN31488820	15.395	20H (Beta, V2)	B.1.351	barcode09	4.258	262,8	98,80	B.1.351	B.1.351	barcode07	1.734	111,1	43,02	A	A_23
CUY17-003849	SAMN31488821	37.374	20I (Alpha, V1)	B.1.1.7	barcode08	147.783	532,5	98,78	B.1.1.7	B.1.1.7_1	barcode06	14.087	461,2	98,78	B.1.1.7	B.1.1.7_1
CUY17-003866	SAMN31488822	13.967	20I (Alpha, V1)	B.1.1.7	barcode20	676	33,8	42,84	A	A_1	barcode18	31.301	174,2	71,49	A	A_1
CUY17-003867	SAMN31488823	11.691	20I (Alpha, V1)	B.1.1.7	barcode32	151	4,2	0,00	Unassigned	Unassigned	barcode30	48.735	313,5	71,49	A	A_1
CUY17-003892	SAMN31488824	38.353	21I (Delta)	AY.26	barcode95	116.610	367,3	83,72	B.1.617.2	B.1.617.2_7	barcode08	145.978	449,5	88,30	B.1.617.2	B.1.617.2_7
CUY17-003893	SAMN31488825	34.808	21A (Delta)	B.1.617.2	barcode59	7.707	330,7	83,85	AY.48	AY.48	barcode20	278.482	471,9	88,28	AY.48	AY.48
CUY17-003901	SAMN31488826	13.986	21J (Delta)	AY.122	barcode22	30.000	486,6	88,26	B.1.617.2	B.1.617.2_2	barcode03	16.000	471,0	88,21	B.1.617.2	B.1.617.2_2
CUY17-003909	SAMN31488827	12.572	21A (Delta)	B.1.617.2	barcode34	3.000	11,4	14,83	A	A_1	barcode15	7.500	210,4	83,15	B.1.617.2	B.1.617.2_2
CUY17-003910	SAMN31488828	14.056	21J (Delta)	AY.122	barcode58	7.500	339,7	82,08	A	A_11	barcode39	10.000	346,1	87,40	B.1.617.2	B.1.617.2_2
CUY17-003939	SAMN31488829	50.819	20J (Gamma, V3)	P.1	barcode33	7.707	488,2	98,93	P.1	P.1_1	barcode31	29.566	418,0	98,91	P.1	P.1_1
CUY18-004019	SAMN31488830	10.835	21H (Mu)	B.1.621	barcode58	37.398	512,4	98,95	B.1.621	B.1.621	barcode56	33.026	387,4	98,93	AY.33.2	AY.33.2
CUY18-004022	SAMN31488831	13.124	21H (Mu)	B.1.621	barcode70	77.392	532,1	98,93	AY.33.2	AY.33.2	barcode68	28.136	251,3	87,82	AY.33.2	AY.33.2
CUY18-004023	SAMN31488832	16.371	21H (Mu)	B.1.621	barcode82	25.329	496,9	98,91	AY.33.2	AY.33.2	barcode80	69.169	417,5	98,89	B.1.617.2	B.1.617.2_2
CUY18-004026	SAMN31488833	4.898	21H (Mu)	B.1.621.1	barcode93	6.056	479,9	98,89	B.1.621.1	B.1.621.1	barcode54	133.724	410,7	98,82	AY.30	AY.30
CUY19-004183	SAMN31488834	73.438	21G (Lambda)	C.37	barcode35	16.120	527,9	98,57	C.37	C.37	barcode33	4.968	295,6	71,53	C.37	C.37
CUY21-004441	SAMN31488835	29.349	21H (Mu)	B.1.621							barcode51	17.000	523,3	98,95	B.1.621	B.1.621
CUY23-004816	SAMN31488836	22.640	21J (Delta)	AY.20	barcode47	76.909	358,5	83,85	AY.20	AY.20	barcode32	56.577	418,7	87,82	AY.20	AY.20
CUY24-004876	SAMN31488837	31.373	21J (Delta)	AY.25.1	barcode46	21.000	371,5	88,24	A	A_11	barcode27	7.000	395,0	88,17	B.1.617.2	B.1.617.2_4
CUY24-004954	SAMN31488838	23.594	21J (Delta)	AY.99.2	barcode46	194.573	468,2	88,28	A	A_11	barcode44	74.859	414,8	87,79	A	A_11
CUY26-005235	SAMN31488839	17.810	21I (Delta)	B.1.617.2	barcode82	500	9,2	7,21	A	A_2	barcode63	15.000	363,5	83,32	B.1.617.2	B.1.617.2_3
CUY4-000486	SAMN31488840	22.255	20B	P.2	barcode12	106.885	534,3	98,97	P.2	P.2	barcode10	79.571	451,7	98,97	P.2	P.2
CUY42-007286	SAMN31488841	73.308	21K (Omicron)	BA.1	barcode36	6.500	157,9	73,26	A	A_1	barcode01	3.500	103,6	55,61	BA.2.12	BA.2.12
CUY42-007288	SAMN31488842	55.059	21K (Omicron)	BA.1	barcode48	8.000	223,7	75,44	BA.1.15	BA.1.15	barcode13	6.500	296,7	98,59	BA.1.15	BA.1.15
CUY42-007289	SAMN31488843	67.186	21K (Omicron)	BA.1	barcode60	15.500	329,6	84,96	BA.1.15	BA.1.15	barcode25	5.000	248,4	98,74	BA.1	BA.1_2
CUY42-007291	SAMN31488844	31.682	21K (Omicron)	BA.1	barcode72	3.500	68,5	49,16	A	A_1	barcode37	4.500	189,9	57,21	BA.2.12	BA.2.12
CUY42-007294	SAMN31488845	90.587	21K (Omicron)	BA.1.1	barcode84	9.500	276,1	84,96	BA.1	BA.1_4	barcode49	8.000	370,7	98,80	BA.1.1	BA.1.1_2
CUY5-000547	SAMN31488846	65.494	20B	P.6	barcode24	59.987	533,9	98,97	P.6	P.6	barcode22	95.108	476,4	98,97	P.6	P.6
CUY6-001233	SAMN31488847	19.595	21G (Lambda)	C.37.1	barcode47	36.472	530,8	98,57	C.37.1	C.37.1	barcode45	5.607	300,2	71,49	A	A_5
CUY8-001646	SAMN31488848	31.632	21G (Lambda)	C.37.1	barcode59	326.563	533,7	98,65	C.37.1	C.37.1	barcode57	46.296	424,9	98,55	C.37.1	C.37.1

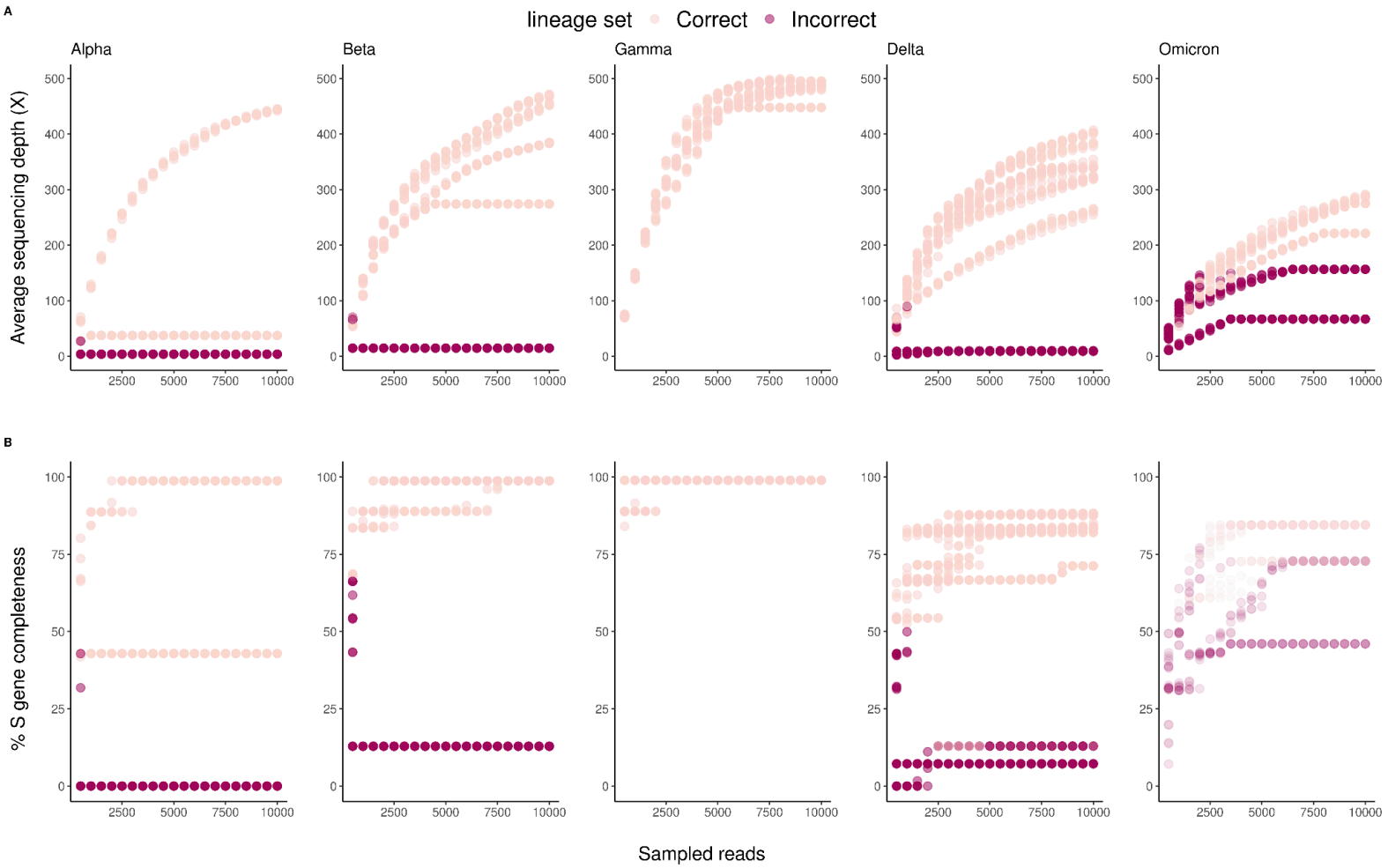

Figura S1: Asignación de linaje PANGO basado en el gen S de SARS-CoV-2 con la metodología estándar. Se realizó un muestreo al azar de lecturas de secuenciación de VOCs. A) Profundidad de secuenciación promedio vs. lecturas de secuenciación. B) Completitud del gen (%) vs. lecturas de secuenciación. Las asignaciones incorrectas fueron detectadas a menor profundidad de secuenciación y menor completitud del gen S.

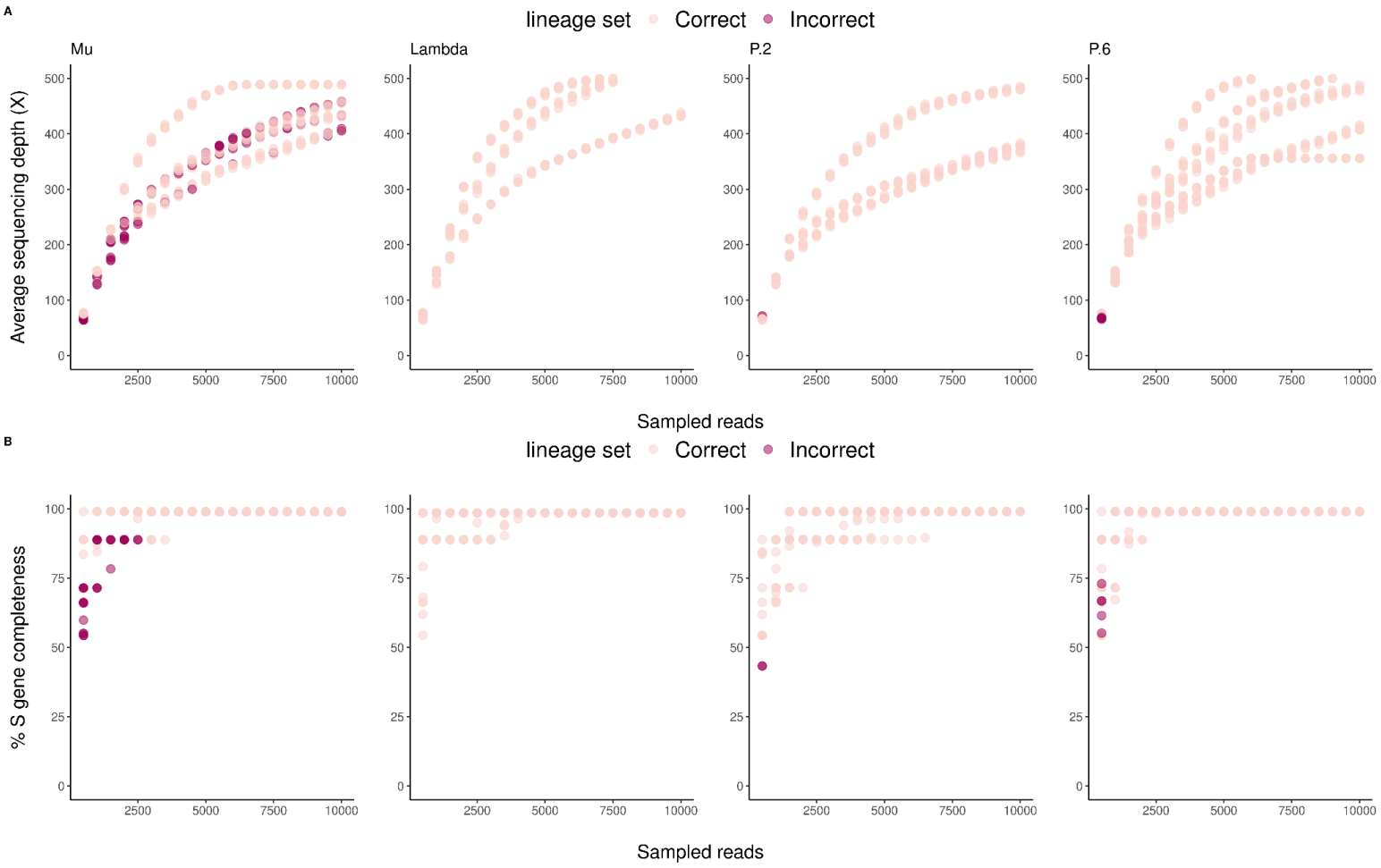

Figura S2: Asignación de linaje PANGO basado en el gen S de SARS-CoV-2 con la metodología estándar. Se realizó un muestreo al azar de lecturas de secuenciación de las muestras correspondientes a VOIs y no VOC/VOIs (P.2 y P.6). A) Profundidad de secuenciación promedio vs. lecturas de secuenciación. B) Completitud del gen (%) vs. lecturas de secuenciación. Las asignaciones incorrectas fueron detectadas a menor profundidad de secuenciación y menor completitud del gen S.

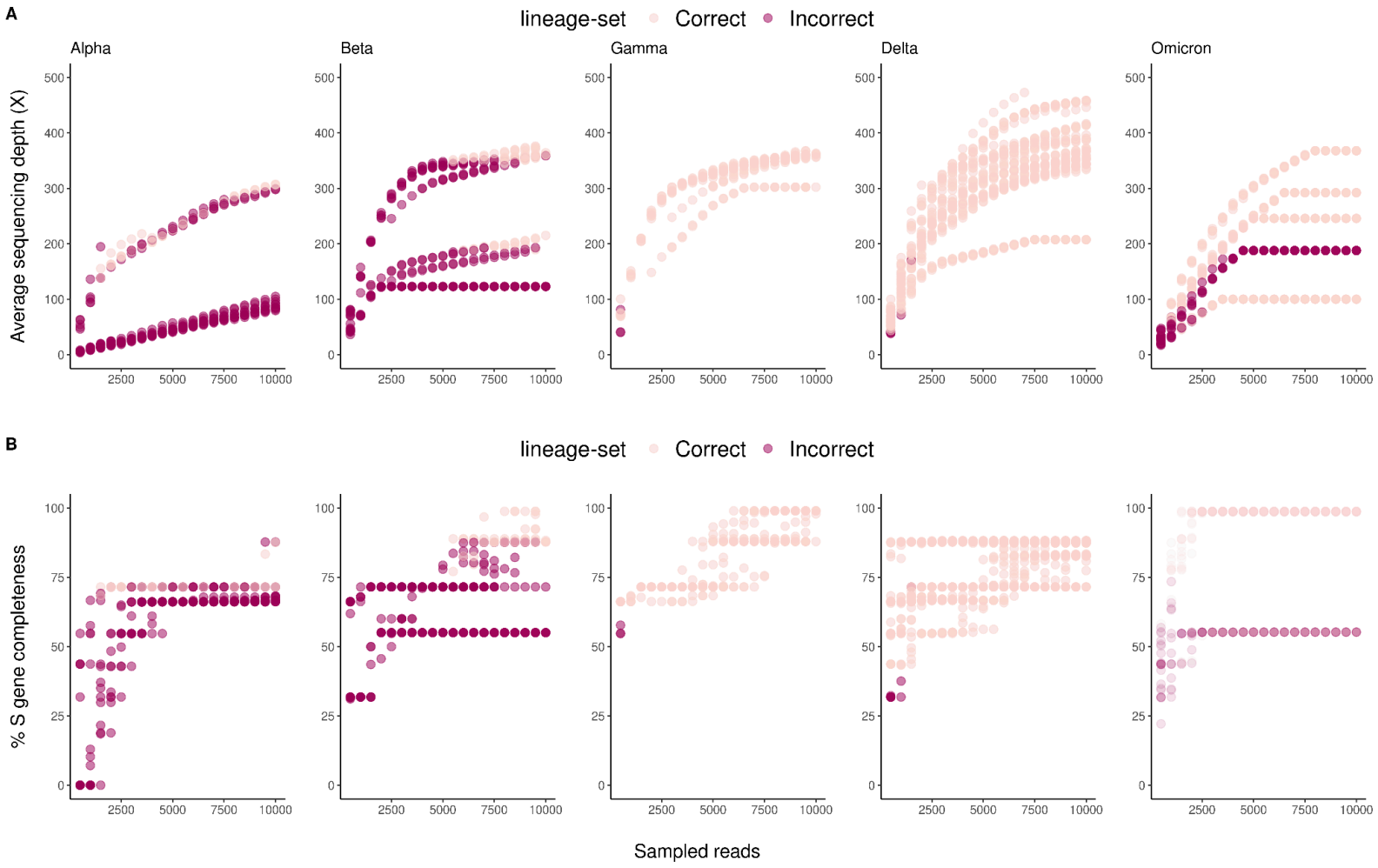

Figura S3: Asignación de linaje PANGO basado en el gen S de SARS-CoV-2 con la metodología rápida. Se realizó un muestreo al azar de lecturas de secuenciación de VOCs. A) Profundidad de secuenciación promedio vs. lecturas de secuenciación. B) Completitud del gen (%) vs. lecturas de secuenciación. Las asignaciones incorrectas fueron detectadas a menor profundidad de secuenciación y menor completitud del gen S.

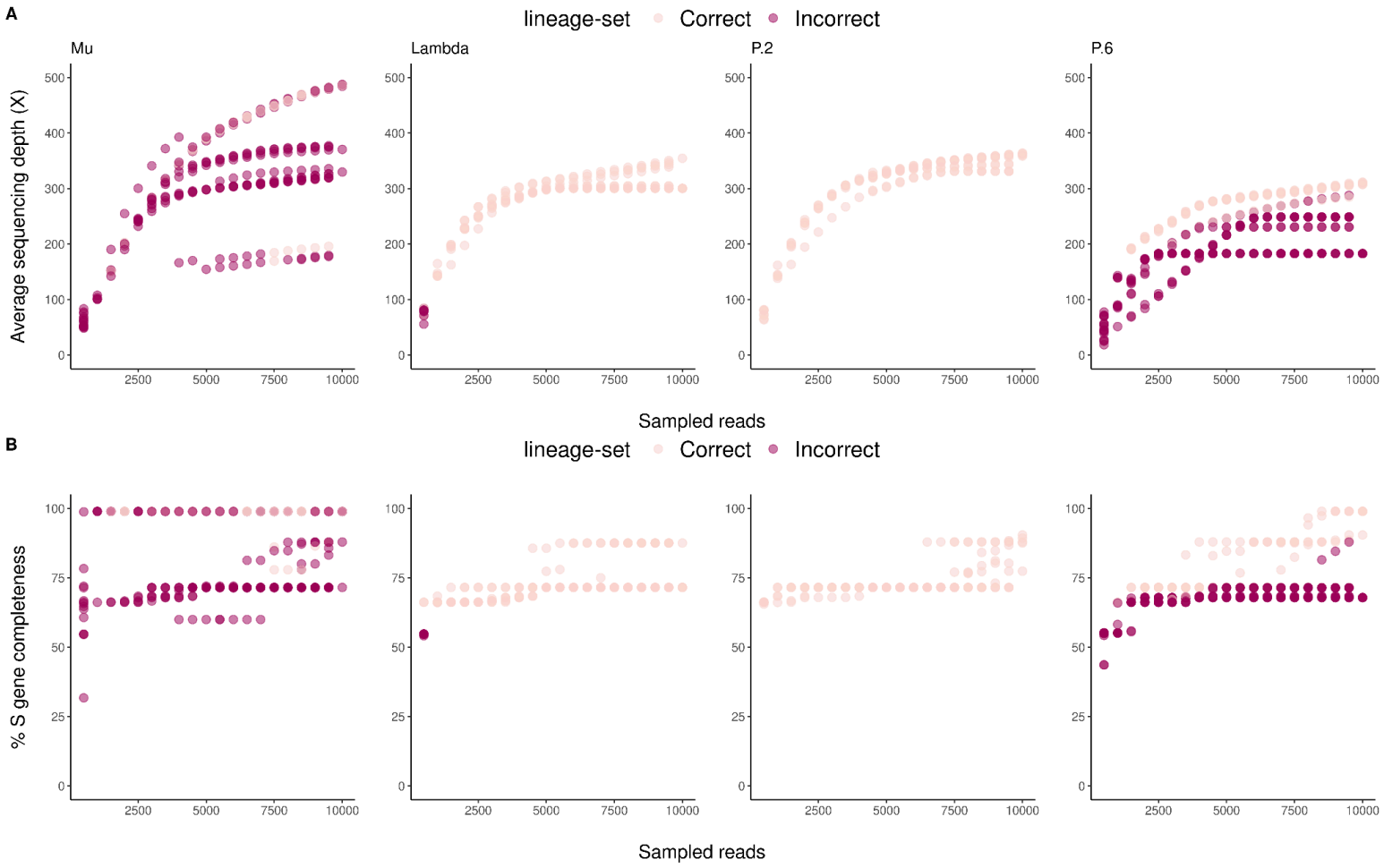

Figura S4: Asignación de linaje PANGO basado en el gen S de SARS-CoV-2 con la metodología rápida. Se realizó un muestreo al azar de lecturas de secuenciación de las muestras correspondientes a VOIs y no VOC/VOIs (P.2 y P.6). A) Profundidad de secuenciación promedio vs. lecturas de secuenciación. B) Completitud del gen (%) vs. lecturas de secuenciación. Las asignaciones incorrectas fueron detectadas a menor profundidad de secuenciación y menor completitud del gen S.

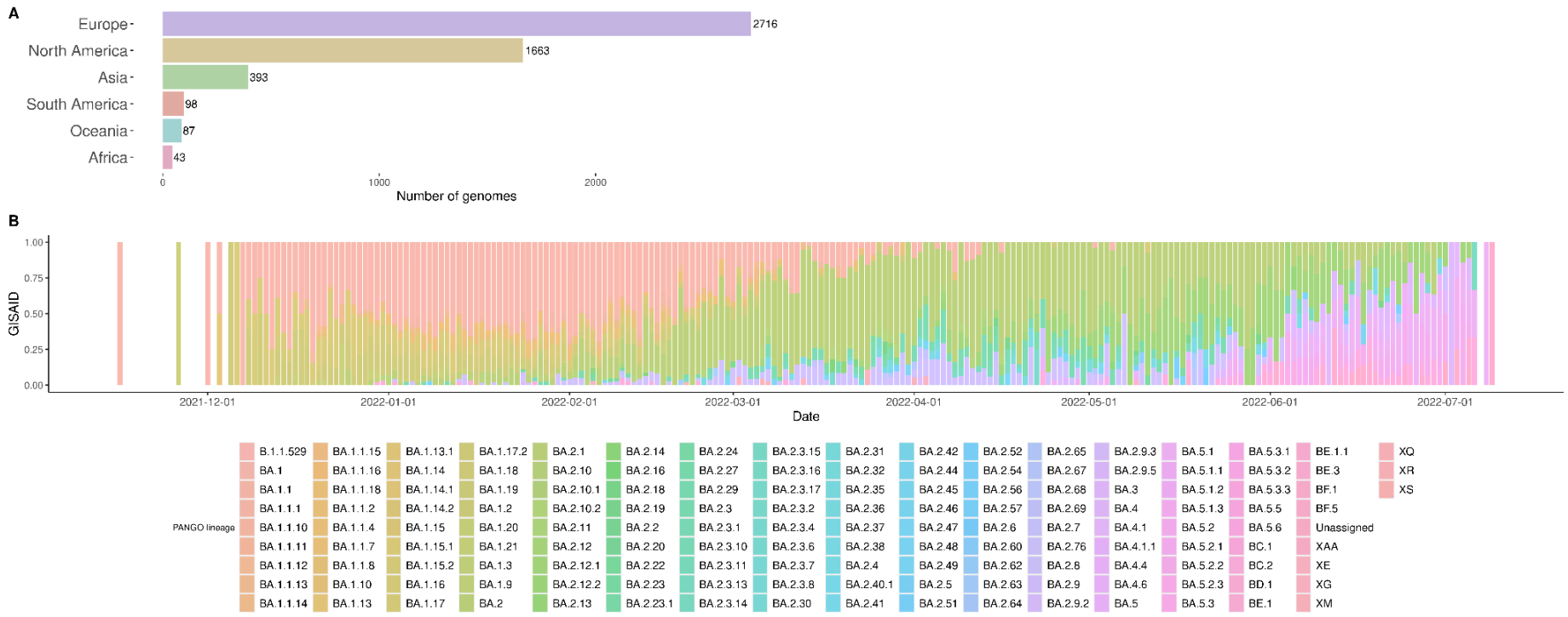

Figura S5: Dataset de Omicron obtenido de EpiCoV/GISAID (n = 5000). A) Representación continental de secuencias en el dataset. B)Abundancia relativa entre los distintos sublinajes muestreados de la base de datos en el intervalo de tiempo desde noviembre del 2020 hasta julio del 2021.

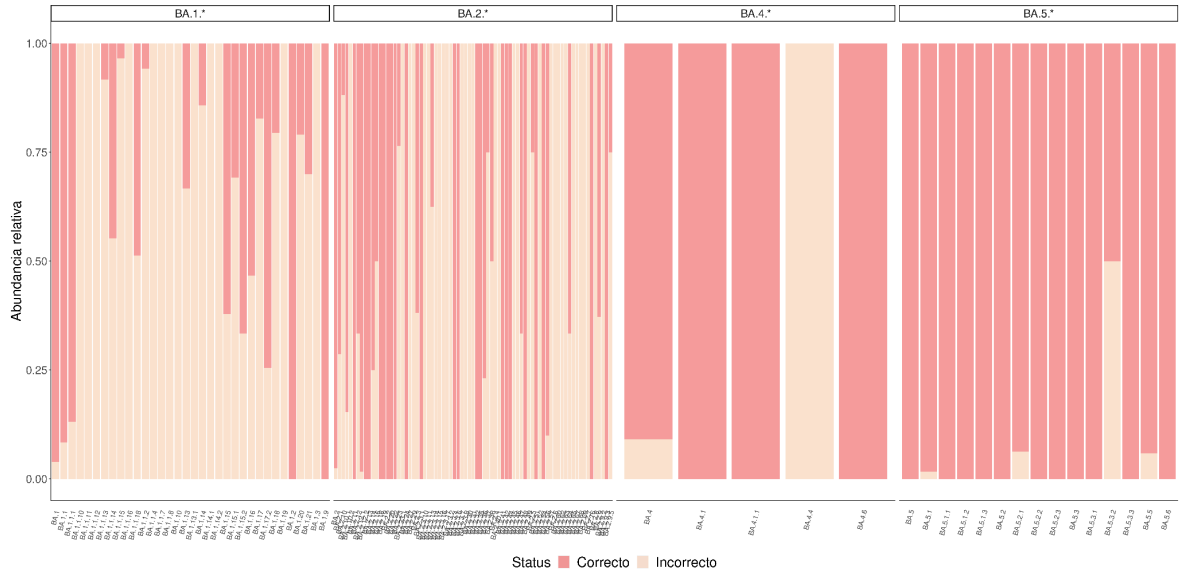

Figura S6: Comparación del linaje exacto basado en el genoma completo de SARS-CoV-2 y set de linaje obtenido con el gen S de un dataset de secuencias de la VOC Omicron (EPI_SET_220929wd). Se muestra las concordancias exactas en rojo y las discordancias en amarillo. Las asignaciones fueron consideradas correctas cuando el linaje del genoma fue encontrado en el rango de linajes PANGO de la nomenclatura de *hedgehog*. Si el linaje del genoma no fue encontrado en el set de linaje, se consideró una asignación incorrecta. Las asignaciones correctas presentaron un porcentaje de completitud de 97.2% ± 5.1% y las asignaciones correctas 94.2% ± 8.6%. Asimismo, 75% de las asignaciones obtenidas con *hedgehog* para BA.1* tuvieron una coincidencia exactas con el linaje PANGO del genoma completo, mientras que 85% para BA.2*, 95% para BA.4* y 97% para BA.5*.

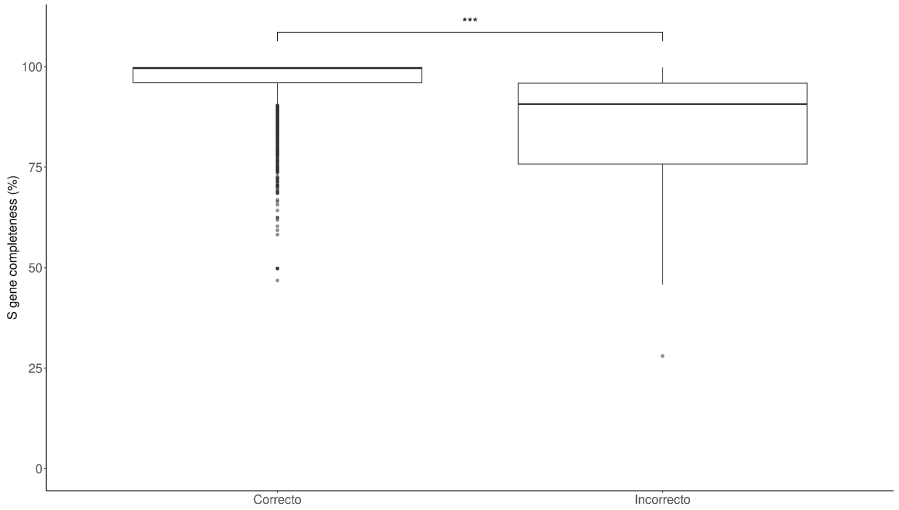

Figura S7: Diferencias entre asignaciones correctas e incorrectas del set de linaje PANGO del dataset de Omicron ($P \leq 0.001$). Se requiere un nivel elevado de completitud del gen S para la asignación correcta de los sublinajes de Omicron.

7. Referencias

1. WHO. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 [Internet]. 2020. Disponible en: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020
2. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 12 de marzo de 2020;579(7798):270-3.
3. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet. febrero de 2020;395(10224):565-74.
4. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 12 de marzo de 2020;579(7798):265-9.
5. Jackson CB, Farzan M, Chen B, Choe H. Mechanisms of SARS-CoV-2 entry into cells. Nat Rev Mol Cell Biol. enero de 2022;23(1):3-20.
6. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance [Internet]. 30 de marzo de 2017 [citado 24 de febrero de 2023];22(13). Disponible en: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2017.22.13.30494
7. Attwood SW, Hill SC, Aanensen DM, Connor TR, Pybus OG. Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. Nat Rev Genet. septiembre de 2022;23(9):547-62.
8. Tosta S, Moreno K, Schuab G, Fonseca V, Segovia FMC, Kashima S, et al. Global SARS-CoV-2 genomic surveillance: What we have learned (so far). Infect Genet Evol. marzo de 2023;108:105405.
9. Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-Stroh S, et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Eurosurveillance [Internet]. 13 de agosto de 2020 [citado 24 de febrero de 2023];25(32). Disponible en: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.32.2001410
10. Konings F, Perkins MD, Kuhn JH, Pallen MJ, Alm EJ, Archer BN, et al. SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse. Nat Microbiol. 9 de junio de 2021;6(7):821-3.
11. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 15 de julio de 2020;5(11):1403-7.

12. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol. 14 de septiembre de 2021;7(2):veab064.
13. WHO. Tracking SARS-CoV-2 variants. 2022.
14. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol. julio de 2021;19(7):409-24.
15. Zhang L, Jackson CB, Mou H, Ojha A, Peng H, Quinlan BD, et al. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. Nat Commun. 26 de noviembre de 2020;11(1):6013.
16. Jackson CB, Zhang L, Farzan M, Choe H. Functional importance of the D614G mutation in the SARS-CoV-2 spike protein. Biochem Biophys Res Commun. enero de 2021;538:108-15.
17. Rego N, Costábile A, Paz M, Salazar C, Perbolianachis P, Spangenberg L, et al. Real-Time Genomic Surveillance for SARS-CoV-2 Variants of Concern, Uruguay. Emerg Infect Dis. noviembre de 2021;27(11):2957-60.
18. Özkan E, Strobl MM, Novatchkova M, Yelagandula R, Albanese TG, Triska P, et al. High-throughput Mutational Surveillance of the SARS-CoV-2 Spike Gene [Internet]. Infectious Diseases (except HIV/AIDS); 2021 jul [citado 1 de septiembre de 2022]. Disponible en: http://medrxiv.org/lookup/doi/10.1101/2021.07.22.21259587
19. Tyson JR, James P, Stoddart D, Sparks N, Wickenhagen A, Hall G, et al. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore [Internet]. Genomics; 2020 sep [citado 1 de marzo de 2023]. Disponible en: http://biorxiv.org/lookup/doi/10.1101/2020.09.04.283077
20. O'Toole Á, Pybus OG, Abram ME, Kelly EJ, Rambaut A. Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. BMC Genomics. diciembre de 2022;23(1):121.
21. Freed NE, Vlková M, Faisal MB, Silander OK. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. Biol Methods Protoc. 1 de enero de 2020;5(1):bpaa014.
22. Salazar C, Ferrés I, Paz M, Costábile A, Moratorio G, Moreno P, et al. Fast and cost-effective SARS-CoV-2 variant detection using Oxford Nanopore full-length spike gene sequencing. Microb Genomics [Internet]. 18 de mayo de 2023 [citado 28 de mayo de 2023];9(5). Disponible en: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.001013
23. Brandt C, Krautwurst S, Spott R, Lohde M, Jundzill M, Marquet M, et al. poreCov-An Easy to Use, Fast, and Robust Workflow for SARS-CoV-2 Genome Reconstruction via Nanopore Sequencing. Front Genet. 28 de julio de 2021;12:711437.
24. Abath FGC, Melo FL, Werkhauser RP, Montenegro L, Montenegro R, Schindler HC. Single-Tube

Nested PCR Using Immobilized Internal Primers. BioTechniques. diciembre de 2002;33(6):1210-4.

25. Oxford Nanopore Technologies. wf-artic [Internet]. Disponible en: https://github.com/epi2me-labs/wf-artic
26. Áine O'Toole. Hedgehog [Internet]. 2023. Disponible en: https://github.com/cov-lineages/hedgehog
27. Martin Hölzer. president [Internet]. 2022. Disponible en: https://gitlab.com/RKIBioinformaticsPipelines/president
28. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 29 de enero de 2021;10(2):giab008.
29. Ahlmann-Eltze C, Patil I. ggsignif: R Package for Displaying Significance Brackets for «ggplot2» [Internet]. PsyArXiv; 2021 mar [citado 1 de marzo de 2023]. Disponible en: https://osf.io/7awm6
30. Aksamentov I, Roemer C, Hodcroft E, Neher R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw. 30 de noviembre de 2021;6(67):3773.
31. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 15 de noviembre de 2014;30(22):3276-8.
32. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd ed. 2016. Cham: Springer International Publishing : Imprint: Springer; 2016. 1 p. (Use R!).
33. CDC. SPHERES [Internet]. 2022. Disponible en: https://www.cdc.gov/coronavirus/2019-ncov/variants/spheres.html
34. COVID-19 Genomics UK. COG-UK [Internet]. Disponible en: https://www.cogconsortium.uk/
35. WHO. Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health [Internet]. 2021. Disponible en: https://www.who.int/publications/i/item/9789240018440
36. Willett BJ, Grove J, MacLean OA, Wilkie C, De Lorenzo G, Furnon W, et al. SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway. Nat Microbiol. 7 de julio de 2022;7(8):1161-79.
37. Fass E, Zizelski Valenci G, Rubinstein M, Freidlin PJ, Rosencwaig S, Kutikov I, et al. HiSpike Method for High-Throughput Cost Effective Sequencing of the SARS-CoV-2 Spike Gene. Front Med. 11 de enero de 2022;8:798130.
38. Lim HJ, Park MY, Jung HS, Kwon Y, Kim I, Kim DK, et al. Development of an efficient Sanger sequencing-based assay for detecting SARS-CoV-2 spike mutations. Lin B, editor. PLOS ONE. 14 de diciembre de 2021;16(12):e0260850.
39. Salles TS, Cavalcanti AC, da Costa FB, Dias VZ, de Souza LM, de Meneses MDF, et al. Genomic surveillance of SARS-CoV-2 Spike gene by sanger sequencing. Ito E, editor. PLOS ONE. 20 de enero de 2022;17(1):e0262170.

40. Bezerra MF, Machado LC, De Carvalho V do CV, Docena C, Brandão-Filho SP, Ayres CFJ, et al. A Sanger-based approach for scaling up screening of SARS-CoV-2 variants of interest and concern. Infect Genet Evol. agosto de 2021;92:104910.
41. Jørgensen TS, Pedersen MS, Blin K, Kuntke F, Salling HK, Marvig RL, et al. SpikeSeq: A rapid, cost efficient and simple method to identify SARS-CoV-2 variants of concern by Sanger sequencing part of the spike protein gene. J Virol Methods. febrero de 2023;312:114648.
42. Daniels RS, Harvey R, Ermetal B, Xiang Z, Galiano M, Adams L, et al. A Sanger sequencing protocol for SARS‐CoV‐2 S‐gene. Influenza Other Respir Viruses. noviembre de 2021;15(6):707-10.
43. Bloemen M, Rector A, Swinnen J, Ranst MV, Maes P, Vanmechelen B, et al. Fast detection of SARS-CoV-2 variants including Omicron using one-step RT-PCR and Sanger sequencing. J Virol Methods. junio de 2022;304:114512.
44. Ko K, Takahashi K, Nagashima S, E B, Ouoba S, Hussain MRA, et al. Mass Screening of SARS-CoV-2 Variants using Sanger Sequencing Strategy in Hiroshima, Japan. Sci Rep. 14 de febrero de 2022;12(1):2419.
45. Stüder F, Petit JL, Engelen S, Mendoza-Parra MA. Real-time SARS-CoV-2 diagnostic and variants tracking over multiple candidates using nanopore DNA sequencing. Sci Rep. 5 de agosto de 2021;11(1):15869.
46. Liou CH, Wu HC, Liao YC, Yang Lauderdale TL, Huang IW, Chen FJ. nanoMLST: accurate multilocus sequence typing using Oxford Nanopore Technologies MinION with a dual-barcode approach to multiplex large numbers of samples. Microb Genomics [Internet]. 1 de marzo de 2020 [citado 1 de marzo de 2023];6(3). Disponible en: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000336
47. Helmy M, Awad M, Mosa KA. Limited resources of genome sequencing in developing countries: Challenges and solutions. Appl Transl Genomics. junio de 2016;9:15-9.
48. Maxmen A. Why some researchers oppose unrestricted sharing of coronavirus genome data. Nature. 13 de mayo de 2021;593(7858):176-7.

Discusión general

La utilización de la genómica en el contexto clínico presenta el potencial de asistir decisiones respecto a la salud y aportar al conocimiento de distintas enfermedades infecciosas. Si bien, la utilización de las tecnologías de secuenciación se ha incrementado en los últimos años, la mayor limitación para su utilización es la necesidad de procedimientos automatizados y/o estandarizados, controles de calidad e incorporación de expertise bioinformática para la interpretación de los resultados (13). Otra de las limitaciones que se enfrenta está relacionada con los costos asociados; si bien no existen grandes requerimientos de infraestructura, ni de grandes inversiones para la adquisición de un dispositivo de secuenciación, los costos de los reactivos, consumibles y procesamiento bioinformático debe justificarse en beneficios adicionales a los pacientes y/o a la eficiencia/calidad en la obtención de resultados por parte de los laboratorios. En cuanto al tipo de plataforma, la utilización de tecnología de tercera generación de ONT tiene el potencial de convertirse en una práctica estándar en los laboratorios de microbiología, debido al bajo costo de inversión inicial y la variedad de herramientas de análisis disponibles (14–16).

En este trabajo se ha evaluado un brote causado por CR-*Kp* en un hospital público de la ciudad de Montevideo utilizando tecnología de secuenciación ONT. Luego del ensamblado *de novo* con la estrategia para lecturas largas (ONT) y un abordaje híbrido (ONT + Illumina), se observó que es posible la detección de determinantes de resistencia y factores virulencia e incluso la clasificación de las secuencias plasmídicas. Sin embargo, solamente con los genomas de alta resolución fue posible determinar a qué grupo clonal pertenecían. Esto se debe a que la designación del ST, requiere el *match* perfecto en los siete loci que se utiliza para la determinación del perfil alélico o MLST. Las actualizaciones más recientes en la química de secuenciación (v14) generan en principio, datos con un menor porcentaje de error asociado a la secuenciación suficientes para la generación de ensamblados de alta calidad. La combinación entre la lectura de ambas hebras de los fragmentos de ADN y la asignación de bases de muy alta precisión (o *super-accurate basecalling*) permitirían prescindir en un futuro de los datos de secuenciación de segunda generación para la obtención de ensamblados de alta resolución y por lo tanto, se espera que la asignación del ST pueda ser determinado con alta confianza solamente con datos ONT (9). Por otra parte, la utilización exclusiva de datos ONT permitió la obtención de ensamblados completos de los plásmidos asociados al brote intrahospitalario. Los plásmidos son los vectores más comunes de transferencia horizontal de genes y por lo tanto tienen un rol protagónico en la diseminación de genes

asociados a RAM. A partir de la obtención de la secuencia de los mismos se pudo detectar los determinantes de resistencia adquiridos en el contexto del brote nosocomial. Esto ejemplifica la potencialidad de la implementación de ONT en el laboratorio clínico, ya sea para la caracterización de microorganismos y/o genes, vigilancia epidemiológica o investigación de brotes (65,66).

Los resultados obtenidos a partir del brote de CR-*Kp* ST-11 fueron integrados a los estudios de RAM en el ambiente urbano de la ciudad de Montevideo. En el mismo se detectó la presencia de un MAG (por *metagenome-assembled genome*) obtenido de la zona oeste de Montevideo filogenéticamente relacionado con el clon de *K. pneumoniae* ST-11 del brote. Mediante una estrategia de mapeo de lecturas de secuenciación, se encontró que la abundancia relativa del plásmido que codifica para la carbapenemasa KPC-2 se encuentra correlacionada con la abundancia relativa del cromosoma de *K. pneumoniae* ST-11 en distintas muestras ambientales. Adicionalmente, se encontró una correlación negativa significativa entre la presencia del plásmido y la densidad de colectores de saneamiento. Esto indicaría una menor co-ocurrencia de *K. pneumoniae* ST-11 y del plásmido conteniendo el gen bla_{KPC-2} en sitios con mejor infraestructura de saneamiento (67).

Otras de las metodologías utilizadas en el laboratorio de microbiología es la aplicación de tecnologías de secuenciación para la generación de perfiles taxonómicos microbianos. Uno de los marcadores genéticos más utilizados para la caracterización de comunidades procariotas es el gen 16S del ARN*r*. Si bien estas metodologías han sido aplicadas desde hace varias décadas, con el surgimiento de las tecnologías de secuenciación masiva, los procedimientos se han simplificado y permitido el relevamiento de mayor cantidad de muestras en un único experimento. Esto vino acompañado del desarrollo de herramientas bioinformáticas para la manipulación de las secuencias generadas y la interpretación de los resultados obtenidos (68–70). Sin embargo, la resolución obtenida con estas tecnologías rara vez incluye niveles taxonómicos más allá del género. La aparición en el mercado de secuenciadores de tercera generación ha permitido la obtención de secuencias completas del gen 16S y el desarrollo de estudios de caracterización de comunidades microbianas de distintos ambientes con una resolución a nivel de especie (46,48,50,71,72). A pesar de las oportunidades que surgen de la obtención de secuencias más largas, aún no se han establecido criterios estandarizados para el análisis de las mismas. En los últimos años han surgido algunas herramientas como *NanoClust* y *EMU* (43,44). La primera realiza un paso de agrupamiento de las secuencias y generación del consenso, la cual es luego contrastada contra una base de datos para la clasificación taxonómica de la misma. En la

segunda, las lecturas son alineadas contra la base de datos y luego se aplican varias iteraciones de corrección de errores a modo de generar perfiles de abundancia relativa a nivel de especie. En este contexto se desarrolló *porefile*, una herramienta de clasificación taxonómica de lecturas de secuenciación obtenidas con plataformas de tercera generación como ONT. La misma presenta un módulo de preprocesamiento de las lecturas que incluye el demultiplexado y filtrado por calidad y tamaño. Luego se realiza una etapa de mapeo contra la última versión de la base de datos SILVA y las clasificaciones obtenidas para cada una de las lecturas de secuenciación son agrupadas utilizando el algoritmo del ancestro común más reciente en base al *score* de alineamiento. A partir del resultado inicial de clasificación a nivel de especie, la base de datos es reducida y el resultado de abundancia relativa es pulido a ese nivel taxonómico. El resultado final incluye las tablas de conteo y taxonomía correspondientes a la clasificación con la base de datos completa y los resultados pulidos a nivel. Los ensayos sobre lecturas de secuenciación simuladas muestran que *porefile* genera resultados similares a los de *EMU* tanto para comunidades de baja como alta complejidad. A su vez, los resultados obtenidos a nivel de género con datos de secuenciación de Illumina son comparables a los obtenidos con *porefile* y *EMU* para un conjunto de datos públicos de la microbiota intestinal humana. *Porefile* se encuentra públicamente disponible en https://github.com/microgenlab/porefile y sus módulos son manejados con el sistema *Nextflow* lo que asegura la automatización y reproducibilidad de cada uno de los procesos. Otra de las características más importantes de *porefile* es que los archivos de salida generados son compatibles con distintas herramientas de visualización y análisis de la diversidad de comunidades microbianas (73–75).

La aplicación de las metodologías de secuenciación ONT ha tenido un rol importante en la caracterización de las distintas variantes de SARS-CoV-2 tanto a nivel internacional como local. A través de la utilización de tecnología ONT se reportaron los primeros genomas de SARS-CoV-2 de Uruguay (76) y la primera detección de la VOC Alfa (77). También ha permitido la implementación de un sistema de vigilancia genómica basado en tecnología ONT que ha reportado a las autoridades la dinámica de circulación de las variantes del virus en el país durante la pandemia de COVID-19 (78,79). Asimismo, ha impulsado la generación de estrategias alternativas para la caracterización rápida de variantes y linajes de SARS-CoV-2 mediante la secuenciación del gen S (80). El desarrollo de estrategias costo-efectivas que asistan a las regiones de bajos recursos a generar información sobre la circulación de linajes o variantes constituye una de las alternativas para mantener la vigilancia de SARS-CoV-2. Si bien existen metodologías basadas en *q*PCR capaces de

distinguir entre distintos sublinajes de VOCs como Omicron, los métodos basados en secuenciación permiten obtener información de la secuencia completa del gen de interés. Esto determina que potenciales cambios de relevancia puedan ser detectados de forma temprana. En este contexto, se desarrolló una metodología de menor costo respecto a la secuenciación del genoma completo como una herramienta alternativa para la vigilancia de SARS-CoV-2. La misma se basa en la secuenciación del gen S utilizando la estrategia de generación de amplicones solapantes y secuenciación ONT (58). Si bien la secuenciación del genoma completo del virus es indispensable para los estudios de la evolución viral, la asignación del linaje basado en el gen S permitiría conocer la circulación del virus en determinado contexto, identificar cambios no sinónimos de forma temprana y generar estrategias para la optimización de los recursos de secuenciación.

En resumen, en este trabajo se han aplicado distintas estrategias basadas en secuenciación de ADN para la caracterización de microorganismos procariotas y virus. También se ha desarrollado una herramienta automatizada para el análisis de datos de secuenciación de tercera generación obtenidas a partir de la amplificación del gen 16S del ARN*r* completo. La misma genera archivos de salida compatibles con otras herramientas para el estudio de la estructura y diversidad de comunidades microbianas. Finalmente, se generó un protocolo de amplificación y secuenciación del gen S para múltiples muestras de SARS-CoV-2 con la finalidad de generar secuencias consenso y posterior detección de linajes y/o variantes del virus utilizando herramientas específicas de asignación de linajes para la interpretación de la información contenida en este gen.

Referencias adicionales

1. Edelstein M, Lee LM, Herten-Crabb A, Heymann DL, Harper DR. Strengthening Global Public Health Surveillance through Data and Benefit Sharing. Emerg Infect Dis. julio de 2018;24(7):1324-30.
2. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet. enero de 2018;19(1):9-20.
3. Nsubuga P, White ME, Thacker SB, Anderson MA, Blount SB, Broome CV, et al. Public Health Surveillance: A Tool for Targeting and Monitoring Interventions. En: Jamison DT, Breman JG, Measham AR, Alleyne G, Claeson M, Evans DB, et al., editores. Disease Control Priorities in Developing Countries [Internet]. 2nd ed. Washington (DC): The International Bank for Reconstruction and Development / The World Bank; 2006 [citado 25 de mayo de 2023]. Disponible en: http://www.ncbi.nlm.nih.gov/books/NBK11770/
4. Declich S, Carter AO. Public health surveillance: historical origins, methods and evaluation. Bull World Health Organ. 1994;72(2):285-304.
5. Groseclose SL, Buckeridge DL. Public Health Surveillance Systems: Recent Advances in Their Use and Evaluation. Annu Rev Public Health. 20 de marzo de 2017;38(1):57-79.
6. Van Belkum A, Tassios PT, Dijkshoorn L, Haeggman S, Cookson B, Fry NK, et al. Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clin Microbiol Infect. 2007;13:1-46.
7. Kwong JC, Mccallum N, Sintchenko V, Howden BP. Whole genome sequencing in clinical and public health microbiology. Pathology (Phila). abril de 2015;47(3):199-210.
8. Köser CU, Ellington MJ, Cartwright EJP, Gillespie SH, Brown NM, Farrington M, et al. Routine Use of Microbial Whole Genome Sequencing in Diagnostic and Public Health Microbiology. Rall GF, editor. PLoS Pathog. 2 de agosto de 2012;8(8):e1002824.
9. Rice LB. Federal Funding for the Study of Antimicrobial Resistance in Nosocomial Pathogens: No ESKAPE. J Infect Dis. 15 de abril de 2008;197(8):1079-81.
10. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, NISC Comparative Sequencing Program, Henderson DK, et al. Tracking a Hospital Outbreak of Carbapenem-Resistant *Klebsiella pneumoniae* with Whole-Genome Sequencing. Sci Transl Med [Internet]. 22 de agosto de 2012 [citado 21 de mayo de 2023];4(148). Disponible en: https://www.science.org/doi/10.1126/scitranslmed.3004129
11. Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM, McDermott PF. Using Genomics to Track Global Antimicrobial Resistance. Front Public Health. 4 de septiembre

de 2019;7:242.

12. Bracing for Superbugs: Strengthening environmental action in the One Health response to antimicrobial resistance [Internet]. 2023. Disponible en: https://www.unep.org/resources/superbugs/environmental-action

13. Aarestrup FM. The Origin, Evolution, and Local and Global Dissemination of Antimicrobial Resistance. En: Aarestrup FM, editor. Antimicrobial Resistance in Bacteria of Animal Origin [Internet]. Washington, DC, USA: ASM Press; 2019 [citado 2 de junio de 2023]. p. 339-59. Disponible en: http://doi.wiley.com/10.1128/9781555817534.ch20

14. Roca I, Akova M, Baquero F, Carlet J, Cavaleri M, Coenen S, et al. The global threat of antimicrobial resistance: science for intervention. New Microbes New Infect. julio de 2015;6:22-9.

15. WHO. WHO publishes list of bacteria for which new antibiotics are urgently needed [Internet]. 2017. Disponible en: https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed

16. Murray CJL, Ikuta KS, Sharara F, Swetschinski L, Robles Aguilar G, Gray A, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. The Lancet. febrero de 2022;399(10325):629-55.

17. Sharma A, Singh A, Dar MA, Kaur RJ, Charan J, Iskandar K, et al. Menace of antimicrobial resistance in LMICs: Current surveillance practices and control measures to tackle hostility. J Infect Public Health. febrero de 2022;15(2):172-81.

18. TACKLING DRUG-RESISTANT INFECTIONS GLOBALLY: FINAL REPORT AND RECOMMENDATIONS [Internet]. 2016. Disponible en: https://amr-review.org/sites/default/files/160518_Final%20paper_with%20cover.pdf

19. WHO. GLASS Whole-genome sequencing for surveillance of antimicrobial resistance [Internet]. 2020. Disponible en: https://apps.who.int/iris/bitstream/handle/10665/334354/9789240011007-eng.pdf

20. Chokshi A, Sifri Z, Cennimo D, Horng H. Global contributors to antibiotic resistance. J Glob Infect Dis. 2019;11(1):36.

21. Dadgostar P. Antimicrobial Resistance: Implications and Costs. Infect Drug Resist. diciembre de 2019;Volume 12:3903-10.

22. Fresia P, Antelo V, Salazar C, Giménez M, D'Alessandro B, Afshinnekoo E, et al. Urban metagenomics uncover antibiotic resistance reservoirs in coastal beach and sewage waters. Microbiome. diciembre de 2019;7(1):35.

23. Karkman A, Do TT, Walsh F, Virta MPJ. Antibiotic-Resistance Genes in Waste Water. Trends Microbiol. marzo de 2018;26(3):220-8.
24. Walker WA. Dysbiosis. En: The Microbiota in Gastrointestinal Pathophysiology [Internet]. Elsevier; 2017 [citado 3 de junio de 2023]. p. 227-32. Disponible en: https://linkinghub.elsevier.com/retrieve/pii/B9780128040249000252
25. Levy M, Kolodziejczyk AA, Thaiss CA, Elinav E. Dysbiosis and the immune system. Nat Rev Immunol. abril de 2017;17(4):219-32.
26. Anthony WE, Burnham CAD, Dantas G, Kwon JH. The Gut Microbiome as a Reservoir for Antimicrobial Resistance. J Infect Dis. 16 de junio de 2021;223(Supplement_3):S209-13.
27. Fox GE, Magrum LJ, Balch WE, Wolfe RS, Woese CR. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proc Natl Acad Sci. octubre de 1977;74(10):4537-41.
28. Brosius J, Palmer ML, Kennedy PJ, Noller HF. Complete nucleotide sequence of a 16S ribosomal RNA gene from Escherichia coli. Proc Natl Acad Sci. octubre de 1978;75(10):4801-5.
29. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci. noviembre de 1977;74(11):5088-90.
30. Santos A, Van Aerle R, Barrientos L, Martinez-Urtaza J. Computational methods for 16S metabarcoding studies using Nanopore sequencing data. Comput Struct Biotechnol J. 2020;18:296-305.
31. Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the rrn operon. F1000Research. 1 de agosto de 2019;7:1755.
32. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 1 de enero de 2012;40(D1):D136-43.
33. Cole JR. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res. 17 de diciembre de 2004;33(Database issue):D294-6.
34. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 27 de noviembre de 2012;41(D1):D590-6.
35. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence

data compatible with ARB. Nucleic Acids Res. 14 de noviembre de 2007;35(21):7188-96.

36. Kozlov AM, Zhang J, Yilmaz P, Glöckner FO, Stamatakis A. Phylogeny-aware identification and correction of taxonomically mislabeled sequences. Nucleic Acids Res. 20 de junio de 2016;44(11):5022-33.

37. Myer PR, McDaneld TG, Kuehn LA, Dedonder KD, Apley MD, Capik SF, et al. Classification of 16S rRNA reads is improved using a niche-specific database constructed by near-full length sequencing. Schierwater B, editor. PLOS ONE. 13 de julio de 2020;15(7):e0235498.

38. Ritari J, Salojärvi J, Lahti L, de Vos WM. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. BMC Genomics. diciembre de 2015;16(1):1056.

39. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science. 2 de abril de 2004;304(5667):66-74.

40. Sunagawa S, Acinas SG, Bork P, Bowler C, Tara Oceans Coordinators, Acinas SG, et al. Tara Oceans: towards global ocean ecosystems biology. Nat Rev Microbiol. agosto de 2020;18(8):428-45.

41. Kinross JM, Darzi AW, Nicholson JK. Gut microbiome-host interactions in health and disease. Genome Med. 2011;3(3):14.

42. Fierer N. Embracing the unknown: disentangling the complexities of the soil microbiome. Nat Rev Microbiol. octubre de 2017;15(10):579-90.

43. Danko D, Bezdan D, Afshin EE, Ahsanuddin S, Bhattacharya C, Butler DJ, et al. A global metagenomic map of urban microbiomes and antimicrobial resistance. Cell. junio de 2021;184(13):3376-3393.e17.

44. Rodríguez-Pérez H, Ciuffreda L, Flores C. NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data. Inanc B, editor. Bioinformatics. 12 de julio de 2021;37(11):1600-1.

45. Curry KD, Wang Q, Nute MG, Tyshaieva A, Reeves E, Soriano S, et al. Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. Nat Methods. julio de 2022;19(7):845-53.

46. Benítez-Páez A, Portune KJ, Sanz Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. GigaScience. diciembre de 2016;5(1):4.

47. Nygaard AB, Tunsjø HS, Meisal R, Charnock C. A preliminary study on the potential

of Nanopore MinION and Illumina MiSeq 16S rRNA gene sequencing to characterize building-dust microbiomes. Sci Rep. 21 de febrero de 2020;10(1):3209.

48. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 6 de noviembre de 2019;10(1):5029.

49. Matsuo Y, Komiya S, Yasumizu Y, Yasuoka Y, Mizushima K, Takagi T, et al. Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION™ nanopore sequencing confers species-level resolution. BMC Microbiol. diciembre de 2021;21(1):35.

50. Kai S, Matsuo Y, Nakagawa S, Kryukov K, Matsukawa S, Tanaka H, et al. Rapid bacterial identification by direct PCR amplification of 16S rRNA genes using the MinION™ nanopore sequencer. FEBS Open Bio. marzo de 2019;9(3):548-57.

51. Urban L, Holzer A, Baronas JJ, Hall MB, Braeuninger-Weimer P, Scherm MJ, et al. Freshwater monitoring by nanopore sequencing. eLife. 19 de enero de 2021;10:e61504.

52. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics. 15 de septiembre de 2018;34(18):3094-100.

53. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 12 de marzo de 2020;579(7798):270-3.

54. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 12 de marzo de 2020;579(7798):265-9.

55. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet. febrero de 2020;395(10224):565-74.

56. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, Gorbalenya AE, Baker SC, Baric RS, De Groot RJ, Drosten C, et al. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2 de marzo de 2020;5(4):536-44.

57. WHO. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 [Internet]. 2020. Disponible en: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

58. Malekpour MR, Abbasi-Kangevari M, Azadnajafabad S, Ghamari SH, Rezaei N,

Rezazadeh-Khadem S, et al. How the scientific community responded to the COVID-19 pandemic: A subject-level time-trend bibliometric analysis. Radfar A, editor. PLOS ONE. 30 de septiembre de 2021;16(9):e0258064.

59. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc. junio de 2017;12(6):1261-76.

60. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. febrero de 2016;530(7589):228-32.

61. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance [Internet]. 30 de marzo de 2017 [citado 24 de febrero de 2023];22(13). Disponible en: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2017.22.13.30494

62. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 15 de julio de 2020;5(11):1403-7.

63. WHO. Tracking SARS-CoV-2 variants. 2022.

64. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Kelso J, editor. Bioinformatics. 1 de diciembre de 2018;34(23):4121-3.

65. Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, Kraemer MUG, et al. Global disparities in SARS-CoV-2 genomic surveillance. Nat Commun. 16 de noviembre de 2022;13(1):7003.

66. Petersen LM, Martin IW, Moschetti WE, Kershaw CM, Tsongalis GJ. Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing. Kraft CS, editor. J Clin Microbiol. 23 de diciembre de 2019;58(1):e01315-19.

67. Sheka D, Alabi N, Gordon PMK. Oxford nanopore sequencing in clinical microbiology and infection diagnostics. Brief Bioinform. 2 de septiembre de 2021;22(5):bbaa403.

68. Salazar C, Giménez M, Riera N, Parada A, Puig J, Galiana A, et al. Human microbiota drives hospital-associated antimicrobial resistance dissemination in the urban environment and mirrors patient case rates. Microbiome. 2 de diciembre de 2022;10(1):208.

69. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. julio de 2016;13(7):581-3.

70. Pichler M, Coskun ÖK, Ortega‐Arbulú A, Conci N, Wörheide G, Vargas S, et al. A 16S rRNA gene sequencing and analysis protocol for the Illumina MiniSeq platform. MicrobiologyOpen. diciembre de 2018;7(6):e00611.

71. Kuczynski J, Stombaugh J, Walters WA, González A, Caporaso JG, Knight R. Using QIIME to Analyze 16S rRNA Gene Sequences from Microbial Communities. Curr Protoc Microbiol [Internet]. noviembre de 2012 [citado 7 de junio de 2023];27(1). Disponible en: https://onlinelibrary.wiley.com/doi/10.1002/9780471729259.mc01e05s27

72. Shin J, Lee S, Go MJ, Lee SY, Kim SC, Lee CH, et al. Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. Sci Rep. 14 de julio de 2016;6(1):29681.

73. Oberle A, Urban L, Falch-Leis S, Ennemoser C, Nagai Y, Ashikawa K, et al. 16S rRNA long-read nanopore sequencing is feasible and reliable for endometrial microbiome analysis. Reprod Biomed Online. junio de 2021;42(6):1097-107.

74. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. Watson M, editor. PLoS ONE. 22 de abril de 2013;8(4):e61217.

75. Lu Y, Zhou G, Ewald J, Pang Z, Shiri T, Xia J. MicrobiomeAnalyst 2.0: comprehensive statistical, functional and integrative analysis of microbiome data. Nucleic Acids Res. 11 de mayo de 2023;gkad407.

76. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. marzo de 2007;17(3):377-86.

77. Salazar C, Díaz-Viraqué F, Pereira-Gómez M, Ferrés I, Moreno P, Moratorio G, et al. Multiple introductions, regional spread and local differentiation during the first week of COVID-19 epidemic in Montevideo, Uruguay [Internet]. Microbiology; 2020 may [citado 27 de mayo de 2023]. Disponible en: http://biorxiv.org/lookup/doi/10.1101/2020.05.09.086223

78. Salazar C, Costabile A, Ferrés I, Perbolianachis P, Pereira-Gómez M, Simón D, et al. Case Report: Early Transcontinental Import of SARS-CoV-2 Variant of Concern 202012/01 (B.1.1.7) From Europe to Uruguay. Front Virol. 28 de mayo de 2021;1:685618.

79. Rego N, Costábile A, Paz M, Salazar C, Perbolianachis P, Spangenberg L, et al. Real-Time Genomic Surveillance for SARS-CoV-2 Variants of Concern, Uruguay. Emerg

Infect Dis. noviembre de 2021;27(11):2957-60.

80. Rego N, Salazar C, Paz M, Costábile A, Fajardo A, Ferrés I, et al. Emergence and Spread of a B.1.1.28-Derived P.6 Lineage with Q675H and Q677H Spike Mutations in Uruguay. Viruses. 10 de septiembre de 2021;13(9):1801.

81. Salazar C, Ferrés I, Paz M, Costábile A, Moratorio G, Moreno P, et al. Fast and cost-effective SARS-CoV-2 variant detection using Oxford Nanopore full-length spike gene sequencing. Microb Genomics [Internet]. 18 de mayo de 2023 [citado 28 de mayo de 2023];9(5). Disponible en: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.001013

Anexo

Contribuciones

Capítulo 1:

Mag. Cecilia Salazar

Preparación de bibliotecas de secuenciación genómica para la plataforma ONT.

Pre y post procesamiento de lecturas de secuenciación de segunda y tercera generación.

Análisis de datos a partir de los ensamblados de genomas.

Visualizaciones.

Resultados parcialmente incluídos en:

Cecilia Salazar, Matias Giménez, Nadia Riera, Andrés Parada, Josefina Puig, Antonio Galiana, Fabio Gril, Mariela Vieytes, Christopher E Mason, Verónica Antelo, Bruno D'Alessandro, Jimena Risso, Gregorio Iraola. Human microbiota drives hospital-associated antimicrobial resistance dissemination in the urban environment and mirrors patient case rates. Microbiome. 2022 Dec 2;10(1):208. doi: https://doi.org/10.1186/s40168-022-01407-8.

Capítulo 2:

Mag. Cecilia Salazar

Conceptualización del pipeline de análisis.

Elección de las estrategias y herramientas para los módulos de preprocesamiento, clasificación taxonómica y pulido a nivel de especie.

Validación del pipeline automático a través de la simulación de lecturas de secuenciación ONT, análisis comparativos con otras herramientas y reanálisis de un dataset obtenido de la literatura. Visualizaciones.

Capítulo 3:

Mag. Cecilia Salazar

Obtención y análisis de secuencias de Uruguay y reportadas en la literatura de la base de datos GISAID.

Participación del grupo de trabajo del Institut Pasteur de Montevideo que obtuvo las primeras secuencias genómicas de SARS-CoV-2 en Uruguay en marzo de 2020.

Grupo de trabajo del Institut Pasteur de Montevideo: Dra. Florencia Díaz-Viraqué, Dra. Marianoel Pereira-Gómez, Mag. Ignacio Ferrés, Dra. Pilar Moreno, Dr. Gonzalo Moratorio y Dr. Gregorio Iraola.

Capítulo 4:

Mag. Cecilia Salazar

Implementación y adaptación de estrategias para la preparación de las bibliotecas de secuenciación de amplicones del gen S para secuenciación ONT.

Análisis de datos de secuenciación genómica y del gen S. Generación de visualizaciones.

Resultados incluídos en:

Cecilia Salazar, Ignacio Ferrés, Mercedes Paz, Alicia Costábile, Gonzalo Moratorio, Pilar Moreno, Gregorio Iraola. Fast and cost-effective SARS-CoV-2 variant detection using Oxford Nanopore full-length spike gene sequencing. Microb Genom. 2023; 9(5): doi: https://doi.org/10.1099/mgen.0.001013.