Procedimiento de Microbiología Clínica

Recomendaciones de la Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica

71. Aplicaciones de las técnicas de secuenciación masiva en la Microbiología Clínica

Editores

Emilia Cercenado Mansilla
Rafael Cantón Moreno

Coordinador

Antonio Oliver Palomo

Autores

Carla López Causapé
Fernando González Candelas
María Tomás Carmona
Antonio Oliver Palomo

ISBN: 978-84-09-31350-1

EDITORES:

Emilia Cercenado Mansilla. Servicio de Microbiología. Hospital General Universitario Gregorio Marañón. Madrid.

Rafael Cantón Moreno, Servicio de Microbiología. Hospital Universitario Ramón y Cajal e Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS). Madrid.

SUGERENCIA DE CITACIÓN:

Carla López Causapé C, González Candelas F, Tomás Carmona M, Oliver Palomo A. Aplicaciones de las técnicas de secuenciación masiva en la Microbiología Clínica. 2021. 71. Antonio Oliver Palomo (coordinador). Procedimientos en Microbiología Clínica. Cercenado Mansilla E, Cantón Moreno R (editores). Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica (SEIMC). 2021.

Procedimientos en Microbiología Clínica

Recomendaciones de la Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica

Editores:

Emilia Cercenado Mansilla

Rafael Cantón Moreno

71. Aplicaciones de las técnicas de secuenciación masiva en la Microbiología Clínica.2021

Coordinador:

Antonio Oliver Palomo[1]

Autores:

Carla López Causapé[1]

Fernando González Candelas[2]

María Tomás Carmona[3]

Antonio Oliver Palomo[1]

[1]Servicio de Microbiología, Hospital Son Espases, Instituto de Investigación Sanitaria Illes Balears (IdISBa), Palma de Mallorca; [2]Unidad Mixta Infección y Salud Pública FISABIO/Universidad de Valencia. Valencia; [3]Servicio de Microbiología-Instituto de Investigación Biomédica A Coruña (INIBIC), Hospital A Coruña (CHUAC), Universidad de A Coruña (UDC), A Coruña.

ÍNDICE

DOCUMENTOS TÉCNICOS

1. INTRODUCCIÓN

De forma general, el genoma es el conjunto de material genético hereditario que posee un organismo vivo y que determina su capacidad de adaptación a las diferentes condiciones ambientales y nutricionales.

A nivel bioquímico, e independientemente de su origen, todas las moléculas de ADN son idénticas. Dichas macromoléculas están compuestas por nucleótidos unidos por medio de enlaces fosfodiéster entre los carbonos de las posiciones 3´ y 5´ de dos residuos de azúcares adyacentes; siendo el orden de la secuencia de nucleótidos que conforma esta macromolécula el determinante de las diferentes características biológicas en una especie. De igual forma, las diferencias fenotípicas observadas entre miembros de una misma especie vienen determinadas por cambios en esta secuencia de nucleótidos.

En los últimos años, el constante desarrollo y la continua mejora de los métodos de secuenciación del ADN ha permitido que la obtención de la secuencia de genoma completo sea accesible en la rutina de los Servicios de Microbiología Clínica, introduciendo así un nuevo paradigma en el diagnóstico microbiológico (1). Este nuevo paradigma implica el manejo de una gran cantidad de información, cuya gestión y análisis demanda inexorablemente el uso de herramientas bioinformáticas lo cual no debiera ser un obstáculo para su implementación en los Servicios de Microbiología Clínica ya que en los últimos años se han desarrollado aplicaciones web y/o interfaces que permiten una fácil utilización de muchas de estas herramientas sin la necesidad de requerir grandes conocimientos en lenguajes de programación como Bash, Perl o Python o del uso de la terminal.

Una de las principales ventajas de la secuenciación masiva es su universalidad, siendo posible obtener la secuencia de genoma completo de cualquier especie microbiana. En este punto, cabe destacar el crucial papel que esta herramienta está teniendo en la pandemia causada por el SARS-CoV-2. En los primeros días, disponer de estas herramientas permitió identificar el agente causal del brote de neumonía de origen desconocido reportado en la ciudad de Wuhan a finales de 2019 así como obtener rápidamente la secuencia del genoma completo del virus, hecho que permitió el inmediato desarrollo de las herramientas diagnósticas necesarias. Posteriormente, la secuenciación de genoma completo se ha perfilado como una herramienta clave en la monitorización y control de la pandemia causada por el SARS-CoV-2, instando los gobiernos y organismos de Salud a su incorporación como herramienta para el control y vigilancia de la pandemia. Dada la relevancia de la secuenciación masiva como herramienta de epidemiología genómica en la pandemia, este tema se tratará en un Procedimiento SEIMC de forma independiente.

Hasta la fecha, y a pesar de su universalidad, el uso de la secuenciación de genoma completo como herramienta de diagnóstico microbiológico se ha enfocado mayoritariamente en el estudio de especies bacterianas, centrándose por tanto el presente procedimiento exclusivamente en la aplicabilidad de la secuenciación masiva para el estudio de genomas bacterianos y metagenomas. No obstante, la secuenciación masiva está adquiriendo también relevancia en otros ámbitos de la Microbiología Clínica, especialmente en la virología, incluyendo aspectos epidemiológicos, diagnósticos, detección de resistencias a antivirales, etc., que serán tratados de forma específica en futuros Procedimientos SEIMC.

El genoma bacteriano se refiere al conjunto total de genes que posee una bacteria y comprende tanto su cromosoma como todos los elementos genéticos extracromosómicos, en caso de poseer alguno. El cromosoma bacteriano contiene toda la información genética esencial para la vida de la bacteria y está formado por una única molécula de ácido desoxirribonucleico (ADN) de doble cadena y circular, cerrado por enlace covalente. Por su parte, el ADN extracromosómico comprende el ADN plasmídico, también circular y cerrado, y los bacteriófagos (virus bacterianos). Académicamente, el genoma bacteriano puede diferenciarse en el denominado genoma core y el denominado genoma accesorio, haciendo referencia el genoma core al conjunto de genes presentes en todos los individuos de una especie bacteriana concreta y el genoma accesorio al conjunto de genes variables entre miembros de una misma especie.

Por otro lado, el término metagenoma, se reserva para hacer referencia al conjunto de genes de origen microbiano presentes en un ecosistema determinado

2. TECNOLOGÍAS DE SECUENCIACIÓN

Desde que en 1995 Fleischmann y colaboradores publicasen la secuencia de genoma completo de *Haemophilus influenzae* (2), se han producido numerosos avances en el campo de la biología molecular que, junto a los numerosos avances tecnológicos, han permitido el desarrollo de nuevas y eficientes tecnologías de secuenciación del ADN (3,4). Así, en la actualidad, conviven diferentes generaciones de secuenciadores cuyas bases científicas y tecnológicas son distintas, resultando complementarias en muchas ocasiones.

2.1. SECUENCIACIÓN DE 1ª GENERACIÓN (SANGER)

La historia de la secuenciación del ADN se remonta a 1977, año en el que, Maxam-Gilbert y Sanger-Coulson describieron, de forma independiente, dos estrategias que permitían obtener secuencias de cientos de nucleótidos en solo unas horas (5,6). No obstante, la complejidad experimental y difícil automatización del método basado en la escisión química diferencial propuesto por Maxam y Gilbert ha hecho que éste haya quedado en desuso en la actualidad.

El método propuesto por Sanger-Coulson, o de terminación de la cadena, se basa en la síntesis secuencial de una hebra de ADN complementaria a la hebra de cadena simple cuya secuencia quiere determinarse, residiendo la clave del método en la adición de dideoxinucleótidos o nucleótidos de parada a la mezcla de la reacción ya que, al carecer estos nucleótidos del grupo 3'-OH, su adición resulta en la interrupción de la síntesis. Así, la secuenciación de Sanger-Coulson consiste en llevar a cabo 4 reacciones independientes conteniendo cada una de las mezclas de reacción la hebra molde, la ADN polimerasa, un cebador marcado, los cuatro 2'-deoxinucleótidos (dNTP) y uno de los cuatro dideoxinucleótidos (ddNTP) en menor proporción. Finalizado el proceso de síntesis, la adición aleatoria de estos nucleótidos de parada resulta en la obtención de 4 mezclas de fragmentos de diferente tamaño interrumpidos por el ddNTP correspondiente que, al ser separados mediante electroforesis en gel de poliacrilamida, permiten elucidar la secuencia de nucleótidos de la cadena de ADN molde gracias al marcaje incorporado al cebador.

El método de terminación de cadena fue bien acogido en la comunidad científica y su uso se extendió casi de forma inmediata. Los avances tecnológicos que acontecieron en años posteriores hicieron que en 1987 saliera al mercado el primer secuenciador Sanger automatizado capaz de generar hasta 1000 pares de bases al día, cuya principal diferencia con el método descrito originalmente radicaba en la sustitución del marcaje radiactivo de los cebadores por el uso de ddNTPs marcados con 4 fluoróforos distintos lo que permitía realizar la secuenciación en una sola mezcla de reacción eliminando además la generación de residuos reactivos. Cabe destacar también, que estos secuenciadores sustituían la electroforesis en geles de poliacrilamida por la electroforesis capilar e incorporaban la técnica de la reacción en cadena de la polimerasa descrita por Mullis en 1983.

La automatización de la secuenciación Sanger supuso una revolución en el campo de la genética humana, actuando como catalizador de proyectos tan ambiciosos para la época como el Proyecto Genoma Humano; proyectos que, a su vez, han propiciado el desarrollo de nuevas generaciones de secuenciadores.

No obstante, a pesar de la fuerte irrupción de estas nuevas generaciones de secuenciadores y de la necesidad de cierto conocimiento previo sobre la secuencia a resolver, los secuenciadores Sanger tuvieron y continúan teniendo un importante papel en numerosas investigaciones y aplicaciones microbiológicas. Así, por ejemplo, la secuenciación Sanger continúa siendo ampliamente utilizada para la identificación taxonómica de aislados microbiológicos mediante secuenciación del gen que codifica la subunidad menor del ARN ribosómico 16S (7).

2.2. SECUENCIACIÓN DE SEGUNDA GENERACIÓN (MASIVA)

La secuenciación masiva engloba toda una generación de secuenciadores que utilizan diferentes tecnologías, estrategias y aproximaciones, cuya característica común es su habilidad para llevar a cabo de forma simultánea millones de reacciones de secuenciación en paralelo a un precio asequible y en un tiempo relativamente corto. Los secuenciadores masivos o de segunda generación difieren en muchos aspectos con los secuenciadores tipo Sanger. Una de las principales diferencias es la realización simultánea de millones de reacciones de secuenciación de fragmentos muy cortos que pueden asignarse a cada una de las muestras que se analizan simultáneamente gracias a la preparación previa de unas librerías etiquetadas (multiplexado) que se inmovilizan sobre una matriz bidimensional.

De forma general, el proceso de secuenciación masiva puede dividirse en tres pasos: (1) preparación de librerías, (2) inmovilización sobre una superficie bidimensional y amplificación *in vitro* de la librería y (3) secuenciación y captación de señal.

La preparación de librerías consiste básicamente en fragmentar y marcar de forma inequívoca las diferentes muestras de ADN cuya secuencia quiere conocerse. Actualmente, existen diferentes estrategias de fragmentación del ADN (física, enzimática o química) cuyo objetivo final es la obtención de una mezcla de fragmentos de un tamaño homogéneo, estando el tamaño deseado determinado por la tecnología de secuenciación empleada. Una vez fragmentado el ADN genómico se procede al marcaje para lo cual se añaden unos identificadores únicos que no sólo permiten identificar la muestra a la que pertenecen, sino que además permiten la posterior fijación de los mismos a la superficie bidimensional donde tendrá lugar la reacción de secuenciación.

Una vez preparadas las librerías, éstas se fijan a una superficie bidimensional donde son amplificadas *in vitro* existiendo también en este punto diferentes estrategias. De todas ellas, la más sencilla es la utilizada por los secuenciadores de Illumina y denominada amplificación puente (*bridge amplification*). En esta estrategia los cebadores necesarios para el inicio de la reacción de amplificación se encuentran inmovilizados sobre una superficie bidimensional; al ser estos cebadores complementarios a una parte de los identificadores incorporados en la preparación de librerías, las muestras quedan inmovilizadas y en disposición de iniciarse la amplificación clonal en paralelo de todas las muestras. Otras estrategias alternativas incluyen la amplificación en emulsión o la amplificación en nanobolas (*rolling-circle amplification*).

Finalmente, una vez amplificados los fragmentos de ADN, se produce la reacción de secuenciación para la cual también se han desarrollado diferentes estrategias. Una de estas estrategias es la pirosecuenciación, cuya base es la utilización de nucleótidos cuya incorporación a la hebra complementaria en síntesis conlleva la liberación de pirofosfatos que, en última instancia, generan luz que puede detectarse permitiendo así elucidar la secuencia de nucleótidos. Otra estrategia ampliamente utilizada es la denominada secuenciación por síntesis (SBS, *sequencing by synthesis*). En esta aproximación la reacción de síntesis consiste en la incorporación secuencial y reversible de nucleótidos marcados con fluoróforos distintos que impiden además la adición posterior de más nucleótidos; de esta forma, en cada paso se produce la adición de un único nucleótido a cada uno de los fragmentos en síntesis cuya señal fluorescente es captada y traducida a nucleótidos. Finalmente, otra aproximación ampliamente utilizada es la denominada secuenciación mediante ligación (SBL; *sequencing by ligation*) que no utiliza ADN polimerasas y cuya base reside en la utilización de una mezcla de sondas marcadas con fluoróforos que se unen mediante ligasas específicas. Por tanto, es importante conocer la estrategia química de secuenciación en que se basa el secuenciador ya que determina en gran medida sus características y prestaciones (Tabla 1).

De entre las diferentes tecnologías de secuenciación masiva actualmente disponibles (Tabla 1), la tecnología Illumina es la que se haya más extendida en el ámbito clínico dado que su alto rendimiento y la precisión de las lecturas generadas la hacen idónea para la mayoría de las aplicaciones de la secuenciación masiva en Microbiología Clínica.

Tabla 1. Características de las distintas plataformas de secuenciación.

Plataforma	Secuenciador	Amplificación de la librería	Tecnología de secuenciación	Máximo de lecturas por carrera (/día*)	Longitud máxima de las lecturas	Tiempo carrera	Output máx. por carrera (/día*)	Precisión por base	Tipos de errores más frecuentes	Coste mín. Gb (/muestra)
ABI Sanger	SeqStudio	No aplica	Sanger	67K*	800 pb	30 min	192 muestras	<0,01%	-	$
	3500 series	No aplica	Sanger	3500: 138K * 3500xL: 403K*	> 850 pb	30 min	3500: 384 muestras 3500xL: 1152 muestras	<0,01%	-	$
	Refreshed 3730 series	No aplica	Sanger	3730: 1.38 M * 3730xL: 2.76 M*	900 pb	20 min	3730: 3456 muestras 3730xL: 6912 muestras	<0,01%	-	$
Illumina	MiSeq	Puente (*bridge*-PCR)	Secuenciación por síntesis	50 M	300x2 pb	4-56 h	15 Gb	0,1-1%	Sustituciones de nucleótidos	$$
	NextSeq	Puente (*bridge*-PCR)	Secuenciación por síntesis	800 M	150x2pb	11-29 h	120 Gb	0,1-1%	Sustituciones de nucleótidos	$$
	HiSeq 4000	Puente (*bridge*-PCR)	Secuenciación por síntesis	2,5 B	150x2pb	1-3,5 días	750 Gb	0,1-1%	Sustituciones de nucleótidos	$$
Ion Torrent	PGM	PCR en emulsión (emPCR)	Secuenciación por ligación	5,5 M	400 pb	2-7 h	1 Gb	1%	InDels	$$
	S5 series	PCR en emulsión (emPCR)	Secuenciación por ligación	S5: 80M S5Plus: 130M S5Prime: 130M	600 pb	3-21,5 h	25 Gb	1%	InDels	$$
PacBio	RS	No aplica	SMRT	50 K	10-15 Kb	4 h (máx.)	1 Gb	10-15%	InDels	$$$
	Sequel	No aplica	SMRT	500 K	10-15 Kb	4 h (máx.)	10 Gb	10-15%	InDels	$$$
Oxford nanopore	MiniON	No aplica	SMRT	1M (máx)	10-20 Kb	1-48 h	5 Gb	5-15%	Sustituciones e Indels	$$$

2.3. SECUENCIACIÓN DE TERCERA GENERACION (DE MOLÉCULA ÚNICA)

Una de las principales limitaciones de la secuenciación masiva es su incapacidad de generar secuencias de gran tamaño, hecho por el cual surge la denominada secuenciación de 3ª generación o de molécula única.

A diferencia de la secuenciación masiva, la secuenciación de molécula única no fragmenta ni amplifica el ADN en pasos previos a la reacción de secuenciación, siendo la estabilización de las largas moléculas de ADN el principal reto al que se enfrenta esta generación de secuenciadores. Otra diferencia destacable de los secuenciadores de tercera generación con respecto a los secuenciadores masivos es que la detección de las señales que se producen durante la reacción de secuenciación ocurre a tiempo real. Al igual que ocurre en la secuenciación masiva, se han desarrollado diferentes aproximaciones.

Una primera aproximación es la basada en la utilización de unos nanopocillos en cuya base existe un orificio, denominado *zero-mode waveguide*, que permite el paso de luz. El pequeño tamaño de estos nanopocillos permite albergar una única molécula de ADN junto con una sola ADN polimerasa; así, la sucesiva adición de nucleótidos, cada uno marcado con un fluoróforo distinto, por la polimerasa provoca la liberación y detección de los mismos. Por otro lado, la principal alternativa al uso de estos nanopocillos es la utilización de nanoporos embebidos en una membrana sintética a la que se le aplica un voltaje. En este caso la desnaturalización de la cadena de ADN se produce en presencia de uno de estos nanoporos produciéndose así la unión de una de las cadenas simples. Posteriormente, conforme la molécula de ADN va atravesando el nanoporo se producen cambios en el voltaje aplicado que pueden medirse y traducirse en una secuencia nucleotídica.

En este punto cabe resaltar que, en la actualidad, las secuencias generadas por estos secuenciadores de molécula única resultan más caras o presentan una calidad insuficiente para la mayoría de las aplicaciones en Microbiología Clínica. No obstante, el coste está disminuyendo de forma significativa en los últimos años y, por tanto, en ciertos escenarios puede valorarse ya el uso de la secuenciación de molécula única, especialmente cuando prime la rapidez en la generación de resultados.

3. FLUJO DE TRABAJO EN LA SECUENCIACIÓN MASIVA

3.1. ANÁLISIS METAGENÓMICO Y SECUENCIACIÓN DE GENOMAS COMPLETOS

La universalidad de las nuevas tecnologías de secuenciación permite su aplicación al estudio de cualquier especie o comunidad microbiana incluso en ausencia de una sospecha o conocimiento previo sobre el origen etiológico del material genético a estudiar.

De forma general, se pueden distinguir dos aproximaciones bien diferenciadas según la secuenciación ocurra en presencia o ausencia de cultivo microbiológico. Así, cuando el ADN objeto de secuenciación haya sido extraído a partir de un microorganismo en cultivo puro hablaremos de secuenciación de genoma completo (*Whole Genome Sequencing,* WGS) mientras que cuando éste haya sido extraído directamente de la muestra clínica en ausencia de cultivo hablaremos de secuenciación metagenómica (*Whole Metagenome Sequencing*, WMS).

La elección de una u otra aproximación dependerá de la cuestión a resolver (8). En su elección, es importante considerar que la secuenciación de genoma completo ofrece, en general, un rendimiento mucho mayor y, por tanto, su coste es menor, al corresponderse la totalidad de las lecturas generadas por el secuenciador con el microorganismo objeto de estudio. Sin embargo, cabe destacar que la aproximación metagenómica nos permitirá resolver cuestiones no abordables con secuenciación de genoma completo. Así, por ejemplo, será de elección en el estudio de microorganismos no cultivables o de crecimiento lento o bien cuando queramos estudiar una comunidad microbiana en su conjunto.

Con independencia de la aproximación escogida, en secuenciación masiva se diferencian dos procesos, en inglés denominados *wet-lab* (Figura 1) y *dry-lab* (Figura 2). Estos procesos requieren de unas habilidades y conocimientos muy diferentes, ya que el primero hace referencia a todos los pasos relacionados con el procesamiento de la muestra hasta la obtención última de las lecturas y el segundo al análisis bioinformático de las mismas.

3.2. PROCESAMIENTO DE LAS MUESTRAS (Figura 1)

3.2.1. Elección de la muestra

La elección de la muestra va a depender principalmente de la localización de la infección o colonización y/o de la sospecha etiológica existente, en caso de que exista. Para su elección se recomienda consultar el Procedimiento 1b Recogida, transporte y procesamiento general de las muestras en el laboratorio de Microbiología, SEIMC 2017 (https://seimc.org/contenidos/documentoscientificos/procedimientosmicrobiologia/seimc-procedimientomicrobiologia1b.pdf) en el que se especifican las muestras clínicas más adecuadas para el diagnóstico microbiológico de los procesos infecciosos más habituales.

Habitualmente, para la secuenciación de genoma completo partiremos de un cultivo clonal puro del aislado microbiológico a estudio ya que así obtendremos un mayor rendimiento. Para metagenómica, existen dos posibilidades, bien podemos partir de la muestra clínica o bien de un cultivo enriquecido de la misma (ejemplo: hemocultivos) (9). En ambos casos, es necesario asegurar que la muestra no contiene sustancias que puedan interferir en alguna de las etapas del proceso de secuenciación, así como valorar su complejidad celular.

3.2.2. Recogida, conservación y transporte

La calidad de los resultados que se obtengan mediante secuenciación dependerá en gran medida de la cantidad y calidad de la extracción de ácidos nucleicos aislados, por lo que tanto la recogida de la muestra como su posterior transporte al laboratorio y conservación, son etapas cruciales en el proceso de secuenciación. La recogida de la muestra debe realizarse en condiciones asépticas. Asimismo, como regla general, y con el fin de preservar la integridad de los ácidos nucleicos, no deben transcurrir más de dos horas entre la toma de la muestra y su recepción en el laboratorio, manteniéndose refrigeradas durante el transporte.

En el caso particular de requerirse la utilización de medios de transporte específicos, es necesario asegurarse de que éstos no contienen sustancias que puedan interferir o inhibir la posterior extracción de ácidos nucleicos y secuenciación, siendo esta consideración especialmente relevante en las aproximaciones metagenómicas.

Una vez la muestra clínica se recepciona en el laboratorio, la muestra se procesará para cultivo microbiológico siguiendo los procedimientos y protocolos existentes en el laboratorio o bien se procederá a la extracción de los ácidos nucleicos totales. Alternativamente, la muestra puede congelarse inmediatamente a -80ºC y ser procesada con posterioridad.

Se recomienda consultar el Procedimiento 1b. Recogida, transporte y procesamiento general de las muestras en el laboratorio de Microbiología, SEIMC 2017. (https://seimc.org/contenidos/documentoscientificos/procedimientosmicrobiologia/seimc-procedimientomicrobiologia1b.pdf).

3.2.3. Extracción de ácidos nucleicos

Como se ha adelantado en el apartado anterior, la extracción de los ácidos nucleicos puede realizarse sobre la muestra clínica, directa o en cultivo de enriquecimiento, o bien a partir de un cultivo clonal puro del microorganismo a estudio.

Cuando la extracción se realice directamente sobre muestra clínica, es preciso considerar su posible contenido en ácidos nucleicos de origen humano ya que la sensibilidad y eficiencia del método podría verse disminuida significativamente. Asimismo, debe considerarse el hecho de que ciertas enfermedades infecciosas, como por ejemplo la diarrea causada por *Shigella flexneri* o la fiebre causada por el virus del Zika, pueden cursar con bajas cargas microbianas. En estos escenarios, se recomienda enriquecer la muestra en el material genético de interés, lo cual puede alcanzarse bien mediante depleción del material genético de origen humano o bien mediante la selección del material genético a estudio (10).

Para la depleción del material genético de origen humano durante la preparación de librerías de ADN existen diferentes alternativas que han demostrado tener una buena eficacia (11-14). Sin embargo, la depleción selectiva de ADN resulta más compleja y su efectividad suele ser menor. Cabe destacar el método descrito por Feehery y colaboradores basado en las diferencias en la metilación de las cadenas de ADN de origen humano y bacteriano (15), así como los métodos basados en la lisis diferencial de las células humanas seguida de la degradación del ADN recién liberado (16). Es importante considerar las limitaciones intrínsecas a estos métodos de lisis diferencial, ya que no sólo deplecionan el ADN de origen humano recién liberado, sino que la depleción afectará a todo el ADN que se encuentre libre independientemente de su origen. Alternativamente, se puede enriquecer la muestra en el material genético de interés mediante sistemas de captura por hibridación o mediante amplificación con polimerasas de alta fidelidad, aproximación de elección cuando existe una sospecha etiológica previa.

La extracción de ácidos nucleicos constituye un paso crítico en el procesamiento de las muestras para secuenciación ya que se necesita un ADN de buena calidad y libre de contaminantes. La calidad del ADN extraído y la presencia de potenciales contaminantes puede evaluarse utilizando la absorbancia diferencial de luz UV: una ratio de absorbancia 260/280 nm comprendido entre 1,8-2 es indicativo de que el extracto únicamente contiene ADN y una ratio 260/230 nm entre 2,0-2,2 es indicativo de la ausencia de contaminantes.

3.2.4. Preparación de librerías

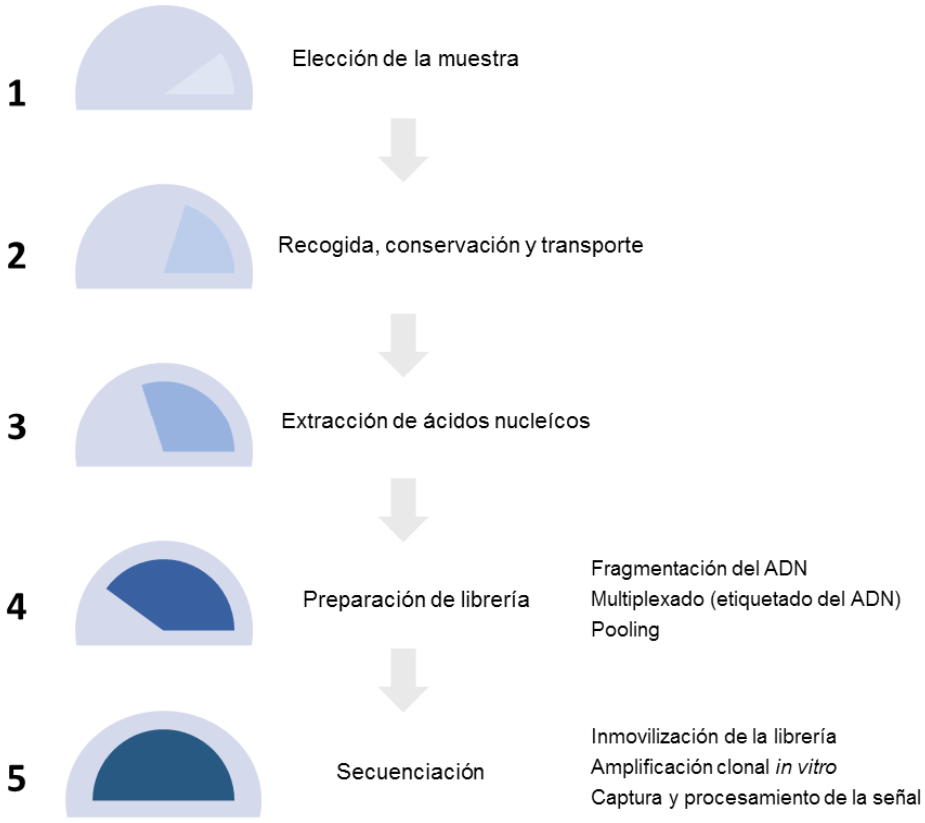
La preparación de las librerías depende de la tecnología que incorpore el secuenciador y de la aplicación última de la secuenciación. No obstante, existen pasos comunes en las preparaciones de librerías para secuenciación masiva: la fragmentación y multiplexado de las muestras.

Como ya se adelantó, las diferentes aproximaciones de secuenciación masiva o de segunda generación comparten el hecho de que la secuenciación ocurre sobre una superficie bidimensional sobre la que se adhieren los fragmentos cuyas secuencias queremos elucidar. Por tanto, para asegurar una cobertura homogénea de la superficie, las muestras deberán tener un tamaño homogéneo y relativamente pequeño que no impida la adición posterior de más hebras de ADN en las zonas adyacentes. La preparación de librerías suele iniciarse con un primer paso de fragmentación del ADN genómico, que puede ser enzimática, química o física. En este punto, es muy importante controlar la concentración inicial de ADN de doble hebra para lo que se recomienda la utilización de métodos de cuantificación fluorométricos.

Fragmentado el ADN genómico, se procede al multiplexado de las muestras. En este paso, las muestras son identificadas de forma inequívoca mediante la adición de unos oligonucleótidos que actúan como identificadores únicos y que además incorporan una secuencia universal que les permite su posterior adición a la superficie bidimensional donde se produce la secuenciación.

Finalmente, antes de introducir las muestras en el secuenciador, éstas deben normalizarse ya que introduciremos una mezcla equimolar de las mimas con el objetivo último de que todas sean secuenciadas con igual cobertura. En este punto, también deben utilizarse métodos de cuantificación fluorométricos.

Figura 1. Flujo de trabajo en el laboratorio (*wet-lab*)



3.3. ANÁLISIS BIOINFORMÁTICO (Figura 2)

3.3.1. Evaluación de la calidad y procesamiento de las lecturas

Dado que en el secuenciador introducimos una mezcla equimolar de un conjunto de muestras, el primer paso en el análisis bioinformático es un proceso conocido como "demultiplexado". Generalmente, este proceso lo realiza el propio secuenciador y consiste simplemente en la clasificación del total de lecturas generadas en función de los oligonucleótidos flanqueantes en las mismas, de tal forma que, a partir de un fichero único de lecturas, se generarán tantos ficheros de lecturas como muestras se hayan secuenciado.

Posteriormente, se debe evaluar la calidad de las lecturas obtenidas; FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) y Prinseq-lite (http://prinseq.sourceforge.net/) son recursos frecuentemente utilizados para este fin. El siguiente paso consiste en eliminar las lecturas erróneas que se hayan generado durante la secuenciación. Los errores, pueden afectar a la totalidad de la lectura o solamente a una parte específica de las mismas siendo necesario filtrar y limpiar estos errores con herramientas como Prinseq-lite o Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic) (17).

Finalmente, antes de proceder a cualquier análisis, se eliminarán todas las lecturas de ADN que no sean de nuestro interés, como el ADN de origen humano en aproximaciones metagenómicas, paso que requiere de un elevado consumo de recursos computacionales. Para ello, primero se identificarán todos los organismos que aportan lecturas con herramientas como Kraken2 (http://ccb.jhu.edu/software/kraken2/) (18) o Metaphlan2 (https://huttenhower.sph.harvard.edu/metaphlan2/) (19) y, posteriormente, se filtrarán las que no sean de interés utilizando herramientas como seqtk (https://github.com/lh3/seqtk).

3.3.2. Alineamiento y análisis de variantes

La disponibilidad de un genoma de referencia adecuado sobre el cual alinear las lecturas procesadas es clave a la hora de realizar un alineamiento y su posterior análisis de variantes. En la actualidad, existen diferentes alineadores y programas para el posterior análisis de variantes que pueden integrarse en *pipelines* propios, así como programas que ofrecen una solución integral, entre los que cabe destacar snippy (https://github.com/tseemann/snippy).

Si bien las herramientas de alineamiento y llamada de variantes disponibles han sido ampliamente evaluadas en genética humana, son pocos los trabajos que han evaluado su uso para el estudio de genomas bacterianos. Entre ellos, cabe destacar un trabajo recientemente publicado por Bush y colaboradores en el que se evalúo el desempeño de 209 *pipelines* (alineador/análisis de variantes) en un total de 10 especies bacterianas clínicamente relevantes (20) y en el que los autores terminan concluyendo que, más allá de las herramientas bioinformáticas utilizadas, es la elección de un genoma de referencia adecuado lo que determina la fiabilidad y precisión final de los resultados. Así, los autores recomiendan basar la elección del genoma de referencia entre los potenciales candidatos en base a las distancias estimadas entre cada genoma y las lecturas obtenidas, utilizando para ello programas como Mash (21).

Conceptualmente, el mapeo o alineamiento es sencillo: primero se indexa el genoma de referencia y posteriormente se alinean las lecturas frente a él. BWA y Bowtie2 se hayan entre los alineadores más populares (22-24). Una vez realizado el alineamiento y, antes de proceder al análisis de variantes, es necesario curarlo eliminando posibles errores de secuenciación y manteniendo la variabilidad genética originalmente presente en la muestra utilizando herramientas disponibles en paquetes como *samtools* o GATK (22,25). Finalmente, lo que se obtiene es un archivo en formato *bam* que contiene, para cada una de las lecturas, información sobre el punto del genoma de referencia donde alinea y la calidad con que lo hace. El siguiente paso consiste en visualizar los resultados a partir del fichero *bam* generado y el archivo fasta del genoma de referencia utilizado, para lo que pueden usarse programas como Tablet o IGV (26,27). Con esta visualización se persigue detectar si hay zonas del genoma de referencia que no han quedado cubiertas o, al contrario, si hay zonas en las que hay un exceso de lecturas alineadas. Cabe destacar, que esta visualización puede, por sí sola, tener aplicaciones interesantes en Microbiología Clínica, pudiendo por ejemplo ser utilizada en la predicción de la resistencia a ciertos beta-lactámicos en *Pseudomonas aeruginosa*, ya que la selección de grandes deleciones en ciertas regiones cromosómicas es un mecanismo de resistencia a beta-lactámicos relativamente frecuente en este patógeno (28,29).

Tras visualizar el alineamiento se puede proceder al análisis de variantes de las lecturas generadas con respecto al genoma de referencia, para lo cual pueden emplearse herramientas incluidas en paquetes como samtools o bcftools (30) o soluciones integrales como la ya mencionada herramienta snippy. No debemos olvidar que el resultado de un mapeo representa exclusivamente aquello que cada muestra tiene en común con el genoma usado como referencia.

3.3.3. Ensamblaje de *novo*

Cuando se quiere explorar el genoma accesorio presente en un aislado microbiológico o en el caso de no disponer de un genoma de referencia adecuado para el alineamiento, procederemos al ensamblado de *novo* de las lecturas.

El genoma accesorio de un microorganismo comprende el conjunto de genes cromosómicos no compartidos por todos los miembros de una misma especie y todo el conjunto de genes adquiridos por transferencia horizontal, genes que con frecuencia codifican determinantes de resistencia a los antibióticos y factores de virulencia.

Existen diferentes herramientas que permiten el ensamblado de *novo* de genomas y metagenomas, siendo SPAdes la herramienta más utilizada en la actualidad (31). SPades es una herramienta de línea de comando de fácil utilización que puede trabajar y producir buenos resultados tanto con lecturas cortas como largas, pudiendo incluso realizar ensamblados híbridos combinando ambos tipos de lecturas. El resultado que se obtiene

de la unión de estas lecturas son una serie de secuencias más largas que se denominan *contigs.* Idealmente, un ensamblado de *novo* debería tener tantos *contigs* como estructuras genéticas independientes (cromosomas, plásmidos, etc) estén presentes en la muestra secuenciada; sin embargo, en la práctica la presencia de regiones repetitivas en dichas estructuras genéticas impide la generación de grandes *contigs.* Para evaluar la calidad del ensamblado obtenido disponemos también de herramientas, siendo Quast la más frecuentemente utilizada (32).

3.3.4. Anotación genómica

La anotación genómica es un proceso mediante el cual podemos identificar los genes contenidos en las secuencias generadas y predecir su funcionalidad por homología con genes/proteínas previamente caracterizadas y depositadas en bases de datos (33). Este proceso, aunque no es habitual en la práctica clínica, puede resultar interesante en situaciones en las que con otras aproximaciones no logramos encontrar una explicación al fenotipo observado.

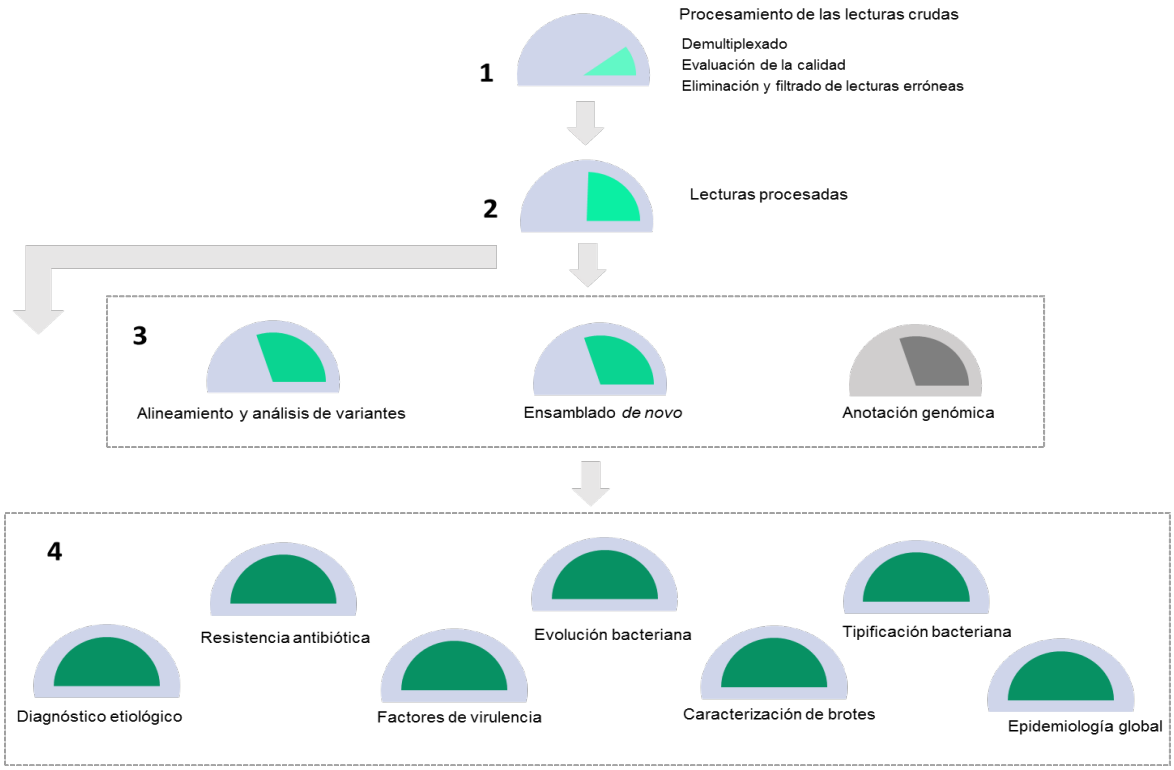
Entre las herramientas más frecuentemente utilizadas para anotación genómica encontramos *pipelines* como Prokka (34) o PGAP (*Prokaryotic Genome Annotation pipeline*) disponible en el *National Center for Biotechnology Information* (35). Estos *pipelines* son versátiles y personalizables, pudiendo combinar la utilización de diferentes estrategias y bases de datos como UniProt/SwissProt, RefSeq, TIGRFAMs (36), FIGfams (37), COG (38) o Pfam (39).

Mediante anotación genómica es posible predecir la función de hasta el 98% de las proteínas codificadas en un genoma bacteriano; no obstante, este porcentaje varía notablemente en función del enfoque escogido (proteínas o dominios), del tamaño del genoma en estudio y de la clasificación taxonómica del mismo (40).

3.4. EMISIÓN DE INFORMES DE RESULTADOS

Es esencial que los resultados se comuniquen de forma clara, consistente y concisa, ya que los informes pueden ser leídos por personal experto e inexperto en secuenciación masiva. Para ello, estos informes sólo deben incluir aquellos resultados que sean clínicamente relevantes e idealmente se deben incluir las bases de datos consultadas.

Figura 2. Flujo de trabajo en el análisis bioinformático (dry-*lab*)



4. APLICACIONES DE LA SECUENCIACIÓN MASIVA A LA MICROBIOLOGÍA CLÍNICA

4.1. DIAGNÓSTICO ETIOLÓGICO DE LAS ENFERMEDADES INFECCIOSAS

La identificación del microorganismo o microorganismos que pueden estar causando una infección es clave para el manejo adecuado del paciente y su ulterior recuperación. En este sentido, los métodos convencionales de diagnóstico microbiológico, directos e indirectos, pueden presentar limitaciones, pudiendo en ocasiones ser demasiado lentos o incluso resultar insuficientes para establecer el diagnóstico etiológico definitivo. Para solventar estas limitaciones, los servicios de Microbiología Clínica han ido paulatinamente incorporando técnicas diagnósticas basadas en la PCR, aunque, al igual que los métodos tradicionales, estos métodos moleculares presentan la desventaja de requerir de una sospecha etiológica previa, hecho que las diferencia de las técnicas de secuenciación masiva.

A grandes rasgos se pueden diferenciar dos estrategias en la aplicación de la secuenciación masiva al diagnóstico etiológico de las enfermedades infecciosas: las estrategias no dirigidas o metagenómicas y las estrategias de secuenciación dirigida (*targeted-sequencing*) (41).

La secuenciación dirigida de amplicones ha sido clásicamente utilizada en Microbiología para la identificación de bacterias y hongos mediante amplificación de regiones universalmente presentes en sus genomas. Así, para la identificación de bacterias se utilizan regiones del gen que codifica la subunidad 16S del ribosoma (7) mientras que para la identificación de hongos se utilizan regiones del ADN espaciador de los genes que codifican las dos subunidades del ribosoma o bien regiones del gen que codifica la subunidad 18S del ribosoma (42).

Como se adelantó, la secuenciación Sanger presenta una baja tasa de error y, por ello, ha sido y es ampliamente utilizada para fines taxonómicos y para el diagnóstico etiológico de las enfermedades infecciosas. No obstante, los electroferogramas generados impiden el diagnóstico etiológico de infecciones polimicrobianas causadas por más de un agente etiológico. Este problema sí puede abordarse mediante técnicas de secuenciación masiva mediante la adición a estos amplicones de los oligonucleótidos necesarios para su identificación y fijación sobre la superficie donde se producirá la secuenciación, estrategia frecuentemente utilizada para estudios de la microbiota (Procedimiento SEIMC 59. Microbiota. https://seimc.org/contenidos/documentoscientificos/procedimientosmicrobiologia/seimc-procedimientomicrobiologia59.pdf). Asimismo, esta estrategia también ha demostrado su potencial en el diagnóstico etiológico de las enfermedades infecciosas. Entre otros, cabe mencionar un trabajo de Salipante y colaboradores en el que se comparan los resultados de esta aproximación con el cultivo microbiológico en una colección de esputos de pacientes con fibrosis quística y abscesos cerebrales, demostrando la utilidad de esta técnica en el diagnóstico de infecciones polimicrobianas en muestras biológicas complejas (43). Como alternativa a la amplificación, la muestra puede enriquecerse capturando las regiones de interés mediante sistemas híbridos de captura. Esta alternativa ha demostrado ser útil en el diagnóstico etiológico de las enfermedades infecciosas, generando resultados compatibles con los hallazgos anatomopatológicos y demostrando una alta sensibilidad en muestras con un ADN altamente degradado (44).

Las estrategias de secuenciación dirigida son más económicas y presentan un menor requerimiento de recursos computacionales aunque pueden ser insuficientes si la carga del microorganismo es muy baja, situación en que la ultrasecuenciación metagenómica sí puede resolver (45). La metagenómica presenta la gran ventaja de no introducir sesgos, teniendo por tanto la capacidad potencial de detectar e identificar cualquier microorganismo presente en la muestra clínica. El elevado coste económico y computacional de la metagenómica puede reducirse mediante la de depleción del ADN de origen humano o aplicándose a muestras clínicas de bajo contenido celular. Respecto a la secuenciación dirigida, la metagenómcia presenta la gran ventaja de permitir estudiar las características asociadas al microorganismo identificado, permitiendo, entre otros, detectar mecanismos de resistencia a los antibióticos o factores de virulencia.

La metagenómica ha demostrado tener un buen rendimiento y ser de gran utilidad en el diagnóstico etiológico de numerosas enfermedades infecciosas como: en infecciones del sistema nervioso central a partir de líquido cefalorraquídeo (46,47), en la infección fúngica invasora en pacientes inmunodeprimidos en muestras de biopsia líquida (48) y en infecciones del tracto respiratorio inferior mediante combinación con datos de la respuesta inmune del hospedador (49).

La implementación de la metagenómica en la rutina diagnóstica requiere de una validación clínica previa. Existen métodos que han sido validados, clínica y analíticamente, como el método Karius, capaz de identificar hasta 1250 microorganismos clínicamente relevantes a partir de los fragmentos de ADN bacteriano presentes en la sangre del paciente (50), método cuyo uso resulta coste-efectivo en el diagnóstico de la infección fúngica invasora (51).

Actualmente, disponemos de herramientas de análisis que nos permiten identificar las especies implicadas en una infección a partir de lecturas procesadas o *contigs*. Algunas de ellas están disponibles en servidores web, resultan muy fáciles de utilizar y apenas requieren de experiencia bioinformática, como KmerFinder (https://cge.cbs.dtu.dk/services/KmerFinder/) (52,53) o megaBLAST (ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/indexed_megablast) (54). Entre los recursos de línea de comando cabe destacar PathoScope 2.0 (https://github.com/PathoScope/PathoScope) (55), que contiene además un programa accesorio llamado Clinical PathoScope diseñado para el análisis de muestras clínicas (55).

4.2. DETECCIÓN DE MECANISMOS DE RESISTENCIA A LOS ANTIBIÓTICOS

Los tratamientos con antibióticos suponen una herramienta fundamental en el manejo de las enfermedades infecciosas. En las últimas décadas, se ha constatado un notable aumento de la resistencia a muchos de los antibióticos disponibles, así como un incremento en el número de microorganismos multirresistentes.

De forma natural, los microorganismos presentan resistencia a algunos antibióticos, pudiendo incrementar su perfil de resistencia mediante la adquisición y selección de mutaciones puntuales, deleciones o inserciones en su cromosoma y/o mediante la adquisición horizontal de determinantes de resistencia a los antibióticos.

La resistencia a los antibióticos está, en general, codificada en el genoma bacteriano, por lo que la aplicación de la secuenciación masiva tiene el potencial teórico de predecir el fenotipo de resistencia en un microorganismo dado. La predicción del fenotipo en base al genotipo es una aplicación especialmente interesante para el estudio de resistencia a los antibióticos en microorganismos de crecimiento lento como *Mycobacterium tuberculosis* o microorganismos no cultivables. Se han explorado diferentes estrategias para predecir la resistencia en base al genotipo. La más sencilla es la basada en reglas, en las que la predicción se basa en la detección de genes y de ciertas mutaciones cromosómicas bien caracterizadas. Estrategias más complejas son las que incorporan modelos estadísticos o el conocido *machine-learning*, estrategias que no precisan de una caracterización y conocimiento previo de los mecanismos que participan en la resistencia antibiótica (56).

Se han desarrollado numerosos modelos para la predicción de la resistencia en *M. tuberculosis*, mostrando especificidades y sensibilidades de hasta el 100% para algunos antibióticos (57-61). Asimismo, se han desarrollado modelos para la predicción de la resistencia en patógenos Gram-negativos como *Escherichia coli*, *Klebsiella pneumoniae, Neisseria gonorrhoeae, Acinetobacter baumannii o Pseudomonas aeruginosa,* demostrando una buena sensibilidad y especificidad (>90%) (62-68). De igual modo, en especies Gram-positivas también encontramos modelos para la predicción de la resistencia en importantes patógenos como *Staphylococcus aureus* o *Streptococcus pneumoniae* (69,70). Si bien es cierto que todavía quedan muchas cuestiones por resolver, especialmente para patógenos como *P. aeruginosa*, con un genoma con alto contenido en genes reguladores y en los que la resistencia mutacional juega un papel fundamental, es probable que, en un futuro no muy lejano, los modelos basados en análisis de secuencias de genoma completo sean utilizados en los servicios de Microbiología Clínica como herramienta para la predicción de la sensibilidad antibiótica.

Por otra parte, la precisión de las técnicas de secuenciación masiva y las múltiples ventajas asociadas a su uso deben invitarnos a valorar la sustitución de los actuales métodos moleculares por la secuenciación de genoma completo como herramienta para la detección y caracterización de genes relacionados con la resistencia antibiótica en microorganismos aislados en la práctica clínica, siendo recomendable su uso como técnica complementaria a las actuales técnicas *gold-standard* (71).

En la tabla 2 se recogen una serie de recursos para la detección y caracterización de genes y mecanismos relacionados con la resistencia antibiótica. Mayoritariamente, estos recursos están enfocados a la detección de determinantes de resistencia adquiridos por vía horizontal aunque algunos de ellos también incluyen discretas colecciones de mutaciones cromosómicas relacionadas con la resistencia a los antibióticos. Estos recursos utilizan distintas herramientas de análisis *(read-based, assembly-based* o *k-mers-based*), no existiendo un consenso en cuanto a qué metodología es mejor y dependiendo su elección mayoritariamente del tipo de análisis (genómico/metagenómico) y de la disponibilidad de recursos computacionales. En general, en análisis de genoma completo suelen preferirse los métodos basados en ensamblaje (*assembly-based*) ya que éstos proporcionan información adicional sobre el entorno en el que se encuentran los genes de resistencia, reservándose los métodos basados en lecturas o k-meros (*read-based* o *k-mers-based*) para el estudio de metagenomas dada su demostrada mayor sensibilidad en la detección de genes presentes en baja abundancia (72).

Tabla 2. Recursos bioinformáticos para la detección de mecanismos de resistencia a los antibióticos.

Recurso	Microorganismo diana	Aplicación	Herramienta de análisis	Base de datos	Detección de mutaciones cromosómicas	Input	Link
ResFinder	General	Genomas Metagenomas	BLAST	Propia	Sí	FASTA (nt) FASTQ (nt)	https://cge.cbs.dtu.dk/services/ResFinder/
CARD	General	Genomas Metagenomas	BLAST, RGI	Propia	Sí	FASTA (nt) FASTA (aa)	https://card.mcmaster.ca/home
NCBI AMRFinderPlus	General	Genomas	BLAST, HMMER	Propia	Sí	FASTA (nt) FASTA (aa) GFF	https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/
ABRICATE	General	Genomas Metagenomas	BLAST	ResFinder CARD ARG-ANNOT NCBI AMRFinder EcOH PlasmidFinder Ecoli_VF VFDB	No	FASTA (nt)	https://github.com/tseemann/abricate
ARIBA	General	Genomas	Minimap, Bowtie2	Derivado de: ARG-ANNOT CARD PlasmidFinder ResFinder VFDB	Sí	FASTQ (nt)	https://github.com/sanger-pathogens/ariba
MEGARes	General	Genomas Metagenomas	BWA	Derivado de: ARG-ANNOT CARD NCBI Lahey Clinic beta-lactamase archive ResFinder	No	FASTQ (nt)	https://megares.meglab.org/
Kmer resistance	General	Genomas Metagenomas	KMA	ResFinder	Sí	FASTQ (nt) FASTA (nt)	https://cge.cbs.dtu.dk/services/KmerResistance-2.2/
LREfinder	Resistencia a linezolid en *Enterococcus* spp.	Genomas	KMA	Propia	Si	FASTA (nt) FASTQ (nt)	https://cge.cbs.dtu.dk/services/LRE-finder/
SCCmec Finder	Búsqueda de elementos SCCmec en *S. aureus*	Genomas	BLAST, KMA	Propia	No	FASTA (nt) FASTQ (nt)	https://cge.cbs.dtu.dk/services/SCCmecFinder/
Mykrobe	Resistencia en *M. tuberculosis* y *S. aureus*	Genomas	Propio (basado en gráficos de De Bruijn)	Propia	Sí	FASTQ (nt)	https://www.mykrobe.com/
DRAGdb	Resistencia mutacional en microorganismos SKAPE y *M. tuberculosis*	Genomas	BLAST	Propia	Sí	FASTA (nt)	http://bicresources.jcbose.ac.in/ssaha4/drag/index.php

4.3. DETECCIÓN DE GENES DE VIRULENCIA

Otro factor que determina la morbilidad y mortalidad en pacientes con una enfermedad infecciosa es la virulencia del microorganismo implicado. De hecho, no todas las especies bacterianas tienen capacidad patogénica en el ser humano y no todos los miembros de una especie se comportan de un mismo modo, estando esta distinta patogenicidad determinada por factores relacionados con el sistema inmune del hospedador y por los factores de virulencia presentes en el genoma del microorganismo causante de la infección.

S. aureus es un patógeno capaz de producir un amplio número de infecciones, estando su patogenia estrechamente relacionada con la presencia de factores de virulencia. La leucocidina de Panton-Valentine es clave en la piomiositis (73) y la presencia de la toxina TSST-1, leucocidinas, enterotoxinas y exfoliatinas lo son para la progresión hacia una osteomiletis en las infecciones de úlceras (74). Asimismo, la invasividad de *S. pneumoniae* y su capacidad de producir infecciones tan graves como la meningitis está en gran parte condicionada por la presencia de ciertos factores de virulencia (75,76).

Durante décadas, numerosas investigaciones se han centrado en la identificación de los factores de virulencia implicados en la patogénesis bacteriana, interés que en los últimos años se ha visto incrementado con el desarrollo de terapias antivirulencia, como la terapia para combatir las bacterias multirresistentes (77). En los últimos años, la generación de millones de secuencias de genomas completos ha revelado una enorme diversidad genética intraespecie e interespecie y su análisis mediante genómica comparativa ha permitido ampliar profundamente el conocimiento existente sobre los elementos genéticos responsables de los fenotipos de virulencia observados.

En la tabla 3 se recogen una serie de recursos de fácil utilización para la detección de elementos genómicos asociados con la virulencia bacteriana.

4.4. DETECCIÓN DE INTEGRONES, TRANSPOSONES, PLÁSMIDOS Y BACTERIÓFAGOS

La evolución bacteriana está en gran parte determinada por la transferencia horizontal de moléculas de ADN entre células procariotas. Esta transferencia depende de una serie de elementos genéticos móviles especializados entre los que se incluyen transposones, plásmidos y ciertos bacteriófagos, y también de estructuras capaces de albergar y reclutar genes de resistencia (integrones). Además de sus genes centrales, estos elementos genéticos móviles contienen una serie de determinantes que proporcionan a la célula huésped cierta ventaja selectiva, portando frecuentemente determinantes de resistencia a los antibióticos y factores de virulencia, entre otros (78).

La secuenciación masiva como herramienta para la detección de los determinantes de resistencia y virulencia supone, por tanto, una gran ventaja frente a otras técnicas moleculares al permitir detectar y conocer el entorno genético de todos los elementos presentes en una sola prueba. En la Tabla 3 se recogen una serie de recursos bioinformáticos que permiten la detección de estos elementos.

Tabla 3. Recursos bioinformáticos para la detección de factores de virulencia y elementos genéticos móviles.

Detección de genes de virulencia						
Recurso	**Microorganismo diana**	**Aplicación**	**Herramienta de análisis**	**Base de datos**	**Input**	**Link**
SPIFinder	*Salmonella* spp.	Genomas	BLAST	Propia	FASTA (nt) FASTQ (nt)	https://cge.cbs.dtu.dk/services/SPIFinder/
VirulenceFinder	*Listeria* spp. *E. coli* *S. aureus* *Enterococcus* spp.	Genomas	BLAST	Propia	FASTA (nt) FASTQ (nt)	https://cge.cbs.dtu.dk/services/VirulenceFinder/
VFanalyzer	General	Genomas	BLAST	Propia	FASTA (nt) FASTA (aa)	http://www.mgc.ac.cn/VFs/main.htm
Detección de genes de elementos genéticos móviles						
Recurso	**Microorganismo diana**	**Aplicación**	**Herramienta de análisis**	**Base de datos**	**Input**	**Link**
IntegronFinder	General	Genomas	HMM	Propia	FASTA (nt) multiFASTA (nt)	https://github.com/gem-pasteur/Integron_Finder
ISseeker	General	Genomas	BLAST	Propia	FASTA (nt)	https://github.com/JCVI-VIRIFX/ISseeker
PlasmidID	General	Genomas	Kmer	Propia	FASTA (nt) FASTQ (nt)	https://github.com/BU-ISCIII/plasmidID
PHAST	General	Genomas	BLAST, GLIMMER	Propia	FASTA (nt)	http://phast.wishartlab.com/

4.5. TIPADO MOLECULAR: EPIDEMIOLOGÍA GLOBAL Y CARACTERIZACIÓN DE BROTES .

La tipificación bacteriana engloba un conjunto de técnicas mediante las cuales se persigue discernir si dos o más aislados están genéticamente relacionados, siendo por tanto una herramienta fundamental para la realización de estudios filogenéticos, vigilancia epidemiológica global e investigación de brotes.

Las primeras técnicas de tipificación bacteriana consistían en la realización de ensayos de caracterización fenotípica, como por ejemplo el estudio de la sensibilidad a los antibióticos, de sensibilidad a fagos o la determinación de los antígenos de superficie presentes. El desarrollo de la biología molecular dio paso a técnicas de tipificación basadas en la restricción del ADN, más reproducibles y resolutivas, y, en general, más rápidas y menos laboriosas. Durante muchos años, la electroforesis de campo pulsado (ECP) ha sido considerada como el *gold-standard* al analizar todo el genoma bacteriano y ofrecer, por tanto, un gran poder de resolución. Sin embargo, la complejidad de la técnica y la difícil comparación de los resultados entre laboratorios promovió el desarrollo de numerosas técnicas de tipificación basadas en la PCR, habiéndose posicionado el *Multi Locus Sequence Typing* (MLST) como la principal técnica de tipificación para la realización de estudios de vigilancia epidemiológica a nivel global.

Además de utilizarse con fines epidemiológicos, y dado que ciertos genotipos se asocian con fenotipos de mayor resistencia a los antibióticos y/o virulencia, la tipificación bacteriana también ha sido ampliamente utilizada como herramienta para la predicción de dichos fenotipos.

Los genomas de bacterias pertenecientes a una misma especie contienen una serie de elementos genéticos comunes y un conjunto de genes que están presentes de forma variable: el genoma central o *core* y el genoma accesorio. Además, dos aislados de una misma especie pueden diferir genéticamente por la adquisición diferencial de mutaciones puntuales (variantes de polimofismo único y pequeñas deleciones e inserciones) en su genoma, así como mediante procesos de recombinación homóloga.

En comparación con las técnicas de tipificación bacteriana clásicas, la secuenciación de genoma completo nos revela la totalidad de genes contenidos, así como las variaciones presentes en ellos, ofreciendo por tanto el máximo poder de resolución y discriminación (79,80). La secuenciación de genoma completo ha sido frecuentemente empleada para la caracterización de brotes de origen alimentario, donde se perfila como la herramienta de vigilancia del futuro (81) al haber demostrado su mayor poder discriminativo en comparación con la ECP y otras técnicas basadas en la PCR en brotes causados por diversas especies como *Clostridium botulinum, Listeria monocytogenes* o *Salmonella* spp., entre otros (82-84). Asimismo, son muchos los trabajos que demuestran la superioridad de la secuenciación del genoma completo en el rastreo de cadenas de transmisión y análisis de brotes en patógenos altamente clonales de gran relevancia clínica como *S. aureus* resistente a la meticilina, *Neisseria meningitidis* o *M. tuberculosis* complex (85-89).

A pesar del cada vez mayor uso de la secuenciación de genoma completo con fines de tipificación, existen todavía cuellos de botella en su aplicación. No obstante, es previsible que en un futuro no muy lejano esta herramienta acabe desplazando a los métodos tradicionales mediante la implementación universal de protocolos y herramientas de análisis estandarizados. De hecho, en los últimos años, se han desarrollado diversas herramientas para el análisis comparativo de genomas entre las que caben destacar Nexstrain, GrapeTree o PopPUNK (90-92) como herramientas para estudios de vigilancia epidemiológica a nivel regional/mundial y PathoSPOT para el rastreo de cadenas de transmisión y caracterización de brotes en el ámbito nosocomial (93), entre otras. Asimismo, en la actualidad se dispone de numerosos recursos que permiten simular *in silico* diversas técnicas clásicas de tipificación bacteriana a partir de secuencias de genoma completo (http://www.genomicepidemiology.org/services/).

4.5.1. Análisis gen a gen

Una de las estrategias más frecuentemente utilizada para la tipificación y comparación de genomas son las basadas en la comparación de alelos. El MLST es un método ampliamente utilizado basado en esta estrategia en el que mediante secuenciación se detectan las variaciones presentes en 7 genes/alelos altamente conservados (genes *housekeeping*), definiendo las combinaciones alélicas encontradas el denominado secuenciotipo. El éxito del MLST como herramienta epidemiológica a nivel global reside principalmente en la fácil comparación de sus resultados. No obstante, este método no es lo suficientemente discriminativo para la caracterización de brotes y cadenas de transmisión, por lo que dada está siendo desplazado por otros métodos como los denominados MLST de genoma central (cgMLST) y MLST de genoma completo (wgMLST) (94,95).

Conceptualmente, estos métodos son semejantes al MLST clásico, aunque ofrecen una resolución más alta al incluirse cientos de *loci* en el análisis (94). Al analizar la totalidad de *loci* presentes, la estrategia gen a gen más resolutiva es el wgMLST, seguido del cgMLST en el que se analiza la totalidad de alelos compartidos por los miembros de una misma especie. Por tanto, el enfoque wgMLST será de elección para el estudio de patógenos muy clonales y se reservará el cgMLST para patógenos de mayor diversidad clonal, en las que no todos los clones comparten la totalidad de los alelos. Al igual que para el MLST, para cada especie bacteriana existe un esquema cgMLST/wgMLST propio. Hasta la fecha se han definido esquemas cg/wgMLST para diversas especies que se pueden encontrar en páginas web como Enterobase (https://enterobase.warwick.ac.uk/), la página web https://www.cgmlst.org/ o en la plataforma taxonómica BIGSdb del Instituto Pasteur (https://bigsdb.pasteur.fr/). Asimismo, también se pueden encontrar estos esquemas en programas de pago, como Ridom seqSphere+ o Bionumerics (95). No obstante, el uso de estos métodos como herramienta de tipificación no está muy extendido en la actualidad.

Además de estos métodos, existen esquemas intermedios como el MLST ribosómico en el que se incluyen 53 alelos codificantes de proteínas ribosomales presentes en la mayoría de las bacterias, que es de uso universal y disponible en https://pubmlst.org/ (94).

4.5.2. Análisis de variantes

Desde el punto de vista biológico, un brote se define cuando en varios individuos se detecta, de forma coincidente en el tiempo y espacio, un mismo agente causal y para los aislados implicados se encuentra una mayor similitud genética que con otros aislados de dicha especie (96). Naturalmente, entre dichos aislados pueden existir diferencias en el número de cambios acumulados cuya extensión depende tanto del tiempo como de la tasa de mutación de los microorganismos implicados, factores por tanto a tener en cuenta en dichos análisis.

Cuando existe una referencia adecuada, de alta calidad y próxima filogenéticamente a los aislados en estudio, la aproximación más sencilla consiste en inferir las diferencias entre aislados mediante comparación y conteo de las variaciones encontradas con respecto al genoma de referencia. Es importante destacar que en la actualidad no existe un valor que sea universalmente válido para ninguna especie que permita afirmar si los aislados estudiados pertenecen o no a un brote; no obstante, cada vez son más las publicaciones y trabajos en los que se proponen puntos de corte para tal fin (Tabla 4). En este punto, es importante también considerar la relevancia de la calidad de la secuenciación para este tipo de análisis al influir notablemente en la posterior llamada de variantes.

Como alternativa a los métodos basados en la comparación con un genoma de referencia pueden utilizarse herramientas como Parsnp (97), en las que el análisis de variantes se focaliza en un alineamiento core que contiene únicamente las regiones del genoma que son comunes en todos los aislados. Asimismo, existen herramientas como kSNP que no utilizan referencia y que basan el análisis de diferencias en k-meros (98). Los k-meros son el conjunto de “palabras” de longitud k presentes en el genoma de un organismo.

Tabla 4. Puntos de corte sugeridos para determinar si dos aislados están o no están genéticamente relacionados. (Adaptado de Schurch *et al.* 2018 (99)).

Microorganismo	Punto de corte	Referencia
Acinetobacter baumannii	≤3	Halachev MR et al, Genome Med 2014 (100)
Campylobacter coli, C. jejuni	≤15	Llarena AK et al, J Clin Microbiol 2017 (101)
Clostridioides difficile	≤4	Kumar N et al, Clin Infect Dis 2016 (102)
Enterococcus faecium	≤16	de Been M et al. J Clin Microbiol 2015 (103)
Escherichia coli	≤10	Roer L et al, J Antimicrob Chemother 2017 (104)
Francisiella tularensis	≤2	Afset J et al, Euro Surveill 2015 (105)
Klebsiella pneumoniae	≤18	Snitkin ES et al, Sci Transl Med 2012 (106)
Legionella pneumophila	≤15	David S et al, J Clin Microbiol 2016 (107)
Listeria monocytogenes	≤3	Kvistholm JA et al, Clin Infect Dis 2016 (108)
Mycobacterium abscessus	≤30	Trovato A et al, Int J Mycobacteriol 2017 (109)
Mycobacterium tuberculosis	≤12	Kohl T et al, J Clin Microbiol 2014 (110)
Neisseria gonorrhoeae	≤14	De Silva D et al, Lancet Infect Dis 2016 (111)
Pseudomonas aeruginosa	≤37	Snyder LA et al, Euro Surveill 2013 (112)
Salmonella enterica	≤4	Bekal S et al, J Clin Microbiol 2016 (113)
Salmonella Typhimurium	≤2	Phillips A et al, BMC Microbiol 2016 (114)
Staphylococcus aureus	≤15	Bartels MD et al, Euro Surveill 2015 (115)

4.5.3. Análisis filogenómico

La filogenómica se entiende como la reconstrucción de la historia evolutiva de una serie de microorganismos a partir del estudio de sus genomas completos, reconstrucción que, junto a los métodos anteriormente expuestos, puede ser de gran utilidad en la caracterización de un brote, ya que si los aislados sospechosos se agrupan formando un grupo monofilético o clado se puede concluir que dichas especies comparten un mismo ancestro común.

Todo análisis filogenómico parte de un alineamiento previo, generalmente obtenido por mapeo frente a una referencia, cuya calidad determinará la validez de los resultados que se obtengan. Por ello, el primer paso del análisis filogenómico consistirá en asegurarnos de que el alineamiento es correcto. Para ello disponemos de herramientas de código libre que nos permiten visualizar y editar los alineamientos como AliView (116), otras que nos permiten eliminar de forma objetiva zonas mal alineadas, como GBlocks o TrimAl (117,118), e incluso algunas que nos permiten detectar la presencia de secuencias erróneas o mal alineadas, como evalmsa (119). Validados los alineamientos de todos los aislados en estudio, generaremos un

archivo multiFasta que contendrá tantas secuencias como aislados estemos analizando, secuencias que tendrán la misma longitud. Idealmente, estas secuencias debieran tener la misma longitud que el genoma utilizado como referencia; no obstante, es frecuente construir secuencias consenso que únicamente contienen las posiciones de la referencia cubiertas en todos los aislados. Posteriormente, procederemos a la construcción del árbol filogenético para lo cual pueden emplearse programas de uso sencillo como MegaX (https://www.megasoftware.net/), o bien otros más complejos, pero que suelen ofrecer mejores resultados, como IQTree (http://www.iqtree.org/). Finalmente, podemos visualizar los árboles filogenéticos generados en visualizadores como FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

5. RETOS Y LIMITACIONES EN LA IMPLEMENTACIÓN DE LA SECUENCIACIÓN MASIVA EN LA RUTINA DEL SERVICIO DE MICROBIOLOGÍA CLÍNICA

A medida que ha aumentado la accesibilidad a la secuenciación de genoma completo, son más los datos que apoyan la implementación de esta tecnología para resolver distintas cuestiones microbiológicas, destacando entre ellas su utilidad para la realización de estudios epidemiológicos a nivel local y global, así como en el estudio y vigilancia de los mecanismos de resistencia antibiótica (120).

Sin embargo, las nuevas tecnologías de secuenciación plantean un cambio de paradigma y, por lo tanto, se deben considerar los retos y desafíos asociados a su implementación (121,122).

El primer punto es decidir para qué se va a utilizar esta tecnología, es decir, es necesario identificar las aplicaciones para las cuales existe suficiente evidencia científica para su aplicación clínica y considerar el valor añadido de esta tecnología frente a los métodos que estén ya implementados en la rutina clínica. En este sentido, el tiempo de respuesta será un factor importante a considerar.

Posteriormente, se deberá decidir qué tecnología resulta más adecuada (Tabla 1), considerando también en la toma de esta decisión las posibles necesidades futuras, dada la gran inversión económica inicial que se requiere. La inversión económica requerida es elevada ya que, además del coste del equipo de secuenciación, son técnicas que precisan de reactivos de coste elevado y además requieren de una infraestructura adicional que incluya ordenadores de alto rendimiento y arreglos de almacenamiento de datos, entre otros. Otro punto clave a considerar en la implementación de la secuenciación masiva es la necesidad de personal altamente cualificado y con cierta experiencia en bioinformática. Así, a la hora de implementar la secuenciación masiva en un Servicio de Microbiología Clínica se deben valorar las capacidades y posibilidades y escoger el modelo operacional (Figura 3) que mejor se adapte a nuestra casuística.

Figura 3. Modelos operacionales de secuenciación masiva implementables en los servicios de Microbiología Clínica

6. CONCLUSIONES

La implementación de la secuenciación masiva en los servicios de Microbiología Clínica está acompañada de numerosos retos y limitaciones. No obstante, su uso está cada vez más extendido, perfilándose como la herramienta del futuro para la vigilancia epidemiológica global. Además, se trata de una herramienta universal aplicable en ausencia de sospecha etiológica previa lo que la hace especialmente interesante en Microbiología Clínica dado el gran número de enfermedades infecciosas de etiología desconocida. Asimismo, la posibilidad de descifrar todo el contenido genómico del agente causal de la infección en una sola prueba la posiciona como la herramienta diagnóstica del futuro.

7. REFERENCIAS

1. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet. 2012; 13:601-612. doi: 10.1038/nrg3226.
2. Fleischmann R, Adams M, White O, Clayton R, Kirkness E, Kerlavage A, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae*. Science 1995; 269:496-512. doi 10.1126/science.7542800.
3. Kchouk M, Gibrat JF, Elloumi M. Generations of sequencing technologies: from first to next generation. Biol Med (Aligarh) 2017; 9:395. doi:10.4172/0974-8369.1000395.
4. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. Nature. 2017; 550:345-353. doi: 10.1038/nature24286.
5. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci USA. 1977; 74:560-4. doi: 10.1073/pnas.74.2.560.
6. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating. Proc Natl Acad Sci. 1977; 74:5463–7. doi: 10.1073/pnas.74.12.5463.
7. Church DL, Cerutti L, Gürtler A, Griener T, Zelazny A, Emler S. Performance and application of 16S rRNA gene cycle sequencing for routine identification of bacteria in the clinical microbiology laboratory. Clin Microbiol Rev. 2020; 33(4):e00053-19. doi: 10.1128/CMR.00053-19.
8. Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. Nat Rev Microbiol. 2015; 13:787-794. doi: 10.1038/nrmicro3565.
9. Food and Drug Administration (FDA) 2016. Infectious disease next generation sequencing based diagnostic devices: microbial identification and detection of antimicrobial resistance and virulence markers. FDA-2016-D-0971.
10. Chiu CY, Miller SA. Clinical metagenomics. Nat Rev Genet. 2019; 20:341-55. doi: 10.1038/s41576-019-0113-7.
11. O'Neil D, Glowatz, H, Schlumpberger M. Ribosomal RNA depletion for efficient use of RNA-seq capacity. Curr Protoc Mol Biol. 2013; 4:4.19.1-4-19.8. doi: 10.1002/0471142727.mb0419s103.
12. Matranga C, Andersen KG, Winnicki S, Busby M, Gladden AD, Tewhey R, et al. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. Genome Biol. 2014; 15(11):519. doi:10.1186.
13. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, et al. Depletion of abundant sequences by hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. Genome Biol. 2016; 17:41. doi: 10.1186/s13059-016-0904-5.
14.Hasan M, Rawat A, Tang P, Jithesh PV, Thomas E, Tan R, et al. Depletion of human DNA in spiked clinical specimens for improvement of sensitivity of pathogen detection by next-generation sequencing. J Clin Microbiol. 2016; 54:919-27. doi: 10.1128/JCM.03050-15
15. Feehery, GR, Yigit E, Oyola SO, Langhorst BW, Schmidt VT, Stewart FJ, et al. A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. PLOS ONE 2013; 8(10):e76096. doi: 10.1371/journal.pone.0076096
16. Thoendel, M, Jeraldo PR, Greenwood-Quaintance KE, Yao JZ, Chia N, Hanssen AD, et al. Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. J Microbiol Methods. 2016; 127:141-145. doi: 10.1016/j.mimet.2016.05.022.
17. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30:2114-2120. doi: 10.1093/bioinformatics/btu170.

18. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014; 15(3):R46. doi: 10.1186/gb-2014-15-3-r46.
19. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015; 12:902-3. doi: 10.1038/nmeth.3589.
20. Bush SJ, Foster D, Eyre DW, Clark EL, De Maio N, Shaw LP, et al. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. Gigascience. 2020; 9(2):giaa007. doi: 10.1093/gigascience/giaa007.
21. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016; 17(1):132. doi: 10.1186/s13059-016-0997-x.
22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754-1760. doi: 10.1093/bioinformatics/btp324.
23. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010; 26:589-595. doi: 10.1093/bioinformatics/btp698.
24. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012; 9:357-9. doi: 10.1038/nmeth.1923
25. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013; 43(1110):11.10.1-11.10.33. doi: 10.1002/0471250953.bi1110s43.
26. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013; 14:178-92. doi: 10.1093/bib/bbs017.
27. Milne I, Bayer M, Stephen G, Cardle L, Marshall D. Tablet: visualizing next-generation sequence assemblies and mappings. Methods Mol Biol. 2016; 1374:253-68. doi: 10.1007/978-1-4939-3167-5_14.
28. Cabot G, Zamorano L, Moyà B, Juan C, Navas A, Blázquez J, Oliver A. Evolution of *Pseudomonas aeruginosa* antimicrobial resistance and fitness under low and high mutation rates. Antimicrob Agents Chemother. 2016; 60:1767-78. doi: 10.1128/AAC.02676-15.
29. Hocquet D, Petitjean M, Rohmer L, Valot B, Kulasekara HD, Bedel E, et al. Pyomelanin-producing *Pseudomonas aeruginosa* selected during chronic infections have a large chromosomal deletion which confers resistance to pyocins. Environ Microbiol. 2016; 18:3482-93. doi: 10.1111/1462-2920.13336.
30. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021; 10(2):giab008. doi: 10.1093/gigascience/giab008.
31. Segerman B. The most frequently used sequencing technologies and assembly methods in different time segments of the bacterial surveillance and RefSeq genome databases. Front Cell Infect Microbiol. 2020; 10:527102. doi: 10.3389/fcimb.2020.527102.
32. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013; 29:1072-5. doi: 10.1093/bioinformatics/btt086.
33. Richardson EJ, Watson M. The automatic annotation of bacterial genomes. Brief Bioinform. 2013; 14:1-12. doi: 10.1093/bib/bbs007.
34. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014; 30:2068-2069. doi: 10.1093/bioinformatics/btu153.
35.Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, et al. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res 2018; 46:D851-D860. doi: 10.1093/nar/gkx1068.
36. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res. 2003; 31:371-3. doi: 10.1093/nar/gkl1043.
37. Meyer F, Overbeek R, Rodriguez A. FIGfams: yet another set of protein families. Nucleic Acids Res. 2009; 37:6643-54. doi: 10.1093/nar/gkp698
38.Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000; 28:33-6. doi: 10.1093/nar/28.1.33.
39. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016; 44:D279–85. doi: 10.1093/nar/gkv1344.
40. Lobb B, Doxey AC. Novel function discovery through sequence and structural data mining. Curr Opin Struct Biol. 2016; 38:53-61. doi: 10.1016/j.sbi.2016.05.017.
41. Dulanto Chiang A, Dekker JP. From the pipeline to the bedside: Advances and challenges in clinical metagenomics. J Infect Dis. 2020; 221(Suppl 3):S331-40. doi: 10.1093/infdis/jiz151.

42. Iwen PC, Hinrichs SH, Rupp ME. Utilization of the internal transcribed spacer regions as molecular targets to detect and identify human fungal pathogens. Med Mycol. 2002; 40:87-109. doi: 10.1080/mmy.40.1.87.109.
43. Salipante SJ, Sengupta DJ, Rosenthal C, Costa G, Spangler J, Sims EH, et al. Rapid 16S rRNA next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. PLoS One. 2013; 8(5):e65226. doi: 10.1371/journal.pone.0065226.
44. Rassoulian Barrett S, Hoffman NG, Rosenthal C, Bryan A, Marshall DA, Lieberman J, et al. Sensitive identification of bacterial DNA in clinical specimens by broad-range 16S rRNA gene enrichment. J Clin Microbiol. 2020; 58(12):e01605-20. doi: 10.1128/JCM.01605-20.
45. Mongkolrattanothai K, Naccache SN, Bender JM, Samayoa E, Pham E, Yu G, et al. Neurobrucellosis: unexpected answer from metagenomic next-generation sequencing. J Pediatric Infect Dis Soc. 2017; 6:393-398. doi: 10.1093/jpids/piw066.
46. Miller S, Naccache SN, Samayoa E, Messacar K, Arevalo S, Federman S, et al. Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. Genome Res. 2019; 29:831-42. doi: 10.1101/gr.238170.118.
47. Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J, et al. Clinical metagenomic sequencing for diagnosis of meningitis and encephalitis. N Engl J Med. 2019; 380:2327-40. doi: 0.1056/NEJMoa1803396.
48. Hill JA, Dalai SC, Hong DK, Ahmed AA, Ho C, Hollemon D, et al. Liquid biopsy for invasive mold infections in hematopoietic cell transplant recipients with pneumonia through next-generation sequencing of microbial cell-free DNA in plasma. Clin Infect Dis. 2020 Oct 19; ciaa1639. doi: 10.1093/cid/ciaa1639.
49. Langelier C, Kalantar KL, Moazed F, Wilson MR, Crawford ED, Deiss T, et al. Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. Proc Natl Acad Sci U S A. 2018; 115:E12353-62. doi: 10.1073/pnas.1809700115.
50. Blauwkamp TA, Thair S, Rosen MJ, Blair L, Lindner MS, Vilfan ID, et al. Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. Nat Microbiol. 2019; 4:663-74. doi: 10.1038/s41564-018-0349-6.
51. MacIntyre AT, Hirst A, Duttagupta R, Hollemon D, Hong DK, Blauwkamp TA. Budget impact of microbial cell-free DNA testing using the Karius® test as an alternative to invasive procedures in immunocompromised patients with suspected invasive fungal infections. Appl Health Econ Health Policy. 2021; 19:231-41. doi: 10.1007/s40258-020-00611-7.
52. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Møller N, et al. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. J Clin Microbiol. 2014; 52:139-46. doi: 10.1128/JCM.02452-13.
53. Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H, et al. Benchmarking of methods for genomic taxonomy. J Clin Microbiol. 2014; 52:1529-39.
54. Chen Y, Ye W, Zhang Y, Xu Y. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res. 2015; 43:7762-68. doi: 10.1093/nar/gkv784.
55. Byrd AL, Perez-Rogers JF, Manimaran S, Castro-Nallar E, Toma I, McCaffrey T, et al. Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. BMC Bioinformatics. 2014; 15:262. doi: 10.1186/1471-2105-15-262.
56. Su M, Satola SW, Read TD. Genome-based prediction of bacterial antibiotic resistance. J Clin Microbiol. 2019; 57(3):e01405-18. doi: 10.1128/JCM.01405-18.
57. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and *Mycobacterium tuberculosis*. Nat Commun. 2015; 6:10063. doi: 10.1038/ncomms10063.
58. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015; 7(1):51. doi: 10.1186/s13073-015-0164-0
59. Pankhurst LJ, Del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J, et al. COMPASS-TB Study Group. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. Lancet Respir Med. 2016; 4:49-58. doi: 10.1016/S2213-2600(15)00466-X.
60. Gygli SM, Keller PM, Ballif M, Blöchliger N, Hömke R, Reinhard M, et al. Whole-genome sequencing for drug resistance profile prediction in *Mycobacterium tuberculosis*. Antimicrob Agents Chemother. 2019; 63(4):e02175-18. doi: 10.1128/AAC.02175-18.

61. Hunt M, Bradley P, Lapierre SG, Heys S, Thomsit M, Hall MB, et al. Antibiotic resistance prediction for *Mycobacterium tuberculosis* from genome sequence data with Mykrobe. Wellcome Open Res. 2019; 4:191. doi: 10.12688/wellcomeopenres.15603.1.
62. Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R, et al. Antimicrobial resistance prediction in PATRIC and RAST. Sci Rep. 2016; 6:27930. doi: 10.1038/srep27930.
63. Moradigaravand D, Palm M, Farewell A, Mustonen V, Warringer J, Parts L. Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data. PLoS Comput Biol. 2018; 14(12):e1006258. doi: 10.1371/journal.pcbi.1006258.
64. Nguyen M, Brettin T, Long SW, Musser JM, Olsen RJ, Olson R, et al. Developing an in silico minimum inhibitory concentration panel test for *Klebsiella pneumoniae*. Sci Rep. 2018; 8(1):421. doi: 10.1038/s41598-017-18972-w.
65. Eyre DW, Golparian D, Unemo M. Prediction of minimum inhibitory concentrations of antimicrobials for *Neisseria gonorrhoeae* using whole-genome sequencing. Methods Mol Biol. 2019; 1997:59-76. doi: 10.1007/978-1-4939-9496-0_4.
66. Demczuk W, Martin I, Sawatzky P, Allen V, Lefebvre B, Hoang L, et al. Equations to predict antimicrobial MICs in *Neisseria gonorrhoeae* using molecular antimicrobial resistance determinants. Antimicrob Agents Chemother. 2020; 64(3):e02005-19. doi: 10.1128/AAC.02005-19.
67. Khaledi A, Weimann A, Schniederjans M, Asgari E, Kuo TH, Oliver A, et al. Predicting antimicrobial resistance in *Pseudomonas aeruginosa* with machine learning-enabled molecular diagnostics. EMBO Mol Med. 2020; 12(3):e10264. doi: 10.15252/emmm.201910264.
68. Pitt ME, Nguyen SH, Duarte TPS, Teng H, Blaskovich MAT, Cooper MA, et al. Evaluating the genome and resistome of extensively drug-resistant *Klebsiella pneumoniae* using native DNA and RNA Nanopore sequencing. Gigascience. 2020; 9(2):giaa002. doi: 10.1093/gigascience/giaa002.
69. Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE Jr, Walker H, et al. Active Bacterial Core surveillance team. Validation of beta-lactam minimum inhibitory concentration predictions for pneumococcal isolates with newly encountered penicillin binding protein (PBP) sequences. BMC Genomics. 2017; 18(1):621. doi: 10.1186/s12864-017-4017-7.
70. Mason A, Foster D, Bradley P, Golubchik T, Doumith M, Gordon NC, et al. Accuracy of different bioinformatics methods in detecting antibiotic resistance and virulence factors from *Staphylococcus aureus* whole-genome sequences. J Clin Microbiol. 2018; 56(9):e01815-17. doi: 10.1128/JCM.01815-17.
71. Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, et al. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. Clin Microbiol Infect. 2017; 23:2-22. doi: 10.1016/j.cmi.2016.11.012.
72. Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. Nat Rev Genet. 2019; 20:356-70. doi:10.1038/s41576-019-0108-4o.
73. Young BC, Earle SG, Soeng S, Sar P, Kumar V, Hor S, et al. Panton-Valentine leucocidin is the key determinant of *Staphylococcus aureus* pyomyositis in a bacterial GWAS. Elife 2019; 8:e42486. doi: 10.7554/eLife.42486.
74. Shettigar K, Murali TS. Virulence factors and clonal diversity of *Staphylococcus aureus* in colonization and wound infection with emphasis on diabetic foot infection. Eur J Clin Microbiol Infect Dis. 2020; 39:2235-46. doi: 10.1007/s10096-020-03984-8.
75. Lees JA, Ferwerda B, Kremer PHC, Wheeler NE, Serón MV, Croucher NJ, et al. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. Nat Commun. 2019; 10(1):2176. doi: 10.1038/s41467-019-09976-3.
76. Cremers AJH, Mobegi FM, van der Gaast-de Jongh C, van Weert M, van Opzeeland FJ, Vehkala M, et al. The contribution of genetic variation of *Streptococcus pneumoniae* to the clinical manifestation of invasive pneumococcal disease. Clin Infect Dis. 2019; 68:61-69. doi: 10.1093/cid/ciy417.
77. Allen JP, Snitkin E, Pincus NB, Hauser AR. Forest and trees: exploring bacterial virulence with genome-wide association studies and machine learning. Trends Microbiol. 2021; Jan 14:S0966-842X(20)30317-6. doi: 10.1016/j.tim.2020.12.002.
78. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol. 2005; 3:722-732. doi: 10.1038/nrmicro1235.
79. Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, García-Cobos S, et al. Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol. 2017; 250:2-10. doi: 10.1016/j.jbiotec.2017.03.035.

80. Neoh HM, Tan XE, Sapri HF, Tan TL. Pulsed-field gel electrophoresis (PFGE): a review of the "gold standard" for bacteria typing and current alternatives. Infect Genet Evol. 2019; 74:103935. doi: 10.1016/j.meegid.2019.103935.
81. World Health Organization (WHO) 2018. Whole genome sequencing for foodborne disease surveillance. ISBN: 978-92-4-151386-9.
82. Kenri T, Sekizuka T, Yamamoto A, Iwaki M, Komiya T, Hatakeyama T, et al. Genetic characterization and comparison of *Clostridium botulinum* isolates from botulism cases in Japan between 2006 and 2011. Appl Environ Microbiol. 2014; 80:6954-64. doi: 10.1128/AEM.02134-14.
83. Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, et al. Prospective whole-genome sequencing enhances national surveillance of *Listeria monocytogenes*. J Clin Microbiol. 2016; 54:333-42. doi: 10.1128/JCM.02344-15.
84. Morganti M, Bolzoni L, Scaltriti E, Casadei G, Carra E, Rossi L, et al. Rise and fall of outbreak-specific clone inside endemic pulsotype of *Salmonella* 4,[5],12:i:-; insights from high-resolution molecular surveillance in Emilia-Romagna, Italy, 2012 to 2015. Euro Surveill. 2018; 23:17-00375. doi: 10.2807/1560-7917.ES.2018.23.13.17-00375.
85. Nikolayevskyy V, Niemann S, Anthony R, van Soolingen D, Tagliani E, Ködmön C, et al. Role and value of whole genome sequencing in studying tuberculosis transmission. Clin Microbiol Infect. 2019; 25:1377-82. doi: 10.1016/j.cmi.2019.03.022.
86. Retchless AC, Fox LM, Maiden MCJ, Smith V, Harrison LH, Glennie L, et al. Toward a global genomic epidemiology of meningococcal disease. J Infect Dis. 2019; 220(220 Suppl 4):S266-73. doi: 10.1093/infdis/jiz279.
87. Coll F, Raven KE, Knight GM, Blane B, Harrison EM, Leek D, et al. Definition of a genetic relatedness cutoff to exclude recent transmission of meticillin-resistant *Staphylococcus aureus*: a genomic epidemiology analysis. Lancet Microbe. 2020; 1(8):e328-e335. doi: 10.1016/S2666-5247(20)30149-X.
88. Itsko M, Retchless AC, Joseph SJ, Norris Turner A, Bazan JA, Sadji AY, et al. Full molecular typing o*f Neisseria meningitidis* directly from clinical specimens for outbreak investigation. J Clin Microbiol. 2020; 58(12):e01780-20. doi: 10.1128/JCM.01780-20.
89. Tagliani E, Anthony R, Kohl TA, de Neeling A, Nikolayevskyy V, Ködmön C, et al. ECDC molecular surveillance project participants. Use of a whole genome sequencing-based approach for *Mycobacterium tuberculosis* surveillance in Europe in 2017-2019: an ECDC pilot study. Eur Respir J. 2021; 57(1):2002272. doi: 10.1183/13993003.02272-2020.
90. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018; 34:4121-3. doi: 10.1093/bioinformatics/bty407.
91. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res. 2019; 29:304-16. doi: 10.1101/gr.241455.118.
92. Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, et al. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Res. 2018; 28:1395-404. doi: 10.1101/gr.232397.117
93. Berbel Caban A, Pak TR, Obla A, Dupper AC, Chacko KI, Fox L, et al. PathoSPOT genomic epidemiology reveals under-the-radar nosocomial outbreaks. Genome Med. 2020; 12(1):96. doi: 10.1186/s13073-020-00798-3.
94. Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013; 11:728-36. doi: 10.1038/nrmicro3093.
95. Vilne B, Meistere I, Grantiņa-Ieviņa L, Ķibilds J. Machine learning approaches for epidemiological investigations of food-borne disease outbreaks. Front Microbiol. 2019;10:1722. doi: 10.3389/fmicb.2019.01722. eCollection 2019.
96. Peacock SJ, Parkhill J, Brown NM. Changing the paradigm for hospital outbreak detection by leading with genomic surveillance of nosocomial pathogens. Microbiology (Reading). 2018; 164:1213-9. doi: 10.1099/mic.0.000700.
97. 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. 2014;15(11):524. doi: 10.1186/s13059-014-0524-x.
98. Gardner SN, Hall BG. When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. PLoS One. 2013; 8(12):e81760. doi: 10.1371/journal.pone.0081760.
99. Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. Clin Microbiol Infect. 2018; 24:350-54. doi: 10.1016/j.cmi.2017.12.016.

100. Halachev MR, Chan JZ, Constantinidou CI, Cumley N, Bradley C, Smith-Banks M, et al. Genomic epidemiology of a protracted hospital outbreak caused by multidrug-resistant *Acinetobacter baumannii* in Birmingham, England. Genome Med 2014; 6:70. doi: 10.1186/s13073-014-0070-x
101. Llarena AK, Taboada E, Rossi M. Whole-genome sequencing in epidemiology of *Campylobacter jejuni* infections. J Clin Microbiol 2017;55:1269e75. doi: 10.1128/JCM.00017-17.
102. Kumar N, Miyajima F, He M, Roberts P, Swale A, Ellison L, et al. Genome-based infection tracking reveals dynamics of *Clostridium difficile* transmission and disease recurrence. Clin Infect Dis 2016;62:746e52. doi: 10.1093/cid/civ1031.
103. de Been M, Pinholt M, Top J, Bletz S, Mellmann A, van SW, et al. Core genome multilocus sequence typing scheme for high-resolution typing of *Enterococcus faecium*. J Clin Microbiol 2015;53:3788e97. doi: 10.1128/JCM.01946-15.
104. Roer L, Hansen F, Thomsen MC, Knudsen JD, Hansen DS, Wang M, et al. WGSbased surveillance of third-generation cephalosporin-resistant *Escherichia coli* from bloodstream infections in Denmark. J Antimicrob Chemother 2017;72:1922e9. doi: 10.1093/jac/dkx092.
105. Afset JE, Larssen KW, Bergh K, Larkeryd A, Sjodin A, Johansson A, et al. Phylogeographical pattern of *Francisella tularensis* in a nationwide outbreak of tularaemia in Norway, 2011. Euro Surveill 2015;20:9e14. PMID: 25990357
106. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, et al. Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. Sci Transl Med 2012;4:148ra116. doi: 10.1126/scitranslmed.3004129
107. David S, Mentasti M, Tewolde R, Aslett M, Harris SR, Afshar B, et al. Evaluation of an optimal epidemiological typing scheme for *Legionella pneumophila* with whole-genome sequence data using validation guidelines. J Clin Microbiol 2016;54:2135e48. doi: 10.1128/JCM.00432-16.
108. Kvistholm Jensen A, Nielsen EM, Björkman JT, Jensen T, Müller L, Persson S, et al. Whole-genome Sequencing Used to Investigate a Nationwide Outbreak of Listeriosis Caused by Ready-to-eat Delicatessen Meat, Denmark, 2014. Clin Infect Dis. 2016; 63:64-70. doi: 10.1093/cid/ciw192.
109. Trovato A, Baldan R, Costa D, Simonetti TM, Cirillo DM, Tortoli E. Molecular typing of *Mycobacterium abscessus* isolated from cystic fibrosis patients. Int J Mycobacteriol 2017;6:138e41. doi: 10.4103/ijmy.ijmy_33_17.
110. Kohl TA, Diel R, Harmsen D, Rothganger J, Walter KM, Merker M, et al. Wholegenomeebased *Mycobacterium tuberculosis* surveillance: a standardized, portable, and expandable approach. J Clin Microbiol 2014;52:2479e86. doi: 10.1128/JCM.00567-14.
111. De Silva D, Peters J, Cole K, Cole MJ, Cresswell F, Dean G, et al. Whole-genome sequencing to determine transmission of *Neisseria gonorrhoeae*: an observational study. Lancet Infect Dis 2016;16:1295e303. doi: 10.1016/S1473-3099(16)30157-8.
112. Snyder LA, Loman NJ, Faraj LA, Levi K, Weinstock G, Boswell TC, et al. Epidemiological investigation of *Pseudomonas aeruginosa* isolates from a six year-long hospital outbreak using high-throughput whole genome sequencing. Euro Surveill 2013;18. doi: 10.2807/1560-7917.es2013.18.42.20611.
113. Bekal S, Berry C, Reimer AR, Van DG, Beaudry G, Fournier E, et al. Usefulness of high-quality core genome single-nucleotide variant analysis for subtyping the highly clonal and the most prevalent *Salmonella enterica* serovar Heidelberg clone in the context of outbreak investigations. J Clin Microbiol 2016;54:289e95. doi: 10.1128/JCM.02200-15.
114. Phillips A, Sotomayor C, Wang Q, Holmes N, Furlong C, Ward K, et al. Whole genome sequencing of *Salmonella Typhimurium* illuminates distinct outbreaks caused by an endemic multi-locus variable number tandem repeat analysis type in Australia, 2014. BMC Microbiol 2016;16:211. doi: 10.1186/s12866-016-0831-3.
115. Bartels MD, Larner-Svensson H, Meiniche H, Kristoffersen K, Schonning K, Nielsen JB, et al. Monitoring meticillin resistant *Staphylococcus aureus* and its spread in Copenhagen, Denmark, 2013, through routine whole genome sequencing. Euro Surveill 2015;20. doi: 10.2807/1560-7917.es2015.20.17.21112.
116. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014; 30:3276-3278. doi: 10.1093/bioinformatics/btu531.
117. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000; 17:540-552. doi: 10.1093/oxfordjournals.molbev.a026334.
118. Di Franco A, Poujol R, Baurain D, Philippe H. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. BMC Evol Biol. 2019; 19(1):21. doi: 10.1186/s12862-019-1350-2.

119. Chiner-Oms A, González-Candelas F, EvalMSA: a program to evaluate multiple sequence alignments and detect outliers. Evol Bioinform Online. 2016;12:277-84. doi: 10.4137/EBO.S40583. eCollection 2016.
120. Black A, MacCannell DR, Sibley TR, Bedford T. Ten recommendations for supporting open pathogen genomic analysis in public health. Nat Med. 2020; 26:832-41. doi: 10.1038/s41591-020-0935-z.
121. Fricke WF, Rasko DA. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. Nat Rev Genet. 2014; 15:49-55. doi: 10.1038/nrg3624.
122. Rossen JWA, Friedrich AW, Moran-Gilad J; ESCMID Study Group for Genomic and Molecular Diagnostics (ESGMD). Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. Clin Microbiol Infect. 2018; 24:355-360. doi: 10.1016/j.cmi.2017.11.001.

Servicio / Unidad de Microbiología Hospital.....	Procesamiento de muestras para secuenciación de genoma completo	PNT-TSM-01	
		Edición Nº 01	Página 1 de 11

PNT-TSM-01
Procesamiento de muestras para secuenciación de genoma completo

ELABORADO		REVISADO Y APROBADO	
Nombre / Firma	Fecha	Nombre / Firma	Fecha

EDICIÓN	FECHA	ALCANCE DE LAS MODIFICACIONES
01	2021	Edición inicial

Servicio / Unidad de Microbiología Hospital.....	Procesamiento de muestras para secuenciación de genoma completo	PNT-TSM-01	
		Edición Nº 01	Página 2 de 11

1. PROPÓSITO Y ALCANCE

En este procedimiento se describe la metodología a seguir para la preparación de librerías para secuenciación de genoma completo de bacterias aisladas en cultivo microbiológico.

El presente procedimiento es aplicable a todos los servicios de Microbiología que dispongan o tengan acceso a la plataforma de secuenciación Miseq de Illumina®.

2. FUNDAMENTO

El genoma bacteriano se refiere al conjunto total de ADN que posee una bacteria, y comprende tanto su cromosoma, de doble cadena, circular y cerrado, como todos sus elementos genéticos extracromosómicos adquiridos. Dicho genoma bacteriano varía entre especies, así como entre miembros de una misma especie y su expresión resulta en fenotipos distintos. En los últimos años, la mayor accesibilidad a las nuevas tecnologías de secuenciación está revolucionando la Microbiología Clínica al permitir éstas obtener la secuencia de genoma completa en una prueba única y universal.

Actualmente existen diferentes plataformas de secuenciación que permiten obtener la secuencia de genoma completo. Estas plataformas incorporan diferentes estrategias y tecnologías que derivan en distintas prestaciones (ver Tabla 1 del documento científico de este procedimiento). Entre ellas, destacan las plataformas Illumina® ya que resultan ideales para la mayoría de las aplicaciones de la secuenciación de genoma completo en Microbiología Clínica (1).

La secuenciación de genoma completo puede dividirse en dos procesos, en inglés denominados *wet-lab* y *dry-lab*. El término *wet-lab* hace referencia a todos los pasos que ocurren en el laboratorio antes de introducir las muestras en el secuenciador, pasos que se detallan en el presente procedimiento. El término *dry-lab* hace referencia al análisis bioinformático posterior, siendo objeto del PNT-TSM-02 de este procedimiento.

3. DOCUMENTOS DE CONSULTA

1. García-Lechuz Moya JM, González López JJ, Orta Mira N, Sánchez Romero MI. Recogida, transporte y procesamiento general de las muestras en el Laboratorio de Microbiología. 1b. Sánchez Romero MI (coordinadora). Procedimientos en Microbiología Clínica. Cercenado Mansilla E, Cantón Moreno R (editores). Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica (SEIMC). 2017.

2. Alados Arboledas JC, Gómez García de la Pedrosa E, Leiva León J, Pérez Sáenz JL, Rojo Molinero E. Seguridad en el laboratorio de Microbiología Clínica. 10a. Pérez Sáenz JL (coordinador). Procedimientos en Microbiología Clinica. Cercenado Mansilla E, Cantón Moreno R (editores). Sociedad Española de Enfermedades Infecciosas y Microbiologia Clínica (SEIMC). 2014.

3. Alados Arboledas JC, Fedele G, Ocete Mochón MD, Rodríguez-Iglesias MA. Gestión de solicitudes e informes en Microbiología y conservación del material biológico. 2018. 63. Rodríguez Iglesias MA (coordinador). Procedimientos en Microbiología Clínica. Cercenado Mansilla E, Cantón Moreno R (editores). Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica (SEIMC). 2018.

4. Qubit™ 1X dsDNA HS Assay Kits. User guide. Thermo Fisher Scientific Inc. 2020. Disponible en: https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2FMAN0017455_Qubit_1X_dsDNA_HS_Assay_Kit_UG.pdf&title=VXNIciBHdWlkZTogUXViaXQgMVggZHNETkEgSFMgQXNzYXkgS2l0

5. Quant-iT™ PicoGreen ® dsDNA Reagent and Kits. User guide. Invitrogen. 2017. Disponible en: https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2Fmp07581.pdf&title=UXVhbnQtaVQgUGljb0dyZWVuIGRzRE5BIFJlYWdlbnQgYW5kIEtpdHM

6. Illumina DNA Prep. Reference Guide. Nº de documento #1000000025416v09. 2020. Disponible en: https://emea.support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/illumina_prep/illumina-dna-prep-reference-guide-1000000025416-09.pdf

7. Index Adapters. Pooling Guide. Nº de documento #1000000041074v10. 2020. Disponible en: https://emea.support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/index-adapters-pooling-guide-1000000041074-10.pdf

8.Illumina Adapters Sequences. Nº de documento #1000000002694v15. 2021. Disponible en: https://emea.support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-adapter-sequences-1000000002694-15.pdf

9. Agilent High Sensitivity DNA Kit Guide. Nº de documento G2938- 90321 Rev. B. 2013
Disponible en: https://www.agilent.com/cs/library/usermanuals/Public/G2938-90321_SensitivityDNA_KG_EN.pdf

10. Miseq. Guía del Sistema. Nº de documento 15027617v06 ESP Nº de material 20000262. 2021. Disponible en: https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/translations/miseq-system-guide-15027617-esp.pdf

11. Miseq system. Denature and dilute library guide. Nº de documento #15039740v10. 2019. Disponible en: https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-denature-dilute-libraries-guide-15039740-10.pdf

12. Cluster optimization. Overview Guide. Nº de documento #1000000071511v00. 2019. Disponible en: https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/cluster-optimization-overview-guide-1000000071511-00.pdf

4. MUESTRAS

El presente procedimiento se realizará sobre cultivos bacterianos clonales y puros. Dichos cultivos podrán guardarse en nevera (4ºC) hasta un máximo de 48 h antes de proceder a la extracción del ADN genómico total o, alternativamente, podrán congelarse a -20ºC o a -80ºC, con o sin adición de criopreservantes, hasta el momento de la extracción.

5. REACTIVOS Y PRODUCTOS

5.1. PREPARACIÓN DE CULTIVOS CLONALES PUROS

Para la obtención del cultivo clonal se precisan medios de cultivo como agar sangre o agar chocolate. La elección del medio de cultivo se realizará en función de los requerimientos nutricionales de la bacteria objeto de secuenciación.

5.2. EXTRACCION DE ADN GENÓMICO

Se pueden utilizar sistemas comerciales de extracción de ADN genómico total, manuales o automatizados. Para obtener un buen rendimiento en la extracción es necesario asegurar la lisis total de la pared bacteriana, para lo que generalmente se incluirá un paso de incubación con lisozima o lisostafina en la extracción de bacterias Gram-positivas. Asimismo, es esencial asegurarse de que el sistema comercial de extracción no incluye ningún reactivo cuya formulación tenga sustancias que puedan interferir con la posterior preparación de librerías y secuenciación. Con frecuencia, los reactivos de elución incorporan EDTA en su formulación por lo que será necesario sustituirlo por Tris 10 mM pH=8,5.

5.3. PREPARACIÓN DE LIBRERÍAS GENÓMICAS

Actualmente existen diferentes sistemas comerciales de preparación de librerías compatibles con la plataforma Miseq de Illumina (2,3). Generalmente, estos sistemas no incluyen los oligonucleótidos necesarios para el marcaje inequívoco de las muestras, sino que estos se adquieren de forma independiente. Estos oligonucleótidos incorporan una secuencia universal que permite la adición última de las librerías sobre la superficie bidimensional donde se produce la secuenciación por lo que es imprescindible asegurarse de la compatibilidad de los mismos con el sistema escogido para la preparación de librerías y con la plataforma y reactivos de secuenciación.

A pesar de poder utilizarse sistemas comerciales de preparación de librerías de diferentes casas comerciales es recomendable trabajar con los *kits* de Illumina ya que el análisis bioinformático posterior resulta más sencillo. Una buena y versátil elección es el *kit* de preparación de librerías *Illumina DNA Prep* ya que es fácil de utilizar, permite trabajar con un rango amplio de ADN inicial (1-500 ng) y ha demostrado buenos resultados con independencia del contenido CG del genoma bacteriano a secuenciar (4). Asimismo, se recomienda utilizar los oligonucleótidos *Nextera™ DNA Combinatorial Dual* (CD) *Indexes*, ya que éstos vienen ya combinados en una placa de 96 pocillos de un sólo uso por lo que se minimizan los posibles errores humanos durante el marcaje de las muestras, así como posibles contaminaciones.

El presente protocolo de trabajo está basado en el kit de preparación de librerías *Illumina DNA Prep* (referencia comercial Illumina 20018704 ó 20018705) y los oligonucleótidos *Nextera™ DNA CD Indexes* (referencia comercial Illumina 20018707 ó 20018708).

Además de dichos reactivos se necesitan de reactivos que permitan la cuantificación precisa de ADN de doble hebra como *Quant-iT™ Picogreen® dsDNA Assay Kit high sensitivity* (referencia comercial ThermoFisher P7589 o P11496) o *Qubit 1X dsDNA HS Assay Kit* (referencia comercial ThermoFisher Q33230 o Q33231), dependiendo su elección de la disponibilidad de equipos y de las muestras a secuenciar en paralelo.

5.4. NORMALIZACIÓN, PREPARACIÓN DE LA MEZCLA EQUIMOLAR DE LIBRERÍAS ("*POOLING*") Y CARGA EN EL SECUENCIADOR

Para la cuantificación final de las librerías se utilizará el mismo reactivo de cuantificación que se ha utilizado en el apartado anterior. Para la normalización y preparación de la mezcla equimolar se necesitan además reactivos para determinar el tamaño de los fragmentos de la librería como, por ejemplo, el *Agilent High Sensitivity DNA Kit* (referencia comercial Agilent 5067-4626).

Para la carga en el Miseq necesitaremos un cartucho de reactivos de secuenciación y *flow-cell* que sean compatibles con el equipo. Para secuenciación de genoma completo disponemos de dos opciones: *MiSeq Reagent kit* v2 de 500 ciclos (referencia comercial Illumina MS-102-2003) o *Miseq Reagent kit* v3 de 600 ciclos (referencia comercial Illumina MS-102-3003), ya que estos generan las lecturas más largas y permiten incluir un número de muestras aceptable. En este protocolo se utilizará *Miseq Reagent kit* v3 de 600 ciclos.

Opcionalmente puede añadirse un control del proceso de secuenciación, el *PhiX Control v3* (referencia comercial Illumina FC-110-3001).

Tabla 1. Resumen del material y reactivos necesarios para el procesamiento de las muestras para secuenciación masiva

PROCESO	REACTIVOS
Preparación de los cultivos clonales	• Placas de agar sangre, agar chocolate o Mueller-Hinton
Extracción de ADN genómico	• Sistema comercial de extracción de ADN genómico total • TRIS 10 mM pH=8,5
Preparación de librerías genómicas	• *Illumina® DNA Prep, (M) Tagmentation* (24 ó 96 muestras) • *Nextera™ DNA CD Indexes* (24 ó 96) • Agua libre de nucleasas • Etanol 80% • Hielo • *Quant-iT™ dsDNA Assay Kit high sensitivity o Qubit 1X dsDNA HS Assay Kit*
Normalización, preparación de la mezcla equimolar ("*pooling*") y carga en el secuenciador	• *Agilent High Sensitivity DNA Kit* • *Quant-iT™ dsDNA Assay Kit high sensitivity o Qubit 1X dsDNA HS Assay Kit* • *MiSeq Reagent Kit v3 (600-cycle)* • *PhiX control v3 (opcional)* • NaOH 0,2N • Hielo

6. APARATOS Y MATERIALES

6.1. APARATOS

- Nevera
- Congelador -20ºC
- Estufa
- Balanza
- Cabina de seguridad
- Agitador orbital tipo vórtex
- Centrífuga con refrigeración para tubos y/o microplacas
- Espectrofotómetro de luz UV Nanodrop
- Espectrómetro de fluorescencia con lector de microplacas (Quant-iT™ dsDNA Assay Kit high sensitivity) o fluorímetro Qubit (Qubit 1X dsDNA HS Assay Kit)
- Agilent 2100 Bioanalyzer
- Termociclador
- Termobloque
- Pipetas calibradas de volumen variable
- Soporte magnético de tubos y/o microplacas
- Secuenciador de mesa Miseq (Illumina)
- Sistema de extracción de ADN automático (opcional)

6.2. MATERIALES

- Asas de siembra
- Puntas para pipetas con filtro de diferentes volúmenes
- Reservorios de reactivos para pipeta multicanal libre de nucleasas
- Tubos de plástico de fondo cónico estériles tipo Eppendorf de 1,7 mL
- Microplacas de 96 pocillos de 1 ó 2 mL
- Microplacas de 96 pocillos o tubos de plástico 0,2 mL para reacción de PCR
- Film para microplacas
- Film aluminio para microplacas
- Microplacas para ensayos basados en fluorescencia de 96 pocillos o tubos 0,5mL para fluorímetro Qubit
- Gradillas
- Contenedores de residuos
- Guantes de acetonitrilo o similares

7. PROCESAMIENTO

Durante el procesamiento de las muestras para secuenciación masiva se debe disponer de 3 áreas diferenciadas de trabajo:

- Área 1: área de preparación de reactivos
- Área 2: área de manipulación de muestras pre-amplificación
- Área 3: área de manipulación de muestras post-amplificación

7.1.PREPARACIÓN DE CULTIVOS CLONALES

Los cultivos clonales se prepararán en el área 2 mediante subcultivo de 1 colonia de la bacteria a secuenciar. Los medios y las condiciones de incubación dependerán de los requerimientos de dicha bacteria.

7.2.EXTRACCIÓN DE ADN GENÓMICO

La extracción del ADN genómico se realizará en el área 2 siguiendo las instrucciones del sistema comercial escogido con excepción del último paso de elución en el que se sustituirá el eluyente por TRIS 10 mM pH=8,5.

Se evaluará la calidad del ADN extraído con un espectrofotómetro de luz UV Nanodrop midiendo los ratios de absorbancia 260/280 nm y 260/230 nm. Un valor de 1,8-2,0 en el ratio 260/280 nm es indicativo de que la muestra únicamente contiene ADN y un valor de 2,0-2,2 en el ratio 260/230 nm es indicativo de la ausencia de contaminantes; cualquier extracto fuera de estos rangos deberá ser excluido.

7.3.PREPARACIÓN DE LIBRERÍAS GENÓMICAS

7.3.1. Número de muestras por carrera de secuenciación

El número de muestras que pueden secuenciarse en paralelo depende básicamente de 3 aspectos: (1) el cartucho para realizar la secuenciación, (2) el tamaño del genoma de las bacterias a secuenciar y (3) la profundidad de secuenciación deseada.

En este protocolo utilizaremos el cartucho de reactivos *MiSeq Reagent Kit* v3 (600-cycle) ya que nos permite secuenciar fragmentos más largos. Por tanto, para calcular el número de muestras que pueden introducirse en la carrera queda saber el tamaño del genoma y la profundidad de secuenciación deseada, siendo recomendable que ésta sea superior a 50x.

Una vez se definen estos parámetros se puede utilizar la herramienta *Sequencing Coverage Calculator de Illumina*® para calcular el número de muestras a secuenciar (https://emea.support.illumina.com/downloads/sequencing_coverage_calculator.html).

7.3.2. Puntos de parada seguros

La preparación de librerías es un proceso que suele requerir de 2-3 días de trabajo, especialmente cuando se trabaja con 96 muestras. Existen ciertos puntos en los que se puede detener el procedimiento sin afectar la calidad final de la carrera y de las secuencias, los denominados “puntos de parada seguros”.

7.3.3. Ajuste inicial de la concentración de ADN genómico

El primer paso de preparación de librerías consiste en medir y ajustar la cantidad del ADN de partida. Para ello, en el área 2, se medirá la concentración del ADN de doble hebra de cada uno de los extractos con *Quant-iT™ dsDNA Assay Kit high sensitivity* o *Qubit 1X dsDNA HS Assay Kit* (según disponibilidad) y se ajustará a una concentración de 3,4 ng/microL con TRIS 10 mM pH=8,5 (volumen final ADN ajustado > 30 microL).

7.3.4. Tagmentación del ADN genómico

En el área 1, se dejan atemperar los reactivos durante 30 minutos y se prepara la mezcla de tagmentación, añadiendo 11 microL de *BLT* previamente vorteadas (transposomas unidas a bolitas magnéticas) y 11 microL del buffer *TB1* (*tagment buffer*) por cada muestra. Se agita bien en vortex la mezcla y se hacen alícuotas de 20 microL de la mezcla en tubos de PCR de 0,2 mL o en microplaca (según el número de muestras).

Se pasa ahora al área de trabajo 2 y se añade a cada tubo/pocillo 30 microL del extracto de ADN ajustado en el paso anterior, mezclando mediante pipeteo (sellar la microplaca con film de aluminio) y dar un golpe de centrífuga. Posteriormente, se incubará a 55ºC durante 15 minutos en termociclador, siendo la temperatura de la tapa de 100ºC, y al final enfriar a 10ºC.

Cuando se alcancen los 10ºC, se sacan los tubos/microplaca del termociclador y se añaden 10 microL de *TSB* (reactivo de parada de la tagmentación) a cada tubo/pocillo, mezclando mediante pipeteo (sellar la microplaca con film de aluminio) e incubar a 37ºC durante 15 minutos en el termociclador, siendo la temperatura de la tapa de 100ºC, y al final enfriar a 10ºC.

Posteriormente, se retiran los tubos o microplaca del termociclador y se colocan en el soporte magnético para separar las bolitas magnéticas del reactivo. Cuando estas se separen por completo (3 minutos aprox.) y el sobrenadante sea transparente se procederá a retirarlo con la pipeta.

Retirar los tubos/microplaca del soporte magnético y lavar las bolitas 2 veces de la siguiente manera: se añaden 100 mciroL de *TWB* (solución de lavado de la tagmentación) y pipetear hasta resuspender las bolitas, colocar en el soporte magnético y retirar el sobrenadante cuando se hayan separado (3 minutos aprox.). Finalmente, añadir 100 microL de *TWB* sobre las bolitas magnéticas para que no se sequen y se dejará en el soporte hasta que sean utilizadas.

7.3.5. Amplificación y marcaje del ADN fragmentado

En el área 1, se descongelará el reactivo *EPM* (mezcla maestra concentrada) en hielo y mezclar por inversión. A continuación, preparar la mezcla maestra añadiendo 22 microL de *EPM* y 22 microL de agua libre de nucleasas por muestra, agitar en vortex y dar un pulso de centrífuga.

Volver al área 2, retirar e*l TWB* de los tubos/microplaca y retirar el imán. Añadir 40 microL de mezcla maestra, descongelar los índices *Nextera* DNA CD Indexes y añadir 10 microL a la mezcla de reacción, mezclar mediante pipeteo (sellar la microplaca con film de aluminio), dar un pulso de centrífuga y llevar al termociclador con el programa indicado en la tabla 2. Es muy importante identificar y anotar qué índices se añaden a cada muestra y no añadir nunca la misma combinación a 2 muestras distintas. La placa *Nextera* DNA CD *Indexes* contiene 96 combinaciones diferentes sirviendo, por lo tanto, para un máximo de 96 muestras. Cuando se trabaje con un número de muestras inferior a 96, se recomienda utilizarlos por columnas.

Tabla 2. Programa de PCR

68ºC, 3 minutos	
98ºC, 3 minutos	
98ºC, 45 segundos	x5 ciclos
62ºC, 30 segundos	
68ºC, 2 minutos	
10ºC, ∞	

PUNTO DE PARADA SEGURO

En este punto se puede detener el proceso. Los productos de la amplificación pueden guardarse hasta un máximo de 3 días a una temperatura de entre 2-8ºC.

7.3.6. Limpieza de los productos de amplificación

En el área 3, colocar los tubos/microplaca, previamente centrifugada, sobre el soporte magnético y, una vez se hayan separado las bolitas magnéticas, transferir 45 microL del sobrenadante a tubos de 1,7 mL o a una placa de pocillos de 1 ó 2 mL.

Añadir a cada tubo/pocillo 81 microL de *SPB* (bolitas de purificación) y mezclar mediante pipeteo suave. Las bolitas de purificación se deben vortear muy bien y utilizarse siempre a temperatura ambiente, por lo que se sacarán de la nevera 30 minutos antes de su uso. Se incubará a temperatura ambiente durante 5 minutos y, transcurrido este tiempo, colocar los tubos/microplaca sobre el soporte magnético.

Transcurridos los 5 minutos retirar el sobrenadante y lavar 2 veces de la siguiente manera: se añaden 200 microL de etanol recién preparado al 80%, esperar 30 segundos y retirar el etanol. Es importante retirar bien el etanol y dejar secar las bolitas durante 5 minutos antes de seguir el proceso. Una vez que las bolitas dejen de brillar, añadir 32 microL de *RSB* (buffer de resuspensión) a cada tubo/pocillo y mezclar bien mediante pipeteo e incubar 2 minutos a temperatura ambiente. Finalmente, volver a colocar en el imán hasta que las bolitas se separen (3 minutos aprox.) y transferir 30 microL del sobrenadante a unos tubos/placa nuevos.

PUNTO DE PARADA SEGURO

En este punto se puede detener el proceso, las librerías pueden guardarse hasta 30 días a una temperatura de -20ºC.

7.3.7. Comprobación de las librerías

Antes de proceder a la normalización y carga en el secuenciador es conveniente comprobar el tamaño de los fragmentos contenidos en las librerías ya que éstos determinan la concentración (ng/microL) a la que se deben ajustar en la posterior normalización. Esta comprobación puede realizarse en un *Agilent* 2100 Bioanalyzer con los *kits* de alta sensibilidad siguiendo las recomendaciones del fabricante.

El tamaño medio de los fragmentos que se obtienen ha de ser de alrededor de 600 pares de bases. Tamaños superiores a 1000 pares de bases disminuirán notablemente el rendimiento y la calidad de la secuenciación, no siendo recomendable seguir adelante en el proceso.

7.4. NORMALIZACIÓN, PREPARACIÓN DE LA MEZCLA EQUIMOLAR ("*POOLING*") Y CARGA EN EL SECUENCIADOR

En el área 3, medir la concentración de ADN de doble cadena de las librerías, con *Quant-iT™ dsDNA Assay Kit high sensitivity* o con *Qubit 1X dsDNA HS Assay Kit*, según disponibilidad, y normalizar con RSB o TRIS 10 mM pH=8,5 a una concentración final de 4 nM. Para transformar de ng/microL a nM, utilizar la siguiente fórmula:

nM = (1500/tamaño medio de los fragmentos) x [ ]muestra en ng/microL

Una vez que las librerías estén normalizadas, coger el mismo volumen (> 3 microL) de cada una de ellas y mezclarlas en un nuevo tubo de 1,7 mL. De esta mezcla de librerías se toman 5 microL y mezclar con 5 microL de NaOH 0,2 N recién preparado, dejar incubar 5 minutos a temperatura ambiente y añadir 990 microL de HT1 frío (solución de hibridación). De esta forma, se obtendrá 1 mL de una mezcla equimolar de librerías a una concentración 20 pM.

En paralelo, se puede preparar el control *PhiX v3* (opcional). Para ello, mezclar 2 microL de PhiX 10 nM con 3 microL de TRIS 10 mM pH=8,5 y añadir 5 microL de NaOH 0,2 N recién preparado, dejar incubar 5 minutos a temperatura ambiente y añadir 990 microL de *HT1* frío (solución de hibridación). De esta forma, se obtendrá 1 mL de PhiX a una concentración 20 pM.

Una vez que se tiene la mezcla de librerías y el control *PhiX* desnaturalizados, mezclar 594 microL de la mezcla de librerías 20 pM con 6 microL de *PhiX 20pM* en un nuevo tubo de 1,7 mL e incubar 2 minutos a 96ºC en termobloque. Dejarlo en hielo durante 5 minutos y proceder a cargarlo en el cartucho (previamente descongelado) y en el secuenciador.

Para cargar en el secuenciador hay que preparar la hoja de trabajo, para lo que es necesario descargar el software *Illumina Experiment Manager* (disponible en https://emea.support.illumina.com/sequencing/sequencing_software/experiment_manager/downloads.html). Seguir una serie de ventanas en las que hay que escoger: la plataforma de secuenciación (*Miseq)*, el tipo de análisis (*Only fastq*), el identificador del cartucho de reactivos, el número de ciclos (301), si deseamos que el *software* del secuenciador procese las lecturas y elimine los índices, y la relación de muestras e índices (posiciones de la placa *Nextera DNA CD Indexes* utilizados).

8. OBTENCIÓN Y EXPRESIÓN DE LOS RESULTADOS

Al finalizar la carrera de secuenciación aparecerá un cuadro resumen del proceso. Se puede visualizar el número de *clusters* formados sobre la superficie bidimensional (*flow-cell*) durante la amplificación clonal *in vitro*, el número de clusters que el sistema de captura de señal es capaz de diferenciar y registrar señal y la calidad de las lecturas generadas en dichos *clusters*.

El número de *clusters* que pasan el filtro determina en gran medida la calidad de las lecturas que se generan. La formación de *clusters* en exceso, como consecuencia de una concentración de carga elevada, dificultará la captación de señal y el número de *clusters* que pasan el filtro disminuirá. De igual forma, si apenas se forman *clusters* se obtendrán pocas lecturas que impedirán el análisis posterior.

9. RESPONSABILIDADES

- Deben estar descritas en el manual general de organización del laboratorio de Microbiología y en las fichas de descripción de los puestos de trabajo.

- El procedimiento deberá llevarse a cabo por personal técnico cualificado con un entrenamiento y formación específica, deberá respetar las normas de seguridad e higiene del laboratorio y realizar una correcta gestión de residuos.

- La supervisión de la técnica y evaluación de resultados deberá realizarla el facultativo especialista responsable.

10. ANOTACIONES AL PROCEDIMIENTO

- Existen ciertos aspectos en la preparación de librerías que determinan notablemente la calidad de los resultados y en los que debe prestarse especial atención. Realizar una correcta cuantificación del ADN de doble cadena, respetar los tiempos y temperaturas durante la tagmentación y una correcta normalización final influye de forma determinante en los resultados.

- No es necesario realizar la comprobación del tamaño de fragmentos en todas las muestras y carreras de secuenciación. Aunque sí es muy recomendable cuando es la primera vez que se secuencia una especie bacteriana en concreto, ya que el tamaño medio podría variar y afectar a la normalización de librerías.

- Es necesario leer con atención los documentos de consulta indicados en el apartado 3; dedicando especial atención a las normas de seguridad, uso correcto de reactivos y gestión de residuos.

11. LIMITACIONES DEL PROCEDIMIENTO

- El presente procedimiento se basa en la preparación de librerías para secuenciación de genoma completo utilizando una serie de reactivos comerciales concretos, no siendo posible su aplicación a ninguna otra combinación sin una validación previa.

12. BIBLIOGRAFÍA

1.- Gray Ravi RK, Walton K, Khosroheidari M. MiSeq: a next generation sequencing platform for genomic analysis. Methods Mol Biol. 2018; 1706:223-232. doi: 10.1007/978-1-4939-7471-9_12.

2.- van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias. Exp Cell Res. 2014; 322:12-20. doi: 10.1016/j.yexcr.2014.01.008.

3.- Hess JF, Kohl TA, Kotrová M, Rönsch K, Paprotka T, Mohr V, Hutzenlaub T, Brüggemann M, Zengerle R, Niemann S, Paust N. Library preparation for next generation sequencing: A review of automation strategies. Biotechnol Adv. 2020; 41:107537. doi: 10.1016/j.biotechadv.2020.107537.

4.- Sato MP, Ogura Y, Nakamura K, Nishida R, Gotoh Y, Hayashi M, et al. Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. DNA Research. 2019; 26:391-398. https://doi.org/10.1093/dnares/dsz017

Servicio / Unidad de Microbiología Hospital.....	Análisis bioinformático de secuencias de genoma completo aplicado a la Microbiología Clínica	PNT-TSM-02	
		Edición Nº 01	Página 1 de 13

PNT-TSM-02

Análisis bioinformático de secuencias de genoma completo aplicado a la Microbiología Clínica

ELABORADO		REVISADO Y APROBADO	
Nombre / Firma	Fecha	Nombre / Firma	Fecha

EDICIÓN	FECHA	ALCANCE DE LAS MODIFICACIONES
01	2021	Edición inicial

1. PROPÓSITO Y ALCANCE

En este procedimiento se describe la metodología a seguir para la obtención de la secuencia de genoma completo a partir de las lecturas crudas generadas en la plataforma de secuenciación Miseq de Illumina® (ver PNT-TSM-01 de este procedimiento).

Asimismo, se describe la metodología a seguir para, partiendo de las secuencias de genoma completo, resolver ciertas cuestiones microbiológicas clínicamente relevantes como son la detección de mecanismos de resistencia a los antibióticos, de factores de virulencia o el tipado molecular.

2. FUNDAMENTO

Durante la secuenciación masiva se generan millones de lecturas (*gigabytes*) que deben ser procesadas para obtener las secuencias de genoma completo. La enorme cantidad de información y datos que se generan suponen un cambio de paradigma al demandar de forma inexorable el uso de herramientas bioinformáticas. En los últimos años, la mayor accesibilidad a las técnicas de secuenciación masiva ha favorecido el desarrollo de numerosas aplicaciones web y/o interfaces que permiten una fácil utilización de muchas de estas herramientas sin la necesidad de requerirse grandes conocimientos en lenguajes de programación.

Para la obtención de la secuencia de genoma completo a partir de las lecturas crudas que se generan en el secuenciador, existen dos estrategias básicas según se utilice, o no, un genoma de referencia durante el proceso. Así, cuando utilicemos un genoma de referencia para obtener la secuencia de genoma completo hablaremos de alineamiento o mapeo y cuando se obtenga en ausencia de referencia hablaremos de un ensamblaje de *novo*. En la elección de una u otra estrategia influye la cuestión microbiológica a resolver, aunque cabe decir que ambas estrategias son válidas para muchos fines resultando, además, complementarias en muchas ocasiones. En este sentido, resulta imprescindible considerar la localización de la parte del genoma que se quiere estudiar (genoma *core* o genoma accesorio) ya que, si se utiliza un genoma de referencia para la construcción de la secuencia de genoma completo únicamente tendremos acceso al estudio de aquellas partes contenidas en la referencia utilizada. Por tanto, siempre que se quiera explorar la presencia de elementos genéticos extracromosómicos adquiridos se realizará un ensamblaje de *novo*.

Independientemente de la estrategia, existen diferentes herramientas y protocolos para la obtención de las secuencias de genoma completo a partir de las lecturas crudas generadas en el secuenciador, más o menos complejos, y múltiples formas de resolver una misma cuestión microbiológica. En la actualidad no existen protocolos que hayan sido validados de forma universal por la comunidad científica, por lo que los resultados derivados de la aplicación de la metodología presentada en este procedimiento deben ser cuidadosamente analizados y validados por el usuario final (1).

3. DOCUMENTOS DE CONSULTA

FastQC Documentation. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

Trimmomatic User manual.
http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf

Bowtie2 User Manual. http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

Samtools User Manual. http://www.htslib.org/doc/samtools.html

Picar-tools User Manual. http://broadinstitute.github.io/picard/command-line-overview.html

GATK tool Documentation. https://gatk.broadinstitute.org/hc/en-us/categories/360002369672

Bcftools User Guide. https://samtools.github.io/bcftools/bcftools.html

MEGA Manual. https://www.megasoftware.net/web_help_10/index.htm#t=First_Time_User.htm

SPAdes Protocols. https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpbi.102

ParSNP Quick Guide. https://harvest.readthedocs.io/en/latest/content/parsnp.html

4. MUESTRAS

Durante la secuenciación en la plataforma Miseq de Illumina® se generan millones de lecturas que se devuelven en formato fastq. En concreto, por cada muestra, el secuenciador nos devolverá 2 archivos: R1.fastq y R2.fastq y éste será el material de partida en este procedimiento.

La plataforma de secuenciación Miseq incorpora un *software* que se encarga de clasificar las lecturas y asignarlas a cada una de las muestras en base a las lecturas de los índices incorporados en el "multiplexado". Además, si durante la preparación de la hoja de trabajo para el secuenciador, seleccionamos la opción *adapter trimming* este *software* también eliminará los índices y adaptadores de los extremos 3' en las lecturas generadas (ver PNT-TSM-01).

5. RECURSOS COMPUTACIONALES

Para la ejecución de los programas y herramientas bioinformáticas utilizadas en este protocolo se precisa disponer de un ordenador con un sistema operativo basado en Unix, como por ejemplo Ubuntu https://ubuntu.com/download/desktop.

Asimismo, para llevar a cabo el análisis de secuencias de genoma completo se precisa de un ordenador potente con una memoria RAM de 16Gb (o superior), un disco duro de 512Gb SSD y 1Tb HDD (o superior) y un procesador de 8 núcleos y 16MB de memoria caché (por ejemplo, Core i9-9900K 3.6GHz).

6. PROGRAMAS Y HERRAMIENTAS BIOINFORMÁTICAS

Para la ejecución de este protocolo se requiere la instalación de los siguientes programas:

- FastQC. Descarga disponible en: https://github.com/s-andrews/FastQC
- Trimmomatic. Descarga disponible en: http://www.usadellab.org/cms/?page=trimmomatic
- Bowtie2. Descarga disponible en: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

- Samtools. Descarga disponible en: https://sourceforge.net/projects/samtools/files/samtools/
- Picard-tools. Descarga disponible en: https://github.com/broadinstitute/picard/
- GATK. Descarga disponible en: https://github.com/broadinstitute/gatk/releases
- Bcftools. Descarga disponible en: http://www.htslib.org/download/
- SPAdes. Descarga disponible en: https://github.com/ablab/spades
- MEGA. Descarga disponible en: https://www.megasoftware.net/
- ParSNP. Descarga disponible en: https://github.com/marbl/parsnp
- Figtree. Descarga disponible en: https://github.com/rambaut/figtree/releases

Además de requerirse la instalación de estos programas se necesita de un ordenador con acceso a la red de Internet.

7. PROCESAMIENTO

7.1. DE LAS LECTURAS CRUDAS A LAS LECTURAS PROCESADAS

En esencia, los archivos fastq que son una concatenación de todas las lecturas generadas en el secuenciador seguidas de la calidad de la llamada de cada una de las bases (Figura 1).

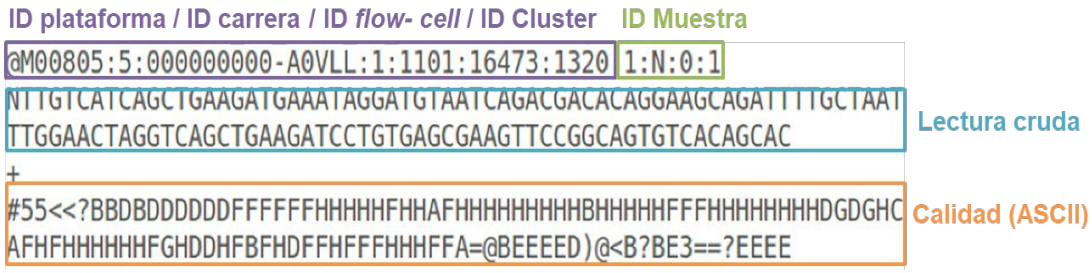


Figura 1. Estructura y formato de un archivo fastq.

Por tanto, el tamaño de estos archivos depende del número de lecturas que el secuenciador haya generado para esa muestra. En el secuenciador las muestras se introducen igualmente representadas (mezcla equimolar), siendo por tanto esperable que todos los archivos generados tengan una tamaño semejante. Estos archivos suelen ser grandes, del orden de megabytes/gigabytes, por lo que del secuenciador salen comprimidos en formato fastq.gz para ahorra espacio en el disco. Para descomprimirlos se puede escribir la siguiente línea de comando en la terminal:

gunzip ~/nombre_archivo_R1.fastq.gz
gunzip ~/nombre_archivo_R2.fastq.gz

El primer paso del análisis consiste en realizar el control de calidad de las lecturas crudas generadas. Para este fin podemos utilizar la herramienta FastQC, herramienta sencilla que puede ejecutarse tanto a través de una interfaz visual como en la terminal. En cualquiera de estas opciones, el resultado principal es un Archivo HTML que contiene diferentes resúmenes sobre la calidad general de las lecturas de secuenciación, resúmenes que nos permiten juzgar si las lecturas son de buena calidad y, por lo tanto, si se pueden utilizar para los pasos posteriores o deben descartarse.

Para los pasos siguientes del análisis de secuencias el control *Per base sequence quality* es especialmente relevante. Normalmente, un *Phred score* >30 es un indicador de buena calidad, ya que esto nos indica que con una probabilidad > 99.9% el nucleótido leído es el correcto. En la Figura 2 se muestra un ejemplo del control *Per base sequence quality*, estando representado el *Phred score* en el eje Y y el ciclo de secuenciación en el eje X.

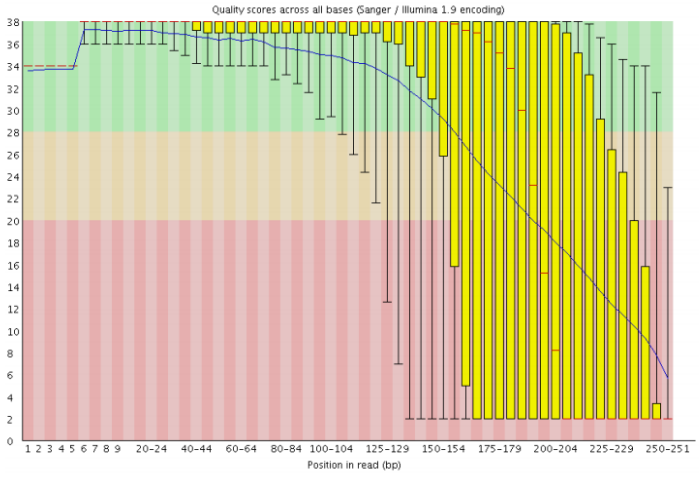


Figura 2. Ventana del control *Per base sequence quality* realizado por FastQC

En el caso particular de la figura 2, se observa un descenso acusado de la calidad a partir del ciclo de secuenciación 150-154 (*Phred score* <20) por lo que hay que recortar las lecturas en pasos posteriores para asegurar la calidad del análisis.

Idealmente todos los parámetros deberían de pasar los controles de calidad y si esto no ocurre se deben analizar las posibles causas y cómo resolverlas antes de seguir adelante con el análisis. La resolución de estos parámetros es más compleja y está fuera del alcance del presente procedimiento. Una explicación detallada de las causas que pueden hacer que estos parámetros no superen el control de calidad puede encontrase en el siguiente enlace https://www.youtube.com/watch?v=bz93ReOv87Y.

Evaluada la calidad de las lecturas crudas generadas por el secuenciador, es necesario recortarlas para eliminar los fragmentos de las mismas que presentan baja calidad. Para este fin se puede utilizar la herramienta de recorte Trimmomatic. Esta herramienta, además de recortar las lecturas en función de su calidad, permite eliminar restos de adaptadores y descartar las lecturas excesivamente cortas que pueden complicar e interferir en los pasos posteriores del análisis. Para ejecutarla, lanzaremos el siguiente comando en la terminal:

java –jar ~/trimmomatic.jar PE –phred33 ~/R1.fastq.gz ~/R2.fastq.gz ~/output.trimmed.R1Paired.fastq.gz ~/output.trimmed.R1Unpaired.fastq.gz ~/output.trimmed.R2Paired.fastq.gz ~/output.trimmed.R2Unpaired.fastq.gz LEADING:"N" TRAILING:"N" SLIDINGWINDOW:"N":30 MINLEN:"N"

sustituyendo las "N" por el número que deseemos y teniendo en cuenta que la opción LEADING corta bases del inicio de la lectura, la opción TRAILING del final de la lectura y que la opción SLIDINGWINDOW realizará un recorte cuando la calidad media de N bases consecutivas sea inferior a 30. Las lecturas procesadas pueden volver a visualizarse en FastQC para comprobar el efecto del recorte sobre la calidad de las lecturas.

7.2. ALINEAMIENTO DE LAS LECTURAS PROCESADAS

La alineación o mapeo de las lecturas procesadas consiste en determinar qué posiciones ocupan estas en el genoma mediante comparación directa con un genoma de referencia. En la actualidad existen diferentes herramientas de alineamiento que funcionan con lecturas cortas como las generadas por las plataformas de Illumina, como Bowtie2.

La primera vez que utilicemos un genoma como referencia para el mapeo necesitaremos indexarlo. Lo indexaremos utilizando las herramientas Picard-Tools, samtools y bowtie2 y para ello únicamente necesitaremos un archivo que contenga el genoma de referencia en formato fasta, archivo que podemos descargar de bases de datos como la disponible en el NCBI.

Para indexar el genoma de referencia lanzaremos los siguientes comandos en la terminal:

mkdir ~/GENOMA_REF
mv ~/GENOMA_REF.fasta ~/GENOMA_REF
java -jar ~/picard.jar NormalizeFasta I=~/GENOMA_REF/GENOMA_REF.fasta O=~/GENOMA_REF/N_ GENOMA_REF.fasta
java -jar ~/picard.jar CreateSequenceDictionary R=~/GENOMA_REF/GENOMA_REF.fasta O=~/GENOMA_REF/GENOMA_REF.dict
~/samtools faidx ~/GENOMA_REF/GENOMA_REF.fasta
~/bowtie2-build ~/GENOMA_REF/GENOMA_REF.fasta ~/GENOMA_REF/GENOMA_REF

Una vez tenemos el genoma de referencia indexado podemos alinear las lecturas procesadas con Bowtie2 ejecutando el siguiente comando:

~/bowtie2 --phred33 -x ~/GENOMA_REF/GENOMA_REF -q -1 ~/Muestra_ID_R1.fastq -2 ~/Muestra_ID_R2.fastq -X 1000 -S ~/Muestra_ID.sam

En este punto tendremos nuestras lecturas alineadas con el genoma de referencia, pero estarán desordenadas, por lo que es necesario ordenarlas:

~/samtools view -b –S ~/Muestra_ID.sam > ~/Muestra_ID.bam
java -jar ~/picard.jar SortSam INPUT=~/Muestra_ID.bam OUTPUT=~/Muestra_ID_sorted.bam SORT_ORDER=coordinate

Finalmente podemos mejorar el alineamiento obtenido. Así se pueden eliminar los duplicados y otros artefactos de la reacción de secuenciación, revisar y mejorar los alineamientos en las zonas donde hay pequeñas inserciones/deleciones (indels), etc. Para este fin se pueden utilizar algunas de las herramientas disponibles en los paquetes Picard-Tools, GATK (*Genome Analysis Tool Kit*) y samtools, ejecutando los siguientes comandos en la terminal:

java -jar ~picard.jar MarkDuplicates INPUT=~/Muestra_ID_sorted.bam OUTPUT=~/Muestra_ID_marked.bam M=~/metrics.txt

java -jar ~picard.jar AddOrReplaceReadGroups INPUT=~/Muestra_ID_marked.bam OUTPUT=~/Muestra_ID_aorrg.bam LB=Read-Group-library PL=Read-Group-platform(=Illumina) PU=Read-Group-platform-unit(=run barcode) SM=Read-Group sample name

java -jar ~/picard.jar BuildBamIndex INPUT=~/Muestra_ID_aorrg.bam

java -jar ~/GenomeAnalysisTK.jar -T RealignerTargetCreator -I ~/Muestra_ID_aorrg.bam -R ~/GENOMA_REF/GENOMA_REF.fasta -o ~/Muestra_ID_realigned_intervals

java -jar ~/GenomeAnalysisTK.jar -T IndelRealigner –I ~/Muestra_ID_aorrg.bam –R ~/GENOMA_REF/GENOMA_REF.fasta --maxConsensuses 60 --maxReadsForConsensuses 240 --maxReadsForRealignment 6000 --targetIntervals ~/Muestra_ID_realigned_intervals -o ~/Muestra_ID_realigned.bam

~/samtools sort ~/Muestra_ID_realigned.bam ~/Muestra_ID_alineamiento_final.bam

Una vez que nuestro alineamiento está mejorado, los pasos posteriores dependerán de la cuestión que queramos resolver.

7.3. ENSAMBLAJE DE *NOVO* DE LAS LECTURAS PROCESADAS

Cuando se desconoce la etiología de la bacteria secuenciada, no se dispone de un genoma de referencia adecuado o cuando se quiere explorar regiones que no están presentes en el genoma de referencia, hay que ensamblar de *novo* las lecturas procesadas.

Para este fin se han desarrollado numerosas herramientas, siendo SPAdes una de las más utilizadas. La forma más sencilla de ejecutar esta herramienta y ensamblar las lecturas procesadas es lanzar el siguiente comando en la terminal:

python ~/spades.py -o ~/Muestra_ID -1 ~/Muestra_ID_R1.fastq -2 ~/Muestra_ID_R2.fastq --careful

SPAdes generará una carpeta que contendrá la secuencia de nuestro genoma en formato fasta. Para evaluar la calidad del ensamblado obtenido nos fijaremos en el número de *contigs* que hemos obtenido y en el tamaño de los mismos; cuánto menos contigs tengamos y mayor tamaño presenten mejor será la calidad del ensamblado.

7.4. APLICACIONES EN MICROBIOLOGÍA CLÍNICA

7.4.1. Diagnóstico etiológico de las enfermedades infecciosas

La universalidad de las técnicas de secuenciación masiva permite obtener la secuencia de genoma completo de la bacteria causante de la infección sin necesidad de conocer la etiología de la misma. Por tanto, se puede realizar el ensamblaje de *novo* a partir de las lecturas procesadas (apartados 7.1 y 7.3) y utilizar el archivo fasta con fines de identificación bacteriana.

Para ello, existen diferentes herramientas bioinformáticas entre las que destaca, por su fácil utilización, KmerFinder. Esta herramienta está disponible en línea de comando, pero también está disponible como servidor web https://cge.cbs.dtu.dk/services/KmerFinder/ (Figura 3).

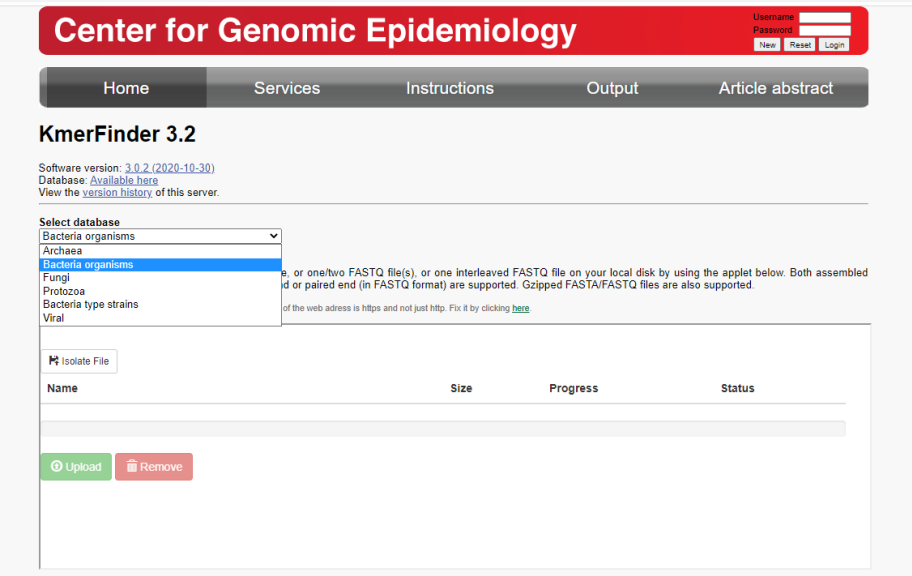


Figura 3. Aspecto de la herramienta web KmerFinder

Los pasos a seguir en el servidor son sencillos: primero se selecciona la base de datos con la que se quiere trabajar y, posteriormente, subimos el archivo fasta de la muestra en estudio. En pocos minutos, obtendremos la identificación de la especie secuenciada. Cabe decir que, además de con fines de identificación bacteriana, se puede utilizar esta herramienta para la selección del genoma de referencia en el alineamiento (apartado 7.2).

7.4.2. Detección de mecanismos de resistencia a los antibióticos

De forma natural, las especies bacterianas presentan resistencia a ciertos antibióticos; resistencia que puede verse incrementada por la adquisición horizontal de nuevos elementos de resistencia o por la adquisición y selección de mutaciones cromosómicas.

La detección de elementos de resistencia adquirida por vía horizontal resulta sencilla. Existen diferentes herramientas y bases de datos para realizar este análisis (ver la tabla 2 del documento científico de este

procedimiento). ResFinder es una de las herramientas más ampliamente utilizada para este fin y puede utilizarse tanto en línea de comando como en servidor web https://cge.cbs.dtu.dk/services/ResFinder/ (Figura 4).

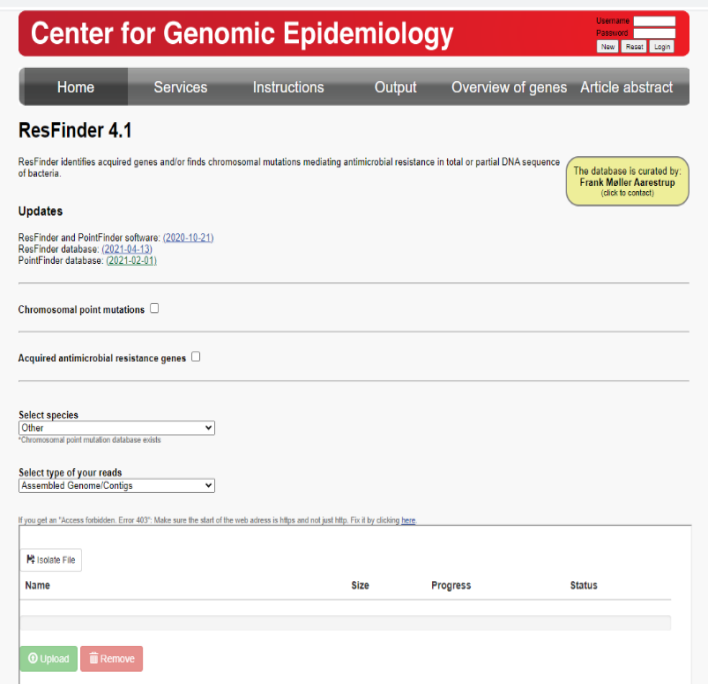


Figura 4. Aspecto de la herramienta web ResFinder

Los pasos a seguir son sencillos: se selecciona la opción *Acquired antimicrobial resistance genes* y se abrirá un desplegable en el que escogeremos los grupos de antibióticos que nos interesa estudiar y los parámetros de búsqueda. En concreto tendremos que decidir que longitud mínima cubren los elementos detectados en nuestro genoma con respecto a los contenidos en la base de datos y el grado de similitud entre ellos. Inicialmente se pueden dejar los parámetros predefinidos, y ajustarlos posteriormente en función de los resultados que se obtengan. Subiremos el ensamblaje de *novo* de nuestra secuencia y en pocos minutos el servidor nos devolverá los elementos de resistencia encontrados en la misma. Las bases de datos contienen elementos de resistencia de todas las especies bacterianas y, por tanto, pueden incluir elementos de resistencia que formen parte de la resistencia intrínseca de la bacteria en estudio. Asimismo, esta herramienta nos puede devolver más de un elemento de resistencia para una misma parte del genoma. Por todo ello, se deben analizar y validar los resultados devueltos por ResFinder ya que, de lo contrario, sobreestimaremos los elementos de resistencia adquiridos en la bacteria secuenciada.

Recientemente, en esta herramienta se ha incorporado una base de datos para la detección de mutaciones cromosómicas relacionadas con la resistencia antibiótica (PointFinder). No obstante, actualmente sólo está disponible para unas pocas especies bacterianas: *Campylobacter* spp., *E. coli*, *Salmonella* spp., *N. gonorrhoeae*, *Enterococcus* spp., *Klebsiella* spp., *S. aureus*, *H. pylori* y *M. tuberculosis.* Previsiblemente, esta lista se irá ampliando en el futuro conforme aumente el conocimiento sobre las bases de la resistencia mutacional en otros microorganismos.

No obstante, siempre que se disponga de un genoma de referencia adecuado y bien anotado se pueden estudiar los mecanismos de resistencia mutacionales mediante el denominado *variant calling* o llamada de

de variantes. Para ello, la primera tarea consistirá en investigar la literatura disponible y construir, en base al conocimiento existente, un catálogo de genes y regiones genómicas relacionados con la resistencia a los antibióticos en la especie con la que estemos trabajando.

Para realizar la llamada de variantes, se parte del archivo final ordenado obtenido en el apartado 7.2, utilizaremos bcftools y ejecutaremos el siguiente comando en la terminal:

java -jar ~picard.jar MarkDuplicates INPUT=~/Muestra_ID_sorted.bam OUTPUT=~/Muestra_ID_marked.bam M=~/metrics.txt

Finalmente, podemos filtrar este archivo vcf aplicando los parámetros de calidad deseados con el comando:

bcftools filter -e '%QUAL<20 || DP<10' ~/Muestra_ID_variantes.vcf -O z -o ~/Muestra_ID_variantes.filtradas.vcf.gz

7.4.3. Detección de genes de virulencia

Para la detección de genes de virulencia en la bacteria secuenciada se puede utilizar una aproximación similar a la utilizada para la detección de mecanismos de resistencia a los antibióticos. Se puede utilizar la herramienta VirulenceFinder en su versión servidor web (https://cge.cbs.dtu.dk/services/VirulenceFinder/), siguiendo las indicaciones y subiendo el ensamblaje de *novo*. No obstante, esta herramienta sólo está disponible para las siguientes especies: *Listeria* spp., *S. aureus*, *E. coli* y *Enterococcus* spp. Alternativamente, se puede utilizar el servidor web VFanalyzer (http://www.mgc.ac.cn/cgi-bin/VFs/v5/main.cgi?func=VFanalyzer) que ofrece la posibilidad de estudiar la virulencia en algunas especies bacterianas más.

7.4.4. Tipado molecular y caracterización de brotes

Una pregunta a la que los servicios de Microbiología Clínica deben responder con frecuencia es si existe, o no, relación entre una serie de aislamientos microbiológicos vinculados epidemiológicamente. Para ello, se han desarrollado una gran variedad de técnicas de tipificación bacteriana cuyo objetivo es comparar la composición de los ácidos nucleicos de dos o más microorganismos para así discernir si derivan de un ancestro común.

Las primeras técnicas de tipificación bacteriana se basaban en el estudio de las características fenotípicas de los microorganismos. En la actualidad, se han desarrollado diferentes herramientas y recursos bioinformáticos que nos permiten predecir el fenotipo in vitro a partir de las secuencias de genomas completo (http://www.genomicepidemiology.org/services/). Entre otros, disponemos de herramientas que nos permiten predecir el serotipo en aislados de *Salmonella* spp. (SeqSero) o *P. aeruginosa* (PAst) de una forma muy sencilla (2).

Posteriormente, con el desarrollo de la biología molecular, fueron apareciendo numerosas técnicas de tipificación bacteriana genotípicas. Durante muchos años, la electroforesis de campo pulsado (ECP) ha sido considerada como la técnica *gold-standard* para el estudio y caracterización de brotes al ofrecer un elevado poder de resolución. No obstante, los resultados de la ECP son difícilmente comparables entre laboratorios, por lo que, para fines de vigilancia epidemiológica a nivel global, el MLST se ha postulado como *gold-standard*. A falta de protocolos universales de trabajo, la secuenciación de genoma completo se

postula como la herramienta de tipificación bacteriana del futuro, al ofrecer el máximo poder de resolución y la posibilidad de generar resultados fácilmente comparables.

En primer lugar, para estudiar si dos o más aislados de una misma especie vinculados epidemiológicamente están, o no, genómicamente relacionados se puede predecir su secuenciotipo (MLST) a partir del ensamblaje de *novo* obtenido en el apartado 7.3 con herramientas bioinformáticas como la disponible en https://cge.cbs.dtu.dk/services/MLST/.

El MLST no ofrece un alto poder resolutivo por lo que para discernir si dos aislados pertenecientes a un mismo secuenciotipo (o complejo clonal) están o no relacionados deberemos ampliar el análisis. Una aproximación sencilla para resolver este problema consiste en inferir las diferencias entre aislados mediante comparación y conteo de las variaciones de nucleótidos encontradas con respecto al genoma de referencia y construir así una matriz de SNPs. Podemos extraer la secuencia de genoma completo del archivo final (Muestra_ID_alineamiento_final.bam) obtenido en el apartado 7.2 con bcftools y lanzando el siguiente comando en la terminal:

bcftools index ~/Muestra_ID_variantes.filtradas.vcf.gz

cat ~/GENOMA_REF.fasta | bcftools consensus ~/Muestra_ID_variantes.filtradas.vcf.gz > ~/Muestra_ID.cns.fa

Este comando nos devolverá nuestras secuencias en un archivo con formato fasta. A partir de aquí, se puede generar un archivo multifasta con la secuencia de referencia y todas las secuencias en estudio y calcular las diferencias de SNPs entre aislados utilizando el programa MEGA. Para ello cargaremos el archivo multifasta y seleccionaremos la opción *Analyze*, nos aparecerá una nueva ventana en la que se puede visualizar la referencia y nuestras secuencias alineadas. A partir de aquí, podremos calcular la matriz de SNPs (*Distance* → *Compare Pairwise Distance* → *Model method* = *No. of differences*) y exportar a formato Excel las regiones variables (*Highlight* → *Variable sites* + *Export file*) para visualizar y validar las diferencias de SNPs encontradas.

Es importante destacar que en la actualidad no existe un valor que sea universalmente válido para ninguna especie que permita afirmar si los aislados estudiados pertenecen, o no, a un mismo brote (ver la tabla 4 del documento científico de este procedimiento) por lo que es recomendable acompañar la matriz de distancias de un análisis filogenómico. MEGA dispone de herramientas para estudiar la relación filogenómica entre los aislados, contenido en el multifasta que le hemos cargado.

En el caso de que no dispongamos de una referencia adecuada, podemos inferir la relación entre aislados construyendo un árbol del genoma *core* a partir de los ensamblajes de *novo* con la herramienta Parsnp. Para ello crearemos un directorio donde alojaremos todos los ensamblajes en estudio y lanzaremos el siguiente comando en la terminal:

./parsnp -c -r! -d ~/Carpeta_secuencias_en_estudio

Para visualizar los resultados y analizar la relación entre aislados podemos utilizar programas como FigTree.

8. OBTENCIÓN Y EXPRESIÓN DE LOS RESULTADOS

Los resultados obtenidos mediante la aplicación de este protocolo deben ser validados por el facultativo especialista responsable.

Al no existir protocolos que hayan sido validados universalmente se recomienda acompañar los resultados de una explicación breve de la metodología empleada para su obtención. Siempre que se trabaje con bases de datos se debe indicar la referencia y/o contenido de las mismas para facilitar una correcta interpretación de los resultados.

9. RESPONSABILIDADES

Deben estar descritas en el manual general de organización del laboratorio de Microbiología y en las fichas de descripción de los puestos de trabajo.

El procedimiento deberá llevarse a cabo por personal técnico cualificado con un entrenamiento y formación específica.

La elección de la aproximación a utilizar, la supervisión de la técnica y evaluación última de resultados deberá realizarla el facultativo especialista responsable.

10. ANOTACIONES AL PROCEDIMIENTO

Para la obtención de las secuencias de genoma completo existen dos aproximaciones, siendo necesario considerar las ventajas y desventajas que cada una de ellas nos ofrecen.

La calidad de la carrera de secuenciación influye notablemente en la calidad de las secuencias de genoma completo obtenidas, así como en la calidad y fiabilidad de los resultados derivados de su análisis. Por tanto, se debe realizar una cuidadosa evaluación de la calidad de la carrera y en consecuencia tomar las medidas que se consideren oportunas con el objeto de asegurar la calidad y validez de los resultados.

Actualmente, existen protocolos bioinformáticos validados y disponibles en plataformas de acceso libre para el análisis de secuencias de genoma completo de especies como *N. meningitidis*, *E. coli* o *M. tuberculosis* (3-5). Asimismo, existen servidores como Galaxy en los que pueden realizarse ciertos análisis de secuencias de genoma completo y que se acompañan de tutoriales bien detallados (https://galaxy-au-training.github.io/tutorials/).

11. LIMITACIONES DEL PROCEDIMIENTO

La aplicación del presente procedimiento únicamente es aplicable a lecturas crudas generadas por plataformas con tecnología Illumina®. Asimismo, en caso de que se hayan utilizado para la preparación de librerías sistemas ofertados por otra casa comercial se deben incorporar pasos adicionales en el procesado inicial.

12. BIBLIOGRAFÍA

1. Carriço JA, Rossi M, Moran-Gilad J, Van Domselaar G, Ramirez M. A primer on mi1. Carriço JA, Rossi M, Moran-Gilad J, Van Domselaar G, Ramirez M. A primer on microbial bioinformatics for nonbioinformaticians. Clin Microbiol Infect. 2018; 24:342-349. doi: 10.1016/j.cmi.2017.12.015.

2. Blake KS, Choi J, Dantas G. Approaches for characterizing and tracking hospital-associated multidrug-resistant bacteria. Cell Mol Life Sci. 2021; 78:2585-2606. doi: 10.1007/s00018-020-03717-2.

3. Bogaerts B, Winand R, Fu Q, Van Braekel J, Ceyssens PJ, Mattheus W, Bertrand S, De Keersmaecker SCJ, Roosens NHC, Vanneste K. Validation of a bioinformatics workflow for routine analysis of whole-genome sequencing data and related challenges for pathogen typing in a European National Reference Center: *Neisseria meningitidis* as a proof-of-concept.
Front Microbiol. 2019; 10:362. doi: 10.3389/fmicb.2019.00362.

4. Bogaerts B, Nouws S, Verhaegen B, Denayer S, Van Braekel J, Winand R, Fu Q, Crombé F, Piérard D, Marchal K, Roosens NHC, De Keersmaecker SCJ, Vanneste K. Validation strategy of a bioinformatics whole genome sequencing workflow for Shiga toxin-producing Escherichia coli using a reference collection extensively characterized with conventional methods.
Microb Genom. 2021; 7(3). doi: 10.1099/mgen.0.000531.

5. Bogaerts B, Delcourt T, Soetaert K, Boarbi S, Ceyssens PJ, Winand R, Van Braekel J, De Keersmaecker SCJ, Roosens NHC, Marchal K, Mathys V, Vanneste K.A bioinformatics WGS workflow for clinical *Mycobacterium tuberculosis* complex isolate analysis, validated using a reference collection extensively characterized with conventional methods and *in silico approaches*. J Clin Microbiol. 2021; JCM.00202-21. doi: 10.1128/JCM.00202-21.