



SECUENCIACIÓN GENÓMICA Y ANÁLISIS DE VARIANTES

Dra. María de Toro



viu

Universidad
Internacional
de Valencia

Secuenciación genómica y análisis de variantes

Dra. María de Toro



Universidad
Internacional
de Valencia

Este material es de uso exclusivo para los alumnos de la Universidad Internacional de Valencia. No está permitida la reproducción total o parcial de su contenido ni su tratamiento por cualquier método por aquellas personas que no acrediten su relación con la Universidad Internacional de Valencia, sin autorización expresa de la misma.

Edita
Universidad Internacional de Valencia

ISBN: 978-84-19314-00-0

Leyendas



Enlace de interés



Ejemplo



Importante



Descarga de archivo



abc Los términos resaltados a lo largo del contenido en color **naranja** se recogen en el apartado **GLOSARIO**.

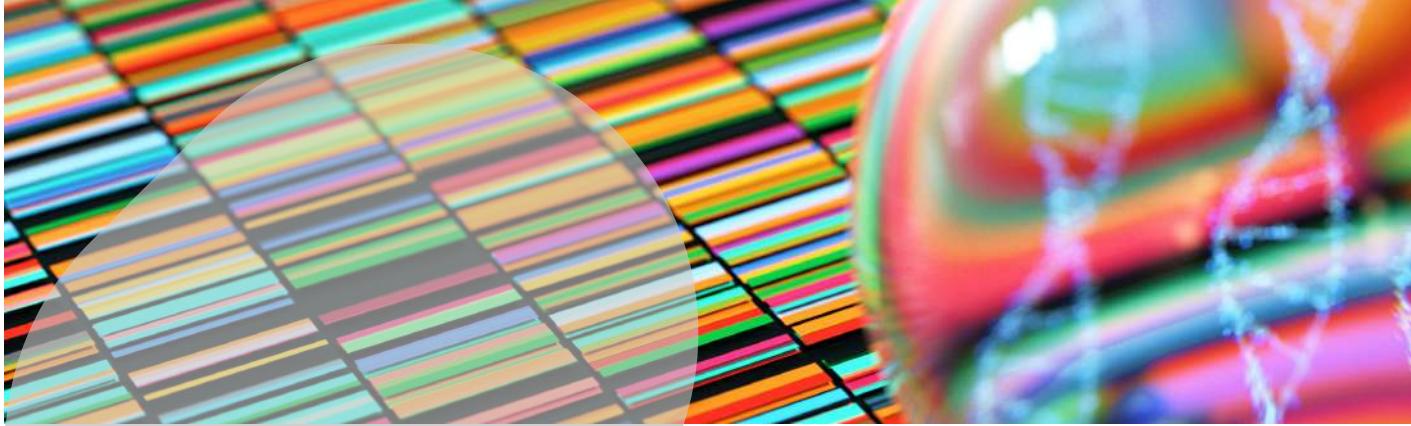
Índice

CAPÍTULO 1. LA ESTRUCTURA DEL GENOMA HUMANO Y PATRONES DE TRANSMISIÓN DE ENFERMEDADES GENÉTICAS.....	8
1.1. Estructura del genoma humano	8
1.1.1. Genoma mitocondrial	8
1.1.2. Genoma nuclear	11
1.1.3. ¿Qué tipos de secuencias encontramos en el genoma humano?	12
1.1.4. Secuencias repetidas y secuenciación masiva	19
1.2. Patrones de transmisión de las enfermedades genéticas	21
1.2.1. Variabilidad del genoma	21
1.2.2. Tipos de enfermedades genéticas	22
1.2.3. Secuencias repetidas y secuenciación masiva	19
1.2. Patrones de transmisión de las enfermedades genéticas	21
1.2.1. Variabilidad del genoma	21
1.2.2. Tipos de enfermedades genéticas	22
CAPÍTULO 2. INTRODUCCIÓN A LA MEDICINA PREVENTIVA PERSONALIZADA, EXPOSOMA Y TIPOS DE DATOS	25
2.1. Medicina preventiva personalizada.....	25
2.1.1. Definición y objetivos.....	25
2.1.2. ¿Qué aporta la genómica en esta área?	26
2.1.3. Aplicaciones de la medicina preventiva personalizada	27
2.1.4. Retos de la medicina preventiva personalizada	28
2.2. Exposoma	29
2.2.1. Definición.....	29
2.2.2. Factores no genéticos	30
2.2.3. ¿Cómo se estudia el exposoma?.....	30
2.2.4. Patologías más comunes y su relación con exposoma	32
2.2.5. Retos asociados	33
2.3. Datos de medicina personalizada/preventiva y de precisión	34
2.3.1. Datos integrados.....	34
2.3.2. Tipos, fuentes y características de los datos	34
2.3.3. Retos asociados	36

CAPÍTULO 3. ¿CÓMO ANALIZAMOS UN GENOMA EUCARIOTA? PANELES DE CAPTURA DE GENES VS. GENOMA COMPLETO	37
3.1. Introducción al análisis de genomas eucariotas	37
3.2. Estrategias generales de análisis	38
3.2.1. Paneles de genes o regiones de interés	38
3.2.2. Secuenciación de exoma (WES).....	39
3.2.3. Secuenciación de genoma completo (WGS)	39
3.3. Protocolo de análisis.....	40
3.3.1. Extracción del ADN	40
3.3.2. Preparación de la genoteca/librería. Genotecas con captura o amplificación de regiones	40
3.3.3. Secuenciación de la genoteca/librería	43
CAPÍTULO 4. ¿CÓMO ANALIZAMOS UN GENOMA EUCARIOTA? ANÁLISIS BIOINFORMÁTICO DE PANELES DE CAPTURA DE GENOMA HUMANO	45
4.1. Análisis primario. Calidad y filtrado de secuencias	46
4.1.1. ¿Por qué es importante evaluar la calidad de las lecturas y filtrarlas?.....	46
4.1.2. Herramientas de control de calidad	47
4.1.3. Análisis de calidad con FastQC	47
4.1.4. Herramientas de filtrado por calidad de las secuencias y eliminación de adaptadores	55
4.2. Mapeo de secuencias. Herramientas de mapeo, visualización y análisis de calidad	57
4.2.1. El proceso de mapeo	57
4.2.2. Herramientas para el mapeo.....	59
4.2.3. Archivos de mapeo: el formato SAM	62
4.2.4. Herramientas para el manejo de los archivos SAM. Generación de archivos BAM.....	66
4.2.5. Visualización del mapeo.....	66
4.2.6. Análisis de calidad del mapeo	68
4.3. Identificación de variantes	70
4.3.1. Preprocesamiento: identificación de secuencias duplicadas.....	72
4.3.2. Identificación de variantes	73
4.3.3. Posprocesado: filtrado de variantes	76
4.4. Anotación de variantes	76
4.4.1. Variant effect predictor (VEP)	78
4.4.2. Annotar	79
CAPÍTULO 5. ANÁLISIS DE GENOMAS BACTERIANOS.....	81
5.1. Genoma procariota	81
5.1.1. Características generales de los genomas procariotas	81

5.1.2. Genomas multipartitos.....	82
5.1.3. Estructura detallada del genoma procariota.....	83
5.1.4. Número de genes y función general.....	84
5.2. Epidemiología genómica. Retos y oportunidades en el análisis de genomas bacterianos	85
5.3. Tipificación de genomas bacterianos	86
5.3.1. Análisis basado en polimorfismos de un único nucleótido (SNP)	87
5.3.2. Análisis gen por gen.....	87
5.3.3. Identidad nucleotídica promedio (<i>average nucleotide identity, ANI</i>)	88
5.4. Reconstrucción del genoma: ensamblaje <i>de novo</i>	88
5.4.1. Genotecas para el ensamblado <i>de novo</i>	89
5.4.2. Conceptos generales del ensamblaje	90
5.4.3. Tipos de algoritmos de ensamblaje	90
5.4.4. Evaluación previa al ensamblaje. Análisis de <i>k</i> -meros	95
5.4.5. Ensamblaje de secuencias cortas con SPAdes.....	97
5.4.6. Evaluación de la calidad del ensamblaje: continuidad y contenido.....	100
5.5. Reconstrucción del genoma: anotación del genoma ensamblado	102
5.5.1. Anotación utilizando Prokka	103
5.5.2. Visualización de las anotaciones.....	105
5.6. Reconstrucción del genoma: elementos genéticos móviles (EGM). Plásmidos	107
5.6.1. Herramientas de ensamblaje y extracción de plásmidos	108
5.6.2. Herramientas de identificación mediante marcadores genéticos.....	108
5.6.3. Herramientas basadas en los grafos de ensamblaje <i>de novo</i>	109
5.7. Detección y anotación de regiones de interés: genes de resistencia a antibióticos y virulencia	112
 CAPÍTULO 6. ANÁLISIS DE GENOMA DE VIRUS. EL CASO DE SARS-COV-2	 116
6.1. Estructura de los virus. Aspectos generales	116
6.1.1. Bacteriófagos	116
6.1.2. Virus que infectan células eucariotas.....	118
6.2. El genoma de SARS-CoV-2	119
6.3. Clasificación de SARS-CoV-2. Variantes de interés y linajes. Bases de datos	124
6.3.1. Clasificación de la OMS: variantes de preocupación, variantes de interés y variantes bajo vigilancia	124
6.3.2. Monitorización de la evolución de SARS-CoV-2 a tiempo real. Nextstrain.....	127
6.3.3. GISAID. El repositorio de todos los genomas SARS-CoV-2 secuenciados	128
6.3.4. Clasificación de linajes y constelaciones de mutaciones: PANGO.....	129
6.3.5. El análisis de mutaciones de interés. Bases de datos de monitorización.....	133
6.4. Análisis bioinformático de linajes de SARS-CoV-2	137

6.4.1. Limpieza de las lecturas crudas. Eliminación del genoma del hospedador	138
6.4.2. Mapeo al genoma de referencia y determinación del linaje.....	139
6.4.3. Ensamblaje <i>de novo</i> y anotación.....	141
6.3.2. Monitorización de la evolución de SARS-CoV-2 a tiempo real. Nextstrain.....	127
6.3.3. GISAID. El repositorio de todos los genomas SARS-CoV-2 secuenciados	128
6.3.4. Clasificación de linajes y constelaciones de mutaciones: PANGO	129
6.3.5. El análisis de mutaciones de interés. Bases de datos de monitorización.....	133
6.4. Análisis bioinformático de linajes de SARS-CoV-2	137
6.4.1. Limpieza de las lecturas crudas. Eliminación del genoma del hospedador	138
6.4.2. Mapeo al genoma de referencia y determinación del linaje.....	139
6.4.3. Ensamblaje <i>de novo</i> y anotación.....	141
GLOSARIO	143
ENLACES DE INTERÉS	146
BIBLIOGRAFÍA	147



Capítulo 1

La estructura del genoma humano y patrones de transmisión de enfermedades genéticas

1.1. Estructura del genoma humano

Llamamos genoma al contenido total de ácido desoxirribonucleico (**ADN**) presente en una célula. Podríamos decir que es el manual de operaciones que contiene todas las instrucciones que ayudan al desarrollo y funcionamiento de un organismo vivo.

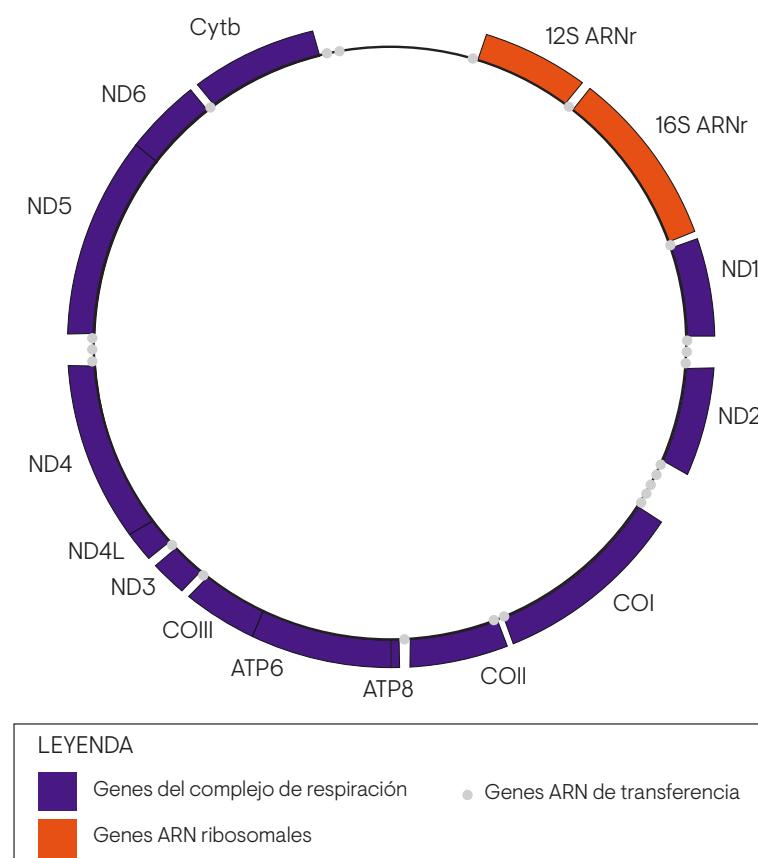
El genoma de los organismos eucariotas consta de dos tipos: genoma nuclear y genoma mitocondrial. En general, para los estudios de genética básicos, nos centramos en el genoma nuclear.

1.1.1. Genoma mitocondrial

El **genoma mitocondrial** fue la primera parte del genoma de la que se conoció la secuencia, en 1981 (Anderson *et al.*, 1981). El **ADN mitocondrial** (ADNmt) codifica para la mayor parte de los componentes que son esenciales en la cadena de fosforilación oxidativa, crucial para la generación de adenosín trifosfato (ATP).

Por lo tanto, salvo en el caso de algunas levaduras que son capaces de sobrevivir en condiciones anaerobias sin la presencia de este ADN, el ADNmt es esencial para la vida (Sharma & Sampath, 2019).

El ADNmt está constituido por una molécula de ADN circular de cadena doble y de 16.6 kb de longitud. Este es un tamaño pequeño, en comparación con el genoma nuclear. Contiene 37 genes, de los cuales dos codifican para ARN ribosómico (ARNr), 22 para ARN de transferencia (ARNr) y 13 para proteínas necesarias para el funcionamiento de la mitocondria. El resto de las proteínas son codificadas por el genoma nuclear. El 97 % del genoma mitocondrial es codificante y tiene una codificación policistrónica, es decir, se transcribe en forma de transcripto continuo (Anderson *et al.*, 1981).

Figura 1*Estructura del genoma mitocondrial humano (ADNmt)*

Nota. Adaptado de “The Human Genome”, por T. A. Brown, 2002, *Genomes* (2.^a ed.).
<https://www.ncbi.nlm.nih.gov/books/NBK21134/figure/A5305/?report=objectonly>



Enlaces de interés

En el siguiente enlace puedes encontrar la base de datos MITOMAP, específica del ADNmt, donde se detalla la estructura del genoma mitocondrial en humanos, así como el compendio de sus polimorfismos y mutaciones.

<https://mitomap.org/MITOMAP>

A continuación, puedes ver un vídeo sobre la estructura del genoma mitocondrial.

https://www.youtube.com/watch?v=_rbcdx5NahQ

Las mutaciones que ocurren en el genoma mitocondrial se descubrieron en 1988, como causantes de enfermedades humanas hereditarias. Existen varias enfermedades mitocondriales hereditarias.

Algunos ejemplos se encuentran en la Tabla 1. La primera de las enfermedades asociadas al ADNmt fue la neuropatía óptica hereditaria de Leber (LHON). Esta es una enfermedad neurodegenerativa que afecta al nervio óptico y provoca, sobre todo en jóvenes, la pérdida brusca de la visión. En este caso, se debe a mutaciones en los genes *MT-ND1*, *MT-ND4* y *MT-ND6*, del complejo I de la cadena respiratoria mitocondrial. Debemos recordar que la transmisión de este tipo de enfermedades es vía materna, ya que el ADNmt se transmite exclusivamente vía materna (Nunnari & Suomalainen, 2012; Sharma & Sampath, 2019). Muchas de estas enfermedades son consideradas “enfermedades raras”, por afectar a una pequeña proporción de la población. Las mutaciones que ocurren en el ADNmt producen daño oxidativo y se han asociado a varios tipos de cáncer, enfermedades neurodegenerativas, como Alzheimer o Parkinson, y enfermedades metabólicas (Nunnari & Suomalainen, 2012; Sharma & Sampath, 2019).

Tabla 1
Enfermedades mitocondriales y su localización genómica

Enfermedad [código OMIM]	Manifestación clínica	Localización genómica	Prevalencia y edad de debut
Neuropatía óptica hereditaria de Leber (LHON) [308905, 535000]	Insuficiencia visual subaguda bilateral e indolora; pérdida súbita	Mutaciones en los genes <i>MT-ND1</i> , <i>MT-ND4</i> o <i>MT-ND6</i>	1-9/100 000 Adolescencia o edad adulta temprana
Epilepsia mioclónica con fibras rojas rasgadas (MERRF) [545000]	Epilepsia mioclónica, ataxia, debilidad muscular y demencia	Mutaciones en <i>MT-LK</i> (codifica para tRNALys, con la mutación A8344G)	Desconocido Infancia o edad adulta
Síndrome de Pearson [557000]	Anemia sideroblástica y disfunción exocrina del páncreas	Deleción en el ADNmt (4977 pb)	< 1/1 000 000 Infancia, neonatal
Síndrome de Kearns-Sayre (KSS) [530000]	Oftalmoplejia externa progresiva (OEP), retinosis pigmentaria, pérdida auditiva, ataxia cerebelosa, bloqueo cardíaco	Deleción en el ADNmt (4977 pb)	1-9/100 000 < 20 años
Encefalopatía mitocondrial, acidosis láctica y episodios similares al ictus (MELAS) [540000]	Encefalomiopatía, acidosis láctica y episodios similares a un accidente cerebrovascular; en algunas ocasiones endocrinopatías, enfermedad cardíaca, diabetes, pérdida auditiva y manifestaciones neurológicas y psiquiátricas	Mutaciones puntuales en el gen codificante de tRNALeu (A3243G)	1-9/1 000 000 Infancia, adolescencia

Nota. Adaptado de “Mitochondrial DNA Integrity: Role in Health and Disease”, por P. Sharma y H. Sampath, 2019, *Cells*, 8(2), p. 100; y
Portal sobre enfermedades raras y medicamentos huérfanos, por Orphanet, 27 de diciembre de 2021, recuperado de
<https://www.orpha.net/consor4.01/www/cgi-bin/?lNg=ES>.

1.1.2. Genoma nuclear

El genoma nuclear, tal y como su nombre indica, es el ADN que se encuentra dentro del núcleo celular. El primer borrador de su secuencia y estructura se obtuvo en el año 2001, tras más de doce años de intenso trabajo en el Proyecto Genoma Humano. Desde ese momento, esta información se ha ido actualizando de manera constante.



Enlace de interés

En esta página web encontrarás toda la información actualizada sobre este proyecto, así como la última versión de la secuencia del genoma humano. En ella podemos buscar información sobre todos los cromosomas humanos, genes de interés, transcritos, exones/intronas, dominios estructurales, etc. Asimismo, la base de datos ENSEMBL contiene otros organismos de interés, algunos de los cuales son utilizados como modelos animales y pueden ser utilizados para estudios de genómica comparativa de genes ortólogos.

http://www.ensembl.org/Homo_sapiens/Info/Index

El genoma nuclear está dividido en un conjunto de moléculas de ADN lineares, cada una de ellas contenida en un cromosoma. No existen excepciones a este patrón en los genomas eucariotas, todos los genomas eucariotas conocidos tienen al menos dos cromosomas y la molécula de ADN es siempre linear. La única variación a este respecto está en el número de cromosomas. Por ejemplo, una levadura tiene 16 cromosomas, mientras que el genoma humano está compuesto por 23 pares de cromosomas.

Por otra parte, este número tampoco está relacionado con el tamaño del genoma. Algunas salamandras tienen un genoma 30 veces superior al genoma humano, pero dividido en la mitad de los cromosomas. Aunque estas comparaciones son muy curiosas, realmente no nos dicen nada sobre los genomas, sino que simplemente nos dan idea de la falta de uniformidad de los eventos evolutivos que han moldeado la arquitectura genómica en los distintos organismos eucariotas.

Tabla 2

Tamaño de distintos genomas eucariotas

Especie	Tamaño genoma (Mb)	Número de cromosomas	Número de genes
Hongos			
- <i>Saccharomyces cerevisiae</i>	12.1	16	6463
- <i>Aspergillus nidulans</i>	25.4	8*	9586
Protozoos			
- <i>Tetrahymena pyriformis</i>	190	**	**
Invertebrados			
- <i>Caenorhabditis elegans</i>	97	6	46 925
- <i>Drosophila melanogaster</i>	180	7	17 864
- <i>Bombyx mori</i> (gusano de seda)	490	28	17 072
- <i>Strongylocentrotus purpuratus</i> (erizo de mar morado)	845	Un	33 520
- <i>Locusta migratoria</i> (langosta migratoria)	5000	**	**

>>>

>>>

Especie	Tamaño genoma (Mb)	Número de cromosomas	Número de genes
Vertebrados			
- <i>Takifugu rubripes</i> (pez puffer/torafugu)	400	22	**
- <i>Homo sapiens</i>	3200	23	61 662
- <i>Mus musculus</i>	3300	20	50 561
Plantas			
- <i>Arabidopsis thaliana</i>	125	5	38 311
- <i>Oryza sativa</i> (arroz)	466	12	35 223
- <i>Zea mays</i> (maíz)	2500	10	49 897
- <i>Pisum sativum</i> (guisante)	4800	7	<i>Un</i>
- <i>Triticum aestivum</i> (trigo)	16 000	21	<i>Un</i>
- <i>Fritillaria assyriaca</i>	120 000	**	**

Nota. Información adaptada de *Genomes 4*, por T. A. Brown, 2017, y *Home - Genome - NCBI*, 2021, recuperado de <https://www.ncbi.nlm.nih.gov/genome/>.

* Aún no hay ningún genoma depositado en <https://www.ncbi.nlm.nih.gov/genome/> que se encuentre cerrado completamente.

** No está contenido ningún genoma ni completo ni parcial en <https://www.ncbi.nlm.nih.gov/genome/>.

Un: no se encuentra ensamblado su genoma.



1.1.3. ¿Qué tipos de secuencias encontramos en el genoma humano?

En el caso del genoma humano ya hemos dicho que la secuencia de ADN está contenida en 23 cromosomas en el núcleo de cada célula humana. De los 23 pares, 22 son cromosomas autosómicos y dos son los determinantes del sexo (dos cromosomas X en mujeres y un cromosoma X y un Y en hombres).

El genoma humano disponible actualmente en la base de datos Ensembl corresponde a la versión GRCh38.p13, cuya última actualización se realizó en marzo de 2021 (Howe et al., 2021). Este genoma haploide (una sola representación por cada par) tiene un tamaño de 3 096 649 726 pb y consta de las características mostradas en la Tabla 3.

Tabla 3

Características del genoma humano recogidas en la base de datos Ensembl

Genoma: GRCh38.p13		
Pares de bases (pb)		3 096 649 726 (3.1 Gb)
Ensamblaje primario	Genes codificantes	20 442
	Genes no codificantes	23 982
	- Cortos	4865
	- Largos	16 896
	- Miscelánea	2221
	Pseudogenes	15 228
	Transcritos	237 081
Ensamblaje alternativo	Genes codificantes	3053
	Genes no codificantes	1555
	- Cortos	297
	- Largos	1071
	- Miscelánea	187
	Pseudogenes	1799
	Transcritos	21 638

Nota. Información recopilada de la base de datos Ensembl (Howe et al., 2021).

A lo largo del genoma nuclear de cualquier organismo eucariota la densidad no es homogénea, lo que hace que no encontremos patrones “típicos” de cada genoma. Aún con esta dificultad, es obvio que entender estos patrones de distribución de genes nos hace comprender la evolución de cada organismo a través de la historia de estos genomas. Veamos en detalle un fragmento de genoma humano, comparándolo con regiones similares de otros genomas de otros organismos.

**Ejemplo**

El fragmento que vamos a revisar (Figura 2) es parte del cromosoma 1, de 200 kb de longitud y va desde la posición 55 000 000 hasta la posición 55 200 000 (genoma de referencia Hg38). Este segmento contiene los siguientes elementos:

- El final del gen **BSND**, que comienza en una posición anterior a la representada (posición 54 998 944). Este gen codifica para una proteína del canal iónico de cloruro, una proteína transmembrana que forma un poro a través del cual varios iones, incluyendo el ion cloro, pueden entrar o salir de la célula.
- El gen **PCSK9**, que codifica para la proteína convertasa subtilisina/kexina de tipo 9, una proteína producida en hígado, intestino o riñón, que participa en la degradación de las lipoproteínas de baja densidad, por lo que desempeña su función en el metabolismo del colesterol.

>>>

>>>

- El inicio del gen ***USP24***, codificante de una ubiquitina peptidasa 24, una proteasa que elimina la ubiquitina de las proteínas que han sido modificadas en el proceso de ubiquitinación. La ubiquitina es una pequeña proteína reguladora cuya adición o eliminación de una proteína controla la localización celular y eventual degradación de esta. Este gen termina en la posición 55 215 366, así que vemos la mayor parte de él contenido en la Figura 2.

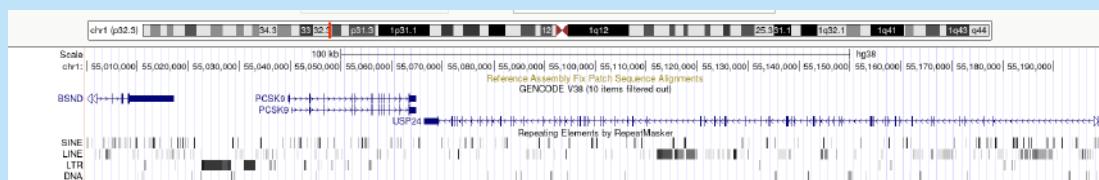
En esta región debemos ver que cada uno de los genes son discontinuos: hay tres intrones para ***BSND***, once en ***PCSK9*** y 73 en ***USP24***. Además, observamos un gran número de secuencias repetidas, que son recurrentes en varios sitios del genoma.

En esta imagen se observan cuatro grandes tipos: SINE (*short interspersed nuclear elements*), LINE (*long interspersed nuclear elements*), LTR (*long terminal repeat elements*) y transposones de ADN. En este mismo fragmento se observan múltiples copias de cada tipo, tanto en regiones intergénicas como en intrones de los genes codificantes de proteínas.

Es notable que, de estos 200 kb de segmento del genoma humano, una muy pequeña proporción de espacio corresponde a la zona codificante de los genes. En este ejemplo, si sumamos las zonas exónicas (que codifican para información biológica) es de 10 644 pb, equivalente al 5.33 % del segmento total de 200 kb. De hecho, podemos considerar que este segmento es bastante rico en genes.

Figura 2

Análisis de un segmento del genoma humano



Nota. Análisis de un segmento del genoma humano que comprende la región 55 000 000 a 55 200 000. Los genes se muestran en azul, con los exones representados como cajas azules y los intrones como líneas azules. En la parte inferior se representan los elementos de repetición más comunes (LINE, SINE, LTR y transposones ADN). Tomada de UCSC Genome Browser, Hg38. Recuperado de <https://genome.ucsc.edu/>

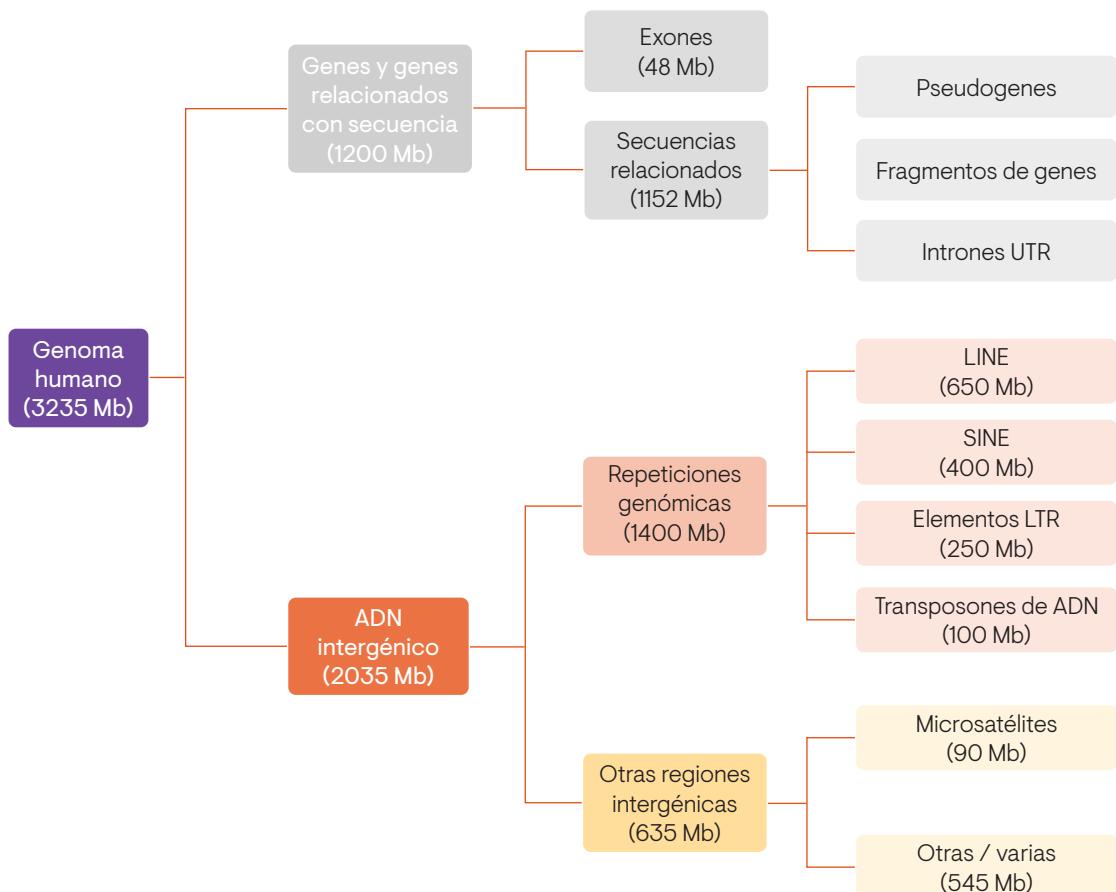
Actualmente conocemos que el genoma humano presenta una densidad de genes muy inferior a lo que se pensaba, solo un 1.5 % de su longitud está compuesto por exones que codifican proteínas. Hay un 70 % compuesto por ADN extragénico y un 30 % por secuencias relacionadas con genes. Del total de ADN extragénico, aproximadamente el 70 % corresponde a repeticiones dispersas, con lo que aproximadamente el 50 % del genoma humano corresponde a secuencias repetidas.

Por otra parte, del ADN relacionado con genes, el 95 % corresponde a ADN no codificante (pseudogenes, fragmentos de genes, intrones o secuencias no traducidas (UTR por las siglas del inglés *untranslated region*).

La Figura 3 muestra la composición del genoma humano, donde los genes y las secuencias relacionadas con genes corresponden a 1200 Mb del total del genoma (aprox. 3235 Mb).

Figura 3

Composición del genoma humano



Nota. Adaptado de *Genomes 4*, por T. A. Brown, 2017.

Si retomamos otros organismos eucariotas, tal y como vimos en la Tabla 2, podemos ver que los vertebrados y las plantas poseen tamaños de genoma superiores que, en general, hongos, protozoos y algunos invertebrados. Esto parece tener sentido, ya que esperaríamos que la complejidad de un organismo esté relacionada con el número de genes en su genoma, así que los eucariotas superiores necesitarían, hipotéticamente, genomas mayores para acomodar los genes extra. Sin embargo, esta correlación no es así.

El genoma humano, de aproximadamente 3235 Mb contiene más de 20 000 genes codificantes. Sin embargo, si analizamos el genoma del hongo *Saccharomyces cerevisiae*, este es 12.2 Mb (0.004 veces el tamaño del genoma humano), por lo tanto, le corresponderían unos 80 genes. Esto, obviamente no es así, ya que el genoma de *S. cerevisiae* contiene entre 6400 y 6700 genes. Este hecho ha sido conocido como la **paradoja del valor C**. Este valor C es la cantidad de ADN por genoma haploide y está claro que cuando se evalúa este valor para las distintas especies, no existe relación entre el contenido de ADN (longitud del genoma) y la complejidad del organismo, entendido como contenido de genes codificados.

Por tanto, tal y como hemos visto, el genoma de *S. cerevisiae* es más compacto que el genoma humano. Esto se debe, en parte, a que en este hongo los genes son más cortos y, por tanto, existen menos intrones por gen que en el caso de los genes humanos; y además, existe menos espacio intergénico.



Existen técnicas, tanto de experimentación como bioinformáticas para identificar y analizar funcionalmente los genes y obtener la anotación de un genoma completo. Sin embargo, obtener esta anotación completa es muy complicado, o incluso imposible, en organismos eucariotas. Esto significa que no conocemos de manera precisa cuántos genes se encuentran en un genoma, y no podemos dar una descripción de todas las proteínas que son codificadas. Aún quedan muchas regiones secuenciadas sin encajar en el genoma y todavía se realizan extrapolaciones en funciones de ciertos genes.

A grandes rasgos, las secuencias contenidas en el genoma humano pueden clasificarse en los siguientes grupos:

1. **Genes que codifican proteínas, intrones y genes que contienen información sobre ARN no-codificante.** Supone menos del 50 % del contenido total genético.

Todos conocemos que el gen es la unidad básica de herencia que porta la información genética necesaria para sintetizar una proteína (en el caso de genes codificantes), o de un ARN no codificante (genes de ARN). Los genes están formados por una secuencia promotora que regula su expresión y una secuencia que se transcribe. Esta secuencia que se transcribe a su vez se compone por secuencias UTR (regiones flanqueantes no traducidas, pero necesarias para la traducción y la estabilidad del ARN mensajero), exones (codificantes) e intrones (secuencias de ADN no traducidas situadas entre dos exones y que serán eliminadas en el procesamiento del ARN mensajero).

Actualmente, y según los resultados arrojados por el proyecto ENCODE (Encyclopedia of DNA Elements), algunos autores consideran que debemos revisitar el concepto clásico de gen, formado por UTR, exones e intrones. Algunos estudios han mostrado que las secuencias UTR 5' se encuentran muy distantes de la secuencia traducida, abarcando secuencias muy largas y dificultando delimitar lo que es el gen. Además, un mismo transcrito puede dar lugar a ARN maduros totalmente diferentes, debido al proceso de *splicing* alternativo. De este modo, un mismo transcrito primario puede dar lugar a proteínas de secuencia y funcionalidad completamente distintas. En este proceso ya no se consideran ni los UTR ni los intrones. De acuerdo a esta observación, en este caso deberían considerarse genes diferentes, según algunos autores (Gerstein et al., 2007).

Finalmente, encontramos los **genes que codifican ARN no-codificantes**. Estos genes se expresan para dar lugar a un ARN funcional, como son los microARN y los piwiARN. Este tipo de ARN se involucran en la regulación de la expresión génica.

2. Por otro lado, los **pseudogenes** son versiones completas o parciales de genes que han ido acumulando mutaciones a lo largo de la evolución y que generalmente no se transcriben. De esta forma, tenemos el gen original activo y una copia inactiva. Aunque tradicionalmente se han pensado que estos pseudogenes son “genoma basura”, recientemente se ha visto que son capaces de regular la expresión de los genes de los que proceden mediante fenómenos de regulación antisentido, competencia de ARN o ARN de silenciamiento. Se estima que el genoma humano contiene unos 19 000 pseudogenes, clasificados entre no-procesados (aprox. 30 %) y procesados (aprox. 70 %). En este grupo encontramos:

- **Pseudogenes no-procesados:** son copias de genes originadas por duplicación que no se transcriben por carecer de una secuencia promotora y tras haber acumulado múltiples mutaciones, algunas de las cuales no tienen sentido originando codones de parada prematuros. Se caracterizan por poseer exones e intrones.
 - **Pseudogenes procesados:** son copias de ARN mensajero que se han retrotranscrito (ADN complementario, ADNc) que se introducen en otra posición del genoma. Este pseudógeno no tiene promotor, pero puede activarse por acción del promotor de un gen cercano (retrogenes). Estos pseudogenes al ser copias insertadas en el genoma carecen de intrones y de secuencia promotora.
- 3. ADN intergénico.** Las regiones intergénicas, o también llamadas extragénicas, comprenden la mayor parte de la secuencia del genoma humano y su función es generalmente desconocida. La mayor parte de estas regiones están compuestas por elementos repetitivos, como veremos a continuación, y el resto de la secuencia no responde a un patrón definido o clasificable. Se piensa que gran parte de este ADN intergénico puede ser un artefacto evolutivo sin función determinada en el genoma actual, sin funcionalidad, por lo que ha sido denominado “ADN basura”. Esta denominación incluye el ADN intergénico, secuencias intrónicas y pseudogenes. La denominación de estas zonas como “ADN no codificante” se debe a que se ha observado que algunas de estas regiones desempeñan una función reguladora y otras están altamente conservadas a lo largo de la evolución y podrían desempeñar otras funciones esenciales aún desconocidas o poco conocidas.
- Nuevamente, el proyecto ENCODE ha arrojado luz en este tema, mostrando que el 15% de la secuencia del genoma humano se transcribe a ARN maduro, y hasta el 90% a transcritos inmaduros, según el tejido analizado. Estos hallazgos ponen nuevamente en el foco de la redefinición del concepto de gen.
- 4.** Las secuencias de **ADN repetido** forman parte de este ADN intergénico, constituyendo más del 50 % del genoma humano. En este grupo de secuencias están aquellas derivadas de procesos de transposición (elementos transponibles), que suponen aproximadamente un 45 % del total del genoma. El otro 5 % corresponde a secuencias repetidas, sencillas y cortas, que se encuentran en miles de copias. Este grupo de secuencias no tiene una función muy conocida, pero es responsable de la variabilidad que da lugar a la evolución (Brown, 2017b).

Si atendemos a estas secuencias repetidas, encontramos los siguientes grupos (Figura 4):

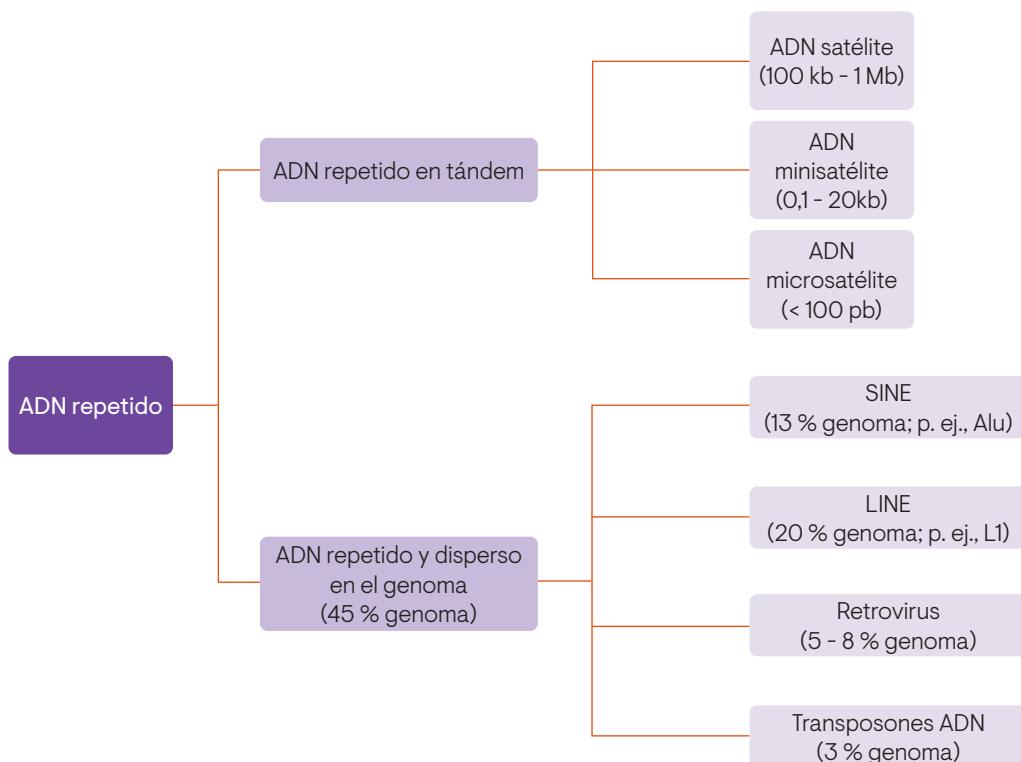
- a. **ADN repetido en tandem.** Está comprendido por el **ADN satélite**, los **minisatélites** y los **microsatélites**, cuya diferencia viene determinada por la longitud de la región que contiene la secuencia repetida en tandem, y no por la longitud de la secuencia que se repite. El ADN satélite ocupa regiones de cientos de kilobases (kb); los minisatélites ocupan entre 0.1-20 kb y los microsatélites son regiones de menos de 100 pb. Su característica principal es que se ordenan de manera consecutiva, de modo que secuencias idénticas, o muy similares, se disponen una detrás de otra.
- **Satélites:** el conjunto de secuencias de este tipo comprende aproximadamente 250 Mb del genoma humano. Su extensión se encuentra entre cinco nucleótidos a varios cientos, que se repiten en tandem miles de veces y generan regiones repetidas de entre 100 kb a varias megabases. Tienen una riqueza en contenido adenina y timina superior a la media del genoma (por lo tanto, son menos densas, propiedad que se utiliza en el laboratorio para su separación por gradiente de densidad). Existen seis tipos de repeticiones satélite, denominadas satélite 1, 2, 3, alfa, beta y gamma. Se diferencian en extensión y posición en el cromosoma.

- **Minisatélites:** compuestas por una unidad básica de secuencia de 6 a 25 nucleótidos que se repiten en tandem para dar estructuras de entre 100 y 200 000 pares de bases. El genoma humano tiene unos 30 000 minisatélites. Han sido relacionados con regulación de la expresión génica en procesos de transcripción, *splicing* alternativo o impronta genética. Son lugares cercanos a puntos de ruptura cromosómica, translocación genética y recombinación, además de ser hipermutables. Por tanto, forman parte de regiones inestables del genoma humano. Algunos minisatélites presentan una alta variabilidad entre individuos, por lo que han sido utilizados como marcadores en genética forense, estableciendo una huella genética característica del individuo.
 - **Microsatélites:** secuencias básicas de 2-4 nucleótidos, que en repetición tandem dan regiones de menos de 150 pb. Son llamados STR (*short tandem repeats*) y se identifican mediante reacción en cadena de la polimerasa (en inglés, *polymerase chain reaction* o PCR) de manera rápida y sencilla. En el genoma humano se estima la presencia de 200 000 microsatélites distribuidos homogéneamente.
- b. **ADN repetido y disperso en el genoma.** Representa el 45 % del genoma humano y se encuentra distribuido a lo largo de todo el genoma. Los elementos más importantes son los LINE y SINE, que se distinguen por el tamaño de la unidad repetida. Su característica principal es que se autopropagan. Para ello se transcriben a un ARNm intermedio, se retrotranscriben y se insertan en otro punto del genoma. Estas secuencias pueden resultar patogénicas bien por producir mutación al insertarse, desregular la expresión de genes próximos o por recombinación.
- **SINE (*short interspersed nuclear elements*).** Son elementos nucleares dispersos cortos. Secuencias cortas (cientos de bases), que aparecen repetidos miles de veces en el genoma humano. Suponen aproximadamente el 13 % del mismo, de los cuales el 10 % son la familia de elementos Alu. Estos son una familia de SINE de 250-280 pb, dímeros casi idénticos, ricos en contenido GC y con cola de poli A (vestigio de su origen ARNm). Poseen promotor de la ARN polimerasa III para su transcripción. Son retrotransposones no autónomos.
 - **LINE (*long interspersed nuclear elements*).** Constituyen el 20 % del genoma humano. La mayor familia de ellos es L1, una secuencia de 6 kb repetida más de 800 000 veces. Gran parte de estas secuencias se encuentran inactivas por metilación de su promotor. Los elementos LINE se encuentran flanqueados por las regiones no codificantes 5'UTR y 3'UTR. Son retrotransposones autónomos, ya que codifican las proteínas necesarias para su autopropagación. Algunos estudios relacionan estas secuencias con la regulación de la expresión génica, ya que se ha visto que genes próximos tienen disminuida su expresión. Esto resulta de interés, ya que el 80 % de los genes del genoma humano tiene un elemento L1 en sus intrones. La inserción aleatoria de los mismos puede ser patógena, dando lugar a enfermedades genéticas por modificación de la expresión génica.
 - **Retrovirus (HERV, *human endogenous retrovirus*).** Virus de ARN capaces de retrotranscribirse e integrar su genoma en el de la célula infectada. En el genoma humano encontramos restos de estos retrovirus integrados en el genoma a lo largo de la evolución, como huellas de infecciones retrovirales antiguas. En torno al 5-8 % del genoma humano está constituido por genomas derivados de estos virus. Su tamaño está entre 6-11 kb, pero habitualmente se presentan incompletos.

- **Elementos transponibles / transposones ADN.** Son el grupo mayoritario de secuencias repetidas, conteniendo el 45 % de las mismas. Derivan del fenómeno de transposición y son capaces de cambiar (saltar) dentro del genoma. Durante este proceso de transposición pueden insertarse en regiones del genoma donde interrumpen genes, provocando reorganizaciones cromosómicas, movilizar exones y, por tanto, crear nuevos genes. Asimismo, pueden interferir en la expresión de genes que estén próximos a este sitio de inserción. Pueden autopropagarse sin un intermediario de ARNm y retrotranscripción. Poseen el gen de la enzima transposasa, flanqueada por repeticiones invertidas. Su mecanismo es un corta y pega, moviendo la secuencia de un lugar a otro del cromosoma. Se estima que constituyen el 3 % de los elementos repetidos en el genoma. (Nota: bajo este nombre a veces se incluyen retrotransposones, pseudogenes procesados, SINE y LINE).

Figura 4

Tipo de secuencias de ADN repetidas en el genoma humano

Nota. Adaptado de *Genomes 4*, por T. A. Brown, 2017.

1.1.4. Secuencias repetidas y secuenciación masiva

Tal y como hemos visto, las secuencias de ADN repetido son abundantes a lo largo del genoma, tanto genomas eucariotas como el humano, como también veremos en temas posteriores, como en genomas bacterianos. Este tipo de repeticiones son un reto técnico para el alineamiento y ensamblaje de los genomas a nivel bioinformático. Desde un punto de vista computacional, estas repeticiones crean ambigüedades, que pueden producir errores en la interpretación de resultados. Debemos tener en cuenta que este problema se ve agravado en la secuenciación con lecturas cortas (tipo Illumina), mientras que la secuenciación de lecturas largas (Oxford Nanopore, PacBio) pueden paliar este efecto si la repetición es de menor tamaño que la zona secuenciada.

Tal y como se expone en el artículo de Treangen y Salzberg (2011), las repeticiones que son suficientemente divergentes no presentan problemas. Sin embargo, consideran que repeticiones de al menos 100 pb de longitud que ocurran dos o más veces en el genoma, que muestren más de un 97 % de identidad con al menos otra copia de sí mismas, son las que muestran un desafío computacional. Estas nos suponen diversos problemas:

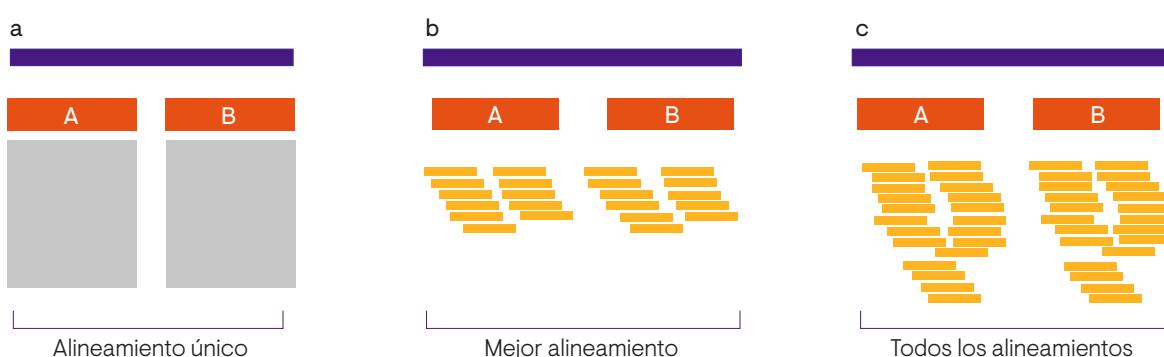
1. **Mapeo de multilecturas:** al alinear las lecturas frente al genoma de referencia, el mayor problema que encontramos son aquellas lecturas que pueden “encajar” en varias localizaciones. Aunque estas lecturas no son un problema para el alineador, sí lo son para análisis posteriores, como puede ser la identificación de mutaciones (SNP). Este problema puede paliarse utilizando lecturas más largas. A longitud mayor, mejor anclaje de la secuencia leída frente al genoma de referencia. Dado que la mayor parte de las secuencias repetidas no son exactamente iguales, muchas de ellas tendrán una única correspondencia (*best match*).

Existen varias estrategias para aminorar este problema (Figura 5):

- Ignorarlo, de manera que eliminemos todas las lecturas multiposición. Esta aproximación es aparentemente la más sencilla, pero limita el análisis no solo de genes de repetición, sino de familias multigen, con alta similitud. Esto puede conllevar que nos perdamos variantes de importancia biológica.
- Aproximación del mejor alineamiento (*best match*). El alineamiento con menor número de desacuerdos (*mismatches*) es el reportado. Si hay múltiples alineamientos igualmente válidos, el alineador elegirá uno al azar, o los informará todos. Esta es quizás la estrategia más aceptada comúnmente y que produce una estimación de cobertura más real.
- Tomar todos los alineamientos hasta un número máximo (X) o bien, ignorar aquellas lecturas alineadas más de X veces. Esta aproximación es la más flexible y hace posible encontrar errores de emplazamiento erróneo de las lecturas.

Figura 5

Estrategias para el manejo de lecturas multimapado



Nota. Los rectángulos anaranjados en la parte superior de la figura representan una región del genoma. Los dos rectángulos azules, etiquetados como A y B representan dos copias idénticas de un gen. Los pequeños rectángulos anaranjados en la parte inferior representan las lecturas de secuenciación. (a) Representa la estrategia en la que solo se reportan las lecturas que mapean únicamente. Dado que A y B son idénticos, ninguna lectura será reportada. (b) En la aproximación del mejor alineamiento se reporta solo el mejor alineamiento para cada lectura, de acuerdo con el algoritmo que clasifica estas lecturas. (c) Estrategia de reporte de todos los alineamientos para cada lectura multimapado, incluyendo aquellas que sean de peor calidad de mapeo. Adaptado de “Repetitive DNA and next-generation sequencing: computational challenges and solutions”, por T. J., Treangen y S. L. Salzberg, 2011, *Nature Reviews Genetics*, 13(1), pp. 36-46. Recuperado de <https://doi.org/10.1038/nrg3117>.

2. **Ensamblaje de genomas *de novo*.** En este proceso, lo que queremos es reconstruir el genoma sin tener una referencia (montar el puzzle sin tener la fotografía final de cómo debe ser). Las lecturas cortas son en sí mismas un problema a la hora de reconstruir un genoma desde cero. Tenemos millones de fragmentos (alta cobertura) pero de un tamaño muy limitado. En este caso, esta alta cobertura puede solventar algunos problemas de las lecturas cortas, sin embargo, las zonas repetidas no se solucionan de esta forma. Nos encontramos dos problemas principales:
- Cuando la repetición es mayor que el tamaño de la lectura se crean huecos (*gaps*) en el ensamblaje. Esto desemboca en genomas altamente fragmentados.
 - Las repeticiones pueden ser colapsadas en una única secuencia, produciendo reorganizaciones erróneas y secuencias quimera.

Los ensambladores, tal y como veremos en un tema posterior, pueden reconstruir el genoma bien por solapamiento de las lecturas completas o por fragmentos de estas. Desde el punto de vista técnico, las secuencias repetidas son un “bucle” o “burbuja”, una bifurcación en el camino de la resolución del genoma. Si este bucle se resuelve de manera errónea se generan uniones falsas (secuencias quimeras). Si el ensamblador es más conservador, romperá el ensamblaje en ese punto de bifurcación, dejando un fragmento corto pero exacto. Cuando manejamos lecturas pareadas, la información de distancia media entre ellas es utilizada por algunos ensambladores para colocarlas y determinar la mejor posición e incluso para llenar con nucleótidos indefinidos (Ns) esos huecos indeterminados.

3. **Alineamiento y ensamblaje de secuencias de ARN.** Aunque no es tema de esta asignatura, el análisis de expresión génica y ensamblaje de transcritos también se ve influenciado por las secuencias repetidas. El análisis de expresión génica se basa en el mapeo de las lecturas frente a un genoma de referencia, con el posterior recuento de estas en cada unidad genómica; asimismo el ensamblaje de transcritos es una situación similar al ensamblaje *de novo*. Por tanto, los problemas que hemos visto en los dos puntos anteriores son similares a los que encontraremos en el análisis de secuencias repetidas de un transcriptoma.

1.2. Patrones de transmisión de las enfermedades genéticas

1.2.1. Variabilidad del genoma

La variabilidad dentro de un genoma puede deberse a una sustitución, delección o inserción. De esta manera, encontramos polimorfismos o alelos genéticos. Estas variaciones pueden darse tanto en regiones codificantes como no codificantes y, obviamente, debido a la degeneración del código genético, no toda alteración supone una alteración en la proteína o su nivel de expresión. Muchos de estos cambios son silenciosos y no tienen repercusión (Trent, 2012). Entre los cambios que pueden darse se encuentran:

- **Single nucleotide polymorphisms (SNP):** son la mayor fuente de variabilidad en los genomas de seres humanos, suponiendo la variación de un solo nucleótido. Se estima que su frecuencia es de un SNP por cada 500-1000 pb. Son utilizados como marcadores genéticos en estudios a gran escala mediante el empleo de chips de ADN (*microarrays*). Actualmente, son la base de nuestros estudios de secuenciación masiva. La identificación de estas variantes de nucleótido único por este método son las SNV (*single nucleotide variants*). Se observan regiones del genoma con mayor grado de conservación, teóricamente por ser regiones con una función altamente esencial para el organismo. Las zonas que codifican para proteínas presentan menor número de SNV que las zonas intergénicas.

- **Variación estructural (duplicaciones, inversiones, inserciones o variantes en número de copias).** Todos estos eventos afectan a una gran proporción del genoma. Este término abarca un gran número de eventos genéticos que implican segmentos de más de 1 kb de longitud.

Aunque pueda parecer sencillo, establecer la relación entre un gen, su polimorfismo y una enfermedad, no es una tarea sencilla.

1.2.2. Tipos de enfermedades genéticas

Se considera enfermedad genética aquella en la que existe un componente genético o hereditario implicado. Se clasifican en cromosómicas, monogénicas y multifactoriales o de herencia compleja (Trent, 2012).

1. **Cromosómicas.** Afectan a 7/1000 nacimientos. Se deben a alteraciones en el número o la estructura de los cromosomas. Son responsables de aproximadamente el 50 % de los abortos espontáneos del primer trimestre. Se pueden clasificar en dos grandes grupos:
 - **Numéricas:** alteración en el número normal de cromosomas. Habitualmente encontramos 23 pares. Puede afectar a un solo par de cromosomas (aneuploidía), habiendo un solo cromosoma (monosomía) o más de dos (trisomía, tetrasomía). Ejemplo es el síndrome de Down (trisomía del cromosoma 21), síndrome de Turner (monosomía del cromosoma X), Edwards (trisomía 18), Patau (trisomía 13), Klinefelter (XXY) o XYY. Si la alteración afecta a todos los cromosomas se habla de euploidías, de manera que el individuo puede tener una sola dotación cromosómica (haploidía, 23 cromosomas totales) o más de dos dotaciones completas (triploidía, 69; tetraploidía, 92). Estas suelen ser letales a nivel embrionario, por lo que habitualmente el embarazo no llega a término.
 - **Estructurales:** alteraciones en la estructura de los cromosomas, como pueden ser grandes inserciones o delecciones, reorganizaciones...
 - **Delecciones:** eliminación de una porción del cromosoma. Ejemplos: síndrome Wolf-Hirschhorn (deleción parcial del brazo corto del cromosoma 4), síndrome de Jacobsen (deleción 11q terminal).
 - **Duplicaciones:** de una región considerable del cromosoma. Enfermedad Charcot-Marie-Tooth con la duplicación del gen *PMP22* en el cromosoma 17.
 - **Translocaciones:** transferencia parcial de un cromosoma a otro.
 - **Inversiones:** una parte del genoma se rompe y se reorienta en dirección opuesta.
2. **Enfermedades monogénicas o mendelianas.** Causadas por mutaciones en genes individuales. Tienen patrones específicos de transmisión y su prevalencia es de 2/100 habitantes. Ejemplos son la fibrosis quística o la enfermedad de Huntington. Son el único tipo de enfermedades que presentan un patrón de herencia claro, al estar causadas por mutaciones en un único gen. Esta mutación tiene un fuerte impacto en el fenotipo, es decir, el riesgo de desarrollar la enfermedad será alto y, además, será igual para todas las familias que presentan la mutación.

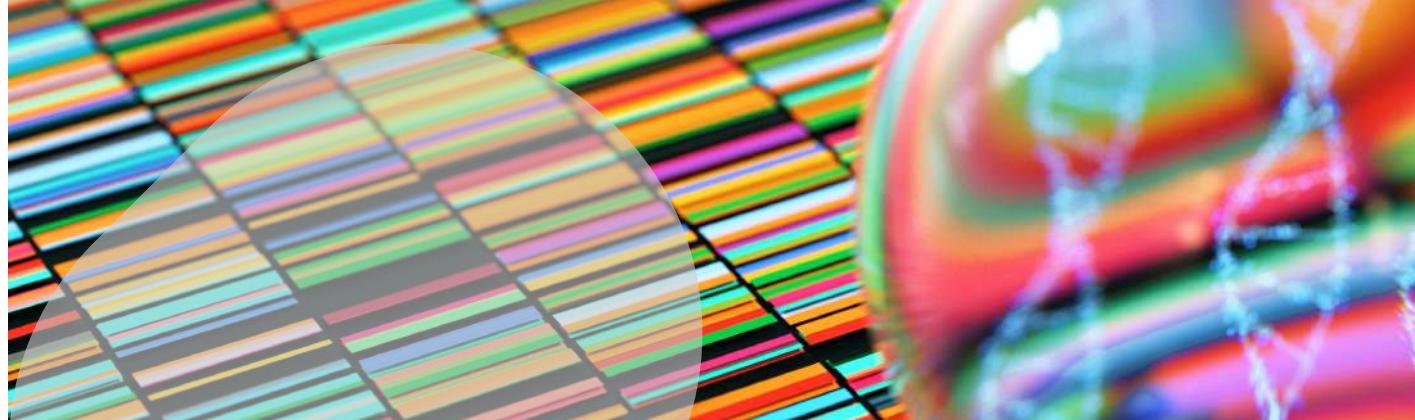
El patrón de herencia que puede mostrar una enfermedad de tipo monogénica o mendeliana se clasifica en:

- **Autosómico dominante:** se manifiesta en individuos heterocigotos. Es suficiente que mute una de las dos copias del cromosoma de un gen para que se manifieste la enfermedad. Los individuos enfermos suelen tener a uno de sus progenitores enfermos y, asimismo, la probabilidad de tener descendencia afectada es de un 50 %. Frecuentemente son mutaciones de ganancia de función (el alelo mutado tiene una nueva función que provoca el desarrollo de la enfermedad) o bien, por pérdida de función. Son enfermedades de baja penetrancia (solo una parte de los individuos que portan la mutación desarrollan la enfermedad). Ejemplos: Huntington, Marfan, algunos tipos de cáncer colorrectal hereditario.
 - **Autosómico recesivo:** la enfermedad solo se manifiesta en individuos homocigóticos recesivos (ambas copias del gen están mutadas). Suelen ser mutaciones que causan pérdida de función, de modo que la causa de la enfermedad es la ausencia de acción de un gen. Habitualmente, el individuo enfermo tiene ambos progenitores sanos, pero son portadores de la mutación. El 25 % de la descendencia queda afectada. Ejemplos: Fibrosis quística, anemia falciforme, Tay-Sachs, atrofia muscular espinal.
 - **Dominante ligado al X:** mutaciones en el cromosoma X. Patrón de herencia especial, siendo poco frecuente. Las mujeres tienen mayor prevalencia de la enfermedad que los hombres. Un varón enfermo tendrá a todos sus hijos varones sanos, mientras que a todas sus hijas mujeres enfermas. Por otra parte, una mujer enferma tendrá un 50 % de su descendencia enferma, independientemente del sexo. Ejemplo: Aicardi.
 - **Recesivo ligado al X:** mutaciones en el cromosoma X, en este caso los varones están más frecuentemente afectados. Un varón portador siempre será enfermo (solo posee un cromosoma X, y este está afectado). Su descendencia serán varones sanos (ya que solo les transmite el cromosoma Y) e hijas portadoras. Una mujer portadora tendrá una descendencia compuesta por un 50 % de hijas portadoras y un 50 % de hijos enfermos. Ejemplo: hemofilia A, Duchenne.
 - **Ligado a Y:** solo pueden manifestarse en varones, cuya descendencia es 100 % de hijas sanas y 100 % de varones enfermos. Habitualmente estas mutaciones causan infertilidad. Ejemplo: infertilidad masculina hereditaria.
 - **Mitocondrial:** mutaciones en el genoma mitocondrial, que solo se transmite por herencia materna. La gravedad de la mutación depende del porcentaje de genomas afectados en la población mitocondrial (heteroplasmia). Ejemplos: neuropatía óptica hereditaria de Leber.
3. **Enfermedades multifactoriales o de herencia compleja.** En este tipo de enfermedades contribuyen tanto factores genéticos como factores de tipo ambiental, o la interacción entre ambos. Son las enfermedades hereditarias más numerosas, responsables de malformaciones congénitas como labio leporino, alteraciones en el tubo neural, diabetes de tipo 2, algunos tipos de hipertensión, cardiopatías o enfermedades psiquiátricas.



La mutación de un único gen no es suficiente para que se manifieste la enfermedad. Se utiliza el concepto de variación génica o polimorfismo. Para que el individuo desarrolle la enfermedad debe tener una combinación concreta de los genes implicados, de esta forma, las mutaciones en genes individuales pueden tener una frecuencia alta en la población sin causar un efecto en el fenotipo. Es, por tanto, complicado rastrear este tipo de enfermedades, dado que además los factores ambientales tienen un factor contribuyente.

Cada mutación tiene una contribución fenotípica distinta y, además, el riesgo para cada familia puede ser distinto en función del conjunto de factores de riesgo que presente. En estas enfermedades se llevan a cabo estudios de asociación, revisando una batería de genes distribuidos por todo el genoma, localizando aquellos polimorfismos más frecuentes en individuos afectados que en controles. Esto nos arroja un resultado de probabilidad estadística y se establece la hipótesis de que ciertos polimorfismos se relacionan con una enfermedad.



Capítulo 2

Introducción a la medicina preventiva personalizada, exposoma y tipos de datos

2.1. Medicina preventiva personalizada

2.1.1. Definición y objetivos

La **medicina preventiva personalizada** es el área de la medicina que se centra en manejar y prevenir problemas de salud incorporando datos de salud pública o datos epidemiológicos generados a partir de tecnologías “-ómicas” (genómica, proteómica, transcriptómica, etc.).

Su objetivo principal es optimizar las estrategias preventivas teniendo en cuenta las características de cada individuo, especialmente cuando este se encuentra en estado de salud o en las etapas tempranas de la enfermedad (Nevado *et al.*, 2018).

En este punto se incluyen las técnicas de base genética orientadas a la prevención en la práctica clínica diaria, incorporando nuevas pruebas genéticas, secuenciación del genoma al nacer o el perfil farmacogenético individual. Es decir, el perfil genético de cada individuo, junto con otros datos ómicos, permitirán identificar y aplicar un abordaje preventivo más efectivo para cada paciente.

De esta manera, la medicina preventiva personalizada permitirá anticiparse al desarrollo de futuras enfermedades de manera individualizada, actuando sobre ellas con anticipación.

2.1.2. ¿Qué aporta la genómica en esta área?

Tenemos que reconocer los grandes avances que se han realizado gracias a la secuenciación del genoma desde la finalización del Proyecto Genoma Humano, que han permitido conocer las bases moleculares de muchas enfermedades de base genética. Paralelamente, esto ha conllevado la disminución de los costes de análisis y la optimización de los tiempos de respuesta, permitiendo su uso como herramienta clínica. Estos avances se han complementado con otras técnicas ómicas, que aportan una información complementaria en lo que está ocurriendo a nivel molecular y celular. Hasta el momento, son la genómica y la proteómica las que han tenido mayor relevancia y desarrollo.



En la asignatura en la que estamos, la identificación de las distintas variantes genéticas de los genes implicados en procesos patológicos puede ser utilizada como biomarcador de la posibilidad de desarrollar una enfermedad, estableciéndose el riesgo estimado y las medidas preventivas que tomar. El análisis de este perfil genético debe contextualizarse con datos relacionados con el entorno de cada persona, como por ejemplo, zona geográfica, exposición a radiación ultravioleta, hábitos saludables, etc., ya que como vimos en el capítulo anterior, muchas enfermedades son complejas y no solo dependen de factores genéticos.

Con el fin de ahondar en este conocimiento se han desarrollado programas dirigidos a la investigación científica traslacional. Algunos ejemplos son:

- **Research Program on Genes, Environment and Health** (Kaiser Permanente Northern California Division, EE. UU.). Estudios a gran escala de factores genéticos y ambientales que influyen en enfermedades cardíacas, cáncer, asma o trastornos psiquiátricos, entre otros.
- **MyCode Community** (Geisinger Health System). Busca la mejora de la atención médica a través del análisis del ADN de los individuos, antes de que aparezcan síntomas de la patología.
- **100,000 Genomes Project** (Genomics England, Department of Health & Social Care, National Health System, NHS). En este proyecto se están secuenciando 100 000 genomas de pacientes del sistema público de salud británico con enfermedades raras y sus familiares, así como pacientes con cáncer. Su propósito es integrar la medicina genómica en el sistema público de salud, creando un servicio de medicina genómica que brinde un acceso amplio y equitativo a este tipo de pruebas a toda la población, de manera que todo el mundo pueda acceder a los sistemas de diagnóstico más sofisticados sin coste alguno.



Enlaces de interés

En el siguiente enlace, podrás acceder a la página web de Kaiser Permanente sobre el Research Program on Genes, Environment and Health.

<https://divisionofresearch.kaiserpermanente.org/genetics/rpgeh/rpgehabout>

En el siguiente enlace, podrás acceder a la página web de Geisinger Health System sobre la iniciativa MyCode Community Health Initiative.

<https://www.geisinger.org/precision-health/mycode>

2.1.3. Aplicaciones de la medicina preventiva personalizada

Como hemos comentado en el primer apartado de este capítulo, la estrategia de medicina preventiva personalizada tiene como objetivo preservar la salud el máximo tiempo posible una vez conocida la posibilidad o probabilidad que tiene un individuo de desarrollar una enfermedad de base genética.

En este contexto existen dos niveles de prevención:

Prevención primaria

Es la prevención que se realiza en un contexto de salud y se basa en las características propias del individuo para evitar factores de riesgo y diseñar medidas preventivas individualizadas para evitar la aparición de una enfermedad.

Las pruebas genéticas que se realizan en este nivel buscan determinar y cuantificar el riesgo del individuo para desarrollar una enfermedad, pudiendo **estratificar en grupos de riesgo**. Las enfermedades de base hereditaria (enfermedades coronarias, presión arterial alta, diabetes, algunos tipos de cáncer) que suelen presentarse con base familiar, son un ejemplo de patologías que podrían ser diana de este tipo de intervención. Por tanto, las patologías diana de este tipo de prevención son:

- **Enfermedades cardiovasculares.**
- **Cáncer**, especialmente colon y mama, para los que se han asociado ya más de 50 y 9 variantes, respectivamente. Cada una de estas variantes por sí solas carecen de utilidad clínica por su limitado poder predictivo, pero la interacción de estas variantes en un mismo individuo, en combinaciones alélicas, nos indica el riesgo relativo para desarrollar esa enfermedad.
- **Enfermedades neurológicas** como, por ejemplo, la enfermedad de Alzheimer, para la que se han descrito más de un centenar de marcadores de SNP; la esclerosis múltiple (más de cien variantes que explican al menos una cuarta parte de la heredabilidad de esta enfermedad) o la enfermedad de Parkinson.
- **Medicina reproductiva** en la detección de enfermedades genéticas fetales, como por ejemplo la detección de anomalías cromosómicas mediante test prenatal no invasivo o, incluso, diagnóstico genético preimplantacional. Estas técnicas permitirán realizar un asesoramiento genético individualizado antes y después de las mismas.
- **Cribado neonatal**. Es muy habitual que cada bebé sea sometido a una serie de pruebas bioquímicas con el objetivo de detectar de manera precoz enfermedades de base genética. Este cribado metabólico es parte de una medicina preventiva. En el futuro, estas pruebas serán realizadas mediante secuenciación del genoma completo del recién nacido. Este tema puede resultar polémico, puesto que existirá la posibilidad de detectar trastornos para los que hoy en día no tenemos tratamiento efectivo, así como mutaciones cuyo significado es desconocido.
- **Genética nutricional**, con el objetivo de establecer un tratamiento nutricional basado en la nutrigenética y la nutrigenómica. La **nutrigenética** estudia la influencia de las variaciones genéticas en la respuesta de un individuo a los nutrientes. Por otro lado, la **nutrigenómica** estudia la influencia de los nutrientes en la expresión de los genes del individuo, lo que puede contribuir a desarrollar alimentos funcionales dirigidos a prevenir o intervenir enfermedades.

Prevención secundaria

Es el conjunto de medidas que se toman una vez que el individuo desarrolla una enfermedad, especialmente en estado presintomático o de difícil detección.



Por tanto, la detección precoz será clave, apoyada de nuevas pruebas genéticas, secuenciación de genoma al nacer y la farmacogenética y farmacogenómica. Estas intervenciones tienen especial interés en cáncer y enfermedades raras.

La detección precoz del **cáncer** es clave para mejorar el pronóstico de este. Por tanto, la prevención secundaria en este campo se centra en la detección precoz mediante programas de cribado dirigido a poblaciones de alto riesgo. Para ello se utilizarán **biomarcadores de la enfermedad**. El desarrollo de la biopsia líquida, prueba para la detección de células cancerosas tumorales o fragmentos de ADN circulante en sangre, ha permitido avanzar en la detección precoz con alta especificidad, así como monitorizar pacientes ante posibles recaídas.

Actualmente, el campo de la biopsia líquida permite detectar biomarcadores genéticos y proteicos en sangre para ocho tipos de tumores sólidos comunes, como ovario, hígado, estómago, páncreas, esófago, colorrectal, pulmón y mama. Hay que destacar que estas pruebas no intentan sustituir a las actuales pruebas de detección clásicas, como la mamografía o colonoscopia, sino proporcionar información adicional e identificar pacientes en estadios muy tempranos.

En el campo de las enfermedades raras, actualmente existen más de 8000 enfermedades raras descritas, las cuales afectan a más del 5 % de la población mundial. Se estima que el 80 % de las mismas tienen origen genético y que el promedio en tiempo hasta su diagnóstico oscila entre cinco y diez años. Por tanto, es de vital importancia buscar biomarcadores que nos permitan disminuir este tiempo de diagnóstico, para proporcionar a los pacientes un tratamiento adecuado en la mayor brevedad posible.

2.1.4. Retos de la medicina preventiva personalizada

Podemos clasificar los retos en tres grupos:

1. Retos de formación, educación y difusión

Integrar las tecnologías ómicas, especialmente la genómica, en nuestro sistema de salud requiere educación y difusión a todos los niveles, no solo a los profesionales sanitarios. Los profesionales que intervienen en este tipo de pruebas deberán conocer cuáles son los tipos de prueba, cómo han de aplicarse y cómo comunicar los resultados, teniendo en cuenta las limitaciones y beneficios para el paciente.

Para ello, es vital contar en el equipo con un profesional genetista, que actúe con la función de asesor genético, para dar la información y asistencia a las familias en riesgo o afectadas por un trastorno genético, asegurando que estas tomen decisiones de manera informada. Además de esta interpretación de los datos, la obtención e integración de los mismos pasa por la incorporación de profesionales bioinformáticos en el sistema nacional de salud.

2. Retos para la implantación de estas estrategias

Quizás el mayor reto en la implantación de estas técnicas en la práctica clínica se encuentra en que sean útiles desde el punto de vista clínico, pero también coste-efectivos. Aunque hay avances en nuestro país en la elaboración de una Estrategia Nacional de Medicina Personalizada de Precisión, la puesta en marcha está siendo lenta y costosa. Es necesario garantizar la continuidad, optimizar los recursos, asegurar su sostenibilidad y, sobre todo, la equidad en el acceso a estas pruebas. Países como Reino Unido, que han apostado fuerte por esta estrategia han creado centros de medicina genómica, dotados de conocimiento técnico y recursos humanos.

3. Retos éticos y legales

Después de analizar este capítulo queda patente que existe una preocupación por la situación ética y legal que estas pruebas genéticas plantean. Es vital contar con una adecuada supervisión médica y asesoramiento genético, para evitar una interpretación errónea de los resultados, conduciendo a situaciones de angustia o ansiedad y toma de decisiones médicas inapropiadas. Por otra parte, es obvio que estos datos son altamente sensibles, lo que requiere implementar medidas que garanticen la total privacidad de la información de los pacientes y su correcta utilización. Debemos prevenir que estos datos sean fuente de una discriminación genética.



Enlaces de interés

En el siguiente enlace, podrás encontrar un podcast donde se explica la medicina personalizada y de precisión.

<https://www.institutoroche.es/jornadas/106-medicina-personalizada-de-precision-datos-que-curan>

En el siguiente enlace encontrarás un documento con estos conceptos detallados y desarrollados.

https://www.institutoroche.es/recursos/publicaciones/185/informes_anticipando_medicina_preventiva_personalizada

2.2. Exposoma

2.2.1. Definición

Hasta ahora hemos centrado nuestra atención en todos aquellos factores genéticos que son responsables de un fenotipo, incluyendo un fenotipo de enfermedad. Pero tal y como explicamos en el Capítulo 1, los factores no genéticos juegan un papel importante en el desarrollo de estas alteraciones.



Todos estos factores no genéticos, elementos que componen el entorno y a los que el individuo está expuesto, son lo que denominamos el **exposoma**. Estos factores incluyen contaminantes ambientales, ámbito socioeconómico, entorno urbano, agentes infecciosos o estilo de vida.

Este término fue acuñado por Wild (2005) haciendo explícitamente referencia a factores ambientales, que pueden determinar que dos individuos con la misma carga genética (gemelos homocigotos) presenten características externas diferentes.

Veremos a lo largo de este apartado cómo ciertos factores ambientales determinan la predisposición a ciertas patologías. Asimismo, determinar esta relación resulta una tarea compleja, por lo que este abordaje debe ser multidisciplinar, incluyendo disciplinas como la epidemiología, la medicina clínica, las ciencias ómicas o las ciencias de datos.

Existe actualmente un gran interés en torno al exposoma, dado que la Organización Mundial de la Salud (OMS) cifra en un 24 % las enfermedades humanas que están condicionadas por factores no genéticos y que, por tanto, podrían evitarse. El proyecto Global Burden of Disease (Gakidou *et al.*, 2017) determinó que nueve millones de muertes al año (16 % de las mundiales) están asociadas exclusivamente a la contaminación del aire, agua y suelo.

2.2.2. Factores no genéticos

Estos factores no genéticos o exposoma pueden definirse en tres dominios: externo general, externo específico e interno.

- **Externo general:** influencias sociales, económicas y psicológicas, como son capital social, educación, situación financiera, estrés psicológico y mental, entorno urbano-rural o el clima.
- **Externo específico:** radiación, agentes infecciosos, contaminantes químicos y ambientales, dieta, factores de estilo de vida (tabaco, alcohol), intervenciones médicas.
- **Interno:** metabolismo, factores hormonales, morfología corporal, actividad física, microbiota, inflamación, estrés oxidativo y envejecimiento.

Todos estos factores pueden tener un impacto en la salud (positivo o negativo) cuando alteran la biología de los organismos, como puede ser mediante modificaciones epigenéticas, alteración de la microbiota o alteraciones metabólicas.

2.2.3. ¿Cómo se estudia el exposoma?

El estudio de todos estos factores resulta altamente complejo, por la enorme variabilidad en cuanto a la naturaleza de estos factores. Además, el exposoma es dinámico, varía a lo largo del tiempo, y también es gradual, ya que el efecto de estos factores sobre el organismo depende de la “dosis” a la que se enfrenta. A estos hay que sumar la heterogeneidad de los individuos, aunque se pueden realizar correlaciones poblacionales de sensibilidad a estos factores no genéticos.

Algunos autores defienden la existencia de **ventanas de susceptibilidad** (Terry *et al.*, 2019; Wright, 2017), aquellas épocas donde el individuo es más susceptible a esta exposición. Estas etapas se identifican con etapas temporales en la vida: prenatal, primera infancia, pubertad, edad reproductiva, periodo de gestación y vejez. En este contexto ha surgido el estudio Lifestage Exposome Snapshots (LEnS) (Shaffer *et al.*, 2017), donde se correlacionan estas ventanas de susceptibilidad en órganos específicos y no en el individuo en su conjunto.

Como vemos, el estudio del exposoma no es sencillo y, como hemos comentado previamente, requiere de una visión multidisciplinar para su abordaje. En este sentido, poder monitorizar las variables que afectan al individuo es un eje fundamental, facilitado por el desarrollo y modernización de sensores ambientales, sistemas de información geográfica, aplicaciones, *wearables*, etc.

Estos **elementos de toma de datos** van parejos a mejoras en los métodos de análisis y desarrollo en el análisis e interpretación de los datos, asistido por **métodos bioestadísticos y bioinformáticos**. Hasta el momento, los estudios de exposoma se basan en la biomonitorización humana (recogida de muestras biológicas) y ambiental (muestras ambientales), así como el estudio de biomarcadores de exposición (análisis de sustancias).

Las ciencias ómicas, especialmente la genómica, proteómica y metabolómica, aportan el enfoque necesario para el descubrimiento de nuevos biomarcadores.

A continuación, se exponen algunos proyectos que han ahondado en el estudio del exposoma. Esta información está modificada del informe de Olea *et al.* (2020) y se presenta en la Tabla 4.

Tabla 4

Proyectos que analizan exposoma

Proyecto	Ámbito	Organización promotora	Web
The Human Early-Life Exposome Project (HELIX)	Europa	Fundació Centre de Recerca en Epidemiología Ambiental (CREAL) y Comisión Europea	https://www.projecthelix.eu/
HBM4EU	Europa	Agenda de Medioambiente Europea y la Comisión Europea	https://www.hbm4eu.eu/
Health and Environment-Wide Associations Based on Large Population Surveys (HEALS)	Europa	Varios	https://cordis.europa.eu/project/ id/603946
EXPOsOMICS	Europa	Comisión Europea	https://www.isglobal.org/en/-/ exposomics?inheritRedirect=true
Health and Exposome Research Center: Understanding Lifetime Exposures (HERCULES)	Europa	Instituto Tecnológico de Georgia e Instituto Tecnológico de Emory	https://emoryhercules.com/
Participative Urban Living for Sustainable Environments (PULSE)	Europa	Comisión Europea	http://www.project-pulse.eu/
OBERON	Europa	Instituto Nacional de Investigación Médica y Salud de Francia	https://oberon-4eu.com/
Nutrition in Early Life and Asthma (NELA)	España	ISCIII	https://nela.imib.es/plataforma/index.jsf
Infancia y Medio Ambiente (INMA)	España	CIBER	https://www.proyectoinma.org/

Nota. Adaptado de *Informes Anticipando. Exposoma*, por N. Olea, M. Casas, A. Castaño, J. Mendiola, M. Vrijheid, J. Arenas, Á. Carracedo, P. Lapunzina y F. Martín-Sánchez, 2020, Fundación Instituto Roche.

Recuperado de <https://www.institutoroche.es/observatorio/exposoma>

2.2.4. Patologías más comunes y su relación con exposoma

Las enfermedades más comunes relacionadas con la exposición a distintos factores ambientales son, según organizaciones como la Organización Mundial de la Salud (OMS) o el Centro para el Control y la Prevención de Enfermedades (en inglés, Centers for Disease Control and Prevention o CDC), las enfermedades que más comúnmente afectan a las sociedades desarrolladas, como son enfermedades cardiovasculares, oncológicas, respiratorias o endocrinas.

Si atendemos a las **enfermedades cardiovasculares**, es ampliamente conocido por todos que los hábitos de dieta no saludables tienen un efecto negativo sobre el sistema cardiovascular, tales como hipertensión o colesterolemia. El estudio del **microbioma intestinal** ha demostrado la relación entre el consumo de ciertos alimentos, especialmente grasas y azúcares, con el riesgo de este tipo de enfermedades. Entre estos compuestos están la fosfatidilcolina o la L-carnitina, que metabolizados por los microorganismos intestinales generan subproductos que promueven la inflamación intestinal. Sin embargo, no debemos alarmarnos, ya que estos problemas aparecen por sobreexposición a estas sustancias, por tanto, se deben a un consumo excesivo y desmedido de ciertos alimentos que los contienen. Por otro lado, factores como **contaminación aérea, ruido o estilo de vida** son también desencadenantes de problemas cardiovasculares, por lo que se insta a la población a disminuir la implicación en todos ellos.

Las **enfermedades oncológicas**, el cáncer, son procesos multifactoriales aún muy desconocidos. Sin embargo, hay ejemplos claros y constatados de factores no genéticos que influyen en el desarrollo de cáncer, como son:

- **Sobreexposición a la luz ultravioleta (UV)** y generación de **cáncer de piel (melanoma)**, por el efecto de la radiación UV como promotor mutagénico en el ADN de las células de la piel. Además de la luz UV se han relacionado compuestos tóxicos, pesticidas, ritmo circadiano o la actividad física como factores que influyen en este tipo de enfermedad (Gracia-Cazaña *et al.*, 2020).
- **Cáncer de mama.** Aunque este tipo de cáncer tiene un marcado carácter genético, la exposición a compuestos químicos como el DDT (dcloro difenil tricloroetano), hidrocarburos policíclicos aromáticos o metales pesados, especialmente en la etapa prenatal, etapa de gestación, pubertad y menopausia, pueden ser desencadenantes (Bessonneau & Rudel, 2020; Jones & Cohn, 2020).
- **Cáncer de pulmón.** Su principal factor de riesgo es el tabaquismo, ya que más del 80 % de los mismos se relacionan con este hábito (Juarez & Matthews-Juarez, 2018; McKeon *et al.*, 2021).

Entre las **enfermedades respiratorias** destaca la enfermedad pulmonar obstructiva crónica (EPOC) o el asma. Ambas están relacionadas con la exposición prenatal al tabaco. Por otro lado, los compuestos organoclorados, metales pesados o perfluorados son los principales factores no genéticos asociados (Benjdir *et al.*, 2021).

Finalmente, la presencia cada vez mayor de **compuestos químicos que actúan como disruptores endocrinos**, impidiendo el normal funcionamiento de las rutas hormonales. Los disruptores hormonales son moléculas que, por su similitud con las hormonas, “engaños” a nuestro sistema, alterando su función.



Ejemplo

Ejemplo de estos compuestos son los ftalatos empleados en plásticos o cosméticos. Sus ventanas de susceptibilidad más acusadas son el embarazo, la lactancia y la infancia (Buck Louis *et al.*, 2019). Es un hecho bastante curioso cómo algunos disruptores, como el bisfenol-A (BPA), actualmente restringido en Europa, pueden afectar al desarrollo de diabetes *mellitus* de tipo II. Está claro que este tipo de diabetes también se ve influenciada por una vida sedentaria o una mala alimentación. Es un ejemplo de cómo varios factores no genéticos pueden sumar para dar un efecto mayor.

2.2.5. Retos asociados

A lo largo de este capítulo se ha visto la complejidad en el estudio del exposoma, lo que plantea algunas limitaciones y retos por solventar. A continuación, se exponen los principales:

- Es necesaria la convergencia de un gran número de disciplinas distintas, cada una de las cuales trabaja con herramientas propias y metodología dispar.
- El número, diversidad y combinación de perfiles que son posibles es un gran desafío para los estudios de asociación.
- No existen datos sistemáticos sobre la exposición a muchos de los factores analizados.
- Armonización en la recogida y análisis de datos.
- Tamaños muestrales en ocasiones insuficientes. Necesidad de bases de datos homogeneizadas y coordinadas para mejorar la armonización y tamaño muestral.
- Necesidad de nuevos métodos de análisis, parejos a recursos digitales potentes y herramientas computacionales más novedosas para la correlación de factores genéticos y no genéticos.
- Métodos computacionales de predicción y correlación.
- Análisis detallado de las vías bioquímicas influenciadas para cada enfermedad.
- Identificación de biomarcadores específicos.



Enlace de interés

En el siguiente enlace encontrarás un explorador de factores de exposoma relacionados con distintas enfermedades.

<http://exposome-explorer.iarc.fr/>



2.3. Datos de medicina personalizada/preventiva y de precisión

Como punto final a este capítulo, veremos qué tipos de datos se utilizan en medicina personalizada de precisión, así como sus principales características.

2.3.1. Datos integrados

Debemos ser conscientes de que vivimos actualmente en una revolución tecnológica fomentada por la creciente capacidad de generar, almacenar y procesar datos de diversos tipos. Estos datos recogidos en el campo de la salud son altamente complejos, heterogéneos (variabilidad de fuentes de información) y sensibles, siendo necesario mantener siempre su **carácter confidencial**. Estas características empujan el diseño y mejora de herramientas e infraestructuras computacionales, como son la minería de datos o la inteligencia artificial, destacando en el ámbito de la salud las **técnicas de aprendizaje automático**.



La medicina del futuro está destinada a basar parte de su carácter predictivo en el análisis de datos ómicos, combinados con información como historia clínica y manejados por herramientas de inteligencia artificial.

La genómica fue la primera de las ciencias ómicas que se desarrolló, pero actualmente el grupo de las ciencias ómicas es heterogéneo y diverso, y está formado por otras disciplinas como epigenómica, transcriptómica, metagenómica, metabolómica, proteómica, secretómica, interactómica, citómica, fenómica, exposómica y farmacogenómica. Todas ellas proporcionan datos valiosos, que combinados con datos de los pacientes conforman los datos que integrar para avanzar en la medicina personalizada o preventiva de precisión.

2.3.2. Tipos, fuentes y características de los datos

Los datos obtenidos pueden clasificarse de distintas maneras, aquí se presentan las principales:

- **Nivel de organización biológica:** indica la complejidad biológica de las entidades sobre las que se recogen los datos. Puede ser molecular (es decir, ADN), celular, tisular, órgano, organismo, población o ecosistema.
- **Grado de procesamiento:** primario, secundario o metadatos. El grado primario hace referencia a los datos originales, en bruto (por ejemplo, la secuencia genómica); mientras que los datos secundarios son los que ya han sido computacionalmente procesados o manualmente curados a partir de datos primarios (por ejemplo, una tabla de mutaciones asociadas a un genoma). Los **metadatos** son el conjunto de datos que nos describen los datos originales (por ejemplo, serían tanto las características del paciente del que hemos recibido el ADN para analizar su genoma, como las versiones de software utilizadas o el puntaje de fiabilidad de cada mutación detectada).
- **Grado de estructuración:** datos estructurados o no estructurados. Los datos estructurados siguen un modelo definido y habitualmente constituyen una base de datos. Sin embargo, los datos no estructurados no son fácilmente accesibles porque no siguen un esquema formal que nos permita analizarlos automáticamente. Son por ejemplo textos libres e imágenes.

- **Enfoque: small data vs. big data.** Los *small data* son datos cuyo enfoque es lograr una mejor descripción a nivel individual, predicción y control de una única unidad, que puede ser un individuo, un hospital, una comunidad o una ciudad. Son asignados como “pequeños” porque solo son utilizados para una unidad. El enfoque *big data* se refiere a la recopilación de datos en un grupo numeroso de individuos, para luego utilizarlos en otro grupo externo. En realidad, la mayor parte de los datos que etiquetamos habitualmente como proyectos *big data*, son en realidad proyectos *small data*.

Debemos destacar la **complejidad y heterogeneidad de las fuentes de datos**, tanto asociadas a su origen como a la manera de recopilarlos. Por ejemplo, si nos centramos en datos que nos permitan analizar genoma, fenoma y exposoma, estos pueden provenir de distintas fuentes como son:

- **Tecnologías ómicas.** Proporcionan datos genómicos y fenómicos a nivel molecular (proteómica, metabolómica). Sus datos nos indican procesos bioquímicos y reguladores.
- **Sistemas de imagen médica.** Ecografías, radiografías, resonancias nucleares, etc. Ayudan en el análisis del fenotipo.
- **Instrumentación médica.** Datos con la situación médica del paciente. Como ejemplo: respiradores, pulsioxímetros, etc.
- **Bases de datos científicas.** Estas pueden ser primarias, conteniendo datos crudos sin analizar y que pueden ser útiles para el reanálisis por parte de distintos grupos con distintas metodologías; o bases de datos secundarias o de datos derivados, donde el dato ya está procesado.
- **Otros orígenes de datos.** Estos pueden proceder de ensayos clínicos regulados, tecnología participativa (aplicaciones de salud digital, de moda en nuestros teléfonos inteligentes y relojes), redes sociales, directamente aportados por los pacientes en cuestionarios o a partir de la historia clínica digital.

Como características generales de los datos deberíamos seguir los **principios FAIR** (Wilkinson et al., 2016):

- **Findable o localizables:** permitir la localización de nuestros datos y metadatos de manera fácil mediante motores de búsqueda.
- **Accessible o accesibles:** datos y metadatos accesibles públicamente para otros investigadores.
- **Interoperable o interoperables:** describir datos y metadatos siguiendo reglas y estándares abiertos para facilitar su intercambio y reutilización, así como su integración junto con otros conjuntos de datos externos.
- **Reusable o reutilizables:** estos datos y metadatos pueden ser reutilizados por otros investigadores, al quedar clara su procedencia y condiciones de uso.

Estos principios FAIR buscan optimizar la reutilización de los datos y para ello, estos deben estar cuidadosamente descritos y organizados. Adicionalmente, estos datos son confidenciales, lo que lleva, inevitablemente a un reto en la anonimización y seguridad de los mismos.

2.3.3. Retos asociados

En esta breve introducción a las características de los datos quedan patentes algunos retos, los cuales es importante tener en cuenta cuando recopilemos y analicemos nuestros propios datos:

- Relacionados con el procesamiento de los datos:
 - No existen criterios claros de estandarización de los datos, que faciliten su intercambio e interoperabilidad.
 - Limitaciones en la búsqueda de datos de manera pública.
 - Ausencia de procesos que nos permitan establecer relaciones entre los datos.
 - Necesidad de evaluar y validar los resultados de estos proyectos.
 - Difícil aplicación del marco legal que permita compartir datos entre instituciones garantizando la seguridad y privacidad de los mismos.
 - Promover la publicación y accesibilidad de los datos de salud de manera pública y gratuita, para favorecer su reutilización.
- De formación, educación y difusión:
 - Necesidad de formar y capacitar al personal clínico, científico y de gestión en la toma, manejo, análisis e interpretación de este tipo de datos.
 - Necesidad de incorporación y formación de personal técnico dedicado al procesamiento digital de esta información.
 - Importancia de la recolección de los datos de una manera adecuada y de calidad.
 - Garantizar la difusión responsable y rigurosa, valorando los mensajes que se difunden en los medios de comunicación y redes sociales.
 - Visibilización del papel de la bioinformática aplicada a la práctica clínica.
- Retos organizativos:
 - Financiación estable y continua para garantizar la sostenibilidad de los datos y sus repositorios.
 - Soporte informático adecuado de las organizaciones para favorecer la interoperabilidad, asegurando la confidencialidad de estos datos.



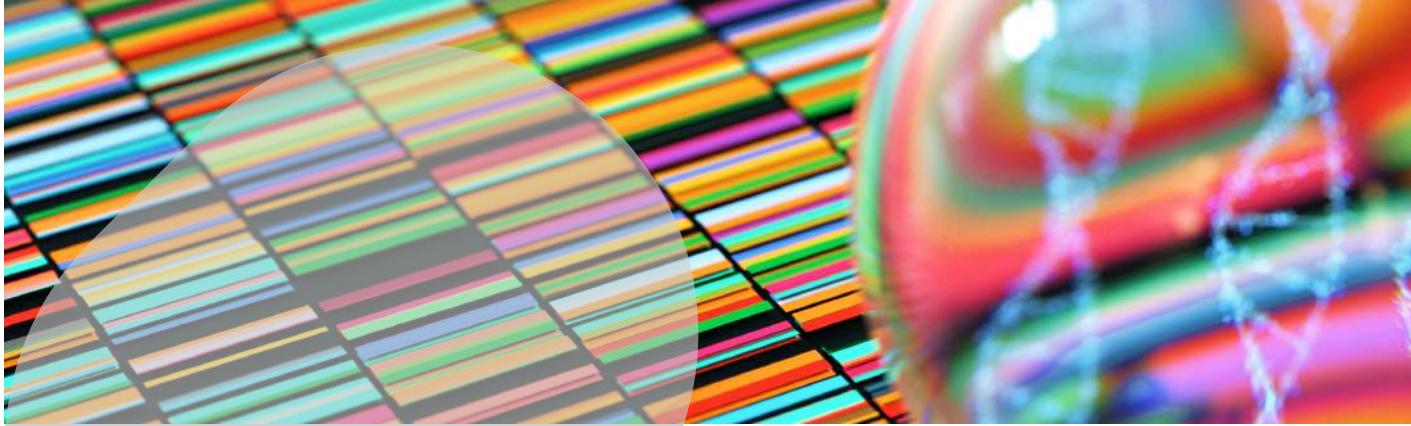
Enlaces de interés

En el siguiente enlace encontrarás un documento con un análisis detallado de las ciencias ómicas.

<https://www.institutoroche.es/observatorio/cienciasomicas>

En el siguiente enlace encontrarás un documento con un informe detallado sobre datos en medicina personalizada de precisión.

https://www.institutoroche.es/recursos/publicaciones/188/Informes_Anticipando_LOS_DATOS_EN_LAERA_DE_LA_MEDICINA_PERSONALIZADA_DE_PRECISION



Capítulo 3

¿Cómo analizamos un genoma eucariota? Paneles de captura de genes vs. genoma completo

Hasta el momento hemos visto las distintas maneras en las que puede variar un genoma, así como las implicaciones y utilidad que tiene su estudio. A continuación, vamos a profundizar en cuestiones más prácticas: ¿cómo podemos analizar un genoma eucariota, especialmente el genoma humano, con fines clínicos para el diagnóstico, pronóstico y tratamiento de enfermedades de base genética?

3.1. Introducción al análisis de genomas eucariotas

Desde que se completó el Proyecto Genoma Humano (HGP) en 2003, la era de la tecnología de secuenciación masiva (NGS) aplicada a su conocimiento despegó y comenzó a revolucionar la práctica médica y el diagnóstico clínico.



Este tipo de técnicas nos ayudan a solventar problemas de resolución que presentan las técnicas moleculares convencionales, como la secuenciación Sanger de genes candidatos, la hibridación comparativa por arrays o el **cariotipado**. Nos proporcionan el poder de detectar nuevas variantes e incrementar el diagnóstico de enfermedades raras y complejas de etiología desconocida.

Su aplicación no solo se ciñe a enfermedades raras o desconocidas, sino que incluye cáncer, enfermedades complejas o análisis de modificaciones en la expresión génica (Pérez & Tolosa, 2017; Petersen *et al.*, 2017; Sun *et al.*, 2015).

En los últimos años ya se ha visto que el coste de secuenciar un genoma, o parte de él, ha caído desde varios miles de euros a precios totalmente asequibles para cualquier organización médica (Wetterstrand, 2021). Por otro lado, también se ha democratizado la utilización de estas técnicas en distintos entornos médicos. Sin embargo, lo más complejo sigue siendo el análisis, anotación e interpretación de estos datos (Pérez & Tolosa, 2017; Seaby *et al.*, 2016).

El análisis del genoma humano se realiza por **resecuenciación**, es decir, secuenciamos el genoma o parte de él, y lo comparamos frente a un genoma que consideramos modelo. El genoma completo mide aproximadamente tres billones de pares de bases, pero solo un 1-2 % codifica para proteínas. Es lo que denominamos **exoma**. La mayor parte de las variaciones genómicas conocidas alteran la secuencia de la proteína. Conocemos aún muy poco sobre la función del ADN no-codificante, aunque el proyecto ENCODE nos acerca poco a poco a este conocimiento.

Dado que se estima que el 85 % de las mutaciones que causan patología residen en el exoma, esta técnica suele ser la elección preferente frente a la secuenciación del genoma completo. Es la opción intermedia entre un coste asequible, cobertura del genoma, rendimiento diagnóstico y utilidad para la interpretación (Seaby *et al.*, 2016).

3.2. Estrategias generales de análisis

De manera general existen **tres estrategias para abordar el diagnóstico molecular empleando técnicas de secuenciación masiva**. A lo largo de este capítulo desgranaremos sus ventajas e inconvenientes.

3.2.1. Paneles de genes o regiones de interés

Los paneles de genes o regiones de interés analizan un número limitado de genes que están asociados a un grupo de enfermedades determinadas o fenotipos.

Está indicado en caso de **enfermedades mendelianas bien descritas genéticamente**, para las que se conoce certamente los genes y mutaciones implicadas.

- **Ventajas:**
 - Menor coste económico, ya que son regiones bien delimitadas.
 - Rapidez.
 - **Gran cobertura de secuenciación**, lo que **permite la detección de variantes de baja frecuencia**.
 - Seleccionar *a priori* regiones de interés facilita el análisis posterior de datos.
- **Limitaciones:**
 - No es posible detectar mutaciones en genes o regiones que no han sido amplificadas o capturadas por el diseño inicial.



3.2.2. Secuenciación de exoma (WES)

La **secuenciación del exoma completo**, también llamado **WES** (*whole exome sequencing*) es un **panel donde se incluyen las regiones codificantes de todos los genes**, aproximadamente un **2 % del genoma completo**. Existen variaciones sobre este panel para reducirlo a zonas de mayor interés clínico, más focalizado, siendo un intermedio entre el panel de genes y el exoma completo.

Esta es la **metodología elegida cuando un panel génico no puede identificar la causa de la enfermedad**, ya que permite detectar mutaciones en nuevos genes implicados en la misma. También está indicado en casos cuyo fenotipo implica varios genes o no se conoce una asociación clara con alguna enfermedad previamente descrita.

- **Ventajas:**
 - Es más rápido y barato, tanto en preparación como en análisis, que el estudio del genoma completo.
- **Limitaciones:**
 - **La cobertura no es uniforme.**
 - No incluye regiones intrónicas ni reguladoras.
 - **En muchos casos no detecta variantes estructurales.**

3.2.3. Secuenciación de genoma completo (WGS)

La secuenciación del genoma completo, llamada WGS (*whole genome sequencing*), cubre teóricamente la información completa codificada en el ADN del individuo, incluyendo regiones intrónicas y reguladoras.

- **Ventajas:**
 - **Permite detectar variaciones estructurales complejas**, como cambios en el número de copias y translocaciones cromosómicas.
 - **Permite detectar mutaciones en genes previamente no asociados a enfermedad**, e incluso en regiones intrónicas o reguladoras.
- **Limitaciones:**
 - Coste de secuenciación. Pese a que el coste de secuenciar un genoma completo humano ha descendido notablemente en los últimos años, actualmente se tasa en aproximadamente 1000 euros por genoma.
 - **Complejidad en el análisis de los datos**, tanto a nivel computacional (necesidad de gran número de recursos) como a nivel de interpretación de los mismos.

3.3. Protocolo de análisis

3.3.1. Extracción del ADN

El primer paso en cualquiera de los tres protocolos consiste en la obtención de un ADN de alta calidad a partir de muestras biológicas. Habitualmente se obtiene de leucocitos de sangre periférica, aunque también son comunes muestras como saliva o incluso tejidos parafinados (*formamin-fixed paraffin-embedded*, FFPE). La saliva presenta problemas de posible contaminación con ADN de la microbiota oral y las muestras de FFPE suelen obtenerse de tejidos cancerosos, dando un ADN de peor calidad debido a alta degradación (Seaby et al., 2016). Este ADN se cuantifica utilizando fluorimetría y se cualifica utilizando electroforesis capilar.

3.3.2. Preparación de la genoteca/librería. Genotecas con captura o amplificación de regiones

En el caso de los paneles de genes y WES la preparación de la genoteca sigue metodologías similares: amplificación mediante PCR de las regiones de interés (sobre las que previamente hemos realizado un diseño con cebadores específicos) o bien, captura mediante sondas de estas regiones. En realidad, podemos pensar en que un análisis WES no es más que un panel de genes muy ampliado, con un mayor número de regiones que capturar y secuenciar. Sin embargo, en el proceso de preparación de un genoma completo no existe una captura de regiones, sino que simplemente se secuencia todo el ADN que tenemos en la muestra.

Estas son las genotecas de paneles o exoma WES. Existen dos metodologías principales de construirlas:

- Secuenciación por amplicones. Se realiza un diseño de cebadores específicos sobre los genes de interés y se amplifican mediante PCR. Al pool amplificado se adicionan los adaptadores e índices, previamente a la secuenciación. Es un proceso rápido y sencillo, permite partir de menor cantidad de ADN, además de tener un menor coste por muestra que otros procedimientos; sin embargo, está limitado a un máximo de unos 10 000 amplicones e inevitablemente el proceso de amplificación por PCR puede producir errores y sesgos. Está indicado para el genotipado mediante secuenciación, detección de variantes ya conocidas a enfermedad y detección de SNP e indels en células germinales.
- Captura por hibridación. Este tipo de captura se realiza mediante sondas de ARN que hibridan con el ADN de interés, previamente fragmentado bien enzimáticamente o por métodos de sonicación focalizada (método mecánico). Tras la captura, se liberan los fragmentos capturados, se adicionan los adaptadores e índices y se secuencia la genoteca. En este procedimiento, dependiendo de la concentración del ADN de partida, la amplificación por PCR para enriquecer los fragmentos capturados es opcional. Aunque es un proceso más largo y laborioso, así como el coste puede ser más elevado que con amplicones, su sensibilidad es mayor que la técnica anterior. Está indicado para genotipado, secuenciación de exoma, análisis de mutaciones oncológicas, descubrimiento de variantes raras, descubrimiento de genes, detección de variaciones somáticas de baja frecuencia y análisis de número de copias de un gen (*copy number variants*, CNV).

Aunque existen muchas soluciones comerciales en el mercado, en la tabla siguiente, adaptada del artículo de Seaby et al., (2016), se detallan algunas características de los kit de captura mediante sondas de exoma disponibles. Como se puede ver en la tabla, las principales diferencias entre los kits disponibles radican en dos cuestiones principales. La primera es el tamaño de las regiones capturadas.

Los diseños en kits de captura de exoma recogen los genes más relevantes en patologías clínicas de interés, pero adicionalmente pueden extenderse a otras regiones, de ahí sus principales diferencias. También, la captura de lugares promotores o UTR, sitios de *splicing* o variantes intrónicas. En cuanto a la preparación de la genoteca, el método de fragmentación determina la cantidad de ADN inicial del que partir. Los métodos mecánicos producen genotecas con una dispersión de tamaños menor, pero necesitan unas cantidades de ADN inicial mayores y, por supuesto, el equipamiento de sonicación mecánica adecuado, equipamiento de alto coste no disponible en todos los laboratorios. Por otra parte, los métodos enzimáticos son más rápidos y menos laboriosos, así como más baratos y parten de cantidades de ADN menores, pero el tamaño medio de la genoteca resultante es más disperso.

Tabla 5*Kits de captura de exoma: comparativa*

	xGen Exome Research Panel v2 (IDT)	SureSelect Focused Exome (Agilent)	SureSelect Human All Exon V8 (Agilent) ¹	SureSelect Clinical Research Exome V2 (Agilent)	Illumina DNA Prep with Enrichment Exome ⁴ (Illumina)	Illumina DNA Prep with Enrichment Exome ⁵ (Illumina)	TruSeq DNA Exome (Illumina) ³
Tamaño regiones (target size)	34 Mb	12 Mb	35.1 Mb	67.3 Mb	12 Mb	16.5 Mb	45 Mb
Número de sondas	415 115	231 855	436 193	843 912	125 395	183 809	429 826
Número de genes / regiones capturadas	19 433	4800	> 20 0002	1009 genes > 800 promotores 75 K sitios splicing no codificantes > 12 000 variantes intrónicas	4813	6704	19 396
Lecturas sobre la diana (reads on target)	> 95 %	> 83 %	> 60 %	> 97 %	> 85 %	> 85 %	> 85 %
% de pb cubiertas ≥ 20x	> 90 %	> 97 %	> 96 %	> 95 %	> 93 %	> 93 %	> 90 %
Método fragmentación	Enzimático	Enzimático/ mecánico	Enzimático/ mecánico	Enzimático/ mecánico	Enzimático	Enzimático	Mecánico
ADN inicial	100 ng	50 ng / 200 ng – 3 ug o 100 ng – 1 ug (QXT, XT, XT2) ³	50 ng / 200 ng – 3 ug o 100 ng – 1 ug (QXT, XT, XT2) ³	50 ng / 200 ng – 3 ug o 100 ng – 1 ug (QXT, XT, XT2) ³	50 ng	50 ng	100 ng

Nota. Adaptado de "Exome sequencing explained: A practical guide to its clinical application", por E. G. Seaby, R. J. Pengelly y S. Ennis, 2016, *Briefings in Functional Genomics*, 15(5), pp. 374–384. Recuperado de <https://doi.org/10.1093/bfgp/elv054>. Datos proporcionados por las casas comerciales.

Todos los diseños se han comparado según el genoma *Homo sapiens* Hg19.

¹ Existen versiones del kit extendidas, donde se cubren también regiones UTR.

² No se encuentra especificado por la casa comercial. Datos calculados por el autor del manual.

³ QXT: método fragmentación enzimática; XT: método fragmentación mecánico y captura en pool individual de muestras; XT2: método fragmentación mecánico y captura en pool de muestras conjunta.

⁴ Conocido anteriormente como Nextera Flex for Enrichment. Sondas TruSight One.

⁵ Conocido anteriormente como Nextera Flex for Enrichment. Sondas TruSight One Expanded.



Ejemplo

En la Figura 6 se observa la representación de las sondas de algunos de estos kits de captura de exoma sobre el genoma de referencia *Homo sapiens* Hg19 en el programa IGV.

- Descarga de archivo**
En *Campus virtual > Aula de la asignatura > Recursos y materiales*, podrás encontrar los archivos de extensión BED correspondientes a las sondas de estos kits de captura de exoma que vamos a exemplificar (archivos: [Tema3_ejemplo.zip](#)).

En el panel de la Figura 6A se ha tomado una región del cromosoma 2 (posiciones 198 226 792 a 198 444 348), donde ya podemos ver que no todos los diseños de panel cubren de la misma manera todos los genes de esta zona.

El gen *COQ10B*, codificador de la proteína coenzima Q10B, no está cubierto por los paneles Sure Select Focused Exome (Agilent) ni por TruSight One (Illumina); mientras que el panel Sure Select Clinical Research Exome (CRE) v2 cubre adicionalmente una región más extendida de su zona 3'-UTR.

Una situación similar ocurre con el gen *HSPD1*, presente también en este cromosoma. Aunque todos los paneles presentan sondas frente a él, no todos lo cubren de igual manera. Este gen, codificador de una chaperonina, está involucrado en leucodistrofias y paraplejias.

Por lo tanto, si es uno de los genes que está dentro del enfoque de nuestro estudio debemos ser cautos en analizar qué sondas debemos adquirir para cubrirlo correctamente.

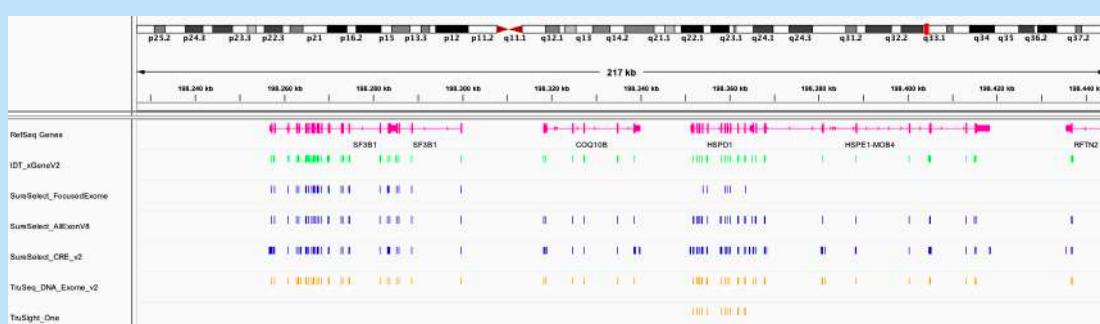
Esta situación es análoga a la que acontece en la región visualizada en la Figura 6B, con el gen *MORC2*, involucrado en un subtipo de la enfermedad de Charcot-Marie-Tooth, denominada axonal. Los genes *JPH1* y *GDAP1*, localizados en la región aproximada chr8:75,127,023-75,303,851; también se correlacionan con esta neuropatía y deben ser analizados conjuntamente con el gen previamente mencionado. Sin embargo, como vemos en la Figura 6C, solo *GDAP1* está cubierto en todos los kits mostrados.

¿Creéis que todos estos kits os permitirían el diagnóstico certero de esta neuropatía de igual manera? Si sospecháis de ella, ¿utilizaríais cualquiera de ellos?

Figura 6

Comparativa de las regiones capturadas en los distintos paneles en distintas zonas del genoma

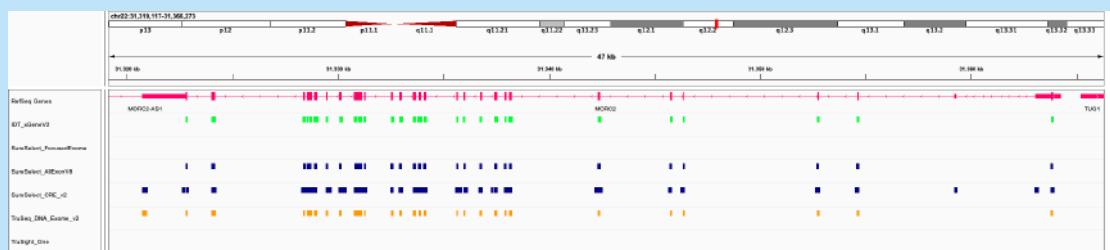
(A) La región del genoma Hg19 representada es chr2:198,226,792-198,444,348.



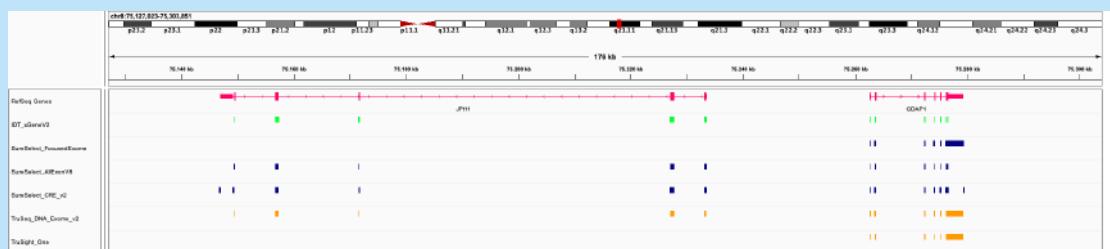
>>>

>>>

(B) Comparativa de sondas que cubren el gen MORC2, involucrado en la enfermedad de Charcot-Marie-Tooth.



(C) Comparativa de sondas que cubren la región chr8:75,127,023-75,303,851, involucrado en la enfermedad de Charcot-Marie-Tooth.



Nota. Imágenes de elaboración propia con el software de distribución libre Integrative Genomics Viewer (IGV) versión v2.11.2.



Enlace de interés

Adicionalmente, en este enlace, se puede visualizar el genoma humano en sus distintas versiones y permite analizar las sondas de los kits comerciales más frecuentemente utilizados.

https://genome-euro.ucsc.edu/cgi-bin/hgTrackUi?hgsid=277312691_qnaKZAVtoze5TjXzl2ZxP3t2eqrw&db=hg38&c=chr1&g=exomeProbesets

De manera análoga a estos paneles de captura de exoma, los paneles de captura de genes aislados siguen la misma metodología. Las casas comerciales principales permiten en sus web diseñar sondas para los genes seleccionados, simplemente incluyendo el identificador de los genes de interés.

3.3.3. Secuenciación de la genoteca/librería

Una vez obtenida la genoteca, se procede a la secuenciación de esta. La elección del equipo de secuenciación depende fundamentalmente de la cobertura que queremos conseguir para el genoma analizado. Los millones de lecturas, o gigabases, secuenciados.



La cobertura de secuenciación (coverage, en inglés), como ya indicamos en capítulos anteriores, es el número de veces que cada base nucleotídica está secuenciada, es decir, el número de lecturas que apoyan dicha base. La cobertura requerida varía en función de la aplicación, teniendo en cuenta que, a mayor cobertura, mayor confianza en la base nucleotídica secuenciada, lo que nos permitirá realizar la determinación de mutaciones con mayor precisión. Sin embargo, debemos tener en cuenta que las lecturas no se distribuyen uniformemente sobre el genoma, sino que lo hacen de manera aleatoria e independiente. Por lo tanto, muchas bases estarán cubiertas por menos lecturas que la cobertura promedio, mientras que otras bases estarán cubiertas por más lecturas que el promedio.

No existen una regla única y estricta para la cobertura requerida, ya que depende del tipo de estudio, tamaño del genoma de referencia y, en el caso de experimentos de expresión génica, el nivel de esta expresión. Sin embargo, actualmente, para experimentos de genómica humana se recomienda para genoma humano completo (WGS) una cobertura recomendada entre 30 y 50x; mientras que para exoma completo (WES) o paneles de genes se recomienda una cobertura 100x. Estos estándares suelen fijarse por las sociedades de genética, basándose en los resultados publicados en las diversas publicaciones científicas.

La ecuación que calcula la cobertura media estimada es la ecuación de Lander-Waterman, expresada como:

$$C = LN / G$$

Donde C es la cobertura, G es el tamaño del genoma haploide en megabases (Mb), L es la longitud de la lectura (si es pareada, se tiene en cuenta la suma de las dos lecturas) y N es el número de lecturas en millones.



Ejemplo

Veamos qué cobertura obtendríamos para la secuenciación en los siguientes casos:

Genoma completo (WGS) de humano (3000 Mb) en un formato 2 × 100 pb, con un total de 60 M de lecturas para cada muestra:

$$C = LN / G = (2 \times 100) \times (60 \times 10^6) / (3 \times 10^9) = 4x$$

Genoma completo (WGS) de humano (3000 Mb) en un formato 2 × 150 pb con un total de 60 M de lecturas por muestra:

$$C = LN / G = (2 \times 150) \times (60 \times 10^6) / (3 \times 10^9) = 6x$$

Genoma completo (WGS) de *Caenorhabditis elegans* (100 Mb) en formato 2 × 100 pb, con un total de 180 M de lecturas por muestra:

$$C = LN / G = (2 \times 100) \times (180 \times 10^6) / (100 \times 10^9) = 360x$$

¿Cuántos millones de lecturas necesitaríamos para llegar a una cobertura 50x en un genoma completo humano? (Nota: lecturas pareadas 150 pb).

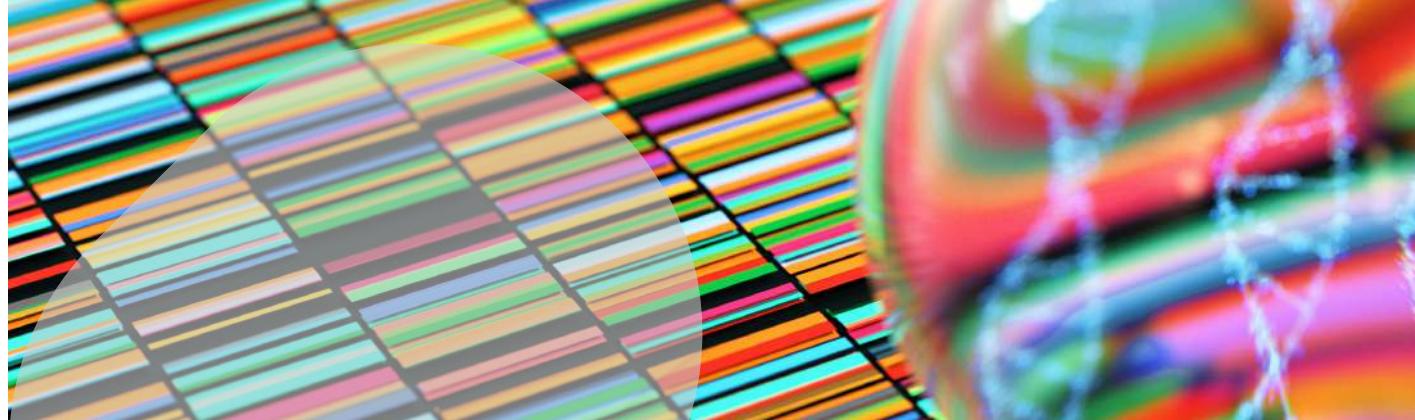
$$C = \frac{LN}{G} \rightarrow N = \frac{CG}{L} = 50 \times (3 \times 10^9) / (2 \times 150) = 500 \times 10^6 \text{ (500 M de lecturas)}$$



Enlace de interés

Para el uso de equipos Illumina, en el siguiente enlace se encuentran unas directrices generales, así como una calculadora de coberturas para distintas aplicaciones.

<https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html>



Capítulo 4

¿Cómo analizamos un genoma eucariota? Análisis bioinformático de paneles de captura de genoma humano

Actualmente el principal cuello de botella que encontramos en proyectos de secuenciación no se encuentra en el propio proceso de secuenciación, ya que como sabemos el coste por base secuenciada sigue bajando, gracias al desarrollo de equipos cada vez más potentes y rápidos.

El cuello de botella es el análisis bioinformático, que resulta aún un proceso complejo que debe automatizarse para obtener el mejor rendimiento en tiempo, pero también adaptarse en función del objetivo de cada proyecto.

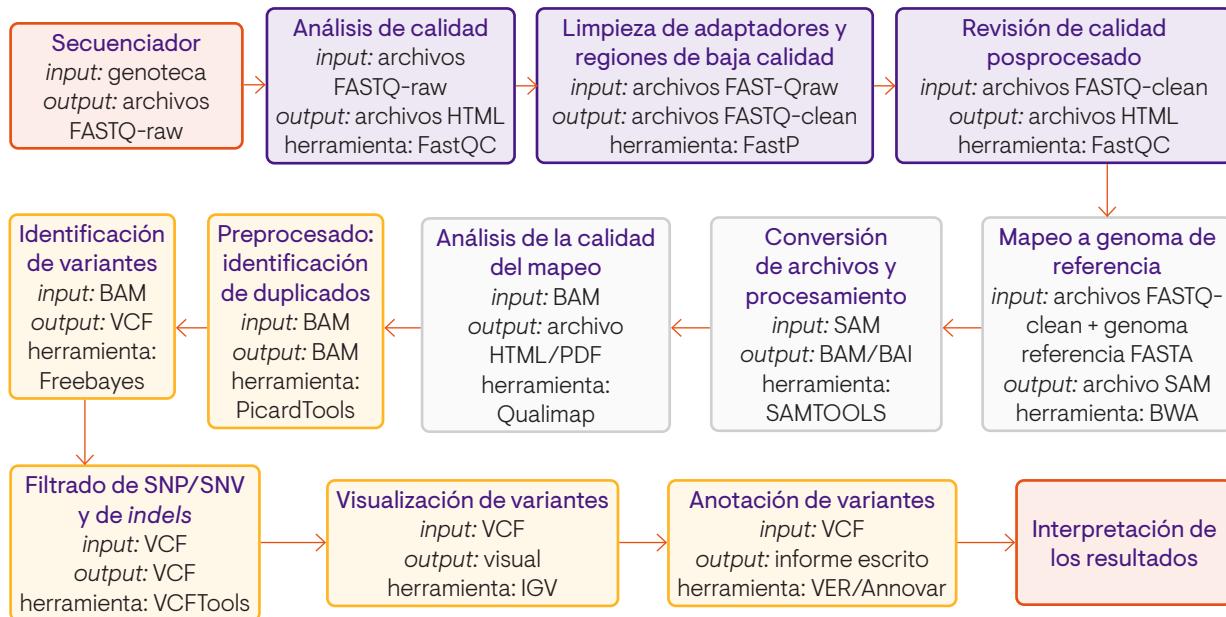
Este análisis está separado en etapas bien diferenciadas, donde cada una de ellas deben diseñarse al detalle, utilizando herramientas y bases de datos específicas. Las etapas variarán en función del tipo de experimento de secuenciación realizado y dependerán de la respuesta que se vaya a responder.

En este capítulo, vamos a ver algunas etapas generales en el análisis de un genoma eucariota, con la finalidad de encontrar variantes de interés para el diagnóstico clínico.

Estas etapas se pueden ver esquematizadas en la Figura 7.

Figura 7

Etapas generales en el análisis simplificado de un panel de genes en un genoma de humano con fines de diagnóstico clínico



Nota. Se detalla cada proceso, con el *input* y *output* de cada proceso, así como la herramienta informática utilizada.

4.1. Análisis primario. Calidad y filtrado de secuencias

El primer paso que cualquiera de nuestros análisis debe contener, independientemente del protocolo de preparación de genoteca u objetivo que queramos alcanzar, es el análisis de calidad de las lecturas obtenidas del secuenciador. La idea es analizar la calidad de las muestras secuenciadas, eliminar o filtrar las de baja calidad (o parte de ellas) y quedarnos con aquellas de alta calidad para las etapas siguientes de análisis.

Ya habéis visto en asignaturas anteriores cómo es el formato de lecturas FASTQ que obtenemos a partir de un secuenciador Illumina. Recordad que cada lectura tiene una cabecera, una línea de secuencia, una tercera línea (en blanco o con la cabecera repetida) y una última línea con caracteres de calidad asociados a cada base secuenciada. Cada uno de estos caracteres de calidad determina la probabilidad de que esa base nucleotídica leída sea errónea. Estos son los índices de calidad que queremos evaluar en los pasos siguientes. Si necesitáis repasar este formato, en el artículo de Cock *et al.* (2010) se definió detalladamente.



Enlace de interés

En el siguiente enlace, encontrarás un vídeo con la explicación de los formatos FASTA y FASTQ detallados.

https://www.youtube.com/watch?v=cJm_BGpjnWg

4.1.1. ¿Por qué es importante evaluar la calidad de las lecturas y filtrarlas?

Como ya hemos dicho, evaluar esta calidad es el primer paso de todo análisis bioinformático y debe llevarse a cabo de manera obligatoria, independientemente del objetivo del estudio y del tipo de muestra secuenciado.

Los puntos que se deben tener en cuenta son:

- Debemos **evaluar la calidad de la carrera de secuenciación** y preguntarnos si hemos obtenido el rendimiento de datos esperado, mayor o menor, así como evaluar si la concentración de las muestras secuenciadas ha sido adecuada. Este es un trabajo conjunto con el departamento de genómica que haya preparado las genotecas y secuenciado las muestras.
- Evaluar la calidad de las lecturas nos permitirá **valorar cómo de fiables serán nuestros resultados o conclusiones del experimento**, incluyendo evaluar la extracción del material genético y la preparación de la librería.
- Determinar si es necesario un **preprocesamiento o filtrado de las secuencias**, que de manera habitual realizaremos rutinariamente, con la finalidad de eliminar secuencias de baja calidad, regiones de baja calidad y adaptadores.
- Debemos **tener en cuenta qué secuenciador se ha utilizado**, evaluando si nuestro secuenciador es de cuatro canales de color (MiniSeq, MiSeq, HiSeq) o de dos canales de color (NextSeq, NovaSeq). En este último caso, la peculiaridad de regiones polyG debe tenerse en cuenta para el procesamiento bioinformático de las muestras.
- Debemos **mantener un compromiso entre la cantidad de datos secuenciados y la calidad de los mismos**. Nuestro nivel de exigencia en el filtrado nos hará perder cantidad de datos, pero aumentar su calidad.

4.1.2. Herramientas de control de calidad

Existen varias herramientas (Chen *et al.*, 2014) que realizan el análisis de control de calidad, e incluso la limpieza, en la literatura, como son Fastx-toolkit (FASTX-Toolkit, s. f.), NGS QC Toolkit (Patel & Jain, 2012), FastqCleaner (Roser *et al.*, 2019), fastqcr (GitHub - kassambara/fastqcr: fastqcr: Quality Control of Sequencing Data, s. f.) o el más conocido, **FastQC** (Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data, s. f.).

Adicionalmente, la herramienta **MultiQC** (Ewels *et al.*, 2016) permite compilar distintos tipos de archivos, como son los archivos de salida de FastQC, pero también archivos de mapeo BAM o de limpieza de lecturas, para proveer informes interactivos a modo de resumen, agrupando todas las muestras en una única visualización.

4.1.3. Análisis de calidad con FastQC

Esta aplicación, bien en su versión de línea de comandos o bien en su versión gráfica, nos ofrece las siguientes métricas para evaluar la calidad de las secuencias crudas FASTQ:

- Estadísticas básicas del conjunto de secuencias.
- Calidad de la secuencia por base.
- Valores de calidad por secuencia.
- Contenido de la secuencia por base.
- Contenido en GC por secuencia.
- Contenido de N por base.
- Distribución de la longitud de las secuencias.
- Secuencias duplicadas.
- Secuencias sobrerepresentadas.
- Contenido de adaptadores.

Enlace de interés

En el siguiente enlace a la web de FastQC podéis encontrar ejemplos explicados sobre la calidad en distintos set de datos, en el apartado *Example Reports*.

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>



Ejemplo

Para ilustrar este análisis vamos a realizar un ejemplo con unas lecturas reales.



Descarga de archivo

En Campus virtual > Aula de la asignatura > Recursos y materiales > Ejemplos en clase > Tema4_Ejemplo1y2_FastQC_FastP, podrás encontrar los FASTQ correspondientes a este ejemplo.

1. Abrimos la máquina virtual WorkSpaces.
2. Activamos el entorno de trabajo: `<conda activate 05MBIN_human>`
3. Ejecutamos el programa FastQC en línea de comandos `<fastqc *.fastq.gz>`, que mostrará en pantalla su evolución.
4. Abrimos los archivos de salida del programa, archivos HTML, en un navegador.

Estadísticas básicas del conjunto de lecturas (*basic statistics*) (Figura 8)

En este apartado se genera una tabla con información acerca del fichero analizado, que incluye:

- Nombre del fichero analizado.
- Tipo de fichero.
- Codificación utilizada en los valores de calidad.
- Número total de secuencias contenidas.
- Secuencias etiquetadas como baja calidad (si se ha indicado que lo haga).
- Longitud de las secuencias.
- Porcentaje de contenido GC medio.

Figura 8
Estadísticas básicas generadas por FastQC

Measure	Value
Filename	SRR15170798_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	290645
Sequences flagged as poor quality	0
Sequence length	20-301
%GC	54

Calidad de las secuencias por base (*per base sequence quality*) (Figura 9)

En este apartado se observa el rango de valores de calidad a lo largo de todas las bases secuenciadas teniendo en cuenta todas las lecturas del archivo FASTQ.

En el eje X se representa la posición de la base, mientras que en el eje Y se representan los valores de calidad (Phred score). Cuanto mayor es este valor, mayor precisión en la lectura de esta base o menor probabilidad de que sea incorrecta. En este eje se representan tres regiones: verde (Phred 28-38, calidad buena), naranja (Phred 20-28, calidad razonable) y rojo (Phred 0-20, mala calidad).

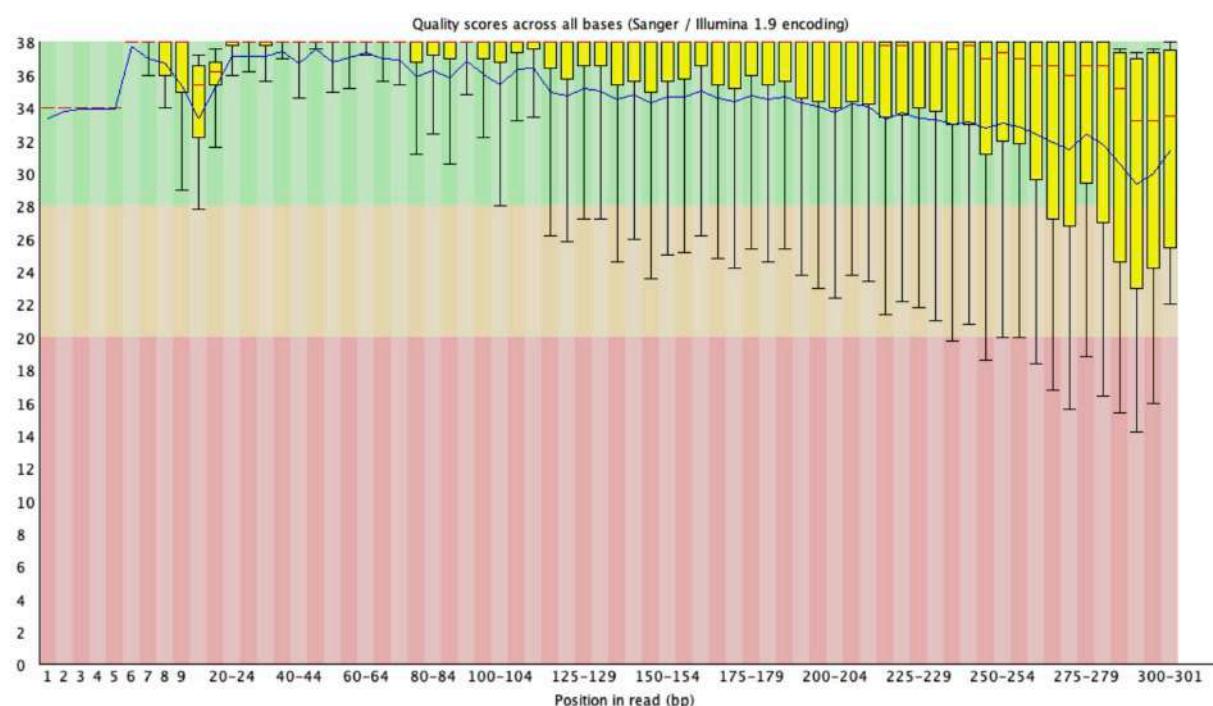


Lo normal en este tipo de lecturas es que la calidad se vaya degradando a medida que nos acercamos al final de la secuenciación. Por tanto, lo normal es que los valores vayan cayendo a regiones naranjas. Este efecto es más acusado cuanto más largas son las lecturas. Este hecho se debe a la degradación de la reacción química de fluorescencia a medida que avanzamos en el proceso cíclico de secuenciación.

Las cajas (boxplot) representadas en amarillo tienen en cuenta todas las secuencias en esa posición de la lectura determinada y tienen una línea central roja correspondiente a la mediana de calidad de todos los valores de calidad, la caja amarilla que es el rango intercuartil (25-75 %), los bigotes superior e inferior (10 y 90 % de calidad, respectivamente) y la línea azul que es el valor medio del índice de calidad.

Figura 9

Calidad de las secuencias por base para el conjunto de secuencias analizado



Valores de calidad por secuencia (*per sequence quality score*) (Figura 10)

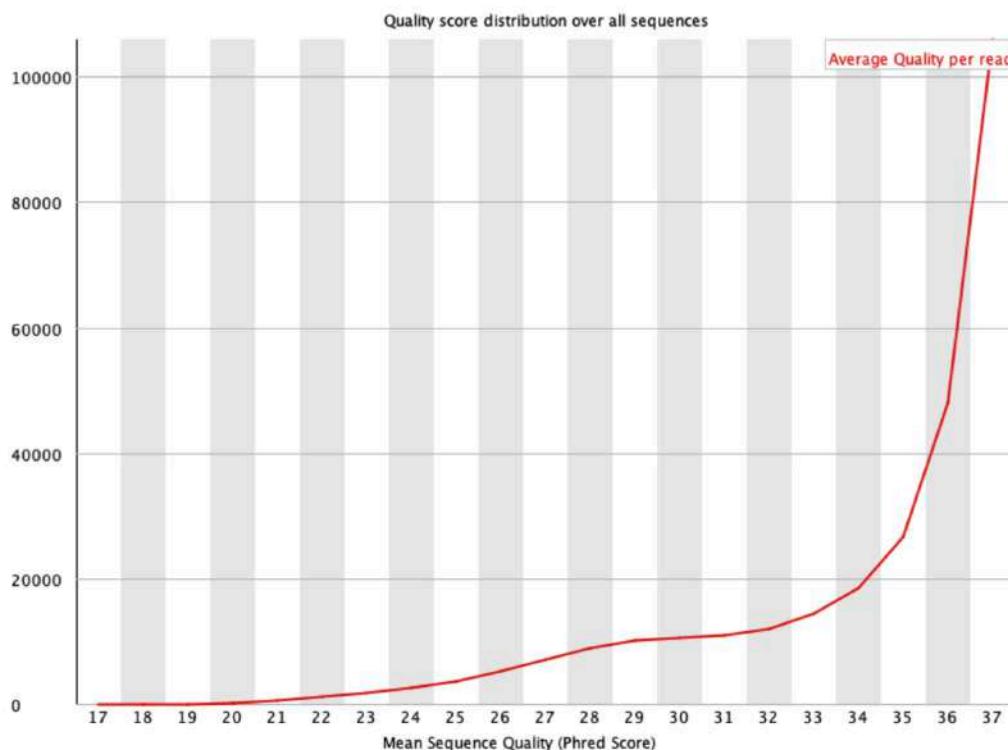
En este apartado se calcula el valor medio de calidad para cada una de las secuencias del archivo FASTQ y se dibujan los valores en una distribución, de manera que podemos analizar si un subconjunto de secuencias tiene un valor medio de calidad bajo en comparación con el resto. En el eje Y se representa la frecuencia de lecturas que tienen un valor de calidad medio dado por el eje X.

En este ejemplo, la mayor parte de nuestras secuencias tienen alto valor de calidad (por encima de 36); sin embargo, hay un subconjunto de secuencias que tienen valores medios de entre 26 y 32. Debemos considerar eliminarlas de nuestro conjunto de datos.

Si una proporción de secuencias tiene un valor de calidad medio-bajo nos puede indicar algún tipo de problema sistemático en una parte de la carrera de secuenciación como, por ejemplo, una celda defecuosa. Veremos este término en un apartado posterior.

Figura 10

Valores de calidad por secuencia

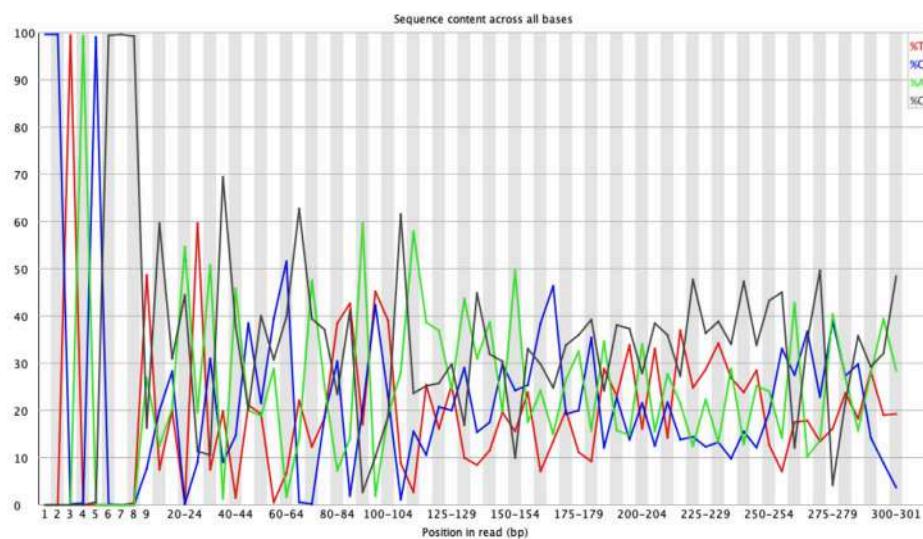


Contenido de secuencia por base (*per base sequence content*) (Figura 11)

En esta gráfica vemos la cantidad relativa de cada nucleótido (A, C, G, T) en porcentaje a lo largo de todas las bases de las secuencias del fichero FASTQ: en una genoteca aleatoria esperaríamos que todas estas bases estuviesen compensadas, de forma que las líneas del gráfico deberían ser paralelas a lo largo de la longitud de la lectura; sin embargo, esto puede verse alterado al inicio de la secuencia si hablamos de experimentos de secuenciación ARN (por el uso de hexámeros en la genoteca); o bien porque sean genotécas de amplicones por ejemplo, de metataxonomía, donde el fragmento es prácticamente idéntico en todas las lecturas.

Figura 11

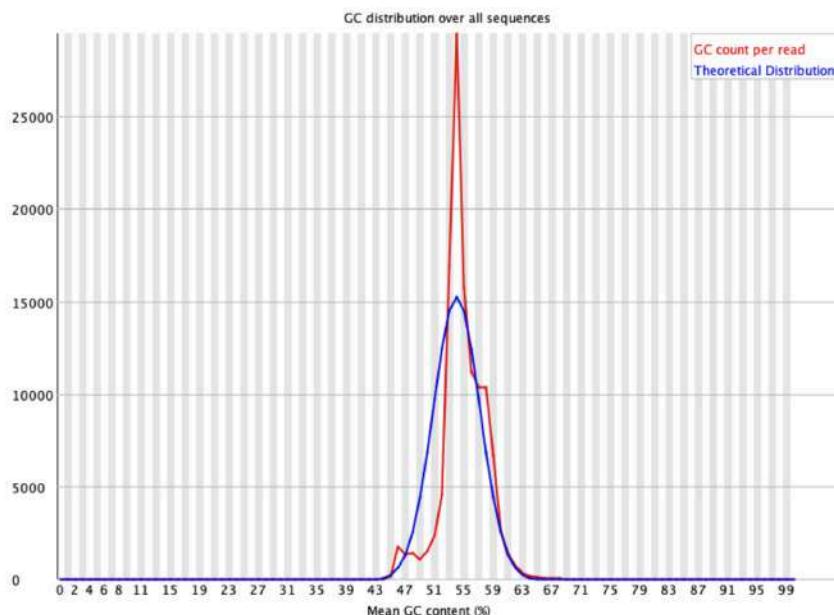
Contenido de secuencias por base en el ejemplo analizado

Contenido GC por secuencia (*per sequence GC content*) (Figura 12)

Este apartado mide el contenido GC medio de cada secuencia contenido en el fichero FASTQ y lo muestra en forma de una distribución, que además se compara con una distribución normal-gaussiana modelada. En una genoteca aleatoria es esperable una distribución normal de este contenido, donde el pico central es el contenido GC medio del genoma del que se han obtenido las lecturas. Una distribución inusual, como la que vemos en nuestro ejemplo, puede indicarnos algún tipo de sesgo o contaminación. Si vemos dos picos, podría tratarse de una contaminación de dos genomas diferentes. Los contaminantes específicos —como pueden ser adaptadores de dímeros, que producen picos puntiagudos en la distribución— pueden evaluarse en el apartado de secuencias sobrerepresentadas.

Figura 12

Contenido GC en el ejemplo analizado



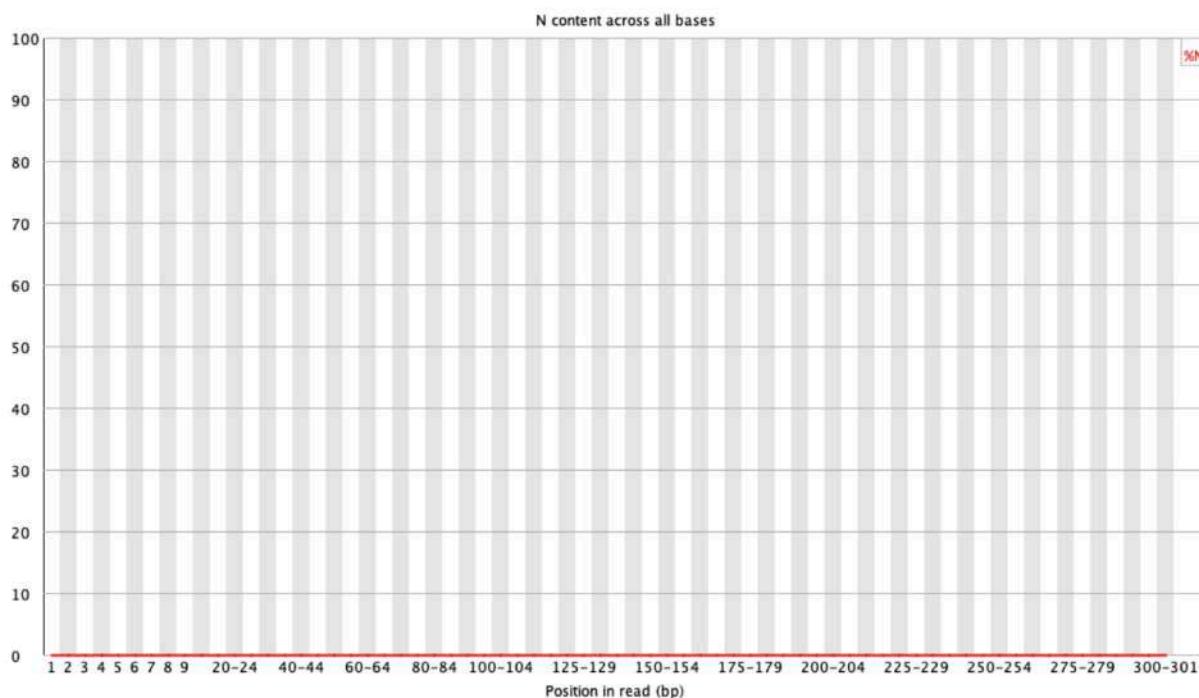


Contenido de N por base (*per base N content*) (Figura 13)

A lo largo del proceso de secuenciación puede ocurrir que alguna de las bases incorporadas no tenga la resolución suficiente para determinar cuál es exactamente. En este caso, el secuenciador asigna un N en esa posición. En este gráfico se muestra el porcentaje de N en cada posición de la lectura. Es habitual detectarlas en baja proporción al final de las lecturas, y pueden ser eliminadas en el proceso de limpieza posterior. En nuestro ejemplo no se observan contenidos en N.

Figura 13

Contenido en secuencias desconocidas (N) por base

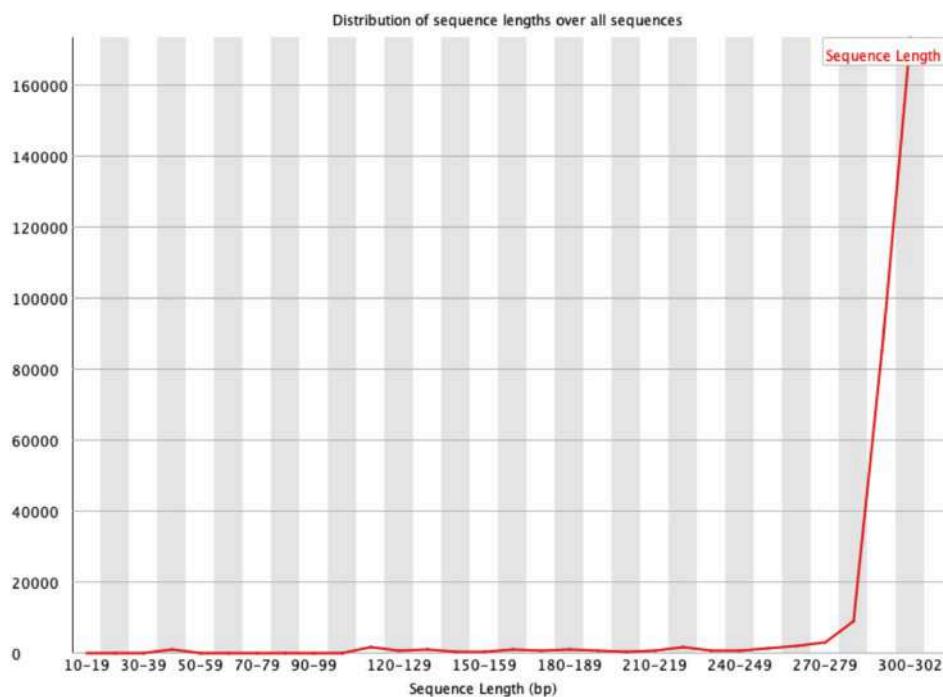


Distribución de la longitud de las secuencias (*sequence length distribution*) (Figura 14)

La gráfica muestra la distribución de la longitud de las secuencias que componen el archivo FASTQ. Habitualmente, si no se ha realizado preprocesamiento de las mismas, todas las lecturas de un secuenciador Illumina tienen la misma longitud. En nuestro ejemplo, que ha sido obtenido de una base de datos pública, las secuencias tienen una distribución de tamaño de entre 10 pb y 300 pb, siendo este tamaño el mayoritario. Esto es lógico, puesto que se trata de secuencias de metataxonomía, secuenciadas en formato pareado de 300 pb. Podemos limpiar las secuencias de bajo tamaño en pasos posteriores.

Figura 14

Distribución de la longitud de las secuencias analizadas



Secuencias duplicadas (*sequence duplication levels*) (Figura 15)

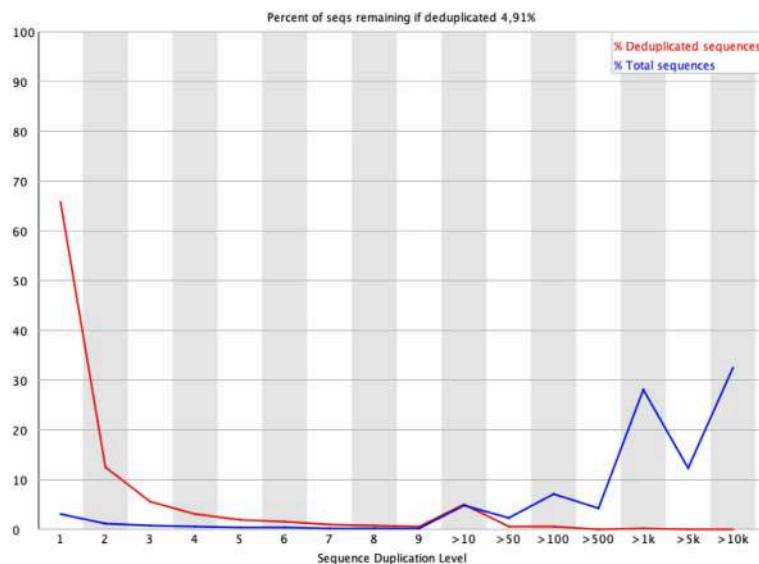
En una genoteca diversa, obtenida a partir de un genoma completo, lo que deberíamos esperar es que la mayor parte de las secuencias aparezcan una sola vez, lo que se observa en un nivel bajo de duplicación. Por otro lado, un nivel alto de duplicación puede deberse a un sesgo de enriquecimiento de la genoteca; sin embargo, puede tener su sentido si hablamos de datos de expresión génica o, como en este caso de un amplicón para análisis metataxonómico.

En la gráfica se observan dos líneas: azul y roja. La línea azul muestra los niveles de duplicidad totales, del conjunto total de datos; mientras que la línea roja se calcula tras eliminar las secuencias duplicadas. Las proporciones mostradas por la línea roja son del conjunto de datos sin duplicidades. El porcentaje superior nos indica la proporción del conjunto de datos original que obtendríamos si conserváramos una sola copia de cada secuencia del conjunto de datos. En nuestro caso, tenemos una alta proporción de secuencias duplicadas, y nuestro conjunto de datos quedaría reducido al 4.91 % si eliminásemos duplicados.

Las duplicaciones pueden surgir de dos fuentes y es importante diferenciarlas para estimar si es un problema de la genoteca o bien esta duplicidad tiene sentido biológico:

- Duplicaciones técnicas surgidas por problemas de PCR durante el enriquecimiento de la genoteca.
- Duplicación biológica, propia de la naturaleza de nuestra genoteca, por el tipo de experimento que estemos analizando, como puede ser un experimento de transcriptómica o el análisis de un amplicón determinado.

Figura 15
Niveles de duplicación en las secuencias



Secuencias sobrerepresentadas (*overrepresented sequences*) (Figura 16)

Una genoteca normal debería ser aleatoria y contener un conjunto diverso de secuencias. Este apartado muestra todas las secuencias que suponen más del 0.1 % de las lecturas totales. Asimismo, el programa buscará en su base de datos de contaminantes comunes, entre los que se encuentran los adaptadores de secuenciación, si existen coincidencias.

Si observamos alguna secuencia sobrerepresentada, esto puede significar que esta secuencia es biológicamente significativa (por ejemplo, es un transcripto altamente expresado, si es una genoteca de ARN), que la librería está contaminada (por ejemplo, con adaptadores o cebadores de amplificación), o que la genoteca no sea tan diversa como se esperaba. En el caso de nuestro ejemplo, se trata de un análisis de un amplicón, ya vemos que hay un tipo del mismo que está presente en el 14.7 % de las secuencias generadas.

Figura 16
Secuencias sobrerepresentadas en nuestro conjunto de datos

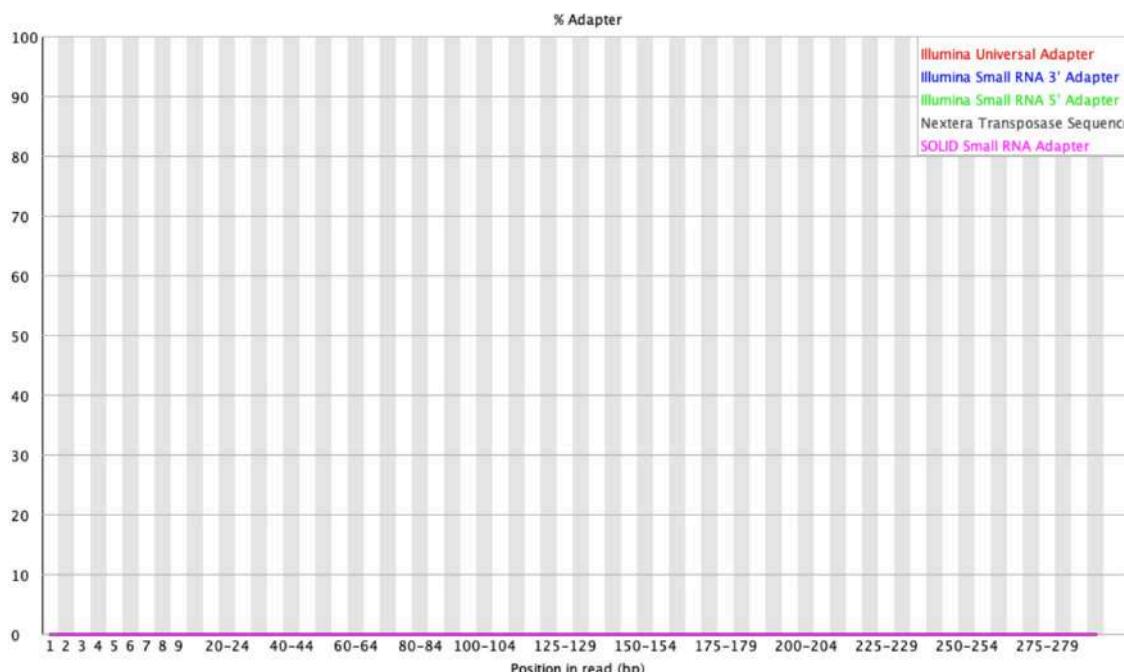
Sequence	Count	Percentage	Possible Source
CCTACGGGTGGCAGCGACTCGAGAATCATTACAATGGGGAAACCCGTAT	42712	14.6955908410604	No Hit
CCTACGGGTGGCTGCAGTCGAGAATCATTACAATGGGGAAACCCGTAT	18007	6.195530630150183	No Hit
CCTACGGGAGGCCTGCAGTCGAGAATCATTACAATGGGGAAACCCGTAT	13635	4.691290061759191	No Hit
CCTACGGGAGGCAGCAGTCGAGAATCATTACAATGGGGAAACCCGTAT	10313	3.548314954669786	No Hit
CCTACGGGGGGCAGCAGTCGAGAATCATTACAATGGGGAAACCCGTAT	10209	3.512532470883724	No Hit
CCTACGGGGGGCAGCAGTCGAGAATCATTACAATGGGGAAACCCGTAT	9297	3.1987476130674883	No Hit
CCTACGGGGGCTGCAGTCGAGAATCATTACAATGGGGAAACCCGTAT	7449	2.5629204011766933	No Hit
CCTACGGGGGCTGCAGTCGAGAATATTGGCAATGGGGAAACCCGTAT	6978	2.400867037107124	No Hit
CCTACGGGTGGCAGCAGTCGAGAATATTGGCAATGGGGAAACCCGTAT	6452	2.2198902441122335	No Hit
CCTACGGGGCGCACAGTCGAGAATCATTACAATGGGGAAACCCGTAT	6111	2.1025649847752415	No Hit
CCTACGGGGAGGCAGCAGTCGAGAATATTGGCAATGGGGAAACCCGTAT	4751	1.63464019680361	No Hit
CCTACGGGTGGCGAGTCGAGAATATTGGCAATGGGGAAACCCGTAT	4457	1.5334858676392162	No Hit
CCTACGGGGAGGCAGCAGTCGAGAATATTGGCAATGGGGAAACCCGTAT	4249	1.461920900667092	No Hit
CCTACGGGTGGCGAGCAGTCGAGAATATTGGCAATGGGGAAACCCGTAT	4160	1.4312993514424812	No Hit
CCTACGGGTGGCTGCAGTCGAGAATATTGGCAATGGGGAAACCCGTAT	3899	1.3414990796332296	No Hit
CCTACGGGGAGGCAGCAGTCGAGAATATTGGCAATGGGGAAACCCGTAT	3677	1.265117239243751	No Hit
CCTACGGGGAGGCCTGCAGTCGAGAATATTGGCAATGGGGAAACCCGTAT	3476	1.195860708080304	No Hit
CCTACGGGGAGGCAGCAGTCGAGAATATTGGCAATGGGGAAACCCGTAT	3262	1.1223313664435997	No Hit

Contenido en adaptadores (*adapter content*) (Figura 17)

En este apartado únicamente se buscan adaptadores de secuenciación que se encuentren en las lecturas analizadas. En el gráfico se muestra la proporción acumulativa de lecturas que tienen adaptadores a lo largo de sus posiciones. Cuando se detecta un adaptador, se cuenta su presencia en todas las posiciones hasta el final de la lectura, de forma que los porcentajes siempre se incrementan a medida que avanzamos en la secuencia. La presencia de estos adaptadores, habitualmente en una abundancia del 5-10 % puede ser normal, y podrán ser eliminados en los pasos posteriores del análisis. En el caso de nuestro ejemplo, no se han detectado adaptadores.

Figura 17

Ejemplo de análisis de contenido de adaptadores en nuestras secuencias



Versiones anteriores de este programa permitían realizar un análisis de contenido de *k*-meros y un análisis de calidad por celda de secuenciación, pero actualmente se encuentran descatalogados de las versiones más actuales.

4.1.4. Herramientas de filtrado por calidad de las secuencias y eliminación de adaptadores

Una vez que hemos revisado la calidad de nuestras secuencias debemos tomar la decisión de limpiarlas y mejorar su calidad, a costa de perder longitud de estas lecturas e incluso lecturas completas. Para las siguientes decisiones debemos observar los datos y elegir los parámetros adecuados. En general, nuestras secuencias serán buenas si todo el proceso previo de preparación de genoteca y secuenciación ha sido adecuado. Habitualmente lo que haremos será: buscar adaptadores y eliminarlos, así como recortar la parte final de las lecturas que decaen en calidad (*trimming* de la región final) y eliminar aquellas cuya calidad media no supere un umbral. Asimismo, durante el proceso de *trimming* por calidad puede que algunas secuencias queden muy cortas, y ya no nos interesen, en ese caso, lo idóneo es eliminarlas. Si estamos trabajando con secuencias pareadas hay que recordar que debemos eliminar también su pareja.

Adicionalmente, en aquellas secuencias procedentes de secuenciadores de dos canales de color, como son NextSeq y NovaSeq, debemos tener en cuenta la presencia de gran contenido en guanina (G, polyG), debido a que la ausencia de color, debida inclusivamente a un problema del secuenciador, se asocia a la base guanina. En este caso, debemos seleccionar un programa de limpieza que realice esta tarea adecuadamente.

En este proceso se recomienda realizar el *trimming*, y posteriormente realizar otro análisis de calidad que nos asegure que la limpieza ha sido adecuada.

Existen varios programas para realizar esta tarea (Chen *et al.*, 2014), casi todos ellos basados en Cutadapt (Martin, 2011) para eliminar adaptadores, pero con ampliaciones a *trimming* y análisis de calidad como Trimmomatic (Bolger *et al.*, 2014), TrimGalore (Babraham Bioinformatics - Trim Galore!, s. f.) o FastP (Chen *et al.*, 2018).

Por su alta versatilidad, vamos a realizar el ejemplo de eliminación de adaptadores y *trimming* con FastP.



Ejemplo

Sobre las lecturas utilizadas en el ejemplo anterior, vamos a realizar la limpieza con el programa FastP.



Enlace de interés

Antes de continuar con el ejemplo, te animo a revisar la documentación del programa, disponible en el siguiente enlace, donde encontrarás detalle de todas las opciones que ofrece.

<https://github.com/OpenGene/fastp>



Descarga de archivo

En Campus virtual > Aula de la asignatura > Recursos y materiales > Ejemplos en clase > Tema4_Ejemplo1y2_FastQC_FastP, podrás encontrar los FASTQ correspondientes a este ejemplo.

Para ejecutar este programa en nuestra máquina virtual, en el entorno conda 05MBIN_Genoma, utilizaremos el siguiente comando:

```
fastp -i reads_1.fastq.gz -I reads_2.fastq.gz -o reads_1.clean.fq.gz -0 reads_2.clean.fq.gz --cut_by_quality3 25 --cut_by_quality5 25 --cut_mean_quality 25 -l 100 --qualified_quality_phred 25 -h out_FastP.html
```

Ahora analizamos la calidad de las lecturas resultantes del análisis.

```
fastqc *.clean.fq.gz
```

>>>

>>>

Discutamos en clase las siguientes cuestiones:

- ¿Cuál es la calidad media que le hemos pedido que retenga?
- ¿Cuál es el criterio para recortar las lecturas?
- ¿Cuántas lecturas teníamos inicialmente?
- ¿Cuántas han quedado finalmente?
- ¿Cuántas han sido filtradas por baja calidad? ¿Cuántas por contener N? ¿Cuántas por ser demasiado cortas?
- ¿Cuál es el tamaño mínimo de lectura?
- ¿Se han detectado adaptadores?

4.2. Mapeo de secuencias. Herramientas de mapeo, visualización y análisis de calidad

4.2.1. El proceso de mapeo



El proceso de alineamiento o mapeo de las lecturas (también denominadas *reads*) es el proceso en el cual un programa llamado mapeador o alineador coloca estas lecturas sobre el genoma de referencia, proporcionando una o más localizaciones probables para cada lectura.

Este es un proceso fundamental y central en multitud de análisis de datos de secuenciación masiva y, por tanto, debemos cuidar su éxito para lograr el éxito del experimento. Sin embargo, un mapeo mal realizado, con parámetros incorrectos, nos dificultará extraer conclusiones.

Ya se ha visto en otras asignaturas —y hemos recalcado en capítulos anteriores— que durante el proceso de secuenciación masiva, especialmente cuando tratamos de tecnología de lecturas cortas, se realiza una fragmentación del genoma y **se secuencian en paralelo millones de veces los mismos fragmentos, obteniendo así una redundancia**. Esta es la manera de tener cada base del genoma secuenciada varias veces y calcular la fiabilidad de esa posición. **El número de veces que cada base nucleotídica está secuenciada, es decir, el número de lecturas que apoyan dicha base, es la cobertura o profundidad (coverage) de secuenciación.**

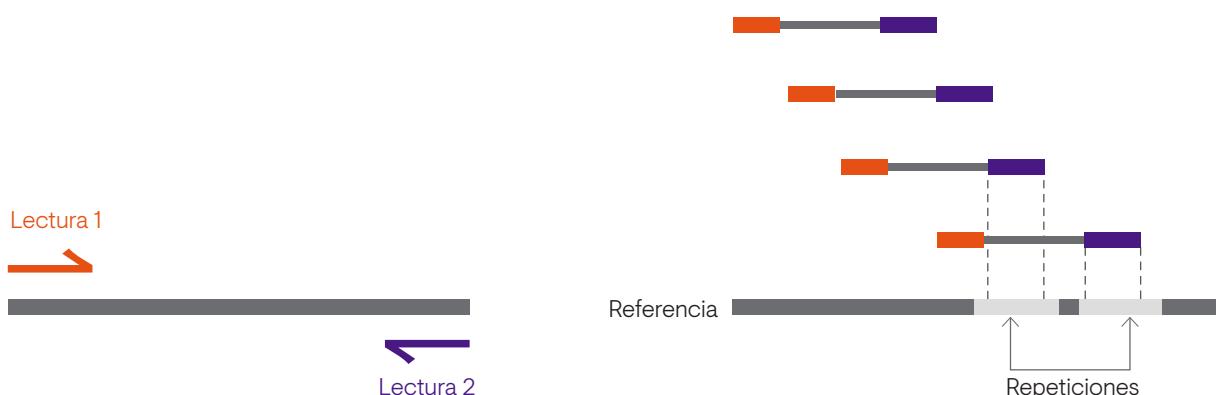
Aún así, teniendo cada base secuenciada cientos o miles de veces, ¿por qué el proceso de mapeo es tan complejo? Más allá de la propia complejidad computacional de manejar millones de datos, habitualmente imaginamos que cada lectura tiene una posición única e inequívoca sobre el genoma de referencia. Pero esto no es cierto, como vimos ya en el Capítulo 1, hay algunos factores:

- **Redundancia genómica.** Los genomas, en especial los genomas eucariotas, son muy redundantes. Tenemos regiones iguales a otras, especialmente si pensamos en pseudogenes y regiones repetitivas. En las zonas de alta redundancia, como son telómeros y centrómeros, ningún mapeador es capaz de encontrar una única posición para cada lectura.
- **Diversidad genética.** En el proceso de mapeo partimos de la base de que todos los genomas de la misma especie son esencialmente iguales. Sin embargo, esto no es estrictamente cierto, ya que sabemos que existen mutaciones que lo hacen peculiar. Esto hará que las lecturas puedan tener divergencias respecto al genoma de referencia. Es el objetivo de nuestro estudio y los mapeadores contemplan estas diferencias.
- **Errores debidos a la técnica.** Tanto durante la amplificación de la genoteca/librería, como la amplificación del material y la secuenciación se pueden producir errores en los fragmentos (cambios de un nucleótido por otro, delecciones o inserciones), que se reflejarán en las lecturas del secuenciador. Sin embargo, debemos recordar que la tasa de error asociada a secuenciación en equipos Illumina de lecturas cortas es < 0.01 %; haciendo que podamos solventar en gran medida este problema con una cobertura del genoma adecuada.

Además de estas complejidades, los mapeadores de lecturas cortas deben lidiar con **secuencias emparejadas (paired-end)** (Figura 18). Podemos secuenciar cada fragmento preparado en la genoteca de manera sencilla (*single-end*) o emparejada, por los dos extremos (*paired-end*). Para cada una de las parejas de lecturas secuenciadas de esta forma, se conoce aproximadamente la distancia entre ellas, de manera que el mapeador debe colocarlas sobre el genoma de referencia respetando esta distancia. Aunque es un reto para la computación y supone mayor complejidad para el alineamiento, las lecturas emparejadas nos permiten sortear secuencias repetitivas, siempre y cuando el tamaño de estas sea menor que el tamaño que separa las lecturas emparejadas. Esta es una de las ventajas de este tipo de lecturas, que nos permiten resolver algunas reestructuraciones cromosómicas (translocaciones e inversiones) y son de ayuda en el ensamblaje *de novo* de genomas, como veremos en capítulos posteriores.

Figura 18

Representación de las lecturas emparejadas



Nota. A la izquierda, se muestra de dónde provienen cada una de las lecturas emparejadas de cada fragmento secuenciado. A la derecha, se muestra cómo se alinean en una zona con repeticiones. La distancia entre estas lecturas se utiliza para el anclaje y solventar los problemas de redundancia.

Adaptada de *Advantages of paired-end and single-read sequencing*, por Illumina, Inc., 2021. Recuperado de <https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html>.

 Enlace de interés

En el siguiente vídeo encontrarás una explicación a la generación de clústeres y secuenciación de lecturas pareadas según tecnología Illumina.

<https://www.youtube.com/watch?v=fCd6B5HRaZ8&t=1s>

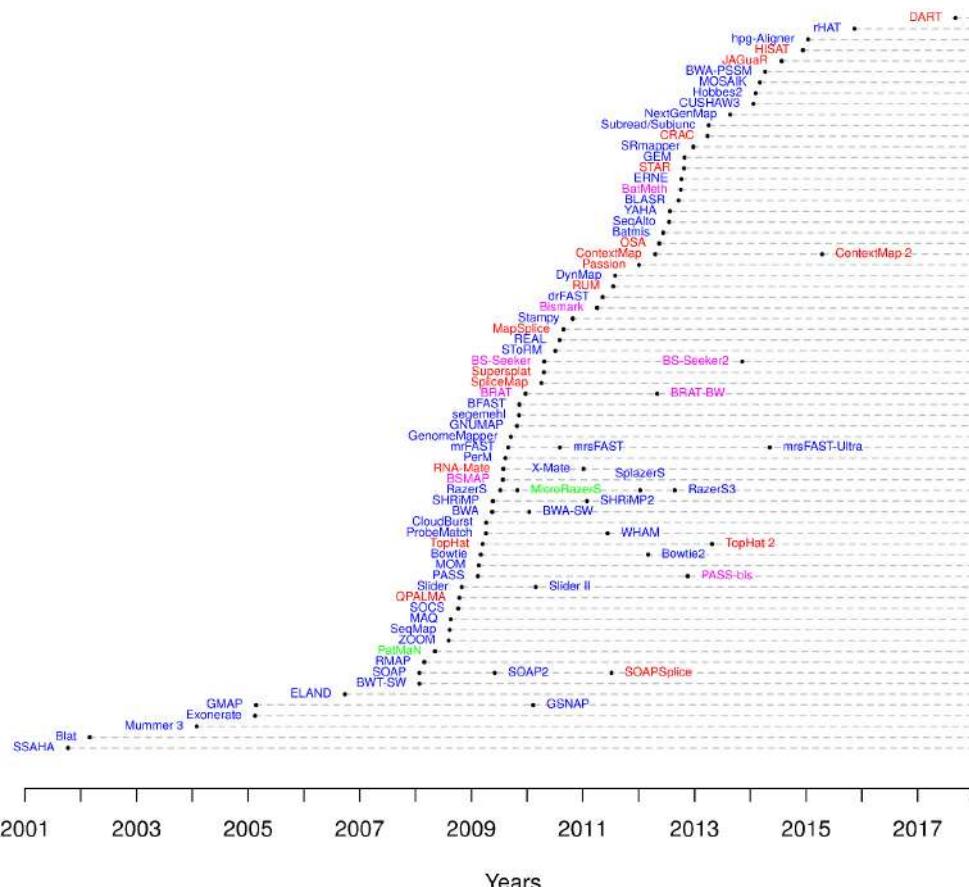
4.2.2. Herramientas para el mapeo

Desde la aparición de la secuenciación masiva y, por tanto, de millones de lecturas obtenidas de distintos genomas, aparecieron diversas herramientas para alinearlas a los genomas de referencia. Cada una de estas herramientas tiene sus peculiaridades, basadas en el algoritmo base subyacente y la aplicación a la que van dirigidas. En este apartado vamos a repasar algunas características generales, aunque en nuestras prácticas seleccionemos solo unos pocos mapeadores en función del objetivo que queramos alcanzar.

En la Figura 19 se muestra una cronología de la aparición de los mapeadores, mostrando en color azul los mapeadores de ADN (los que nos van a interesar en esta asignatura), en rojo los diseñados para ARN, en verde los diseñados para miARN y en púrpura para el análisis de secuenciación con bisulfito (análisis de metilación). La eclosión de su desarrollo fue a partir de 2008, pero en 2015 parece haber una ralentización.

Figura 19

Gráfico temporal de la aparición de mapeadores diseñados para análisis de lecturas generadas por secuenciación masiva



Nota. Tomada de HTS Mappers. 2020 (consultado el 1 de diciembre de 2021).

Nota. Tomada de HTS Mappers, 2020 (consultado el 14/07/2020).
Recuperado de http://cracs.fc.up.pt/~nf/hts_mappers/



Enlaces de interés

En el siguiente enlace se puede ver un listado de herramientas de mapeo generales actualizadas y clasificadas en función de su objetivo: búsqueda en base de datos, alineamiento pareado, alineamiento múltiple, análisis genómico, búsqueda de motivos conservados; así como un listado de editores de secuencias y visualizadores.

https://en.wikipedia.org/wiki/List_of_sequence_alignment_software

En el siguiente enlace se encuentra un listado actualizado y altamente detallado de herramientas de mapeo para secuenciación masiva.

http://cracs.fc.up.pt/~nf/hts_mappers

La clasificación de los programas de mapeo atiende a tres criterios: según tipo y tamaño de las lecturas que alinear, según el tipo de algoritmo base y según la procedencia de la genoteca/librería.

1. Segundo tipo y tamaño de las lecturas que alinear

Esta es la clasificación más general, ya que atiende a principalmente los dos tipos de lecturas que podemos encontrar:

- **Alineadores de lecturas largas.** Estos algoritmos son más lentos, pero más específicos. Son más antiguos, porque son los que tradicionalmente trabajaban con secuencias Sanger o con aquellas que provenían de secuenciadores 454 de Roche. Van a manejar de mejor manera los errores y nos van a permitir mapear de manera más precisa lecturas que vengan de regiones muy variables. Entre ellos están mapeadores como BLAT, LAST, LASTZ, BLASTZ, SMALT o MUMmer. Los utilizaremos cuando tengamos lecturas largas (> 300 pb) y pocas secuencias (< 1 M). Actualmente, dada la extensión del uso de secuenciadores tales como PacBio y Oxford Nanopore, con lecturas de > 10 kpb, existen mapeadores específicos para este tipo de lecturas, como son Minimap2 (Li, 2018), Ira (Ren & Chaisson, 2021) o LRA (GitHub - ChaissonLab/LRA: Long read aligner, s. f.).
- **Alineadores de secuencias cortas.** Aquí fundamentalmente hacemos referencia a lecturas procedentes de tecnología Illumina o análoga. Se caracterizan por ser mapeadores muy rápidos, pero muy sensibles a los errores. Son los más numerosos en la Figura 8, y comprenden el grupo de mapeadores bien conocidos como son Bowtie/Bowtie2, BWA, SOAP, TopHat/TopHat2, HISAT/HISAT2 o STAR.

Hay mapeadores híbridos, que actualmente nos permiten alinear tanto lecturas cortas como largas. Es el caso de BWA en sus últimas versiones, con su algoritmo *swy mem* (opciones *-x pacbio* o *-x ont2d*).

2. Segundo tipo de algoritmo base

- **Mapeadores basados en hashing.** El proceso de *hashing* (concepto computacional) es, de manera sencilla, crear un índice a partir de la referencia para encontrar rápidamente la posición de cualquier lectura. Este es un proceso muy general en el proceso de mapeo: generar la secuencia índice (*index*) del genoma de referencia. Este paso permite un mapeo posterior con extrema rapidez y muy sensible a errores. Se utiliza el inicio de cada lectura (*semilla, seed*) para consultar el índice. Ejemplo de estos mapeadores son STAR, SOAP y BLAT.

- **Mapeadores basados en el algoritmo de Smith-Waterman.** Este es un algoritmo de programación dinámica que garantiza que el alineamiento local es óptimo, basándose en un sistema de puntuación que utiliza una matriz de sustitución. Por tanto, dadas dos secuencias y una tabla de puntuación que actúa como reglas para puntuar el alineamiento, este algoritmo siempre encontrará el mejor alineamiento. Este tipo de mapeadores no son tan rápidos como los anteriormente descritos, pero sí son más precisos y menos sensibles a los errores. Ejemplo: BFAST.
- **Mapeadores basados en la transformación de Burrows-Wheeler (BWT).** Estos mapeadores utilizan la transformada de Burrows-Wheeler para optimizar el uso de la memoria. Se suelen utilizar junto con mapeadores basados en *hashes*. Son los preferidos para lecturas cortas, ya que ofrecen un buen balance entre eficacia, sensibilidad y especificidad. Ejemplo: BWA, Bowtie o HISAT2.

Los mapeadores más actuales, e incluso las nuevas versiones de mapeadores clásicos como BWA, combinan más de una de estas estrategias para sacar el mejor partido de ellas. Ejemplo es HISAT2, que combina BWT junto con índices de tipo GFM. Asimismo, el alineador BWA con sus funciones SW y MEM realiza una combinación, por lo que hoy en día sigue siendo uno de los mapeadores más ampliamente utilizados.

Enlaces de interés

En el siguiente enlace, podrás encontrar un vídeo para profundizar en el algoritmo de Smith-Waterman.

<https://www.youtube.com/watch?v=lu9ScxSejSE>.

En el siguiente enlace, podrás encontrar un vídeo para profundizar en la transformación de Burrows-Wheeler.

<https://www.youtube.com/watch?v=4n7NPk5lwbl>

En el siguiente enlace se encuentra la descarga para el mapeador BWA, así como su manual de utilización.

<http://bio-bwa.sourceforge.net/>

3. Segundo la procedencia de la librería

Por último, los mapeadores pueden clasificarse según su optimización para distintos tipos de librerías genómicas. Detallamos algunos de ellos:

- **Mapeadores para secuencias cromosómicas (DNAseq).** Incluye aquellos capaces de mapear lecturas obtenidas de librerías genómicas de ADN. La mayor parte de los mapeadores pueden realizarlo. Los ejemplos más conocidos: BWA, Bowtie/Bowtie2.
- **Mapeadores para secuencias de ADN complementario (RNAseq).** Cuando la librería se ha obtenido a partir de ADN complementario de una secuencia de ARN (bien total o mensajero), lo normal es que no tenga intrones, al proceder de transcriptos maduros. Bien es cierto que debido a la profundidad de la técnica encontraremos algunos transcriptos sin procesar totalmente. La eliminación de los intrones hace que el mapeo sea más difícil, ya que los alineadores deben saltar estos intrones que están en el genoma de referencia y pueden incluso medir varios kb. Si la lectura solapa con dos exones adyacentes del ARNm, el alineador debe fragmentar la lectura para colocarla, sin que esto suponga una penalización en el puntaje del mapeo. En este grupo están los alineadores TopHat/TopHat2 y HISAT/HISAT2.

Las **características** que le pedimos a un mapeador son las siguientes:

- **Capacidad de manejar errores de secuenciación y diferencias genéticas (polimorfismos).** Debido a los errores de secuenciación y a que cada individuo posee mutaciones propias respecto al genoma de referencia, nuestro mapeador debe poder alinear estas lecturas pese a no ser exactamente iguales a la referencia.
- **Especificidad.** Como hemos dicho, existe redundancia genética y una lectura puede alinearse con más de una región del genoma. El alineador debe encontrar la posición más probable.
- **Eficacia.** El mapeo masivo de millones de lecturas debe realizarse en un tiempo razonable y consumiendo unos recursos computacionales razonables.



Ejemplo

Vamos a mapear unas lecturas de ADN, procedentes de un experimento de secuenciación de ADN genómico sobre el genoma de referencia. Para ello vamos a utilizar BWA como alineador.

 **Descarga de archivo**
En Campus virtual > Aula de la asignatura > Recursos y materiales > Ejemplos en clase > Tema4_Ejemplo3_Mapeo, podrás encontrar los archivos necesarios para este ejemplo.

Vamos a mapear unas lecturas de ADN, procedentes de un experimento de secuenciación de ADN genómico de un panel sobre el genoma de referencia. Para ello vamos a utilizar BWA como alineador.

Necesitaremos el genoma de referencia humano en versión Hg19/GRCh37:

http://ftp.ensembl.org/pub/grch37/current/fasta/homo_sapiens/dna/

Para agilizar el proceso podemos utilizar un único cromosoma para el mapeo. Fecha del archivo: 27/Nov/2015.

1. Indexar el genoma de referencia: `bwa index genoma_ref.fa` [aprox. 178 sg].
2. Mapear las lecturas limpias de pasos anteriores al genoma de referencia: `bwa mem -a genoma_ref.fa reads_R1.fastq.gz reads_R2.fastq.gz -o sample.sam 2> sample.out`

4.2.3. Archivos de mapeo: el formato SAM

Los archivos resultantes del proceso de mapeo son archivos SAM (*sequence alignment map*). En este formato se representan cada una de las lecturas sobre el genoma de referencia, indicando su secuencia, posición exacta de inicio, de final, hebra sobre la que mapea y parámetros de calidad del mapeo.

 Enlace de interés

En el siguiente enlace, podrás encontrar la documentación oficial sobre el archivo estándar de tipo SAM.

<https://samtools.github.io/hts-specs/SAMv1.pdf>

El formato SAM es un archivo de texto plano delimitado por tabuladores, con dos secciones definidas (Figura 20), la sección cabecera y los alineamientos.

Figura 20

Ejemplo de un fichero de formato SAM

1. Sección de cabecera

La sección de cabecera comienza con un carácter @, seguido de dos letras mayúsculas que codifican el tipo de registro, tras las que vienen las etiquetas (TAG) separadas por tabuladores, en formato TAG:VALOR. Cada TAG son dos letras mayúsculas que definen el formato y contenido del VALOR. Existen pocas TAG y solo son válidas las definidas en las directrices del formato SAM. En la Figura 20 se encuentran dos líneas correspondientes a la cabecera:

- **@SQ**: este registro indica la secuencia de referencia.
 - **SN:12**: nombre de la secuencia de referencia, en este caso del cromosoma 12.
 - **LN:133275309**: longitud de la secuencia de referencia.
 - **@PG**: este registro hace referencia al programa utilizado para mapear/alinear las lecturas a la secuencia de referencia.
 - **ID:bwa**: identificador del programa. BWA en este caso
 - **PN:bwa**: nombre del programa para alinear.
 - **VN:0.7.17-r1188**: versión del programa utilizado.
 - **CL :...:** es la línea que indica la línea de comandos utilizada para ejecutar este programa.

2. Sección de alineamientos

La sección de alineamientos va desde la primera línea que no comienza por @ hasta la última del fichero. Contiene la información del alineamiento o mapeo de las lecturas con el genoma de referencia. En esta sección cada línea corresponde al alineamiento de un segmento. Deben aparecer todos los campos obligatorios, siempre en el mismo orden, aunque su valor sea "0" o "*" cuando la información no está disponible. Tras estos valores vienen los campos opcionales.

Las columnas separadas por tabulador son las siguientes:

- **QNAME (query name):** nombre de la lectura en el alineamiento.
- **FLAG:** describe las propiedades del alineamiento. Es un campo fundamental, en el que ahondaremos más adelante.
- **RNAME:** nombre de la secuencia de referencia.
- **POS:** posición del inicio del alineamiento.
- **MAPQ:** calidad del mapeo. Este valor lo otorga el programa de mapeo en función de la calidad del alineamiento, y su valor es más alto cuanto más perfecto sea. Valora también la presencia de mapeos secundarios de esa lectura.
- **CIGAR:** este campo muestra las operaciones realizadas para alinear las dos secuencias (lectura y referencia).
- **RNEXT:** nombre de la secuencia en al que ha alineado su pareja, cuando las lecturas son pareadas. El símbolo “=” indica que la pareja está mapeada sobre la misma referencia.
- **PNEXT:** posición de la pareja en el alineamiento.
- **TLEN:** longitud del alineamiento, calculada según las coordenadas de la secuencia de referencia.
- **SEQ:** secuencia de la lectura que participa en el alineamiento.
- **QUAL:** calidad de la lectura, extraída del archivo FASTQ original.

Los campos opcionales aparecen a continuación, separados por tabuladores, en el formato TAG:TIPO:VALOR.



Enlace de interés

En el siguiente enlace, podrás acceder a un documento donde se pueden encontrar las TAG que hay para los camposopcionales.

<https://samtools.github.io/hts-specs/SAMtags.pdf>

3. El campo FLAG

Anteriormente se ha comentado que este es un campo obligatorio que se utiliza para describir con precisión las propiedades del alineamiento. Cada una de las flag está definida en la Tabla 6, pero en el valor que se observa en el archivo SAM (ejemplificado en la Figura 20) es la suma total de los valores en base decimal asignados para cada una de las flag.

Tabla 6

Detalle de las flag utilizadas en el formato SAM

Número	Decimal	Hexadecimal	Descripción
1	1	0x1	La lectura forma parte de un par
2	2	0x2	La lectura forma parte de un emparejamiento correcto

>>>

>>>

Número	Decimal	Hexadecimal	Descripción
3	4	0 × 4	Lectura no mapeada
4	8	0 × 8	Pareja no mapeada
5	16	0 × 10	Lectura mapeada en la cadena complementaria (negativa) del genoma de referencia
6	32	0 × 20	Pareja mapeada en la cadena complementaria (negativa) del genoma de referencia
7	64	0 × 40	Primera lectura del par
8	128	0 × 80	Segunda lectura del par
9	256	0 × 100	Alineamiento secundario
10	512	0 × 200	La lectura no ha pasado los filtros de calidad
11	1024	0 × 400	La lectura es un duplicado óptico o de PCR
12	2048	0 × 800	Alineamiento suplementario

Nota. Adaptado de SAM format, consultado en diciembre de 2021, disponible en <https://www.samformat.info/sam-format-flag>



Ejemplo

Veamos algunos ejemplos de *flags*:

- Flag 83 ($1 + 2 + 16 + 64$): la lectura forma parte de un par (1) correctamente alineado (2), es la primera lectura del par (64) y se ha mapeado en la cadena complementaria (antisentido) del genoma de referencia (16).
- Flag 99 ($1 + 2 + 32 + 64$): la lectura forma parte de un par (1) correctamente alineado (2), es la primera lectura del par (64) y se ha mapeado en la cadena positiva (sentido) del genoma de referencia. Este último punto lo sabemos porque el código 32 indica que su pareja se ha mapeado en la cadena antisentido del genoma de referencia.
- Flag 73 ($1 + 8 + 64$): la lectura forma parte de un par (1), es la primera lectura (64) del par, pero su pareja no ha sido mapeada.



Enlaces de interés

Si deseáis ahondar en el formato SAM, especialmente en las *flag*, cabeceras y etiquetas de alineamiento, en el siguiente enlace se encuentra mucha información.

<https://www.samformat.info/sam-format-flag>

Para poder realizar el cálculo automático de las *flag* podéis utilizar el recurso del siguiente enlace:

<https://broadinstitute.github.io/picard/explain-flags.html>

4.2.4. Herramientas para el manejo de los archivos SAM. Generación de archivos BAM

Hasta ahora hemos visto archivos de mapeo SAM, de texto plano, que habitualmente tienen un peso alto, computacionalmente hablando. Además, manejar archivos de texto no es computacionalmente rentable. Para ahorrar espacio y agilizar los cálculos posteriores, los **archivos SAM se convierten en archivos binarios, denominados BAM**. Ambos archivos contienen la misma información. Para realizar esta conversión se utiliza un paquete de herramientas, llamadas Samtools, que nos permiten leer, filtrar, transformar y extraer información de los ficheros BAM y SAM.



Enlace de interés

En el siguiente enlace se encuentra información detallada de la herramienta Samtools.

<http://www.htslib.org/>

Se van a ejemplificar algunos usos de esta herramienta a través de varios ejemplos.



Ejemplo



Descarga de archivo

En *Campus virtual > Aula de la asignatura > Recursos y materiales > Ejemplos en clase > Tema4_Ejemplo4_Indexado*, podrás encontrar el archivo SAM sobre el que replicar este ejemplo.

- “samtools view” para leer un fichero SAM/BAM. Esta herramienta nos permite leer un fichero SAM y convertirlo en un fichero BAM. `<samtools view -bS mapping.sam > mapping.bam>`
- “samtools view” para visualizar un archivo BAM `<samtools view -h mapping.bam>`
- “samtools view” para seleccionar los alineamientos con una calidad superior a la especificada. `<samtools view -q 30 mapping.bam>`
- “samtools sort” permite ordenar un fichero por las coordenadas genómicas. Este paso es imprescindible para visualización en programas específicos. `<samtools sort mapping.bam -o mapping.sorted.bam>`
- “samtools index” genera un archivo índice (fichero con extensión BAI) para un fichero BAM ordenado. `<samtools index mapping.sorted.bam>`
- “samtools flagstat” nos da unas estadísticas básicas pero útiles para tener una visión general sobre los alineamientos en el archivo BAM. `<samtools flagstat mapping.sorted.bam>`

4.2.5. Visualización del mapeo

Antes de continuar en el análisis del mapeo, se puede realizar una inspección visual del alineamiento de las lecturas contra el genoma de referencia, siendo un paso complementario a cualquier proceso de análisis de variantes genómicas. Para este fin se utiliza comúnmente el programa Integrative Genomics Viewer (IGV).

Enlace de interés

En el siguiente enlace se encuentra la información de este software, así como su descarga.

<https://software.broadinstitute.org/software/igv/>

Este visualizador funciona en distintos sistemas operativos, como Windows, Linux o Mac. Se encuentra instalado en la máquina virtual de trabajo y puede abrirse desde la terminal, tecleando `igv` en el *prompt*.

Los puntos más importantes para tener en cuenta cuando revisemos la visualización de un genoma, tal y como se ejemplifica en la Figura 21, son:

- Cargar en el programa el genoma de referencia adecuado a nuestro análisis. IGV incluye algunos genomas de referencia estándar, pero pueden descargarse un gran número de ellos e incluso cargar un genoma de referencia propio, a partir de un archivo FASTA.
 - Tener un archivo BAM del mapeo previamente ordenador por coordenadas (tal y como se ha visto en el apartado 4.2.4 de este documento) y con su índice BAI, localizado en la misma carpeta del archivo.
 - Es opcional cargar en el programa el archivo de sondas de captura, que se trata de un archivo en formato BED donde se especifica para cada cromosoma, qué región cubre cada sonda.

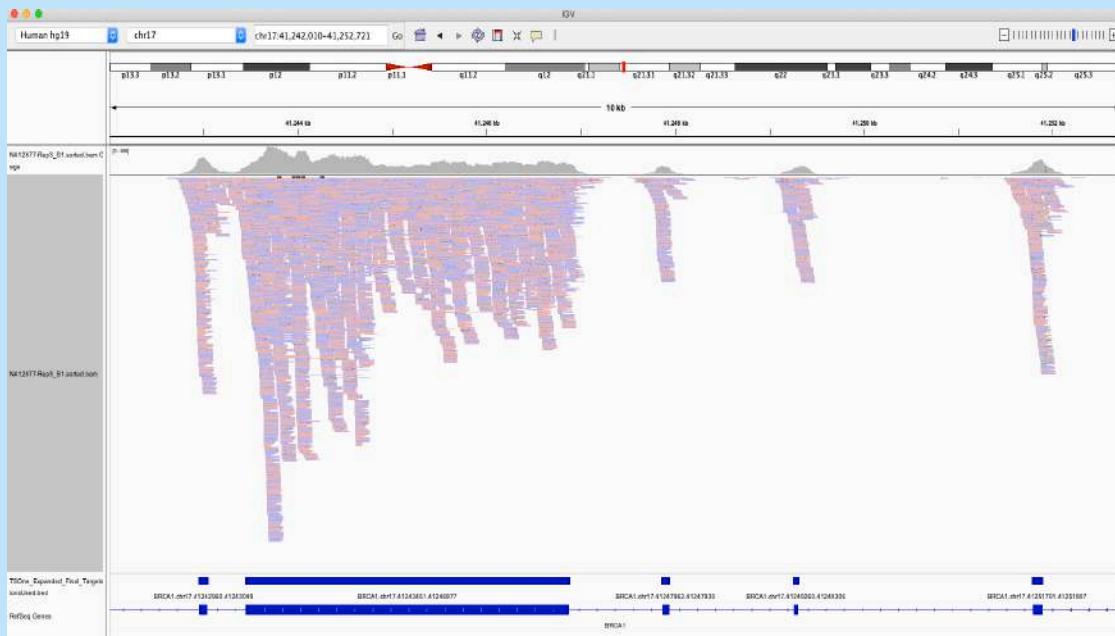
Ejemplo

Descarga de archivo

En Campus virtual > Aula de la asignatura > Recursos y materiales > Ejemplos en clase > Tema4_Ejemplo5_IGV, podrás encontrar los archivos BAM y de sondas (BED) para poder practicar.

Figura 21

Visualización de una región del genoma humano mapeado, junto con las sondas de captura utilizadas en este experimento



Enlace de interés

Para ahondar en la visualización de archivos BAM en el explorador IGV, en el siguiente enlace podrás acceder a un tutorial.

https://www.youtube.com/watch?v=E_G8z_2gTJM

4.2.6. Análisis de calidad del mapeo

El proceso de análisis de la calidad del mapeo es fácil si disponemos de unas pocas secuencias, pero en este caso, donde se han manejado millones de ellas, puede resultar muy complejo. Una persona altamente entrenada en visualizar los mapeos tiene una idea bastante aproximada de la calidad de este, simplemente realizando una inspección visual del mismo.

En este apartado analizaremos la calidad del mapeo utilizando Qualimap2 (Okonechnikov *et al.*, 2015), una herramienta con interfaz gráfica y por línea de comandos que nos permite la evaluación de experimentos de secuenciación de genoma o exoma, o incluso experimentos de secuenciación de ARN.

Enlaces de interés

La última versión del programa, así como una completa guía de uso está disponible en los siguientes enlaces.

<http://qualimap.conesalab.org/>

http://qualimap.conesalab.org/doc_html/index.html

El módulo BamQC de Qualimap genera un **informe** en el que se evalúa la calidad de los datos de alineamiento del archivo BAM. Nos genera la siguiente información:

- **Resumen.** Información y estadísticas básicas de alineamiento, como son datos globales sobre el número de lecturas, número de lecturas mapeadas, eficacia del mapeo emparejado, distribución de la longitud de las lecturas, número de lecturas recortadas, tasa de duplicación, contenido en nucleótidos y porcentaje GC, cobertura media y desviación típica de la profundidad, calidad media de los mapeos, tamaño medio de inserto, visión general de la tasa de error y estadísticas sobre el cromosoma (número de bases mapeadas, media y desviación estándar para cada uno de los cromosomas).
- **Profundidad a lo largo de la referencia.** La gráfica muestra una distribución de la profundidad, así como la desviación sobre la referencia. También muestra el porcentaje GC a lo largo de la referencia.
- **Histograma de profundidad.** Representa el número de regiones genómicas que tienen un determinado intervalo de profundidad.
- **Histograma de profundidad 0-50x.** Es un histograma similar al anterior, pero solo se representa la profundidad hasta 50x.
- **Profundidad de la fracción del genoma.** Muestra la fracción del genoma cubierto al menos a una determinada profundidad. Resulta útil para saber si la secuenciación ha cubierto nuestras expectativas.

- **Histograma de la tasa de duplicación.** Nos muestra la tasa de duplicación.
- **Contenido nucleotídico de las lecturas mapeadas.**
- **Distribución GC en las lecturas mapeadas.** Nos muestra el contenido GC solo de las lecturas mapeadas. Si se provee de un genoma de referencia, realizará la comparativa con el mismo.
- **Perfil de recorte de las lecturas mapeadas.** Nos muestra el porcentaje de bases recortadas a lo largo de las lecturas en el proceso de mapeo. El número total de lecturas recortadas aparece en el apartado de resumen.
- ***Indels* en homopolímeros.** Nos muestra el número de *indels* que están dentro de homopolímeros de cualquiera de los nucleótidos. Un número elevado de homopolímeros indica un problema de secuenciación.
- **Calidad del mapeo a lo largo de la referencia.** Distribución de la calidad a lo largo de la referencia analizada.
- **Histograma de la calidad del mapeo.** Histograma del número de localizaciones genómicas con una calidad de mapeo dada.
- **Tamaño de los insertos a lo largo de la referencia.**
- **Histograma del tamaño de insertos.**



Ejemplo

Ahora puedes crear tu análisis Qualimap a partir del archivo mapping.bam obtenido en apartados anteriores y analizar cada apartado, utilizando como guía los puntos anteriores.

 **Descarga de archivo**
En Campus virtual > Aula de la asignatura > Recursos y materiales > Ejemplos en clase > Tema4_Ejemplo6_Qualimap, podrás encontrar el archivo BAM sobre el que practicar con este ejemplo.

Además del modo interactivo de Qualimap, al que se puede acceder simplemente invocando al programa desde la terminal; la opción por la línea de comandos es deseable si pensamos en automatizar y escalar el proceso:

```
qualimap bamqc -bam mapping.bam -c -nt 2 -gd HUMAN -gff genome.gff -  
java-mem-size=32G
```



Enlaces de interés

Los siguientes programas que se encuentran en los siguientes enlaces son herramientas muy útiles para determinar la calidad de un mapeo sobre un archivo BAM, siendo programas más sencillos en cuanto a resultados ofrecidos.

SAMstat:

<http://samstat.sourceforge.net/>

ngsCAT (especializado en evaluar capturas dirigidas):

<http://ngscat.clinbioinfosspa.es/start>

4.3. Identificación de variantes

Como ya se explicó en el Capítulo 1 de este manual, la variación dentro del genoma humano es un proceso habitual y normal. Estas variaciones incluyen la alteración en el número de cromosomas de la célula o la alteración de estos o de los genes.

A lo largo de esta sección se desgranará un protocolo bioinformático base para analizar las mutaciones germinales. El análisis de mutaciones somáticas sería similar, variando alguno de los pasos de la llamada de variantes, para tener en cuenta el mosaicismo genético.



La identificación y anotación de variantes es el proceso que comprende la extracción de los sitios variables (inserciones, delecciones, cambios nucleotídicos) o diferencias respecto al genoma de referencia a partir del alineamiento de las secuencias obtenido en pasos anteriores.

El proceso de identificación de variantes no es sencillo, ya que existen numerosos factores que complican la identificación de estas y que pueden hacernos ver variantes (SNV, *indels*, CNV) cuando realmente no las hay. Entre los factores que nos conducen a error están:

- **Sesgos en el proceso** de amplificación cuando se preparan las genotecas en el laboratorio.
- **Errores de secuenciación** durante el proceso de lectura que realiza la máquina. Para paliar este efecto, cada base nucleotídica leída lleva asociado un índice de calidad que nos da la fiabilidad determinada de que ese nucleótido sea real.
- **Cuestiones relacionadas con el software** de alineamiento o mapeo, especialmente en aquellas regiones altamente repetidas. Cada software calcula una probabilidad de que esa lectura esté asignada a esa posición.

Por ello, los programas de identificación de variantes deben tratar de distinguir entre variaciones reales y artefactos debidos a estos errores generados durante el proceso de la generar la genoteca, secuenciación y análisis.



Aunque existen en la literatura numerosos programas para realizar esta tarea de identificación de variantes, los más conocidos son **GATK** y **FreeBayes**. GATK es un paquete de software que ofrece una amplia variedad de herramientas focalizadas a la identificación de variantes y genotipado en células somáticas y germinales. Su arquitectura permite implementarlo en sistemas operativos Linux a gran escala, como por ejemplo el proyecto 1000 genomas. Desde 2018 este software es abierto. Además, es altamente modular, permitiendo combinar herramientas para abordar distintos problemas. El flujo de trabajo recomendado se encuentra en el apartado de “buenas prácticas” de este enlace. En este tema veremos un procedimiento sencillo pero robusto para la identificación de variantes.

Por otra parte Freebayes es un detector de variantes bayesiano desarrollado para la detección de pequeños polimorfismos, como SNP, *indels*, polimorfismos múltiples de nucleótido y eventos complejos (como inserciones y sustituciones complejas) (Garrison & Marth, 2012). Se basa en haplotipos, ya que llama variantes basadas en las lecturas literales alineadas, no en su alineamiento exacto sobre el genoma de referencia. Su funcionamiento es más sencillo que otros programas como GATK. Este software presenta una mejor complejidad en su manejo, así que para introducirnos, utilizaremos este en las prácticas.



Enlaces de interés

En el siguiente enlace encontrarás amplia documentación sobre el uso de GATK.

<https://gatk.broadinstitute.org/hc/en-us>

En el siguiente enlace encontrarás el software FreeBayes, así como un manual de utilización.

<https://github.com/freebayes/freebayes>

En los siguientes artículos se encuentran varias revisiones de software utilizados para identificación de variantes, donde se indica para qué tipo de variantes están indicados y la tecnología de secuenciación más apropiada.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., & Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, 15(2), 256–278. <https://doi.org/10.1093/bib/bbs086>
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3956068/>

Zhao, S., Agafonov, O., Azab, A., Stokowy, T., & Hovig, E. (2020). Accuracy and efficiency of germline variant calling pipelines for human genome data. *Scientific reports*, 10(1), 20222.

<https://doi.org/10.1038/s41598-020-77218-4>

<https://www.nature.com/articles/s41598-020-77218-4>

Después del mapeo o alineamiento (Figura 7), la **etapa de preprocesamiento** tiene como objetivo preparar las secuencias alineadas contra el genoma de referencia para su posterior uso en la identificación de variantes. Esta preparación consiste en realizar ajustes sobre las secuencias, de forma que la posterior identificación de variantes trate de localizar el mayor número de ellas (minimizar los falsos negativos), pero a su vez, evitar los sesgos inherentes a los datos de secuenciación que inducen a errores en la identificación.

Por tanto, estas etapas de preprocesamiento tienen en cuenta **tres factores principales que nos pueden inducir a errores:**

- Presencia de secuencias exactamente iguales (duplicadas).
- Errores introducidos en la etapa de alineamiento.
- Errores generados por la asignación de calidad por parte del secuenciador a cada una de las bases.

Una vez realizados estos tres ajustes, en tres pasos diferenciados, se procede a la **identificación de variantes**, tratando de hallar el mayor número posible de ellas (maximizar la sensibilidad). Finalmente se realiza un filtrado de estas variantes identificadas (etapa de posprocesado) para eliminar aquellas que no cumplen unos criterios mínimos de calidad, etiquetándolas adecuadamente.

Esta **etapa de posprocesado o filtrado** trata de minimizar el número de falsos positivos. Por tanto, la etapa de identificación de variantes está diseñada para maximizar la sensibilidad, mientras que el posprocesado lo está para lograr un nivel de especificidad que pueda personalizarse en cada proyecto.

A continuación, se detallan cada una de estas etapas.

4.3.1. Preprocesamiento: identificación de secuencias duplicadas

Tras el alineamiento podemos encontrarnos secuencias duplicadas, es decir, cuya posición de inicio y final sobre el genoma de referencia es el mismo lugar. Estas lecturas duplicadas pueden deberse a artefactos en el paso de amplificación de PCR de la librería/genoteca, o a la lectura por parte del secuenciador de ese fragmento de ADN más de una vez. Esto puede suponer un problema cuando intentamos detectar variantes. Imaginemos que un error de secuenciación o de amplificación es propagado en distintas secuencias duplicadas, dando lugar a variantes falsas. Por tanto, el primer paso que realizamos es identificar las lecturas cuyas coordenadas externas mapeen en la misma posición del genoma de referencia y nos quedamos con una de ellas.

El proceso de eliminación de lecturas duplicadas se realiza con la herramienta PicardTools, que identifica todas las lecturas que tienen la misma posición de inicio, final y orientación. Para ello se fija en el valor CIGAR del archivo SAM o BAM. Marca todas las lecturas duplicadas excepto una, la que posee mayor calidad, y marca como duplicadas las demás.



Enlaces de interés

En el siguiente enlace se encuentra una descripción más detallada del valor CIGAR, presente en los archivos de alineamiento SAM o BAM.

<https://sites.google.com/site/bioinformaticsremarks/bioinfo/sam-bam-format/what-is-a-cigar>

Puedes encontrar la herramienta PicardTools y su manual de uso en el siguiente enlace.

<https://broadinstitute.github.io/picard/>



Ejemplo

Para este ejemplo vamos a utilizar el archivo BAM proporcionado en ejemplos anteriores (Tema 4 – Ejemplo 6). Podéis encontrarlo en:

>>>



Descarga de archivo

En Campus virtual > Aula de la asignatura > Recursos y materiales > Ejemplos en clase > Tema4_Ejemplo7_duplicados, podrás encontrar el archivo BAM sobre el que practicar con este ejemplo.

```
</>  
picard MarkDuplicates --INPUT sample.sorted.bam --OUTPUT sample.  
dedup.bam --METRICS_FILE markDuplicatesMetrics.txt --ASSUME_SORTED True  
samtools index sample.dedup.bam
```

Dado que no incluimos en el mapeo la información de grupos de lecturas (ReadGroups), los incluimos ahora:

```
</>  
picard AddOrReplaceReadGroups I=sample.dedup.bam O=Sample.dedup.  
RG.bam RGID=M02899 RGLB=lib1 RGPL=ILLUMINA RGPU=unit1 RGSM=20
```

4.3.2. Identificación de variantes

En este punto tenemos un archivo BAM de alineamiento donde hemos introducido una serie de mejoras para evitar sesgos producidos durante pasos anteriores y, con el objetivo de identificar todas las variantes potenciales de la manera más precisa posible, minimizando el número de falsos positivos.

En este punto identificaremos las **variantes**, determinando en qué posición al menos una base difiere con la referencia, e identificaremos el **genotipo** para el individuo en esas posiciones modificadas. Recuentaremos para cada posición modificada el número de ocurrencias de cada nucleótido teniendo en cuenta todas las lecturas alineadas en esa posición (Figura 7).



Para cada gen podemos encontrar versiones alternativas que se diferencian en la secuencia, es lo que denominamos **alelo**. Por otra parte, la combinación particular de alelos para un locus es el **genotipo**. Se denomina genotipo **homocigoto** si los alelos que lo componen son exactamente iguales o **heterocigoto** si los alelos difieren.

Para la detección de las distintas variantes o alelos que podemos encontrar en un locus, deben tenerse en cuenta los sesgos intrínsecos a los datos de secuenciación y análisis bioinformático. Por ello, se siguen métodos probabilísticos que incorporan estas incertidumbres en la detección de variantes, como es el software GATK, donde se tienen en cuenta:

- La **calidad de todas las bases** (valor Phred) en la posición objeto de estudio, teniendo en cuenta todas las lecturas que apoyan esa posición.

- Proximidad a una región de inserciones, delecciones o regiones homopolímeras.
- Calidad del mapeo de las lecturas que apoyan la variante. Valores bajos de calidad pueden indicar región repetida.
- Longitud de las secuencias que soportan la variante. Secuencias cortas tienen más posibilidades de alinear en múltiples localizaciones del genoma.
- Profundidad de la secuenciación. Cuando una variante está soportada por un número muy elevado de secuencias, esto apoya que sea real.



Ejemplo

Utilizamos la herramienta FreeBayes, indicándole que aquellas posiciones soportadas por menos de 25 lecturas (-C) sean descartadas del análisis.

```
freebayes -C 25 -f  
./mapeo/Homo_sapiens.GRCh37.dna.chromosome.7.fa  
chr7.dedup.RG.bam > chr7.vcf
```



Una vez realizado el procesamiento bioinformático para la detección de variantes, donde GATK toma el archivo BAM preprocesado, el archivo resultante es un archivo en formato **variant call format (VCF)**. Este tiene un **formato de texto plano, tabular, donde están presentes todas las variantes que han sido encontradas en el genoma**, con numerosa información sobre cada una de ellas. Este es un archivo estándar que puede ser visualizado en programas como IGV, que ya hemos visto.

El formato VCF es bastante complejo, así que vamos a describir las características principales. Son ficheros de texto plano, pero pueden llegar a ser muy pesados. Es importante no abrir un archivo de este tipo en un procesador Word, puesto que elimina el formato del archivo, sino en editores de texto plano.



Enlaces de interés

En el siguiente enlace podéis encontrar una descripción detallada del formato VCF en sus dos últimas versiones.

<https://samtools.github.io/hts-specs/VCFv4.2.pdf>

<https://samtools.github.io/hts-specs/VCFv4.3.pdf>

Estructura básica del archivo VCF

Este fichero está formado por dos partes principales:

- **Cabecera:** contiene información acerca del conjunto de datos y otro tipo de información, como organismo y versión del genoma de referencia, como las definiciones o descripciones de todas las anotaciones utilizadas.
 - Versión del fichero VCF: `##fileformat=VCFv4.1`

- Líneas FILTER, que nos indican el tipo de filtros que se han utilizado para filtrar las variantes detectadas que no cumplen los parámetros de calidad. `##FILTER=<ID=LowQual,Description="Low quality">`
- Líneas FORMAT e INFO. Definen las anotaciones contenidas en las columnas FORMAT e INFO de la sección de variantes identificadas.
- Líneas de contigs y referencia. En estas líneas están los nombres de los **contigs** o cromosomas, sus longitudes y qué versión de referencia se utilizó en el fichero BAM de entrada.
- Apartado “Records” (Figura 22): tras la cabecera, aparecen las variantes identificadas, donde tenemos una línea de cabecera y a continuación una línea por cada variante identificada.

La línea de cabecera define las columnas de las siguientes líneas, que tienen un formato separado por tabuladores. Las primeras ocho columnas, hasta el campo INFO inclusive, representan las propiedades observadas a nivel de variante identificada. La información específica de la muestra (genotipo) se muestra en la columna FORMAT (novena columna). Además, hay una columna adicional de información por cada muestra representada en el archivo VCF. [#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 ...]

Existen dos tipos de anotaciones en el apartado “Records”:

- Anotaciones a nivel de variante. Se refiere a las siete primeras columnas, que son obligatorias en todo archivo VCF. Si el valor para ese dato no se conoce, se indica con un punto ('.'). De esta forma:
 - CHROM: define el contig o cromosoma donde está la variante.
 - POS: define la posición sobre el genoma de referencia. Si esta variación es una delección, la posición corresponde a la base anterior a la delección.
 - ID: identificador opcional de la variante, si hemos utilizado una base de datos de variantes como dbSNP.
 - REF: alelo de referencia.
 - ALT: alelo alternativo. (Nota: tanto REF como ALT se proporcionan respecto a la hebra sentido 5'-3' -forward-; adicionalmente, para las inserciones, el alelo ALT contiene la base insertada y su precedente; mientras que para delecciones, el alelo ALT es la base anterior a la delección).
 - QUAL: probabilidad en escala Phred de que el polimorfismo REF/ALT exista.
 - FILTER: este campo tiene el nombre o nombres de los filtros en los que la variante no ha pasado. Si ha cumplido con todos los filtros, indicará “PASS”. Si el valor FILTER es “.”, significa que no se ha aplicado ningún filtro a las variantes.
 - INFO: este campo no es obligatorio y corresponde a diferentes anotaciones realizadas sobre la variante. Estas anotaciones nos proporcionan información variada de las muestras. La información descrita en este campo siempre está definida en la cabecera del archivo VCF (líneas ##INFO), lo que facilita su comprensión. Las anotaciones más frecuentes son: AC (número de alelos para el genotipo, para cada alelo ALT), AF (frecuencia alélica para cada alelo ALT), AN (Número total de alelos identificados), DP (profundidad combinada de todos los alelos). Para otras anotaciones adicionales, podéis consultar la guía de formato VCF indicada anteriormente en enlaces de interés.

- **Anotaciones a nivel de muestra.** El resto de las columnas situadas a la derecha de la columna INFO nos muestran información a nivel de muestra. Como mínimo esta información indica el genotipo inferido en la muestra.

Figura 22

Estructura básica de un archivo VCF

```
##fileformat=VCFv4.0
##fileDate=20220114
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="Hapmap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 . T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
```

Nota. En azul se encuentra la zona de cabecera y en rojo la zona de información de cada variante para cada muestra.

4.3.3. Posprocesado: filtrado de variantes

Una vez identificadas todas las potenciales variantes, que en nuestro caso son SNV, inserciones y delecciones, es recomendable realizar un filtrado de todas las variantes. En la etapa anterior de identificación de variantes hemos tratado de maximizar la sensibilidad (minimizar falsos negativos), pero en la etapa de filtrado vamos a maximizar la especificidad (minimizar falsos positivos).

Esta etapa debe adaptarse a cada tipo de proyecto.



Ejemplo

Vamos a filtrar separadamente los SNP de los *indels* a partir del archivo VCF obtenido en el proceso anterior, para lo que utilizamos la herramienta VCFTools.

```
</>
vcftools --vcf chr7.vcf --keep-only-indels --recode
--recode-INFO-all --out chr7_indels.vcf

vcftools --vcf chr7.vcf --remove-indels --recode --recode-INFO-all
--out chr7_snps.vcf
```

4.4. Anotación de variantes

El último paso, bioinformáticamente hablando, en el estudio de variantes es la anotación de las variantes preprocesadas y filtradas. Debemos tener en cuenta algunos datos, que nos dan idea de que en las etapas previas tenemos un listado de miles de variantes, cuando analizamos un exoma, o millones si estamos analizando un genoma completo.

Pero no todas ellas, sino un número muy reducido, serán las responsables de la enfermedad que estamos analizando.



Ejemplo

Una persona sana, en promedio, tiene los siguientes números de variantes en su exoma:

- > 10 000 variantes **no sinónimas**, donde la variación de un solo nucleótido hace que el codón correspondiente codifique para otro aminoácido.
- > 10 000 variantes **sinónimas**, donde la variación del nucleótido no implica una variación en el aminoácido codificado.
- > 250 variantes con **pérdida de función** en el gen anotado.
- > 50 variantes que poseen **relación** con enfermedades previamente descritas.



En esta etapa identificaremos un subconjunto pequeño de variantes potencialmente interesantes, a partir del total de variantes candidatas. Esta anotación nos ayudará a reducir el número de variantes y focalizar nuestra atención y esfuerzos en las más interesantes.

Este proceso de anotación es el proceso de asignar información biológica relevante a las variantes detectadas, haciendo uso de diversas bases de datos. Este proceso puede ser iterativo, y ser necesarias varias bases de datos para recabar información suficiente para determinar las variantes sospechosas.

Lo primero que nos va a interesar de una variante es su consecuencia, ¿qué consecuencia tiene esa variante? Puede tener las siguientes **consecuencias**:

- **Variante en una región codificante**, cuyo cambio nucleotídico no supone un cambio en la proteína. Es una variante sinónima (*synonymous variant*).
- **La variante cambia una o más bases**, produciendo una secuencia de aminoácidos diferente, es una variante no sinónima o *missense variant*.
- **La variante provoca la aparición de un codón de parada prematuro**, produciendo un transcripto de menor longitud (*stop gained*).
- **La variante provoca el desplazamiento del marco de lectura**, habitual cuando la variante es una inserción o delección (*frameshift variant*).

Adicionalmente, podremos preguntarnos cuestiones como si la variante ha sido reportada previamente, qué frecuencia alélica tiene en población general, qué impacto tiene en la función de la proteína y en su estructura, si existen enfermedades asociadas a dicha variante o si se encuentra en una región alta o bajamente conservada. En este capítulo veremos la anotación con las bases de datos VEP y Annovar.



Enlaces de interés

En los siguientes enlaces encontrarás bases de datos que proporcionan información sobre distintos aspectos de las variantes encontradas.

Descripción de todos los tipos de consecuencia, descritos por ENSEMBL, donde se categorizan por orden de importancia de la consecuencia que provocan.

https://m.ensembl.org/info/genome/variation/prediction/predicted_data.html

Estudio de la relación del gen con distintas enfermedades según la base de datos OMIM (NCBI). Esta base de datos incluye enfermedades de base genética y ofrece información sobre su manifestación genotípica.

<https://www.omim.org/>

La base de datos ClinVar (NCBI) proporciona un repositorio de relación entre las variaciones en humano y los fenotipos observados.

<https://www.ncbi.nlm.nih.gov/clinvar/>

4.4.1. Variant effect predictor (VEP)

Esta es una herramienta de software gratuita que se puede ejecutar *on line* o bien por vía de comandos, así como utilizarlo en el programa R. VEP soporta varias especies y determina el efecto de una variante de tipo SNV, *indel*, CNV o variación estructural en genes, transcritos y secuencias de proteínas. De esta forma nos proporciona la siguiente información:

- Genes y transcritos afectados.
- Localización de la variante (en secuencia codificante, región reguladora, etc.).
- Consecuencia de la variante en la secuencia de la proteína.
- Información sobre si esta variante se ha descrito anteriormente o su frecuencia alélica en el proyecto 1000 genomas.
- Predicción del cambio de aminoácido, mediante la utilización de las herramientas SIFT y Polyphen. Ambas se basan en métodos de aprendizaje supervisado para predecir el nivel de malignidad de una variante. SIFT nos dirá si el cambio de aminoácido afecta a la función de la proteína, basándose en homología de secuencias; mientras que Polyphen realizará una predicción sobre el impacto de una variación de aminoácido en la estructura y función proteica.



Ejemplo

Accedemos vía web a la herramienta VEP:

http://grch37.ensembl.org/Homo_sapiens/Tools/VEP

Siendo cuidadosos de seleccionar el genoma de referencia correspondiente a nuestro mapeo, indicamos:

>>>

>>>

- *Species*: por defecto humano, cuidando que sea la versión de genoma adecuada (GRCh37 o GRCh38).
- *Name for the job*: Tema4_Ejemplo10.
- Subimos el archivo VCF generado en la identificación de *variants*.
- Bases de datos que utilizar Ensembl/Gencode.
- *Identifiers and frequency data*: podemos añadir identificadores adicionales para genes, transcritos y variantes, así como información relacionada con la frecuencia de las variantes en bases de datos como los 1000 genomas o el de los 6000 exomas. Vamos a seleccionar:
 - Identificador de proteína Ensembl.
 - Frecuencia alélica global de los 1000 genomas.
 - El resto lo dejamos por defecto.
- Opciones extra:
 - Información miscelánea de la variante: biotipo del transcripto (miRNA, secuencia codificante, etc.). Dejamos por defecto.
 - Predicciones acerca de la patogenicidad de la variante. Elegimos SIFT y Polyphen: son dos algoritmos que predicen cómo un cambio de aminoácido afecta a la función de la proteína. Los dejamos activados (predicción y score).
- Opciones de filtrado: podemos añadir opciones para filtrar las variantes. Por ejemplo, por frecuencia de aparición en los 1000 genomas: podríamos eliminar aquellas variantes que son frecuentes en la población y, por tanto, no tienen significado relevante en patogenicidad.

4.4.2. Annotar

Annotar es una herramienta software de línea de comandos gratuita para la anotación funcional de varios genomas estándar. Puedes realizar esta anotación de la misma manera en la página web wANNOVAR, de manera análoga. De esta forma, suministrada una lista de variantes (cromosoma, posición, referencia y nucleótido alternativo), puede darnos la información a tres niveles principales:

- **Anotaciones basadas en gen**: identifica si una variante de tipo SNP, *indel* o CNV causa cambios en la proteína codificada y los aminoácidos que están afectados. En este paso puede elegirse utilizar las bases de datos de genes como RefSeq, UCSC, ENSEMBL o Gencode, entre otras.
- **Anotaciones basadas en región**: identifica variantes en regiones genómicas específicas, por ejemplo regiones conservadas entre especies, lugares de unión de factores de transcripción, etc.
- **Anotaciones basadas en filtros**: identifica variantes que están previamente descritas en bases de datos como 1000 genomas, 6500 exomas, Exome Aggregation Consortium (ExAC), etc.
- **Otras funcionalidades**: identificar una lista de genes candidatos para enfermedades mendelianas a partir de datos de exoma.



Enlaces de interés

En los siguientes enlaces encontrarás la herramienta VEP para anotación de variantes tanto en vía de comandos, como en su versión en R.

<https://www.ensembl.org/info/docs/tools/vep/index.html>

<https://bioconductor.org/packages/release/bioc/html/ensemblVEP.html>

En este enlace encontrarás la herramienta SIFT.

<https://sift.bii.a-star.edu.sg/>

Este enlace te conducirá hasta la herramienta Polyphen.

<http://genetics.bwh.harvard.edu/pph2/>

En este enlace se encuentra la herramienta software Annovar, para la anotación de variantes.

<https://annovar.openbioinformatics.org/en/latest/>

En esta página web puedes utilizar la herramienta wANNOVAR.

<https://wannovar.wglab.org/>

El proceso de anotación de variantes es muy amplio, por lo que existen varios artículos en literatura con revisiones profundas sobre herramientas disponibles, que nos ayudan a guiarnos en el proceso de anotación y correcta interpretación de las variantes. A continuación, se proporcionan los enlaces a algunos de estos artículos:

Whole-genome sequencing in personalized therapeutics (Cordero & Ashley, 2012):

<https://pubmed.ncbi.nlm.nih.gov/22549284/>

A survey of tools for variant analysis of next-generation genome sequencing data (Pabinger et al., 2014):

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3956068/>

Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (Richards et al., 2015):

<https://pubmed.ncbi.nlm.nih.gov/25741868/>

Exome sequencing explained: A practical guide to its clinical application (Seaby et al., 2016):

<https://academic.oup.com/bfg/article/15/5/374/2240049>

Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists (Li et al., 2017):

<https://pubmed.ncbi.nlm.nih.gov/27993330/>

ACGS Best Practice Guidelines for Variant Classification in Rare Disease 2020 (Ellard et al., 2020):

<https://www.acgs.uk.com/media/11631/uk-practice-guidelines-for-variant-classification-v4-01-2020.pdf>



Capítulo 5

Análisis de genomas bacterianos

A lo largo de este capítulo se ahondará en el análisis global de genomas bacterianos, especialmente enfocado desde el punto de vista de la **epidemiología genómica** y la monitorización de bacterias de interés bien por sus características de resistencia, virulencia o patogenia.

La metodología que se describirá será aplicable a cualquier genoma bacteriano que queramos analizar.

5.1. Genoma procariota

Antes de continuar, vamos a recalcar las diferencias significativas que existen entre un genoma eucariota, como el que hemos visto en temas anteriores, y el genoma bacteriano en particular.

5.1.1. Características generales de los genomas procariotas

De manera general, los **procariotas** son organismos cuyas células carecen de compartimentos internos. Diferenciamos dos tipos de procariotas que varían en sus características genéticas y bioquímicas. Las **bacterias** incluyen a la mayor parte de los procariotas como son las bacterias gramnegativas (por ejemplo, *Escherichia coli*), grampositivas (*Bacillus subtilis*), cianobacterias (*Anabaena*), y otras más. Por otra parte, las **Archaea** son un grupo mucho menos estudiado que ha sido encontrado principalmente en ambientes extremos.

De manera general, el genoma procariota típico está contenido en una única molécula de ADN circular localizada en el nucleoide, una región físicamente no separada del citosol. No existe una membrana formando un núcleo separado, a diferencia de lo que ocurre en eucariotas. Esta diferencia principal es muy habitual en un gran número de bacterias, sobre todo aquellas más estudiadas, como *E. coli*. Sin embargo, el conocimiento creciente nos hace cuestionar algunos de estos conceptos. Al igual que ocurre en eucariotas, existe una serie de proteínas nucleares que participan en el empaquetamiento del genoma de una manera organizada con forma superenrollada (*supercoiled*).

A medida que se conocen más genomas procariotas sabemos que la amplia mayoría de bacterias cumplen esta estructura de cromosoma. Sin embargo, cada vez conocemos más estructuras de **cromosomas lineales**, como son los genomas de *Borrelia burgdorferi* (causante de la enfermedad de Lyme), *Streptomyces coelicolor* y *Agrobacterium tumefaciens*. Estas moléculas de ADN lineales tienen extremos libres, con una terminación similar a lo que son los telómeros en los cromosomas eucariotas.

5.1.2. Genomas multipartitos

Podemos destacar que existen procariotas que poseen **genomas multipartitos**, es decir, genomas divididos en dos o más moléculas de ADN. En estos genomas lo complicado radica en distinguir entre lo que corresponde al cromosoma, a lo que corresponde a elementos genéticos independientes como son los plásmidos. Brevemente, un **plásmido** es un elemento genético pequeño, habitualmente circular, que coexiste con el cromosoma principal en la célula bacteriana y que codifica para factores como genes de resistencia a antibióticos, metales pesados o factores de virulencia. Además, estos elementos son transferibles de manera horizontal entre bacterias. Hablaremos en profundidad sobre ellos en un apartado posterior, ya que van a ser una parte fundamental de nuestro estudio bioinformático. La Tabla 7 muestra ejemplos de organización genómica en distintas bacterias.

Tabla 7
Ejemplo de organización genómica en distintas bacterias

Espece bacteriana	Molécula ADN	Tamaño (Mb)	Número de genes
<i>Escherichia coli</i> K12	Cromosoma	4.639	4405
<i>Borrelia burgdorferi</i> B31	Cromosoma linear	0.911	853
	Plásmido circular cp9	0.009	12
	Plásmido circular cp26	0.026	29
	Plásmido circular cp32	0.032	No conocido
	Plásmido lineal lp17	0.017	25
	Plásmido lineal lp25	0.024	32
	Plásmido lineal lp28-1	0.027	32
	Plásmido lineal lp28-2	0.030	34
	Plásmido lineal lp28-3	0.029	41
	Plásmido lineal lp28-4	0.027	43
	Plásmido lineal lp36	0.037	54
	Plásmido lineal lp38	0.039	52
	Plásmido lineal lp54	0.054	76
	Plásmido lineal lp56	0.056	No conocido

>>>

>>>

Especie bacteriana	Molécula ADN	Tamaño (Mb)	Número de genes
<i>Vibrio cholerae</i> El Tor N16961	Cromosoma Megaplásmido	2.961 1.073	2770 1115
<i>Deinococcus radiodurans</i> R1	Cromosoma 1 Cromosoma 2 Megaplásmido Plásmido	2.649 0.412 0.177 0.046	2633 369 145 40

Nota. Adaptado de “Genomes 3”, por T. A. Brown, 2017, *Yale Journal of Biology and Medicine*, 90(4).



Enlaces de interés

Para explorar los genomas procariotas junto con sus características, se puede encontrar la información actualizada en los siguientes enlaces:

<https://www.ncbi.nlm.nih.gov/genome/microbes/>

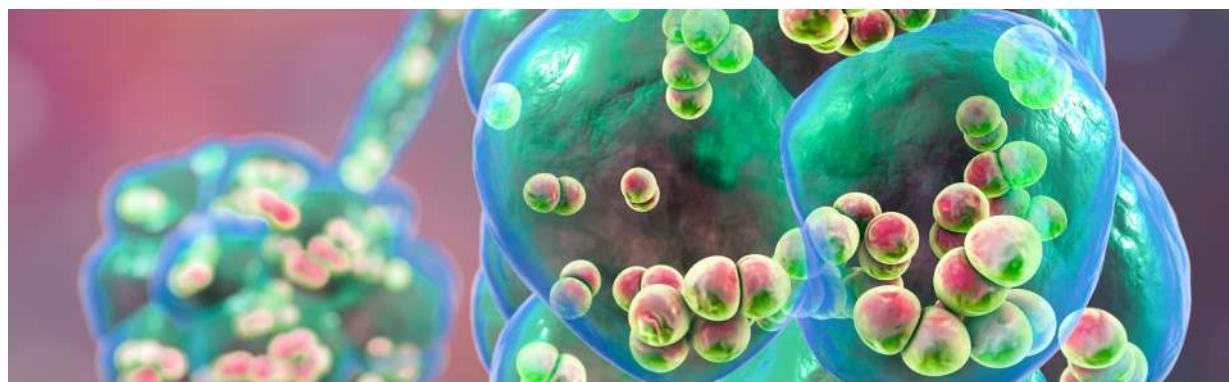
<https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/>

5.1.3. Estructura detallada del genoma procariota

Si nos fijamos en la organización de los genes en los genomas procariotas, encontramos que la característica principal es que los genes no contienen intrones; así como que existe muy poco espacio entre los distintos genes, haciendo un genoma muy compacto. Se estima que solo el 11 % del genoma del organismo modelo *E. coli* K12 es no codificante. Esto permite que la replicación de los genomas procariotas sea muy rápida.

Si centráramos nuestro interés en el genoma de esta bacteria y comparásemos un fragmento de su genoma frente a un fragmento del genoma humano, observaríamos:

- Entre los genes de una bacteria existe menos espacio que entre los genes eucariotas, y que incluso algunos de ellos son solapantes, formando organizaciones llamadas **operones**. Un operón es un grupo de genes que están involucrados en una única ruta bioquímica.
- Los genes bacterianos son más cortos en longitud que los eucariotas, incluso sin tener en cuenta los intrones de estos últimos.
- En general, no existen intrones en los genes bacterianos, aunque existen algunas excepciones en el grupo de las Archaea.
- Aunque en el genoma procariota existen regiones repetidas, como son las secuencias de inserción, estas no tienen la extensión, la frecuencia, ni el alto número de copias de las que se presentan en los genomas eucariotas. Se considera que la mayor parte de los genomas procariotas poseen muy pocos elementos repetidos, como puede ser el genoma de *Campylobacter jejuni* NCTC11168, donde en sus 1.64 Mb de genoma no tiene ningún elemento repetido. Sin embargo, *Neisseria meningitidis* Z2491 tiene más de 3700 copias de 15 tipos diferentes de secuencias repetidas, lo que agrupa un 11 % de su genoma (2.18 Mb).



5.1.4. Número de genes y función general

Los genomas procariotas tienen en general un **tamaño de genoma** menor que un organismo eucariota, aunque el rango entre el genoma procariota más pequeño (*Nanoarchaeum equitans*, 491 kb) y el más grande secuenciado (*Bradyrhizobium japonicum*, 9.1 Mb) es amplio.

En cuanto al **número de genes**, la densidad de genes media es de aproximadamente 950 genes por cada megabase (Mb) de genoma (Tabla 8). Los **genomas más grandes** tienden a pertenecer a especies de vida libre, encontrados en ambientes como suelos o aguas, donde necesitan un arsenal de funciones para hacer frente a situaciones físicas y biológicas muy cambiantes (p. ej., *E. coli* ambiental). Por otro lado, los **genomas más pequeños** se piensa que son aquellos de especies que son parásitos obligados, tales como *Mycoplasma genitalium* (Tabla 9). Esta capacidad limitada de codificar funciones en estos genomas pequeños queda suplida por la obtención de nutrientes de sus hospedadores o incluso de sus compañeros microbianos en ese ambiente. De acuerdo con este tipo de análisis se ha establecido que el rango de genes indispensables para la vida procariota está en torno a 250-350, dependiendo de la especie.

Tabla 8

Tamaños de genoma y número de genes para varios procariotas

Especie	Tamaño de genoma (Mb)	Número de genes (aproximado)
Bacteria		
<i>Mycoplasma genitalium</i>	0.58	500
<i>Streptococcus pneumoniae</i>	2.16	2300
<i>Vibrio cholerae</i> El Tor N16961	4.03	4000
<i>Mycobacterium tuberculosis</i> H37Rv	4.41	4000
<i>Escherichia coli</i> K12	4.64	4400
<i>Yersinia pestis</i> CO92	4.65	4100
<i>Pseudomonas aeruginosa</i> PAO1	6.26	5700
Archaea		
<i>Methanococcus jannaschii</i>	1.66	1750
<i>Archaeoglobus fulgidus</i>	2.18	2500

Nota. Adaptado de “Genomes 3”, por T. A. Brown, 2017, *Yale Journal of Biology and Medicine*, 90(4).

Tabla 9Catálogo parcial de genes de *Escherichia coli K12* y *Mycoplasma genitalium*

Categoría	Número de genes	
	<i>E. coli</i> K12	<i>M. genitalium</i>
Total de genes codificantes	4288	470
Biosíntesis de aminoácidos	131	1
Biosíntesis de cofactores	103	5
Biosíntesis de nucleótidos	58	19
Proteínas de envuelta celular	237	17
Metabolismo energético	243	31
Metabolismo intermediario	188	6
Metabolismo de lípidos	48	6
Replicación, recombinación y reparación de ADN	115	32
Plegamiento de proteínas	9	7
Proteínas regulatorias	178	7
Trascipción	55	12
Traducción	182	101
Captación de moléculas del ambiente	427	34

Nota. Adaptado de “Genomes 3”, por T. A. Brown, 2017, *Yale Journal of Biology and Medicine*, 90(4).

5.2. Epidemiología genómica. Retos y oportunidades en el análisis de genomas bacterianos

Las técnicas de secuenciación de genoma completo es actualmente la herramienta más potente en análisis de genomas bacterianos con fines epidemiológicos. La epidemiología genómica es la disciplina que ha evolucionado desde la epidemiología molecular, y que nos permite utilizar el genoma completo de un organismo como marcador epidemiológico de su evolución y trazado. De esta manera, el análisis de los genomas completos nos permite trazar la transmisión de los microrganismos con la mayor resolución posible.

Aunque ha sido la pandemia de SARS-CoV-2 la que ha mostrado al mundo el potencial de estas técnicas, la epidemiología genómica ya ha sido ampliamente utilizada para la monitorización de patógenos de interés clínico, como son bacterias resistentes a los antibióticos (Comas, 2017; Deng *et al.*, 2016).



Enlace de interés

En el siguiente vídeo, elaborado por el Centro para la Prevención y el Control de Enfermedades (CDC) de EE. UU. se explican conceptos básicos de epidemiología genómica.

<https://www.youtube.com/watch?v=njSOTw4uQUo>

La secuenciación del genoma completo de los patógenos nos ha ayudado a resolver brotes epidémicos, al otorgar un alto poder discriminatorio para diferenciar aislados estrechamente relacionados, así como a rastrear las tendencias epidemiológicas mediante la monitorización de resistencia a antibióticos, como eje central en la epidemiología genómica bacteriana.

El desarrollo de la epidemiología genómica ha venido de la mano del desarrollo en las tecnologías de secuenciación masiva. El primer genoma bacteriano secuenciado completamente fue el de *Haemophilus influenzae* Rd en 1995, mediante tecnología Sanger (Fleischmann *et al.*, 1995). Desde ese momento y hasta la actualidad (25 noviembre 2021), existe en la base de datos RefSeq de National Center for Biotechnology Information (NCBI) un total de 372 481 genomas procariotas, 46 594 genomas víricos y 34 223 plásmidos, a distintos niveles de ensamblaje y anotación.



Enlace de interés

En este enlace tenéis acceso a la base de datos RefSeq de NCBI.

<https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/>

No cabe duda de que esta explosión en el número de genomas ha venido dada por el rápido avance en el desarrollo de tecnologías de secuenciación como Illumina, SOLiD-Ion Torrent o Roche 454. Estas tecnologías, actualmente siendo el mayor exponente la tecnología Illumina, tienen una limitación fundamental: las lecturas de secuenciación cortas no son suficientes para cubrir los elementos repetidos, como operones ARN ribosómicos o transposones multicopia. Consecuentemente, los genomas que están ensamblados a partir de este tipo de lecturas están fragmentados y consisten en decenas o centenas de fragmentos de ADN, llamados *contigs*.

Aquellas tecnologías, como Oxford Nanopore o PacBio, que producen lecturas de secuenciación más largas pueden resolver estas repeticiones y permiten, en combinación con lecturas cortas de mayor resolución, ensamblajes híbridos menos fragmentados. Sin embargo, como punto negativo a estas tecnologías está aún su alta tasa de error, que nos dificulta la identificación y tipificación genómica con detalle, como veremos en apartados posteriores.

5.3. Tipificación de genomas bacterianos

Tradicionalmente, el subtipado de bacterias con fines de trazado de brotes se ha realizado mediante técnicas moleculares como la determinación de perfiles de campos pulsados (*pulse field gel electrophoresis*, PFGE) o secuenciación de locus genéticos (*multi locus sequence typing*, MLST). En los últimos años se ha impuesto el análisis de genoma completo sobre estos métodos que analizan tan solo una parte del genoma, siendo ya una rutina en países como Reino Unido, Dinamarca, Francia, EE. UU. y Canadá (Álvarez-Molina *et al.*, 2021; Jagadeesan *et al.*, 2019).

Para determinar la similitud genética entre patógenos se pueden seguir dos aproximaciones: basado en polimorfismos de un único nucleótido (*single nucleotide polymorphism*, SNP) o un análisis gen por gen. Finalmente, podríamos considerar que el análisis de identidad nucleotídica promedio (*average nucleotide identity*, ANI), es un método de tipado e identificación bacteriano.

5.3.1. Análisis basado en polimorfismos de un único nucleótido (SNP)

Esta técnica se basa en el mapeo o alineamiento de las lecturas secuenciadas sobre un genoma de referencia bien conocido (resecuenciación). Las diferencias nucleotídicas entre cada par de genomas se analizan como variantes de secuencia, incluyendo cambios nucleotídicos, inserciones y delecciones. Estos cambios se utilizan para cuantificar la relación genética y por tanto para inferir una relación filogenética.

Aunque este tipo de análisis es muy potente, su debilidad y su punto clave es la selección del genoma de referencia adecuado, que debe estar completamente secuenciado y estar genéticamente relacionado con los genomas de estudio, para no infraestimar la similitud genética entre ellos. Una de las mayores debilidades de esta técnica es la baja reproducibilidad entre laboratorios diferentes, ya que no solo debe usarse el mismo genoma de referencia y el mismo protocolo bioinformático, sino también los mismos parámetros de mapeo y llamada de variantes. Además, la interpretación de un árbol filogenético para evaluar un brote puede ser complejo y la metodología requiere pericia en el manejo de datos bioinformáticos.



Una buena guía de interpretación de análisis filogenéticos se encuentra en la referencia Pightling *et al.* (2018); adicionalmente, si queréis ver un ejemplo de cómo el genoma de referencia determina el resultado del análisis de un brote epidémico, podéis consultar el artículo de Francés-Cuesta *et al.* (2021) y Valiente-Mullor *et al.*

A diferencia del caso que vimos en el apartado de genomas humanos, donde los protocolos de análisis de SNP están bien definidos, en genomas bacterianos aún no hay un consenso. En este punto, los mapeadores más ampliamente utilizados son Bowtie2 (Langmead & Salzberg, 2012) o BWA (Li & Durbin, 2010), en combinación con programas de llamada de variantes como SAMtools/bcftools (Li *et al.*, 2009) o Freebayes (Garrison & Marth, 2012). Existen algunos protocolos integrados desarrollados como Snippy (Seemann, s. f.), uno de los más populares.

Otro de los métodos que se basan en la identificación de SNP conlleva generar alineamientos del **genoma core** o del **pangenoma**. En este caso primero hay que identificar los genes ortólogos o bloques de secuencias, alinearlos e identificar las localizaciones variables. En este punto podemos analizar regiones altamente variables que son sospechosas de recombinación. Este proceso puede llevarse a cabo con Snippy (Seemann, s. f.), Parsnp/Harvesttools (Treangen *et al.*, 2014) o con Roary (Page *et al.*, 2015).

Para completar la información sobre los programas utilizados en el tipado genético podéis consultar la Tabla 3 del artículo Álvarez-Molina *et al.* (2021).

5.3.2. Análisis gen por gen

Por otra parte, el análisis gen por gen (*gene-by-gene*) se basa en la aproximación tradicional de tipado de genes multilocus (*Multi Locus Sequence Typing*, MLST). En esta técnica se selecciona un grupo pequeño de genes (tradicionalmente siete genes), que son considerados *housekeeping*, es decir, que cumplen funciones vitales. Se amplifican mediante reacción en cadena de la polimerasa (PCR) y se secuencian mediante técnica Sanger.

El esquema inicial de 7 *loci*, definido para varias especies bacterianas individualmente, se ha extendido para un esquema ribosomal (rMLST), un esquema core-genome (cgMLST), de genoma completo (wgMLST) y de genoma accesorio (agMLST). Estos incluyen más regiones variables esenciales o accesorias para cada especie.

El punto fuerte de estos esquemas de tipado es que no dependen de la selección de un genoma de referencia cercano, sino que los alelos definidos para cada gen son fácilmente accesibles en bases de datos públicas. En el caso de la utilización de datos de genoma completo, las lecturas se pueden ensamblar previamente (como veremos en los apartados posteriores), para posteriormente localizar estos genes y tipificarlos mediante un alineamiento frente a la base de datos de referencia (BLAST). Estas bases de datos permiten la estandarización y comparación entre diferentes laboratorios, además implementados en soluciones comerciales de uso sencillo.

Enlaces de interés

En el siguiente enlace se encuentra la base de datos pública de MLST:

<https://pubmlst.org/>

En el siguiente vídeo se explica la tipificación MLST:

<https://www.youtube.com/watch?v=GMv5sFWRqFU>

5.3.3. Identidad nucleotídica promedio (*average nucleotide identity, ANI*)

Una de las cuestiones fundamentales en microbiología es conocer si existe un continuo de diversidad genética entre las distintas especies o, si en caso contrario, prevalecen unos límites claros entre especies. En este punto, medir la similitud del genoma completo en forma de identidad nucleotídica promedio (ANI) ayuda a abordar esta cuestión y facilita el análisis taxonómico de alta resolución de miles de genomas de diversos linajes filogenéticos. Esta medida, dada como porcentaje de similitud, nos permite comparar genomas completos dos a dos y generar matrices de distancias y árboles filogenéticos. El límite de especie se ha determinado en diversos estudios (Goris *et al.*, 2007; Jain *et al.*, 2018; Kim *et al.*, 2014) en un valor de ANI > 95 % la definición de especie (Figueroa *et al.*, 2014).

Actualmente, existen herramientas prediseñadas para el análisis de genomas completos a partir de sus ensamblajes como son FastANI (Jain *et al.*, 2018), JSpecies (Richter & Rosselló-Móra, 2009) u OrthoANI (Lee *et al.*, 2016).

Enlaces de interés

A continuación, algunos vídeos donde se explica el cálculo de valores ANI y su interpretación.

<https://www.youtube.com/watch?v=zNsetoW7USc>

<https://www.youtube.com/watch?v=FDmJALmmPck>

5.4. Reconstrucción del genoma: ensamblaje de novo

El ensamblaje *de novo* es la técnica que consiste en unir fragmentos de secuencias para reconstruir el genoma de un organismo, que ha sido previamente fragmentado y secuenciado. Hasta el momento, lo que habíamos visto en esta asignatura era la resecuenciación, la secuenciación basándonos en un molde o referencia sobre la cual reconstruir el genoma. El ensamblaje *de novo* no toma referencia ninguna para reconstruir el genoma, lo que podría asimilarse a construir un puzzle de millones de pequeñas piezas sin tener la fotografía final como referencia.

Este proceso no resulta sencillo, debido fundamentalmente a los problemas que genera la fragmentación: cuanto más fragmentado haya sido el genoma del organismo, más difícil es reconstruirlo. Si volvemos a la analogía con un puzzle, imaginaros que tenemos 30 millones de piezas que, a diferencia de los puzzles comerciales, sería difícil de encajar porque:

1. Pueden faltar piezas (debido a errores en secuenciación).
2. Existe un gran número de piezas iguales (regiones repetidas).
3. Algunas de las piezas contienen errores (errores de secuenciación).

Para paliar algunos de estos problemas, los ensambladores necesitan entre 5-10 copias de cada pieza para poder ensamblar, lo que sería tener una cobertura media de entre 5 a 10x.

El proceso de ensamblaje no está exento de errores y tiene varias explicaciones:

- Habitualmente descartamos fragmentos que pensamos contienen errores, en ocasiones de manera incorrecta, con lo que perdemos regiones secuenciadas.
- La unión de fragmentos en el sitio incorrecto o en orientación equivocada, formando regiones quimera. La única manera en la que podríamos paliar este tipo de problemas es tener lecturas muy largas y sin errores. Esto es parcialmente posible hoy en día, ya que las tecnologías de lecturas largas (PacBio y Oxford Nanopore) cada vez contienen menos errores, pero aún superan con creces los errores que comete Illumina.

Por eso, para concluir un ensamblaje con éxito: menor número de errores y maximizando la longitud ensamblada, aún requiere de ensamblajes híbridos entre lecturas largas y cortas.

5.4.1. Genotecas para el ensamblado de novo

Antes de entrar de lleno en el ensamblaje computacional, debemos tener en cuenta el tipo de librerías que podemos construir para nuestro genoma. En la Tabla 10, se muestran el tipo de librerías junto con sus ventajas, inconvenientes y tipo de plataforma en la que se aplican. Hoy en día, para las librerías de tipo Illumina *paired-end* se puede seleccionar distintos protocolos que se diferencian en dos cuestiones: fragmentación del ADN de partida (fragmentación enzimática o fragmentación mecánica) y enriquecimiento (enriquecimiento por PCR o libres de enriquecimiento).

Tabla 10

Tipos de librerías utilizadas para la reconstrucción de un genoma

Tipo de librería	Ventajas	Inconvenientes	Plataforma
Lecturas cortas <i>single-end</i>	Aportan profundidad, lo que determina mayor fiabilidad y precisión. Hoy en día están en desuso.	No sirven para reconstruir grandes regiones del genoma. Propensas a artefactos de ensamblaje.	Illumina y SOLiD
Lecturas cortas emparejadas/ <i>paired-end</i>	Dan profundidad al ensamblaje, pero además es mucho más exacto al sortear los elementos repetidos de longitud menor a la del inserto entre las dos lecturas.	No se resuelven la mayor parte de las repeticiones, ya que los insertos suelen ser pequeños (< 500 pb).	Illumina y SOLiD

>>>

>>>

Tipo de librería	Ventajas	Inconvenientes	Plataforma
Lecturas cortas conjugadas (<i>mate-pair</i>)	Muy usadas en conjunto con los dos tipos anteriores, ya que permiten reconstrucción de fragmentos mayores. Incorporan insertos de unos 10 kb.	No aportan profundidad al ensamblaje, son bastante costosas y no siempre se obtienen buenos resultados.	Illumina
Lecturas largas	Resuelven las repeticiones más largas, ya que las lecturas suelen ser de entre 10-50 kb.	Tasa de error de secuenciación muy alta. No aportan profundidad.	PacBio y Oxford Nanopore

5.4.2. Conceptos generales del ensamblaje

Un ensamblaje clásico suele estar compuesto por varios conjuntos de lecturas sueltas o emparejadas. La **unión de lecturas sueltas (single-end) por solapamiento** de una secuencia continua llamada ***unitig***. Los ***unitig*** a su vez, **pueden unirse entre ellos formando secuencias continuas** llamadas ***contigs***. Las secuencias repetidas, los polimorfismos, los fragmentos perdidos y los errores provocan el acortamiento de la longitud de los ***contigs*** resultantes.

Habitualmente utilizamos el término ***contig*** directamente, para hacer referencia a una secuencia continua. La diferencia entre ***unitig*** y ***contig*** radica en cómo se generan dentro del algoritmo de ensamblaje. Es una diferencia sutil: el ***unitig*** ha sido creado con la certeza de que es correcto porque es la única posibilidad, mientras que el ***contig*** se obtiene tras decidir entre varias opciones según el criterio del ensamblador.

Por otro lado, el uso de lecturas emparejadas o conjugadas (*mate-pair*) es extremadamente útil en las últimas fases del ensamblaje. Cada par de estas lecturas se encuentra a una distancia conocida, que varía en función del protocolo de preparación de la genoteca empleado (200 pb hasta decenas de kilobases). Como las parejas de lecturas se generaron a partir del mismo fragmento de ADN, la formación de ***contigs*** se simplifica y nos permite unir estas estructuras en ***scaffolds***. Los ***scaffolds*** son un conjunto de ***contigs*** ordenados entre sí, con huecos (***gaps***) de tamaño conocido entre ellos.

5.4.3. Tipos de algoritmos de ensamblaje

Existen tres tipos de algoritmos de ensamblaje *de novo* (Miller et al., 2010):

1. **Algoritmo de solapamiento-composición-consenso (OLC, overlap layout consensus)**, que fueron los primeros en utilizarse. Son muy intuitivos, pero computacionalmente muy costosos, por lo que no son apropiados para las lecturas brutas de NGS. Siguen teniendo su campo de aplicación, dada la enorme fiabilidad de sus resultados.
2. **Algoritmos “voraces” (greedy)**. Aparecieron con los primeros datos de NGS, cuando las lecturas eran muy cortas, de entre 25-35 pb, y no se podían aplicar los algoritmos OLC. Debido a su “voracidad”, se comen, como mínimo, las secuencias repetitivas. Hoy en día ya no se utilizan.
3. **Algoritmos basados en grafos de De Bruijn (DBG)**. Son los más utilizados hoy en día porque son muy eficaces computacionalmente.

Vamos a ahondar en los algoritmos OLC y De Bruijn, por ser los más comunes en ensamblaje *de novo*, incluyendo una breve reseña sobre teoría de grafos.

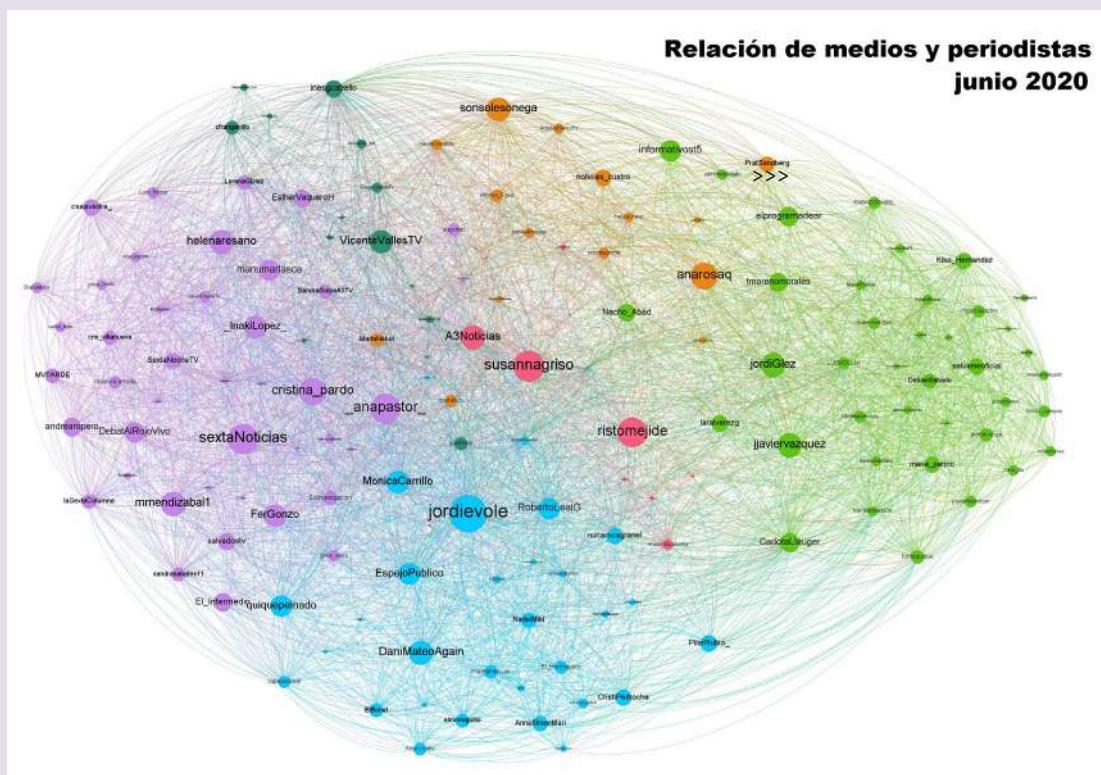


Breve reseña sobre la teoría de grafos

Los grafos son una estrategia muy utilizada en informática para el desarrollo de algoritmos. Cada grafo se define por un conjunto de nodos y de enlaces (aristas) entre ellos. Si estos enlaces tienen dirección asignada, hablamos de grafos dirigidos. Estos enlaces forman caminos (*paths*), que son los recorridos entre nodos unidos por enlaces. Un camino simple es el que solo tiene nodos distintos. En la Figura 23 se muestra un ejemplo de grafo, donde los nodos son círculos y los enlaces son las líneas que los unen. Un grafo puede representar cualquier concepto, como por ejemplo en esta figura, los nodos representan periodistas o medios de comunicación y los enlaces las interacciones que se dan entre ellos.

Figura 23

Ejemplo de grafo



Nota. Recuperado de https://pbs.twimg.com/media/EafV_ThXgAAD9JT?format=jpg&name=4096x4096

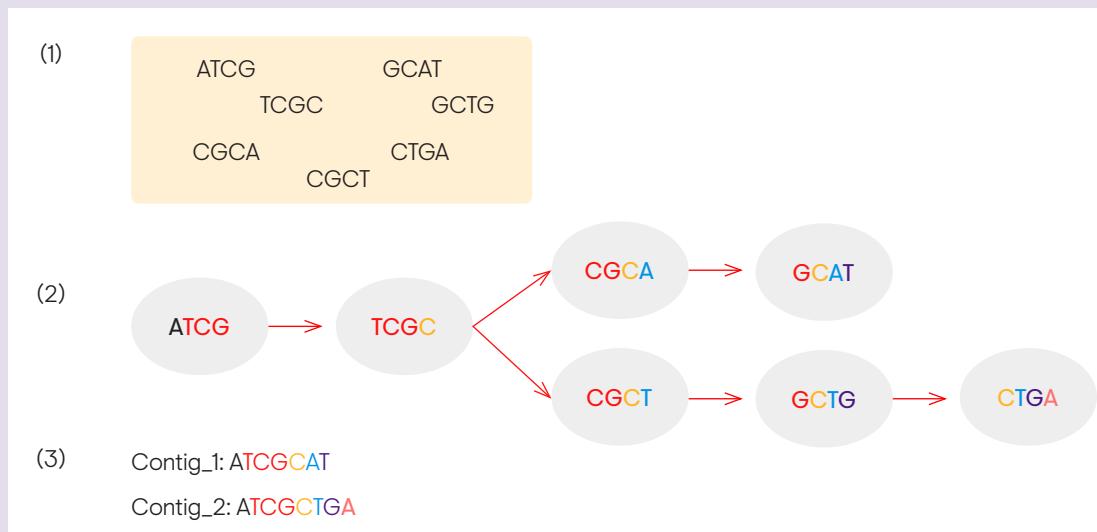
En el ensamblaje de novo se utilizan grafos de solapamiento, donde los nodos son secuencias y los enlaces son solapamientos entre las secuencias. Este es el concepto básico, ya que en la práctica el grafo tiene atributos para diferenciar entre el extremo 3' y 5' de la secuencia, si es secuencia directa o complementaria, la longitud de la lectura y la longitud del solapamiento, entre los principales. Un camino a través del grafo sería un posible *contig*. Cada lectura se divide en k -meros de longitud menor (k) y con un solapamiento de longitud $k-1$ (Figura 24).

>>>

>>>

Figura 24

Grafo de solapamiento sencillo, construido a partir de secuencias de longitud 4 (k -mero = 4) y un solapamiento de $k-1$ nucleótidos



Nota. (1) Dado un conjunto de secuencias, de longitud $k = 4$, se construye el grafo de solapamiento. (2) El resultado del grafo con un solapamiento $k-1$, en este caso 3. (3) Para resolver el grafo, recorremos todos los caminos posibles, donde se extraerán los *contigs*.



Enlaces de interés

En los siguientes vídeos podéis ahondar en la teoría de grafos:

¿Qué son los grafos?

<https://www.youtube.com/watch?v=wKeg6tOG7ql>

Conceptos básicos de la teoría de grafos:

<https://www.youtube.com/watch?v=pzca71UtH-A>

Ensamblaje genómico:

<https://www.youtube.com/watch?v=dyGuXMyQEy8>

[https://www.youtube.com/watch?v=OY9Q_rUCGDw.](https://www.youtube.com/watch?v=OY9Q_rUCGDw)

Algoritmos OLC

El proceso se divide en tres pasos (Miller *et al.*, 2010):

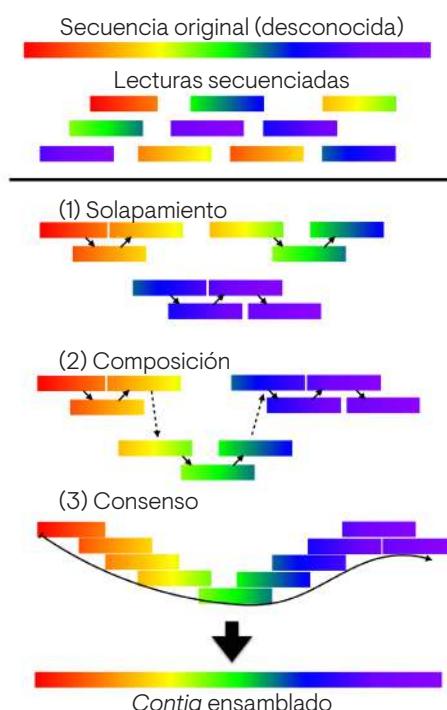
1. **Solapamiento.** Se buscan solapamientos entre las lecturas comparándolas todas contra todas para saber qué lecturas solapan entre sí. El proceso es computacionalmente muy intensivo, puesto que se dispara el número de comparaciones con cada lectura que entra en juego ($2n$). Lo habitual es buscar lecturas que solapen como mínimo 40-60 pb en sus extremos (por eso no eran útiles en las primeras lecturas obtenidas, que no llegaban a esta longitud).

Un solapamiento considerado aceptable sería el perfecto (100 %), pero debido a polimorfismos y errores de secuenciación, se considera aceptable si ronda el 95-99 % de la región solapante. Por esto, parámetros como la longitud mínima del solapamiento y el porcentaje mínimo de identidad requerido en dichos solapamientos son dos parámetros configurables y de importancia en este paso. Tras la selección de los mejores solapamientos se genera un grafo (Figura 25-1). Originalmente, muchos de los programas OLC utilizaban versiones optimizadas del algoritmo de Smith-Waterman, para alineamientos locales, que dividía cada lectura en palabras, lo que sería equivalente a los k -meros en teoría de grafos.

2. **Composición.** El grafo resultante de los solapamientos no necesita incluir todas las secuencias de partida, sino tan solo un conjunto representativo (Figura 25-2). Los caminos extraídos de este grafo darán lugar a los contigs, creciendo con el número de lecturas. Puede volverse inmanejable y complicado. Cada uno de estos caminos en busca de los contigs se denomina camino de Hamilton o hamiltoniano.
3. **Consenso.** Una vez obtenidos los contigs, hay que incorporar las secuencias que no formaron parte del grafo para corregir los posibles errores que tengan las lecturas previamente seleccionadas. Esto se consigue con un alineamiento múltiple de los *contigs* y las lecturas de partida para establecer el consenso (Figura 25-3). Debemos ser cuidadosos y conscientes de que este paso puede generar secuencias químéricas que en realidad no existen.

Figura 25

Proceso de reconstrucción mediante un algoritmo OLC, desglosado en sus etapas:
(1) solapamiento, (2) composición y (3) consenso



Nota. Adaptada OLCOverlap-Layout-consensus.png, por Dr.who what what, 2017, [Imagen]. Bajo licencia CC BY-SA 4.0. Recuperado de <https://commons.wikimedia.org/wiki/File:OLCOverlap-Layout-consensus.png>

Los algoritmos OLC son los más precisos y consiguen mejores resultados, pero solo pueden ser usados con lecturas largas y de gran calidad, básicamente restringiéndose a lecturas Sanger o **pirosecuenciación** 454 de Roche (discontinuada en 2016). No son algoritmos recomendables para secuencias Illumina por ser cortas, ni para PacBio u Oxford Nanopore, por la elevada tasa de error. Como ejemplo de ensambladores basados en OLC están los clásicos CAP, Newbler y MIRA (Miller *et al.*, 2010).

Algoritmos basados en grafos de De Bruijn (DBG)

Estos algoritmos surgieron para las lecturas cortas de las plataformas SOLiD e Illumina, pero hoy en día son de uso universal, por ser muy eficaces desde el punto de vista computacional. Su eficacia computacional se debe a que se basan en grafos de k -meros que no requieren una comparación de todos contra todos, a diferencia de los algoritmos OLC. A cambio, no se mantienen las secuencias originales, lo que provoca que no sean tan precisos.

Otra diferencia fundamental es que los grafos de De Bruijn hay que resolverlos mediante caminos de Euler (estrategia euleriana) que consiste en buscar el camino que cubre todo el grafo y pasa por cada nodo una única vez. Los polimorfismos y los errores de secuenciación, así como las regiones repetidas, hacen que no se pueda encontrar un único camino y que el ensamblador se detenga en ciertas bifurcaciones.

Las etapas por las que pasa un ensamblador de este tipo son (Figura 24):

- 1. Fragmentación.** Se fragmentan las lecturas en k -meros de longitud (k) constante y solapantes de nucleótido en nucleótido. Esto evita comparar todas las lecturas contra todas ellas. k suele oscilar entre 19 pb y la longitud total de la lectura. Cuanto más pequeño es k , más computación es necesaria; mientras que a mayor k -mero, más fiable es el solapamiento y menos sobrecarga computacional, siendo más sensibles a los errores. Hay que ser cautos porque muchos ensambladores admiten un margen pequeño de k -meros y algunos incluso tenemos que compilarlos para cambiar el único valor de k -mero que admiten.
- 2. Construcción del grafo** en el que se conectan los k -meros que solapan $k-1$ nucleótidos entre sí. Al reducir a un k -mero todas las lecturas que lo contienen, se suprime la redundancia, con lo que el grafo se reduce y simplifica, volviéndose más manejable.
- 3. Búsqueda de caminos eulerianos** que reconstruyan la secuencia original.

El ejemplo de reconstrucción de este tipo está esquematizado en la Figura 24. La construcción del grafo se realiza rápidamente con la búsqueda de los k -meros en una tabla *hash*. Una tabla *hash* o matriz asociativa, mapa *hash*, tabla de dispersión o tabla fragmentada, es una estructura de datos, muy habitual en programación, que asocia claves con valores.

Los factores que complican el ensamblaje a partir de grafos basados en k -meros son:

- Las repeticiones directas generan bucles en los grafos de los k -meros, por lo que impiden una única reconstrucción de la secuencia original.
- El ADN tiene dos cadenas, por lo que la secuencia de una lectura seguro solapa con su complementaria. Para solventar este problema, los ensambladores incorporan en cada nodo enlaces para las dos cadenas, para impedir ensamblar la secuencia dos veces.
- Las repeticiones complejas (tándem, invertidas, imperfectas, etc.) generan estructuras en el grafo altamente ramificadas, complejas y difíciles de resolver, que solo se pueden solventar con k -meros superiores a la longitud de la repetición, y eso por ahora solo es posible con tecnología PacBio u Oxford Nanopore.
- Los errores de secuenciación se intentan soslayar con un preprocessamiento de las lecturas para eliminarlos y ponderarlos en los enlaces del grafo, en función del número de lecturas que los soportan.

Los ejemplos más conocidos de ensambladores de este tipo, utilizados actualmente son Velvet (Zerbino, 2010; Zerbino & Birney, 2008) y SPAdes (Bankevich *et al.*, 2012).

5.4.4. Evaluación previa al ensamblaje. Análisis de *k*-meros

Previo al ensamblaje y posterior al análisis de calidad de las lecturas mediante FastQC y limpieza de adaptadores y regiones de baja calidad con programas como Trimmomatic, TrimGalore o FastP, podemos realizar un análisis de *k*-meros para localizar posibles errores o contaminaciones de otros genomas, vectores, adaptadores. Este es un paso opcional, que se encuentra incluido en algunas versiones de programas como FastQC y FastP.



Ejemplo

Imaginemos una secuencia de nucleótidos de 50 pb que se ha generado en la secuenciación masiva, que vamos a dividir en *k*-meros de 4 nucleótidos (con lo que la longitud que utilizaremos de solapamiento será de $k-1 = 3$ pb). Para ello, recorremos la secuencia con ventanas de longitud 4 desplazadas nucleótido a nucleótido, con lo que serán solapantes. Vemos un ejemplo en la Figura 26. Cada *k*-mero se guarda en una tabla hash, junto con el número de veces que aparece ese *k*-mero en la secuencia. Si representamos el número de veces que aparece un *k*-mero en la secuencia, vemos que en este ejemplo tan sencillo, la mayoría de ellos aparecen una única vez, mientras que solo unos pocos aparecen dos o más veces.

Figura 26

Ejemplo de extracción de *k*-meros de longitud 4 mediante ventanas solapantes en $k-1$ nucleótidos

ATCGCTATGTTGAATAGCTTATCGATGGATCAATCGTATCGATCCTGG

ATC	6
TCG	4
GAT	3
TAT	3
AAT	2
ATG	2
CGA	2
CTT	2
GCT	2
TGG	2
AGC	1
ATA	1
CAA	1
CGC	1
CGT	1
CTA	1
CTG	1
GAA	1
GGA	1
GGG	1
GTA	1
GTT	1
TAG	1
TCA	1
TCC	1
TGA	1
TGT	1
TTA	1
TTG	1
TTT	1



Este tipo de curvas se mantiene en los datos reales, de lo que deducimos que la generación aleatoria de lecturas representará con fidelidad el contenido de la secuencia original. Todas las librerías de un mismo genoma deberían presentar el mismo espectro de k -meros, y el genoma ensamblado también. Cualquier desviación de esta situación nos ayudará a estimar la representatividad de las librerías y del ensamblaje con respecto a la secuencia original que queremos reconstruir.



Enlaces de interés

Aunque existen varias herramientas bioinformáticas que nos proporcionan análisis de k -meros la más indicada para realizarlo es KAT.

Disponible en el siguiente enlace:

<https://github.com/TGAC/KAT>

Con una documentación extensa en:

<https://kat.readthedocs.io/en/latest/>

Y DSK:

<https://github.com/GATB/dsk>

<https://gatb.inria.fr/software/dsk/>



Ejemplo

En este ejemplo vamos a descargar unas lecturas de la base de datos Bioproject, para posteriormente analizar su calidad, calcular su cobertura teórica, limpiar adaptadores y regiones de baja calidad y analizar su contenido en k -meros.



Enlace de interés

Las secuencias se han obtenido de la web:

<https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR17002549>



Descarga de archivo

En *Campus virtual > Aula de la asignatura > Recursos y materiales > Ejemplos en clase > Tema5_Ejemplo1*, podrás encontrar los FASTQ correspondientes a este ejemplo.

1. Descarga de las secuencias desde esta web en una carpeta en nuestra máquina. Si no puedes hacerlo de esa manera, tienes disponibles las lecturas en el aula virtual.
2. Realizamos un FastQC para analizar su calidad `<fastqc *.fastq.gz>`. Respondemos a las siguientes preguntas:
 - a. ¿Cuántas lecturas tiene el genoma secuenciado?
 - b. ¿Qué cobertura teórica tendría este genoma, sabiendo que corresponde a *E. coli* (aprox. 5 Mb)?
 - c. ¿Contiene adaptadores?
 - d. ¿Cuál es la longitud de lectura?
 - e. ¿Consideráis necesario hacer una limpieza previa al ensamblado?

>>>

>>>

3. Realizamos un análisis con FastP.

```
fastp -i input_1.fastq.gz -I input_2.fastq.gz -o out_1.clean.  
fastq.gz -O out_2.clean.fastq.gz --cut_by_quality3 25 --cut_by_  
quality5 25 --cut_mean_quality 25 -l 151 --qualified_quality_  
phred 25 -h out_FastP.html
```

</>

- ¿Qué significan los parámetros que estamos utilizando?
- ¿Cuántas lecturas obtenemos tras el filtrado? ¿Qué porcentaje de bases Q20 y Q30? ¿Cuál es la tasa de duplicación? ¿Qué tamaño estimado de inserto ha detectado? ¿Cuál es el tamaño medio antes de la limpieza? ¿Y después? ¿Cuál es el contenido GC?
- Analicemos el contenido en *k*-meros obtenido con la herramienta FastP. ¿Qué conclusiones puedes sacar al revisar la tabla de *k*-meros?

5.4.5. Ensamblaje de secuencias cortas con SPAdes

En este apartado vamos a ver cómo realizar el ensamblaje de unas lecturas que ya han sido evaluadas en calidad. Para ello utilizaremos el ensamblador SPAdes (Bankevich *et al.*, 2012; Prjibelski *et al.*, 2020), por ser uno de los más utilizados para genomas bacterianos.

Siempre es deseable, antes de lanzarse a utilizar un ensamblador, echar un vistazo a las evaluaciones publicadas donde se compara el rendimiento y resultado de cada uno de los ensambladores, especialmente en genomas que sean similares al nuestro.

En el caso de SPAdes (<https://cab.spbu.ru/software/spades/>), se trata de un kit de ensambladores del tipo de De Bruijn, donde se albergan varios protocolos, que han sido descritos en (Prjibelski *et al.*, 2020), y que contemplan los siguientes, aunque se encuentra en constante evolución:

- **SPAdes:** ensamblador general para genomas (el que utilizaremos).
- **metaSPAdes:** para metagenomas.
- **plasmidSPAdes:** extracción y ensamblaje de plásmidos en secuencias de genoma completo.
- **metaplasmidSPAdes:** extracción y ensamblaje de plásmidos en metagenomas.
- **rnaSPAdes:** ensamblaje de transcriptomas *de novo* a partir de datos de RNAseq.
- **biosyntheticSPAdes:** ensamblaje de clústeres de genes biosintéticos a partir de lecturas pareadas.
- **coronaSPAdes:** ensamblaje *de novo* de SARS-CoV-2.
- **rnaviralSPAdes:** ensamblaje *de novo* para sets de datos de ARN viral (transcriptoma, **metatranscriptoma** y metaviroma).
- **metaviralSPAdes:** ensamblaje de metaviroma.

Las versiones más actuales permiten el uso de lecturas Illumina o IonTorrent, en solitario o combinadas con lecturas PacBio, Oxford Nanopore o Sanger, para realizar ensamblajes híbridos, así como *contigs* adicionales que pueden ser utilizados como lecturas largas. No se recomienda utilizar lecturas largas en solitario, sino en combinación con las cortas. Asimismo, este ensamblador puede utilizar lecturas pareadas, no pareadas y conjugadas, de manera agrupada o sencilla. En todo caso, SPAdes fue **diseñado para genomas pequeños**, como son genomas bacterianos, fúngicos o virus; no se recomienda su aplicación en genomas grandes.

Además de sus distintas formas de ensamblaje, en función del tipo de muestra del que provengan las lecturas, se proporcionan *scripts* para el conteo de *k*-meros (spades-kmercounter), construcción del gráfico de ensamblaje (spades-gbuilder) y el alineamiento de lecturas largas a un grafo (spades-gmapper).

El procedimiento que sigue SPAdes consta de tres pasos:

1. **Corrección de errores.** Este paso es opcional y puede realizarse fuera del protocolo SPAdes, con otras herramientas. SPAdes proporciona dos de ellas integradas: BayesHammer y IonHammer, para lecturas Illumina y IonTorrent, respectivamente. Este es un proceso largo (aprox. 25-30 minutos para un genoma bacteriano de *E. coli*, con 25-30 M de lecturas pareadas de 100 pb).
2. **Ensamblaje.** Este segundo paso es un proceso iterativo, donde se seleccionan unos valores de *k*-mero automáticamente basados en la longitud de la lectura y tipo de lecturas. De este proceso se obtienen un conjunto de *contigs* y de *scaffolds*.
3. **Corrección de mismatches.** Esta herramienta mejora la detección de *mismatches* e *indels* cortos en los *contigs* y *scaffolds* resultantes. Para ello utiliza la herramienta BWA. Aunque esta opción está desactivada por defecto, los autores recomiendan utilizarla.



Enlace de interés

Recordad visitar el manual de SPAdes en el siguiente enlace o bien, en la terminal tecleando `spades.py -h`.

<https://cab.spbu.ru/files/release3.15.3/manual.html>

A continuación, vamos a ver algunos de los parámetros más sensibles que se debe tener en cuenta cuando utilicemos este programa.

- **Opciones básicas.** En opciones básicas es obligatorio determinar una carpeta de salida para los archivos (`-o output_dir`). Opcionalmente, podremos seleccionar alguno de los protocolos “especiales” de SPAdes para el ensamblaje de metagenomas, viroma, ARN o plásmidos, entre otros.
- **Datos de entrada.** Dada la diversidad de datos de entrada que podemos utilizar, SPAdes tiene una larga lista de comandos opcionales en este punto. Para un conjunto de lecturas pareadas Illumina, el método más estándar y sencillo, los indicaremos como `-1 read_1.fastq.gz` y `-2 read_2.fastq.gz`. Las lecturas de ensambladores como PacBio o Nanopore se indican como `-pacbio` y `-nanopore`. Adicionalmente, si tenemos un conjunto de *contigs* ensamblados de un genoma similar (referencia) o del mismo, de ensamblajes anteriores, podemos utilizarlos como “molde” con la opción `-trusted-contigs`.

- Opciones adicionales:

- **Corrección de errores.** Como se ha comentado anteriormente, se puede realizar un primer paso de corrección de errores. Por defecto está habilitado en el protocolo, pero si queremos realizar solo la corrección, sin realizar ensamblaje utilizaremos **--only-error-correction**; o si por el contrario queremos deshabilitar la corrección de errores y solo ensamblar utilizaremos **-only-assembler**.
- **Minimizar mismatches e indels.** Asimismo, la minimización de *mismatches* e *indels* cortos sobre los *contigs* y *scaffolds* finales debe ser habilitada manualmente con la opción **-careful**.
- **Tamaño de *k*-mero.** La opción **-k**, seguida de uno o varios números impares separados por comas, nos permite elegir el tamaño de *k*-mero. Por defecto, SPAdes hace una selección automática de estos valores, utilizando como dato la longitud de la lectura, con un número máximo de 128 nucleótidos. Este programa no permite *k*-meros superiores a este tamaño, aunque en la compilación del programa puede modificarse. Sin embargo, los desarrolladores no lo recomiendan. Este programa determina automáticamente de entre los *k*-meros seleccionados la mejor opción. **Este paso es de extrema importancia.** La determinación del mejor *k*-mero define el ensamblaje. Es imposible saber de antemano el valor idóneo, ya que depende de la calidad, longitud y tipo de lecturas, así como del organismo que estemos secuenciando. Para su elección, lo mejor es ensamblar con varios *k*-meros y evaluar. Debemos tener en cuenta que a mayor valor *k*, menos *contigs* generamos y más largos, pero nos arriesgamos a perder una parte importante del genoma y a forzar ensamblajes híbridos (quimera) no reales. Por otra parte, una *k* baja nos lleva a ensamblajes con más *contigs* y más cortos, con una mayor representación del genoma, pero más errores.
- **Cobertura mínima (-- cov-cutoff):** con este parámetro, por defecto desactivado en el protocolo, se puede controlar la cobertura mínima que debe tener un *unitig* para ser considerado como tal.

Para poner en práctica estos parámetros y ver los archivos de salida que obtenemos con SPAdes, vamos a utilizar un ejemplo.



Ejemplo

En este ejemplo utilizaremos las lecturas generadas en el ejemplo anterior con FastP para realizar el ensamblaje *de novo* de ese genoma bacteriano. Evaluaremos distintos *k*-meros y analizaremos los archivos de salida que nos proporciona el programa.



Descarga de archivo

En Campus virtual > Aula de la asignatura > Recursos y materiales > Ejemplos en clase > Tema5_Ejemplo2, podrás encontrar los FASTQ correspondientes a este ejemplo.

El comando que utilizaremos para SPAdes es el siguiente, utilizando la selección de *k*-meros por defecto (auto):

```
spades.py -1 out_1.clean.fastq.gz -2 out_2.clean.fastq.gz --careful
-t 32 -k auto
```



>>>

>>>

Una vez finalizado el proceso (aprox. 60 minutos) podemos encontrar los siguientes archivos en la carpeta de salida (más información en <https://cab.spbu.ru/files/release3.12.0/manual.html>):

- **spades.log y params.txt:** archivos que nos muestran todo el proceso realizado y los parámetros utilizados. En él podemos encontrar si han existido errores, así como los pasos realizados y las versiones de los programas y dependencias.
- **Lecturas corregidas:** se encuentran en la carpeta *corrected*. En ellas están las lecturas corregidas de errores pareadas y no pareadas.
- **Carpetas K21, K33, K55, K77, K99 y K127:** corresponde al proceso para cada *k*-mero seleccionado. En su interior están los *contigs* ensamblados y los grafos del ensamblaje.
- **Archivo contigs.fasta:** *contigs* ensamblados con el *k*-mero seleccionado como óptimo por el programa.
- **Archivo before_rr.fasta:** archivo de *contigs* antes de resolver repeticiones. No se recomienda trabajar con él.
- **Archivo scaffolds.fasta:** *scaffolds* reconstruidos con el *k*-mero seleccionado como óptimo por el programa. Son los recomendados para trabajar posteriormente.
- **Archivo assembly_graph.fastg, assembly_graph_with_scaffolds.gfa:** grafos de ensamblaje. Pueden ser visualizados en herramientas como Bandage (<http://rrwick.github.io/Bandage/>).
- **Archivos de “path” en el grafo de ensamblaje correspondientes a los contigs y los scaffolds:** *contigs.paths* y *scaffolds.paths*.

5.4.6. Evaluación de la calidad del ensamblaje: continuidad y contenido

Una vez realizado el ensamblaje debemos analizar la calidad de dicho ensamblaje en función de dos criterios:

- **Continuidad:** para valorar la fragmentación del ensamblaje.
- **Contenido:** se estudia la cantidad de información de las lecturas originales que no se ha incorporado al ensamblaje.

Según la **continuidad del ensamblaje**, mediante la cual evaluamos la fragmentación, podemos determinar:

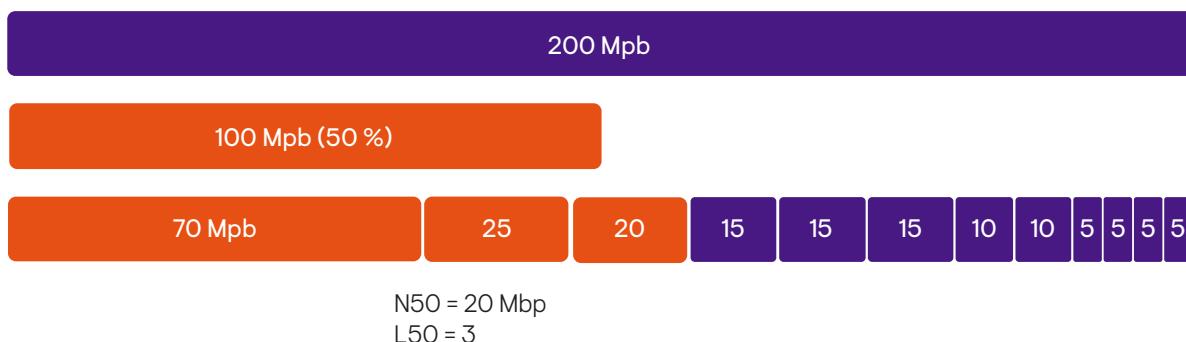
- **Número de bases del ensamblaje.** Aunque sea una métrica muy sencilla, nos da la estimación de cuánto hemos conseguido ensamblar, sobre todo si conocemos la longitud aproximada del organismo de referencia. Una desviación muy por encima de la longitud que esperamos, puede indicar contaminación o bien secuencias inesperadas (como por ejemplo fagos o plásmidos). En cambio, una desviación muy por debajo del número esperado nos indicará que una parte del genoma se ha quedado sin ensamblar.
- **Número de contigs/scaffolds.** Se suele utilizar este número para describir un ensamblaje. Un número muy elevado nos indicará que el ensamblaje está muy fragmentado y que será difícil trabajar con él. Sin embargo, forzar un ensamblaje a que tenga bajo número puede ser contraproducente, ya que podríamos estar forzando a la generación de secuencias quimera.

Quizás lo más importante no sea el número, sino la longitud de dichos *contigs/scaffolds*. No es lo mismo trabajar con 1000 *contigs* de tamaño medio 500 pb, que hacerlo con 1000 *contigs* donde una parte de ellos superen, por ejemplo, los 5000 pb mientras que otros tantos sean pequeños y, por tanto, no nos aporten demasiada información.

- **Valor N50.** Esta métrica es la más utilizada porque es poco sensible a los valores extremos. Nos indica el tamaño del *contig* de forma que el 50 % del genoma ensamblado está soportado por bloques de este tamaño o mayores. Para calcularlo, se ordenan de mayor a menor longitud los *contigs* del ensamblaje, repitiendo cada elemento en la distribución tantas veces como sea su valor. La Figura 27 muestra un ejemplo de su cálculo.
- **Valores N20, N60, N90.** En la misma idea que la métrica N50, pero para los percentiles 20 %, 60 % y 90 %.
- **Valor L50.** Número de *contigs* que comprenden el 50 % del genoma ensamblado.

Figura 27

Cálculo de la métrica N50



Nota. Si tenemos 12 *contigs* de longitudes (Mpb) 70, 25, 20, 15, 15, 15, 10, 10, 5, 5, 5 y 5, para calcular N50 primero los ordenaremos de mayor a menor, calculamos su suma (200 Mpb). Si revisamos qué *contig* sería el que abarca el 50 % del tamaño ensamblado (100 Mpb), sería 20 ($70 + 25 + 20 = 115$ Mpb), por tanto N50 = 20.

Si atendemos al **contenido** podemos evaluar el ensamblaje en función de la cantidad de información de las lecturas originales que no se han incorporado al mismo. Existen dos estrategias:

- **Mapeo de las lecturas.** Se remapean las lecturas de partida sobre el ensamblaje considerado definitivo. La situación ideal sería que todas las lecturas mapeasen sin errores, sin embargo, siempre existen lecturas que no se mapean debido a errores. Esta cantidad nos da una idea aproximada.
- **Evaluación del espectro de *k*-meros.** Para ello se comparan las secuencias incorporadas en el ensamblaje, las totales, y las que no se mapean. Se puede evaluar su contenido en *k*-meros de ambos grupos, de manera que cuanto menos concuerden, más información se estará perdiendo.

La evaluación de los ensamblajes viene dada en muchos casos por el propio ensamblador, que muestra sus estadísticas. En el caso de SPAdes, esto no es así, de manera que podemos utilizar otro programa, como GAGE (Salzberg *et al.*, 2012), Quast (Gurevich *et al.*, 2013), CheckM (Parks *et al.*, 2015), SQUAT (Yang *et al.*, 2019) o GenomeQC (Manchanda *et al.*, 2020), para evaluar todos estos parámetros.



Enlace de interés

Una buena guía para comprender algunos de estos conceptos es el artículo *Do it yourself guide to genome assembly*, de Wajid y Serpedin (2016):

<https://academic.oup.com/bfg/article/15/1/1/1741842>



Ejemplo

En este ejemplo vamos a evaluar la calidad de nuestro ensamblaje. Para ello utilizaremos el programa Quast. Este programa, a diferencia de los demás, no ha podido instalarse en el entorno conda general, sino que necesitamos un entorno con Python 2.7, de manera que lo instalamos de la siguiente manera:

```
conda create -n quast python=2.7
conda activate quast
conda install -c bioconda quast
```



Nos ubicamos en la carpeta donde tenemos nuestros *contigs* y *scaffolds* ensamblados con SPAdes y ejecutamos Quast. En una primera prueba no vamos a utilizar un genoma de referencia para evaluar los parámetros, pero vamos a contemplar todos los *contigs* y *scaffolds* disponibles.

```
quast.py -o quast_result -m 0 -t 32 -k --k-mer-size 127 --circos
--pe1 ../out_1.clean.fastq.gz --pe2 ../out_2.clean.fastq.gz contigs.
fasta scaffolds.fasta
```



Analicemos el resultado:

- ¿Cuántos *contigs* y cuántos *scaffolds* hemos obtenido? De estos, ¿cuántos son mayores de 1kb?
- ¿Cuál es el tamaño del *contig* más largo?
- ¿Cuál es el total de nucleótidos ensamblados? ¿Y en *contigs* mayores de 1 kpb?
- ¿Cuál es el valor de N50? ¿L50?
- ¿Cuál es el contenido GC?
- ¿Cuántas N se encuentran por cada 100 kpb?

5.5. Reconstrucción del genoma: anotación del genoma ensamblado

Tras el ensamblaje de lecturas, el siguiente paso es la anotación del genoma. Este es el proceso de identificar la localización y el papel biológico de los genes encontrados en el ensamblaje. Los programas utilizados para la anotación de los genomas llevan el uso de programas externos que nos ayuden en la predicción de las secuencias codificantes de proteínas (CDS), genes de ARN de transferencia, genes de **ARN ribosomal** y otros, como operones o **elementos CRISPR**.

El programa de anotación clásico de genomas es RAST Server, que ha evolucionado desde una página de anotación web a un protocolo de anotación para múltiples genomas (Aziz *et al.*, 2008; Brettin *et al.*, 2015; Meyer *et al.*, 2008; Overbeek *et al.*, 2014). Hoy en día los protocolos más utilizado son NCBI Prokaryotic Genome Annotation Pipeline (PGAP) (Tatusova *et al.*, 2016) o Prokka (Seemann, 2014). Siempre debemos tener en cuenta tras el uso de estas herramientas que es necesario un paso de revisión manual para corregir errores potenciales.

5.5.1. Anotación utilizando Prokka

El flujo de trabajo Prokka facilita el proceso de anotación completa de un genoma bacteriano, Archaea o viral.



Enlace de interés

Prokka está disponible en su última versión en el siguiente enlace:

<https://github.com/tseemann/prokka>

Este flujo de trabajo coordina una serie de herramientas prediseñadas como Blast, predictores de CDS y proteínas como Prodigal, programas de búsqueda de CRISPR, HMMER, etc., para realizar una anotación sencilla de un genoma bacteriano en aproximadamente diez minutos en cualquier ordenador de sobremesa. Este proceso puede complementarse con bases de datos adicionales para refinar la anotación.

Prokka trabaja con los *contigs* o *scaffolds* ensamblados en formato FASTA. Idealmente, la secuencia debería no tener huecos, pero habitualmente siempre utilizamos un multi-FASTA que procede de nuestro ensamblaje *de novo*. Este es el único archivo obligatorio.

El proceso de anotación se realiza basado en **herramientas de predicción externas** para identificar las coordenadas de los patrones genéticos en los *contigs/scaffolds*. Estas herramientas son:

- **Prodigal** (Hyatt *et al.*, 2010), para la predicción de secuencias codificantes (CDS).
- **RNAmmer** (Lagesen *et al.*, 2007), para la búsqueda de genes ribosomales (rARN).
- **Aragorn** (Laslett & Canback, 2004), para la búsqueda de genes de transferencia de ARN.
- **SignalIP** (T. N. Petersen *et al.*, 2011), para la búsqueda de péptidos señal.
- **Infernal** (Nawrocki *et al.*, 2009), para la búsqueda de ARN no codificante (ncARN).

La anotación de los genes codificantes de proteínas se realiza en dos pasos. Inicialmente, Prodigal identifica las coordenadas de los genes candidatos, pero no describe el producto de ese gen putativo. La forma tradicional de predecir qué codifica un gen es compararlo con una gran base de datos de secuencias conocidas, generalmente a nivel de secuencia de proteínas, y transferir la anotación de la coincidencia más significativa. Prokka utiliza este método, pero de una manera jerárquica, comenzando por una base de datos confiable más pequeña, pasando a una base de datos de tamaño mediano pero específica del dominio, y finalmente a modelos seleccionados de familias de proteínas.

De manera predeterminada se utiliza un umbral para el e-valor de esta coincidencia de 106, incluyendo las siguientes **bases de datos**:

- **Base de datos de proteínas provista por el usuario (opcional).** Se espera un archivo multifasta de proteínas anotadas que sean muy fiables. Serán utilizadas como base de datos primaria. Se busca en ellas utilizando BLAST+/Blastp.
- **Todas las proteínas bacterianas contenidas en UniProt,** que tienen evidencia de ser una proteína real o transcripto y que no son un fragmento. Esta base de datos contiene entre 16 000 a 20 000 proteínas y contiene aproximadamente más del 50 % de los genes core de la mayor parte de los genomas. A esta se accede vía BLAST+/Blastp.
- **Todas las proteínas de genomas bacterianos cerrados en RefSeq para un género específico.** Si incluimos el género bacteriano entre los comandos de entrada, podrá utilizarse. Las bases de datos varían en tamaño y calidad dependiendo de lo poblado que esté ese género en las bases de datos. Se utiliza BLAST+/Blastp para este análisis.
- **Bases de datos basadas en perfiles de Markov (hidden Markov model, HMM)** incluyendo la base de datos Pfam y TIGRFAMs. Esta búsqueda se realiza utilizando hmmsearch del paquete HMMER.

Si no se encuentra ninguna correspondencia en estas bases de datos, la proteína se etiqueta como “proteína hipotética”.

Tras el procesamiento obtendremos los siguientes archivos de salida:

- **.fna:** archivo original FASTA que utilizamos como entrada del programa (nucleótidos).
- **.faa:** archivo de genes codificantes traducidos en formato FASTA (proteínas).
- **.ffn:** archivo de regiones genómicas en formato FASTA (nucleótidos).
- **.fsa:** secuencias de los *contigs* preparadas para ser enviadas a las bases de datos como NCBI (nucleótidos).
- **.tbl:** tabla con las regiones genómicas y CDS para el envío a las bases de datos.
- **.sqn:** archivo en formato SEQUIN editable para su envío a las bases de datos.
- **.gbk:** archivo de tipo Genbank que contiene la secuencia y las anotaciones.
- **.gff:** archivo de tipo GFF v3 que contiene la secuencia y las anotaciones.
- **.log:** archivo de registro del procesamiento de Prokka.
- **.txt:** resumen de las estadísticas de la anotación.



Ejemplo

En este ejemplo vamos a realizar la anotación de los *contigs* ensamblados en el paso anterior mediante Prokka. Para ello, vamos a utilizar una base de datos de proteínas de referencia. Aunque existen varias maneras de realizarla, para simplificar el proceso, utilizaremos las proteínas de un genoma de *E. coli* referencia, como es *E. coli* K12. Exploraremos las opciones disponibles.

>>>

>>>

1. Descargamos el archivo de GenBank para *E.coli* K12, desde el subapartado de:

https://www.ncbi.nlm.nih.gov/genome/167?genome_assembly_id=161521

Nota: recordad bajar el archivo GenBank (Full).

2. Extraemos las proteínas desde este archivo, con la función de Prokka siguiente:

```
</>
prokka-genbank_to_fasta_db GCF_000005845.2_ASM584v2_genomic.gbff >
ref_prots.faa
```

3. Ejecutamos Prokka con los siguientes parámetros:

```
</>
prokka --outdir prokka_Ecoli --addgenes --addmrna --genus
Escherichia --species coli --kingdom Bacteria --usegenus
--proteins ref_prots.faa --mincontiglen 0 .../spades_auto/
contigs.fasta
```

4. Revisamos los archivos de salida.



Enlace de interés

Recordad que para más opciones debéis consultar el manual de Prokka o la web:

<https://github.com/tseemann/prokka>

5.5.2. Visualización de las anotaciones

Una vez anotado el genoma de manera automática, es importante realizar un proceso de curado manual de la anotación. Este proceso es largo y tedioso, por lo que el paso de anotación con bases de datos específicas puede facilitar sustancialmente los pasos posteriores.

Se pueden visualizar los archivos de anotación en programas como Artemis (Carver et al., 2012) o CLC Sequence Viewer/CLC Genomics Workbench (Latest improvements for CLC Sequence Viewer - Current line - Qiagen Bioinformatics, s. f.), donde se pueden modificar las regiones detectadas. Existen programas más avanzados, pero que requieren una suscripción de pago. Los archivos habitualmente utilizados para revisar las anotaciones son los archivos GenBank (GBK) o GFF/GTF. Ambos tipos de archivos son proporcionados por Prokka.



Ejemplo

En este ejemplo vamos a visualizar las anotaciones del genoma de *E. coli* del ejemplo anterior, para analizar las regiones detectadas y anotadas.

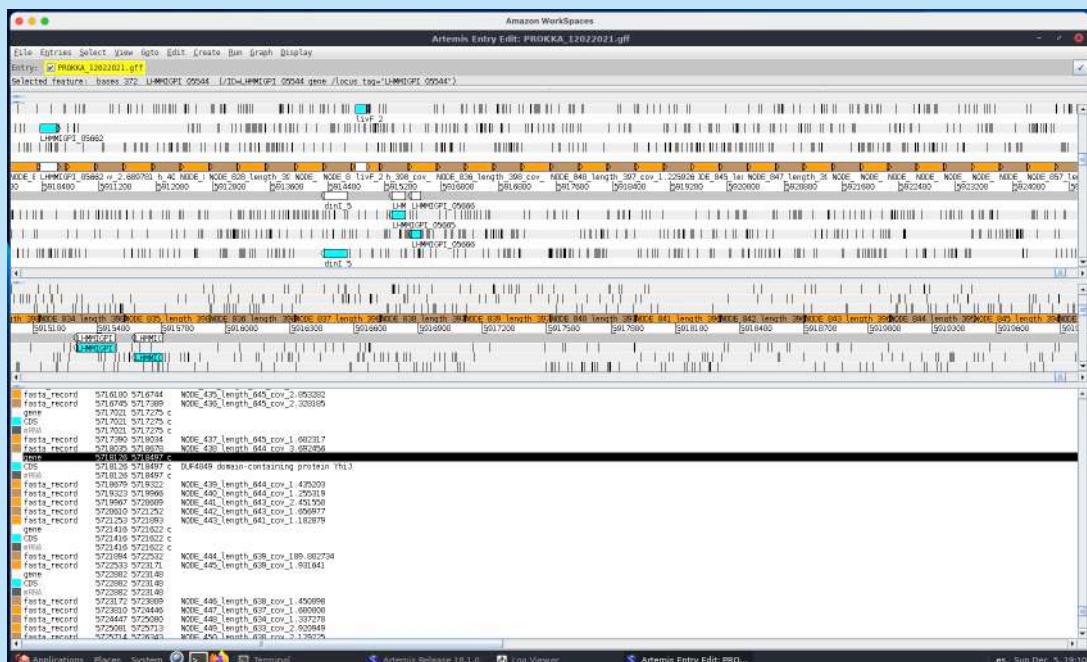
La apertura del programa se realiza desde la terminal, simplemente indicando art. Para una mayor sencillez podemos abrir directamente el archivo GFF indicando `art file.gff`. Se desplegará el programa como se observa en la Figura 28.

>>>

>>>

Figura 28

Visualización de los archivos de anotación GFF en el programa Artemis



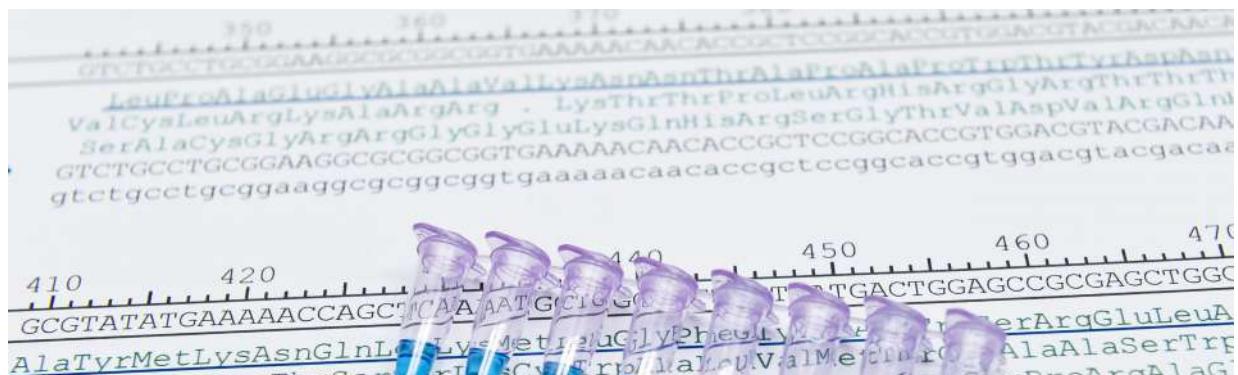
Aunque es una herramienta bastante antigua (Berriman & Rutherford, 2003; Carver *et al.*, 2012; Rutherford *et al.*, 2000) es una de las pocas herramientas de visualización que nos permite editar los archivos de manera gratuita sin pagar licencias. Además de la lectura de archivo GFF, se pueden leer archivos de tipo GenBank (siempre que contengan la secuencia en su interior) o FASTA. Se pueden añadir anotaciones a estos archivos en diferentes capas de información, incluso realizar una predicción de genes desde el propio programa, aunque no es una opción muy depurada.



Enlace de interés

Para aprender el funcionamiento de Artemis en detalle, se recomienda la lectura del manual *Viewing and annotating sequence data with Artemis* de Berriman y Rutherford (2003), así como la visualización del videotutorial:

<https://www.youtube.com/watch?v=3m84K7l7Omk>



5.6. Reconstrucción del genoma: elementos genéticos móviles (EGM). Plásmidos



Los elementos genéticos móviles (EGM) son todos aquellos elementos genéticos que juegan un papel importante en la transferencia horizontal de material genético entre bacterias.

Los organismos procariotas intercambian ADN de manera horizontal mediante tres procesos, donde la transducción y la conjugación dependen de EGM especializados, donde se incluyen los plásmidos y algunos bacteriófagos:

- **Transformación**, donde el paso del material genético está mediado por la captación de ADN libre.
- **Conjugación**, donde la transferencia del material genético está mediada por plásmidos o elementos integrativos conjugativos (ICE). El contacto entre células, a través de un *pilus* de conjugación es necesario para esta transferencia.
- **Transducción**, mediada por la presencia de un bacteriófago que encapsula el ADN entre su material genético y lo libera a una célula nueva.

Los organismos procariotas poseen una tercera clase de EGM que son los **transposones**, capaces de moverse y reorganizar el ADN en la propia célula. Los transposones se mueven entre distintas células a través de los plásmidos, fagos o los elementos integrativos conjugativos (ICE).

Los EGM pueden mediar en el tráfico de ADN tanto a nivel intracelular como intercelular, ya que tienen genes propios que les confieren la capacidad de autorreplicación, de manera independiente al cromosoma celular, para alcanzar la recombinación homóloga o no homóloga, así como para comunicarse mediante un *pilus* con células cercanas y transferir este material genético. Esta capacidad se encuentra en sus genes centrales (*core*); mientras que disponen de genes accesorios que proporcionan a la célula receptora una ventaja adaptativa. Los más comunes incluyen genes de resistencia a antibióticos, genes de virulencia o rutas metabólicas inusuales. De hecho, los fenotipos más preocupantes desde el punto de vista de la implicación clínica están codificados por EGM, por lo que son elementos muy destacados en el estudio de genomas bacterianos.

Una revisión sobre EGM como agentes de evolución bacteriana se puede encontrar en el artículo de Frost *et al.* (2005), así como en estas otras referencias, focalizadas en los últimos avances en este campo: De Toro *et al.* (2014), Redondo-Salvo *et al.* (2020).

Tradicionalmente, los EGM se han examinado por métodos moleculares tradicionales como PCR sobre regiones de interés (región de replicación o conjugación) o hibridación. Las técnicas de secuenciación masiva nos permiten llegar a un nivel de detalle superior. Sin embargo, no están exentas de limitaciones, debido principalmente a la presencia de elementos repetidos y regiones redundantes (por ejemplo, inserciones), tan comunes en este tipo de elementos. Estas regiones son confusas a la hora de la resolución de los ensamblajes, causando reorganizaciones irreales y huecos, especialmente con tecnología de lecturas cortas.

No se debe desestimar tampoco que la preparación de las genotecas para la tecnología de lecturas largas provoca la pérdida de los EGM de pequeño tamaño, así como los ensamblajes basados en jerarquía de tamaños. Estos son los motivos por los que la reconstrucción de los EGM es todavía un reto a nivel bioinformático (Álvarez-Molina *et al.*, 2021; Arredondo-Alonso *et al.*, 2017).

Existe un gran número de herramientas bioinformáticas para el análisis específico de plásmidos, los EGM más relevantes desde el punto de vista epidemiológico. Estas herramientas pueden dividirse en tres grupos, que veremos a continuación.

5.6.1. Herramientas de ensamblaje y extracción de plásmidos

Estas son herramientas desarrolladas para ensamblar y extraer los plásmidos directamente de las lecturas de secuenciación o los *contigs* ensamblados. En ellas se asume que los plásmidos tienen características diferenciales del cromosoma. Ejemplo de ello es asumir (no siempre de manera correcta) que los plásmidos están en mayor número de copias que el cromosoma, de manera que su cobertura al ser secuenciados es mayor que la del cromosoma. También se asume, que debido a su naturaleza circular, el ensamblaje de los mismos puede circularizarse. Herramientas que están en este grupo son cBar (Zhou & Xu, 2010), plasmidSPAdes (Antipov *et al.*, 2016), Plasmid Profiler (Zetner *et al.*, 2017), Recycler (Rozov *et al.*, 2017), PlasmidSeeker (Roosaare *et al.*, 2018), PlaScope (Royer *et al.*, 2018), PlasFlow (Krawczyk *et al.*, 2018) o PlasmidTron (Page *et al.*, 2018).

Las asunciones sobre las que se basan no siempre son ciertas. En primer lugar, solo los plásmidos multicopia presentan una cobertura media sustancialmente superior al cromosoma. De hecho, los plásmidos considerados grandes, que son principalmente aquellos que atraen nuestro interés por portar genes de resistencia a antibióticos, virulencia u otras funciones de interés epidemiológico, se suelen encontrar en una única copia respecto al cromosoma (De Toro *et al.*, 2014; Fernandez-Lopez *et al.*, 2017; Redondo-Salvo *et al.*, 2020). Además, no debemos olvidar que el método de preparación de la genoteca puede afectar a la cobertura media final observada. En este sentido, los protocolos que conllevan un Enriquecimiento por PCR de la genoteca pueden producir sesgos, enriqueciendo aquellas regiones o elementos que son más comunes en el genoma. Por este motivo, las librerías sin este tipo de enriquecimiento son deseables para la detección de EGM. Por otra parte, no todos los plásmidos son circulares, sino que existen (aunque son poco numerosos) plásmidos que son moléculas de ADN lineal, como aquellos encontrados en *Borrelia* spp., *Streptomyces* spp. o *Nocardia opaca*, por poner algunos ejemplos.

5.6.2. Herramientas de identificación mediante marcadores genéticos

Estas son las herramientas que permiten la identificación de plásmidos a través de marcadores genéticos. En este grupo de herramientas se encuentran:

- PLACNET (Lanza *et al.*, 2014) o PLACNETw (Vielva *et al.*, 2017), donde se combina el ensamblaje de los plásmidos, la información de cobertura y la visualización de marcadores genéticos de plásmidos (como son las proteínas de inicio de replicación o relaxasas, encargadas de la formación del *pilus*).
- Por otra parte, PlasmidFinder (Carattoli *et al.*, 2014) o MOBscan (Garcillán-Barcia *et al.*, 2020) actúan como base de datos y motor de búsqueda para regiones de replicación o relaxasa, respectivamente, permitiendo la tipificación de *contigs* o proteínas, tras el paso de ensamblaje.
- La herramienta MOB-suite (Robertson & Nash, 2018) expande el esquema de tipado con la reconstrucción de las secuencias de plásmidos a partir de ensamblajes.

- Otra de las herramientas de interés es **mlplasmids** (Arredondo-Alonso *et al.*, 2018), donde se aplica un sofisticado sistema de clasificación binaria de especies para predecir si un *contig* deriva de un plásmido o un cromosoma, basándose en un entrenamiento previo mediante técnicas de *machine learning* con genomas bien anotados.
- Finalmente, aplicando índices de similitud nucleotídica ANI modificados, se encuentra **COPLA** (Redondo-Salvo *et al.*, 2021), que es capaz de clasificar los plásmidos en grupos taxonómicos (*plasmid taxonomic units*, PTU), previamente descritos en una base de datos (Redondo-Salvo *et al.*, 2020).
- **Bases de datos con información sobre plásmidos y otros EGM**, donde se permite la inspección de secuencias. Se incluyen **ACLAME** (Leplae *et al.*, 2004; Leplae *et al.*, 2010), que contiene fagos, plásmidos y transposones; **plasmidATLAS (pATLAS)** (Jesus *et al.*, 2019), donde vía web se explora la relación entre plásmidos, genes de resistencia y virulencia; y finalmente la base de datos de PTU (Redondo-Salvo *et al.*, 2020; <https://castillo.dicom.unican.es/PlasmidID/>).

5.6.3. Herramientas basadas en los grafos de ensamblaje de novo

Estas herramientas no están especialmente diseñadas para el análisis de plásmidos, pero ha encontrado su utilidad al examinar los *contigs* y su relación entre ellos, para la detección de plásmidos y para integrar la información de lecturas cortas y largas. Entre esas herramientas están Contiguity (Sullivan *et al.*, 2015), Bandage (Wick *et al.*, 2015), Circlator (Hunt *et al.*, 2015) y Unicycler (Wick *et al.*, 2017).



Ejemplo

En este ejemplo vamos a realizar un ensamblaje diferencial de plásmidos utilizando la herramienta **plasmidSPAdes**. Revisaremos el grafo de ensamblaje obtenido con Bandage y tipificaremos los plásmidos mediante **PlasmidFinder** y **COPLA**. Para ello utilizaremos las lecturas del Ejemplo 1 (Capítulo 1), obtenidos de Bioproject que fueron filtrados y que se han utilizado en los ejemplos anteriores.

Anotación con plasmidSPAdes

Comando:

```
plasmidspades.py -1 out_1.clean.fastq.gz -2 out_2.clean.fastq.gz
--careful -t 2 -k 127 -o plasmidSPAdes
```

</>

Revisión de los archivos de salida

Lo primero que podemos notar al revisar el archivo *contigs.fasta* o *scaffolds.fasta*, es que las cabeceras de los *contigs/scaffolds* muestran una información adicional de “componente”, que se refiere a la distinción entre cromosoma o plásmidos realizada en función de la cobertura.

Recuenta el número de *contigs* y *scaffolds* obtenidos. ¿Es coherente con el ensamblaje de genoma completo?

>>>

>>>

Visualización del grafo de ensamblaje

Cargamos el archivo assembly_graph_with_scaffolds.gfa que es el grafo de ensamblaje en Bandage.

Sacamos las estadísticas del ensamblaje en la opción: *Graph information / More info* (Figura 29).

Dibujamos el grafo, con los parámetros por defecto: grafo completo en estilo sencillo. (Figura 30). A la vista de esta figura, ¿cuántos plásmidos crees que podrías tener? ¿De qué tamaños y coberturas?

Figura 29

Visualización de las estadísticas del ensamblaje realizadas con Bandage

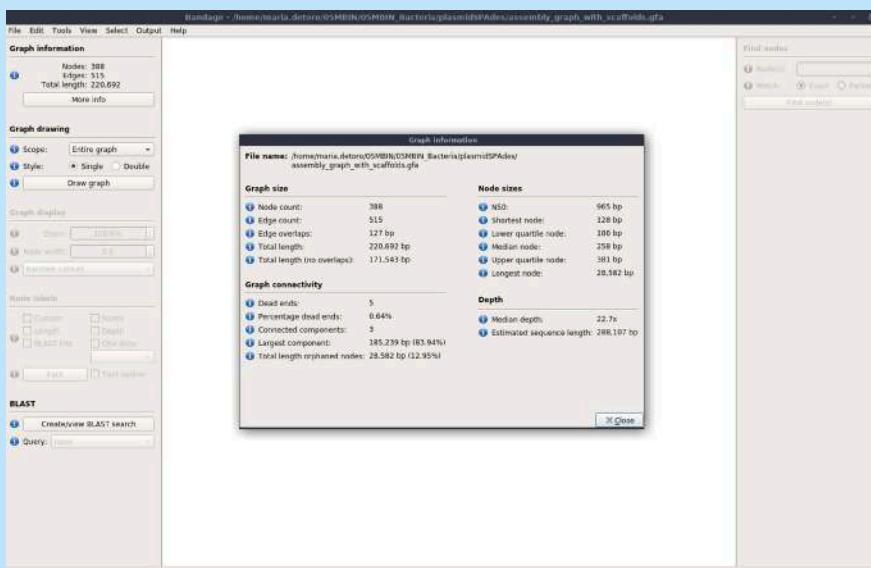
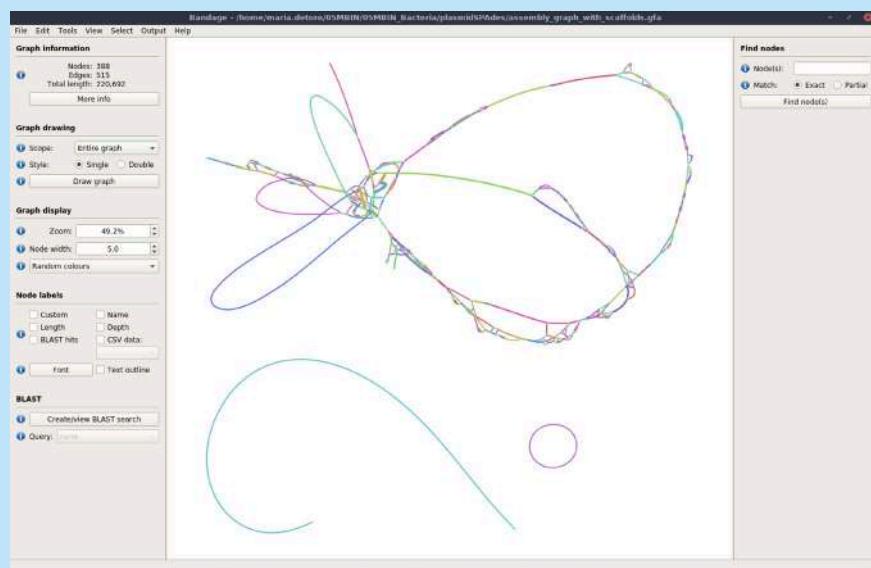


Figura 30

Visualización del grafo realizada con Bandage



>>>

>>>

Podemos analizar lo que contiene cada uno de los nodos utilizando la función BLAST incluida en el programa (*Output / Web BLAST selected nodes*), que nos redireccionará a la página BLAST de NCBI. Con esta opción, realiza el BLAST de algunos de los componentes. ¿Qué información puedes obtener? ¿Puedes verificar que se trata de un plásmido?



Enlaces de interés

En los siguientes enlaces puedes encontrar más información sobre grafos y su visualización en Bandage.

<https://github.com/rrwick/Bandage/wiki>

<https://github.com/rrwick/Bandage/wiki/Graph-paths>

<https://towardsdatascience.com/visualising-assembly-graphs-fb631f46bbd1>

<https://www.youtube.com/watch?v=cierloa5hS0>

Tipificación de plásmidos mediante COPLA

Utilizamos el clasificador COPLA en su versión web (<https://castillo.dicom.unican.es/copla/>), donde incluiremos un identificador del trabajo y los scaffolds obtenidos del ensamblaje. Dado que conocemos la taxonomía de nuestro organismo, incluiremos la información apropiada al mismo (Figura 31). ¿Qué tipo de plásmido nos indica que es?

Busca información sobre ese tipo de PTU en la web de la base de datos:

https://castillo.dicom.unican.es/PlasmidID/RefSeq84_MOB/

Figura 31

Carga de datos en la web de tipificación de plásmidos COPLA

Nota: si quieres separar cada uno de los plásmidos en Bandage, pueden seleccionarse y guardarse como un archivo FASTA en *Output / Save selected node sequences to FASTA*. A partir de los archivos FASTA, puede repetirse esta tipificación, para una mayor resolución.

>>>

>>>

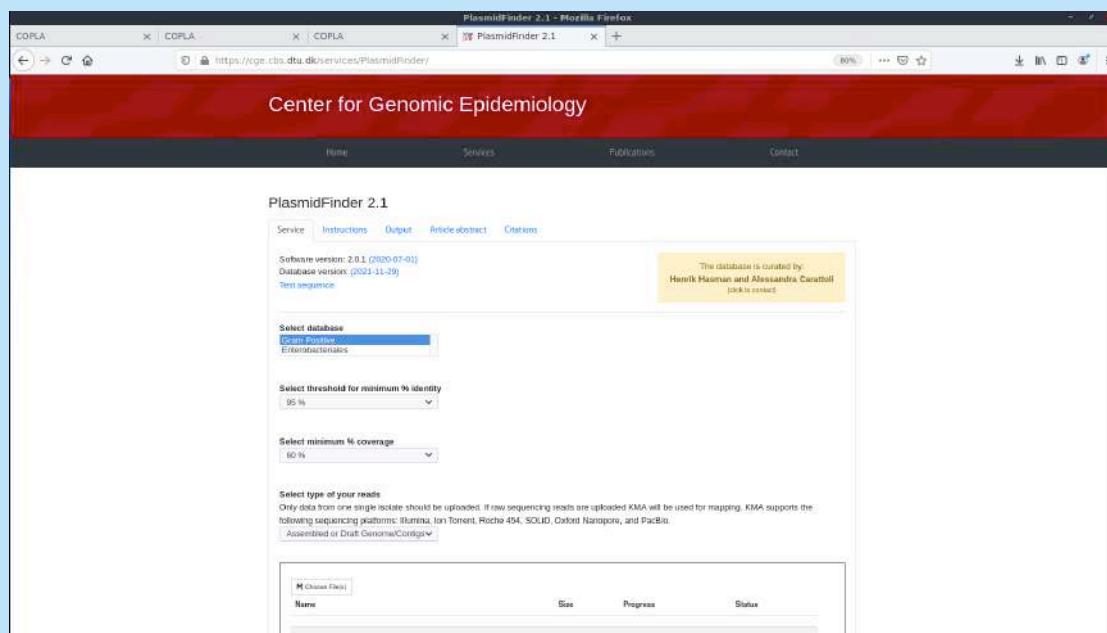
Tipificación de plásmidos mediante PlasmidFinder

En esta primera aproximación a PlasmidFinder utilizaremos su versión web, disponible en (Figura 32):
<https://cge.cbs.dtu.dk/services/PlasmidFinder/>

Donde incluiremos los *scaffolds* o bien, los plásmidos separados previamente con Bandage en formato FASTA. Nota: seleccionad apropiadamente el tipo de organismo que estamos analizando (grampositivo o enterobacteriales). ¿Qué tipos de plásmidos localiza aquí?

Figura 32

Carga de datos en la web de tipificación de plásmidos PlasmidFinder



5.7. Detección y anotación de regiones de interés: genes de resistencia a antibióticos y virulencia

Como hemos visto hasta este momento, la secuenciación del genoma completo de los organismos tiene grandes utilidades. En esta sección, veremos que la predicción de los genes de resistencia a antibióticos o virulencia a partir de la secuencia del genoma completo es una herramienta muy útil en la predicción de fenotipos, monitorización de la dispersión de este tipo de genes y en epidemiología genómica. Para esta finalidad, existen diferentes procedimientos:

- Mapear las lecturas secuenciadas directamente sobre una base de datos de referencia, que contenga las secuencias que queremos detectar. Esta aproximación se utiliza habitualmente con datos de metagenómica, puesto que no es necesario ensamblar el genoma completo para poder realizar esta predicción. El mapeo puede realizarse directamente con las lecturas completas, como la que ejecutan los programas SRST2 (Inouye *et al.*, 2014) o GROOT (Rowe & Winn, 2018).

Por otra parte, otra aproximación más rápida aún es machear *k*-meros de las lecturas secuenciadas frente a la base de datos de referencia, siendo un método rápido y evitando los pasos de ensamblaje y anotación, pero incrementando el riesgo de falsos positivos debido a errores de secuenciación o contaminación de ADN. Esta última aproximación la utilizan programas como PhenotypeSeeker (Aun *et al.*, 2018) o KmerResistance (Clausen *et al.*, 2016). Ambas aproximaciones son muy útiles cuando tenemos un gran número de genomas y queremos seleccionar aquellos que portan alguna característica especial para su procesamiento posterior.

- Mapear los *contigs/scaffolds* ensamblados previamente utilizando el algoritmo Basic Local Alignment Search Tool (BLAST) para encontrar todos los posibles genes de interés entre esta secuencia y la base de datos. Los *contigs/scaffolds* pueden compararse frente a bases de datos públicas disponibles o bien frente a bases de datos diseñadas a medida. Este procedimiento es más lento, puesto que requiere un paso previo de ensamblaje e, incluso, de anotación de secuencias codificantes. Esto puede ser costoso en tiempo y en recursos computacionales. Asimismo, este método solo es fiable para genomas sencillos, mientras que ve limitada su aplicación en poblaciones de genomas complejos.

Independientemente del método utilizado, es crucial utilizar una base de datos fiable y curada para obtener resultados fidedignos sobre el fenotipo de resistencia y virulencia. Entre estas bases de datos encontramos:

- ResFinder (E. Zankari *et al.*, 2012) (para genes de resistencia a antibióticos adquiridos), PointFinder (Ea Zankari *et al.*, 2017) (para mutaciones puntuales) y VirulenceFinder (Joensen *et al.*, 2014) (para factores de virulencia). Todas ellas han sido implementadas vía web, pero también es posible descargar las secuencias y utilizarlas por vía de comandos:

<https://bitbucket.org/genomicepidemiology/workspace/projects/DB>

- Comprehensive Antibiotic Resistance Database (CARD) (Jia *et al.*, 2017) y ARG-ANNOT (Gupta *et al.*, 2014) son dos de las bases de datos más utilizadas para genes de resistencia a antibióticos; mientras que Virulence Factor Data Base (VFDB) (Liu *et al.*, 2019; VFDB: Virulence Factors Database, s. f.) lo es para genes de virulencia.
- La base de datos AMRFinder, perteneciente a NCBI, es una herramienta para detectar genes de resistencia a antibióticos utilizando la anotación de proteínas o secuencias nucleotídicas obtenidas directamente de NCBI y una colección curada de modelos de Markov:

<https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/>

Aunque en algunos casos estas bases de datos disponen de su propio programa de búsqueda, en todos los casos basados en BLAST, en todo caso podemos descargar estas bases de datos, manipularlas si fuese necesario y utilizar BLAST de manera local. Obtendremos una tabla manipulable y fácil de interpretar. Cabe destacar que aquellos genes de interés que no están presentes en la base de datos utilizada, obviamente, no podrán ser detectados.



Ejemplo

En este ejemplo vamos a utilizar la anotación Prokka realizadas anteriormente, donde tenemos los archivos de CDS y proteínas calculados. Utilizaremos AMRFinder para la detección de genes de resistencia a antibióticos y un BLAST frente a la base de datos VirulenceFinder para la detección de genes de Virulencia.

Utilización de AMRFinder para la detección de genes de resistencia a antibióticos y factores de virulencia

Para su utilización tenemos un *environment* conda propio para este programa. Toda la información de instalación y uso está en la web:

<https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/>

En primer lugar actualizaremos la base de datos a su última versión:

```
amrfinder --update
```



Seleccionamos el organismo que utilizar preferentemente, buscando entre los disponibles como: `amrfinder -list_organisms`. Aquí veremos que en nuestro caso tenemos que seleccionar “Escherichia” con el parámetro `-organism`.

Utilizamos el programa:

```
amrfinder --organism Escherichia -n PROKKA_12022021.ffn -c 0.8  
-i 0.9 --plus --log amrfinder.log > AMRFinder_out.csv
```



Hay que prestar atención a los parámetros de identidad y cobertura utilizados.

- ¿Qué tipos de genes se detectan con esta base de datos?
- ¿Qué genes de resistencia encuentras? ¿Y de virulencia?

Utilización de ResFinder y VirulenceFinder para la detección de genes de resistencia a antibióticos y factores de virulencia

Para la utilización de estas bases de datos usaremos BLAST como motor de búsqueda (instalado en el *environment* conda de la asignatura). Descargamos las bases de datos correspondientes del repositorio Bitbucket:

https://bitbucket.org/genomicepidemiology/resfinder_db/src/master/

https://bitbucket.org/genomicepidemiology/virulencefinder_db/src/master/

>>>

>>>

En el caso de la base de datos de resistencias, **ResFinder**, está desglosada en el tipo de antibióticos diana. Debemos descargar cada uno de los archivos FASTA (.fsa) y concatenarlos para tener todos los genes en un único archivo. Si se prefiere examinar por familias, podríamos utilizarlos independientemente, pero resulta más laborioso.

```
git clone https://mdtorohernando@bitbucket.org/genomicepidemiology/
resfinder_db.git

cd resfinder_db

cat *.fsa > ResFinder_all.fsa

grep '>' ResFinder_all.fsa #recontamos el número de genes que tenemos:
3161

makeblastdb -dbtype nucl -in ResFinder_all.fsa #indexamos la base de
datos de referencia

blastn -query ../PROKKA_12022021.ffn -db ResFinder_all.fsa -outfmt 6
-max_target_seqs 5 > resfinder_out.csv
```

- ¿Qué genes de resistencia se han encontrado?
- ¿Te parece suficiente esta base de datos?
- ¿Qué fenotipo de resistencia corresponde al gen principal detectado? ¿Cómo podrías saberlo?

La base de datos de factores de virulencia, **VirulenceFinder**, se encuentra desglosada en función del organismo. Bajaremos la base de datos desde el siguiente enlace:

https://bitbucket.org/genomicepidemiology/virulencefinder_db/src/master/

```
git clone https://mdtorohernando@bitbucket.org/genomicepidemiology/
virulencefinder_db.git

cd virulencefinder_db

makeblastdb -dbtype nucl -in virulence_ecoli.fsa

blastn -query ../PROKKA_12022021.ffn -db virulence_ecoli.fsa -outfmt 6
-max_target_seqs 5 > virulencefinder_out.csv
```

- ¿Cuáles son los genes de virulencia detectados?
- ¿A qué factores corresponden? ¿Cómo puede extraer esta información?



Capítulo 6

Análisis de genoma de virus. El caso de SARS-CoV-2

6.1. Estructura de los virus. Aspectos generales

Los virus constituyen el grupo de los organismos más simples que se conocen. De hecho, en términos biológicos, los virus no son considerados organismos vivos, por depender de una célula hospedadora y su maquinaria biológica para poder replicarse. Algunos virus poseen genes que codifican su propio ADN o ARN polimerasa, pero algunos dependen de las enzimas del hospedador para replicarse y transcribirse. Todos los virus utilizan los ribosomas y la maquinaria de translación para la síntesis de los péptidos que construyen la envuelta de su progenie. Esto nos indica que los virus deben encajar con la maquinaria de su hospedador, lo que hace que sean bastante específicos de organismos particulares, y tipos de virus concretos no infectan un amplio espectro de especies.

Existen una multitud de tipos de virus, que podemos clasificar en dos grandes grupos, en función del tipo de célula al que infectan.

6.1.1. Bacteriófagos

Los **bacteriófagos** son aquellos virus que infectan bacterias. Están compuestos por dos componentes básicos que son las proteínas (que forman la envuelta o cápside) y el ácido nucleico, contenido dentro de ella.

Este ácido nucleico puede ser tanto ADN como ARN y, a su vez, puede ser tanto de doble hélice como de hélice sencilla. La cápside puede tener forma isosaédrica (fago MS2 que infecta *E. coli* o PM2 que infecta *P. aeruginosa*), filamentosa o con forma de hélice (M13 – *E. coli*) o con forma cabeza-cola, con una cabeza icosáédrica que contiene el ácido nucleico unida a una cola filamentosa que facilita la entrada del ácido nucleico en la célula infectada (fagos T4 y lambda). Por lo tanto, existe una gran variedad de tipos de bacteriófagos, que podéis ver generalizada y resumida en la Tabla 11.

Tabla 11

Características generales de algunos bacteriófagos

Fago	Hospedador	Estructura cápside	Estructura del genoma	Tamaño genoma (kb)	Número de genes
Lambda	<i>Escherichia coli</i>	Cabeza-cola	ADN linear doble cadena	49.5	48
X174	<i>E. coli</i>	Icosaédrico	ADN circular cadena sencilla	5.4	11
f6	<i>Pseudomonas phaseolicola</i>	Icosaédrico	ARN linear segmentado doble cadena	2.9; 4.0; 6.4	13
M13	<i>E. coli</i>	Filamentoso	ADN circular cadena sencilla	6.4	10
MS2	<i>E. coli</i>	Icosaédrico	ARN linear cadena sencilla	3.6	3
PM2	<i>Pseudomonas aeruginosa</i>	Icosaédrico	ADN linear cadena doble	10.0	21
SPO1	<i>Bacillus subtilis</i>	Cabeza-cola	ADN linear doble cadena	150	> 100
T2, T4, T6	<i>E. coli</i>	Cabeza-cola	ADN linear doble cadena	166	> 150
T7	<i>E. coli</i>	Cabeza-cola	ADN linear doble cadena	399	> 55

Nota. Adaptado de “Genomes 3”, por T. A. Brown, 2017, *Yale Journal of Biology and Medicine*, 90(4).

Una de las características generales de los fagos es su capacidad de empaquetar un gran número de genes en poco espacio, gracias a contener genes solapantes. Estos comparten secuencia genética, pudiendo estar un gen contenido dentro de otro, pero codificando para proteínas diferentes, ya que los transcritos comienzan en puntos de traducción diferentes y en marcos de lectura diferentes. Esta característica es común a muchos virus, que exhiben mayor complejidad cuantos más genes contienen.

Por otro lado, los bacteriófagos se clasifican en dos grupos según su ciclo de vida: líticos y lisogénicos. Los fagos de tipo lítico extienden a la bacteria hospedadora muy rápidamente tras la infección inicial, mientras que los fagos lisogénicos pueden permanecer quiescentes en la célula huésped por un tiempo sustancial, incluso varias generaciones. Durante el ciclo lisogénico algunos fagos pueden integrarse en el genoma del hospedador, dando lugar a un **profago**. Esta integración se da mediante recombinación sitio-específica, en un lugar concreto del genoma del hospedador. Este profago integrado puede mantenerse durante generaciones y es replicado junto con el genoma bacteriano, pasando a las células hijas.

Estímulos químicos o físicos hacen que este profago se induzca a una **fase lítica**, comenzando por una recombinación que lo escinde del genoma del hospedador, comienza su replicación y se sintetizan las proteínas de la envuelta. Finalmente, la célula es lisada y los nuevos fagos son liberados al medio, listos para infectar nuevas bacterias. Ejemplo de este proceso es el fago lambda de *E. coli*.

Los bacteriófagos han tomado un especial interés porque pueden arrastrar en su escisión genes de resistencia a antibióticos o virulencia, así como empaquetar elementos genéticos móviles como plásmidos de pequeño tamaño (Colavecchio *et al.*, 2017; Navarro & Muniesa, 2017; Prieto *et al.*, 2016; Rodríguez-Rubio *et al.*, 2020). Por otra parte, están siendo ampliamente estudiados para combatir el problema de la resistencia a antibióticos (Bragg *et al.*, 2014; Brives & Pourraz, 2020; Moghadam *et al.*, 2020; Principi *et al.*, 2019).



Enlace de interés

Para comprender las diferencias entre el ciclo lítico y lisogénico, este vídeo explica perfectamente los conceptos.

<https://www.youtube.com/watch?v=vX2D5zJbuMU>

6.1.2. Virus que infectan células eucariotas

La cápside de estos virus puede ser icosaédrica o filamentosa, pero no se han detectado estructuras de cabeza-cola, que permanecen exclusivas de bacteriófagos. Una característica diferencial de los virus eucariotas, especialmente aquellos que infectan células animales, es que la cápside puede estar rodeada de una membrana lipídica, formando un componente adicional de la estructura vírica. Esta membrana deriva del hospedador cuando la nueva partícula vírica se libera y abandona la célula y puede estar modificada por la inserción de proteínas específicas del virus.

Los virus de eucariotas muestran una gran variedad de estructuras (Tabla 12). Nuevamente, pueden ser virus de ADN o ARN, cadena doble o sencilla, o incluso parte doble con regiones de cadena sencilla, lineares o circulares, segmentados o no segmentados.

Tabla 12

Características generales de algunos virus que infectan células eucariotas

Virus	Hospedador	Estructura del genoma	Tamaño genoma (kb)	Número de genes
Adenovirus	Mamíferos	ADN linear doble cadena	360	30
Hepatitis B	Mamíferos	ADN circular parcialmente doble cadena	32	4
Virus Influenza	Mamíferos	ARN linear segmentado cadena sencilla	220	12
Parvovirus	Mamíferos	ADN linear cadena sencilla	16	5
Poliovirus	Mamíferos	ADN linear cadena sencilla	76	8
Reovirus	Mamíferos	ARN linear segmentado doble cadena	225	22

>>>

>>>

Virus	Hospedador	Estructura del genoma	Tamaño genoma (kb)	Número de genes
Retrovirus	Mamíferos, pájaros	ARN linear cadena sencilla	6.0-9.0	3
SV40	Monos	ADN circular cadena doble	5.0	5
Virus mosaico del tabaco	Plantas	ARN linear cadena sencilla	6.4	6
Virus <i>Vaccinia</i> (viruela)	Mamíferos	ADN circular cadena doble	240	240

Nota. Adaptado de "Genomes 3", por T. A. Brown, 2017, *Yale Journal of Biology and Medicine*, 90(4).

La mayor parte de estos virus siguen un ciclo de infección lítico, pero unos pocos pueden comportarse con ciclo lisogénico. Algunos coexisten con las células hospedadoras durante largos períodos de tiempo (años), incluso integrándose en el genoma de la célula. Si son virus ARN tienen un paso adicional de conversión a ADN, para integrarse y ser retroelementos virales. Gracias a una enzima llamada **transcriptasa reversa**, estos virus ARN se copian a ADN complementario, crean la hebra complementaria de ADN para generar la estructura de doble cadena y posteriormente son integrados en un lugar del genoma. A diferencia del bacteriófago lambda, los retrovirus no tienen similitud genética con el ADN del hospedador. Esta integración es un prerequisito para la expresión de los genes del retrovirus que codifican para poliproteínas. Cada una de ellas es escindida en dos o más productos, incluyendo las proteínas de la envuelta viral y la transcriptasa reversa.



Enlace de interés

En el siguiente enlace, podrás encontrar un resumen de los tipos de virus que podemos encontrar.

<https://www.youtube.com/watch?v=jX3MhWWi6n4>

6.2. El genoma de SARS-CoV-2

La enfermedad COVID-19 es una enfermedad infecciosa causada por el virus SARS-CoV-2. La primera datación de este virus se tiene en la provincia de Wuhan (China) en otoño-invierno de 2019 (estas dataciones están aún bajo estudio). Al igual que otros coronavirus, SARS-CoV-2 afecta al sistema respiratorio, causando una enfermedad respiratoria con síntomas como tos, fiebre y, en casos más graves, disnea. La mortalidad es alta en grupos de edad elevada (> 80 años), así como en personas de distintas edades con factores de riesgo (inmunosupresión, obesidad, hipertensión, diabetes, etc.). Además del síndrome respiratorio, hoy en día conocemos que COVID-19 puede dar un proceso inflamatorio, conduciendo a sepsis, daño cardiaco agudo, fallo cardiaco y disfunción multiorgánica en pacientes de alto riesgo; así como persistir con secuelas graves en algunos pacientes.

Los coronavirus (CoV) son un grupo de virus con envuelta, donde su material genético es una hebra de ARN de cadena única. Este grupo de virus pertenecen a la familia *Coronaviridae* del orden de los *Nidovirales*. Han sido clasificados en cuatro géneros que incluyen los alfa, beta, gamma y delta coronavirus. Entre ellos, alfa y beta infectan mamíferos, mientras que gamma infecta especies aviares y delta infectan ambos tipos, mamíferos y aves. Los coronavirus como SARS-CoV, el coronavirus de la hepatitis murina (MHV), MERS-CoV, el coronavirus bovino (BCoV), el coronavirus de murciélagos HKU4 y los coronavirus humanos como OC43 y SARS-CoV-2 pertenecen al grupo de los beta coronavirus. Los más conocidos, que han dado lugar a epidemias o incluso pandemia, son SARS-CoV, MERS y SARS-CoV-2, todos ellos transmitidos a partir de un evento zoonótico.

Los coronavirus tienen un genoma de ARN que comprende entre 26 a 32 kb de longitud. En el caso de SARS-CoV-2, comparte un 82 % de identidad en su secuencia con los genomas de SARS-CoV y MERS-CoV, y más de un 90 % de identidad en proteínas estructurales y con potencial enzimático. Este alto nivel de homología en su secuencia revela un mecanismo similar de patogénesis y está siendo clave en el desarrollo de tratamientos antivirales y, por supuesto, de la vacuna. De manera principal, SARS-CoV-2 contiene cuatro proteínas estructurales que incluyen la espícula (*spike*, S), envuelta (*envelope*, E), la membrana (*membrane*, M) y la nucleocápside (*nucleocapsid*, N). Estas proteínas comparten alta homología con las correspondientes de los coronavirus asociados SARS-CoV y MERS-CoV.

Los CoV dependen de sus proteínas de espícula (S) para unirse al receptor localizado en la superficie de las células huésped. El receptor de tipo angiotensina ACE-2 es la cerradura, donde la espícula (la llave) abre la puerta para entrar en la célula y provocar la infección. El dominio de unión a receptor (RBD) está compuesto por las subunidades S1 y S2.

El primer genoma disponible de SARS-CoV-2, originario de Wuhan, se encuentra disponible en la base de datos NCBI con el acceso NC_045512.2 (acceso a fecha 12/12/2021). Tiene una longitud de 29 903 pb y está compuesto por 13-15 marcos de lectura abiertos (ORF), doce de ellos funcionales. Existen once genes codificantes de doce proteínas totales que se expresan. La organización génica es muy similar a SARS-CoV y MERS-CoV, donde en sentido 5' a 3' contienen en primer lugar los genes codificantes de las proteínas no estructurales ORF1a y ORF1b, parcialmente solapantes. Estas codifican para las poliproteínas pp1a y pp1ab, respectivamente, que se procesan mediante las proteinasas codificadas por el virus y producen 16 proteínas, que están bien conservadas en todos los CoV pertenecientes a la misma familia (Figura 33). Posteriormente, se encuentra la codificación para las proteínas estructurales S, E, M y N, que son las dianas principales de las vacunas y tratamientos farmacológicos. Sus productos son vitales en la entrada del virus a la célula, la fusión y supervivencia en la célula hospedadora (Tabla 13).

Figura 33

Estructura del genoma de SARS-CoV-2

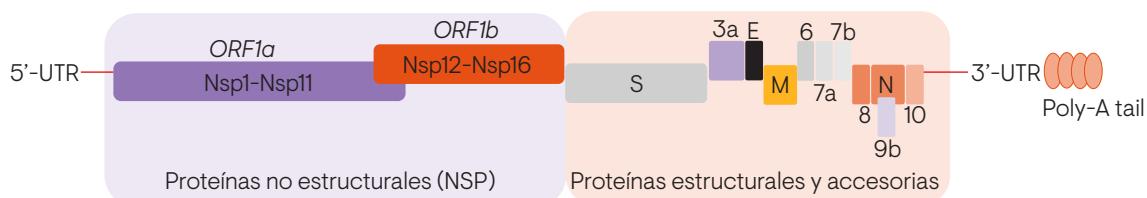


Tabla 13

Lista de proteínas del virus SARS-CoV-2 y sus funciones moleculares

Gen	Proteína	Localización*	ID NCBI	Descripción	Función propuesta
ORF1a	pp1a	266..13483	YP_009725295.1	Poliproteína	
	Nsp1	[1..180]	YP_009725297.1	Producto N-terminal de la replicasa vírica	Media en la replicación del ARN y su procesado. Involucrada en la degradación de ARNm. Inhibición de la traducción.

>>>

>>>

Gen	Proteína	Localización*	ID NCBI	Descripción	Función propuesta
	Nsp2	[181..818]	YP_009725298.1	Producto de replicasa esencial para corrección de la replicación viral	Modulación de la ruta de señalización de supervivencia de la célula hospedadora.
	Nsp3	[819..2763]	YP_009725299.1	Proteasa similar a la papaína.	Proteasa de tipo cisteína esencial para la replicación del virus.
	Nsp4	[2764..3263]	YP_009725300.1	Proteína transmembrana	Proteína de anclaje del complejo de replicación-transcripción.
	Nsp5	[3264....3569]	YP_009725301.1	Proteinasa de tipo 3C	Relacionada con el procesamiento viral de las poliproteínas durante la replicación viral.
	Nsp6	[3570..3859]	YP_009725302.1	Dominio transmembrana	Jugá un papel en la inducción inicial del autofagosoma desde el retículo endoplasmático de la célula hospedadora.
	Nsp7	[3860..3942]	YP_009725303.1	ARN polimerasa dependiente de ARN	Forma un complejo hexadecamerico junto con Nsp8. Adoptan la estructura de un cilindro hueco implicado en la replicación.
	Nsp8	[3943..4140]	YP_009725304.1	Replicasa. ARN polimerasa multimérica.	Forma un complejo hexadecamérico junto con Nsp7. Adoptan la estructura de un cilindro hueco implicado en la replicación.
	Nsp9	[4141..4253]	YP_009725305.1	Proteína viral de unión al ARN de cadena sencilla	Participa en la replicación viral actuando como una proteína de unión al ARN de cadena sencilla.
	Nsp10	[4254..4392]	YP_009725306.1	Proteína de tipo factor de crecimiento	Actúa en la transcripción viral, estimulando la función exoribonucleasa de Nsp14 y actividad metiltransferasa de Nsp16. Además, juega un papel esencial en la metilación de los ARNm.

>>>

>>>

Gen	Proteína	Localización*	ID NCBI	Descripción	Función propuesta
	Nsp11	[4393..4405]	YP_009725312.1	Proteína pequeña (13 aa) idéntica al primer segmento de Nsp12	Desconocido
<i>ORF1ab</i>	pp1ab	266..21555	YP_009724389.1	Poliproteína	
	Nsp12	[4393..5324]	YP_009725307.1	Polimerasa ARN, dependiente de ARN (Pol/RdRp)	Responsable de la replicación y transcripción del genoma del virus.
	Nsp13	[5325..5925]	YP_009725308.1	Dominio de unión Zinc, NTPasa/helicasa, ARN 5'- trifosfatasa	Dominio helicasa que se une a ATP. Dominio de unión a Zinc que se involucra en la replicación y trnascripción.
	Nsp14	[5926..6452]	YP_009725309.1	Dominio exoribonucleasa de corrección de errores (ExoN/nsp14)	Actividad exoribonucleasa actuando en dirección 3' a 5' y con actividad N7-guanina metiltransferasa.
	Nsp15	[6453..6798]	YP_009725310.1	EndoRNAsa	Endoribonucleasa dependiente de Manganese (Mn^{2+}).
	Nsp16	[6799..7096]	YP_009725311.1	Metiltransferasa 2'-O-ribosa	Metiltransferasa que media en el proceso de metilación del ARNm.
S		21563..25384	YP_009724390.1	Proteína estructural	Proteína de superficie, espícula. Unión al receptor ACE2 de la superficie celular para promover la fusión del virus y la célula infectada.
<i>ORF3a</i>		25393..26220	YP_009724391.1		
E		26245..26472	YP_009724391.1	Proteína estructura	Proteína estructural de la envuelta. Formación de poros para el transporte de iones.
M		26523..27191	YP_009724393.1	Proteína estructural	Glicoproteína de membrana. Empaquetamiento del ARN.
<i>ORF6</i>		27202..27387	YP_009724394.1		
<i>ORF7a</i>		27394..27759	YP_009724395.1		

>>>

>>>

Gen	Proteína	Localización*	ID NCBI	Descripción	Función propuesta
ORF7b		27756..27887	YP_009725318.1		
ORF8		27894..28259	YP_009724396.1		
N/ORF9		28274..29533	YP_009724397.2	Proteína estructural	Fosfoproteína nucleocápside
ORF10		29558..29674	YP_009725255.1		

Nota. Los colores de la tabla corresponden a la anotación de la Figura 33. Adaptada y ampliada de “Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach”, por A. A. T. Naqvi, K. Fatima, T. Mohammad, U. Fatima, I. K. Singh, A. Singh, S. M. Atif, G. Hariprasad, G. M. Hasan y M. I. Hassan, 2020, *Biochimica et Biophysica Acta. Molecular Basis of Disease*, 1866(10), p. 165878, recuperado de <https://doi.org/10.1016/J.BBADIS.2020.165878>. Anotada según NCBI referencia NC_045512.2.

*Las regiones indicadas entre corchetes se refieren a las regiones escindidas de la poliproteína correspondiente.

Además de las proteínas estructurales que forman la cápside, el genoma viral codifica para varias proteínas no estructurales, comúnmente denominadas *non-structural proteins* (NSP) que realizan diversas funciones en la replicación y ensamblaje del virus, así como en su patogénesis, modulando la transcripción temprana, función helicasa, inmunomodulación, activación génica y contrarrestar la respuesta antiviral. Las funciones de estas proteínas están especificadas en la Tabla 13.

La **proteína estructural de la espícula (spike, S)** es una glicoproteína vital para la patogénesis, ya que es la proteína de unión al dominio RBD del receptor. En este punto es donde se inicia la infección, cuando el virión se introduce en la célula hospedadora. Se compone de 1273 aminoácidos y contiene tres subunidades: S1, S2 y S2'. El dominio S1 es el encargado en la unión de los viriones a la membrana celular, interaccionando con el receptor de tipo ACE2. En este proceso, la proteína S cambia de conformación. Conocemos que mutaciones en esta proteína provocan cambios conformacionales y, por tanto, una mejor o peor afinidad al receptor al que debe unirse. La subunidad S2 está involucrada en la fusión del virión con la membrana celular de la célula hospedadora. La región RBD de la proteína de la espícula es la región más variable y su comparación con otros coronavirus sugiere que comparten un perfil evolutivo similar. En este caso, SARS-CoV-2 se muestra similar a los coronavirus procedentes de murciélagos HKU3 y SARS-CoV; mientras que MERS-CoV muestra divergencia, quizás por su origen no asociado a murciélagos. Los residuos identificados como críticos en esta región RBD para la unión al receptor ACE2 son Leu455, Phe486, Gln493, Ser494, Asn501 y Tyr505.



Enlace de interés

En el siguiente enlace al artículo *Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach* (Naqvi et al., 2020) observamos en la figura 4 alineamiento múltiple de las regiones (A) RBD de la glicoproteína de la espícula; (B) proteína de envuelta; (C) proteína de membrana; (D) nucleoproteína.

<https://www.sciencedirect.com/science/article/pii/S092544392030226X?via%3Dihub#f0020>

Las **proteínas de membrana (envelope, E)** son un grupo de pequeñas proteínas virales que asisten en el ensamblaje y liberación de los viriones. Es considerada una diana para fármacos ideal, por su pequeño tamaño (75 aa) y su papel fundamental en la morfogénesis y el ensamblaje del virus. Actúa formando unos poros lipídicos que actúan en el transporte de iones. La secuencia de la proteína de membrana entre los cuatro coronavirus principales está altamente conservada.

La **proteína de membrana** (*membrane*, M) tiene una longitud de 222 aa y juega un papel principal en el empaquetamiento del ARN, conteniendo tres dominios transmembrana.

La **proteína de nucleocápside** (N) se encarga del empaquetamiento del ARN viral en la nucleocápside. Media el ensamblaje viral interaccionando con el genoma del virus y la proteína M. Es también considerada una buena diana para los fármacos. Además, si inspeccionamos su conservación en distintos coronavirus, se observa que es una de las proteínas más conservadas entre ellos (aproximadamente 90 % de identidad en la secuencia). Basándose en esta alta similitud de secuencia se piensa que los anticuerpos frente a la proteína N de SARS-CoV deberían reconocer la proteína N de SARS-CoV-2.



Enlaces de interés

En el siguiente enlace podéis encontrar la estructura de cada una de las proteínas de SARS-CoV-2 descritas.

<https://www.ncbi.nlm.nih.gov/Structure/SARS-CoV-2.html>

En el siguiente vídeo se encuentra una explicación del genoma de SARS-CoV-2.

<https://www.youtube.com/watch?v=tOlurlmjwu4>

6.3. Clasificación de SARS-CoV-2. Variantes de interés y linajes. Bases de datos

Basándonos en la secuencia nucleotídica del virus, se han desarrollado varios métodos de tipificación y clasificación de los distintos linajes que han ido apareciendo desde la irrupción de SARS-CoV-2.

6.3.1. Clasificación de la OMS: variantes de preocupación, variantes de interés y variantes bajo vigilancia

Todos los virus cambian con el paso del tiempo, en cada uno de los pasos de replicación, debido fundamentalmente a errores en este proceso. La mayoría de estos cambios tienen escaso o nulo efecto sobre las propiedades del virus. Sin embargo, algunos de ellos pueden influir sobre algunas de ellas, como por ejemplo su facilidad de propagación, la gravedad de la enfermedad asociada o la eficacia de las vacunas, los fármacos para el tratamiento, los medios de diagnóstico u otras medidas de salud pública y social.

La OMS, en colaboración con redes de expertos, autoridades nacionales, instituciones e investigadores, ha estado vigilando y evaluando la evolución del SARS-CoV-2 desde enero de 2020. La aparición de variantes que suponían un mayor riesgo para la salud pública mundial, especialmente a finales de 2020, hizo que se empezaran a utilizar las categorías específicas de variante de interés (VOI) y variante de preocupación (VOC), con el fin de priorizar el seguimiento y la investigación a escala mundial y, en última instancia, orientar la respuesta a la pandemia de COVID-19.

El seguimiento de estos cambios tiene importancia, para que en caso de que se detecten mutaciones significativas, se pueda informar de manera rápida y precisa a los distintos países, para implementar las medidas de contención adecuadas y evitar su propagación. Las medidas recomendadas actualmente por la OMS siguen funcionando, independientemente de la variante circulante, demostrando en países con amplia transmisión de variantes preocupantes que las medidas sociales y de salud pública, como las de prevención y control de la infección, reducen eficazmente el número de casos, hospitalizaciones y muertes por COVID-19.

Los sistemas de nomenclatura establecidos para nombrar y rastrear los linajes genéticos del SARS-CoV-2 por GISAID, Nextstrain y PANGO, que veremos en apartados posteriores, se utilizan en círculos científicos y en la investigación científica. Sin embargo, a nivel de comunicación con la población y para el debate no científico, se ha desarrollado un sistema de nomenclatura basado en el alfabeto griego, con el fin de facilitar su identificación y eliminar los estigmas que puedan derivar de la utilización del lugar de detección de una variante para designarla.



Enlace de interés

Semanalmente la OMS publica una actualización de la clasificación de SARS-CoV-2, la distribución geográfica de las variantes preocupantes y los resúmenes de sus características fenotípicas (transmisibilidad, gravedad de la enfermedad, riesgo de reinfección e impactos en el diagnóstico y la eficacia de la vacuna) basada en los estudios más recientes publicados.

<https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>

Existen dos grupos de variantes con especial interés:

Variante preocupante (*variant of concern*, VOC)

Se trata de una variante que ha mostrado uno o varios cambios significativos en los siguientes criterios:

- Aumento de la transmisibilidad o cambio perjudicial en la epidemiología de la COVID-19.
- O aumento de la virulencia o cambio en la presentación clínica de la enfermedad.
- O disminución de la eficacia de las medidas sociales y de salud pública o de los medios de diagnóstico, las vacunas y los tratamientos disponibles.

En la Tabla 14 se muestran las VOC descritas actualmente (diciembre 2021).



Enlace de interés

Para una actualización a tiempo real se puede consultar:

<https://www.who.int/es/activities/tracking-SARS-CoV-2-variants/tracking-SARS-CoV-2-variants>

Tabla 14

Listado de variantes preocupantes (VOC) actualmente designadas

Denominación OMS	Linaje PANGO	Linaje GISAID	Clado Nexstrain	Primera documentación	Fecha de designación
Alpha	B.1.1.7	GRY	20I (V1)	Reino Unido, septiembre 2020	18/12/2020
Beta	B.1.351	GH/501Y.V2	20H (V2)	Sudáfrica, mayo 2020	18/12/2020
Gamma	P.1	GR/501Y.V3	20J (V3)	Brasil, noviembre 2020	11/01/2021
Delta	B.1.617.2	G/478K.V1	21A, 21I, 21J	India, octubre 2020	VOI: 4/04/2021
VOC: 11/05/2021					
Omicron	B.1.1.529	GR/484A	21K	Varios, noviembre 2021	VUM: 24/11/2021
VOC: 26/11/2021					

Nota. Adaptada de *Tracking SARS-CoV-2 variants*, por Organización Mundial de la Salud, consultado el 12 diciembre 2021. Recuperado de <https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/>

Variante de interés (*variant of interest*, VOI)

Son variantes del SARS-CoV-2 que:

- Presentan cambios en el genoma que, según se ha demostrado o se prevé, afectan a características del virus como su transmisibilidad, la gravedad de la enfermedad que causa y su capacidad para escapar a la acción del sistema inmunitario, ser detectado por medios diagnósticos o ser atacado por medicamentos.
- Y además, según se ha comprobado, dan lugar a una transmisión significativa en medio extrahospitalario o causan varios conglomerados de COVID-19 en distintos países, con una prevalencia relativa creciente y ocasionando números cada vez mayores de casos con el tiempo, o bien que presentan, aparentemente, otras características que indiquen que pueden entrañar un nuevo riesgo para la salud pública mundial.

En la Tabla 15 se muestran las VOC descritas actualmente (diciembre 2021).

Tabla 15

Listado de variantes de interés (VOI) actualmente designadas

Denominación OMS	Linaje PANGO	Linaje GISAID	Clado Nexstrain	Primera documentación	Fecha de designación
Lambda	C.37	GR/452Q.V1	21G	Perú, diciembre 2020	14/06/2021
Mu	B.1.621	GH	21H	Colombia, enero 2021	30/08/2021

Nota. Adaptada de *Tracking SARS-CoV-2 variants*, por Organización Mundial de la Salud, consultado el 12 diciembre 2021. Recuperado de <https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/>

Variante bajo vigilancia (*variant under monitoring*, VUM)

Cualquier variante del SARS-CoV-2 que presente modificaciones en el genoma que, según se sospeche, puedan afectar a las características del virus y parezcan indicar que la variante puede entrañar riesgos en el futuro, a pesar de que no se disponga de pruebas claras de los cambios que pueda causar en el fenotipo o en las características epidemiológicas del virus y sea necesario mantener el seguimiento y continuar estudiándola hasta que no se disponga de más información.



Enlaces de interés

En el siguiente enlace se encuentra un vídeo sobre las variantes de interés.

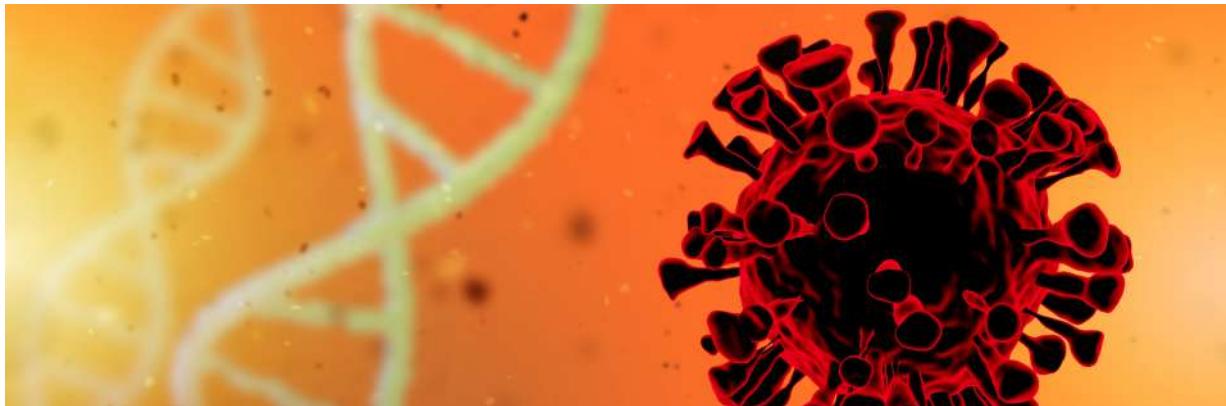
<https://www.youtube.com/watch?v=hxkoePuNUJw>

Un vídeo sobre las bases de datos que veremos en apartados posteriores.

<https://www.youtube.com/watch?v=nuecT8RMCyM>

Un vídeo explicando la importancia de la epidemiología genómica.

<https://www.youtube.com/watch?v=njSOTw4uQUo>



6.3.2. Monitorización de la evolución de SARS-CoV-2 a tiempo real. Nextstrain

Nexstrain (<https://nextstrain.org/>) es un proyecto de código abierto que fue diseñado para aprovechar el potencial científico y de salud pública de los datos de genomas de distintos patógenos. Se basa en la actualización continua de los genomas que están disponibles de manera pública, ofreciendo una visualización y análisis potente. Su objetivo es facilitar la comprensión epidemiológica y mejorar la respuesta frente a brotes epidémicos. Contienen conjuntos de datos para los patógenos de tuberculosis, dengue, ébola, enterovirus, gripe, virus de Nilo y COVID-19, entre otros.



Todo su trabajo se basa en una premisa: durante el curso de una infección y una epidemia, los patógenos acumulan de manera natural mutaciones aleatorias en sus genomas. Esta es una consecuencia inevitable de la replicación del genoma, que es propenso a sufrir errores. Distintos genomas acumulan distintas mutaciones, por lo que estas pueden utilizarse como marcador de transmisión, donde los genomas estrechamente relacionados indican infecciones estrechamente relacionadas.

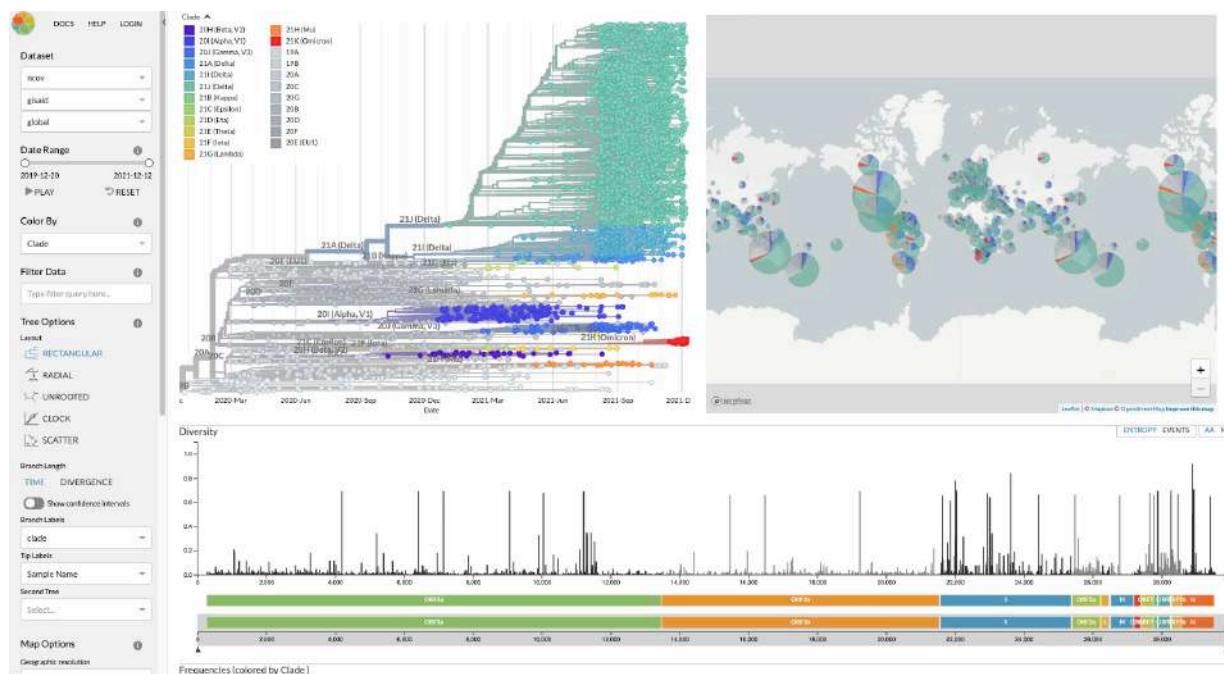
La reconstrucción del árbol filogenético, relacionando los distintos genomas, nos permite estudiar fenómenos epidemiológicos importantes como la dispersión espacial, los tiempos de introducción y la tasa de crecimiento epidémico.

Si queremos que este análisis de secuencias del genoma del patógeno sirva para informar sobre las intervenciones de salud pública, debe realizarse rápidamente y los resultados deben difundirse ampliamente. Las prácticas actuales de publicación científica dificultan la rápida difusión de resultados epidemiológicamente relevantes. Por ello, esta web implementa un sistema en línea abierto con protocolos bioinformáticos sólidos para extraer datos de todos los grupos de investigación, mejorando la capacidad de realizar inferencias epidemiológicamente relevantes.

Por lo tanto, esta web (Figura 34) tiene como objetivo proporcionar una instantánea en tiempo real de la evolución de las poblaciones de patógenos y proporcionar visualizaciones de datos interactivas a virólogos, epidemiólogos, funcionarios de salud pública y científicos. A través de visualizaciones de datos interactivas, su objetivo es permitir la exploración de conjuntos de datos en actualización continua, proporcionando una nueva herramienta de vigilancia a las comunidades científicas y de salud pública.

Figura 34

Captura de la web Nextstrain, focalizada en el análisis y trazado de genomas de SARS-CoV-2



Nota. Tomado de *Genomic epidemiology of novel coronavirus - Global subsampling*, por Nextstrain, consultado el 12 diciembre 2021. Recuperado de <https://nextstrain.org/ncov/gisaid/global>

La nomenclatura de los linajes de SARS-CoV-2 es variable en función de la base de datos utilizada. Como se puede observar en el árbol filogenético de la Figura 34, está basado en un código que determina el año (19, 20 o 21), seguido de una letra que determina el linaje exacto. En algunos casos se especifica su correlación con la nomenclatura proporcionada por la OMS, determinada por letras griegas (Alpha-Omicron, en la fecha actual). Para más información sobre el sistema de nomenclatura de Nextstrain, puedes consultar: <https://nextstrain.org/blog/2021-01-06-updated-SARS-CoV-2-clade-naming>.



Enlace de interés

En el siguiente enlace se encuentra un vídeo sobre el uso de Nextstrain.

<https://www.youtube.com/watch?v=aubtBo-dAhw>

6.3.3. GISAID. El repositorio de todos los genomas SARS-CoV-2 secuenciados

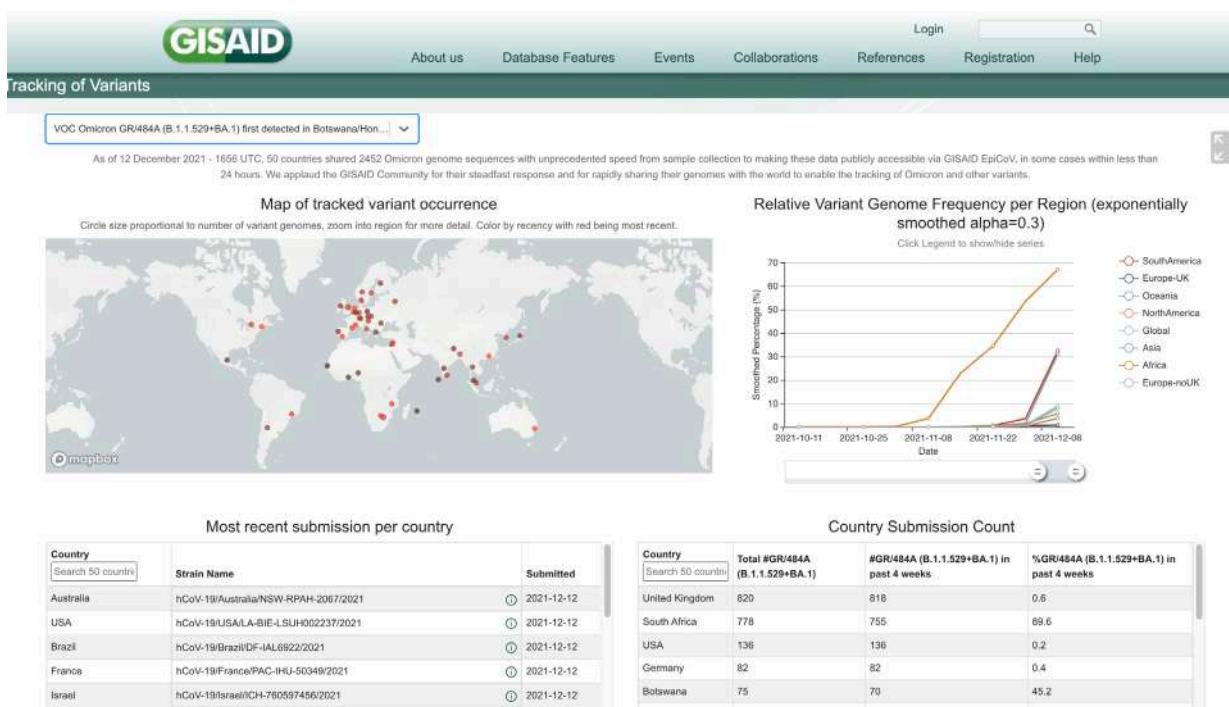
La iniciativa GISAID (<https://www.gisaid.org/>) promueve el intercambio rápido de datos de todos los virus Influenza y SARS-CoV-2. Esto incluye la secuencia genética y los datos clínicos y epidemiológicos asociados con virus humanos, y datos geográficos y específicos de especies asociados con virus aviares y otros virus animales, para ayudar a los investigadores a comprender cómo evolucionan y se propagan los virus durante epidemias y pandemias. En la actualidad (12 diciembre 2021), GISAID alberga más de seis millones de secuencias de SARS-CoV-2, que sirven de fuente de información para otras bases de datos como Nextstrain, anteriormente citada.

Una vez que se accede con registro a la base de datos, cada investigador puede remitir la secuencia nucleotídica del genoma de SARS-CoV-2 secuenciado en su laboratorio, junto con datos epidemiológicos básicos como la edad, sexo, condición y localización del paciente del que ha sido extraído el virus. Asimismo, se pueden depositar secuencias de SARS-CoV-2 aisladas de animales. Existe una gran cantidad de información en el apartado siguiente: <https://www.epicov.org/epi3/frontend#326ae6>

Adicionalmente, se pueden consultar todas las secuencias depositadas junto con estos datos epidemiológicos, seleccionando por localización, linaje, fecha, etc.; así como la dinámica de aparición y transmisión de variantes de interés (Figura 35).

Figura 35

Captura de la web GISAID, focalizada en el genoma y datos epidemiológicos de SARS-CoV-2



Nota. En esta captura se muestra la monitorización para la variante ómicron, detectada por primera vez en Botsuana a finales de noviembre de 2021. Tomado de *Tracking of Variants*, por GISAID, consultado el 12 diciembre 2021. Recuperado de <https://www.gisaid.org/hcov19-variants/>

6.3.4. Clasificación de linajes y constelaciones de mutaciones: PANGO

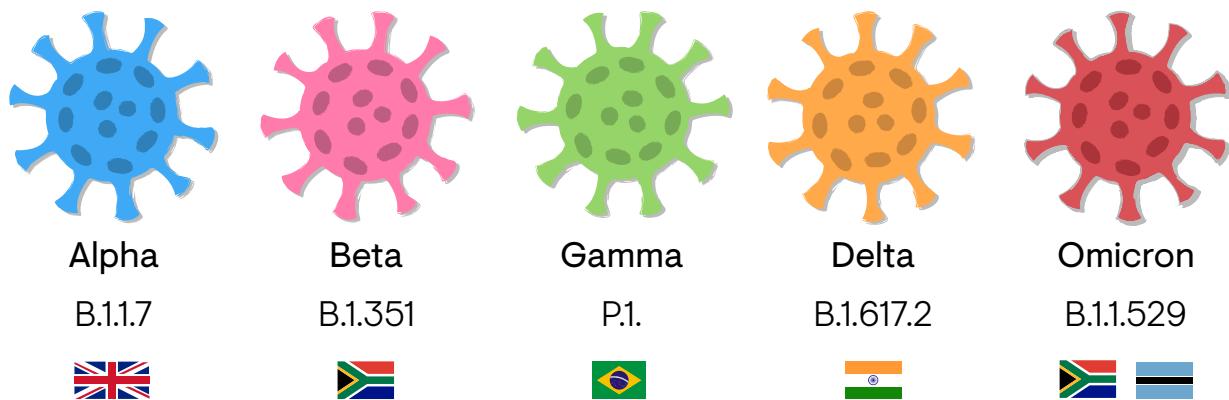
El sistema de clasificación más extendido para la monitorización de linajes de SARS-CoV-2 es el sistema PANGO, basado en un sistema de asignación filogenética y una nomenclatura basada en la combinación correlativa de letras y números.



Enlace de interés

La siguiente página web documenta todos los linajes actuales de SARS-CoV-2, con parámetros relacionados con su aparición y propagación, así como varias herramientas de software que los investigadores pueden utilizar para realizar análisis sobre la secuencia del virus.

<https://cov-lineages.org/>



Enlace de interés

La explicación del sistema de clasificación de linajes está detallada, junto con las últimas novedades en él, en la web:

<https://www.pango.network/>

Este sistema fue propuesto por primera vez en abril 2020 (Rambaut *et al.*, 2020). El sistema es dinámico y flexible, pudiéndose adaptar a la naturaleza cambiante de esta pandemia y crecer conforme crezcan las evidencias genómicas de SARS-CoV-2. Cada linaje de PANGO define un grupo de secuencias del genoma de SARS-CoV-2 y se crea de acuerdo a dos principios:

1. Los linajes PANGO significan **grupos o clústeres de infecciones** que comparten un ancestro común. Si pensamos en la pandemia como un árbol que va ramificándose en cada transmisión, los linajes PANGO representan las ramas individuales dentro de ese árbol.
2. Los linajes PANGO están destinados a resaltar **eventos epidemiológicamente relevantes**, como la aparición del virus en una nueva ubicación, un rápido aumento de casos o la evolución del virus hacia fenotipos distintos.



Enlace de interés

Los criterios actuales para la creación de nuevos linajes PANGO constan de una serie de criterios, definidos en el siguiente enlace y que incluyen unos estándares mínimos de tamaño de linaje, calidad del genoma, distinción genética e importancia epidemiológica. Estos criterios son cambiantes en el tiempo, para adaptarse a las necesidades y circunstancias cambiantes de esta pandemia.

<https://www.pango.network/the-pango-nomenclature-system/statement-of-nomenclature-rules/>

Este sistema jerárquico de nomenclatura se refleja en la forma de nombrar cada linaje. Cada uno de ellos recibe un código alfanumérico único que incluye información parcial, pero no completa, sobre la historia filogenética del mismo. Esta nomenclatura es un compromiso entre la comprensión humana y la legibilidad para los sistemas informáticos.



Enlace de interés

Los linajes pango actuales se muestran en el apartado (Figura 36):

https://cov-lineages.org/lineage_list.html

Figura 36

Captura del listado de linajes PANGO

Lineage List

Lineage	Most common countries	Earliest date	# designated	# assigned	Description	WHO Name
A	United States of America 27.0%, United_Arab_Emirates 13.0%, China 9.0%, Germany 8.0%, Japan 5.0%	2019-12-30	1698	2224	Root of the pandemic lies within lineage A. Many sequences originating from China and many global exports; including to South East Asia Japan South Korea Australia the USA and Europe represented in this lineage	
A1	United States of America 81.0%, Australia 9.0%, Canada 6.0%, United Kingdom 1.0%, Iceland 1.0%	2020-01-01	2699	3035	USA lineage	
A2	Spain 71.0%, United Kingdom 6.0%, Panama 5.0%, United States of America 4.0%, Portugal 2.0%	2020-02-17	1107	1255	Mostly Spanish lineage now includes South and Central American sequences, other European countries and Kazakhstan.	
A2.2	Australia 92.0%, New_Zealand 3.0%, Canada 2.0%	2020-	473	547	Australian lineage	

Nota. Tomado de Lineage List, en cov-lineages.org, consultado el 12 diciembre 2021.

Recuperado de https://cov-lineages.org/lineage_list.html

La clasificación PANGO propone un **sistema de constelaciones de mutaciones**, donde se entiende por constelación un conjunto de mutaciones que son significativas y que han podido emerger de manera independiente, en procesos evolutivos independientes, varias veces.

**Enlaces de interés**

La definición y detalles de este sistema se encuentra en:

<https://github.com/cov-lineages/constellations>

Las constelaciones definidas se encuentran en el siguiente enlace (Figura 37) y comprenden aquellos linajes de preocupación y regiones RBD de interacción con el receptor ACE2 de interés por sus mutaciones.

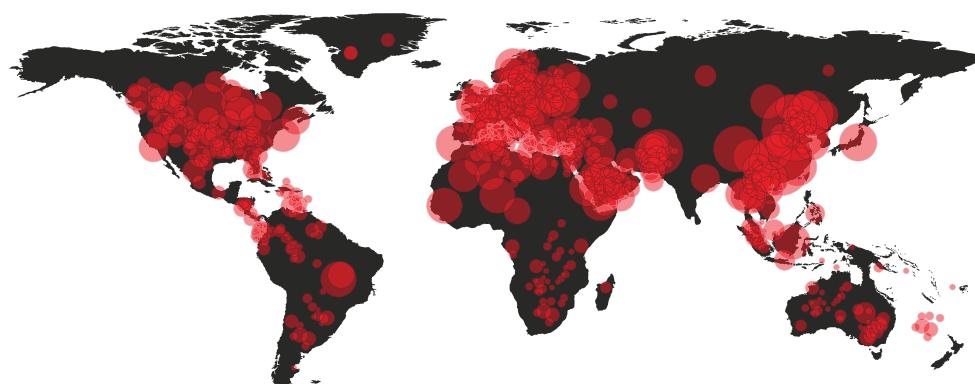
<https://cov-lineages.org/constellations.html>

Figura 37

Captura del listado de constelaciones PANGO definidas

Summary of Currently Defined Constellations							
Label	Description	Sources	Type	Variant	Tags	Sites	Rules
A.23.I-like	A.23.I lineage defining mutations	https://doi.org/10.1101/2021.02.08.21251393	variant	Pango_lineage: A.23.I mrca_lineage: A.23.I representative_genome:	A.23.I	nsp3:L74IF, nsp6:M86I, nsp6:L98F, nsp6:M183L, S:R102L, S:F157L, S:V367F, S:Q613H, S:P681R, ORF8:L84S; ORF8:E92K, N:S202N	min_alt: 5 max_ref: 3
B.1.1.318-like	Defining of lineage B.1.318	https://github.com/cov-lineages/pango-designation/issues/15	variant	Pango_lineage: B.1.1.318 mrca_lineage: B.1.1.318 PHE_label: B.1.1.318-04 representative_genome:	B.1.1.318, VUI-21FEB-04, VUI202102/04	nuc:C3961T, nsp3:K1693N, nsp4:I173L, nsp4:A446V, nsp5:L21L, del:L287:9, nsp15:V320M, S:796L, del:219903, S:484K, nuc:T23287C, S:P681H, S:D796H, nuc:C25276A, M:182T, del:27897:15, ORF8:E108*, nuc:A28271G, del:28895:3	min_alt: 11 max_ref: 3
Beta (B.1.351-like)	Defining of lineage B.1.351	https://www.medrxiv.org/content/10.1101/2020.12.21.20248640v1	variant	Pango_lineage: B.1.351 mrca_lineage: B.1.351 PHE_label: VOC-20DEC-02 WHO_label: Beta representative_genome:	B.1.351, VOC-20DEC-02, VOC202012/02, Beta, S:O1.V2, 20H, GH	NSP2:T85I, ORF1ab:K1655N, ORF1ab:K3353R, S:D80A, S:D216G, S:E484K, S:N501Y, S:A701V, ORF3a:Q57H, ORF3a:S17L, E:P71L, N:T205I, del:L22280:9, del:L287:9	min_alt: 6 max_ref: 3

Nota. Tomado de *Constellations*, en cov-lineages.org, consultado el 12 diciembre 2021.Recuperado de https://cov-lineages.org/lineage_list.html**Enlaces de interés**

Herramientas software que incluye PANGO y que ayudan a los investigadores en el análisis y clasificación de sus secuencias:

<https://cov-lineages.org/resources.html>

Finalmente, PANGO incluye una serie de herramientas software que ayudan a los investigadores en el análisis y clasificación de sus secuencias. Entre ellas destacan:

- **Pangolin** (Phylogenetic Assignment of Named Global Outbreak Lineages) (O'Toole *et al.*, 2021). Esta herramienta fue diseñada para implementar esta nomenclatura dinámica, bien en una herramienta por la línea de comandos (CLI) como con una aplicación web. Asiste al usuario en la asignación del linaje más probable. Disponible en:

<https://cov-lineages.org/resources/pangolin.html>

- **Scorpio** (Serious Constellations of Reoccurring Phylogenetically-Independent Origin). Es una herramienta, contenida en el programa Pangolin por línea de comandos que permite analizar los SNP de las variantes de preocupación y asignar la constelación a la que pertenecen. Disponible en:

<https://github.com/cov-lineages/scorpio>

- **Pando**. Permite la interacción con el árbol filogenético de SARS-CoV-2 global. Disponible en:

<http://pando.tools/>

- **Civet** (Cluster Investigation and Virus Epidemiology Tool). Este software se ha diseñado pensando en un análisis genómico a tiempo real. Utilizando una filogenia de fondo, como la disponible a través del consorcio COG-UK en CLIMB, este software genera un informe para un conjunto de secuencias de interés, permitiendo el análisis de un brote en detalle. Civet coloca las nuevas secuencias en el contexto de la diversidad de fondo, que es conocida. Disponible en:

<https://cov-lineages.org/resources/civet.html>

- **Polecat** (Phylogenetic Overview & Local Epidemiological Cluster Analysis Tool). De manera similar a Civet, utilizando una filogenia de fondo, identificará y marcará los grupos o clústeres. Disponible en:

<https://github.com/artic-network/polecat>

6.3.5. El análisis de mutaciones de interés. Bases de datos de monitorización

Existen varios recursos web que actúan como cuadros de mando para la monitorización de los linajes de interés VOC, VOI y VUM, así como de los datos de rastreo epidemiológico disponibles. Algunas de ellas se mencionan a continuación, como referencia para los estudios de los genomas:

- **Outbreak.info**. Especialmente útil para visualizar la comparativa entre las mutaciones de distintos linajes de interés, así como datos sobre su expansión en todo el mundo (Figura 38; Figura 39). Disponible en:

<https://outbreak.info/>

- **CoVariants** (Figura 40). Además de la monitorización ofrecida por país y por variante, muestra una tabla interesante sobre la conversión de las nomenclaturas de los linajes entre los clados Nextstrain, linaje PANGO y la etiqueta de la OMS (Figura 41). Asimismo, se encuentra una tabla muy útil sobre las mutaciones compartidas entre distintos linajes (Figura 42). Disponible en:

<https://covariants.org/>

- **covSPECTRUM**. Es un cuadro de mandos interactivo donde se muestran las variantes del virus circulantes en cada país (Figura 43). Disponible en:

<https://cov-spectrum.org/explore/Spain/AllSamples/AllTimes>

- **JHU Covid-19 Dashboard**. Cuadro de mandos desarrollado por la universidad Johns Hopkins, donde se encuentran datos epidemiológicos actualizados de todo el mundo, incluyendo el acumulado de casos totales, fallecimientos totales, vacunas administradas e incidencia (Figura 44). Disponible en:

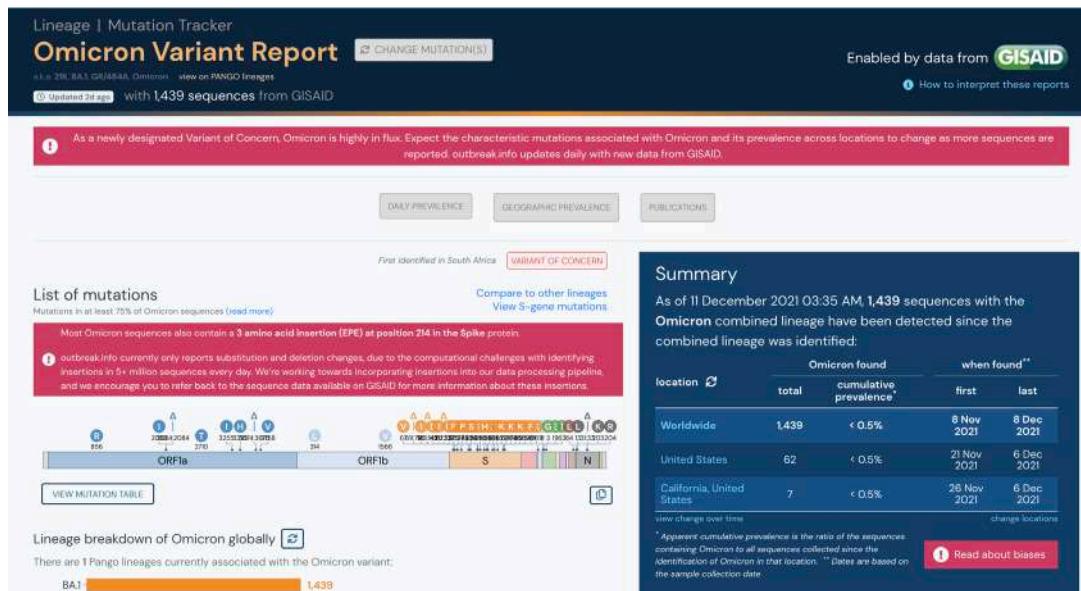
<https://coronavirus.jhu.edu/map.html>

- **COG Mutation Explorer**. Desarrollado por el consorcio británico de secuenciación de SARS-CoV-2, líder indiscutible en la monitorización genómica del virus, proponen una herramienta visual para explorar las variantes y mutaciones relevantes (Figura 45). Disponible en:

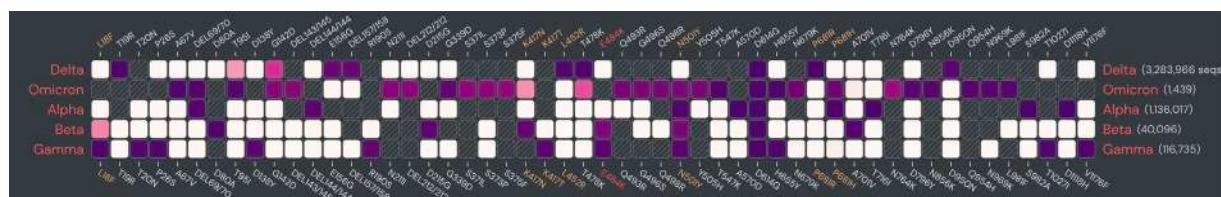
<https://sars2.cvr.gla.ac.uk/cog-uk/>

Figura 38

Página web Outbreak.info; visualización general

**Figura 39**

Comparativa de las mutaciones sobre la proteína de la espícula (S) en las VOC, según Outbreak.info

**Figura 40**

Página web CoVariants; visualización general

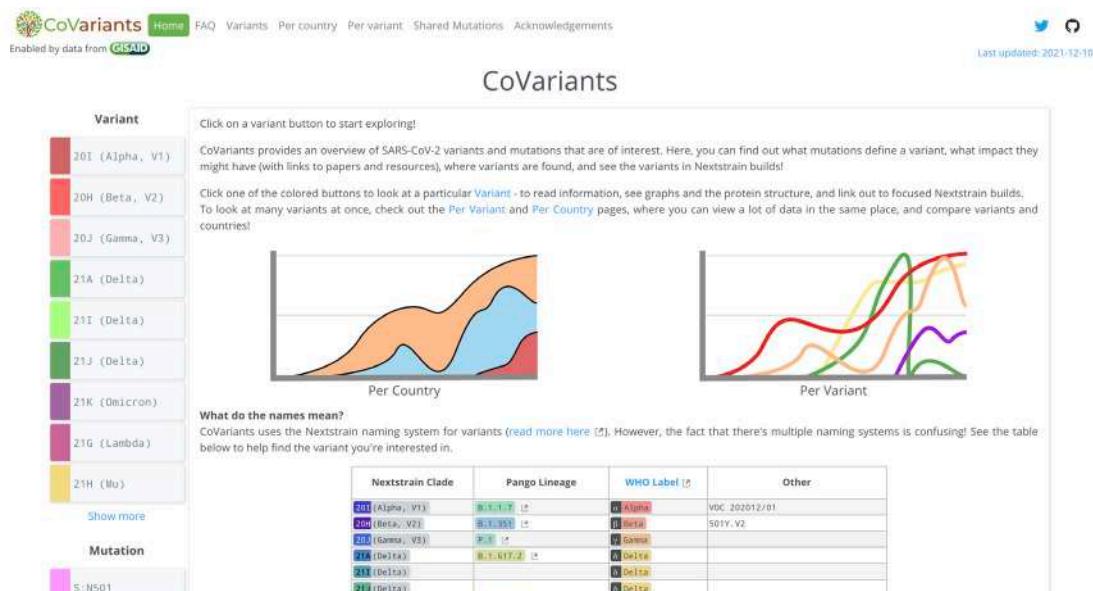


Figura 41

Comparativa de la nomenclatura de distintos linajes en varios sistemas diferentes, proporcionado por CoVariants

Nextstrain Clade	Pango Lineage	WHO Label	Other
20I (Alpha, V1)	B.1.1.7	α Alpha	VOC 202012/01
20H (Beta, V2)	B.1.351	β Beta	S01Y.V2
20J (Gamma, V3)	P.1	γ Gamma	
21A (Delta)	B.1.617.2	δ Delta	
21I (Delta)		δ Delta	
21J (Delta)		δ Delta	
21B (Kappa)	B.1.617.1	κ Kappa	
21C (Epsilon)	B.1.427, B.1.429	ε Epsilon	CAL.20C
21D (Eta)	B.1.525	η Ηta	
21F (Iota)	B.1.526	ι Iota	(Part of Pango lineage)
21G (Lambda)	C.37	λ Lambda	
21H (Mu)	B.1.621	μ Mu	
21K (Omicron)	B.1.1.529	ο Omicron	
20E (EU1)	B.1.177		EU1
20B/S: 732 A	B.1.1.519		
20A/S: 126 A	B.1.620		
20A/EU2	B.1.160		
20A/S: 439 K	B.1.258		
20A/S: 98 F	B.1.221		
20C/S: 80 Y	B.1.367		
20B/S: 626 S	B.1.1.277		
20B/S: 1122 L	B.1.1.302		

Nota. Recuperado de web CoVariants.

Figura 42

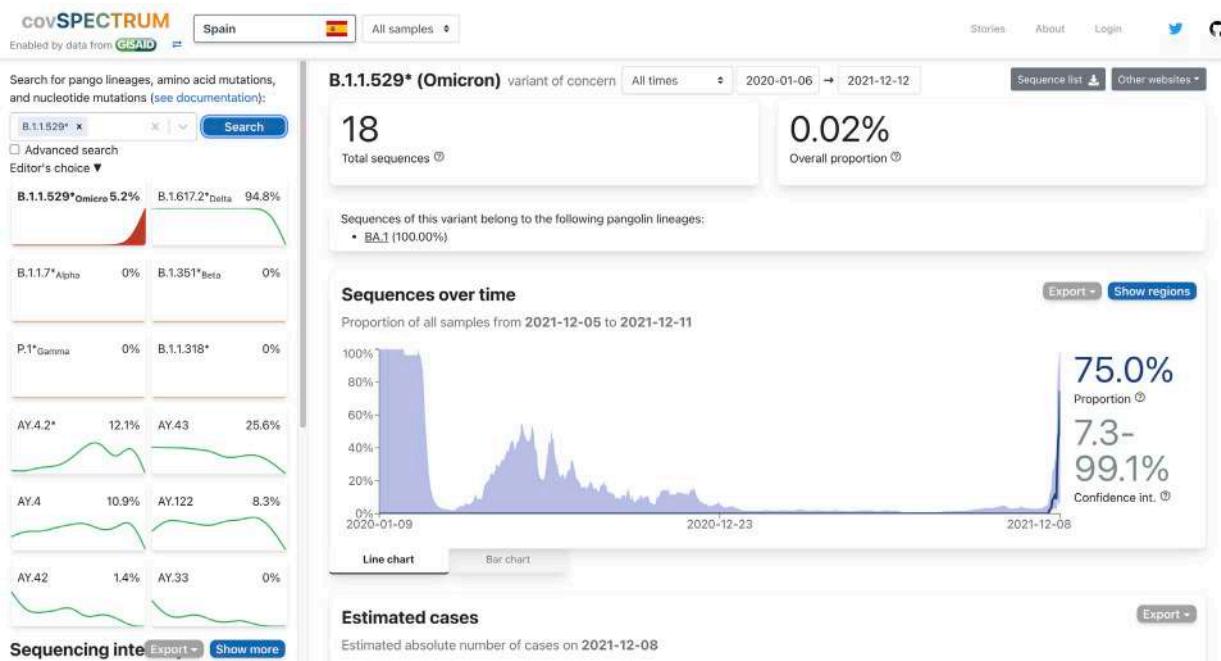
Comparativa de las mutaciones compartidas entre distintos linajes de interés, proporcionado por CoVariants



Nota. Recuperado de la web CoVariants.

Figura 43

Página web covSPECTRUM; visualización general

**Figura 44**

Página web JHU; visualización general

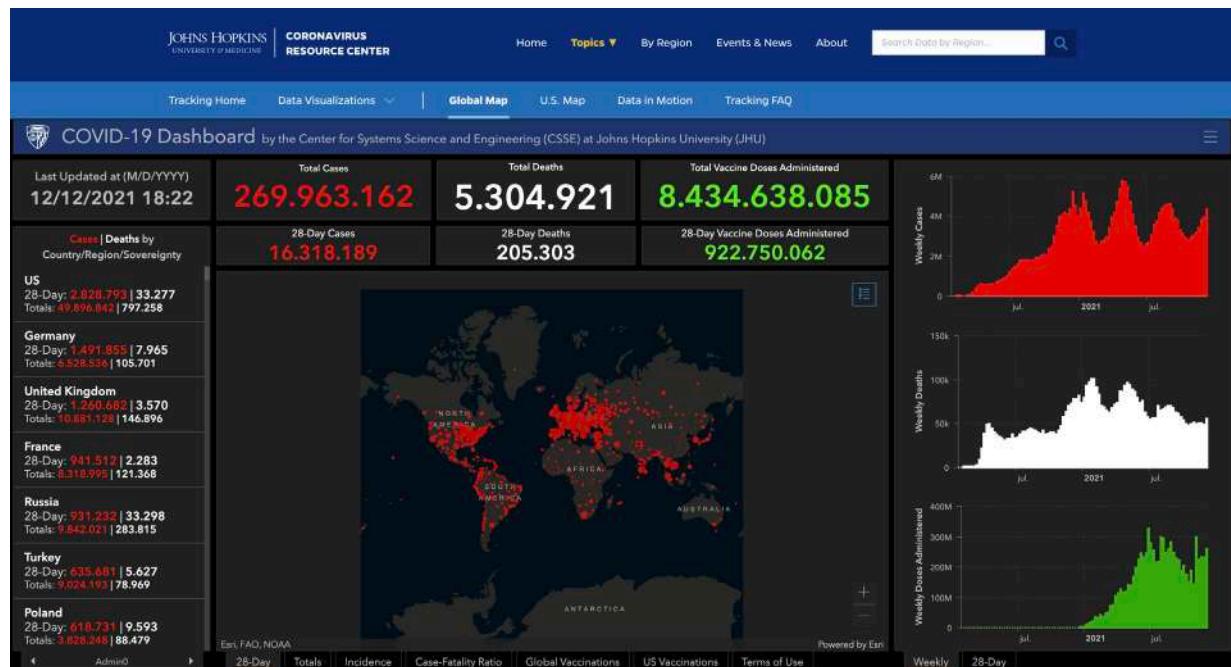
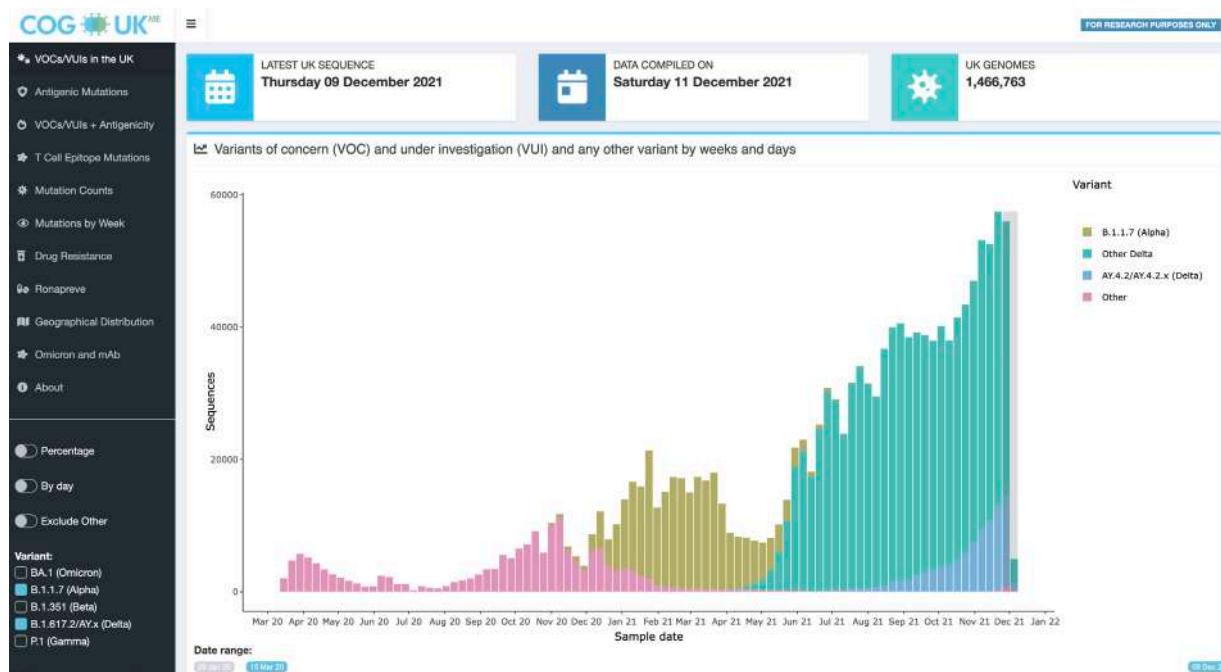


Figura 45

Página web COG Mutation Explorer; visualización general



6.4. Análisis bioinformático de linajes de SARS-CoV-2

En este último apartado se va a realizar el análisis bioinformático de dos maneras alternativas del genoma de SARS-CoV-2.

Lo primero que debemos tener en cuenta es que existen dos maneras principales de analizar una muestra nasofaríngea de un paciente infectado por SARS-CoV-2.

- En primer lugar, se trataría de realizar una secuenciación completa del ARN total contenido en la muestra, para posteriormente ensamblar *de novo* e identificar los *contigs* ensamblados. En este método, se llama **secuenciación shotgun** o en este caso, **metatranscriptoma**, por analizar el ARN total de una muestra. La mayor problemática es que la mayor parte de la muestra analizada corresponderá a ARN del hospedador (humano), e incluso a otros virus que puedan estar contenidos en la muestra y formar parte del viroma del individuo. Esta fue la técnica que se empleó para el análisis de las primeras muestras de SARS-CoV-2 detectadas, a partir de las cuales se obtuvo este nuevo patógeno.
- Actualmente, y de manera rutinaria, sometemos el ARN total extraído del hisopo nasofaríngeo a un preprocessamiento para enriquecer nuestra secuenciación sobre el virus de interés. A grandes rasgos, se realiza en primer lugar una retrotranscripción para obtener el ADN complementario, seguido de dos PCR independientes, con cebadores diseñados sobre el genoma de referencia modelo. De esta forma, se amplifica el genoma del virus en dos *pools* independientes que son purificados y unidos en un único *pool*. En este paso se procura enriquecer la mezcla en el genoma diana de estudio, para evitar contaminaciones por ARN humano u otros virus. Finalmente, se adicionan adaptadores y los índices de secuenciación. Este es conocido como **protocolo ARTIC** y está siendo ampliamente utilizado por equipos de análisis genómico de todo el mundo, por su bajo coste, versatilidad y adaptabilidad (se van adaptando los cebadores para no perder sensibilidad con las nuevas mutaciones que van apareciendo).



Enlace de interés

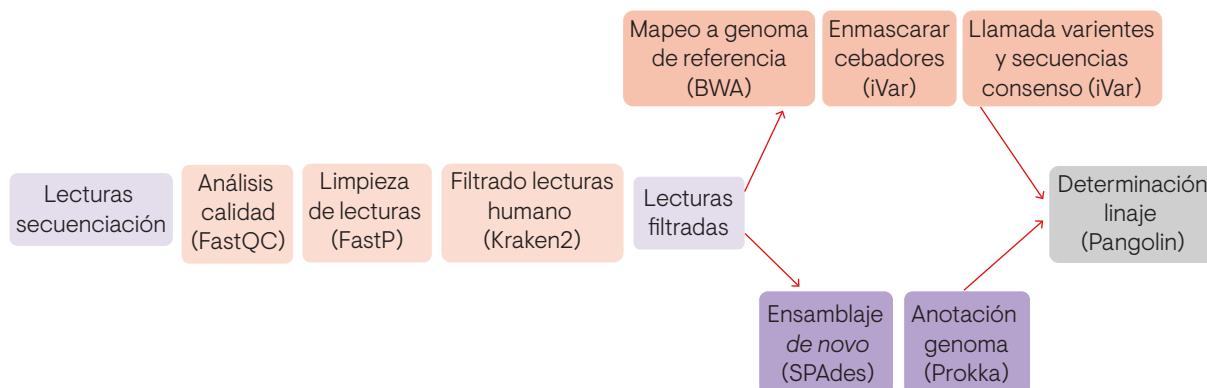
El protocolo ARTIC de secuenciación está disponible en el siguiente enlace, con actualizaciones semanales.

<https://artic.network/ncov-2019>

El análisis bioinformático que a continuación se expone está representado en la Figura 46.

Figura 46

Protocolo de análisis de SARS-CoV-2



6.4.1. Limpieza de las lecturas crudas. Eliminación del genoma del hospedador

Como en cualquier protocolo bioinformático, los pasos iniciales constan del análisis de calidad de las lecturas, así como su limpieza de adaptadores y regiones de baja calidad. En este caso, debido a la naturaleza de la muestra, en la que se ha podido arrastrar ARN del hospedador (humano), se realiza un paso de eliminación de este, mediante la herramienta bioinformática Kraken2 (Kraken2, s. f.; Wood et al., 2019) (Figura 46).



Ejemplo

Los pasos iniciales para realizar son los siguientes:

1. Análisis de calidad con FastQC sobre las lecturas crudas obtenidas del secuenciador.

```
fastqc *.fastq.gz
```

</>

2. Limpieza de adaptadores y regiones de baja calidad con FastP.

```
fastp --in1 SARS_CoV_2_Sample1.R1.fastq.gz --in2 SARS_CoV_2_
Sample1.R2.fastq.gz --out1 Sample1.out.R1.fq.gz --out2 Sample1.
out.R2.fq.gz --detect_adapter_for_pe --cut_tail --cut_window_size
10 --cut_mean_quality 30 --thread 2 --html Sample1.out.fastp.html
```

>>>

>>>

3. Filtrado con Kraken2 del genoma humano. La base de datos que utilizar es la que está precompilada en la web de Kraken2 (Kraken2, s. f.), disponible a través del enlace de descarga. Su tamaño es de 5.5 Gb.

- a. Mapeo Kraken2 a la base de datos, donde se clasifican todas las lecturas de la muestra.

```
kraken2 --db ../minikraken2_v2_8GB_201904_UPDATE/ --paired
Sample1.out.R1.fq.gz Sample1.out.R2.fq.gz --output Sample1.
kraken --threads 2 --gzip-compressed
```

Nota: este paso es computacionalmente intensivo. Si no es posible realizarlo en local, se proporciona el archivo de salida para su uso.

- b. De este archivo de salida de texto, extraemos las **lecturas que no son de humano**, con la herramienta awk.

```
awk '$3 != "9606" { print $2 }' Sample1.kraken > Sample1.
kraken.nohuman.ids
```

- c. Ahora sacamos las lecturas a un nuevo archivo FASTQ y comprimimos.

```
seqtk subseq Sample1.out.R1.fq.gz Sample1.kraken.nohuman.ids
> Sample1.R1.nonhuman.fq

seqtk subseq Sample1.out.R2.fq.gz Sample1.kraken.nohuman.ids
> Sample1.R2.nonhuman.fq

gzip *.fq
```

Estas lecturas sin restos de humano son nuestros archivos limpios para continuar el proceso.

6.4.2. Mapeo al genoma de referencia y determinación del linaje

La metodología a continuación expuesta trata de un mapeo sobre el genoma de referencia de SARS-CoV-2, para posteriormente realizar una llamada de variantes y extraer una secuencia consenso mediante el programa iVar (iVar: Manual, s. f.). Finalmente, se utilizará la herramienta Pangolin (GitHub - cov-lineages/pangolin: Software package for assigning SARS-CoV-2 genome sequences to global lineages., s. f.; O'Toole et al., 2021) para determinar el linaje del virus secuenciado (Figura 46).



Ejemplo

1. Descarga del genoma de referencia, desde NCBI en versión FASTA.
https://www.ncbi.nlm.nih.gov/assembly/GCF_009858895.2/

>>>

>>>

2. Indexado del genoma de referencia:

```
bwa index GCF_009858895.2_ASM985889v3_genomic.fna
```



3. Mapeo de las lecturas limpias al genoma de referencia. En este comando complejo se incluye la reorganización del archivo SAM de salida y su conversión a archivo BAM.

```
bwa mem -Y -M -R '@RG\tID:\tSM:.' -t 2 GCF_009858895.2_ASM985889v3_genomic.fna Sample1.R1.nonhuman.fq.gz Sample1.R2.nonhuman.fq.gz | samtools sort | samtools view -F 4 -b -@ 2 -o Sample1.sort.bam
```



4. Indexado del archivo BAM.

```
samtools index Sample1.sort.bam
```



5. Descargamos los cebadores ARTIC v3 (utilizados en esta versión de la preparación del protocolo). Se descarga un archivo de tipo BED desde:

https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019/V3

6. Con el programa iVar se realiza la eliminación/enmascaramiento de estos *primers* para que no interfieran en la determinación de variantes.

```
ivar trim -i Sample1.sort.bam -p Sample1.trim -m 30 -q 20 -s 4 -b nCoV-2019.primer.bed
```



7. Orden e indexado de los archivos de salida.

```
samtools sort Sample1.trim.bam -o Sample1.trim.sort.bam
```



```
samtools index Sample1.trim.sort.bam
```



8. Realizamos la llamada de variants con iVar. Necesitaremos el archivo FASTA de la referencia, pero también su anotación en formato GFF.

```
samtools mpileup -A -d 0 --reference GCF_009858895.2_ASM985889v3_genomic.fna -B -Q 0 Sample1.trim.sort.bam | ivar variants -p Sample1 iVar
```



>>>

>>>

9. Extraemos la secuencia consenso con iVar.

```
samtools mpileup -aa -A -d 0 -B -Q 0 Sample1.trim.sort.bam
| ivar consensus -p Sample1.ivar -q 20 -t 0.8 -m 30 -n N
```

10. Determinamos el linaje con Pangolin.

```
pangolin Sample1.ivar.fa -o Sample1.pangolin --outfile Sample1.csv -t 2 --max-ambig 0.3
```

Importante: si se instala Pangolin vía Anaconda como (bioconda -c bioconda pangolin) la versión instalada es antigua (v1.1.14). Debemos instalar un nuevo environment conda con Pangolin actualizada (v3.1.17) según se especifica en <https://cov-lineages.org/resources/pangolin/installation.html> y actualizarlo según se indica en <https://cov-lineages.org/resources/pangolin/updating.html>, como sigue:

```
git clone 'https://github.com/cov-lineages/pangolin.git'
cd pangolin
conda env create -f environment.yml
conda activate pangolin
pip install .
pangolin -update
```

6.4.3. Ensamblaje de novo y anotación

Finalmente, de manera alternativa, se va a realizar el ensamblaje *de novo* y anotación del genoma secuenciado. Para ello se utilizará SPAdes en su versión para coronavirus para el ensamblaje y Prokka para la anotación. En ambos casos realizaremos ensamblaje y anotación dirigidas (Figura 46).



Ejemplo

1. Utilizando las lecturas limpias de regiones de baja calidad y restos de ARN de hospedador, realizamos su ensamblaje con SPAdes. Utilizamos como genoma de confianza el de referencia de SARS-CoV-2.

```
spades.py --corona -1 Sample1.R1.nonhuman.fq.gz -2 Sample1.R2.nonhuman.fq.gz -t 2 --trusted-contigs GCF_009858895.2_ASM985889v3_genomic.fa -o covid.spades
```

>>>

>>>

2. Realizaremos la anotación de los *scaffolds* obtenidos con Prokka. Para poder dirigir la anotación, utilizamos las proteínas del genoma de referencia, que podemos descargar desde NCBI:

https://www.ncbi.nlm.nih.gov/assembly/GCF_009858895.2/

```
prokka --outdir Sample1.prokka --addgenes --kingdom Viruses  
--compliant --proteins GCF_009858895.2_ASM985889v3_protein.faa  
covid.spades/scaffolds.fasta
```

</>



Glosario

ADN

Sigla relativa a ácido desoxirribonucleico. Contiene las instrucciones genéticas usadas en el desarrollo y funcionamiento de todos los organismos vivos y algunos virus; también es responsable de la transmisión hereditaria.

ADN mitocondrial (ADNmt)

Material genético presente en las mitocondrias, los orgánulos responsables de generar energía para la célula.

Ácido ribonucleico (ARN)

Ácido nucleico formado por una cadena de ribonucleótidos. Está presente tanto en las células procariotas como en las eucariotas y es el único material genético de ciertos virus (los virus ARN).

ARN ribosomal (ARNr)

ARN que forma parte de los ribosomas y es esencial para la síntesis proteica en todos los seres vivos.

ARN de transferencia o ARN trasnferente (ARNt)

Ácido ribonucleico que tiene la función de síntesis de proteínas. Es aquel que transfiere las moléculas de aminoácidos a los ribosomas, para posteriormente ordenarlos en la molécula de ARN mensajero. Estos aminoácidos transferidos se unirán ordenadamente para formar las proteínas.

Bacteriófago

Virus que infecta exclusivamente a bacterias. Son también llamados fagos.

Cariotipado

Prueba llevada a cabo en el ámbito de la genética clínica para examinar cromosomas en una muestra de células. Este examen ayuda a identificar problemas genéticos como la causa de un trastorno o enfermedad.

Contig

Segmento de ADN superpuesto, que representa una región consenso de ADN ensamblado.

Elementos CRISPR

CRISPR es la sigla inglesa para “Clustered Regularly Interspaced Short Palindromic Repeats” (repeticiones palindrómicas cortas, agrupadas y regularmente interespaciadas). Son fragmentos de ADN repetitivos que las bacterias utilizan como sistema de defensa de los virus que las invaden. Por tanto, puede considerarse el sistema inmunitario de la bacteria.

Exoma

Fracción del ADN del genoma que codifica la producción de proteínas. El exoma clínico se refiere al estudio de las variantes genéticas de los genes con implicación clínica, comprendiendo aproximadamente 4500 genes, y obteniendo un número de variantes aproximado entre 3000 a 4000, comparado con el genoma de referencia. Es el método con mayor rendimiento diagnóstico de enfermedades genéticas.

Genoma core

El genoma core o *core genoma* contiene todos los genes comunes a todos los genomas de una especie estudiada.

Metatranscriptoma

Disciplina que estudia la expresión génica de las bacterias en su entorno natural, obteniendo el perfil completo de expresión génica, y por tanto de genes activos, de comunidades microbianas complejas.

Microarray (ADN)

Es un chip que contiene material genético en una superficie sólida de vidrio, plástico o silicona a la que se une una colección de fragmentos de ADN. Se utiliza para estudiar la expresión diferencial de genes o la presencia de los mismos. Su funcionamiento consiste en medir el nivel de hibridación de una sonda específica y la molécula diana, habitualmente mediante fluorescencia, y a través de un análisis de imagen, lo que indica la presencia/ausencia del gen o su expresión.

Pangenoma

Colección de los genes de una especie. Adquiere importancia en un contexto evolutivo, incluyendo los genes localizados en el genoma-núcleo, comunes a todos los genomas estudiados, así como los genes no esenciales de estos genomas, que resultan particulares de cada uno de ellos.

Pili/Pilus

Pilus (singular) o *pili* (plural) son apéndices muy cortos en forma de pelo que se encuentran en la superficie de un gran número de bacterias. Su función es permitir establecer contacto y/o intercambiar material genético con el exterior.

Pirosecuenciación

Tecnología que permite determinar el orden de una secuencia de ADN mediante luminiscencia. Se basa en el principio de “secuenciación por síntesis”, en el que se detecta la incorporación de bases nucleotídicas por acción de una enzima ADN polimerasa. En este método se detecta la liberación de pirofosfato cuando los nucleótidos son incorporados. Este pirofosfato genera luz por medio de diversas reacciones enzimáticas en cadena.

Profago

Genoma de un fago que se ha perpetuado en la bacteria hospedadora al integrarse en su cromosoma.

Scaffold

Unión de segmentos de ADN en forma de *contigs* para construir un esqueleto del genoma (*scaffold*) conforme con un genoma de referencia.

Operón

Unidad genética funcional formada por un grupo complejo de genes capaces de ejercer una regulación de su propia expresión mediante sustratos con los que interactúan las proteínas codificadas por ellos mismos.

Transcriptasa reversa

También conocida como transcriptasa inversa o retrotranscriptasa. Es una enzima de tipo ADN polimerasa cuya función es sintetizar ADN de doble cadena utilizando como molde ARN monocatenario. Cataliza la reacción de retrotranscripción o transcripción inversa. Se encuentra presente de manera natural en los retrovirus. En biología molecular es utilizada para la conversión *in vitro* de ARN hasta ADN complementario.



Enlaces de interés

MITOMAP: A human mitochondrial genome database

Base de datos que agrupa el compendio de polimorfismos y mutaciones presentes en el ADN mitocondrial.

<https://mitomap.org/foswiki/bin/view/MITOMAP/WebHome>

Ensembl

Ensembl es un buscador de genomas para genomas de vertebrados que apoya la investigación en genómica comparada, evolución, variación de secuencia y regulación transcripcional. Ensembl anota genes, calcula múltiples alineaciones, predice la función reguladora y recopila datos de enfermedades. Las herramientas de Ensembl incluyen BLAST, BLAT, BioMart y Variant Effect Predictor (VEP) para todas las especies admitidas.

<http://www.ensembl.org/index.html>

Bibliografía



Álvarez-Molina, A., de Toro, M., Alexa, E. A., & Álvarez-Ordóñez, A. (2021). Applying Genomics to Track Antimicrobial Resistance in the Food Chain. *Comprehensive Foodomics*, 188-211.

<https://doi.org/10.1016/b978-0-08-100596-5.22700-5>

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R., & Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806), 457-465.

<https://doi.org/10.1038/290457a0>

Antipov, D., Hartwick, N., Shen, M., Raiko, M., & Pevzner, P. A. (2016). plasmidSPAdes: Assembling Plasmids from Whole genome sequencing data. *Bioinformatics*, 32(22).

<https://doi.org/10.1093/bioinformatics/btw493>

Arredondo-Alonso, S., Rogers, M. R. C., Braat, J. C., Verschuren, T. D., Top, J., Corander, J., Willems, R. J. L., & Schürch, A. C. (2018). mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microbial Genomics*, 4(11). <https://doi.org/10.1099/mgen.0.000224>

Arredondo-Alonso, S., Willems, R. J., van Schaik, W., & Schürch, A. C. (2017). On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microbial Genomics*, 3(10). <https://doi.org/10.1099/mgen.0.000128>

Aun, E., Brauer, A., Kisand, V., Tenson, T., & Remm, M. (2018). A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLOS Computational Biology*, 14(10), e1006434. <https://doi.org/10.1371/journal.pcbi.1006434>

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., ... Zagnitko, O. (2008). The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics*, 9(1), 75. <https://doi.org/10.1186/1471-2164-9-75>

Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. (s. f.).
<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Babraham Bioinformatics - Trim Galore! (s. f.).
http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Baker, M. (2012). De novo genome assembly: What every biologist should know. *Nature Methods*, 9(4), 333-337. <https://doi.org/10.1038/nmeth.1935>

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5), 455-477. <https://doi.org/10.1089/cmb.2012.0021>

Benjdir, M., Audureau, E., Beresniak, A., Coll, P., Epaud, R., Fiedler, K., Jacquemin, B., Niddam, L., Pandis, S. N., Pohlmann, G., Sandanger, T. M., Simmons, K., Sorensen, M., Wagner, P., & Lanone, S. (2021). Assessing the impact of exposome on the course of chronic obstructive pulmonary disease and cystic fibrosis: The REMEDIA European Project Approach. *Environmental Epidemiology*, 5(4), pe165.
<https://doi.org/10.1097/EE9.0000000000000165>

- Berriman, M., & Rutherford, K. (2003). Viewing and annotating sequence data with Artemis. *Briefings in Bioinformatics*, 4(2), 124-132. <https://doi.org/10.1093/BIB/4.2.124>
- Bessonneau, V., & Rudel, R. A. (2020). Mapping the human exposome to uncover the causes of breast cancer. *International Journal of Environmental Research and Public Health*, 17(1), 1-7. <https://doi.org/10.3390/ijerph17010189>
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15), 2114-2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Bragg, R., van der Westhuizen, W., Lee, J. Y., Coetsee, E., & Boucher, C. (2014). Bacteriophages as potential treatment option for antibiotic resistant bacteria. *Advances in Experimental Medicine and Biology*, 807, 97-110. https://doi.org/10.1007/978-81-322-1777-0_7
- Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., Shukla, M., Thomason, J. A., Stevens, R., Vonstein, V., Wattam, A. R., & Xia, F. (2015). RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports*, 5(1), 8365. <https://doi.org/10.1038/srep08365>
- Brives, C., & Pourraz, J. (2020). Phage therapy as a potential solution in the fight against AMR: obstacles and possible futures. *Palgrave Communications*, 6(1), 1-11. <https://doi.org/10.1057/s41599-020-0478-4>
- Brown, T. A. (2017a). Genomes 3. In The Yale Journal of Biology and Medicine (Vol. 90, Issue 4). Yale Journal of Biology and Medicine.
- Brown, T. A. (2017b). Genomes 4. (4th Ed.). Taylor & Francis.
- Buck Louis, G. M., Yeung, E., Kannan, K., Maisog, J., Zhang, C., Grantz, K. L., & Sundaram, R. (2019). Patterns and Variability of Endocrine-disrupting Chemicals During Pregnancy: Implications for Understanding the Exposome of Normal Pregnancy. *Epidemiology (Cambridge, Mass.)*, 30(Suppl 2), S65-S75. <https://doi.org/10.1097/EDE.0000000000001082>
- Carattoli, A., Zankari, E., García-Fernández, A., Voldby Larsen, M., Lund, O., Villa, L., Møller Aarestrup, F., & Hasman, H. (2014). In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy*, 58(7), 3895-3903. <https://doi.org/10.1128/AAC.02412-14>
- Carver, T., Harris, S. R., Berriman, M., Parkhill, J., & McQuillan, J. A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics (Oxford, England)*, 28(4), 464-469. <https://doi.org/10.1093/bioinformatics/btr703>
- Chen, C., Khaleel, S. S., Huang, H., & Wu, C. H. (2014). Software for pre-processing Illumina next-generation sequencing short read sequences. *Source Code for Biology and Medicine*, 9, 8. <https://doi.org/10.1186/1751-0473-9-8>
- Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), i884-i890. <https://doi.org/10.1093/BIOINFORMATICS/BTY560>

Clausen, P. T. L. C., Zankari, E., Aarestrup, F. M., & Lund, O. (2016). Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. *Journal of Antimicrobial Chemotherapy*, 71(9), 2484-2488. <https://doi.org/10.1093/jac/dkw184>

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6), 1767-1771. <https://doi.org/10.1093/NAR/GKP1137>

Colavecchio, A., Cadieux, B., Lo, A., & Goodridge, L. D. (2017). Bacteriophages contribute to the spread of antibiotic resistance genes among foodborne pathogens of the Enterobacteriaceae family - A review. *Frontiers in Microbiology*, 8(JUN), 1108. <https://doi.org/10.3389/fmicb.2017.01108>

Comas, I. (2017). Genomic epidemiology of tuberculosis. *Advances in Experimental Medicine and Biology*, 1019, 79-93. https://doi.org/10.1007/978-3-319-64371-7_4

Cordero, P., & Ashley, E. (2012). Whole-genome sequencing in personalized therapeutics. *Clinical Pharmacology and Therapeutics*, 91(6), 1001-1009. <https://doi.org/10.1038/CLPT.2012.51>

De Toro, M., Garcillán-Barcia, M. P. P., De La Cruz, F., M, de T., Garcilláon-Barcia, M. P., F, D. L. C., De Toro, M., Garcillán-Barcia, M. P. P., & De La Cruz, F. (2014). Plasmid Diversity and Adaptation Analyzed by Massive Sequencing of Escherichia coli Plasmids. *Microbiology Spectrum*, 2(6), 219-235. <https://doi.org/10.1128/microbiolspec.PLAS-0031-2014>

Deng, X., Den Bakker, H. C., & Hendriksen, R. S. (2016). Genomic Epidemiology: Whole-Genome-Sequencing-Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. *Annual Review of Food Science and Technology*, 7(January), 353-374. <https://doi.org/10.1146/annurev-food-041715-033259>

Ewels, P., Magnusson, M. M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England)*, 32(19), 3047-3048. <https://doi.org/10.1093/bioinformatics/btw354>

FASTX-Toolkit. (s. f.). Consultado el 28 de junio de 2019. http://hannonlab.cshl.edu/fastx_toolkit/

Fernandez-Lopez, R., Redondo, S., Garcillan-Barcia, M. P., & de la Cruz, F. (2017). Towards a taxonomy of conjugative plasmids. *Current Opinion in Microbiology*, 38, 106-113. <https://doi.org/10.1016/j.mib.2017.05.005>

Figueras, M. J., Beaz-Hidalgo, R., Hossain, M. J., & Liles, M. R. (2014). Taxonomic affiliation of new genomes should be verified using average nucleotide identity and multilocus phylogenetic analysis. *Genome Announcements*, 2(6), 6-7. <https://doi.org/10.1128/genomeA.00927-14>

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., ... Venter, J. C. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, 269(5223), 496-512. <https://doi.org/10.1126/SCIENCE.7542800>

Francés-Cuesta, C., Sánchez-Hellín, V., Gomila, B., & González-Candelas, F. (2021). Is there a widespread clone of *Serratia marcescens* producing outbreaks worldwide? *The Journal of Hospital Infection*, 108, 7-14.
<https://doi.org/10.1016/J.JHIN.2020.10.029>

Frost, L. S., Leplae, R., Summers, A. O., & Toussaint, A. (2005). Mobile genetic elements: The agents of open source evolution. *Nature Reviews Microbiology*, 3(9), 722-732. <https://doi.org/10.1038/nrmicro1235>

Gakidou, E., Afshin, A., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abdulle, A. M., Abera, S. F., Aboyans, V., Abu-Raddad, L. J., Abu-Rmeileh, N. M. E., Abyu, G. Y., Adedeji, I. A., Adetokunboh, O., Afarideh, M., Agrawal, A., Agrawal, S., Ahmad Kiadaliri, A., ... Murray, C. J. L. (2017). Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100), 1345-1422.
[https://doi.org/10.1016/S0140-6736\(17\)32366-8](https://doi.org/10.1016/S0140-6736(17)32366-8)

Garcillán-Barcia, M. P., Redondo-Salvo, S., Vielva, L., & de la Cruz, F. (2020). MOBscan: Automated Annotation of MOB Relaxases. *Methods in Molecular Biology* (Clifton, N.J.), 2075, 295-308.
https://doi.org/10.1007/978-1-4939-9877-7_21

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing.
<http://arxiv.org/abs/1207.3907>

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., & Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research*, 17(6), 669-681. <https://doi.org/10.1101/gr.6339607>

GitHub - ChaissonLab/LRA: Long read aligner. (s. f.). Consultado el 25 de octubre de 2021.
<https://github.com/ChaissonLab/LRA>

GitHub - cov-lineages/pangolin: Software package for assigning SARS-CoV-2 genome sequences to global lineages. (s. f.). Consultado el 13 de diciembre de 2021. <https://github.com/cov-lineages/pangolin>

GitHub - kassambara/fastqcr: fastqcr: Quality Control of Sequencing Data. (s. f.). Consultado el 26 de octubre de 2021. <https://github.com/kassambara/fastqcr>

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, 57(1), 81-91.
<https://doi.org/10.1099/ijst.0.64483-0>

Gracia-Cazaña, T., González, S., Parrado, C., Juarranz, Á., & Gilaberte, Y. (2020). Influence of the Exposome on Skin Cancer. *Actas Dermo-Sifiliográficas (English Edition)*, 111(6), 460-470.
<https://doi.org/10.1016/j.adengl.2020.04.011>

Gupta, S. K., Padmanabhan, B. R., Diene, S. M., Lopez-Rojas, R., Kempf, M., Landraud, L., & Rolain, J.-M. (2014). ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrobial Agents and Chemotherapy*, 58(1), 212-220. <https://doi.org/10.1128/AAC.01310-13>

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), 1072-1075. <https://doi.org/10.1093/bioinformatics/btt086>

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaoui, B., Fatima, R., Gall, A., ... Flicek, P. (2021). GRCh38.p13 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.28. Ensembl 2021. *Nucleic acids research*, 49(1), 884–891.
<https://doi.org/10.1093/nar/gkaa942>

Hunt, M., Silva, N. De, Otto, T. D., Parkhill, J., Keane, J. A., & Harris, S. R. (2015). Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biology*, 16(1), 294.
<https://doi.org/10.1186/s13059-015-0849-0>

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1), 119.
<https://doi.org/10.1186/1471-2105-11-119>

Inouye, M., Dashnow, H., Raven, L.-A., Schultz, M. B., Pope, B. J., Tomita, T., Zobel, J., & Holt, K. E. (2014). SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Medicine*, 6(11), 90.
<https://doi.org/10.1186/S13073-014-0090-6>

iVar: Manual. (s. f.). Consultado el 13 de diciembre de 2021.

<https://andersen-lab.github.io/ivar/html/manualpage.html>

Jagadeesan, B., Gerner-Smidt, P., Allard, M. W., Leuillet, S., Winkler, A., Xiao, Y., Chaffron, S., Van Der Vossen, J., Tang, S., Katase, M., McClure, P., Kimura, B., Ching Chai, L., Chapman, J., & Grant, K. (2019). The use of next generation sequencing for improving food safety: Translation into practice. *Food Microbiology*, 79, 96–115. <https://doi.org/10.1016/J.FM.2018.11.005>

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9(1), 1–8.
<https://doi.org/10.1038/s41467-018-07641-9>

Jesus, T. F., Ribeiro-Gonçalves, B., Silva, D. N., Bortolaia, V., Ramirez, M., & Carriço, J. A. (2019). Plasmid ATLAS: plasmid visual analytics and identification in high-throughput sequencing data. *Nucleic Acids Research*, 47(D1), D188–D194. <https://doi.org/10.1093/nar/gky1073>

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., Lago, B. A., Dave, B. M., Pereira, S., Sharma, A. N., Doshi, S., Courtot, M., Lo, R., Williams, L. E., Frye, J. G., Elsayegh, T., Sardar, D., Westman, E. L., Pawlowski, A. C., ... McArthur, A. G. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 45(D1), D566–D573.
<https://doi.org/10.1093/nar/gkw1004>

Joensen, K. G., Scheutz, F., Lund, O., Hasman, H., Kaas, R. S., Nielsen, E. M., & Aarestrup, F. M. (2014). Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *Journal of Clinical Microbiology*, 52(5), 1501–1510.
<https://doi.org/10.1128/JCM.03617-13>

Jones, D. P., & Cohn, B. A. (2020). A vision for exposome epidemiology: The pregnancy exposome in relation to breast cancer in the Child Health and Development Studies. *Reproductive Toxicology*, 92(March), 4–10.
<https://doi.org/10.1016/j.reprotox.2020.03.006>

Juarez, P. D., & Matthews-Juarez, P. (2018). Applying an exposome-wide (ExWAS) approach to cancer research. *Frontiers in Oncology*, 8(AUG), 8-11. <https://doi.org/10.3389/fonc.2018.00313>

Kim, M., Oh, H. S., Park, S. C., & Chun, J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, 64(PART 2), 346-351.
<https://doi.org/10.1099/ij.s.0.059774-0>

Kraken2. (s. f.). Consultado el 13 de diciembre de 2021. <https://ccb.jhu.edu/software/kraken2/index.shtml>

Krawczyk, P. S., Lipinski, L., & Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Research*, 46(6), e35.
<https://doi.org/10.1093/nar/gkx1321>

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35(9), 3100.
<https://doi.org/10.1093/NAR/GKM160>

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357-359.
<https://doi.org/10.1038/nmeth.1923>

Lanza, V. F., de Toro, M., Garcillán-Barcia, M. P., Mora, A., Blanco, J., Coque, T. M., & de la Cruz, F. (2014). Plasmid flux in Escherichia coli ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. *PLoS genetics*, 10(12), e1004766. <https://doi.org/10.1371/journal.pgen.1004766>

Laslett, D., & Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*, 32(1), 11. <https://doi.org/10.1093/NAR/GKH152>

Latest improvements for CLC Sequence Viewer - Current line - Qiagen Bioinformatics. (s. f.). Consultado el 28 de junio de 2019.
<https://www.qiagenbioinformatics.com/products/clc-sequence-viewer/latest-improvements/current-line/>

Lee, I., Kim, Y. O., Park, S. C., & Chun, J. (2016). OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *International Journal of Systematic and Evolutionary Microbiology*, 66(2), 1100-1103. <https://doi.org/10.1099/ijsem.0.000760>

Leplae, R., Hebrant, A., Wodak, S. J., & Toussaint, A. (2004). ACLAME: A CLAssification of Mobile genetic Elements. *Nucleic Acids Research*, 32(90001), 45D-49. <https://doi.org/10.1093/nar/gkh084>

Leplae, Raphaël, Lima-Mendez, G., & Toussaint, A. (2010). ACLAME: A CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Research*, 38(Database issue), D57-61.
<https://doi.org/10.1093/nar/gkp938>

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094-3100.
<https://doi.org/10.1093/BIOINFORMATICS/BTY191>

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5), 589-595. <https://doi.org/10.1093/bioinformatics/btp698>

- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup, 1000 Genome Project Data Processing. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Li, M. M., Datto, M., Duncavage, E. J., Kulkarni, S., Lindeman, N. I., Roy, S., Tsimberidou, A. M., Vnencak-Jones, C. L., Wolff, D. J., Younes, A., & Nikiforova, M. N. (2017). Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *The Journal of molecular diagnostics*, 19(1), 4–23. <https://doi.org/10.1016/j.jmoldx.2016.10.002>
- Liu, B., Zheng, D., Jin, Q., Chen, L., & Yang, J. (2019). VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Research*, 47(D1), D687–D692. <https://doi.org/10.1093/nar/gky1080>
- Manchanda, N., Portwood, J. L., Woodhouse, M. R., Seetharam, A. S., Lawrence-Dill, C. J., Andorf, C. M., & Hufford, M. B. (2020). GenomeQC: A quality assessment tool for genome assemblies and gene structure annotations. *BMC Genomics*, 21(1), 1–9. <https://doi.org/10.1186/s12864-020-6568-2>
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, 17(1), 10. <https://doi.org/10.14806/EJ.17.1.200>
- McKeon, T. P., Hwang, W. T., Ding, Z., Tam, V., Wileyto, P., Glanz, K., & Penning, T. M. (2021). Environmental exposomics and lung cancer risk assessment in the Philadelphia metropolitan area using ZIP code-level hazard indices. *Environmental Science and Pollution Research*, 28(24), 31758–31769. <https://doi.org/10.1007/s11356-021-12884-z>
- Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., & Edwards, R. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9(1), 386. <https://doi.org/10.1186/1471-2105-9-386>
- Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly Algorithms for Next-Generation Sequencing Data. *Genomics*, 95(6), 315. <https://doi.org/10.1016/J.YGENO.2010.03.001>
- Moghadam, M. T., Amirmozafari, N., Shariati, A., Hallajzadeh, M., Mirkalantari, S., Khoshbayan, A., & Jazi, F. M. (2020). How Phages Overcome the Challenges of Drug Resistant Bacteria in Clinical Infections. *Infection and Drug Resistance*, 13, 45. <https://doi.org/10.2147/IDR.S234353>
- Naqvi, A. A. T., Fatima, K., Mohammad, T., Fatima, U., Singh, I. K., Singh, A., Atif, S. M., Hariprasad, G., Hasan, G. M., & Hassan, M. I. (2020). Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochimica et Biophysica Acta. Molecular Basis of Disease*, 1866(10), 165878. <https://doi.org/10.1016/J.BBADIS.2020.165878>
- Navarro, F., & Muniesa, M. (2017). Phages in the Human Body. *Frontiers in Microbiology*, 8, 566. <https://doi.org/10.3389/FMICB.2017.00566>
- Nawrocki, E. P., Kolbe, D. L., & Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics (Oxford, England)*, 25(10), 1335–1337. <https://doi.org/10.1093/BIOINFORMATICS/BTP157>

- Nevado, J., Arribas, J. y Pérez, L. A. (2018). *Informes Anticipando. Medicina Preventiva Personalizada.* Fundación Instituto Roche.
https://www.institutoroche.es/recursos/publicaciones/185/informes_anticipando_medicina_preventiva_personalizada
- Nunnari, J., & Suomalainen, A. (2012). Mitochondria: In Sickness and in Health. *Cell*, 148(6), 1145-1159.
<https://doi.org/10.1016/J.CELL.2012.02.035>
- O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J. T., Colquhoun, R., Ruis, C., Abu-Dahab, K., Taylor, B., Yeats, C., du Plessis, L., Maloney, D., Medd, N., Attwood, S. W., Aanensen, D. M., Holmes, E. C., Pybus, O. G., & Rambaut, A. (2021). Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolution*, 7(2). <https://doi.org/10.1093/VE/VEAB064>
- Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2015). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, 28(2), btv566.
<https://doi.org/10.1093/bioinformatics/btv566>
- Olea, N., Casas, M., Castaño, A., Mendiola, J., Vrijheid, M., Arenas, J., Carracedo, Á., Lapunzina, P. y Martín-Sánchez, F. (2020). *Informes Anticipando. Exposoma.* Fundación Instituto Roche.
<https://www.institutoroche.es/observatorio/exposoma>
- Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., & Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, 42(Database issue), D206-14. <https://doi.org/10.1093/nar/gkt1226>
- Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., & Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, 15(2), 256. <https://doi.org/10.1093/BIB/BBS086>
- Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics (Oxford, England)*, 31(22), 3691-3693. <https://doi.org/10.1093/bioinformatics/btv421>
- Page, A. J., Wailan, A., Shao, Y., Judge, K., Dougan, G., Klemm, E. J., Thomson, N. R., & Keane, J. A. (2018). PlasmidTron: assembling the cause of phenotypes and genotypes from NGS data. *Microbial Genomics*, 4(3).
<https://doi.org/10.1099/mgen.0.000164>
- Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7), 1043-1055. <https://doi.org/10.1101/GR.186072.114>
- Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLOS ONE*, 7(2), e30619. <https://doi.org/10.1371/JOURNAL.PONE.0030619>
- Pérez, M. y Tolosa, A. (Eds.). (2017). Genómica. En: *Medicina. Una guía práctica.* Medigene Press.
- Petersen, B. S., Fredrich, B., Hoeppner, M. P., Ellinghaus, D., & Franke, A. (2017). Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genetics*, 18(1), 1-13.
<https://doi.org/10.1186/s12863-017-0479-5>

- Petersen, T. N., Brunak, S., Von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, 8(10), 785-786. <https://doi.org/10.1038/NMETH.1701>
- Pightling, A. W., Pettengill, J. B., Luo, Y., Baugher, J. D., Rand, H., & Strain, E. (2018). Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations. *Frontiers in Microbiology*, 9, 1482. <https://doi.org/10.3389/fmicb.2018.01482>
- Prieto, A., Urcola, I., Blanco, J., Dahbi, G., Muniesa, M., Quirós, P., Falgenhauer, L., Chakraborty, T., Hüttner, M., & Juárez, A. (2016). Tracking bacterial virulence: global modulators as indicators. *Scientific Reports*, 6(1), 25973. <https://doi.org/10.1038/srep25973>
- Principi, N., Silvestri, E., & Esposito, S. (2019). Advantages and limitations of bacteriophages for the treatment of bacterial infections. *Frontiers in Pharmacology*, 10(MAY), 513. <https://doi.org/10.3389/fphar.2019.00513>
- Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics*, 70(1), e102. <https://doi.org/10.1002/CPBI.102>
- Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., & Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, 5(11), 1403-1407. <https://doi.org/10.1038/s41564-020-0770-5>
- Redondo-Salvo, S., Bartomeus-Peñalver, R., Vielva, L., Tagg, K. A., Webb, H. E., Fernández-López, R., & de la Cruz, F. (2021). COPLA, a taxonomic classifier of plasmids. *BMC Bioinformatics*, 22(1), 1-9. <https://doi.org/10.1186/S12859-021-04299-X>
- Redondo-Salvo, S., Fernández-López, R., Ruiz, R., Vielva, L., de Toro, M., Rocha, E. P. C., Garcillán-Barcia, M. P., & de la Cruz, F. (2020). Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nature Communications*, 11(1), 1-13. <https://doi.org/10.1038/s41467-020-17278-2>
- Ren, J., & Chaisson, M. J. P. (2021). Ira: A long read aligner for sequences and contigs. *PLOS Computational Biology*, 17(6), e1009078. <https://doi.org/10.1371/JOURNAL.PCBI.1009078>
- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W., Hedge, M., Lyon, E., Spector, E., Voelkerding, K., & HL, R. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 17(5), 405-424. <https://doi.org/10.1038/GIM.2015.30>
- Richter, M., & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences of the United States of America*, 106(45), 19126-19131. <https://doi.org/10.1073/pnas.0906412106>
- Robertson, J., & Nash, J. H. E. (2018). MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial Genomics*, 4(8). <https://doi.org/10.1099/mgen.0.000206>
- Rodríguez-Rubio, L., Serna, C., Ares-Arroyo, M., Matamoros, B. R., Delgado-Blas, J. F., Montero, N., Bernabé-Balas, C., Wedel, E. F., Mendez, I. S., Muniesa, M., & Gonzalez-Zorn, B. (2020). Extensive antimicrobial resistance mobilization via multicopy plasmid encapsidation mediated by temperate phages. *The Journal of Antimicrobial Chemotherapy*, 75(11), 3173-3180. <https://doi.org/10.1093/JAC/DKAA311>

Roosaare, M., Puustusmaa, M., Möls, M., Vaher, M., & Remm, M. (2018). PlasmidSeeker: identification of known plasmids from bacterial whole genome sequencing reads. *PeerJ*, 6, e4588.
<https://doi.org/10.7717/peerj.4588>

Roser, L. G., Agüero, F., & Sánchez, D. O. (2019). FastqCleaner: an interactive Bioconductor application for quality-control, filtering and trimming of FASTQ files. *BMC Bioinformatics*, 20(1), 1-7.
<https://doi.org/10.1186/S12859-019-2961-8>

Rowe, W. P. M., & Winn, M. D. (2018). Indexed variation graphs for efficient and accurate resistome profiling. *Bioinformatics*, 34(21), 3601-3608. <https://doi.org/10.1093/bioinformatics/bty387>

Royer, G., Decousser, J. W., Branger, C., Dubois, M., Médigue, C., Denamur, E., & Vallenet, D. (2018). PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microbial Genomics*, 4(9). <https://doi.org/10.1099/mgen.0.000211>

Rozov, R., Kav, A. B., Bogumil, D., Shterzer, N., Halperin, E., Mizrahi, I., & Shamir, R. (2017). Recycler: An algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics*, 33(4), 475-482.
<https://doi.org/10.1093/bioinformatics/btw651>

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics (Oxford, England)*, 16(10), 944-945.
<https://doi.org/10.1093/BIOINFORMATICS/16.10.944>

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T. J., Schatz, M. C., Delcher, A. L., Roberts, M., Marcxais, G., Pop, M., & Yorke, J. A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, 22(3), 557-567.
<https://doi.org/10.1101/GR.131383.111>

Seaby, E. G., Pengelly, R. J., & Ennis, S. (2016). Exome sequencing explained: A practical guide to its clinical application. *Briefings in Functional Genomics*, 15(5), 374-384. <https://doi.org/10.1093/bfgp/elv054>

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14), 2068-2069.
<https://doi.org/10.1093/bioinformatics/btu153>

Seemann, T. (s. f.). Snippy: Rapid haploid variant calling and core genome alignment.
<https://github.com/tseemann/snippy>

Shaffer, R. M., Smith, M. N., & Faustman, E. M. (2017). Developing the regulatory utility of the exposome: Mapping exposures for risk assessment through lifestage exposome snapshots (LEnS). *Environmental Health Perspectives*, 125(8), 1-8. <https://doi.org/10.1289/EHP1250>

Sharma, P., & Sampath, H. (2019). Mitochondrial DNA Integrity: Role in Health and Disease. *Cells*, 8(2), 100.
<https://doi.org/10.3390/CELLS8020100>

Sullivan, M. J., Zakour, N. L. Ben, Forde, B. M., Stanton-Cook, M., & Beatson, S. A. (2015). Contiguity: Contig adjacency graph construction and visualisation. *PeerJ PrePrints*, 3, e1037v1.
<https://doi.org/10.7287/peerj.preprints.1037v1>

- Sun, Y., Ruivenkamp, C. AL, Hoffer, M. J., Vrijenhoek, T., Kriek, M., van Asperen, C. J., den Dunnen, J. T., & Santen, G. W. (2015). Next Generation Diagnostics: gene panel, exome or whole genome? *Human Mutation*, 36(6), 648-655. <https://doi.org/10.1002/humu.22783>
- Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., & Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, 44(14), gkw569. <https://doi.org/10.1093/nar/gkw569>
- Terry, M. B., Michels, K. B., Brody, J. G., Byrne, C., Chen, S., Jerry, D. J., Malecki, K. M. C., Martin, M. B., Miller, R. L., Neuhausen, S. L., Silk, K., Trentham-Dietz, A., McDonald, J., Oskar, S., Knight, J., Toro-Campos, R., Cai, X., Rising, C. J., Afanaseva, D., ... Fisher, C. (2019). Environmental exposures during windows of susceptibility for breast cancer: A framework for prevention research. *Breast Cancer Research*, 21(1), 1-16. <https://doi.org/10.1186/s13058-019-1168-2>
- Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews. Genetics*, 13(1), 36-46. <https://doi.org/10.1038/nrg3117>
- Treangen, T. J., Ondov, B. D., Koren, S., & Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, 15(11), 524. <https://doi.org/10.1186/PREACCEPT-2573980311437212>
- Trent, R. J. (2012). Genes to Personalized Medicine. En: R. J. Trent (ed.), *Molecular Medicine* (pp. 1-37). Academic Press. <https://doi.org/10.1016/b978-0-12-381451-7.00001-3>
- Valiente-Mullor, C., Beamud, B., Ansari, I., Frances-Cuesta, C., Garcia-Gonzalez, N., Mejia, L., Ruiz-Hueso, P., & Gonzalez-Candelas, F. (2021). One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLOS Computational Biology*, 17(1), e1008678. <https://doi.org/10.1371/JOURNAL.PCBI.1008678>
- VFDB: Virulence Factors Database. (s. f.). Consultado el 14 de septiembre de 2018. <http://www.mgc.ac.cn/VFs/>
- Vielva, L., Toro, M. de, Lanza, V. F., Cruz, F. de la, Berger, B., de Toro, M., Lanza, V. F., & de la Cruz, F. (2017). PLACNETw: a web-based tool for plasmid reconstruction from bacterial genomes. *Bioinformatics (Oxford, England)*, 33(23), 0-0. <https://doi.org/10.1093/bioinformatics/btx462>
- Wajid, B., & Serpedin, E. (2016). Do it yourself guide to genome assembly. *Briefings in Functional Genomics*, 15(1), 1-9. <https://doi.org/10.1093/BFGP/ELU042>
- Wetterstrand, K. A. (1 de noviembre de 2021). *The Cost of Sequencing a Human Genome*. National Human Genome Research Institute. <https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>
- Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*, 13(6), e1005595. <https://doi.org/10.1371/journal.pcbi.1005595>
- Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics (Oxford, England)*, 31(20), 3350-3352. <https://doi.org/10.1093/bioinformatics/btv383>

Wild, C. P. (2005). Complementing the genome with an “exosome”: The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers and Prevention*, 14(8), 1847-1850. <https://doi.org/10.1158/1055-9965.EPI-05-0456>

Wilkinson, M. D., Dumontier, M., Aalbersberg, IJ. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 1-9.
<https://doi.org/10.1038/sdata.2016.18>

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1). <https://doi.org/10.1186/S13059-019-1891-0>

Wright, R. O. (2017). Environment, susceptibility windows, development, and child health. *Current opinion in pediatrics*, 29(2), 211-217. <https://doi.org/10.1097/MOP.0000000000000465>

Yang, L. A., Chang, Y. J., Chen, S. H., Lin, C. Y., & Ho, J. M. (2019). SQUAT: A Sequencing Quality Assessment Tool for data quality assessments of genome assemblies. *BMC Genomics*, 19(9), 1-12.
<https://doi.org/10.1186/S12864-019-5445-3>

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., & Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, 67(11), 2640-2644. <https://doi.org/10.1093/jac/dks261>

Zankari, Ea, Allesøe, R., Joensen, K. G., Cavaco, L. M., Lund, O., & Aarestrup, F. M. (2017). PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *Journal of Antimicrobial Chemotherapy*, 72(10), 2764-2768. <https://doi.org/10.1093/jac/dkx217>

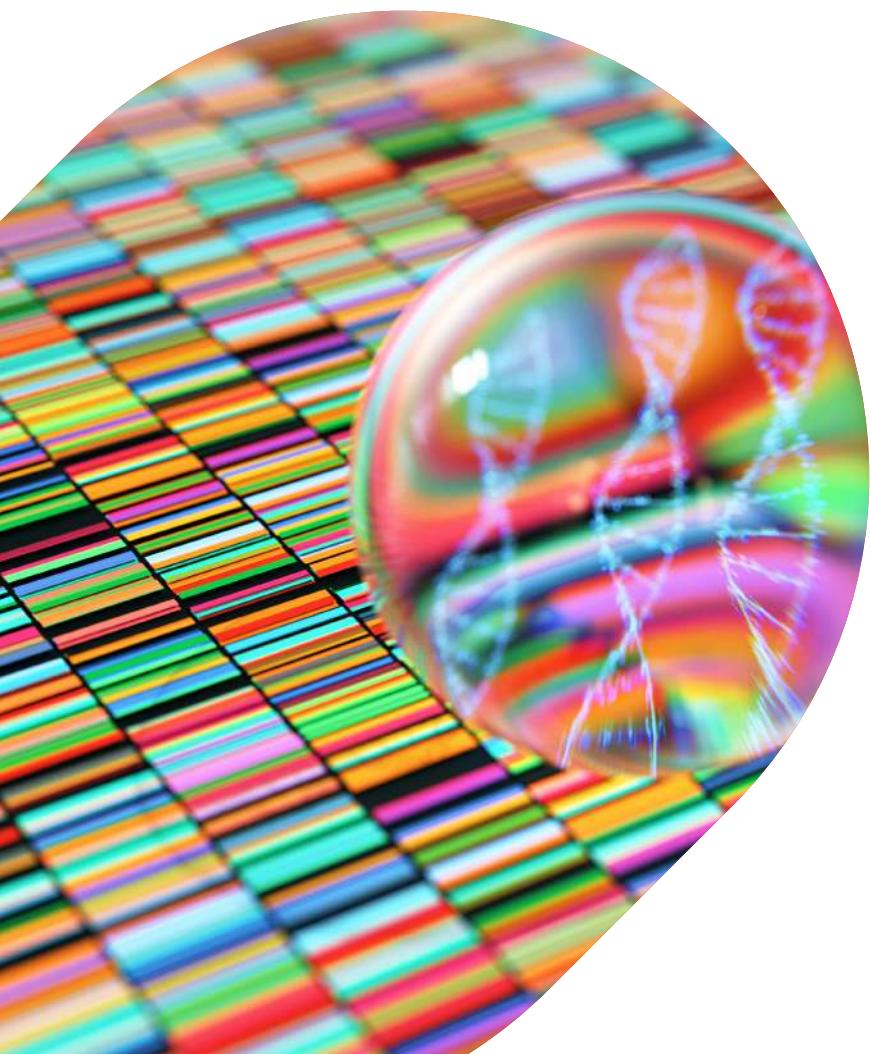
Zerbino, D. R. (2010). Using the Velvet de novo assembler for short-read sequencing technologies. *Current Protocols in Bioinformatics*, Chapter 11, Unit 11.5. <https://doi.org/10.1002/0471250953.bi1105s31>

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), 821-829. <https://doi.org/10.1101/gr.074492.107>

Zetner, A., Cabral, J., Mataseje, L., Knox, N. C., Mabon, P., Mulvey, M., & Domselaar, G. Van. (2017). Plasmid Profiler: Comparative Analysis of Plasmid Content in WGS Data. *BioRxiv*, 121350.
<https://doi.org/10.1101/121350>

Zhao, S., Agafonov, O., Azab, A., Stokowy, T., & Hovig, E. (2020). Accuracy and efficiency of germline variant calling pipelines for human genome data. *Scientific reports*, 10(1), 20222.
<https://doi.org/10.1038/s41598-020-77218-4>

Zhou, F., & Xu, Y. (2010). cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics*, 26(16), 2051-2052.
<https://doi.org/10.1093/bioinformatics/btq299>



Autora
Dra. María de Toro

ISBN: 978-84-19314-00-0
Reservados todos los derechos©
Universidad Internacional de Valencia - 2022