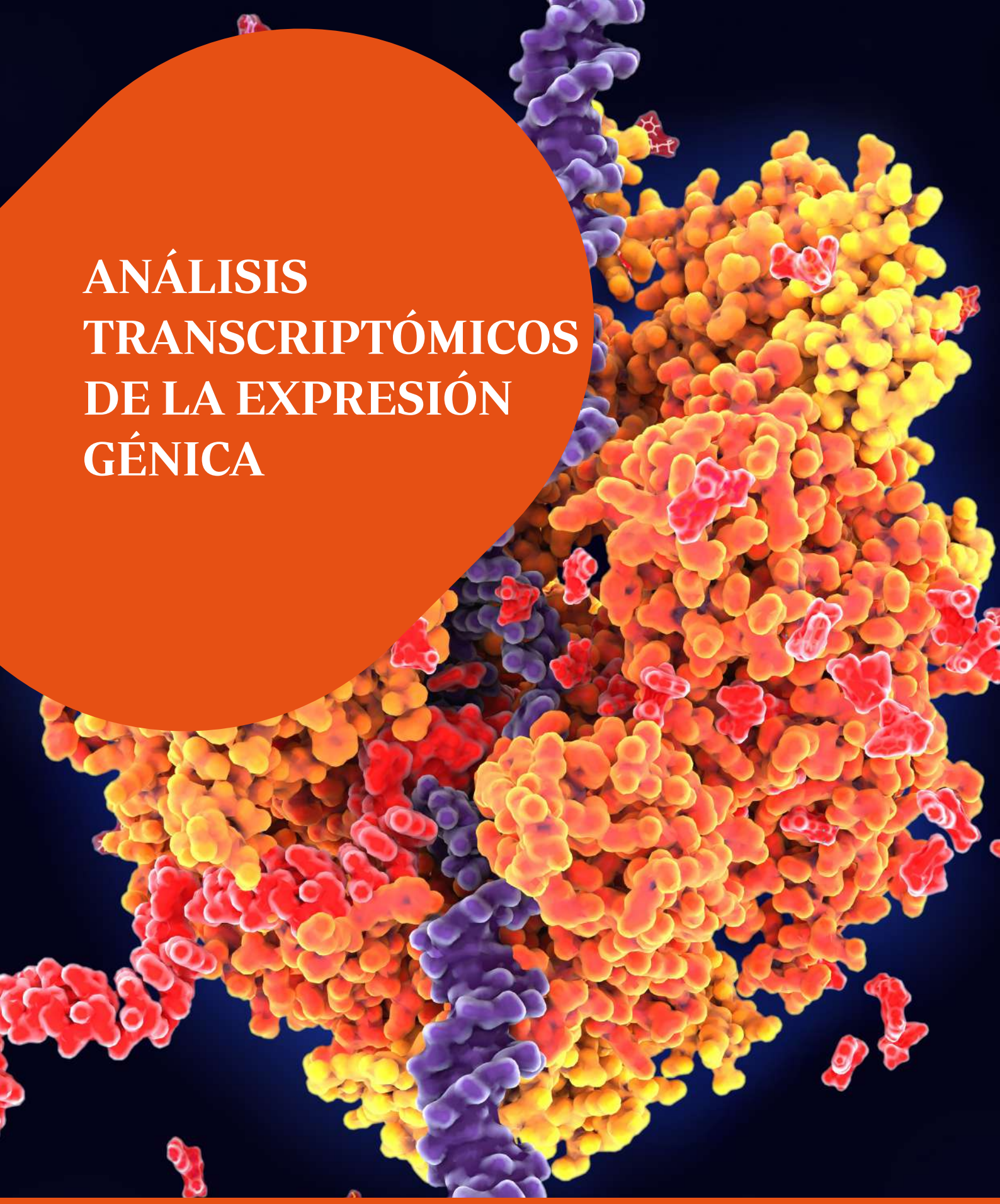
ANÁLISIS TRANSCRIPTÓMICOS DE LA EXPRESIÓN GÉNICA



Dra. Nuria Mauri

viu Universidad Internacional de Valencia

Análisis transcriptómicos de la expresión génica

Dra. Nuria Mauri

Edita
Universidad Internacional de Valencia

Leyendas

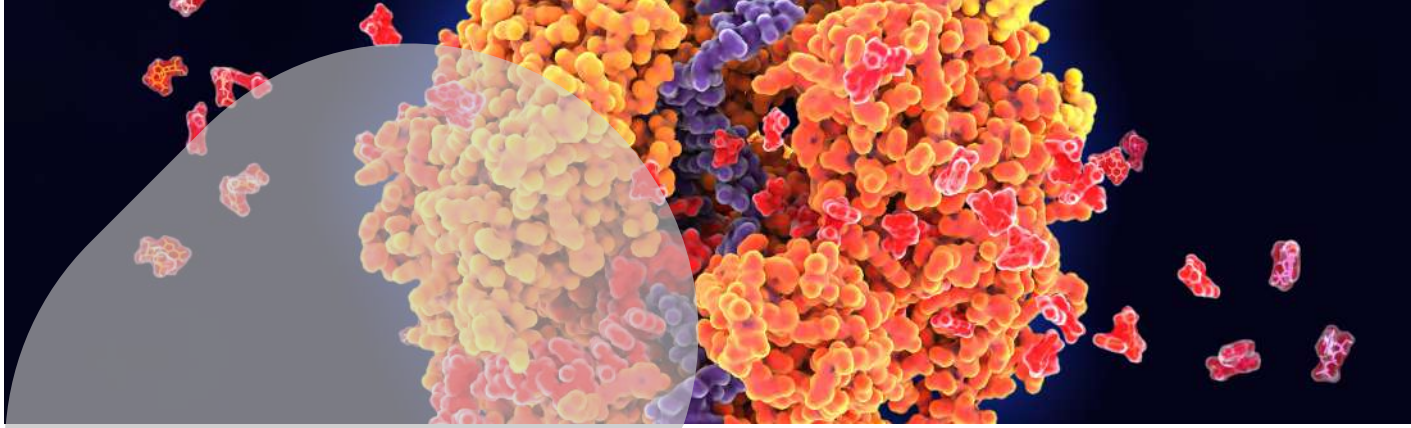
Enlace de interés

Ejemplo

Importante

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

Índice



Capítulo 1

Introducción a las técnicas transcriptómicas actuales y emergentes

Una vez desvelada la estructura del ADN en los experimentos con rayos X de R. Franklin, J. Watson y F. Crick, surgieron una serie de planteamientos acerca de su expresión en formas funcionales.

Ya entonces se indicaba la posible responsabilidad de un "mensajero de ARN" en el traslado de esta información genética a los ribosomas dentro de lo que se propuso como el "dogma central de la biología".

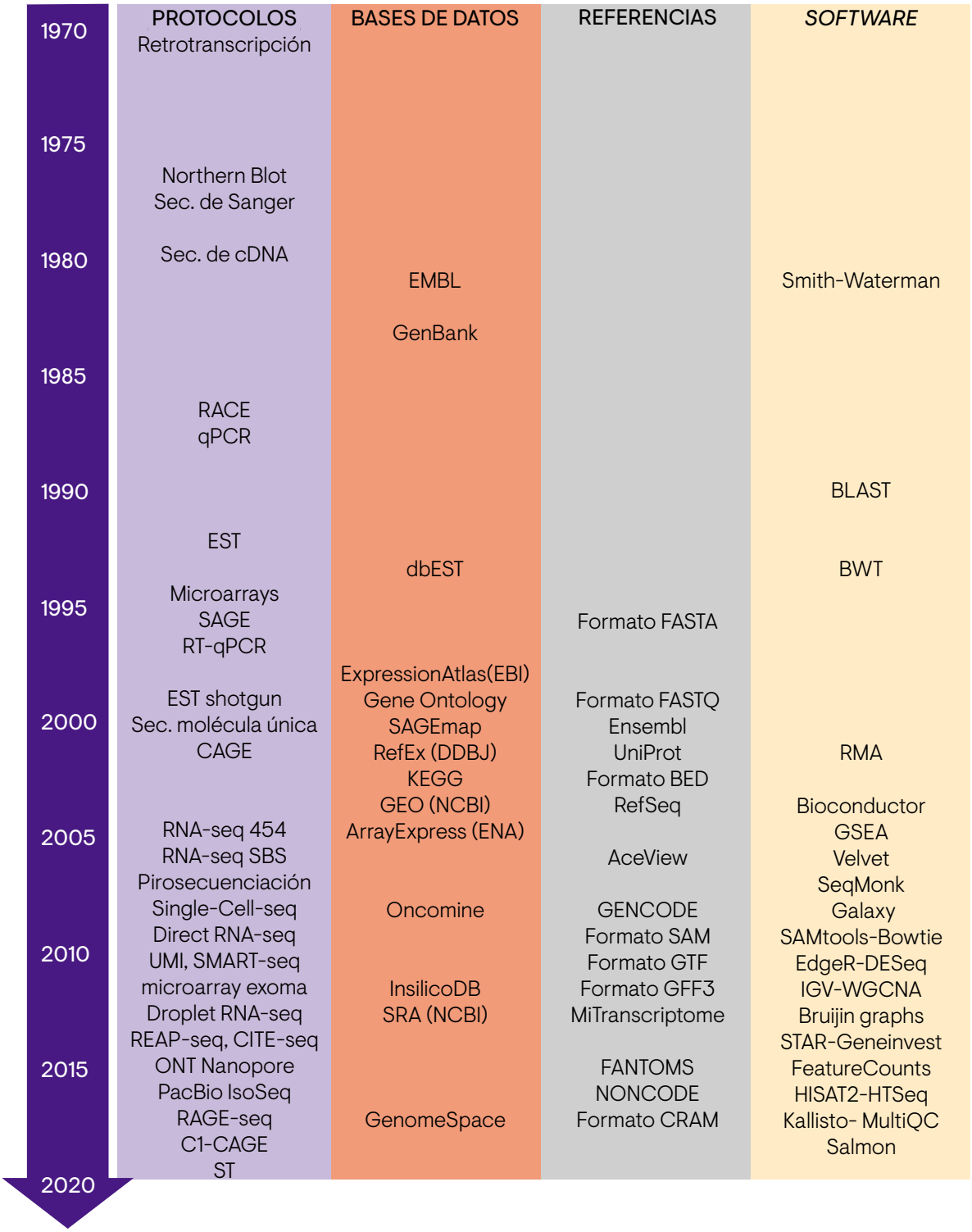
Desde aquel momento, nuestro conocimiento acerca de la expresión génica ha avanzado enormemente. Los mecanismos de la transcripción a ARN y de la traducción a proteína han sido desentramados en organismos tanto procariotas como eucariotas.

El código de los ácidos nucleicos y su correspondencia con los aminoácidos de las proteínas ha sido entendido. Y cada vez sabemos más acerca de las unidades funcionales o dominios de genes, moléculas de ARN y proteínas.

En este primer capítulo haremos un repaso a los diferentes enfoques planteados para el estudio de la expresión génica, identificando algunos de los hitos más importantes en las tecnologías de detección de ARN, la aparición de las primeras bases de datos y los desafíos computacionales de cada momento (Figura 1).

Figura 1

Línea de tiempo mostrando los principales hitos en el desarrollo de la transcriptómica con la aparición de diferentes protocolos de detección de ARN, bases de datos, referencias de secuencia y software



Desde la segunda mitad del siglo xx hasta el momento actual podemos diferenciar dos etapas tecnológicas: la pretranscriptómica y la transcriptómica, separadas por su capacidad de estudiar individualmente o en el conjunto de genes los cambios de expresión.

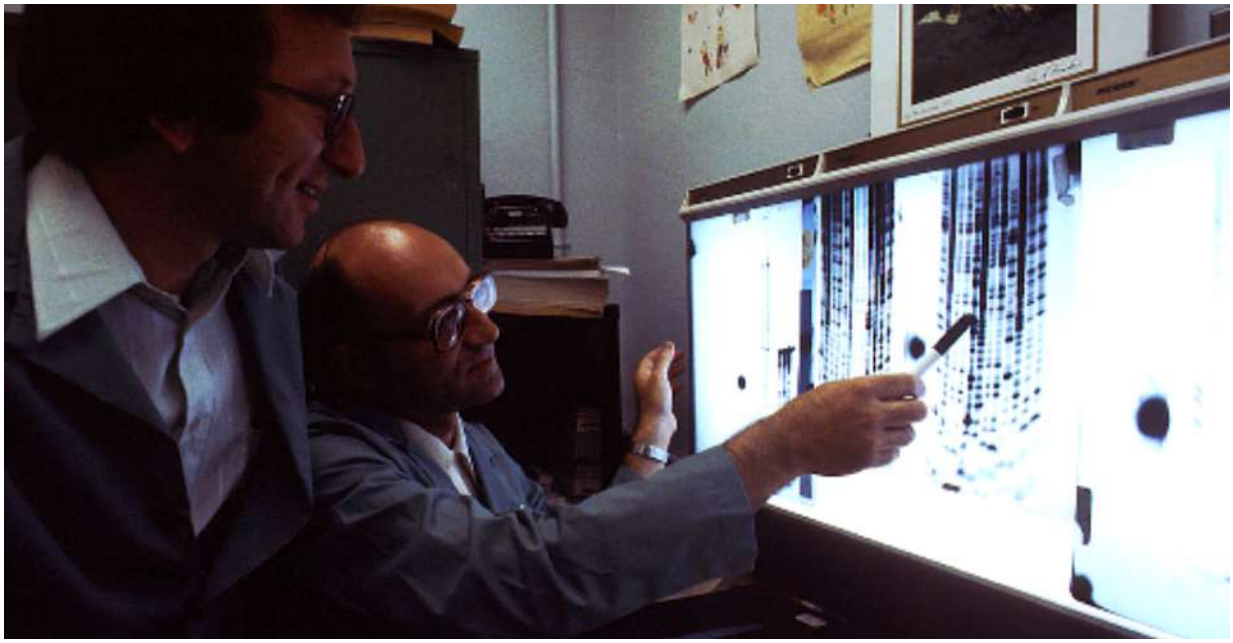
1.1. Pretranscriptómica

Esta primera etapa podría abarcar desde los años 70 hasta los 90, aunque los límites son relativos. El primer hito podría considerarse el descubrimiento de la enzima transcriptasa inversa en los retrovirus (1970). Sin esta reacción de retrotranscripción (RT) no hubiera sido posible convertir la molécula del mensajero de ARN (mRNA) en una molécula más estable: el ADN complementario (cDNA).

Pero, sin duda, la revolución tecnológica más importante para la transcriptómica (aún desde el punto de vista actual) sería la invención de la secuenciación cuyo mecanismo básico fue desarrollado por Frederick Sanger (Sanger *et al*., 1977). Esta primera forma de secuenciación usaba una polimerasa de ADN para sintetizar una segunda cadena de ADN a partir del cDNA cuyo sustrato modificado, el nucleótido de terminación, impedía la continuidad de su reacción generando fragmentos de diferente longitud. El Laboratorio Europeo de Biología Molecular (**EMBL**) fue la primera institución de investigación, en 1980, en empezar a recoger esta información de secuencia (Figura 2).

Figura 2

Dos científicos leyendo la secuencia nucleotídica de un fragmento en la fotografía de un gel en los inicios de la secuenciación de Sanger (años 80)



Nota. Por Linda Bartlett para el National Cancer Institute (National Institutes of Health). Dominio público. Recuperado de https://commons.wikimedia.org/wiki/File:DNA_sequencing.jpg

Esta misma técnica, modernizada mediante el uso de nucleótidos marcados con fluorocromos en lugar de radioactivo, se asentaría unos años como el método mayoritariamente utilizado para la extracción de secuencias en un período llamado de secuenciación de primera generación. La productividad acumulada de los diferentes laboratorios alrededor del mundo aumentó considerablemente el catálogo de secuencias de genes o etiquetas de secuencias expresadas (EST). Estas secuencias eran de un tamaño generalmente menor a 1000 pares de bases (pb) y se producían en grandes lotes, lo que impulsó la apertura de las primeras bases de datos para recopilarlas. Estas primeras librerías de transcritos provenientes de diferentes organismos se organizaron en la base de datos dbEST, una división de GenBank (Boguski *et al*., 1993).

Enlace de interés

El siguiente enlace lleva a la biblioteca de secuencias dbEST usada para la elaboración de modelos génicos.
https://www.ncbi.nlm.nih.gov/genbank/dbest/

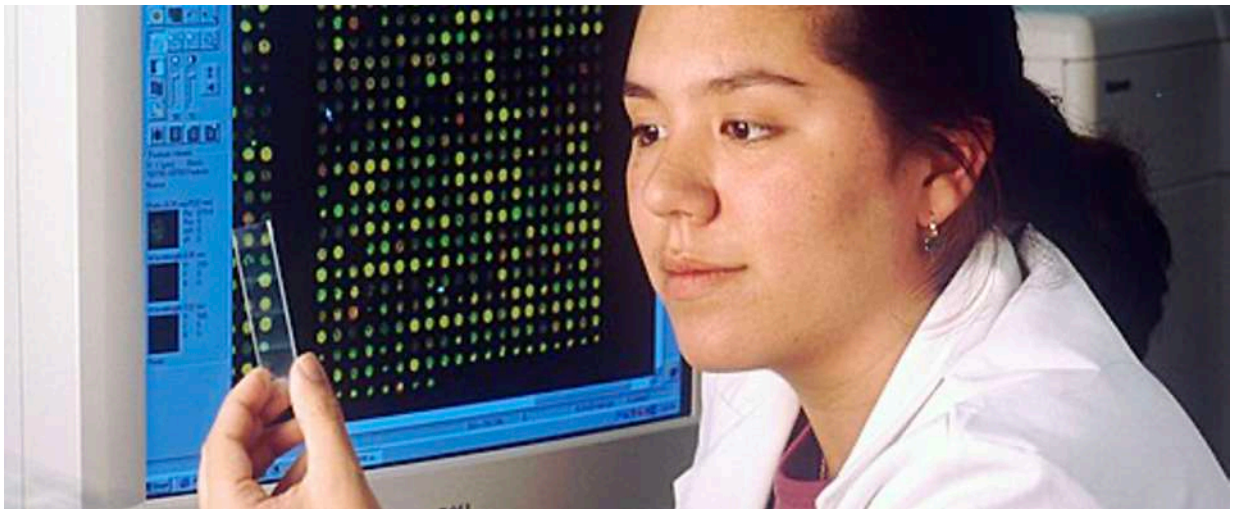
También durante la etapa pretranscriptómica, se afrontaron problemas de carácter matemático como el alineamiento de dos secuencias. Uno de los primeros algoritmos propuestos a principios de los 80 para resolver el alineamiento local fue el algoritmo de Smith-Waterman. En este se basaba la herramienta de computación BLAST (1990) que fue desarrollada por científicos del Centro Nacional de Información Biotecnológica (**NCBI**) para cuantificar la similitud de una secuencia.

1.2. Microarrays

La anotación detallada de los transcritos en los catálogos de EST favoreció la llegada en los 90 de los primeros microarrays. Estas nuevas plataformas se basaban en la hibridación por similitud de dos moléculas de ADN de cadena sencilla. En esta unión por enlaces de hidrógeno, una parte eran las sondas de secuencia conocida previamente fijadas a la superficie de un portaobjetos. En segundo lugar, el cDNA de la muestra era desnaturalizado, marcado con fluorescencia e hibridado con las sondas (Shena *et al*., 1995). Tras varios lavados para retirar los transcritos no hibridados, se escaneaba la imagen que sería analizada más tarde para calcular la tasa de fluorescencia de cada uno de los grupos de sondas o *spots* que representaban diferentes genes (Figura 3).

Figura 3

Una científica portando un microarray en la mano cuya imagen analiza en la computadora (años 90)



Nota. Por Bill Branson para el National Cancer Institute (National Institutes of Health). Dominio público. Recuperado de https://commons.wikimedia.org/wiki/File:Computer_with_microarray.jpg

Por primera vez se monitorizaban centenas de transcritos en un solo experimento, lo que dio lugar a un alcance global de los estudios de expresión génica en lo que llamamos **transcriptómica**. El bajo coste de esta tecnología, como la ofrecida por la empresa Affymetrix, la hizo muy popular y promovió el estudio de manera comparativa de un amplio abanico de condiciones experimentales.

Así, en 2002, el NCBI de los Estados Unidos abrió el primer repositorio destinado a datos de microarrays, el Gene Expression Omnibus (GEO) (Edgar *et al.*, 2002). Esta base de datos recogía los perfiles de expresión de diferentes tejidos, tratamientos o puntos temporales en diferentes organismos, una información muy relevante para la investigación en redes de regulación génica.

Enlace de interés

El siguiente enlace corresponde con la base de datos GEO dentro del portal del NCBI. Este repositorio público mantiene ahora tanto datos basados en array como en secuencia.
https://www.ncbi.nlm.nih.gov/geo/

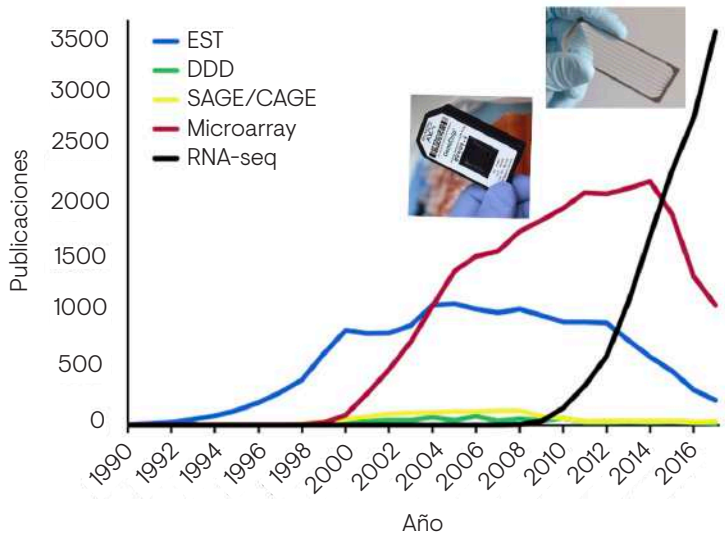
Con el nacimiento de los microarrays surgieron también los primeros análisis de expresión génica diferencial (DGE), ya que permitían el uso de dos colores de fluorocromo. Dos muestras de diferente origen eran comparadas en una misma plataforma, lo que dio pie a la aplicación de principios de probabilidad estadística. La variabilidad técnica que la plataforma sufría debido al ruido de fondo y a la hibridación cruzada obligó además al desarrollo de algoritmos como RMA para el preprocesamiento en los datos.

El uso tan extendido de los microarrays cambió la forma en que los investigadores abordaban sus experimentos y la misma figura del biólogo molecular se tuvo que formar en estadística y computación. En este sentido, un grupo de bioestadísticos tomó la iniciativa creando el primer repositorio de *software* libre destinado a la biología, Bioconductor (Gentleman *et al.*, 2004). Esto dio mayor acceso de otros biólogos a la computación, unificó el uso de métodos de análisis y aumentó la reproducibilidad en las publicaciones.

Con la llegada de las tecnologías de alto rendimiento, el uso de los microarrays dentro del campo de la expresión génica ha disminuido mucho. Nota de ello es el descenso en las publicaciones que usan los microarrays en su metodología, a favor de los RNA-seq (Figura 4).

Figura 4

Artículos publicados que se refieren a RNA-seq (negro), microarrays de RNA (rojo), DDD (verde), EST (azul) y SAGE (amarillo) desde 1990 a 2017



Nota. Recuperado de "Transcriptomics technologies", por R. Lowe, N. Shirley, M. Bleackley, S. Dolan y T. Shafee, 2017, *PloS Computational Biology*, *11*(8), e1004393.

Actualmente sigue usándose en combinación con la secuenciación como el microarray de captura de exoma más enfocado al diagnóstico clínico (Clark, 2013). Esta técnica permite identificar, por ejemplo, variantes funcionales involucradas en enfermedades genéticas.

La transcriptómica, como el estudio de la expresión génica a escala genómica, empezó a desarrollarse a principios de la década de 1990 con la llegada de los datos generados por los microarrays.

1.3. Secuenciación de ARN de alto rendimiento (RNA-seq)

Fue a partir de los 2000 cuando varias empresas retomaron la estrategia de la secuenciación para impulsar la paralelización del proceso en nuevas plataformas, que se denominó secuenciación de alto rendimiento (HTS). La HTS ha producido dos generaciones de secuenciación de ARN (**RNA-seq**) que actualmente tienen diferentes aplicaciones en transcriptómica: la secuenciación de segunda y de tercera generación. Los productos de ambas se diferencian en su rendimiento, el tamaño de la lectura y su tasa de error (Tabla 1).

Tabla 1
Plataformas de secuenciación utilizadas para RNA-seq

HTS	Plataforma	Año	Tipo	Longitud lectura	Rendimiento máximo por carrera	Exactitud por lectura (%)	Depositados en SRA (2016)
2.ª	454 Life Sciences	2005	pirólisis	700 bp	0.7 Gbp	99.9	3548
	Illumina	2006	síntesis	50–300 bp	900 Gbp	99.9	362903
	SOLiD	2008	ligación	50 bp	320 Gbp	99.9	7032
3.ª	Ion Torrent	2010	semiconducción	400 bp	30 Gbp	98	1953
	PacBio	2011	molécula única tiempo-real	10,000 bp	2 Gbp	87	160

Nota. Adaptado de "Transcriptomics technologies", por R. Lowe, N. Shirley, M. Bleackley, S. Dolan y T. Shafee, 2017, *PLoS Computational Biology, 13*(5), e1005457. https://doi.org/10.1371/journal.pcbi.1005457
Legenda: HTS - High throughput sequencing; NCBI SRA - National Center for Biotechnology Information Sequence Read Archive.

1.3.1. Secuenciación de segunda generación ("lectura corta")

La secuenciación de segunda generación (NGS) se caracteriza por el uso de la fragmentación generando **insertos** cortos de cDNA que representan a los transcritos originales, es por ello que también se la denomina secuenciación de "lectura corta". Además, la amplificación en grupos de copias o clústeres que utiliza asegura una señal amplia que se detecta con un alto nivel de seguridad estadística. La señal de la secuenciación es detectada por una cámara que registra la emisión de luz generada por el nucleótido incorporado. Es en este punto donde se diferencian las dos principales estrategias de la secuenciación de segunda generación: la pirosecuenciación y la secuenciación por síntesis (SBS).

La primera versión de RNA-seq en implementarse fue la RNA-seq 454, desarrollada por la compañía Roche en 2006 basándose en el sistema de secuenciación por pirólisis (Brainbridge *et al*., 2006). Esta plataforma detecta la emisión de la luz proveniente indirectamente del pirofosfato generado tras la unión del nucleótido (dNTP) a la cadena creciente. Cuatro rondas de reacción son necesarias para saber de cuál se trata: A, C, G o T.

Poco más tarde, aparecieron nuevas soluciones a través de la secuenciación por síntesis usada en los diferentes sistemas que desarrolló Illumina como MiSeq o HiSeq. Este tipo de RNA-seq se estableció como la más popular hasta el día de hoy para el estudio de expresión génica. A diferencia de la pirólisis, este sistema añade los nucleótidos marcados con cuatro tipos de fluoróforo y, por lo tanto, detecta en una sola ronda el color de fluorescencia correspondiente. Sus lecturas son típicamente de 30 a 400 pb.

Enlace de interés

En el siguiente enlace podrás encontrar un vídeo explicativo de las diferentes etapas de la secuenciación por síntesis (SBS) de Illumina.
https://www.youtube.com/watch?v=fCd6B5HRaZ8&t=7s&ab_channel=Illumina

Para el análisis de estos nuevos datos, muchas de las herramientas bioinformáticas desarrolladas para el análisis de la expresión diferencial de microarrays fueron sencillamente implementadas para su uso con la RNA-seq. Un ejemplo es el paquete de R edgeR (2010), cuyo código se ha basado en el original "limma" para adaptarse a la distribución binomial negativa que presentan estos datos.

Algunos nuevos retos, como el mapeo de secuencias cortas a un genoma de referencia de gran tamaño, fueron enfrentados mediante colaboraciones entre matemáticos, biólogos computacionales y bioinformáticos. Así, por ejemplo, fue como se desarrolló el mapeador de TopHat (2009) cuyo algoritmo utiliza la transformación de Burrows-Wheeler (BWT) creada anteriormente y utilizada también en la compresión de archivos de texto. Gracias a los esfuerzos para secuenciar mRNA, exones y diferentes formas de transcritos del mismo gen, estas **características de secuencia** empezaron a anotarse en florecientes bases de datos de secuencias de referencia curadas y de acceso libre. Entre ellas, las colecciones más importantes son la europea Ensembl (2002) y la estadounidense RefSeq del NCBI (2005).

Enlace de interés

Ensembl anota genes, calcula múltiples alineaciones, predice la función reguladora y recopila datos de enfermedades.
https://www.ensembl.org/index.html

Enlace de interés

Un conjunto completo, integrado, no redundante y bien anotado de secuencias de referencia que incluyen genómica, transcripción y proteína.
https://www.ncbi.nlm.nih.gov/refseq/

La secuenciación por síntesis (SBS) desarrollada por Illumina es la tecnología más usada hoy en día para el estudio de expresión génica debido a su alto rendimiento y bajo coste.

1.3.2. Secuenciación de tercera generación ("lectura larga")

Ya en 2010, dos empresas principalmente, Pacific Biosciences (PacBio) y Oxford Nanopore Technologies (ONT), competían por conseguir el mejor dispositivo para la llamada secuenciación de tercera secuenciación. Estas nuevas tecnologías prometen una mejora en la transcriptómica al evitar el paso de la fragmentación y ser capaces de secuenciar transcritos enteros por lo que se denomina también como secuenciación de "lectura larga".

El método propuesto por ONT en su secuenciación de nanoporos (2010) nace de la idea de la secuenciación directa de ARN (*direct* RNA-seq) en la cual los ARN poliadenilados individuales pueden secuenciarse directamente, sin necesidad de retrotranscripción (Ozsolak *et al.*, 2009).

El principio en el que se basa es la semiconducción, en la que cada nucleótido provoca una fluctuación de la corriente iónica al paso de su secuencia a través de un poro diseñado especialmente.

Aunque la aplicación de esta nueva tecnología está todavía en desarrollo, ya ha conseguido longitudes de lectura de hasta de 2,3 Mb (Amarasinghe *et al.*, 2020) y se comercializa además en una plataforma portátil llamada MinION.

Enlace de interés

En el siguiente enlace podrás ver un vídeo acerca del diseño de la plataforma MinION desarrollada para la secuenciación de lectura larga a partir de RNA o cDNA.
https://www.youtube.com/watch?v=1_mER5qmaVk&ab_channel=OxfordNanoporeTechnologies

Por otro lado, PacBio siguió la línea de secuenciación por síntesis en PacBio IsoSeq (2011) en la que incluye una variación para su uso a partir de una molécula única, idea que fue publicada anteriormente (Braslavsky *et al.*, 2003). Este método monitoriza la fluorescencia emitida con un nuevo sistema que circulariza la molécula de cDNA.

La lectura así generada, en lugar de integrar las señales de cientos de copias de un clúster como hacía la secuenciación de segunda generación, es la secuencia consenso de varias rondas de síntesis consecutivas de la misma molécula. Su longitud está, por lo tanto, limitada por la durabilidad de la polimerasa, aunque actualmente ya podemos obtener lecturas de hasta 60 000 nucleótidos (60 kpb).

Enlace de interés

En el siguiente enlace podrás ver un vídeo explicativo de la tecnología de PacBio:
https://www.youtube.com/watch?v=_lD8JyAbwEo&t=3s&ab_channel=PacBio

En cuanto a la precisión, estos nuevos métodos de RNA-seq tienen tasas de error más altas en comparación con la secuenciación de lectura corta. Sin embargo, los métodos más nuevos como ONT directo RNA-seq limitan los errores al evitar la fragmentación y la conversión de cDNA.

La secuenciación de ARN de lectura larga permite hoy secuenciar transcritos en la totalidad de su longitud, lo que permite, por ejemplo, caracterizar nuevas **isoformas**.

Enlace de interés

En el siguiente enlace puedes encontrar un manual de análisis de datos RNA-seq de lectura larga elaborado por el proyecto ENCODE:
https://www.encodeproject.org/pipelines/ENCPL075LUJ/

1.3.3. Transcriptómica espacial

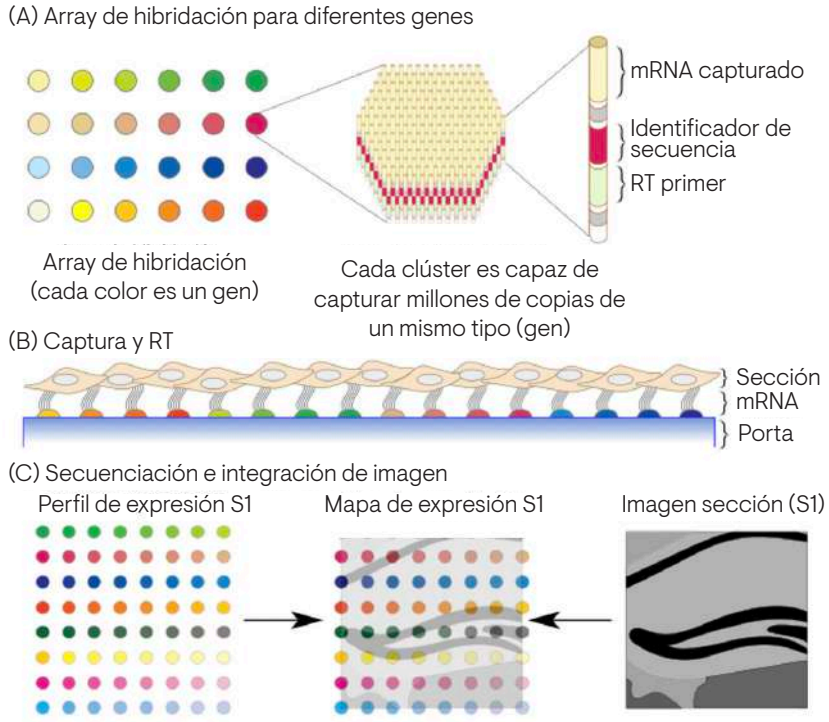
Desde la década del 2010, hubo diferentes prototipos que intentaron trasladar las ventajas del rendimiento de la secuenciación a un nivel celular. La secuenciación de célula única o *single-cell* RNA-seq es un ejemplo y, aunque su nombre lleve a error, todavía no puede aplicarse a partir de una sola célula. Para enriquecer el cDNA de la muestra en una línea celular se deben usar monocultivos en placa. Esta técnica ha conseguido descubrir transcritos de muy baja expresión, nunca descritos anteriormente (Tang *et al*., 2009).

Con este objetivo de entender cómo se relacionan entre sí las poblaciones celulares en un organismo, emergió hace relativamente poco una técnica que aglutina los principios de la captura *in situ* y la secuenciación. Esta tecnología, llamada transcriptómica espacial (ST), fue publicada por primera vez en 2016 (Ståhl *et al*., 2016) y tiene una prometedora carrera en la investigación fisiológica.

El protocolo consiste en la preparación histológica del tejido de interés en diferentes secciones que son escaneadas para conocer la posición de las diferentes células que se distinguen. Cada una de estas secciones es hibridada con una plataforma de nanoporos donde están previamente fijados los cebadores de los genes que se quieren conocer. Estos cebadores han sido colocados en el portaobjetos siguiendo las coordenadas detectadas y se distinguen entre ellos por un código de barras (Figura 5).

Figura 5

Descripción general de la tecnología transcriptómica espacial



Mientras la sección de tejido está unida al portaobjetos, se inicia la transcripción inversa del mRNA capturado y el cDNA resultante incorpora el código de barras del cebador. Todo el material es extraído después para la preparación de la biblioteca y su lectura mediante NGS. El código de barras permite rastrear cada secuencia generada por secuenciación y mapearla en su posición original en el corte histológico. La reconstrucción de varias secciones genera una imagen tridimensional muy informativa de la expresión del tejido a detalle celular, o casi.

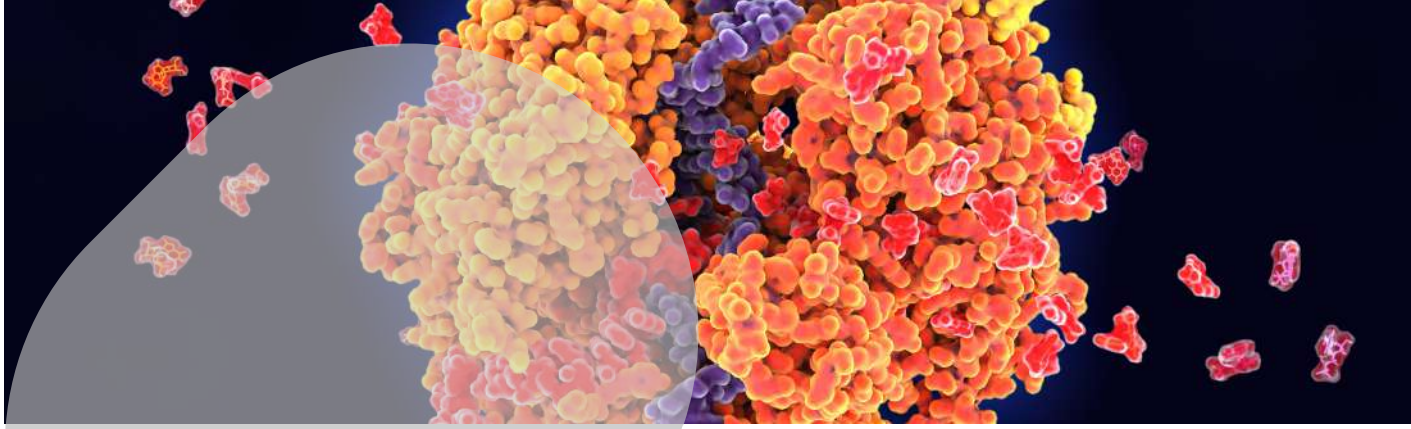
Enlace de interés

En el siguiente enlace puedes encontrar un manual de análisis para datos de ST publicados recientemente (Maynard *et al*., 2021) con la herramienta de Bioconductor spatialLIBD.
http://research.libd.org/spatialLIBD/

A finales de 2018, la tecnología ST fue adquirida por la empresa 10X Genomics bajo el nombre Visium Spatial Gene Expression que implementó algunas mejoras en su resolución y tiempo de ejecución. La captura puede ser aleatoria o dirigida, llegando a alcanzar una resolución de 18 000 genes. Esta emergente tecnología ya se ha empezado a aplicar con éxito en investigación anatómica, por ejemplo, sobre el estudio de órganos como el cerebro o el corazón. Algunos **flujos de trabajo** con estos datos empiezan a funcionar para desarrollar mapas de expresión tridimensional.

Enlace de interés

El siguiente enlace corresponde con la página web de la compañía 10xgenomics, donde se muestra un vídeo explicativo de la tecnología ST que ofrecen:
https://www.10xgenomics.com/products/spatial-gene-expression



Capítulo 2

Estudios de expresión génica con datos de NGS

En este segundo capítulo nos detendremos en la técnica más importante hasta ahora por la popularidad de su uso y producción científica dentro de las ciencias transcriptómicas, la NGS. Veremos qué características la hacen tan atractiva y tomaremos nota de algunas recomendaciones para su petición al servicio de secuenciación. Repasaremos en detalle la metodología de su protocolo y los puntos clave para obtener unos datos de calidad. También veremos las opciones para descargar NGS en las bases de datos.

2.1. ¿Por qué elegir la NGS y qué tipo?

Actualmente, las lecturas cortas que hoy encontramos principalmente distribuidas por el sistema de SBS de Illumina son la opción más usada dentro de toda la gama de RNA-seq o de otras basadas en arrays. Aunque la secuenciación de tercera generación manifiesta un auge en su uso, estos sistemas se destinan más bien a los estudios de caracterización genómicos donde disminuyen la necesidad de ensamblaje. En los estudios de niveles de expresión, el sistema de NGS presenta una serie de ventajas a diferencia de los anteriores microarrays:

- Cantidades de ARN de partida muy bajas del orden de nanogramos frente a los microgramos que requerían los microarrays.
- Detección no-dirigida que sucede sin necesidad de sondas específicas. Con NGS podemos detectar, en principio, cualquier ARN, lo que permite identificar nuevos transcritos.

- Alto nivel de sensibilidad cuya tasa de error es muy baja, lo que hace a la NGS capaz de detectar variantes de un solo nucleótido (SNP) además de pequeñas inserciones y deleciones.
- Alta especificidad evitando los problemas de hibridación cruzada de los microarrays.
- Un rango dinámico cinco veces mayor que los microarrays. Esto significa que sus niveles de saturación y de ruido son muy altos y muy bajos, respectivamente. Con este amplio umbral es capaz de recoger la expresión de genes muy activos al mismo tiempo que genes con muy poca expresión.
- Reproducibilidad técnica tan alta que no es necesario usar réplicas técnicas (biológicas sí).

La NGS se ha convertido en la herramienta de generación de datos mayoritariamente elegida para el estudio de la expresión génica, entre otras cosas, porque permite detectar transcritos en un amplio rango de expresión.

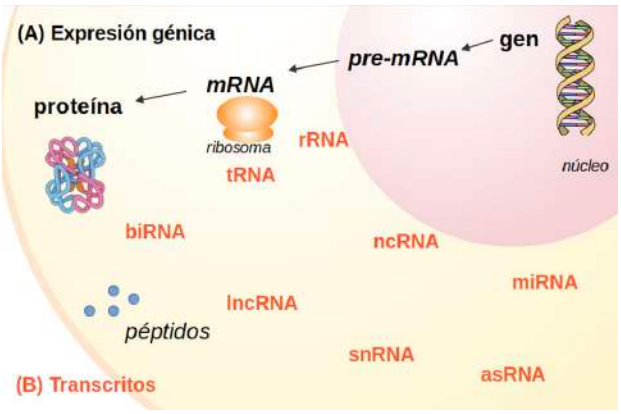
La tecnología de secuenciación para un análisis como este suele ser la HiSeq (3000/4000) de Illumina, recomendándose utilizar el mismo carril en la celda de flujo para las muestras a comparar. En cuanto a la preparación de la librería, las parejas de lecturas de un tamaño entre 75 y 150 pb con un protocolo específico de cadena son las más adecuadas para un análisis óptimo. La profundidad necesaria depende de la actividad génica de las muestras y del número de genes, pero en general se suelen solicitar entre 30 y 50 millones de pares de lecturas finales.

El costo de una secuenciación como esta ronda normalmente los 200-250 euros por muestra (IVA incluido) a día de hoy. La mayoría de los servicios, además, ofrecen un análisis de la calidad de la muestra de RNA (Bioanalyzer) y de la calidad de las secuencias obtenidas (Library Quality Control) dentro del precio establecido.

Una de las revoluciones científicas alcanzadas con el uso de esta tecnología es el descubrimiento de otras formas de ARN o biotipos no directamente vinculadas con la expresión de proteína. Algunos ejemplos son la existencia de nuevos transcritos no codificantes (ncRNA) o no directamente codificantes, como algunos ARN de tipo largo (lncRNA) que generan péptidos pequeños con función reguladora de la propia transcripción (Figura 6).

Figura 6

Esquema representando la expresión génica en la célula como una de las formas de transcripción de ARN entre otras, a veces, no codificantes



Estos descubrimientos han cambiado en los últimos años el escenario de los procesos postranscripcionales y del mismo papel del ARN en el funcionamiento de la célula dentro del organismo, cuya función anterior estaba ligada únicamente a la expresión génica. Para mayor claridad, en las siguientes páginas nos referiremos al estudio de estos datos de manera enfocada al estudio en la expresión génica, aunque el estudiante debe conocer que tanto protocolos como análisis pueden ser adaptados para la identificación específica de otras formas de ARN.

2.2. Protocolo básico de NGS

Previamente al protocolo más común para la obtención de datos de NGS, generalmente llevado a cabo por un servicio de secuenciación, deberemos extraer la muestra en el laboratorio. En general, existen métodos estandarizados de extracción como los kits comerciales de RNeasy (Kiagen) que facilitan la tarea. El proceso consiste básicamente en romper la membrana celular con algún detergente para liberar las moléculas de ARN. En el caso de que el tejido sea vegetal será necesaria también la rotura mecánica de las paredes con mortero o *tissuelyser*.

La integridad del ARN es determinante para el éxito de la secuenciación por lo que deberá ser comprobada antes. Para ello usaremos, por ejemplo, un aparato Bioanalyzer o Qubit que cuantifica por fluorometría el nivel de integridad. Si las muestras cumplen un valor mínimo de 7 RIN (el máximo es 10), estarán listas para su envío. En cuanto a la cantidad, necesitaremos un mínimo de unos 1000 ng de ARN total por muestra.

Antes de secuenciar, la calidad del ARN en el material de partida es decisiva para lograr una representación ajustada de los transcritos expresados en la muestra. Por ello se recomienda evitar la acción de las enzimas que degradan el ARN con frío y guantes además de eliminar el ADN restante.

Una vez en el servicio de secuenciación, podemos distinguir tres etapas principales en el proceso de secuenciación de lecturas cortas:

2.2.1. Preparación de la biblioteca

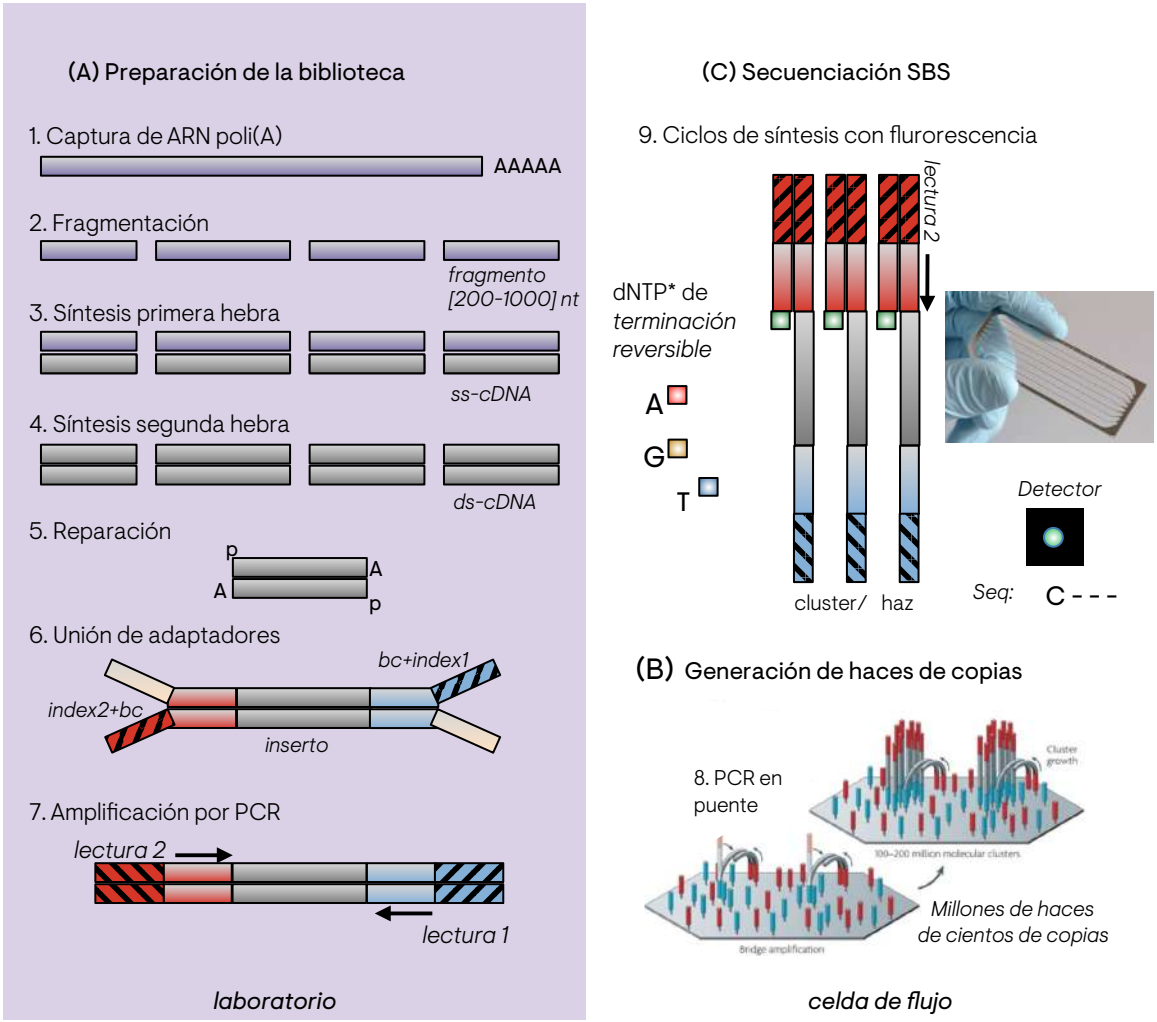
Paso opcional - Un paso previo en la preparación de la librería que muchos servicios ofrecen es el enriquecimiento de mRNA. Dado que el ARN ribosomal (rRNA) constituye la mayor parte del ARN de la célula, se supone que puede causar sobresaturación especialmente si la muestra tiene bajos niveles de ARN. Para evitarlo, es posible seleccionar el mRNA maduro por la cola de poli(A) o eliminar el rRNA (*depletion*) si la muestra proviene de organismo procariota (sus ARN no se poliadenilan). Sin embargo, en los últimos años se empieza a recomendar evitar este paso por los sesgos que puede introducir.

El ARN es fragmentado por ultrasonicación generalmente y los segmentos resultantes de un tamaño entre 200 a 1000 nt son seleccionados. Estos fragmentos son usados como molde para la síntesis de la primera hebra mediada por RT y seguida de la síntesis de la segunda hebra por una polimerasa de ADN (Figura 7).

Esta segunda reacción da como resultado la pérdida de la hebra, lo que puede evitarse con el marcaje químico en la modalidad de lecturas “específicas de hebra” o *strand-specific*.

Figura 7

Descripción general del protocolo de secuenciación de lecturas cortas y parejas (SBS de Illumina)



El producto de estas dos reacciones es una doble cadena que llamamos el inserto para diferenciarlo del fragmento original. El inserto es adenilado y reparado en sus extremos para poder ligar eficazmente los adaptadores en "Y". Estos adaptadores contienen tres tipos de secuencia: un *index*, un identificador o "código de barras" (bc en la figura) y un cebador. El *index* es complementario a los oligos que están unidos covalentemente a la superficie de la celda de flujo, mientras que el cebador permite a la polimerasa empezar su extensión cuando la reacción de síntesis empiece. En un sistema de lecturas emparejadas como el que se muestra en la figura 7, se utilizan dos tipos de cebadores para distinguir una lectura y la otra. El identificador permitirá distinguir el origen de la muestra dentro de la misma celda.

2.2.2. Generación de los haces de copias

Esta segunda etapa ocurre ya en el soporte sólido de la celda de flujo o *flow cell* que la plataforma ha desarrollado cuya superficie está nanométricamente horadada distribuida en diferentes carriles o *lanes* en los que la reacción de polimerización sucede de manera independiente.

Enlace de interés
https://www.youtube.com/watch?v=pfZp5Vgsbw0&ab_channel=Illumina

En cada nanopozo, a ambas caras de la superficie, hibrida una sola molécula de inserto de una biblioteca total de miles de moléculas. Una vez unidos los insertos a través de sus adaptadores, la cámara se somete a varias rondas de desnaturalización-polimerización en un sistema llamado de "amplificación en puente". Esto genera cadenas complementarias de longitud completa que van quedando unidas a través de los oligos a la superficie en grupos de haces. Cuando estos grupos de secuencias clonadas alcanzan un tamaño de diámetro de aproximadamente 1 micra empiezan a ser ópticamente visibles y se detiene la amplificación.

2.2.3. Secuenciación SBS

Tras la amplificación de la biblioteca, cada transcrito original está representado por cerca de mil copias de su cDNA. En este momento, se retira el medio de reacción y se añaden los cuatro tipos de **nucleótidos de terminación reversible** (A, C, T, G) marcados con un fluoróforo distintivo. En cada ciclo, se une un nuevo dNTP que detiene la síntesis de la nueva cadena por la polimerasa. En este lapso de tiempo, una cámara recoge la fluorescencia correspondiente y se continúa con un lavado para dar paso a la siguiente ronda de reacción. Este tipo de proceso permite la secuenciación de regiones homopoliméricas (por ejemplo: GGGG) con un alto nivel de confianza a diferencia de otras. Durante la identificación de bases, los valores de intensidad de fluorescencia son transformados en una matriz de valores de confianza. La tasa de error aumenta progresivamente con un número creciente de ciclos. Actualmente, en la SBS de Illumina se admiten hasta 150 ciclos con una tasa de error global de solo el 0,2 %.

Durante la secuenciación por síntesis, cada punto o *spot* de fluorescencia detectado por la cámara dentro de la celda de flujo representa cerca de 1000 reacciones de unión de nucleótido de terminación reversible.

2.3. Experimentos de NGS en las bases de datos

Las bases de datos destinadas a la información de secuencia de la biología genómica existen desde hace ya 30 años, desde que en 1980 el EMBL empezara a recoger estos datos de la primera generación de secuenciación (ver Figura 1). Desde entonces, el organismo internacional de colaboración entre bases de datos públicas (INSDC) se encarga de su coordinación, asegurando el mantenimiento de los datos y su distribución pública. Además, desde este organismo se unifican métodos de almacenaje de datos, formatos de distribución y normas de identificación o metadatos. Este nodo centraliza las aportaciones a nivel mundial de los tres portales principales:

- El Banco de Datos de DNA de Japón (DDBJ) del Instituto Nacional de Genética en Mishima (Japón).
- El Archivo Europeo de Nucleótidos (ENA) del Instituto Europeo de Bioinformática del Laboratorio Europeo de Biología Molecular (EMBL-EBI) en Hinxton (Reino Unido).
- El GenBank del Centro Nacional de Información Biotecnológica (NCBI) en Maryland (EE. UU.).

Esta colaboración internacional garantiza la estandarización y el acceso a los datos de toda la comunidad científica, cuya información es comunicada y actualizada regularmente (cada mes). Los archivos depositados pueden ser tanto "datos crudos" como "datos procesados".

Este volumen de datos almacenados ha sufrido un crecimiento exponencial en los últimos años debido principalmente a la aparición de la NGS. El tamaño total, que ya ha superado los 9 petabytes, se espera que crezca 10 veces cada 4 años (Arita *et al.*, 2021). Esto, aunque es una ventaja para el usuario investigador pues representa un mayor número de muestras disponibles con las que comparar sus experimentos, sin embargo, empieza a ser un problema para los equipos de almacenaje. De hecho, el NCBI y las API del EMBL-EBI están comenzando a ofrecer datos de lectura en un formato menos pesado (CRAM) y para liberar espacio el NCBI ha trasladado los datos de la SRA a entornos comerciales "en la nube". Mientras tanto, EMBL-EBI y DDBJ están más dedicados a las soluciones locales.

Dado que los tres nodos reflejan los datos completos, los usuarios se benefician de múltiples opciones dependiendo de su entorno informático. Las principales bases de datos de RNA-seq corresponden con estas direcciones (Tabla 2).

Tabla 2
Bases de datos de NGS

Nombre	Institución	Datos	Descripción
Gene Expression Omnibus (GEO)	NCBI		Primera base de datos de transcriptómica que acepta datos de cualquier fuente. Introdujo los estándares comunitarios MIAME y MINSEQE que definen los metadatos necesarios del experimento para asegurar una interpretación y repetibilidad efectivas.
Expression	EBI	RNA-Seq	Base de datos de expresión génica específica de tejidos para animales y plantas. Muestra análisis y y visualización secundarios, como el enriquecimiento funcional de términos de ontología, dominios InterPro o rutas. Enlaces a datos de abundancia de proteínas cuando estén disponibles.
RefEx	DDBJ	Todos	Transcriptomas humanos, de ratón y de rata de 40 órganos diferentes. Expresión genética visualizada como mapas de calor proyectados en representaciones 3D de estructuras anatómicas.
NONCODE	noncode.org	RNA-Seq	RNA no codificantes (ncRNA) excluyendo ARNt y ARNr.

Nota. Adaptado de "Transcriptomics technologies", por R. Lowe, N. Shirley, M. Bleackley, S. Dolan y T. Shafee, 2017, *PLoS Computational Biology, 13*(5), e1005457. https://doi.org/10.1371/journal.pcbi.1005457
Legenda: NCBI – National Center for Biotechnology Information; EBI – European Bioinformatics Institute; DDBJ – DNA Data Bank of Japan; ENA – European Nucleotide Archive; MIAME – Minimum Information About a Microarray Experiment; MINSEQE – Minimum Information about a high-throughput nucleotide SEQuencing Experiment.

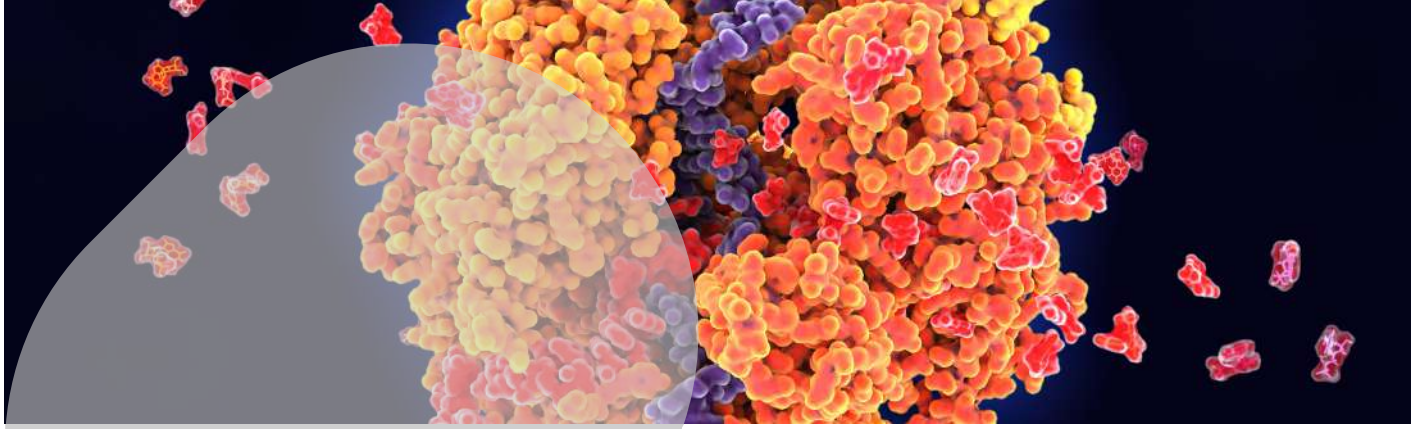
En ellas podemos encontrar secuencias de una gran cantidad de organismos vivos, algunos son organismo modelo y otros tienen un especial interés para la investigación, como el humano. Además de estos portales que ponen en conjunto la información de secuencia, han surgido iniciativas independientes dentro de cada comunidad de investigación en un organismo en particular, son las bases de datos específicas del organismo. En caso de pertenecer a una comunidad, este puede ser el mejor recurso por su estado de actualización y detalle de depuración.

Durante las clases utilizaremos la herramienta SRA Toolkit para la descarga de muestras de datos de NGS de una de las principales bases de datos. Estos archivos nos servirán como material de inicio para el procesado y análisis de la expresión génica.

Enlace de interés

En el siguiente enlace puedes encontrar el manual que seguiremos en clase para la descarga de datos transcriptómicos utilizando la herramienta SRA Toolkit.
https://github.com/ncbi/sra-tools/wiki/02.-Installing-SRA-Toolkit

Hablamos de metadatos para referirnos a la información asociada a la muestra de origen, como, por ejemplo: el organismo al que pertenece, el tejido, el momento de recogida o el método de secuenciación (Principios FAIR).



Capítulo 3

Análisis de datos de NGS

En este tercer capítulo, haremos una introducción del flujo de trabajo más común para el análisis de la expresión génica a partir de datos de NGS y de los diferentes tipos de archivo que se requieren. Además, abordaremos los primeros pasos del análisis que corresponden con el manejo de los datos de secuencia, pasando por el preprocesado de las lecturas, su mapeo y su cuantificación hasta obtener una matriz de recuentos.

3.1. Introducción al flujo de trabajo con datos NGS y sus archivos resultantes

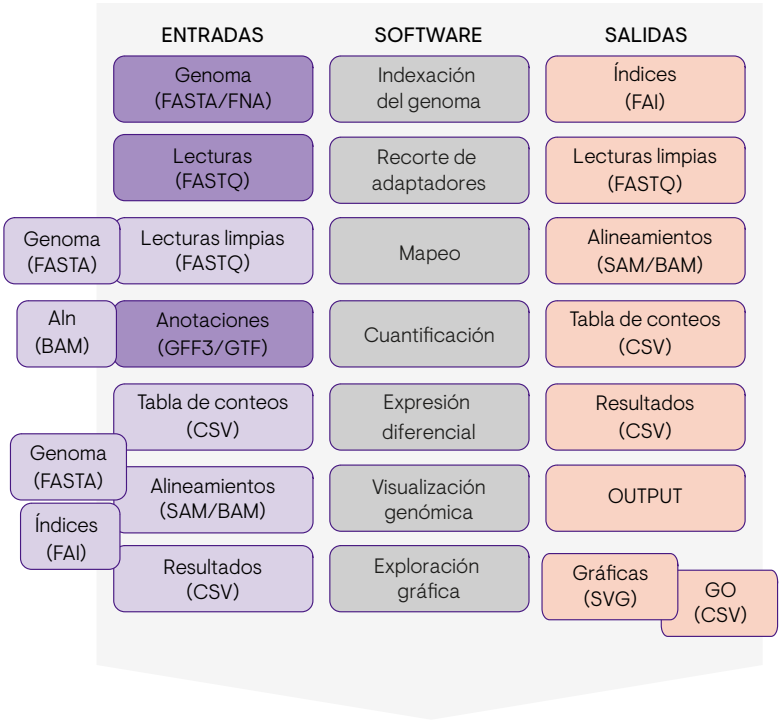
Dentro del procesado bioinformático de datos de NGS para la obtención de resultados de expresión génica se suceden una serie de procesos que corresponden con: el mapeo de las lecturas en una posición genómica dentro de un genoma de referencia, la estimación de su abundancia, el análisis estadístico de su expresión diferencial y la exploración de sus resultados.

Para su ejecución, el usuario deberá tener un cierto nivel de conocimiento en computación, sistema de archivos y programación (Bash y R). Si bien existen algunas herramientas disponibles listas para usar tipo *user friendly*, la mayoría de los programas de código libre desarrollados por la comunidad se usan directamente en entornos de línea de comandos. Estos entornos existen, por ahora, en sistemas operativos Linux (Unix, Bash) o Mac (tipo Unix).

El análisis de datos NGS es un proceso computacionalmente intensivo (especialmente el mapeo) que requiere el almacenamiento, la transferencia y el procesamiento de archivos de datos muy grandes. Por ello se recomienda tener acceso a un computador avanzado o acceder a través de alguna red. Trabajar con estos archivos requiere estar familiarizado con varios de los formatos que se han estandarizado para su trabajo en genómica. Los diferentes pasos pueden diferenciarse por el tipo de archivo de "entrada" y de "salida" de cada uno de los programas que se utilizan (Figura 8).

Figura 8

Flujo de trabajo para el análisis de la expresión génica con datos de NGS indicando las entradas y salidas en el procesado



3.1.1. Archivo de datos de lecturas FASTQ

Corresponde con el archivo de salida del llamado "análisis primario" cuyo proceso está automatizado en el *software* del secuenciador. Este programa analiza la imagen detectada por la cámara del secuenciador y establece la composición y calidad de las bases de la lectura. En la SBS, cada ciclo de PCR genera una imagen del conjunto de las señales lumínicas en la celda de flujo de manera que 100 imágenes darán las lecturas de 100 nt. Su calidad de base es cuantificada en relación con la intensidad de la fluorescencia total de los cuatro fluorocromos utilizados.

El formato FASTQ organiza la información de cada lectura en cuatro líneas:

- Línea 1: empieza con el carácter '@' y va seguida del identificador de secuencia.
- Línea 2: secuencia FASTA de la lectura.
- Línea 3: comienza con el carácter '+'.
- Línea 4: secuencia de la calidad de base de la lectura.

El carácter que codifica la calidad de base en la línea 4 corresponde con una puntuación Phred (Q) que se define como: $Q=10\log_{10}P$, donde P es la probabilidad de error en la identificación de esa base. Por ejemplo, 10, 20 y 30 Phred corresponden con 1 error de cada 10, 100 y 1000 designaciones de base, respectivamente.

Ejemplo

Línea de comando para ver las primeras líneas de un archivo comprimido GZ como FASTQ:

```
$ zcat < file.fastq.gz > | head -4
@D00733:474:CE41UANXX:6:2216:1233:2233 1:N:0:ACAGTG
CTCATGAGTCAGCCCCTTGGTTTGGCTTGGCTGAGCCGATTTCCCAGCGCCGAAG
+
CBB@BCC=1>@FF@G>GGG=FGGGBCGGGGGGEGGGFA>EGGGGG>DGDGG><
```

Enlace de interés

En el siguiente enlace podrás encontrar una tabla que muestra la correspondencia entre el carácter de codificación, su código ASCII y la puntuación de calidad Phred.
https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm

3.1.2. Archivo de datos genómicos FASTA

Este formato se caracteriza por usar una línea de cabecera señalada por el símbolo '>' que describe la secuencia que le sigue a continuación (el mismo que para proteínas). En el flujo de trabajo con datos de NGS es el archivo de entrada que recoge la secuencia del genoma de referencia.

Ejemplo

Línea de comando para seleccionar las primeras que empiezan por '>' en un archivo FASTA:

```
$ grep -i '>' Zea_mays.B73_RefGen_v4.dna.toplevel.fa | head -3
>1 dna:chromosome chromosome:B73_RefGen_v4:1:1:307041717:1 REF
>2 dna:chromosome chromosome:B73_RefGen_v4:2:1:244442276:1 REF
>3 dna:chromosome chromosome:B73_RefGen_v4:3:1:235667834:1 REF
```

3.1.3. Archivo de alineamientos SAM/BAM

En el flujo de trabajo, los archivos anteriores sirven como entrada para procesar el mapeo generando un archivo de alineamientos en formato SAM, o su forma comprimida, BAM. Este archivo contiene la posición genómica, la dirección y la calidad de mapeo (MAPQ) de cada una de las lecturas.

Ejemplo

Línea de comando para ver las primeras líneas de un archivo comprimido BAM:

$ samtools view <file.bam>| head -2

D00733:271:CBKNNANXX:3:2314:6568:85769 99 chr1 102955 60 125M = 103007 177
CATCTATAAACAACTTAGTCAAAAACCAAAAATTACATACAAAAGAATTTATCTTTTGAATCGAAAGCTC
CTCATTTTCCCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGE
MC:Z:125M MD:Z:125 PG:Z:MarkDuplicates RG:Z:alldnaseqNM:i:0 MQ:i:60 UQ:i:0 AS:i:125 XS:i:54

D00733:271:CBKNNANXX:3:2314:6568:85769 147 chr1 103007 60 125M = 102955 -177
CTTTTGAATCGAAAGCTCCTCATTTTCAAAAGCAATCCTGTTTTTGTCATCCCAAACCGACCAAAAAA
GGCATAAGGGGGGGGGGGGGGGEGGFGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGG
MC:Z:125M MD:Z:125 PG:Z:MarkDuplicates RG:Z:alldnaseqNM:i:0 MQ:i:60 UQ:i:0 AS:i:125 XS:i:21

Enlace de interés

En el siguiente enlace podrás encontrar más información acerca del formato SAM.
https://samtools.github.io/hts-specs/SAMv1.pdf

3.1.4. Archivo de anotaciones GFF/GFT

El archivo de anotaciones es usado junto con el archivo SAM/BAM de las lecturas mapeadas en el paso de cuantificación de la expresión. Su información se organiza en 9 columnas que indican, entre otras cosas, las posiciones de inicio y fin de cada uno de los exones que forma un gen y su hebra correspondiente.

Ejemplo

Línea de comando para ver las primeras líneas de un archivo comprimido GFF3:

$ head -300 *gff3 | sed 's/\t/ /g' | tail -n 23

1 gramene five_prime_UTR 44289 44350 . + . Parent=transcript:Zm00001d027230_T001

1 gramene exon 44289 44947 . + . Parent=transcript:Zm00001d027230_T001;
Name=Zm00001d027230_T001.exon1;constitutive=1;ensembl_end_phase=0;
ensembl_phase=-1;exon_id=Zm00001d027230_T001.exon1;rank=1

1 gramene CDS 44351 44947 . + 0 ID=CDS:Zm00001d027230_P001;Parent=transcript:
Zm00001d027230_T001;protein_id=Zm00001d027230_P001

Enlace de interés

Esta página describe las características y reglas de anotación para el formato GFF3 o GTF.
https://www.ncbi.nlm.nih.gov/genbank/genomes_gff/

En general, se establece que el archivo resultante mantenga al menos una línea silenciada (#) con la información del comando o paso anterior que lo generó. Esto permite hacer un seguimiento de los distintos procesos realizados sobre los datos y favorece una mayor reproductibilidad en los resultados.

3.2. Preprocesado de las lecturas

El primer paso del análisis de las secuencias de NGS es el preprocesado de las lecturas. Su objetivo es disminuir los errores de mapeado posteriores identificando aquellas secuencias exógenas al transcriptoma de la muestra como son los propios adaptadores u otras posibles contaminaciones.

3.2.1. Control de calidad

Una vez recibidos los datos del servicio de secuenciación, se debe proceder a una evaluación general de la calidad de lecturas con una herramienta de control como FastQC. Esta herramienta toma al azar un porcentaje de la muestra y analiza parámetros de calidad como:

- El contenido de adaptadores.
- La calidad de base a lo largo de la lectura.
- Los niveles de duplicación.
- El contenido en bases (A, C, G, T).
- El contenido de bases no identificadas o "N".
- La proporción de GC.

Enlace de interés

En el siguiente enlace podrás encontrar una descripción en profundidad de los diferentes análisis de calidad de las secuencias utilizados por la herramienta FastQC.
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

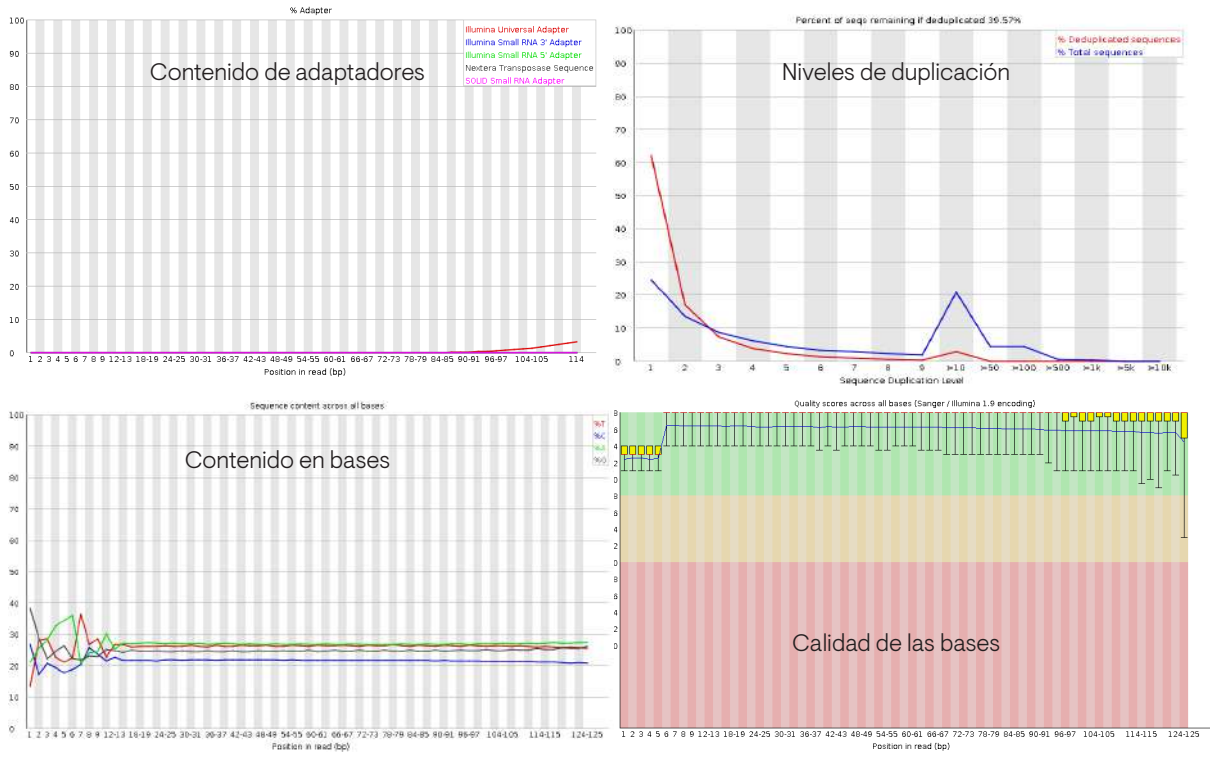
En general, el contenido en adaptadores depende del tamaño de inserto, el cual está definido por el tratamiento de fragmentación y su selección durante la preparación de la librería (ver sección 2.2.1.).

En principio, este tamaño es mayor que la longitud de la lectura (entre 75 y 150 pb), de manera que durante su amplificación la cadena creciente no superará al siguiente adaptador.

Sin embargo, siempre queda un pequeño porcentaje de fragmentos cortos que suelen representar menos del 5 % en el total de las lecturas (Figura 9).

Figura 9

Resultados típicos del control de calidad de lecturas de NGS producidos por FASTQC para el contenido de adaptadores, niveles de duplicación, contenido y calidad de bases



Por otro lado, otro efecto de la técnica es la disminución de la calidad de la base (*phasing*) hacia el final de la lectura. Esto es debido en la secuenciación por síntesis a un desfase acumulado entre el bloqueo del nucleótido que emite la señal y el siguiente que lo reemplaza. Al principio de la polimerización, el haz de cadenas crecientes hace este recambio al unísono y genera una señal luminosa inequívoca de la base. Pero, hacia el ciclo 75-100, la probabilidad de que el anterior nucleótido se encalle es mayor, lo que provoca una señal en conjunto cada vez menos homogénea. En general, la calidad de base cae a 10 Phred hacia el ciclo 150, por eso se recomienda una longitud de lectura dentro de este rango.

El contenido en bases nos da una idea de la diversidad de secuencia que tenemos en las muestras junto con el nivel de duplicación. Estas duplicaciones se suponen poco frecuentes teniendo en cuenta que la fragmentación y los procesos enzimáticos siguientes se dan al azar dentro del conjunto de moléculas originales. Sin embargo, los transcritos muy abundantes y el posible sesgo de la actividad polimerasa, con menor afinidad catalítica por ciertas secuencias, pueden provocar alguna desviación.

3.2.2. Eliminación de adaptadores

En cuanto a la eliminación o marcaje de adaptadores, existen diferentes herramientas para realizarlo. Entre ellas las más populares son Cutadapt (Trim Galore!) y MarkIlluminaAdapters (GATK). La segunda opción permite no eliminar la secuencia del adaptador, sino marcarla con una calidad de base de 2 Phred, con lo que no será tenida en consideración para procesos posteriores al ser demasiado baja.

Durante las clases aprenderemos a usar Trim Galore! (Babraham Bioinformatics), con la que eliminaremos las secuencias de:

- Adaptadores.
- Baja calidad (< 30 Phred).
- Poca longitud (< 20 nt).
- Lecturas impares.

Enlace de interés

En el siguiente enlace de GitHub podrás encontrar documentación y el archivo de programa de la herramienta Trim Galore!
https://github.com/FelixKrueger/TrimGalore/blob/master/Docs/Trim_Galore_User_Guide.md

3.3. Mapeo de las lecturas

Una vez descartadas las secuencias de los adaptadores y de baja calidad de las lecturas, es necesario mapear las lecturas a una secuencia de referencia. Esta secuencia puede ser un genoma o un transcriptoma de referencia, en el que los sitios de inicio y fin de cada gen estén previamente anotados. Si no se dispone de tal, es posible realizar un ensamblaje *de novo* en el que cada transcrito será reconstruido sobre una secuencia consenso de las representadas en las lecturas (no es objeto de este flujo de trabajo).

3.3.1. Genoma de referencia

La elección de un genoma de referencia adecuado es importante para tener una representación completa y actualizada de las regiones exónicas. Para ello acudiremos en primer lugar a la base de datos específica del organismo de interés, o, en su lugar, a cualquiera de los tres portales principales (ver sección 2.3.). En ellas encontraremos el archivo FASTA con estos datos genómicos.

Algunos genomas pueden alcanzar tamaños de varias gigabases (Gb), por lo que es necesario indexar su secuencia antes de alinear. La indexación genera varios archivos FAI que contienen sencillamente un diccionario con las posiciones de cada segmento en unidades menores que permiten al mapeador encontrar con mayor facilidad el patrón de la lectura buscado.

3.3.2. Mapeadores de empalme

La mayoría de los mapeadores genómicos utilizan el algoritmo de transformación de Burrows-Wheeler (BWT) para almacenar los datos de secuencia en paquetes comprimidos definidos por índices y distribuidos de manera jerárquica. De esta manera, la búsqueda de millones de lecturas puede dirigirse de manera más eficiente.

Adicionalmente, los mapeadores de secuencia de ARN deben enfrentar el problema del alineamiento entre intrones. A diferencia de la secuencia genómica, el transcriptoma no contiene regiones intrónicas, por lo que una misma lectura puede cubrir dos o más exones separados por distancias, a veces, de hasta 10 000 nt. Los mapeadores de empalme están especialmente diseñados para realizar este tipo de alineamiento. Entre los principales mapeadores utilizados para datos de NGS se encuentran TopHat, STAR y Salmon.

Esta última herramienta es considerada un pseudomapeador debido a que simplifica el proceso de búsqueda de alineamiento a la secuencia exónica del transcriptoma, cuyo archivo de anotaciones es requerido como entrada. Esto ofrece ventajas de uso de memoria y rapidez, pero por otro lado limita el estudio de expresión a los transcritos anotados.

En cambio, los programas de TopHat y STAR pueden utilizarse de manera "no guiada" y además son capaces de generar su propio archivo de anotaciones. Ambas herramientas han sido ampliamente validadas por su alta eficiencia y fiabilidad. Durante las clases, utilizaremos la versión mejorada de TopHat que existe actualmente: Hisat2 (Pertea *et al.*, 2016), cuyo consumo de memoria es significativamente menor. Otra de las ventajas de Hisat2 es que sus parámetros son regulables, lo que permite adaptar el proceso a las características del genoma de referencia que usemos.

Ejemplo

Aumentar la permisividad del mapeo permite usar genomas de referencia más distantes filogenéticamente a las muestras secuenciadas. En cambio, disminuirla mejora la correcta identificación de genes homólogos o familias multigénicas cuyas secuencias son muy parecidas entre sí.

Enlace de interés

En el siguiente enlace podrás encontrar documentación e instrucciones para la instalación del mapeador de empalmes Hisat2.
http://daehwankimlab.github.io/hisat2/manual/

El archivo de salida del mapeador es un mapa de alineamiento en texto (SAM) cuya compresión (BAM) puede realizarse con SAMtools (Li *et al.*, 2009). Este kit de herramientas está especialmente diseñado para trabajar con este tipo de archivos y es increíblemente versátil. Los resultados del mapeo pueden visualizarse con un navegador genómico (ver sección 5.4.).

Las lecturas provenientes de la NGS de ARN deben ser mapeadas con herramientas capaces de procesar alineaciones interexónicas debido a que la secuencia del transcriptoma no presenta intrones, a diferencia de la genómica.

3.4. Recuento de alineamientos sobre exones

El siguiente paso en el flujo de trabajo es la cuantificación de las lecturas sobre las zonas exónicas de los genes anotados. Para ello existen multitud de herramientas, entre las cuales las más usadas son featureCounts y HTSeq-count. Básicamente, recogen el recuento de las lecturas alineadas sobre cada exón en un archivo de texto de salida (CSV), el cual servirá para la posterior evaluación de la expresión diferencial.

Estas herramientas tienen diferentes parámetros que deberemos ajustar a las características de nuestro experimento de secuenciación. Por ejemplo, si se trata de lecturas solas (*single reads*) o emparejadas (*paired-end*) y si el ensayo es específico de hebra.

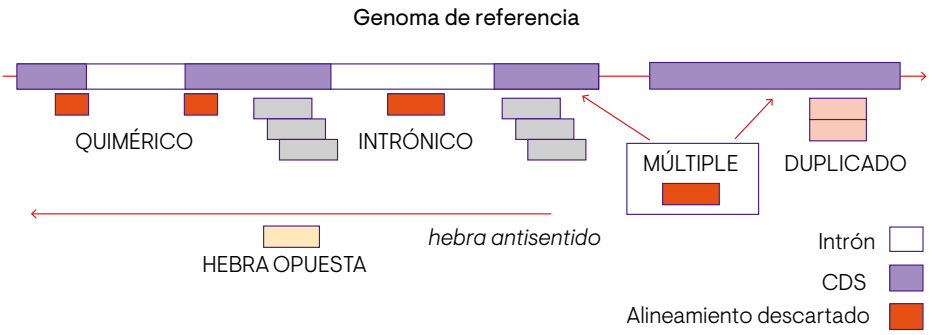
Enlace de interés

Puedes encontrar más información acerca de la instalación e instrucciones para utilizar la herramienta de cuantificación HTSeq-count en el siguiente enlace.
https://htseq.readthedocs.io/en/release_0.11.1/count.htmlttp://daehwankimlab.github.io

Por defecto, la mayoría de los programas diseñados para el análisis de la expresión diferencial descartan todo aquel alineamiento considerado "ambiguo" y que engloba los diferentes casos (Figura 10).

Figura 10

Representación de los diferentes tipos de alineamiento posibles producidos con lecturas cortas de NGS indicando en naraja los que se descartan durante la cuantificación



- **Alineamiento intrónico**

  Es aquel que sucede en los intrones y su presencia en organismos eucariotas puede deberse a la presencia de estadios inmaduros de transcritos o a la expresión de ARN no codificante (ncRNA).

- **Alineamiento en la hebra opuesta**

  Algunas lecturas pueden alinear con la cadena contraria al gen cuya expresión está anotada. Esta representación puede ocurrir debido a eventos de superposición génica o a la presencia de transcritos naturales antisentido (NAT) cuya expresión está vinculada a funciones regulatorias del gen que se transcribe en la hebra opuesta.

- **Alineamiento múltiple**

  Al contrario de lo que ocurre con los duplicados, la secuencia repetida en el alineamiento múltiple se encuentra en el genoma de referencia, dando lugar a dos o más alineamientos con diferente posición genómica (*locus*) de la misma lectura. Esto es debido, especialmente en organismos eucariotas superiores, a la naturaleza repetitiva del genoma, como ocurre en las regiones del centrómero y del telómero de los cromosomas. Generalmente, los genomas de referencia disponibles en las bases de datos suelen estar previamente enmascarados reemplazando estas secuencias de ADN de baja complejidad por Ns. Sin embargo, cuando las lecturas de NGS son cortas sigue existiendo un 15-30 % de alineamiento múltiple que corresponde con secuencias compartidas por dominios comunes, familias multigénicas o pseudogenes.

- Alineamiento solapante o quimérico

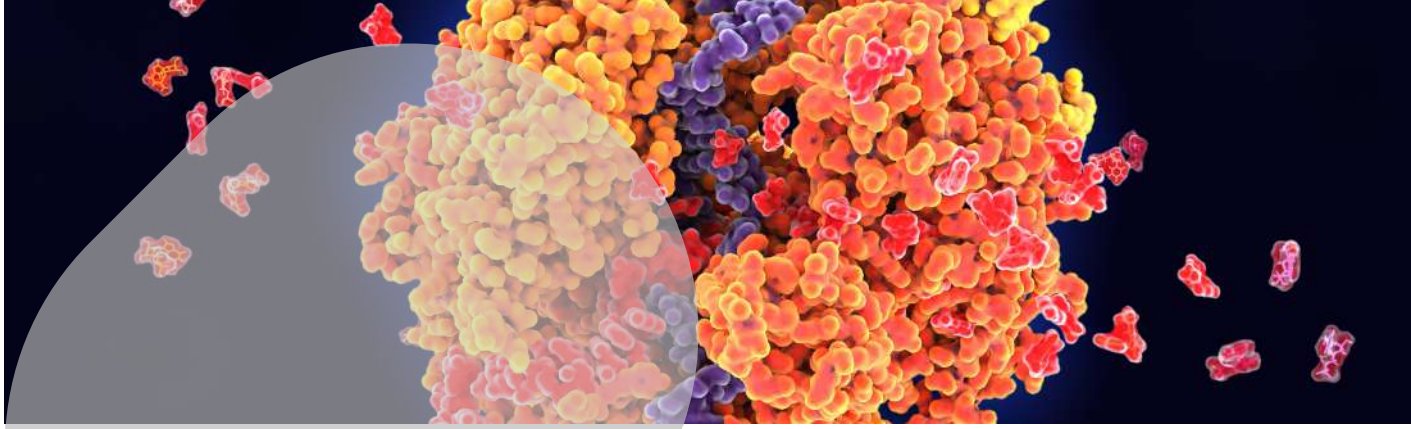
  Es aquel que ocurre cuando una misma lectura alinea parcialmente con dos regiones genómicas diferentes. En la NGS de ARN, este tipo de alineamiento es relativamente frecuente y corresponde con los límites exón-exón. Este tipo de alineamiento es tomado como una evidencia de los sitios de *splicing* por los mapeadores de empalme para la elaboración de anotaciones propias. Sin embargo, es descartado por los programas de recuento si no se indica.

Un último caso considerado, a veces, "ambiguo" es el alineamiento duplicado que ocurre cuando dos lecturas tienen exactamente la misma secuencia y alinean en la misma localización del genoma de referencia. Esto ha sido ampliamente debatido en los foros, y las publicaciones concluyen que la frecuencia de duplicados por PCR en NGS no es significativa siempre y cuando la diversidad de la librería es suficiente (Parekh *et al.*, 2016). Para comprobarlo, podemos evaluar los niveles de duplicación con la herramienta de control FastQC, preferiblemente tras la eliminación de adaptadores.

Enlace de interés

En el siguiente enlace puedes encontrar el post: "Should we remove duplicated reads in RNA-seq" de la comunidad Biostars, donde se discute este tema.
https://www.biostars.org/p/55648/

La secuenciación de lecturas cortas específica de hebra permite diferenciar la cadena de origen de la expresión génica. Para distinguirlas, deberemos indicar al programa de cuantificación de lecturas la modalidad de nuestra secuenciación.



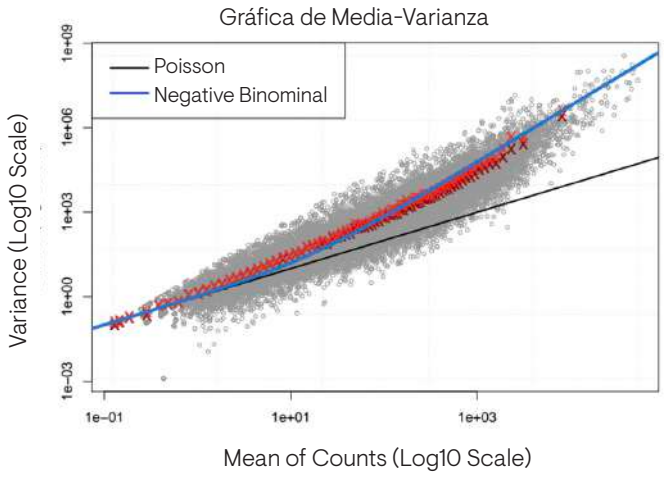
Capítulo 4

Análisis estadístico de la diferencia de expresión

Con el objetivo de identificar aquellos transcritos cuya expresión es diferente entre las muestras de interés necesitaremos realizar un estudio estadístico. Los datos de NGS, a diferencia de lo que ocurría con los datos de microarrays (distribución de Poisson), se caracterizan por una distribución binomial negativa (Figura 11).

Figura 11

Gráfica que representa la relación entre la media y la varianza de los datos de NGS cuya distribución se ajusta más a una función binomial negativa



El origen de esta distribución es que el rango de genes representado en la biblioteca de cDNA abarca toda la complejidad del transcriptoma, mientras que los microarrays solo representan los perfiles de aquellos genes predefinidos en la plataforma de hibridación. Esto significa que una gran parte de los transcritos secuenciados en la NGS no presentan diferencias en su nivel de expresión. Con esta premisa, los algoritmos paramétricos existentes hacen una estimación de la significancia dentro de la variabilidad observada. Entre los paquetes de evaluación estadística más utilizados se encuentran edgeR (Robinson *et al.*, 2010) y DESeq2 (Love *et al.*, 2014). Durante las clases seguiremos los pasos del análisis con las herramientas de edgeR para determinar cuáles son los genes diferencialmente expresados (DEG).

Enlace de interés

En el siguiente enlace puedes encontrar un manual detallado de las distintas funciones del paquete estadístico edgeR.
https://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

Para aquellos casos en los que no es posible asumir este comportamiento de los datos o el número de **réplicas biológicas** es bajo existen otras opciones menos heurísticas que han demostrado tener una baja tasa de **falsos positivos**, como SAMseq (Li y Tibshirani, 2013) o NOISeq (Tarazona *et al.*, 2011).

4.1. Normalización de los recuentos

Los recuentos de lecturas mapeadas de las bibliotecas de cDNA comparadas en el estudio de su expresión génica deben ser escalados previamente para corregir las posibles diferencias debidas a una profundidad de secuenciación irregular. El método más ampliamente utilizado para esta normalización por tamaño de biblioteca es la media recortada de los valores M (TMM) que utiliza una media ponderada de los valores de la expresión logarítmica (Robinson *et al.*, 2010).

Una segunda normalización enfrenta las diferencias de longitud de los transcritos que provocan que un transcrito largo tenga más probabilidad de ser secuenciado y, por lo tanto, sea más fácil de detectar como diferencialmente expresado. Esta normalización por longitud de transcrito se realiza durante la conversión a fragmentos (FPKM) o lecturas (RPKM) por kilobase de transcrito por millón de lecturas mapeadas, para parejas o lecturas solas, respectivamente.

Las funciones *calcNormFactors* y *rpkm* de la herramienta edgeR pueden usarse para completar estos dos pasos. La corrección para el tamaño de los transcritos necesitará un archivo con un listado de las longitudes de los diferentes transcritos en el parámetro "gene.length" de la función *rpkm*.

4.2. Eliminación de genes de baja expresión

El filtrado de genes de recuento bajo es un paso opcional en el análisis de datos de expresión de ARN. La idea es eliminar aquellos genes cuyo nivel de recuento es cercano a 0 (entre 0 y 10), puesto que pequeñas variaciones entre réplicas pueden ser consideradas erróneamente como un cambio de expresión. Eliminar esta fracción de genes en el límite de detección o dentro del ruido experimental puede mejorar la sensibilidad y la precisión de los DEG después del filtrado (Bourgon *et al.*, 2010).

En el lenguaje de programación R existen varias opciones para llevar a cabo este paso, aunque algunos paquetes han sido especialmente diseñados para ello, como HTSFilter, cuya implementación está basada en el índice de similitud de Jaccard global. Identifica, además, aquellos genes cuya expresión es constante en las diferentes condiciones experimentales.

4.3. Estudio de la variabilidad entre muestras

4.3.1. Variabilidad técnica

Aunque la variabilidad técnica de la NGS se acerca a cero (especialmente cuando las muestras a comparar se preparan en el mismo carril de la celda de flujo), existen otras fuentes relacionadas con el muestreo y la extracción de ARN que introducen otras variables independientes diferentes a la "variable dependiente" que se analiza en el DGE (Liu *et al.*, 2014).

Tal y como hemos visto, la normalización de las bibliotecas y el filtraje de los alineamientos durante la cuantificación tratan de minimizar esta carga. Sin embargo, la variabilidad técnica relacionada con el muestreo puede ser evitada con un buen diseño experimental que implica la recogida del mayor número posible de réplicas biológicas de una manera aleatoria. Un caso especial es el reloj circadiano, cuya influencia es tan generalizada sobre los genes que se recomienda seguir un muestreo estrictamente homogéneo para no dejar que esta otra variable se confunda con la biológica (a no ser que sea el motivo de estudio).

En general, cuando se usan herramientas con enfoque paramétrico como edgeR para el análisis estadístico se recomienda un mínimo de 3 réplicas biológicas por muestra para evitar desviaciones del modelo que utilizan.

Aumentar el número de réplicas biológicas en nuestro experimento de NGS mejorará la estimación de los niveles medios de expresión y podrá aumentar el número de DEG identificados.

4.3.2. Variabilidad biológica

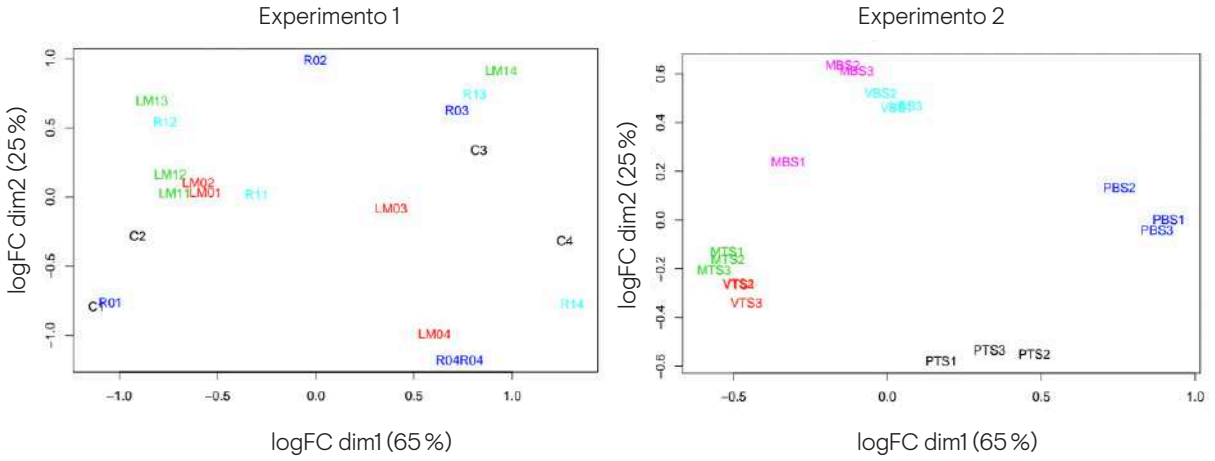
La variabilidad biológica "dependiente" de la condición impuesta en el experimento (por ejemplo, sequía) es lo que esperamos encontrar al realizar el estudio de DGE. Para extraerla, deberemos poder distinguirla de la variabilidad biológica "no dependiente" que las muestras presenten entre sí, de ahí la necesidad de usar las réplicas biológicas.

En primer lugar, se recomienda hacer una estimación de la dispersión general que existe entre todas las muestras y/o réplicas del conjunto de datos. Esta estimación, llamada análisis de los componentes principales (PCA), nos permitirá conocer las relaciones de distancia entre las muestras señalando cuáles son las mayores fuentes de variación en nuestros datos. El escalado multidimensional (MDS) es un tipo de PCA aplicado al estudio de dispersión en datos de genes como los de NGS.

La función *plotMDS* (edgeR) sirve para hacer estimación de dispersión de genes con MDS de varias maneras. La que ejecuta automáticamente calcula las distancias entre las diferentes muestras basándose en los 500 genes (modificable) más dinámicos. Para ello, recoge provisionalmente el nivel de cambio de cada gen en cada una de las comparaciones posibles entre muestras y con estos datos representa un gráfico de logFC frente a logFC en los ejes (Figura 12).

Figura 12

Gráfico de MDS de dos experimentos distintos de NGS cuya variabilidad biológica se muestra independiente (izquierda) y dependiente (derecha) de los tipos de muestra recogidos en el experimento



Esta estimación nos permitirá reconocer una distancia o variabilidad biológica suficiente para extraer DEG del experimento. Si la dispersión es igual para todas las muestras, es decir, tanto las muestras control como las experimento quedan a una distancia parecida y no forman grupos (Figura 12, izquierda), podemos concluir que al menos entre este grupo de genes no se encuentran DEG. Además, podremos comprobar si existe alguna muestra muy desviada del comportamiento de las otras réplicas (*outliers*) en alguno de los grupos y evaluar cuál es la razón.

El anterior análisis de componentes principales asume que todos los genes tienen una misma relación media-varianza (datos univariados). Sin embargo, los genes no tienen este comportamiento regular en la mayoría de experimentos debido a la presencia de variables independientes (datos multivariados).

Entonces, el cálculo de dispersión de los genes debe tener un tratamiento específico, cada gen debe analizarse por separado, para lo cual se acude al ajuste a modelos lineales generalizados (GLM). Aunque no es perfecta, esta aproximación permite hacer una estimación válida de la dispersión de los genes de manera paramétrica. Existen otros enfoques no-paramétricos, pero no son objeto de este capítulo.

Tras calcular la dispersión con *estimateDisp*, aplicaremos este enfoque estadístico con la última de las versiones metodológicas implementadas o *quasi-likelihood* en la función *glmQLFit* (edgeR).

4.4. Comparativas entre grupos

Una vez recogido el comportamiento de cada gen frente a nuestra variable dependiente mediante la regresión de modelos lineales generalizados (GLM), queremos comparar las medias de expresión de ambos grupos y saber si la diferencia es suficientemente significativa para considerarlos como DEG.

Para ello, deberemos haber indicado previamente en la función *glmQLFit* cuáles son los grupos de réplicas en el parámetro "group" y el diseño experimental de la matriz de datos en "design". El diseño experimental no es más que una tabla donde se indica la identidad de cada muestra como réplica y su pertenencia al grupo control o experimento en cada comparativa.

Durante las clases aprenderemos a usar la modalidad comparativa más sencilla "uno contra uno", que corresponde con un análisis de varianza clásico tipo ANOVA (prueba ANOVA). Sin embargo, la herramienta que usaremos es versátil como para poder aplicar la prueba en más de una comparativa (*multi-way* ANOVA-*like*) o para estimar más de una variable dependiente como ocurre en los experimentos de tipo temporal (*time-course*).

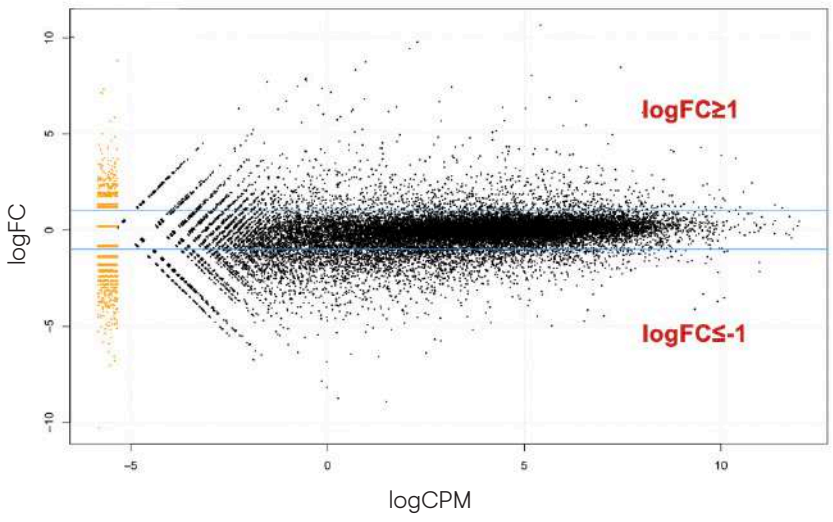
4.4.1. Nivel de cambio

El nivel de cambio (FC) es la magnitud de la diferencia de expresión entre las muestras comparadas. Se calcula sencillamente como la relación entre las medias de las réplicas de un grupo frente a las del otro: FC = media B / media A (en general, A es el grupo control). Por sus características de distribución, el nivel de cambio en los datos de NGS se expresa en la escala logarítmica en base 2 (log2FC). Esta escala tiene la ventaja de ser simétrica, donde 0 es el valor del "no cambio" y, a ambos lados, cambios positivos y negativos tienen el mismo valor absoluto.

El gráfico de Smear que dibuja la función *PlotSmear* (edgeR) a partir de los datos de recuentos en CPM es un tipo de gráfico de diferencia de medias (gráfico MA). Con él podemos representar el nivel de cambio (logFC) general, donde cada punto es un gen, entre las muestras comparadas (Figura 13). Este nivel de cambio se dibuja con respecto a la cantidad de recuentos normalizados (logCPM), lo que nos permite apreciar una nube de puntos cada vez menos densa cuando la media de expresión es alta (hacia la derecha).

Figura 13

Gráfico de Smear de datos de NGS que representa los niveles de cambio (logFC) entre dos muestras respecto al número de recuentos (logCPM). Las líneas azules marcan los límites de corte en el doble (logFC≥1) o la mitad (logFC≤1) expresión



En general, se usan los **límites de corte** (*cutoff point*) del doble o la mitad en la selección de DEG, pero esto es algo que puede ser ajustado a la biología del experimento. Este filtraje por encima y por debajo de la media recogerá aquellos "genes sobreexpresados" e "infraexpresados" en la muestra experimento con respecto al control. Cuando el experimento es un tratamiento que genera una respuesta de señalización como la adición de hormonas o las condiciones de estrés, suelen utilizarse los términos "inducción" y "represión", respectivamente.

4.4.2. Pruebas de significancia

La prueba de significancia que se aplica en una prueba ANOVA, la prueba p (*p-test*), no es más que un estudio de probabilidad de que nuestra hipótesis (que el gen cambia) sea errónea. Para calcularla, se debe comparar la desviación normal de la media de expresión con respecto a la del gen cuestionado. Además, se ha de tener en cuenta que la probabilidad de error es proporcional al número de genes testeados.

La prueba de significancia desarrollada en la función *glmQLFTest* (edgeR) es análoga a la ANOVA solo que un poco más moderada en el cálculo de la varianza. Antes de realizar el test necesitamos estudiar la desviación normal con *estimateDisp* (edgeR). Después, *glmQLFTest* computará niveles de cambio y pruebas de significancia entre las muestras que indiquemos en el parámetro "contrast". (Es posible cambiar a comparativas múltiples modificando el contraste en el diseño. Para poder hacerlo, deberás usar "~0 + grupos" en lugar de "~grupos" en el parámetro "design").

Pero una interpretación más estricta del *p-value* es considerada obligatoria en el análisis de datos de RNA-seq debido a que el elevado número de comparaciones (tantas como genes anotados) aumenta la probabilidad de falsos positivos. Para ello, se ha establecido el uso de un valor p ajustado (p.adj) mediante la corrección de Benjamini-Hochberg o también llamado FDR (*false discovery rate*). En general, se exige un FDR mínimo de 0.05, es decir, se acepta un 5 % de probabilidades de obtener un resultado significativo por azar después de hacer esta corrección. Sin embargo, esta condición es ajustable y se recomienda ser más estricto cuando se dispone de pocas réplicas biológicas.

Enlace de interés

En el siguiente enlace puedes encontrar una explicación más detallada del FDR.
https://www.youtube.com/watch?app=desktop&v=K8LQSvtjcEo&ab_channel=StatQuestwithJoshStarmer

En el paquete de edgeR, la función topTags computará esta corrección sobre el valor p y añadirá una columna a la tabla con los resultados. Además, pueden seleccionarse los genes (n) con mejores puntuaciones de logFC con el parámetro "sort.by" y, por ejemplo, dentro de una significancia segura con el parámetro "p.value".

Ejemplo

Salida de topTags (edgeR):

	logFC	logCPM	F	PValue	FDR
Zm00001d014840	9.98	2.62	70.37	4.55E-08	3.52E-07
Zm00001d036366	3.41	2.54	72.01	3.78E-08	2.97E-07
Zm00001d036370	3.85	4.33	163.57	3.05E-11	4.71E-10
Zm00001d026422	-4.22	-1.91	60.81	1.44E-07	1.00E-06
Zm00001d044585	-4.35	-3.20	62.02	1.21E-07	8.55E-07
Zm00001d050082	-11.13	1.87	303.58	1.17E-10	1.57E-09

En el análisis estadístico, el nivel de significancia de la expresión de un gen depende de la cantidad de genes totales observados.

4.4.5. Simulación

Si todavía no tienes datos de NGS propios puedes practicar los pasos descritos en este capítulo partiendo de una simulación y usando la consola de Rstudio.

Ejemplo

```
# Instala los paquetes necesarios para este capítulo

BiocManager::install(c("stats", "edgeR"))

# Simula unos datos de NGS siguiendo una distribución binomial negativa (NB) con una dispersión de 0.2 y un valor promedio de 1000 unidades de expresión.

library(stats)
genes ← 5000
muestras ← 8
sim ← matrix(rnbinom(genes*muestras, size=1/0.2, mu=1000), genes, muestras)
rownames(sim) ← paste("gen",1: genes, sep=".")
grupos ← gl(2, 4, labels=c("Control","Experimento"))

# Dibuja la distribución en un histograma
hist(sim, breaks = 100)

# Introduce un cambio de expresión de 500 unidades entre Experimento y Control.
sim[1:500,grupos=="Experimento"] ← sim[1:500,grupos=="Control"] + 2500

# Observa el cambio en otro histograma
hist(sim, breaks = 100)

# Normaliza el tamaño de biblioteca
library(edgeR)
y ← DGEList(sim,group=grupos)
y ← calcNormFactors(y)

# Elimina los genes con baja expresión
library(HTSFilter)
y ← HTSFilter(y)

# Visualiza variabilidad entre muestras
col ← as.numeric(grupos)
plotMDS(y, top=500, labels=grupos)
```

>>>

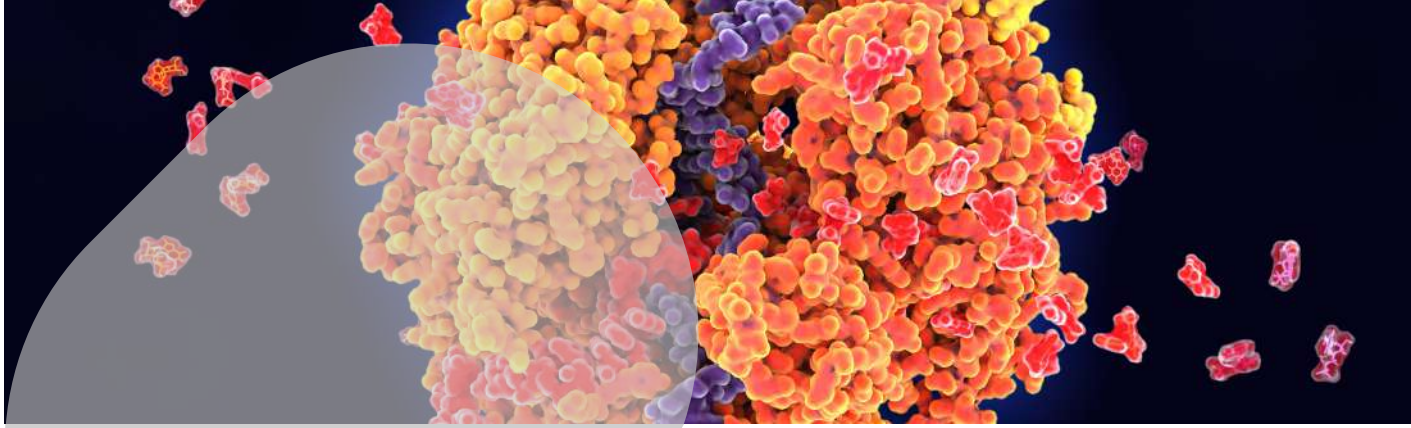
>>>

```
# Estima varianza
diseño ← model.matrix(~0 + grupos)
varianza ← estimateDisp(y, diseño)

# Aplica modelos lineales generalizados
glm ← glmQLFit(varianza, diseño)

# Calcula genes significativos
contraste ← makeContrasts("gruposExperimento-gruposControl", levels = diseño)
test ← glmQLFTest(glm, contrast = contraste)

# Calcula el FDR
test2 ← topTags(test, n=5000)
deg ← topTags(test, p.value = 0.05)
```



Capítulo 5

Exploración y visualización de resultados

Conseguir identificar cuál es el grupo de genes cuya expresión cambia entre las muestras comparadas puede significar llegar al final de nuestro recorrido en el análisis, habiendo respondido a la pregunta inicial para la cual diseñamos el experimento de NGS. Sin embargo, la amplitud de los datos de transcriptómica y la complejidad de la misma expresión génica (especialmente en organismos eucariotas) dificulta muchas veces la extracción de conclusiones. En este paso adicional para "dar sentido" a los resultados priorizaremos aquellos genes directamente vinculados a la variable de nuestro experimento, visualizaremos posibles nodos de regulación y podremos revisar si existen cambios en la secuencia de algunos genes candidato.

5.1. Visualización por significancia y nivel de cambio: gráfico Volcano

El gráfico Volcano permite "separar el grano de la paja" cuando tenemos listas muy largas de genes significativos. Es, sencillamente, un gráfico de dos ejes donde se representan la magnitud del cambio (log2FC) frente a la significación estadística (FDR), que se representa en -log10, lo que le confiere su característica forma de "U". Para ello, se usa la tabla de resultados de los genes DE y se visualizan los rangos entre los cuales se mueven dichas variables para estimar posibles puntos de corte.

Además, los genes más significativos, es decir, aquellos con menor FDR y mayor nivel de cambio, pueden identificarse mediante etiquetas. La importancia de estos genes puede estar relacionada, por ejemplo, con núcleos de regulación alrededor de determinados factores de transcripción.

Existen varias funciones en R para representarlo, como ggplot (paquete ggplot2), VolcanoPlot (paquete limma), o enhancedVolcano, que genera gráficos con muchas posibilidades de configuración para usar en publicaciones.

Enlace de interés

En el siguiente enlace puedes encontrar un manual para elaborar este tipo de gráficos en un formato de publicación.
https://bioconductor.org/packages/release/bioc/vignettes/EnhancedVolcano/inst/doc/EnhancedVolcano.html

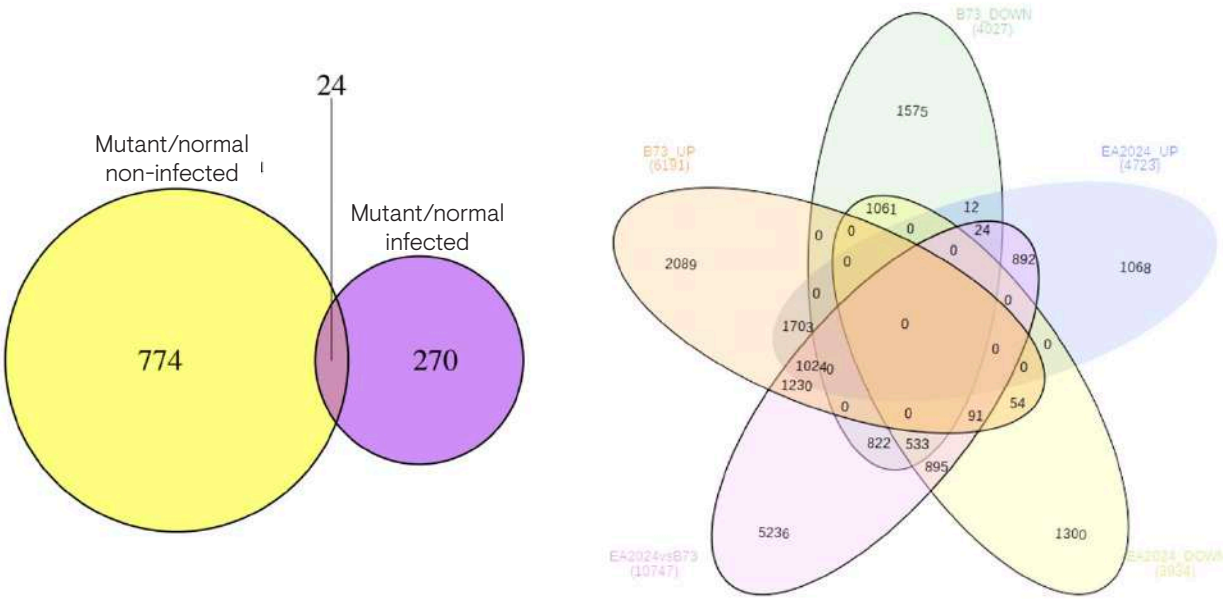
5.2. Comparación de conjuntos de genes: diagrama de Venn

El diagrama de Venn recoge la coincidencia de identificadores génicos (ID) entre diferentes grupos de genes. Este tipo de gráficos de conjuntos es muy útil para estudiar relaciones entre los DEG, por ejemplo, entre diferentes comparativas para identificar parcelas comunes de regulación.

También podemos buscar relaciones entre DEG sobreexpresados e infraexpresados. Si el mismo gen aparece inducido y reprimido en dos estados de una condición (p. ej., sequía/irrigación) tendrá muchas posibilidades de pertenecer a su vía reguladora.

Figura 14

Dos ejemplos de diagramas para representar grupos de conjuntos de DEG



Existen diversas funciones en R para dibujarlos como Venn (limma), sin embargo, las opciones que ofrecen algunas herramientas web son muy fáciles de usar, solo hay que subir el archivo de cada lista de identificadores en un formato de texto.

Enlace de interés

En el siguiente enlace puedes encontrar la herramienta web que proporciona el grupo de Bioinformática y Genómica Evolutiva de la Universidad de Gante:
hhttp://bioinformatics.psb.ugent.be/webtools/Venn/

Algunas ofrecen incluso la posibilidad de tener un diagramasde Venn no simétrico, es decir, cuyo tamaño del conjunto represente proporcionalmente el número de genes coincidentes. La gráfica de salida puede guardarse en formato PNG, aunque si la elegimos con un formato SVG tendremos la posibilidad de editar la imagen a nuestro gusto con un *software* compatible como InkScape.

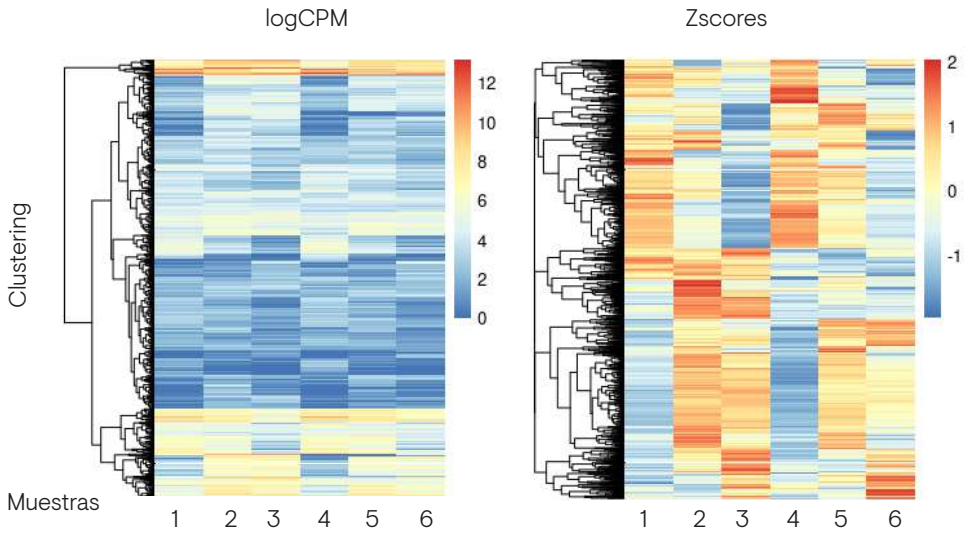
5.3. Agrupamiento de perfiles de expresión: gráfico *heatmap*

El gráfico *heatmap* es una representación por color de los valores en un conjunto de datos. Es extremadamente útil para la exploración de datos de DEG, especialmente por su implementación de funciones de agrupación o *clustering* capaz de abordar grandes cantidades de datos (10 000 genes o más). Este *clustering* es el ordenamiento por patrones de comportamiento similares y tiene grandes implicaciones en la interpretación de los resultados puesto que permite detectar grupos o clústeres de genes con una regulación común. La presencia de patrones podrá ser respaldada, por ejemplo, por un estudio de enriquecimiento funcional posterior.

Si el método de agrupamiento no es guiado, hablamos de agrupación jerárquica (*hierarchical clustering*), cuyo proceso primero calcula las distancias entre cada par de genes con la función *dist* (stats) y, en segundo lugar, aglomera con diferentes opciones metodológicas. Siempre que la lista de DEG esté libre de ruido, podremos escalar los datos con la función *scale* (stats) para convertir la media en 0 y cada dato en su desviación estándar alrededor de dicha media. Si la aplicamos tanto en columnas (genes) como en filas (muestras), transformaremos las unidades de expresión a **Z-score**, lo que eliminará la diferencia de escala en la expresión global y hará destacar las distancias entre muestras (Figura 15).

Figura 15

Gráfico heatmap de DEG realizado con pheatmap mostrando su representación en logCPM (izquierda) y su transformación a Z-score (derecha)



Durante las clases trabajaremos con *pheatmap*, cuya métrica de distancia "euclidiana" y método de agrupación "completo" son los predeterminados y suelen funcionar bien en este tipo de datos. Por otro lado, podemos definir previamente el número de grupos con el parámetro "kmeans_k". Esto es aconsejable si el número de filas es tan grande que R ya no puede manejar su agrupación jerárquica (mayor de 1000). Por ejemplo, la transformación en Z-scores seguida de una agrupación en 2 kmeans será muy útil para separar en los estados de sobreexpresión e infraexpresión del conjunto de DEG.

Enlace de interés

En el siguiente manual puedes encontrar diferentes métodos de agrupación usados en R para exploración de la expresión génica (autores: Hugo Tavares, Georg Zeller):
https://tavareshugo.github.io/data-carpentry-rnaseq/04b_rnaseq_clustering.html

5.4. Enriquecimiento funcional: términos GO

La base de datos de Gene Ontology (GO) mantiene y desarrolla un sistema universal de clasificación de los productos génicos. Para ello, recoge tres tipos de características: la función molecular (F), el proceso biológico al que pertenece (P) y su localización celular (C). Para ello, GO utiliza unos indicadores con código de siete cifras (GO:XXXXXXX) que se ordena jerárquicamente y que corresponde con un nombre, por ejemplo, el término GO:0005524 es la "unión de ATP" en cualquier organismo que tenga esta función molecular.

Cada anotación incluye, además, un código de evidencia que nos indica cuál es la fuente en la que se apoya: experimental, filogenética, computacional o automática. Las anotaciones respaldadas por anotación automática (IEA) se encuentran generalmente disponibles en todos los organismos publicados y recogen información de diferentes fuentes: (1) la proteína codificada se identifica mediante homología con las firmas de InterPro; (2) se le asocia una actividad enzimática (números de la comisión de enzimas: EC) y una predicción de ubicación subcelular, ambas desde UniProt; (3) se añade la **ortología** con los genes de otros organismos publicados en Ensembl.

Dada una lista de genes DE, conocer su representación funcional puede ayudarnos a comprender mejor los procesos biológicos subyacentes a la condición del análisis. Para ello, se compara estadísticamente la frecuencia de cada término GO dentro del conjunto de genes de entrada (lista de genes DE) con la frecuencia en el conjunto del transcriptoma. Puedes hacer una prueba con la herramienta web alojada en el portal GO.

Enlace de interés

El siguiente enlace corresponde con la base de datos Gene Ontology (GO):
http://geneontology.org/

5.5. Navegadores genómicos

Dada la popularidad de la secuenciación masiva a escala genómica, se han hecho necesarias nuevas herramientas destinadas a hacer el seguimiento de las secuencias resultantes por parte de usuarios con mayor o menor preparación en bioinformática. Estos navegadores o visualizadores genómicos permiten integrar grandes conjuntos de datos y metadatos en diferentes pistas, explorar rápidamente diferentes posiciones del genoma y enfocar en regiones de interés.

Ya existen varias opciones de distribución libre como aplicación para el escritorio o como aplicación web, y, en algunos casos, como un componente de JavaScript que se puede incrustar en páginas web. Son los casos de las bases de datos de organismos ampliamente estudiados, que ya implementan esta aplicación para visualizar los datos que recopilan. En general, todos requieren un genoma anotado.

Los diferentes conjuntos de datos que pueden ser cargados a la plataforma y su formato de archivo correspondiente son:

- Secuencias genómicas, en archivo FASTA. Por ejemplo, el genoma del organismo de referencia.
- Alineamientos de las lecturas, en archivo BAM o SAM.
- Anotaciones transcriptómicas, en archivo GFT o GFF. Incluyen formas de *splicing* o secuencias de RNA no codificante.
- Resultados de variación en secuencia, en archivo VCF. Por ejemplo: SNPs o INDELs.

Dos de los navegadores de uso local más populares a día de hoy son: IGV (Robinson *et al*., 2011) y SeqMonk. A diferencia de IGV, SeqMonk visualiza los transcritos *antisense* de manera clara, con dos colores, por lo que se recomienda más para este tipo de estudios. Por contra, SeqMonk no registra la secuencia subyacente a cada lectura mapeada, por lo que no es útil para la identificación de SNP.

Enlace de interés

Este es el enlace de descarga de la página web del Babraham Institute (Cambridge) cuyos bioinformáticos desarrollaron y mantienen SeqMonk y un gran número de otras aplicaciones.
https://www.bioinformatics.babraham.ac.uk/projects/download.html#seqmonk

Durante las clases de esta asignatura utilizaremos la versión para escritorio de IGV que puede ser utilizada bajo diferentes sistemas operativos.

Enlace de interés

Este es el enlace de descarga de la página web del Broad Institute (California) que distribuye y mantiene el *software* de IGV.
https://software.broadinstitute.org/software/igv/download

5.6. Simulación en R

Si todavía no tienes datos de NGS propios, puedes practicar los pasos descritos en este capítulo partiendo de una simulación y usando la consola de Rstudio.

Ejemplo

```
# Dibuja un gráfico Volcano con la tabla de resultados de DEG que generaste anteriormente.
library(ggplot)
ggplot(test2$table,aes(x=logFC,y=-log10(FDR)))+geom_point()
```

>>>

```
# Dibuja un gráfico heatmap
library(pheatmap)
pheatmap(y$counts, cluster_col = FALSE, show_rownames = F)
pheatmap(y$counts, cluster_col = F, show_rownames = F, clustering_distance_rows = "correlation")

# Selecciona solo aquellos genes con FDR ≤ 0.05
test2_deg ← y$counts[unlist(lapply(rownames(y$counts), function(x) grep(x, rownames(deg), fixed = TRUE))),]
pheatmap(test_deg, cluster_col = FALSE)

# Escala a Z-score
zscore ← t(scale(t(test_deg))
pheatmap(zscore, cluster_col = FALSE)
```

Glosario

A

Adenina

Affymetrix

Empresa estadounidense con sede en Santa Clara (California) especializada en el diseño de micromatrices de ADN.

ANOVA

Por sus siglas en inglés *analysis of variance*.

API

Del acrónimo en inglés *application programming interface*. En la distribución de archivos de secuenciación es el recurso abierto para su descarga.

BWT

De las siglas en inglés *Burrows–Wheeler transform*.

C

Citosina

Características de secuencias (*features*)

En el modelo de un gen, se refiere a las subunidades de secuencias de tipo exón, intrón, 3'UTR, 5'UTR, etc.

Celda de flujo

Plataforma o portaobjetos horadada en nanoporos que las tecnologías de secuenciación utilizan como soporte sólido para las diferentes reacciones de polimerización que se dan durante el proceso.

cDNA

Del acrónimo en inglés *complementary DNA*. El ADN complementario es una molécula de ADN de doble cadena, en la que una de sus hebras constituye una secuencia totalmente complementaria al ARN mensajero a partir del cual se ha sintetizado.

Cola de poli(A)

Modificación por adición de un polímero de adenosín monofosfatos que se da en eucariotas durante el proceso de maduración del ARN mensajero.

CSV

De las siglas en inglés *comma separated values*. Archivo de texto con separación por comas de las columnas.

CPM

De las siglas en inglés *count per million*. Son recuentos normalizados por el número de fragmentos secuenciados multiplicado por un millón.

Desviación estándar

Raíz de la varianza.

DEG

De las siglas en inglés *differentially expressed genes*.

DGE

Del acrónimo en inglés *differential gene expression*.

dNTP

Del acrónimo en inglés *deoxynucleotide triphosphate*.

Duplicación

En el contexto de NGS, la duplicación es la presencia de lecturas exactamente iguales en la misma muestra.

edgeR

Paquete de Bioconductor diseñado para el análisis de la expresión diferencial de RNA-seq.

Empalme alternativo

Traducido del inglés *alternative splicing*. Modificación postranscripcional del mRNA que genera diferentes formas de transcrito y proteína con funciones, a veces, opuestas.

EST

Del acrónimo en inglés *expressed sequence tags*.

Exoma

Fracción de genoma que codifica para la producción de proteínas.

Expresión génica

Procesos de transcripción a partir de regiones genómicas determinadas que la célula realiza para producir transcritos o proteínas que permitan su funcionamiento.

Falso positivo

Error de tipo I que sucede cuando se rechaza la hipótesis nula siendo verdadera. O, dicho de otra forma, se correlacionan dos observaciones sin tener relación alguna.

Familia multigénica

Duplicaciones de un gen evolutivamente original que se mantienen funcionales y que con el tiempo de evolución pueden desarrollar especificidad de función. Su ocurrencia en genomas de organismos superiores es generalizada, por ejemplo, el caso de las inmunoglobulinas de mamíferos.

Flujo de trabajo

Pasos de ejecución en análisis computacional para el cálculo, la transformación y la integración de un grupo específico de datos.

FC

De las siglas en inglés *fold change*. En DGE, es el nivel de cambio o abundancia relativa de un transcrito entre dos muestras comparadas.

FPKM

De las siglas en inglés *fragments per kilobase per million mapped reads*. Unidades de expresión génica normalizadas por tamaño de biblioteca y por tamaño de gen para lecturas emparejadas.

Fragmentación

Técnica de ruptura de ADN en segmentos mediante hidrólisis química, nebulización, sonicación o transcripción inversa con nucleótidos de terminación.

Fragmento

Segmento de ARN o ADN que se genera por fragmentación durante la fase de preparación de la biblioteca en la secuenciación de lectura corta.

G

Guanina

Gen candidato

Aquel gen que tiene una localización cromosómica asociada a una enfermedad concreta u otro fenotipo.

Genoma de referencia

Secuencia de ADN representativa de un organismo generada por secuenciación mediante el ensamblaje de *scaffolds* hasta un nivel cromosómico.

GitHub

GitHub es una plataforma que aloja proyectos para la creación de código en grupo utilizando el sistema de control de versiones Git.

GLM

De las siglas en inglés *generalized lineal models*.

Hibridación

Proceso por el cual se combinan dos cadenas de ácidos nucleicos antiparalelas y con secuencias de bases complementarias en una única molécula de doble cadena.

Hisat2

Mapeador de uniones de empalme para lecturas de RNA-seq de lectura corta basado en TopHat y con mejorado rendimiento.

HTS

Del acrónimo en inglés *high-throughput sequencing*.

Illumina

Compañía estadounidense que desarrolla y comercializa sistemas de análisis genético y genómico, entre ellos, la popular secuenciación de lecturas corta que podemos encontrar entre los servicios de secuenciación más habituales hoy en día.

Inserto

ADN de doble cadena que se genera durante la preparación de la biblioteca a partir de los fragmentos de ácidos nucleicos y cuya secuencia es ligada entre dos adaptadores que identifican la muestra y permiten su secuenciación por síntesis.

Isoforma

Cada una de las distintas formas de un transcrito o proteína generados a partir del mismo gen a través del proceso del empalme alternativo o la maduración diferencial.

Lectura

Secuencia nucleotídica de una molécula real de DNA o RNA inferida a través de su secuenciación.

Mapeador

Herramienta bioinformática que compara y alinea una secuencia corta sobre una secuencia larga.

MDS

De las siglas en inglés *multidimensional scaling*.

mRNA

Del acrónimo en inglés *messenger ribonucleic acid*. Molécula producida por la enzima polimerasa de ARN de la célula para expresión de los genes.

mRNA maduro

Transcrito de ARN eucariota procesado mediante el empalmado de sus exones y la adición de una cola de poli(A) para ser reconocido por la maquinaria celular de traducción a proteína.

NGS

Del acrónimo en inglés *next generation sequencing*.

nt

Abreviado de "nucleótido".

Nucleótido de terminación reversible

Nucleótido modificado con tecnología de Illumina cuyo grupo 3'-OH está bloqueado para impedir la incorporación del siguiente nucleótido durante la actividad polimerasa y cuya base une un fluoróforo.

Organismo modelo

Especie usada ampliamente por la comunidad científica para estudiar procesos biológicos específicos por sus características de estabilidad genética, reproducibilidad o interés en un campo.

pb

Abreviado de "pares de bases".

PCA

De las siglas en inglés *principal component analysis*.

PCR

Del acrónimo en inglés *polymerase chain reaction*. Reacción de polimerización de ADN de doble cadena a partir de ADN monocatenario usando como sustrato dNTP y un cebador para la unión a la cadena. Fue

descubierta en las enzimas polimerasas de *Thermus aquaticus* y es actualmente extensamente utilizada en transcriptómica.

Pseudogen

Duplicación de un gen cuya secuencia carece de intrones y de otras secuencias de ADN esenciales para su función. Los pseudogenes, aunque son genéticamente similares al gen funcional original, no se expresan y frecuentemente acumulan numerosas mutaciones.

p-value

En la prueba de hipótesis nula, es la probabilidad de tener el resultado obtenido por azar.

Rango dinámico

Escala en el que una tecnología es capaz de detectar la presencia/ausencia de algo.

Réplica biológica

En bioestadística, medición o dato recogido paralelamente a otro de diferente individuo o identidad biológica para eliminar la variación biológica aleatoria.

RIN

Del acrónimo en inglés *RNA integrity number*. Rango de calidad usado para la estimación de la integridad del ARN por fluorometría (va del 1 al 10 siendo el máximo 10).

RNA-seq

Del acrónimo en inglés *ribonucleic acid sequencing*.

rRNA

Del acrónimo en inglés *ribosomal ribonucleic acid*.

Roche

Empresa de la industria farmacéutica con sede en Basilea y París que también distribuye y desarrolla productos para la investigación como enzimas y sistemas de microarrays o secuenciación.

RPKM

De las siglas en inglés *reads per kilobase per million mapped reads*. Unidades de expresión génica normalizadas por tamaño de biblioteca y por tamaño de gen para lecturas no emparejadas.

RT

Del acrónimo en inglés *reverse transcription*. Reacción de polimerización de ADN desde ARN usando como sustrato dNTP y un cebador para la unión a la cadena. Fue descubierta en las enzimas transcriptasa inversa de los virus de ARN y actualmente se utiliza extensamente en transcriptómica.

SBS

De las siglas en inglés *sequencing by synthesis*.

ST

Del acrónimo en inglés *spatial transcriptomics*. Tecnología surgida en los 2010 para la caracterización de la expresión génica de manera tridimensional en un tejido o grupo de células.

SVG

De las siglas en inglés *scalable vector graphics*.

Superposición génica

Expresión génica que se da en la secuencia que comparten ambas hebras de ADN generando dos productos génicos diferentes.

T

Timina

TopHat

Mapeador de uniones de empalme para lecturas de RNA-seq de lectura corta con rendimiento ultraalto.

Transcripción

En general, proceso celular por el cual se generan copias en ARN de un gen o producto génico a partir de la información del ADN.

Transcriptoma

Grupo de moléculas de ARN o transcritos generados por la célula a partir de su secuencia de ADN las que responde a las diferentes condiciones externas o etapas de desarrollo.

Transcriptómica

Disciplina surgida de la biología molecular que estudia el dinamismo de los transcritos en su conjunto, comparando sus diferentes estados entre muestras mediante el análisis de sus perfiles de expresión.

TPM

De las siglas en inglés *transcripts per million*. Unidades de expresión génica normalizadas por tamaño de gen y por tamaño de biblioteca.

TMM

De las siglas en inglés *trimmed mean of M values*.

Varianza

Medida de dispersión o variabilidad que representa la desviación de la media en un conjunto de datos. Calculada como la suma de las distancias a la media de cada observación dividido entre el total de observaciones.

Variabilidad biológica

En DGE, es la fluctuación de expresión entre muestras individuales alrededor de la expresión media.

Enlaces de interés

Bioconductor

Bioconductor es un proyecto de código abierto para el análisis de datos en genómica con el objetivo de desarrollar e integrar *software*. Está basado en el lenguaje de programación R.

http://www.bioconductor.org/

BioStar

Comunidad de bioinformática, genómica computacional y análisis de datos biológicos con un foro para consultas donde un elevado número de expertos responde a las dudas de usuarios principiantes y se comparte información.

https://www.biostars.org/

Ensembl

Ensembl es un proyecto de investigación bioinformática que trata de desarrollar un sistema de *software* que produzca y mantenga anotaciones automáticas en los genomas eucariotas seleccionados.

http://www.ensembl.org/index.html

The Global Alliance for Genomics and Health (GA4GH)

Comunidad internacional surgida para crear marcos y estándares que permitan el intercambio responsable, voluntario y seguro de datos genómicos y relacionados con la salud.

https://www.ga4gh.org/cram/

GeneOntology

Portal del consorcio GO (Estados Unidos) que mantiene anotaciones funcionales con identificadores propios de manera universal. Esta base de datos integra las anotaciones de diferentes fuentes y establece relaciones de ortología funcional proporcionando descripciones comparables filogenéticamente.

http://geneontology.org/

GitLab

GitLab es una herramienta web de control de versiones y desarrollo de *software* colaborativo basado en Git. A diferencia de GitHub después de su compra por Microsoft, su repositorio de código es privado con posibilidad de ser compartido públicamente.

https://gitlab.com/

INSDC

Organismo internacional de colaboración entre bases de datos públicas.

http://www.insdc.org/

InterPro

Base de datos integrada y herramienta de diagnóstico para la clasificación de proteínas. InterPro utiliza modelos predictivos conocidos como *firmas*, fundamentados en la información de varias bases de datos para predecir y anotar dominios y sitios importantes para la función de la proteína.

https://www.ebi.ac.uk/interpro/

KEGG

Portal del consorcio KEGG (Japón) que mantiene anotaciones funcionales con identificadores propios de manera universal. Esta base de datos integra las anotaciones de diferentes fuentes y establece relaciones de ortología funcional proporcionando descripciones comparables filogenéticamente. Tiene un gran potencial en su uso en biología de sistemas.

https://www.genome.jp/kegg/

RNA-seqBlog

Portal especializado en la secuenciación de ARN con noticias acerca de los avances tecnológicos en la materia, eventos como congresos y manuales para realizar diferentes análisis. También posee un espacio para preguntas y respuestas: SEQanswers.

https://www.rna-seqblog.com/

UniProt

Base de datos universal de referencia que mantiene y centraliza los recursos de secuencia de proteína y los datos de anotación. Es una colaboración entre el Instituto Europeo de Bioinformática (EMBL-EBI), el Instituto Suizo de Bioinformática (SIB) y el Protein Information Resource (PIR).

https://www.uniprot.org/

Bibliografía

Arita, M., Karsch-Mizrachi, I., & Cochrane, G. (2021). The international nucleotide sequence database collaboration. *Nucleic Acids Research*, *49*(D1), D121-D124. https://doi.org/10.1093/nar/gkaa967

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 30. https://doi.org/10.1186/s13059-020-1935-5

Boguski, M. S., Lowe, T. M., & Tolstoshev, C. M. (1993). dbEST--database for "expressed sequence tags". *Nature Genetics*, *4*(4), 332-333. https://doi.org/10.1038/ng0893-332

Bourgon, R., Gentleman, R., & Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(21), 9546-9551. https://doi.org/10.1073/pnas.0914005107

Braslavsky, I., Hebert, B., Kartalov, E., & Quake, S. R. (2003). Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(7), 3960-3964. https://doi.org/10.1073/pnas.0230489100

Clark, M. J., Chen, R., & Snyder, M. (2013). Exome sequencing by targeted enrichment. *Current protocols in molecular biology*, *Chapter 7*, Unit7.12. https://doi.org/10.1002/0471142727.mb0712s102

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*, 13. https://doi.org/10.1186/s13059-016-0881-8

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207-210. https://doi.org/10.1093/nar/30.1.207

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., ... Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, *323*(5910), 133-138. https://doi.org/10.1126/science.1162986

Gentleman, R. C., Carey, V. J., Bates, D. M. *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol 5*, R80. https://doi.org/10.1186/gb-2004-5-10-r80

Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P., & Minghim, R. (2015). InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics*, *16*(1), 169. https://doi.org/10.1186/s12859-015-0611-3

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J., & Tibshirani, R. (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research*, *22*(5), 519-536. https://doi.org/10.1177/0962280211428386

Liu, Y., Zhou, J., & White, K. P. (2014). RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics (Oxford, England)*, *30*(3), 301-304. https://doi.org/10.1093/bioinformatics/btt688

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology*, *13*(5), e1005457. https://doi.org/10.1371/journal.pcbi.1005457

Maynard, K. R., Collado-Torres, L., Weber, L. M. *et al.* (2021) Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci 24*, 425-436. https://doi.org/10.1038/s41593-020-00787-0

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., Jarosz, M., & Milos, P. M. (2009). Direct RNA sequencing. *Nature*, *461*(7265), 814-818. https://doi.org/10.1038/nature08390

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., & Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. *Scientific Reports*, *6*, 25533. https://doi.org/10.1038/srep25533

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, *11*(9), 1650-1667. https://doi.org/10.1038/nprot.2016.095

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139-140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24-26. https://doi.org/10.1038/nbt.1754

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463-5467. https://doi.org/10.1073/pnas.74.12.5463

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, *270*(5235), 467-470. https://doi.org/10.1126/science.270.5235.467

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195-197. https://doi.org/10.1016/0022-2836(81)90087-5

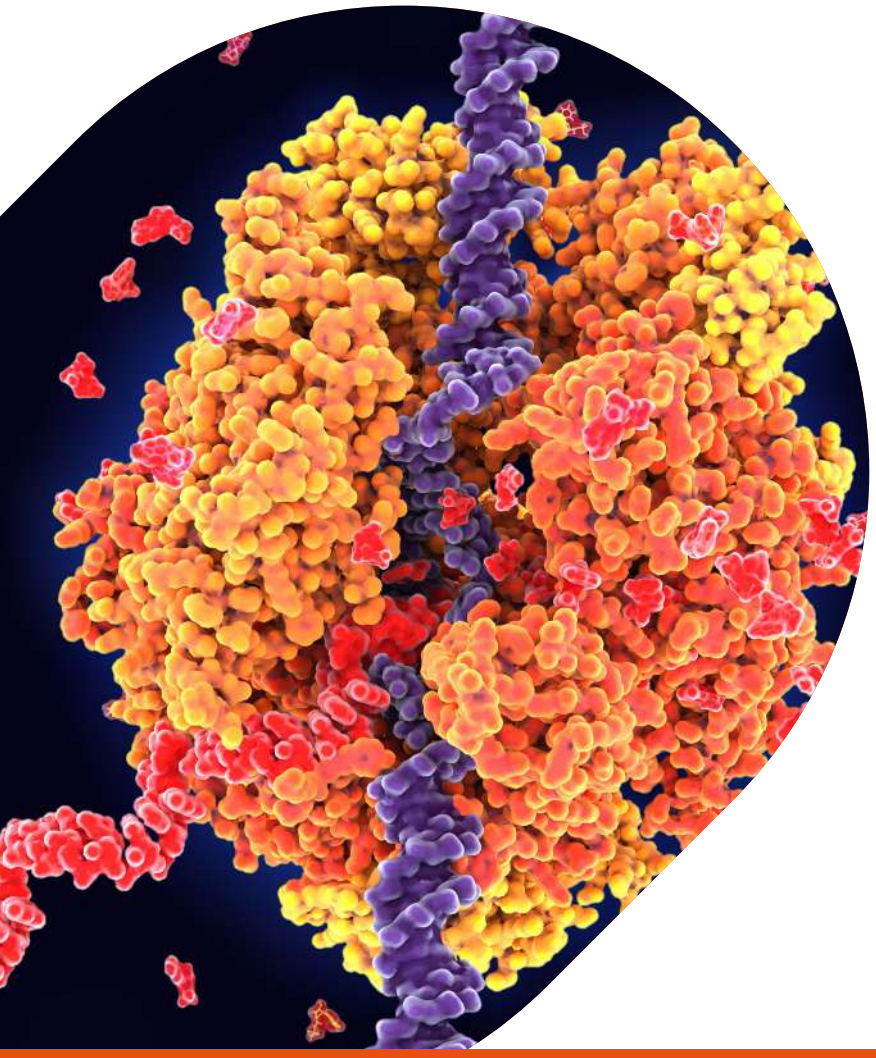
Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., ... Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science (New York, N.Y.)*, *353*(6294), 78-82. https://doi.org/10.1126/science.aaf2403

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377-382. https://doi.org/10.1038/nmeth.1315

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Research*, *21*(12), 2213-2223. https://doi.org/10.1101/gr.124321.111

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, *25*(9), 1105-1111. https://doi.org/10.1093/bioinformatics/btp120

Wang, Z., Gerstein, M. & Snyder M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet 10*, 57-63 . https://doi.org/10.1038/nrg2484



Autora
Dra. Nuria Mauri

viu Universidad Internacional de Valencia