A. Expresión genética midiendo mRNA: RNA-seq

La secuenciación del ARN (RNA-seq) es una técnica de secuenciación que nos permite estudiar la cantidad de ARN asociada a tránscritos de distintos genes que existe en una muestra biológica en un momento determinado. En este apartado expondremos la elaboración, almacenaje y análisis de los datos característicos de esta técnica.

A.1. Extracción del ARN

Antes de realizar la secuenciación del ARN, debe extraerse y seguidamente separarse de su entorno celular. Hay varios métodos de extracción, cuya discusión queda más allá de los objetivos de este trabajo. Lo que debemos tener en cuenta es la importancia de procesar el ARN de forma extremadamente controlada y estandarizada, así podemos tener un pleno conocimiento del proceso de aislamiento que se llevó a cabo, aprovechando esa información para analizar de una forma más satisfactoria los datos más adelante. Cabe destacar, que los métodos de extracción a menudo son imperfectos, obteniendo pequeñas cantidades de contaminación de ADN.

A.2. Preparación de librerías

Las librerías son una colección de fragmentos de ADN, sintetizados a partir del contenido en ARN de una muestra en un contexto dado, los cuales están listos para ser secuenciados siguiendo un proceso específico. Uno de los protocolos más populares para la preparación y secuenciación de librerías a partir del ARN se atiene a un método llamado "Secuenciación por síntesis" desarrollado por la firma bio-tecnológica estadounidense Illumina [1]. Este proceso de secuenciación depende originalmente del proceso biológico por el cual las enzimas como la ARN polimerasa "leen" el ADN continuamente. Son capaces de descifrar la secuencia exacta de bases nitrogenadas de ARN (Adenina,Citosina,Uracilo,Guanina) que hace falta seguir para construir una copia perfecta complementaria de una cadena individual de ADN a la que denominamos ADNcomplementario (ADNc).

El primer paso para obtener las librerías, es aislar el ARN del tejido de una muestra y mezclarlo con desoxirribonucleasa (DNasa), una enzina que se caracteriza por catalizar la rotura de los enlaces del ADN presente en la muestra. La DNasa reduce la cantidad de ADN genómico. La cantidad de degradación de ARN se verifica con electroforesis en gel y capilar y se usa para asignar un número de integridad de ARN a la muestra, que describe la calidad de ARN extraído. El ARN se transcribe inversamente a ADNc porque el ADN es más estable y permite la amplificación. La fragmentación y la selección del tamaño, gracias a la amplificación, se realizan para purificar secuencias que tienen la longitud adecuada para la máquina de secuenciación. La fragmentación del ADN se consigue gracias a la transposasa, que es una enzima que se une al extremo de un transposón, siendo éste una secuencia de ADN capaz de cambiar su posición dentro de un genoma. Lo que hace esta enzima es catalizar ese movimiento en base de un mecanismo de corta y pega. En este contexto se utiliza para cortar el ADN en fragmentos de manera aleatoria. Después, se pasa a la colocación de secuencias adaptadoras [11], que son secuencias de oligonucleótidos (secuencia de fracción de ADN), que sirven para identificar todos los fragmentos correctamente y tener la posibilidad de secuenciar diferentes muestras al mismo tiempo, ya que cada muestra puede llevar ligados diferentes adaptadores. Se realiza una amplificación con ayuda de unos cebadores afines (secuencias de nucleótidos de ARN que constituyen el inicio de la replicación de la cadena de ADN complementario), teniendo como resultado dos sitios de unión, donde se añadirán ésos adaptadores a los dos lados, que serán complementarios a los oligonucleótidos de las celdas de flujo, que describiremos a continuación.

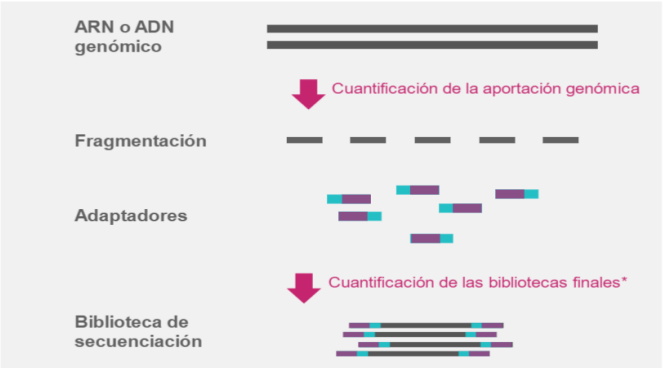

Figura 1: Preparación de librerías

A.3. Producción de conglomerados

El siguiente paso es la producción de los conglomerados. Para ello, contamos con la ayuda de las celdas de flujo. Estas celdas son análogas a un portaobjeto de vidrio donde se han tallado millones de nano pozos que contienen dos tipos de oligonucleótidos en su interior, donde los diferentes fragmentos se amplificarán de manera isotérmica. Una única hebra de la muestra entrará a cada pozo y una polimerasa creará la secuencia complementaria del fragmento hibridado. Como resultado, obtenemos una hebra bicatenaria formada por una hebra complementaria y nuestra hebra de interés. Esta última se libera y se retira. Mas tarde, nuestra hebra se dobla, adhiriéndose al segundo tipo de oligonucleótido de nuestro nano pozo. En ese momento, una segunda polimerasa volverá a crear una secuenciación de nuestra hebra (proceso de amplificación en puente). Esta estructura se desnaturaliza teniendo dos tipos de hebras una forward y otra reverse, la reverse se corta y se lleva a la secuenciación con cuidado de proteger sus extremos para que no ocurra una hibridación indeseada. Este proceso se repite varias veces y se crea un conglomerado de hebras (amplificación clonal) [1].

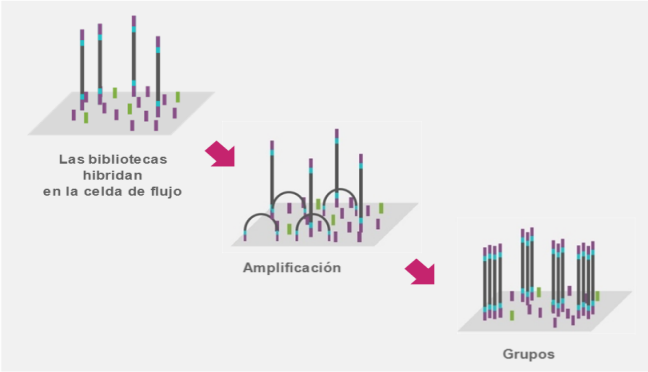

Figura 2: Producción de conglomerados

A.4. Secuenciación

Una vez obtenidos los conglomerados de las hebras, el siguiente paso consiste en que a cada hebra se le adhiere un cebador y además se le añade una solución que contenga los distintos nucleótidos a los que hemos añadido un fluoróforo a cada uno que emitirá luz de un color tras su adhesión. Cada nucleótido está

marcado con un fluoróforo distinto, por tanto, cada vez que la ADN polimerasa añada uno, llegará una señal que se quedará grabada en una serie de cámaras automáticas utilizando microscopios sofisticados. El número de adhesiones de los nucleótidos determina la longitud de lectura y la longitud de onda y la intensidad determinará el nucleótido correspondiente a cada adhesión.

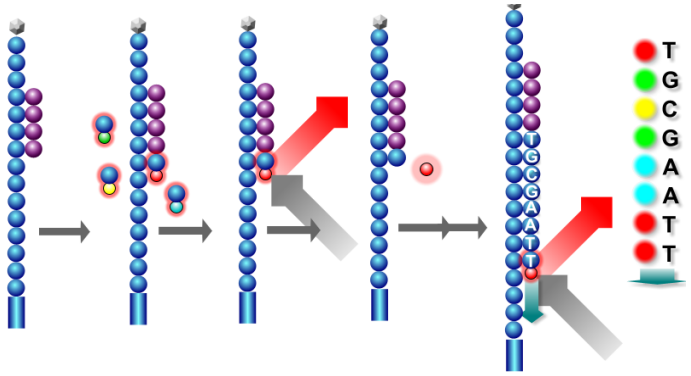

Figura 3: Secuenciación

Profundidad de secuenciación y cobertura

Cuando hablamos de cobertura hablamos del número de lecturas que se llevan a cabo en relación con el tamaño del genoma secuenciado. Siendo una estimación de cuántas veces se secuencia cada base del genoma. La ecuación que nos da la cobertura fue creada por Lander-Waterman [1]:

$$Cobertura = \frac{tamaño\ de\ lectura \cdot número\ lecturas}{tamaño\ del\ genoma}$$

Para evitar errores y lecturas erróneas, debemos de procesar cada base más de una vez, así poder distinguir los errores de secuencia de las variantes genómicas. Cabe destacar que el valor de la cobertura siempre será una estimación. Para el RNA-seq el valor de la cobertura no contiene una gran importancia práctica, ya que el tamaño del transcriptoma no se conoce con tanta precisión como el tamaño del genoma, ya que este último corresponde con el material genético total presente en los cromosomas, común para todas las células, mientras que el transcriptoma constituye la lectura de algunos de esos genes, que se transcriben en ARNm para hacer diferentes funciones, depende del momento y de la célula estudiada, además la cobertura por base variará drásticamente entre las distintas transcripciones dependiendo de su expresión.

Además tenemos que destacar dos tipos de lecturas, *Single-Read* (SR) y *Paired-End* (PE). La secuencia de lectura única (SR) determina la secuencia de ADN de solo un extremo de cada fragmento de ADN, produce la secuenciación en una única dirección, mientras que la secuenciación de extremo emparejado (PE) produce la secuencia de ambos extremos de cada fragmento de ADN. En este trabajo utilizaremos las primeras.

A.5. Datos

Almacenamiento de lecturas formato FASTQC

Las lecturas se almacenan principalmente con un formato llamado FASTQC, pero puede variar según la plataforma de secuenciación. Estos archivos agrupan la secuencia de cada una de las lecturas que tienen lugar en cada secuenciación con su valor de calidad, sobre el que hablaremos con más detalle más adelante.

Como hemos dicho anteriormente, Illumina se basa en la secuenciación gracias a la fluorescencia liberada por cada nucleótido en el proceso de incorporación a la hebra correspondiente. Cuando se completa la secuenciación las imágenes captadas durante todos los pasos de la síntesis de ADN, son analizadas según el color y la intensidad de la fluorescencia en el espectro obtenido, traduciéndolas a las letras de los nucleótidos correspondientes. Este proceso de traducción del espectro a las letras, se denomina "llamada de bases " o "base call" en inglés.

Para corregir las carencias de este método, cada lectura tiene asociada el valor de calidad anteriormente mencionado, que hace referencia a la probabilidad de error, es decir, representa la seguridad con la que podemos afirmar que hemos asignado la base correctamente. Este valor se denomina *Phred* y tiene asociada la letra Q, es logaritmicamente proporcional a la probabilidad (p) de que la lectura de la base individual haya sido incorrecta.

$$Q = -10 \log_{10}(p)$$

Por ejemplo, si obtenemos $Q = 10$ sabemos que tenemos un error en cada 10 bases, si obtenemos $Q = 20$ obtendríamos un error en cada 100. Cuanto mayor es el valor de calidad, mayor probabilidad de que hayamos nombrado la base correctamente.

Para determinar los valores de calidad, *Phred* primero calcula varios parámetros relacionados con la forma y la resolución de los picos del espectro de cada base. Después, usa estos parámetros para buscar el valor de calidad correspondiente en tablas de búsqueda, que se generaron a partir de porciones de secuencia donde se conocía la correcta sucesión de bases.

A.6. Mapeo y alineación

Uno de los pasos más importantes de los experimentos de secuenciación de alto rendimiento se llama mapeo y consiste en asociar las lecturas secuenciadas al lugar de origen más probable del genoma de referencia [1]. Averiguando dónde coinciden con mayor eficacia una secuencia corta de nucleótidos y el genoma, obtenemos la información de su localización y por tanto podemos utilizar esa información para localizar y cuantificar los genes expresados, así como descubrir nuevos genes. Como esto lleva a la alineación de nucleótidos de dos o más hebras, también se le llama alineación de lectura.

El principal desafío de la alineación es mapear millones de muestras en un tiempo razonable a pesar de los errores de secuenciación, elementos repetitivos y variaciones genómicas. Se emplean diferentes estrategias por los programas de alineación para mantener un equilibrio entre la tolerancia de errores y la fidelidad del mapeo.

Una de las principales ventajas de la secuenciación de ARN con respecto a metodologías anteriores de medida de la expresión genética es la posibilidad de medir la cantidad asociada a las múltiples isoformas

de un mismo gen. Para saber qué es una isoforma debemos profundizar en la naturaleza de los genes.

> En la secuencia del ADN el contenido codificante de un gen no está distribuido de forma continua a lo largo de dicho gen, sino que tiene discontinuidades, llamadas intrones, cuya secuencia no codifica proteínas. Las partes de la secuencia que codifican proteínas son los exones. Tras la transcripción, el ARNm resultante es procesado mediante un mecanismo llamado *splicing* y los intrones son eliminados, resultando en un ARNm maduro que contiene únicamente la información de los exones.
> Aunque en el pasado se creía que cada gen tenía una única posibilidad de *splicing* y por lo tanto resultaba en una única proteína, actualmente sabemos que se produce el fenómeno de *splicing* alternativo, en que no todos los exones del gen están presentes en el ARNm maduro, dando lugar a varias proteínas diferentes en función de los exones presentes en la secuencia final (Sabater-Tobella, 2018) [10].

Las isoformas son formas alternativas de un mismo gen, producidas por splicing alternativos o en el uso de diferentes exones transcritos en el ARNm. Cada isoforma de un gen, tiene una secuencia diferente de ARNm. Cabe destacar que pueden existir isoformas en el ARNm con la misma secuenciación de bases ya que utilizan el mismo conjunto de exones de codificación. Por todo ello, los experimentos de secuenciación del ARN intentan mitigar este error alineando las semillas en vez de al genoma, al trascriptoma, ya que no están presentes los intrones, teniendo únicamente la información codificadora. De este modo, se intenta evitar que varias lecturas se alineen con más de una isoforma introduciendo ambigüedad a nuestro mapeo.

Los programas más importantes para llevar a cabo esta alineación como pueden ser STAR, TopHat, GSNAP, utilizan el genoma entero de referencia y anotan la existencia de genes. Más adelante, describiremos con más detalle, el programa STAR, dado que es el que utilizaremos en este trabajo

Aunque se han conseguido grandes avances en relación a la velocidad de computación a la hora de alinear las muestras, sigue siendo el paso computacional más costoso de llevar a cabo.

Una vez obtenidos los datos de alineación, es decir, una vez tenemos la información del origen de cada lectura individual, debemos hacer el recuento de las lecturas que se superponen en genes y tránscriptos conocidos. No solo nos interesa presentar un catálogo de los genes que están expresados en nuestra muestra, sino que también hacer un recuento de las tránscriptos relativas entre genes que se han llevado a cabo.

A.7. Alineación utilizando STAR (Spliced Transcripts Alignment to a Reference)

Existen varios métodos de alineación de las secuencias obtenidas a un genoma de referencia, siendo el objetivo del estudio la principal premisa de elección entre ellos. Para el análisis diferencial de la expresión genética, está demostrado que uno de los métodos más eficientes es la alineación mediante la utilización de STAR [5] [6]. Los pasos más característicos de esta herramienta son:

- Generación del índice del genoma: este paso debe llevarse a cabo una única vez por tipo de genoma. En el archivo que contiene este índice, debe almacenarse la información del genoma de referencia de forma comprimida y optimizada asegurando el rápido acceso a la información para su correcta comparación con la lectura de las secuencias. Incluye la longitud y el nombre de los distintos cromosomas, así como representaciones de la secuenciación del genoma y las coordenadas de la localización de los distintos genes.

- Alineación: en este paso se acopla cada lectura a su secuencia complementaria del genoma de referencia.

Método de alineación

Se llevan a cabo una serie de pasos fundamentales, la búsqueda de las "semillas", la agrupación, la adhesión y la puntuación. Para tener una mejor visión del proceso, nos centraremos en el análisis de una sola lectura.

Búsqueda de semillas

Se busca la coincidencia de mayor tamaño entre una parte de la secuenciación obtenida y una o más posiciones del genoma de referencia. A estas coincidencias se les denomina Maximal Mappable Prefixes (MMPs).

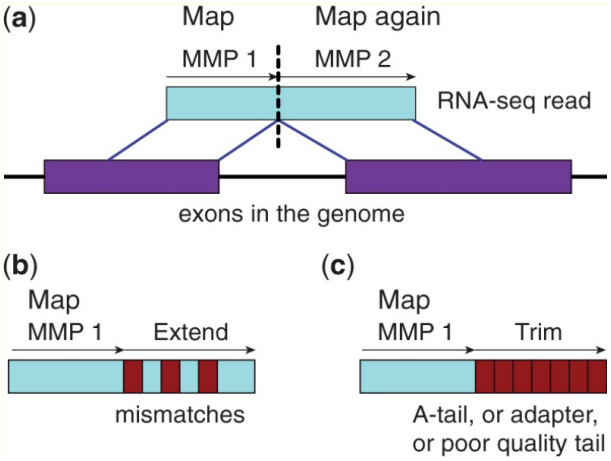

Figura 4: Alineación STAR

Una vez encontrada la coincidencia de esa primera parte, se busca la coincidencia de la parte que no ha sido colocada, esta es la principal ventaja de este algoritmo. Es que es capaz de separar una lectura en varias partes encontrando su mayor coincidencia con el genoma, obteniendo de una misma lectura varios MMPs, denominándolos "seed1", "seed2"...etc, que pueden aparecer en distintos exones, generando así valiosa información sobre la caracterización de los eventos de splicing asociados a cada tránscripto.(*Figura(a)*).

Si el algoritmo no es capaz de encontrar una perfecta coincidencia, para todas las partes de la lectura, debido a un desajuste en alguna base, alarga el MMP inmediatamente anterior a esa discordancia, permitiendo alineaciones con algunos desajustes utilizando las semillas como anclaje, como podemos observar en la *Figura(b)*. En algunas ocasiones esas elongaciones producen malas alineaciones y pasan a ser lo que se denomina como colas de mala calidad, descartándose (*Figura(c)*). La búsqueda de estas MMP, se realiza tanto en la dirección directa como inversa de la lectura, teniendo también la opción de elegir el punto de partida.

A.8. Lectura de datos/recuento de lecturas

Existen dos tipos de recuento de lectura [1] uno basado en la expresión de los genes y otro en la trascripción. El primero corresponde a contar los fragmentos que mapan a cada tránscripto independientemente, el

segundo con sumar todos las cuentas mapadas a tránscriptos de un mismo gen, que es el método utilizado en este trabajo.

A.9. Normalización y transformación del recuento de lecturas

El valor que obtiene un único gen en una única muestra viene definido por el número de lecturas registradas que se superponen coincidiendo con ese gen, pero como hemos visto anteriormente, hay diferentes factores que interfieren en la amplificación y secuenciación de las muestras, es decir, de los fragmentos originales de ARN. El número de lecturas que coinciden con un gen depende de:

- El propio perfil de expresión (es la medida de la actividad (de la expresión génica) de miles de genes simultáneamente, para crear una imagen global de la función celular), es decir, el número de moléculas de ARNm que hay en la muestra asociadas a cada gen.
- El tamaño de esas lecturas(si partimos las moléculas en fragmentos más pequeños, capturaremos más).
- La profundidad de la secuenciación, factor que varía también entre muestras.

Para comparar la diferencia de expresión genética de cara a dos condiciones diferentes, debemos comparar el número total de lecturas por gen, teniendo en cuenta el número total de lecturas y teniendo en cuenta también el ARN presente en cada muestra. Esto último puede variar de forma significativa para cada muestra. Así como el número total de muestras es un dato conocido, la librería que contiene la información del ARN estudiado puede ser muy extensa y compleja y puede sufrir contaminaciones biológicas de una muestra a otra. Para afrontar estas dificultades, se recurre a la normalización, de este modo podemos descartar las diferencias que no sean biológicas y utilizar la estadística para el estudio diferencial de la expresión de los genes.

Para la normalización de la profundidad de secuenciación, se utiliza generalmente la siguiente forma de expresión de los datos [7]:

Counts per million (CPM)

Para el recuento de lecturas, en este estudio utilizaremos las llamadas Counts Per Million (CPM), que son cuentas que van escaladas con el número total de lecturas de la secuenciación(N). Se contabilizan todas las lecturas de una muestra y se divide por un millón, ese valor corresponde a nuestro factor de escala. Después, se divide el recuento de las lecturas correspondientes a un gen (X_i) por ese factor, obteniendo una normalización de profundidad de secuenciación.

$$CPM_i = \frac{X_i}{\frac{N}{10^6}} = \frac{X_i}{N} \cdot 10^6$$

Para hacer comparaciones del nivel de expresión de los genes, necesitamos transformar los datos a una escala adecuada, y normalizar las diferencias técnicas entre muestras para aislar la variabilidad biológica de la técnica [8] [9]. Para ello, utilizamos las herramientas *edgeR* y *limma-voom*. Normalizar únicamente con el tamaño de la librería puede ser demasiado simple o escueta para algunas aplicaciones biológicas. La clasificación no depende solo del tamaño de cada librería, sino también de la composición de todo el ARN que se está testando. Para normalizar hace falta tener en cuenta este factor.

TMM (Trimed Mean of M-values)

La normalización de TMM (Trimed Mean of M-values) es un método simple y efectivo para estimar niveles relativos de producción de ARN a partir de datos de RNA-seq. El método TMM estima factores de escala entre muestras que se pueden incorporar a los métodos estadísticos utilizados para el análisis de expresión diferencial de los genes. En este apartado estudiaremos el método para la comparación de dos muestras, ya que es el utilizado en este estudio.

Haciendo una explicación más formal del requerimiento de otra normalización, definimos Y_{gk} como las cuentas observadas de un gen g en la librería k y siendo μ_{gk} el verdadero y desconocido nivel de expresión de ese gen (número de transcripciones), L_g la longitud del gen y N_k el número total de lecturas de la librería. Podemos calcular el valor esperado de Y_{gk} como:

$$E[Y_{gk}] = \frac{\mu_{gk} L_g}{S_k} N_k \quad ; \quad S_k = \sum_{g'} \mu_{g'k} L_{g'} \tag{A.1}$$

El principal problema es que mientras N_k es conocido S_k es desconocido y varía notablemente de muestra a muestra, dependiendo de la composición del ARN. La ecuación anterior nos acerca aunque no perfectamente, a estimar unívocamente el nivel de expresión genuino del gen g en la librería k, que no es otro que despejándolo (aunque no lo estamos despejando del todo, pues aún aparece a ambos lados de la ecuación):

$$\mu_{gk} \approx \frac{Y_{gk} S_k}{N_k L_g} \tag{A.2}$$

Como hemos dicho S_k no puede hallarse directamente, ya que no sabemos el nivel de expresión de los genes, por lo que calculamos la producción relativa de ARN de dos muestras $\rho(k, k') = \frac{S_k}{S'_k}$, es decir, estimamos ratios relativos a pares de muestras que nos permiten, ante la imposibilidad de obtener directamente μ_{gk} y $\mu_{gk'}$, al menos compararlos entre sí. Estimando el ratio $f_k = \frac{S_k}{S'_k}$ podemos estimar también $\mu_{gk}/\mu_{gk'}$:

$$\frac{\mu_{gk}}{\mu_{gk'}} = \frac{\frac{Y_{gk}}{N_{gk}}}{\frac{Y_{gk'}}{N_{gk'}}} \rho_g(k, k') \tag{A.3}$$

Cabe destacar, que escribimos $\rho_g(k, k')$ ya que es la estimación del ratio obtenida a partir de la expresión entre muestras relativas de un gen g determinado. Cada gen producirá una estimación de la ratio diferente.

Como típicamente vamos a tener más de dos muestras, lo que haremos será elegir una muestra r de referencia, y estimar los cocientes del resto de las muestras con respecto a la referencia, es decir, las ratios $\rho(r, k) = \frac{S_r}{S_k}$. Para calcularlos, explotamos una propiedad que se verifica en la mayoría de los experimentos, y es la existencia de una mayoría de genes G^* que no van a estar diferencialmente expresados entre las distintas muestras, para los cuales:

$$\mu_{g*,k} = \mu_{g*,r} \tag{A.4}$$

Para los genes donde se cumpla esta condición, el cociente de más a la izquierda de la ecuación (4.3) sería igual a uno, por tanto obtendríamos la siguiente fórmula:

$$\frac{\frac{Y_{g*r}}{N_{g*r}}}{\frac{Y_{g*k}}{N_{g*k}}} = \rho_{g*}(r, k) \tag{A.5}$$

Como hemos mencionado anteriormente, para cada gen obtendremos un ratio estimado. Algunos genes nos darán una estimación sesgada (desviada) de $\rho(r, k) = \frac{S_r}{S_k}$. La razón es que para cietos genes, la asunción

de no expresión diferencial es falsa, y por tanto, la ecuación (4.4) no se cumple, y cuando la utilizamos, sesgamos el estimador de $\rho(r,k) = \frac{S_r}{S_k}$.

Estos valores sesgados, típicamente nos darán los valores más extremos de la distribución de estimaciones de $\rho(r,k) = \frac{S_r}{S_k}$, por lo que para evitar dicho sesgo, se procede a eliminarlo, o "recortarlo" (en inglés, trim). El algoritmo, tal como está implementado en *edgeR*, recorta el 30 % de valores de alejados de la media en cada caso.

De ese modo, nos hacemos cargo del sesgo, pero debemos contar también con el ruido. Para tratarlo, la clave consiste en admitir que la calidad de algunas de las estimaciones de nuestro ratio, va a ser mejor que otras, por lo que propondremos pesos de manera acorde, para basar la estimación final en un promedio pesado, donde los genes que contribuyen a dar estimaciones de mejor calidad (menor ruido), pesan más.

Para calcular el ruido se recurre a asimilarlo a la incertidumbre de la estimación de los ratios, entendidos como un ensayo de Bernouilli, en el que $\frac{Y_{g*k}}{N_k}$ representa la probabilidad de éxito (de tener un fragmento asociado al gen g* en a muestra k), tras N_k intentos. Resulta trivial que la varianza asociada a $\frac{Y_{g*k}}{N_k}$ es:

$$Var(\frac{Y_{g*k}}{N_k}) = \frac{\frac{Y_{g*k}}{N_k}(1 - \frac{Y_{g*k}}{N_k})}{N_k} = \frac{Y_{g*k}(N_k - Y_{g*k})}{N_k^2} \tag{A.6}$$

La idea final es obtener el promedio el logaritmo de los estimadores de :

$$log(\rho(r,k)) = \log(\frac{Y_{g*k}}{N_k}) - log(\frac{Y_{g*k}}{N_k}) \tag{A.7}$$

Y promediar este estadístico a través de los genes G* que han sobrevivido al "trimming", aplicándoles unos pesos basados en la estimación de la varianza del logaritmo del estimador asociado a cada gen, esto es, $Var(log(\rho(r,k)))$. Estas varianzas se pueden propagar desde (4.7) como sigue:

$$Var(log(\rho(r,k))) = (\frac{1}{E(\rho(r,k))})^2 Var(\rho(r,k)) = \frac{N_k - Y_{g*k}}{N_k \cdot Y_{g*k}} + \frac{N_r - Y_{g*r}}{N_r \cdot Y_{g*r}} \tag{A.8}$$

Y en función de ellas, los pesos se definen como:

$$w_{g,k} = Var(log(\rho(r,k)))^{-1} \tag{A.9}$$

Para terminar por definir el promedio definitivo de todas las $\rho(r,k)$ integrando la información de todos los genes útiles, como sigue:

$$TMM_k = \frac{\sum_{g \epsilon G*} w_{g,k} \cdot \rho_g(r,k)}{\sum_{g \epsilon G*} w_{g,k}} \tag{A.10}$$

Estos factores de renormalización se terminan integrando, por ejemplo, al correr *voom*, en los denominadores de la normalización estándar, esto es, el número de fragmentos de cada librería. Al calcular el $log(CPM)$ tal como hace en *voom*, tenemos:

$$log_2(CPM)_{gk} = log_2(10^{-6} \cdot \frac{Y_{gk} + 0,5}{N_k \cdot TMM_k + 1}) \tag{A.11}$$

Como hemos dicho, hallamos factores que serán utilizados en los métodos estadísticos para el análisis de los datos, por lo que se conservan las propiedades de la secuenciación. Como no se modifican los datos se pueden utilizar más adelante para el estudio de la comparación de expresión entre genes.

Referencias

[1] Dündar, F., Skrabanek, L., y Zumbo,P.(2015-2019). Introduction to differential gene expression analysis using RNA-seq. *Applied Bioinformatics Core—Weill Cornell Medical College*,1-97.

[2] RNAseq-an introduction. (2020). Consultado el 1 Junio 2020, de https://galaxyproject.org/tutorials/rb_rnaseq/

[3] Secuenciación: tecnología de Illumina. (2020). Consultado 1 Junio 2020, de https://support.illumina.com/content/dam/illumina-support/courses/sequencing-illumina-technology-wbt-esp/story_html5.html?iframe

[4] Zhang, C., Zhang, B., Vincent, M., y Zhao, S.(2016). Bioinformatics Tools for RNA-seq Gene and Isoform Quantification. *Journal Of Next Generation Sequencing Applications, 03*(03). doi: 10.4172/2469-9853.1000140

[5] Meeta Mistry, M. (2020). Alignment with STAR. Retrieved 1 June 2020, from https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

[6] Dobin, A., Davis, C., Schlesinger, F., Drenkow, J., Zaleski, C., y Jha, S. et al. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21. doi: 10.1093/bioinformatics/bts635

[7] What the FPKM? A review of RNA-Seq expression units. (2020). Consultado 1 Junio 2020, de https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/

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

[9] Zhou, Y., Lin, N., y Zhang, B. (2014). Erratum to: An iteration normalization and test method for differential expression analysis of RNA-seq data. *Biodata Mining, 7*(1). doi: 10.1186/s13040-014-0030-4

[10] Exón e intrón. (2018, Noviembre 21). Consultado el Mayo 29, 2020, de https://www.eugenomic.com/es/home/genomica/glossary/e/Exon.html

[11] Fuchs RT, Sun Z, Zhuang F y Robb GB.(2015) Bias in Ligation-Based Small RNA Sequencing Library Construction Is Determined by Adaptor and RNA Structure. *PLoS ONE, 10*(5): e0126049.

B. Genotipado: SNP (Single Nucleotide Polymorfism)

Los SNPs (Single Nucleotide Polymorfism) [2] son pequeñas variaciones genéticas en la secuencia del ADN encontradas en individuos de la misma especie que afectan a la identidad de un solo nucleótido en una posición determinada del genoma. Es decir, son lugares del genoma donde distintos individuos presentan nucleótidos con diferentes bases nitrogenadas. Al conjunto de los distintos nucleótidos que se presentan en la población para un determinado SNP se les llama alelos. Al conjunto de las variaciones específicas registradas para un determinado individuo en todo su genoma (conjunto de SNPs, junto a otras formas de variación genética) se le denomina genotipo. Para caracterizar el genotipo de un individuo de manera global, se utilizan chips de genotipado, que son dispositivos experimentales con forma de matriz diseñados para detectar SNPs en la secuencia del genoma de un grupo de individuos.

En cada posición, en un cromosoma, hay dos posibles alelos, que van codificados como A y B. El alelo A puede ser un par de bases conjugadas A-T (o su reverso equivalente T-A), y el B sería en este caso, el par de bases CG (o GC). Esta asociación puede ser la contraria también (que la letra A represente un par conjugado CG (o su equivalente GC) y el B represente AT (o equivalente, TA). Por convenio, se puede asignar la letra A al alelo más frecuente en la población, sea el par AT o el par CG. En organismos diploides reproducción sexual, tenemos dos copias de cada cromosoma, y por tanto, al menos en lo que concierne a los autosomas, de cada gen, heredados de cada uno de nuestros progenitores, y por tanto, también de cada SNP, el genotipo con respecto a un SNP da cuenta del alelo que tenemos en ambas copias: AA, AB o BB. En el ejemplo anterior, AA significa tener AT-AT en los cromosomas materno y paterno (genotipo homocigoto); AB correspondería a AT-CG (genotipo heterocigoto) y BB a CG-CG (genotipo homocigoto).

B.1. Detección

Para la realización de estas matrices se utilizan los mismos principios bioquímicos utilizados por el RNA-seq [4]. Utilizando que cada nucleótido se adhiere con su base nitrogenada complementaria (AT, GC).

Se utiliza un chip de ADN (pequeña pieza de vidrio de silicio a la que se adhiere químicamente fragmentos de ADN), donde los más avanzados tienen varios compartimentos capaces de analizar el ADN de un sujeto distinto cada uno. Cada compartimento es una matriz microscópica compuesta por cientos de miles de celdas dentro de un centímetro cuadrado. Cada celda es tan pequeña como una bacteria y está cubierta con fragmentos de ADN que han sido sintetizados para que complementen un determinado corte de ADN. Las hebras del ADN que se adhieren a la celda están fragmentadas de tal forma que llegan hasta un polimorfismo concreto del genoma pero no lo incluyen, estas hebras se adhieren a una hebra complementaria que se encuentra suspendida en la muestra a analizar, que contiene la región que hibridiza, más una región adicional donde se encontrará el polimorfismo de interés. Por lo que la matriz con las distintas celdas se llena de fragmentos de ADN que se está analizando. Además, se unirá a una posición determinada de la celda, dependiendo del polimorfismo contiguo.

Centrándonos en el análisis de un fragmento de una celda en particular, para ejemplificar el procedimiento, veremos los tres casos diferentes mencionados anteriormente. Los tres tipos de polimorfismos el AA, el AB o el BB. En este ejemplo vemos tres sujetos distintos, donde en una posición concreta de la celda observamos que la última base nitrogenada de los fragmentos adheridos cambia dependiendo del sujeto.

En las dos primeras imágenes, vemos que todos los fragmentos acaban en Adenina, como vimos con el RNA-seq, el proceso seguido, consiste en que con ayuda de una encima el ADN polimerasa, se adhiere el nucleótido complementario al que se ha añadido un fluoróforo concreto, en nuestro ejemplo, a la Timina se le añade uno rojo. En nuestro primer ejemplo obtendremos la celda de color rojo, ya que todos los

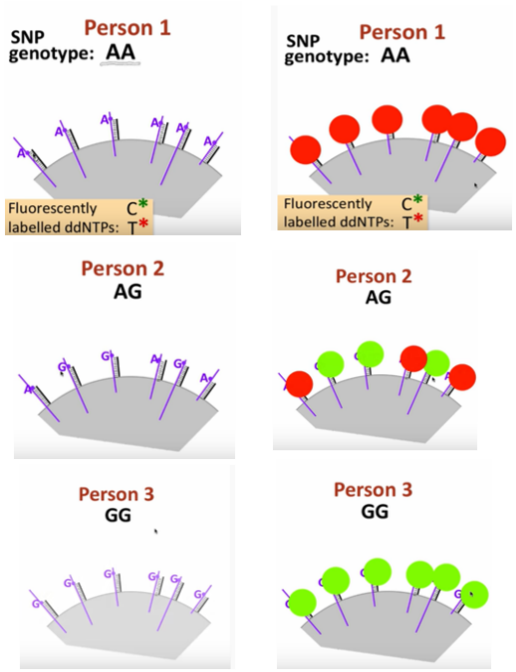

Figura 5: Genotipos

fragmentos acaban en Adenina, por tanto la Timina es el único nucleótido que se unirá. Teniendo este color, sabemos que el polimorfismo es AA, es decir homocigoto. En el segundo ejemplo obtenemos una celda de color amarillo, siendo combinación de los dos colores verde y rojo, correspondientes a la Guanina y a la Timina, ya que en esta posición de la celda existen fragmentos acabados tanto en Citosina como en Adenina. Siendo en este ejemplo, el polimorfismo heterocigoto AB. Y por último, obtenemos una celda verde ocasionada por la Citosina. También homocigoto, siendo esta vez el ejemplo BB. Obteniendo al final algo como esto:

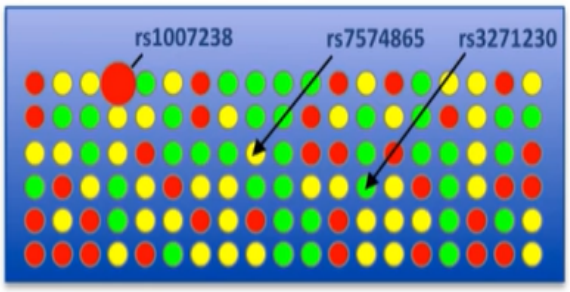

Figura 6: Matriz genotipos

Simplemente escaneando nuestro chip utilizando un láser de alta resolución podremos obtener la información del tipo de polimorfismo que tenemos. Cabe destacar el principal inconveniente de este tipo de análisis, permitiéndonos únicamente el reconocimiento del genotipo de los SNPs definidos a priori, es decir, se requiere ese conocimiento previo de dónde se encuentran los SNPs relevantes para la investigación. Cabe la posibilidad de que muchos de ellos no sean interesantes en las muestras, mientras que al mismo tiempo se pueden estar perdiendo SNPs importantes en otras partes del genoma que no es posible

determinar así. La alternativa es el llamado Whole Genome Sequencing (WGS) la secuenciación completa del ADN del genoma de un organismo.

B.2. Fiabilidad de la medida

Debemos tener en cuenta que la intensidad asociada a cada medida [3], depende de la afinidad que existe entre la muestra y el ADN "diana" presente en el chip. Esto puede llevar a medidas erróneas y es la principal fuente de ruido del experimento.

La afinidad mencionada va ligada a la intensidad de fluorescencia registrada en las celdas. Se han desarrollado a lo largo de los años algoritmos, como por ejemplo el llamado *birdseed*, que puedan registrar para cada pozo la intensidad de la fluorescencia roja, verde o amarilla, e inferir, de esas medidas, qué SNP es un AA, un AB o un BB (proceso denominado "SNP calling"), al cual se añade un score que representa el grado de incertidumbre del call. Se representa la intensidad del alelo AA frente a la intensidad del alelo BB, donde cada punto corresponde a un SNP concreto. Ejemplificamos este proceso mostrando la gráfica del análisis de intensidades de una muestra para las posiciones de las sondas con ADN diana acabado en Adenina. Los datos que se obtienen presentan la forma representada en la *Figura 7*.

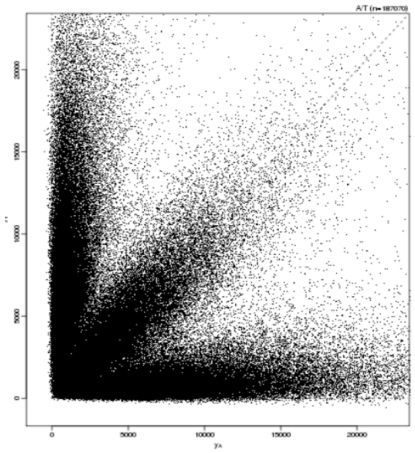

Figura 7: Ejemplo: Datos fluorescencia.

Diferenciamos en la gráfica tres brazos correspondientes en sentido horario con los genotipos AA, AB y BB. Como podemos observar, existen varios SNP difíciles de clasificar que se encuentran entre cada uno de esos "brazos". Por lo que a cada SNP se le asigna un genotipo dependiendo de su posición en la gráfica, con una puntuación que representa la fiabilidad de la medida en función de la distancia de su brazo correspondiente. Para cada cual se calcula una relación con su intensidad, en función de su posición:

$$\theta = \frac{2}{\pi} \cdot \arctan \frac{X}{Y}$$

Cada uno de los tres genotipos (AA,AB,BB) representa una región en el espacio unidimensional θ. La proximidad de cada SNP a cada una de estas, determina el genotipo y su separación representa la puntuación de calidad, con la fiabilidad con la que podemos asignar ese genotipo.

Concluyendo, los datos finales obtenidos de este análisis corresponderían con una lista completa de los SNPs de la muestra, además de su posición cromosómica y su genotipo más probable con una puntuación de calidad asociada.

B.3. Imputación de SNPs

La imputación en genética hace referencia a la inferencia estadística de genotipos que no han sido observados experimentalmente. En este estudio hablaremos concretamente de las variaciones genéticas correspondientes con SNPs, siendo estas las más comunes. Este proceso se consigue gracias a la utilización de haplotipos, siendo éstos secciones de ADN contiguas, conteniendo dos o más variantes pertenecientes a loci distintos, que se transmiten juntos, conocidos en una población. Para posibilitar la imputación de genotipos en muestras humanas, proyectos como HapMap [1], que se completó en 2005, han sido clave. Hapmap, por ejemplo, generó el primer mapa de haplotipos del genoma humano. Lo que da la posibilidad de hacer estudios como este, donde se utilizan además de los datos de genotipado obtenidos en el laboratorio, SNPs no observadas experimentalmente, pero cuyos genotipos han sido como ya hemos dicho inferidos estadísticamente, es decir, "imputados".

Referencias

[1] Acerca Del Proyecto Internacional Hapmap. (2020). Consultado el 10 Julio 2020, de https://www.genome.gov/acerca-del-proyecto-internacional-hapmap

[2] LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research, 37*(13), 4181-4193. doi: 10.1093/nar/gkp552

[3] Genotyping, g. (2020). Genotyping, genotype calling or SNP calling?. Consultado 1 Junio 2020, de https://www.biostars.org/p/277927/

[4] Redfield, R.[Useful Genetics].(2015, 10) 6G-How SNP-typing works[Archivo de vídeo]. Recuperado de https://www.youtube.com/watch?v=Naona1y_I2Ut=343s

C. Datos

En este TFG, se pretende estudiar las relaciones genotipo-fenotipo en macrófagos humanos procedentes de un panel de 90 individuos de ancestralidad europea, para las cuales, previamente al desarrollo de este trabajo, se recolectaron, en paralelo, datos genotípicos (SNPs arrays) y transcriptómicos (RNA-seq).

Es importante destacar que el proyecto de investigación del que proceden los datos analizados, a modo de ejemplo, en este TFG, superó los requerimientos concernientes a aspectos éticos, bajo criterio del comité de ética del Hospital Sainte Justine, en Montreal, Canadá, centro donde se centralizó la coordinación del proyecto (protocolo #4023). La información obtenida de los donantes fue completamente anónima, resultando imposible para ninguno de los investigadores la identificación de los donantes de cada muestra. En el contexto de la realización de este trabajo, se usaron como inputs dos matrices, conteniendo datos de expresión y caracterizaciones genotípicas, donde la identidad de cada donante es simplemente un índice mudo, pareado entre ambos elementos.

Específicamente, los macrófagos se obtuvieron de acuerdo al siguiente proceso experimental:

Las llamadas Peripheral Blood mononuclear cells (PBMCs) fueron aisladas por un método de centrifugación llamado Ficoll-Paque centrifugation. Los PBMCs son una fracción de células sanguíneas que presenta un núcleo redondo único, que incluye monocitos y linfocitos pero no eritrocitos, plaquetas o granulocitos [1]. Los monocitos fueron purificados de los PBMCs por selección positiva con microesferas CD14 magnéticas. La pureza de los monocitos aislados fue verificada utilizando un anticuerpo que actúan sobre el antígeno CD14 (el BD Biosciences), que es un marcador celular de superficie característico para los monocitos. Solo las muestras de las que se obtiene un 90 % de pureza fueron utilizadas para diferenciar macrófagos. Los monocitos fueron cultivados a lo largo de 7 días en el medio de cultivo RPMI-1640, enriquecido con un 10 % de un suero fetal bovino y un factor estimulante de colonias de macrófagos. Este tratamiento estimula a los monocitos y dirige su diferenciación en macrófagos. Una vez los macrófagos están purificados, justo antes de la secuenciación, comprobamos su estatus de diferenciación/activación de las celdas por citometría de flujo para para impedir la inclusión de células insuficientemente diferenciadas (es decir, monocitos) o macrófagos activados inmunológicamente, los cuales se sabe que presentan una divergencia de perfil transcriptómico. Con ese fin, solo conservamos muestras que presentan el fenotipo esperado para macrófagos inactivos (CD1a +, CD14 +, CD83 y HLA-DRlow), que luego se utilizaron en experimentos posteriores.

Más adelante, el total del ARN presente en los macrófagos fue extraído y espectrofotométricamente evaluado, de tal forma que solo se conservaron las muestras que no mostraban evidencia de degradación de ARN. Esto se determinó utilizando el número de integridad de ARN con un umbral de $RIN > 8$. A continuación, las librerías de secuenciación de ARN fueron preparadas. Una vez preparadas, fueron codificadas y agrupadas (6 bibliotecas por grupo) en cantidades equimolares y se secuenciaron con un single-read de 100 pb en Illumina HiSeq2500. Las secuencias adaptadoras y bases de puntuación de baja calidad (Puntuación de calidad $Phred$ <20) se eliminaron primero y las lecturas resultantes se asignaron a la secuencia de referencia del genoma humano (Ensembl GRCh37 versión 75).

Los 90 individuos que se incluyeron en el sample-set utilizado en este estudio, fueron genotipados en una matriz de genotipato Illumina HumanOmni5 con éxito, obteniéndose ∼4.3 millones de SNPs. Después de aplicar el control de calidad para eliminar los SNP con asociaciones de genotipo deficientes, se realizó el proceso de imputación utilizando un panel de referencia de haplotipos de individuos con ascendencia europea.

La recopilación y el procesamiento previo de los datos de ARN y ADN fue realizado de acuerdo a lo

expuesto anteriormente de una manera breve. Proceso previo al trabajo realizado en el contexto de este estudio, cuyo punto de partida eran las limpias y cualitativamente controladas matrices de recuento de lecturas, para la secuenciación del ARN (58051 genes provenientes de 90 individuos) y genotipos de SNP (6159736 SNP autosómicos de los mismos 90 sujetos).

Referencias

[1] Peripheral blood mononuclear cell. (2020). Consultado el 4 Julio 2020, de https://en.wikipedia.org/wiki/Peripheral_blood_mononuclear_cell