Reconstructing pedigrees using RNA-seq data

Natàlia Blay Magriñá
Máster en Bioinformática y Bioestadística
Genómica comparativa y reguladora

Tanya Vavouri
Carles Ventura Royo

5 de junio de 2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Reconstructing pedigrees using RNA-seq data*
Nombre del autor:	*Natàlia Blay Magriñá*
Nombre del consultor/a:	*Tanya Vavouri*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	06/2018
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Genómica comparativa y reguladora*
Idioma del trabajo:	*Castellano*
Palabras clave	*RNA-seq, pedigrí*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

El objetivo de este trabajo es crear un método para determinar y representar relaciones de parentesco entre diferentes individuos a partir de datos de RNA-seq. Actualmente no existe ningún método descrito para este tipo de datos, mientras que sí existen para otros tipos de datos como arrays de genotipado. La reconstrucción de pedigríes tiene aplicaciones diversas como la detección de problemas de etiquetado.

La metodología consiste en (i) hacer un control de calidad de los datos crudos, (ii) mapear los reads en el genoma, (iii) filtrar los datos de los reads mapeados eliminando duplicados y reads con mapeo múltiple, (iv) seleccionar los SNPs comunes que no se encuentren en un sitio de repeticiones o en genes improntados, (v) filtrar los SNPs en los reads y obtener los genotipos, (vi) determinar y representar la relación entre parejas de individuos y (vii) representar el pedigrí completo.

Tras seleccionar los filtros y sus valores de corte se ha obtenido un método que permite diferenciar distintas relaciones de parentesco y representarlas. Las relaciones de primer grado (padre-hijo y hermanos completos) son las que más fácilmente se detectan, mientras que las relaciones más lejanas (abuelo-nieto o tío-sobrino) son más difíciles de diferenciar de individuos no relacionados.

Abstract (in English, 250 words or less):

The aim of this work is to develop a method to determine and represent kinship relationships between individuals using RNA-seq data. Currently, there are methods to reconstruct pedigrees based on data from genotyping arrays, although there is no effective method based on RNA-seq. Pedigree reconstruction has several applications such as detecting labeling mistakes.

The methodology is as follows: (i) perform a quality control of raw data, (ii) map the reads to the reference genome, (iii) filter the mapped reads by eliminating duplicates and multi-mapping reads, (iv) select common SNPs that are not in repeats nor imprinted genes, (v) filter the SNPs in reads and obtain the genotypes, (vi) determine and represent the relationship between pairs of individuals and (vii) represent the whole pedigree.

After the selection of filters and their thresholds, a method that allows to discriminate between different kinship relationships and to represent them has been obtained. First-degree relationships (parent-offspring and full siblings) are the most easily detected, while more distant relationships (grandparent-grandchild or uncle-nephew) are more difficult to differentiate from unrelated individuals.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del Trabajo

El pedigrí es un esquema que indica las relaciones de parentesco entre diferentes individuos, englobando relaciones cercanas como padre-hijo o hermanos completos hasta relaciones más distantes y complejas como primos segundos o bisabuelo-bisnieto. La reconstrucción de pedigríes puede ser utilizada tanto para establecer relaciones de parentesco entre individuos previamente desconocidas como para confirmar relaciones conocidas. Sus aplicaciones son diversas, como la inferencia genealógica, la identificación de víctimas o la detección de problemas en el etiquetado de las muestras [1]. Actualmente, existen métodos para reconstruir un pedigrí basados en datos de arrays de genotipado [2][3], aunque no hay ningún método efectivo basado en RNA-seq. La problemática con este tipo de datos es que puede existir un desequilibrio en los niveles de expresión de algunos alelos (debido, por ejemplo, al imprinting), pudiendo clasificar como homocigotos sitios que en realidad son heterocigotos [4], y esto puede dar incongruencias a la hora de reconstruir un pedigrí.

El objetivo de este trabajo es crear un método para determinar y representar relaciones de parentesco entre diferentes individuos a partir de datos de RNA-seq en base al genotipo de unos SNPs previamente seleccionados.

1.2 Objetivos del Trabajo

El trabajo se va a dividir en dos objetivos generales y cuatro objetivos específicos (Figura 1):

1. Determinar el genotipo de SNPs a partir de los datos de RNA-seq: esta parte va a consistir en el mapeo de los datos de RNA-seq en el genoma, el filtrado de los datos para minimizar tanto errores de secuenciación o mapeo como el desequilibrio en los niveles de expresión de algunos alelos, y finalmente el genotipado de un conjunto de SNPs previamente seleccionados.

    1.1. Mapear los reads en el genoma: a partir de los datos de RNA-seq, obtener un archivo BAM para cada individuo como resultado de mapear los reads con el software seleccionado [5].

    1.2. Filtrar los datos: en esta parte se van a filtrar los datos de los reads mapeados para eliminar duplicados y para eliminar aquellos reads que mapean en más de un sitio [6], y también se van a filtrar los SNPs para eliminar aquellos que están en zonas de repeticiones [6] o en genes improntados [4], así como aquellos SNPs que no tengan una buena calidad de genotipado, demasiado pocos reads por SNP, o una frecuencia del alelo menos común demasiado baja [7]. Los

criterios de filtrado se irán modificando dependiendo de la calidad de los resultados obtenidos.

2. Reconstruir el pedigrí: una vez obtenidos los genotipos de los SNPs, se procederá a representar las relaciones de parentesco que puedan existir entre los diferentes individuos.

   2.1. Representar las relaciones entre parejas de individuos: para cada pareja de individuos con relación de parentesco conocido se va a representar la probabilidad de compartir 0, 1 y 2 alelos idénticos por descendencia (IBD) [8]. Los resultados serán evaluados para decidir si es necesario modificar los criterios de filtrado, con el fin de mejorar el ajuste del modelo.

   2.2. Representar el pedigrí completo: representar las relaciones de parentesco entre todas las muestras en un pedigrí [2][3].

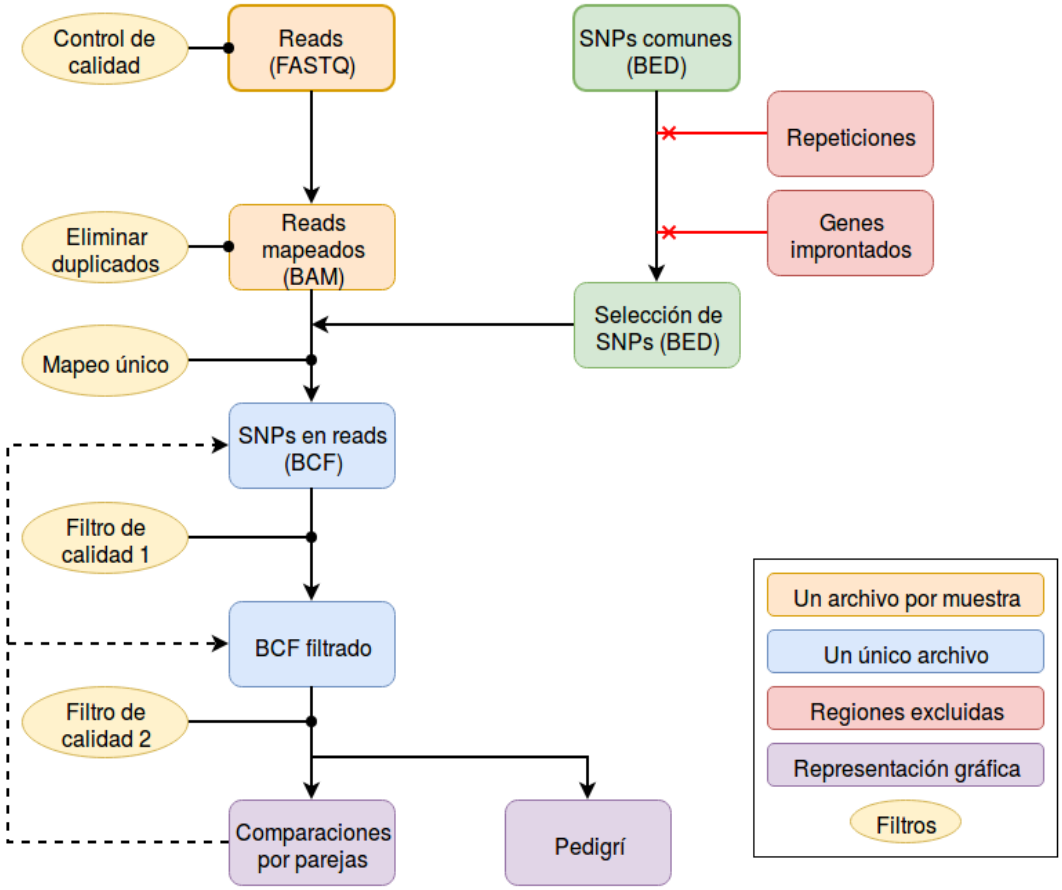

Figura 1: Diagrama de flujo de las diferentes etapas del proyecto.

1.3 Enfoque y método seguido

El método a seguir viene determinado por los programas que se utilizan, por los parámetros de filtrado y por los datos utilizados. A continuación se explican los criterios que se han seguido para la selección de cada uno de ellos:

Programas elegidos:

- TopHat2 [9]: programa elegido para mapear los reads en el genoma. Se ha elegido por su capacidad de mapear datos de RNA-seq, ya que tiene en cuenta el splicing. Existen programas alternativos como STAR [10] o BWA [11] que también tienen esta capacidad, pero los miembros del equipo tienen más experiencia usando TopHat2. Más adelante se puede valorar el uso de estos programas alternativos.
- SAMtools (mpileup/BCFtools call) [12]: Programas elegidos para identificar las variaciones. Existen programas alternativos pero estos son los más comúnmente usados.
- PLINK [13]: programa elegido para determinar la relación entre individuos. Se ha elegido porque es un programa abierto y gratuito, además de ser el más comúnmente usado.
- R [14]: programa elegido para la representación gráfica. Se ha elegido por ser el más comúnmente usado y por la experiencia personal.

Parámetros de filtrado: se empieza con un conjunto de parámetros de filtrado basados en la bibliografía consultada [4][6][7], pero posteriormente van a ser valorados y se cambiarán si es necesario.

Datos elegidos: se han elegido los datos de “CEPH/UTAH family 1463” [5]. Se trata de una gran familia (17 individuos) con relaciones de parentesco conocidas (Figura 2), de la que se tienen datos de RNA-seq (de linfocitos B de sangre periférica) y WGS de todos los individuos (estos últimos sólo son públicos para los abuelos y padres). Posteriormente se utilizará otro conjunto de datos para ver si los parámetros seleccionados son válidos para otros conjuntos de datos. Para dicho propósito utilizaremos datos de 15 ratones macho de 5 familias diferentes (Figura 3) en las que encontramos relaciones de parentesco que no aparecían en el anterior conjunto de datos, como la relación tío-sobrino, mientras que no hay ninguna relación de hermanos completos. Los datos de RNA-seq de dichos individuos proceden de testículo en el caso de F0 y F1, mientras que para la F2 provienen de hígado. Estos datos pertenecen a nuestro grupo.

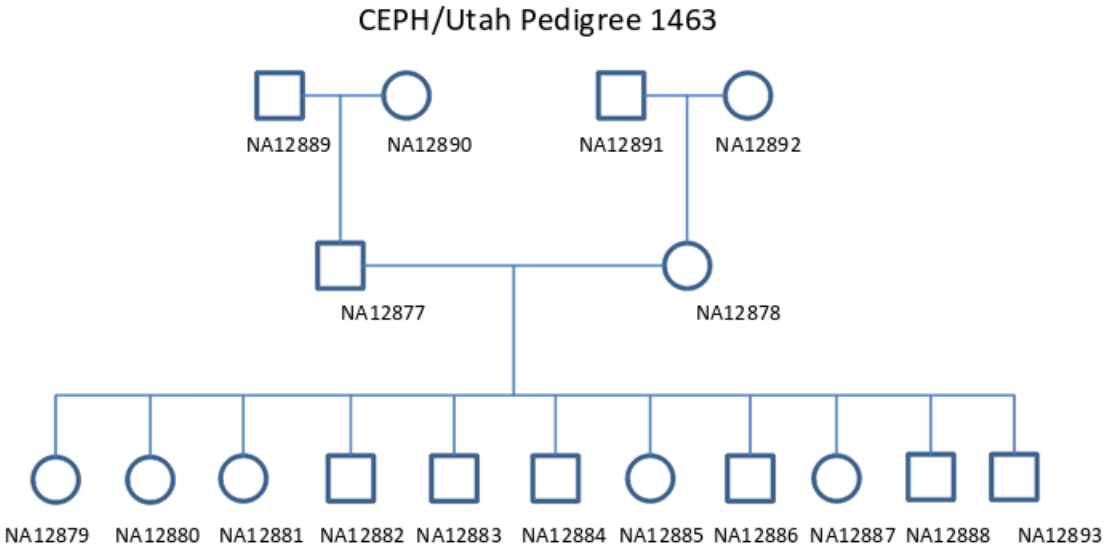

Figura 2: Estructura de la familia “CEPH/UTAH family 1463” [5].

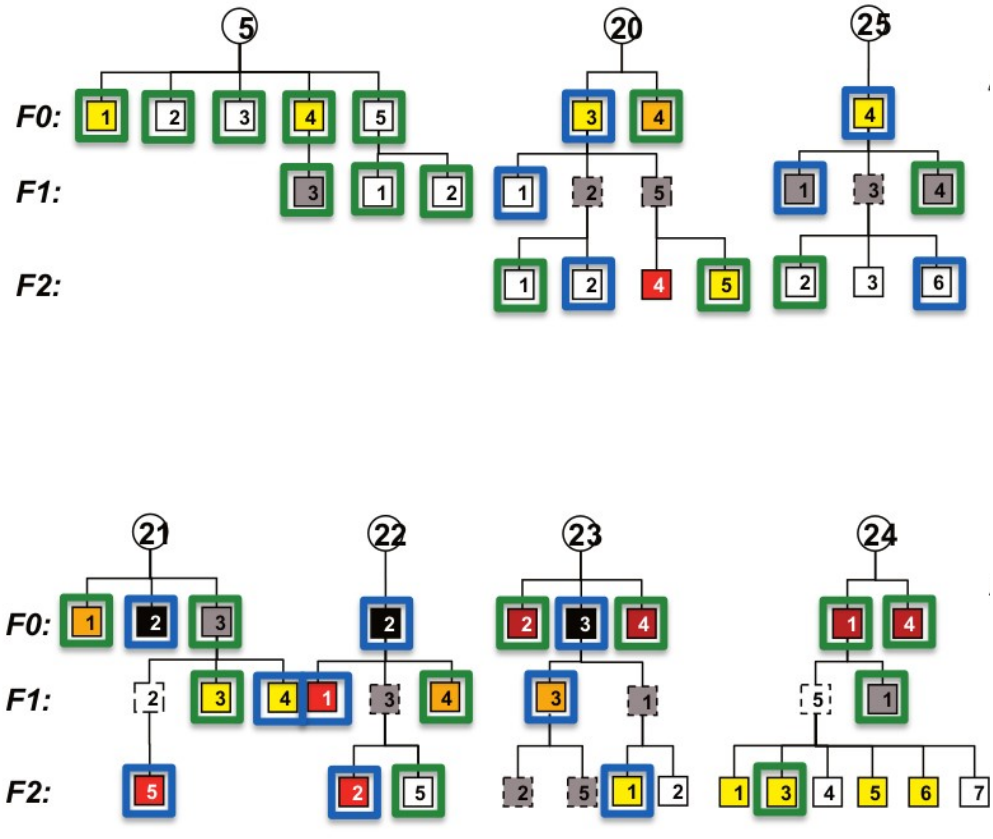

Figura 3: Estructura de las familias de ratones. Los datos de RNA-seq utilizados proceden de los individuos encuadrados en azul.

1.4 Planificación del Trabajo

Se han dividido los objetivos específicos en distintas tareas de la siguiente manera:

1. Determinar el genotipo de SNPs a partir de los datos de RNA-seq

   1.1. Mapear los reads en el genoma
   - Tarea 1: Comprobar la calidad de los datos crudos.
   - Tarea 2: Mapear los reads en el genoma con el programa seleccionado.

   1.2. Filtrar los datos
   - Tarea 3: Filtrar los reads mapeados para eliminar duplicados y reads que mapeen en más de un sitio del genoma, así como los reads que no hayan mapeado.
   - Tarea 4: Hacer una selección de SNPs conocidos para quedarnos con los SNPs comunes que no estén en zonas de repeticiones o en genes improntados.
   - Tarea 5: Filtrar los SNPs en los reads según su calidad, profundidad y frecuencia.

2. Reconstruir el pedigrí

   2.1. Representación de las relaciones entre parejas de individuos
   - Tarea 6: Utilizar un programa para determinar la asociación entre las distintas muestras.
   - Tarea 7: Representar gráficamente la relación entre parejas de individuos y evaluar los datos visualmente.
   - Tarea 8: Decidir si es necesario o no hacer cambios en el proceso de filtrado.

   2.2. Representación del pedigrí completo
   - Tarea 9: Utilizar un programa para hacer la representación gráfica del pedigrí.

Se plantean tres hitos:

1. Tener los reads mapeados: Tener un fichero BAM para cada individuo con los reads de RNA-seq mapeados en el genoma.

2. Tener resultados que concuerden con el pedigrí: Tener los parámetros de filtrado ajustados de tal manera que proporcionen unos resultados que concuerden con las relaciones de parentesco conocidas.

3. Tener la manera de representar gráficamente los datos: Reproducir gráficamente las relaciones de parentesco conocidas a partir de los datos filtrados.

Las tareas, los hitos y los objetivos generales se han calendarizado en el siguiente diagrama (Figura 4). Se ha añadido la Tarea 0, correspondiente al Plan de trabajo. Las tareas 5 a 8 tienen una duración muy extensa, ya que se prevé que se realizarán varias veces hasta conseguir ajustar los parámetros.

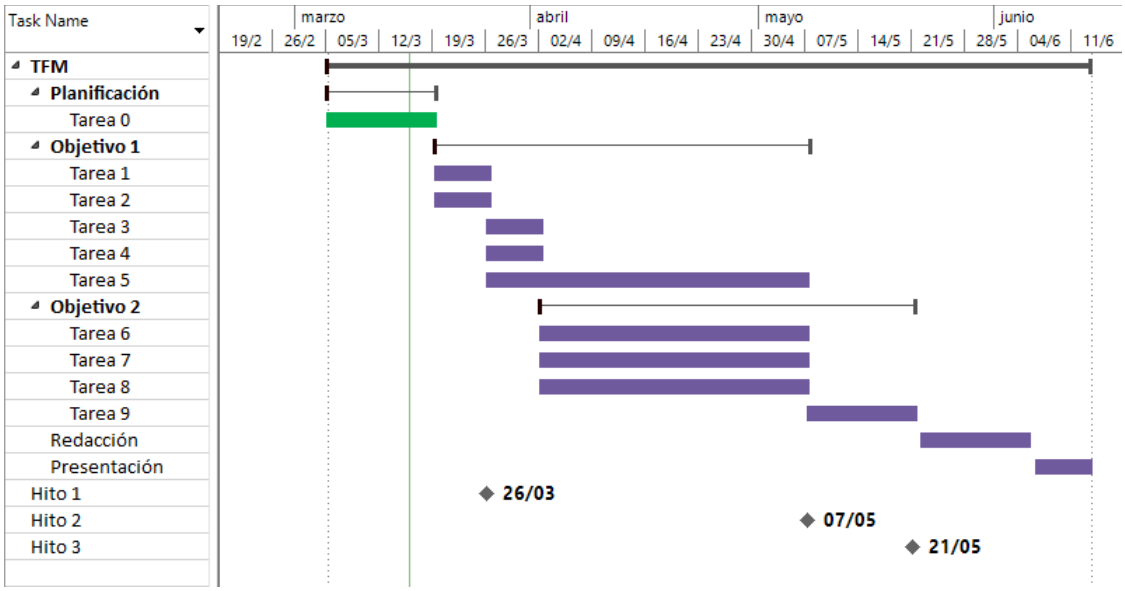

Figura 4: Diagrama de Gantt del proyecto (realizado en MS Project).

A cada una de las PEC le corresponden la siguientes tareas:

- PEC0 - Definición de los contenidos del trabajo: Propuesta de TFM
- PEC1 - Plan de trabajo: Tarea 0
- PEC2 - Desarrollo del trabajo - Fase 1: Tareas 1, 2, 3 y 4, y primera parte de las tareas 5, 6, 7 y 8.
- PEC3 - Desarrollo del trabajo - Fase 2: Fin de las tareas 5, 6, 7 y 8, y tarea 9.
- PEC4 - Redacción de la memoria
- PEC5a - Elaboración de la presentación
- PEC5b - Defensa pública

1.5 Breve sumario de productos obtenidos

- Plan de trabajo: engloba la descripción del TFM, los objetivos, el método elegido, la planificación y la estructuración del proyecto.
- Memoria: documento que contiene la información detallada de todo el trabajo realizado.
- Artículo que describe el método para reconstruir un pedigrí con datos de RNA-seq.
- Presentación virtual: contiene información del trabajo realizado de una forma muy resumida, que servirá de soporte para la defensa pública del TFM.
- Autoevaluación del proyecto: informe que evalúa el cumplimiento de los requisitos.

1.6 Breve descripción de los otros capítulos de la memoria

Métodos: en este capítulo se hace una descripción detallada de los programas utilizados, las opciones usadas en cada uno de ellos y el producto que se obtiene en cada uno de los pasos.

Resultados: se explican y/o muestran los resultados obtenidos en cada apartado y los criterios que se han tenido en cuenta a la hora de escoger los filtros.

Discusión: se hace una valoración de los resultados y se plantean hipótesis de porqué se han obtenido ciertos resultados que difieren con lo esperado a priori.

Conclusiones: se hace una valoración del trabajo global, así como de la manera en la que se ha llevado a cabo y se mencionan las futuras perspectivas.

Glosario: se definen los términos y acrónimos que se han utilizado en la memoria y no han sido explicados.

Bibliografía: en este capítulo se proporciona una relación de las referencias bibliográficas citadas en el texto.

Anexos: se incluyen los scripts utilizados para poder reproducir el método.

2. Métodos

2.1. Obtención de datos

Los datos SRA de los 17 individuos CEPH/UTAH family 1463 se obtienen de NCBI [15] mediante la herramienta GNU Wget (versión 1.16) [16]. Una vez descargados los datos SRA, se han obtenido los ficheros FASTQ mediante la herramienta fastq-dump (versión 2.8.1) de SRA Toolkit [17], con la opción para reads emparejados, obteniendo dos ficheros para cada muestra, uno para las primeras parejas y uno para las segundas.

Los datos de las familias de ratones pertenecen a nuestro grupo y partimos de ficheros FASTQ emparejados.

2.2. Control de calidad de los datos crudos

El control de calidad de los datos crudos (ficheros FASTQ emparejados) se ha llevado a cabo con el programa FastQC (versión 0.11.4) [18], obteniendo un informe de control de calidad para cada fichero FASTQ (dos informes por individuo).

Cuando la calidad de las bases no es buena se aplica un filtro que consiste en cortar las bases finales que no lleguen a una calidad mínima (trimming). Esto se lleva a cabo con la herramienta FASTQ Quality Trimmer de FASTX Toolkit (versión 0.0.14) [19]. La calidad mínima se establece en 20 y también se eliminan aquellos reads con una longitud inferior a 20 bases después de realizar el trimming, para evitar tener reads demasiado cortos que ralentizarían el proceso de mapeado.

Para determinar la calidad y necesidad de este filtro, se mantienen también los datos crudos (sin filtrar por la calidad de las bases) en el proceso para poder hacer comparaciones.

2.3. Mapeo de los reads en el genoma

Para mapear los reads en el genoma se utiliza el programa TopHat (versión 2.1.0) [9] por su capacidad de mapear datos de RNA-seq, ya que tiene en cuenta el splicing. Para los datos humanos se utiliza el genoma de referencia hg19, mientras que para los datos de ratones el mm10. Tras el mapeo se obtiene un fichero BAM para cada individuo.

2.4. Eliminación de duplicados

Para eliminar aquellos reads idénticos que provienen de la amplificación por PCR y podrían dificultar el proceso de genotipado, se utiliza el comando

markdup de samtools (versión 1.7) [12], obteniendo un fichero BAM para cada individuo sin los reads duplicados.

2.5. Eliminación de reads con mapeo múltiple

Para identificar aquellos reads que mapean en un único sitio hay que fijarse en la calidad de mapeo (MAPQ). En la versión utilizada de TopHat, los reads con mapeo único tienen una MAPQ de 50, mientras que los reads con mapeo múltiple tienen una MAPQ de 0, 1 o 3 dependiendo de cuantas veces mapeen [20].

Para eliminar los reads con mapeo múltiple se ha utilizado una opción del comando mpileup de samtools (versión 1.7) [12] que elimina aquellos reads con MAPQ < 4 (ver sección 2.7.).

2.6. Selección de SNPs

A partir de un conjunto de SNPs inicial, se eliminan aquellos SNPs que están en una zona de repeticiones o en un gen improntado, con el subcomando intersect de bedtools (versión 2.26.0) [21].

El archivo de los SNPs ha sido obtenido de UCSC Genome Browser [22], seleccionando la última versión de "Common SNPs" (SNP150 en humanos y SNP142 en ratones) del assembly correspondiente a cada especie. El archivo de las repeticiones ha sido obtenido del RepeatMasker de UCSC Genome Browser. Los genes improntados se han obtenido de geneimprint [23], seleccionando los genes humanos o de ratones con "Status: Imprinted" e introduciéndolos en Biomart para obtener las regiones de los genes improntados. Tras este proceso se obtiene un archivo BED con los SNPs comunes que no están en zonas de repeticiones ni en genes improntados.

2.7. Obtención de SNPs en los reads

Se cruzan los datos de los reads mapeados y filtrados con los datos de los SNPs seleccionados para obtener aquellos SNPs que se encuentran en los reads. Esto se lleva a cabo con comando mpileup de samtools (versión 1.7) y con el comando call de bcftools (versión 1.4) [12], y se obtiene un fichero BCF por conjunto de datos.

Para evaluar la eficacia y/o necesidad del trimming, se incluye una opción en samtools mpileup para no tener en cuenta las bases con una calidad inferior a 20, en los datos en los que no se ha realizado el trimming, manteniendo también los datos sin filtrar por calidad de base.

2.8. Filtro de SNPs en los reads

Para una determinación de los genotipos más precisa, se filtran los SNPs de cada individuo por profundidad (DP) y calidad del genotipo (GQ). Este filtro se lleva a cabo con el programa vcftools (versión 0.1.13) [24]. Diferentes valores de DP y GQ han sido usados para poder hacer comparaciones. Se obtiene un archivo BCF donde los SNPs que no superen el filtro de calidad serán clasificados como genotipo faltante para ese individuo.

2.9. Determinación de las relaciones entre individuos

Para determinar la asociación entre parejas de individuos se usa el programa PLINK (versión 1.90b3.29) [13], que calcula z0, z1 y z2 (la probabilidad de que dos individuos compartan 0, 1 o 2 alelos idénticos por descendencia). Posteriormente estos parámetros son representados gráficamente en R (Figura 5) [8][14]. En PLINK se utilizan diferentes filtros con distintos valores para poder hacer comparaciones, se filtra por frecuencia del alelo menos común, desequilibrio de ligamiento y por proporción de genotipos faltantes por SNP (ya sea porque el individuo no tiene reads en ese SNP o porque no ha pasado el filtro de DP y GQ).

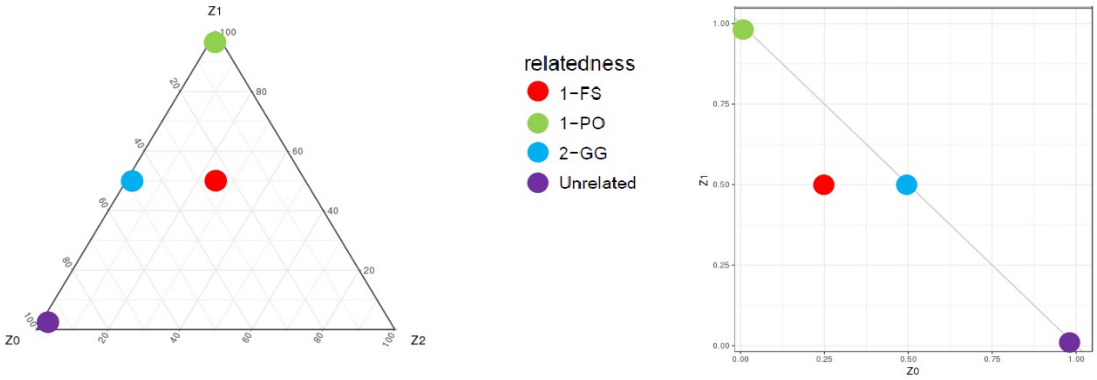

Figura 5: Gráficos usados para representar las relaciones entre parejas de individuos. A la izquierda, diagrama ternario en el que se representan los tres parámetros z0, z1 y z2, a la derecha, gráfico z0-z1. Los puntos representan los valores teóricos de z0, z1 y z2 de cada relación de parentesco (1-FS: hermanos completos, 1-PO: padre-hijo, 2-GG: abuelo-nieto, Unrelated: no relacionados).

2.10. Representación del pedigrí

La representación gráfica del pedigrí se ha llevado a cabo con el paquete sequoia [2] y kinship2 [25] de R [14]. El primero permite determinar los padres de cada individuo y el segundo utiliza estos datos para dibujar el pedigrí.

3. Resultados

3.1. Resultados de los datos de CEPH/UTAH family 1463

3.1.1. Control de calidad de los datos crudos

El control de calidad de los datos crudos con el programa FastQC muestra resultados compatibles con la normalidad o con datos de RNA-seq en la mayoría de pruebas, excepto en la calidad de la secuencia por base (Figura 6) y en el nivel de secuencias duplicadas (Figura 7).

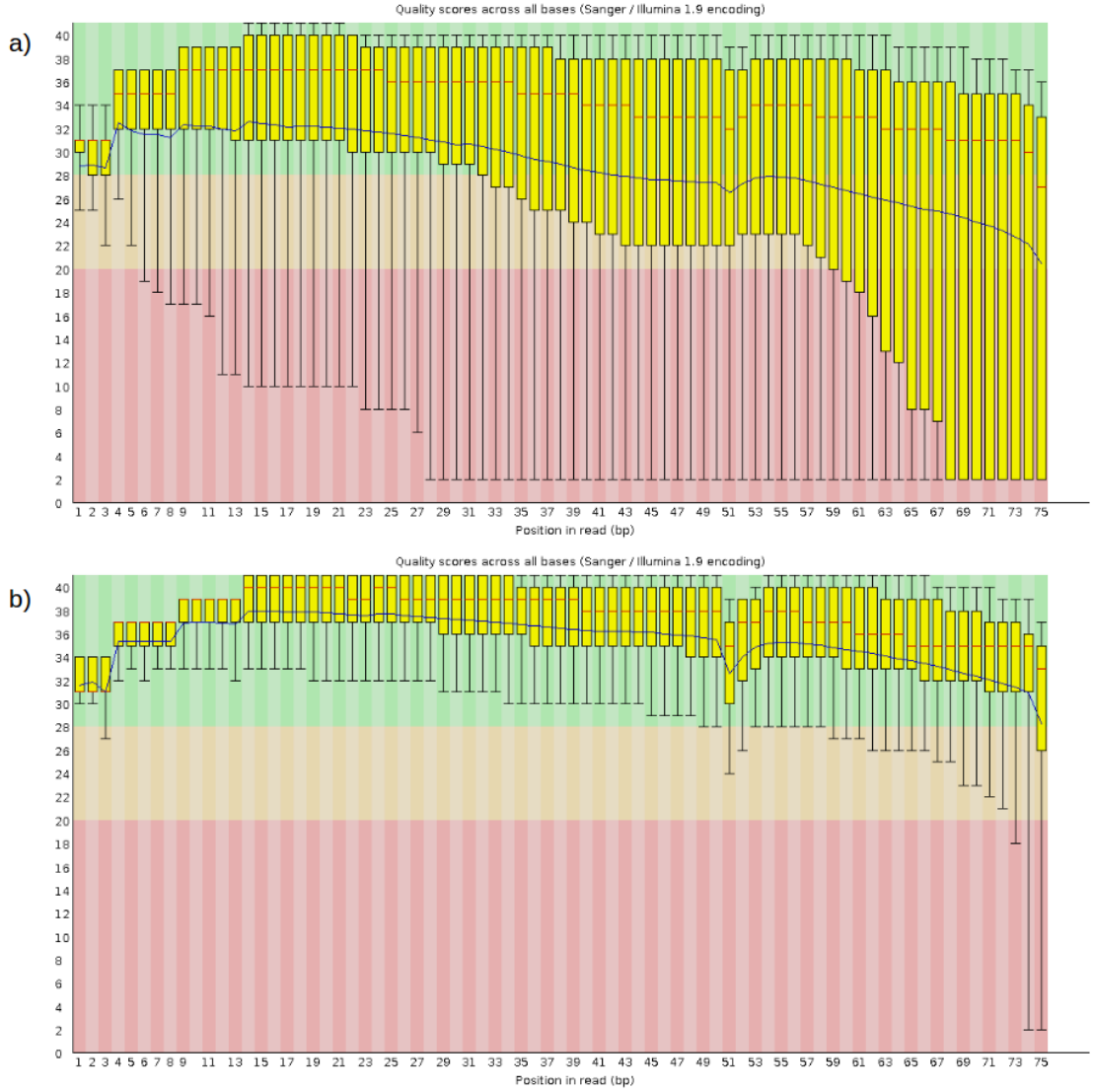

Figura 6: Calidad de las secuencias por base. En el eje x se encuentra la posición de los reads en pares de bases, y en el eje y la calidad de las bases [18].

En la calidad de la secuencia por base, se produce un descenso marcado de la calidad de las bases al final del read en algunas muestras (Figura 6a), mientras

que en otras muestras, casi todas las bases tienen una calidad por encima de 20 (Figura 6b).

Por lo que respecta al nivel de secuencias duplicadas (Figura 7), aproximadamente el 50% de las secuencias tienen algún duplicado, encontrando secuencias con un bajo nivel de duplicación (sólo un duplicado) y otras con más de 100 duplicados. Los resultados son parecidos en todas las muestras. Este nivel de secuencias repetidas podría alterar el proceso de genotipado magnificando posibles errores de secuenciación. Por este motivo se eliminarán las secuencias duplicadas.

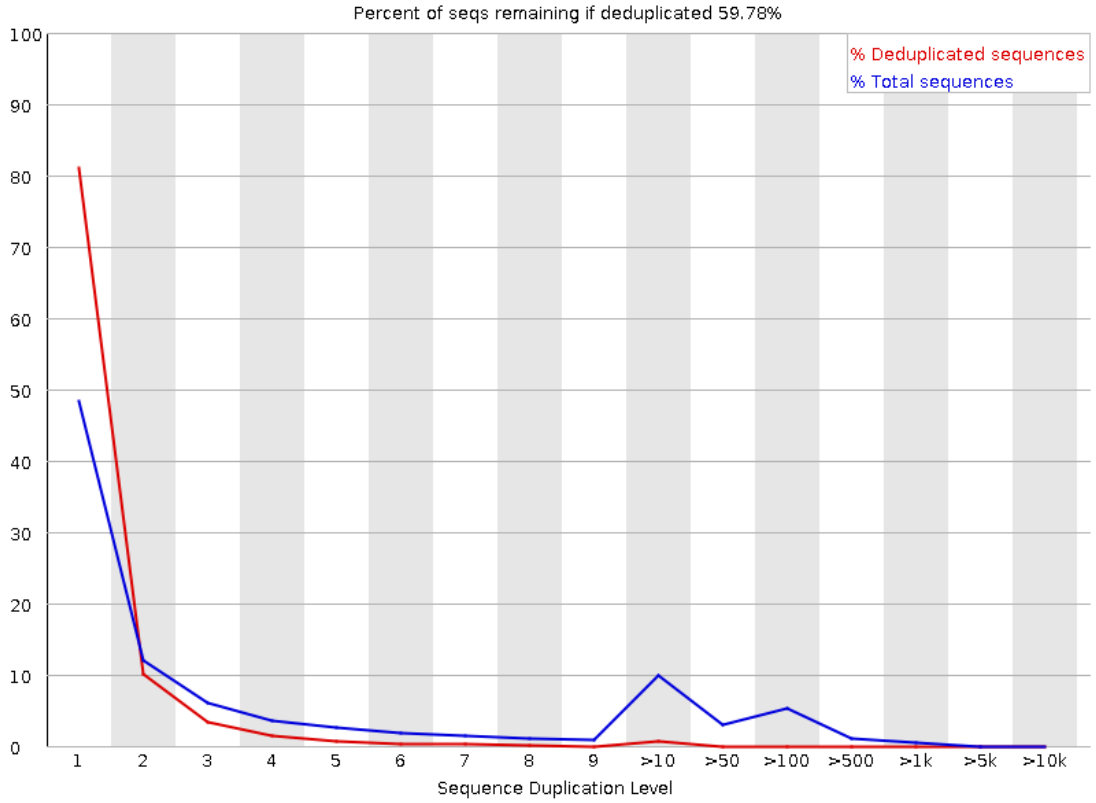

Figura 7: Nivel de secuencias duplicadas, en el eje x se encuentra el número de veces que aparece cada secuencia, y en el eje y el porcentaje de secuencias [18].

3.1.2. Mapeo de los reads en el genoma

En el proceso de mapeo con TopHat2 se eliminan entre un 9 y un 34% de los reads (Figura 8), ya que estos no consiguen mapear en el genoma de referencia.

3.1.3. Eliminación de duplicados

El proceso de eliminación de duplicados con el comando markdup de samtools, elimina entre un 18 y un 33% de los reads (Figura 8).

3.1.4. Eliminación de reads con mapeo múltiple

El proceso de eliminación de reads con mapeo múltiple descarta entre un 4 y un 6% de los reads (Figura 8).

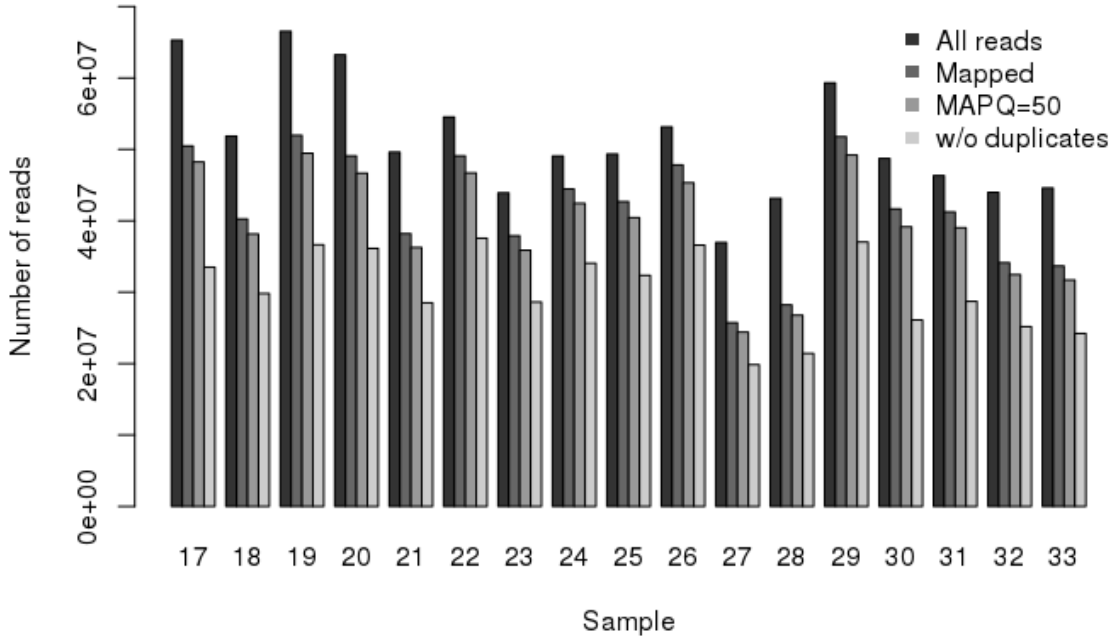

Figura 8: Número de reads antes y después de los diferentes filtros [14]. Cada filtro se aplica a los reads que han pasado el filtro anterior. All reads: reads iniciales, Mapped: reads que han mapeado en el genoma, MAPQ = 50: reads con mapeo único, w/o duplicates: reads sin duplicados.

Una vez aplicados todos los filtros, se mantienen entre un 50 y un 70% de los reads iniciales.

3.1.5. Selección de SNPs

El fichero inicial de SNPs comunes de la anotación snp150 de hg19, contenía 14.8 millones de SNPs, de los cuales, en el proceso de selección, se eliminan 8.2 millones de SNPs en repeticiones o en genes improntados (Figura 9).

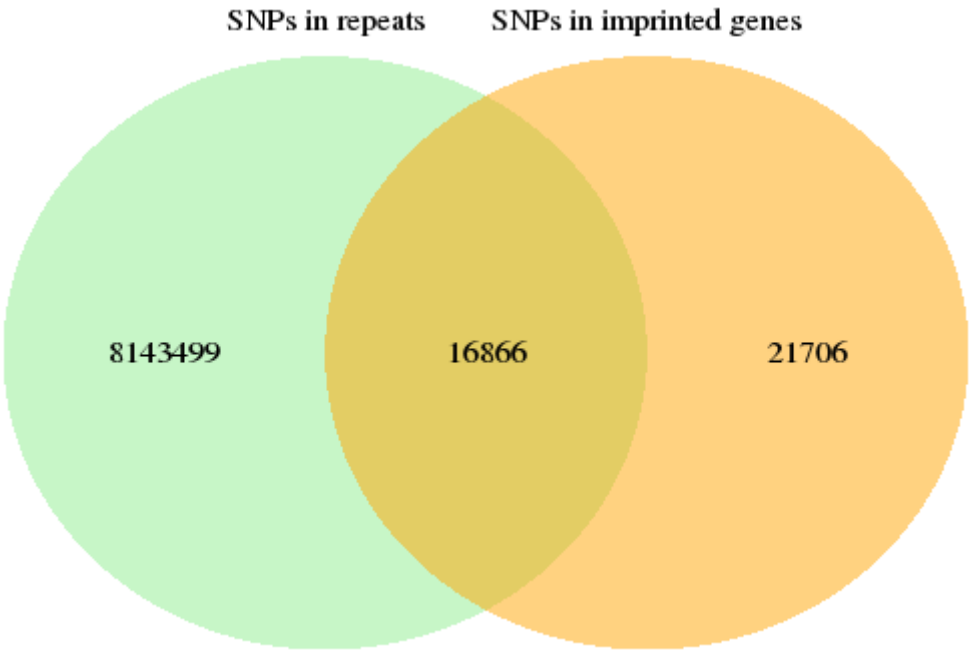

Figura 9: Diagrama de Venn de los SNPs eliminados según su localización [14].

3.1.6. Obtención de SNPs en los reads

De los 6.2 millones de SNPs que han pasado la selección, 2.6 millones se encuentran en los reads, lo que era de esperar, ya que la selección de SNPs se encuentra localizada por todo el genoma, mientras que nuestros datos son de mRNA.

3.1.7. Filtro de SNPs en los reads

Distintos parámetros de filtrado han sido considerados para probar su eficacia y en algunos parámetros, se han probado distintos valores de corte para encontrar el que permita determinar las relaciones de parentesco de una manera más precisa.

3.1.7.1. Profundidad (DP) y calidad del genotipo (GQ)

La profundidad indica el número de reads filtrados que se encuentran en cada SNP, mientras que la calidad del genotipo indica la probabilidad de que el genotipo sea correcto en escala Phred [26].

Para decidir los posibles valores de corte para cada uno de estos parámetros, se observa la relación que existe entre ellos (Figura 10), observando que existe una relación lineal, con pendientes distintas para los genotipos homocigotos y para los heterocigotos (Figura 10a). La mayoría de puntos se encuentran en el

origen de coordenadas (Figura 10b), indicando que para este individuo no existen reads en dicho SNP.

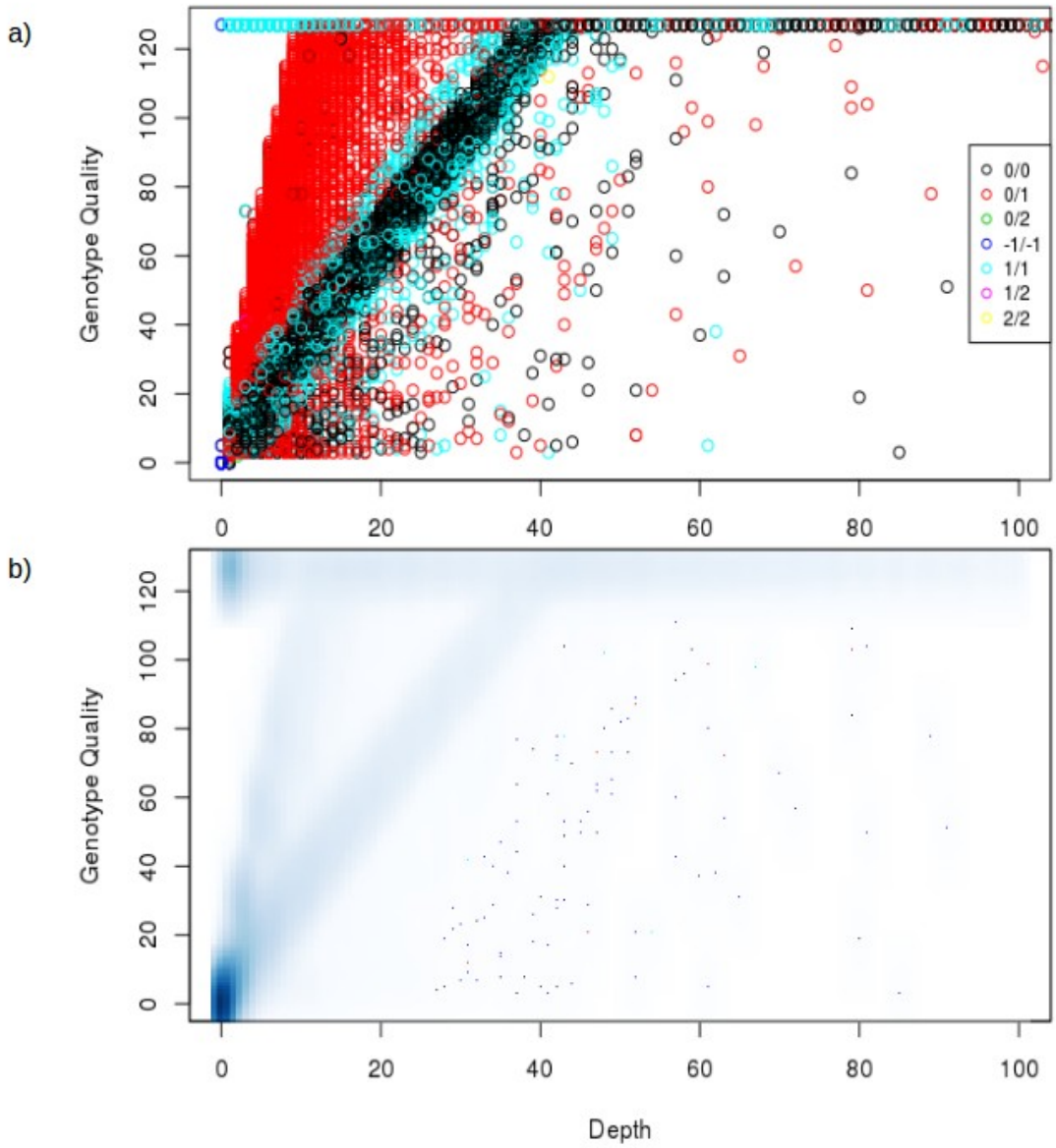

Figura 10: Gráfico de dispersión (a) y de dispersión suavizada (b) de la profundidad y la calidad del genotipo de los SNPs en los reads. Los diferentes colores representan los posibles genotipos (0/0: homocigoto para el alelo de referencia, 0/1: heterocigoto, 0/2: heterocigoto con un alelo de referencia y el segundo alelo alternativo, -1/-1: genotipo faltante, 1/1: homocigoto por el alelo alternativo, 1/2: heterocigoto con un alelo alternativo y el segundo alelo alternativo, 2/2: homocigoto para el segundo alelo alternativo) [14].

Para tener una idea de cuantos SNPs pasarían el filtro para cada una de las posibles combinaciones de profundidad y calidad del genotipo, se representan estas combinaciones (Figura 11), viendo que la calidad del genotipo tiene mucho efecto en profundidades bajas, mientras que pierde efecto a partir de DP 15. Las calidades de filtrado 10, 20, 30 y 40 representan un 10, 1, 0.1 y 0.01% de error respectivamente [26].

Una vez aplicado el filtro por DP y GQ, los SNPs que no pasen el filtro se les asigna genotipo faltante.

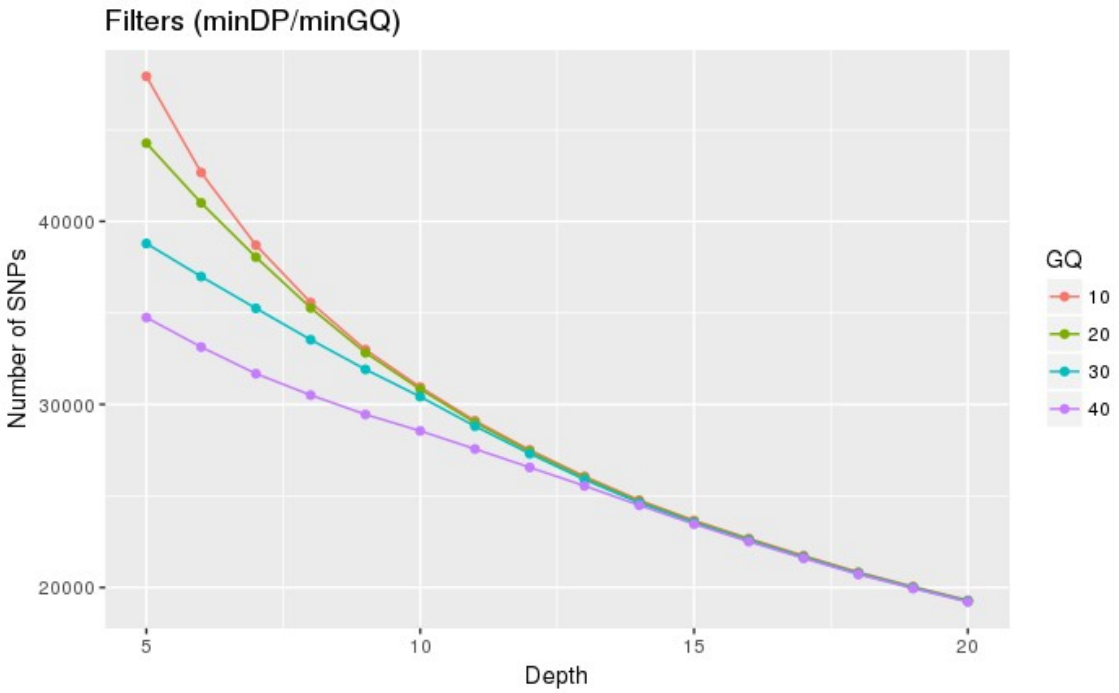

Figura 11: Número de SNPs que pasan el filtro de profundidad mínima y calidad del genotipo mínima. Los SNPs en los que no se encuentra el alelo alternativo en ninguno de los individuos no tienen etiqueta de GQ y por tanto no han sido representados [14].

3.1.7.2. Frecuencia del alelo menos común (MAF) y genotipos faltantes

Para evitar errores de secuenciación y variaciones poco comunes que no nos proporcionarían información de parentesco se decide aplicar un filtro para el alelo menos común. El problema con este parámetro es que se calcula en base a los genotipos existentes, pudiendo obtener valores elevados en SNPs con pocos individuos genotipados (ya sea porque los otros individuos no tienen reads en ese SNP o porque no han pasado el filtro de DP y GQ).

El 74% de los SNPs tienen una MAF = 0, seguidos de lejos por los SNPs con MAF = 0.5, que representan un 4% del total de los SNPs (Figura 12a). Si tenemos en cuenta el número de genotipos faltantes (Figura 12b) vemos como un porcentaje elevado de los SNPs con MAF = 0.5 tienen un solo individuo genotipado.

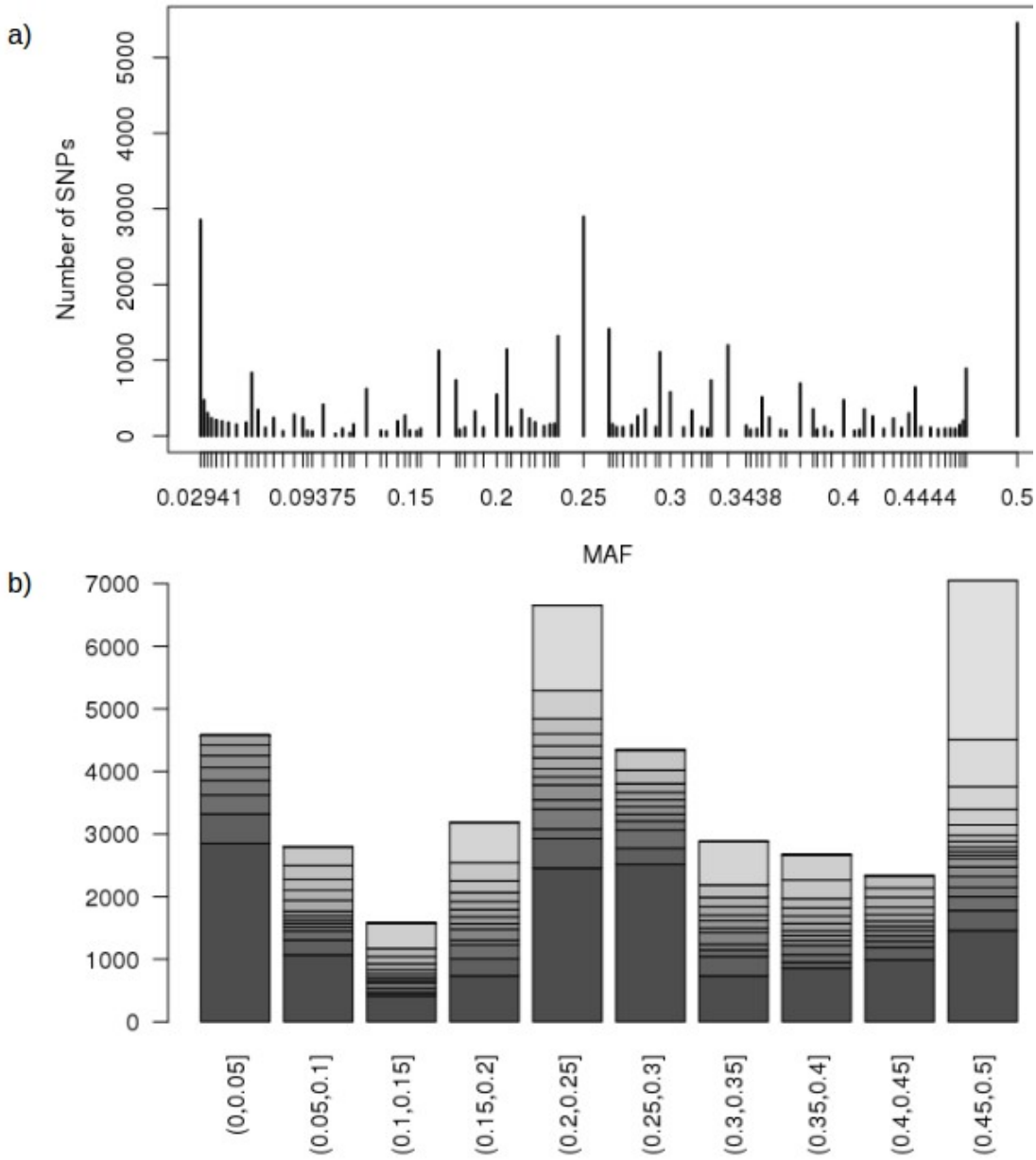

Figura 12: Distribución de los SNPs según su MAF (los SNPs con MAF = 0 han sido excluidos). En la figura (a) se representan todas las MAF existentes, mientras que en la figura (b) se agrupan por intervalos y se representan los distintos genotipos faltantes apilados (el color más oscuro indica que todos los individuos han sido genotipados, mientras que el color más claro indica que solo se ha genotipado un individuo) [14].

3.1.7.3. Desequilibrio de ligamiento

Los SNPs afectados por desequilibrio de ligamiento no nos proporcionan información adicional respecto al SNP con el que están ligados, ya que al no producirse recombinación funcionan como un bloque, y el exceso de SNPs ligados podría alterar los resultados de parentesco.

El parámetro usado para determinar el desequilibrio de ligamiento entre parejas de SNPs es r^2, el cuadrado de la correlación de las frecuencias alélicas [27], en el que los valores cercanos a 0 indican equilibrio de ligamiento y los valores cercanos a 1 indican que los dos SNPs están en desequilibrio de ligamiento.

Al analizar los datos se observa como la mayoría de SNPs están en total desequilibrio de ligamiento (Figura 13).

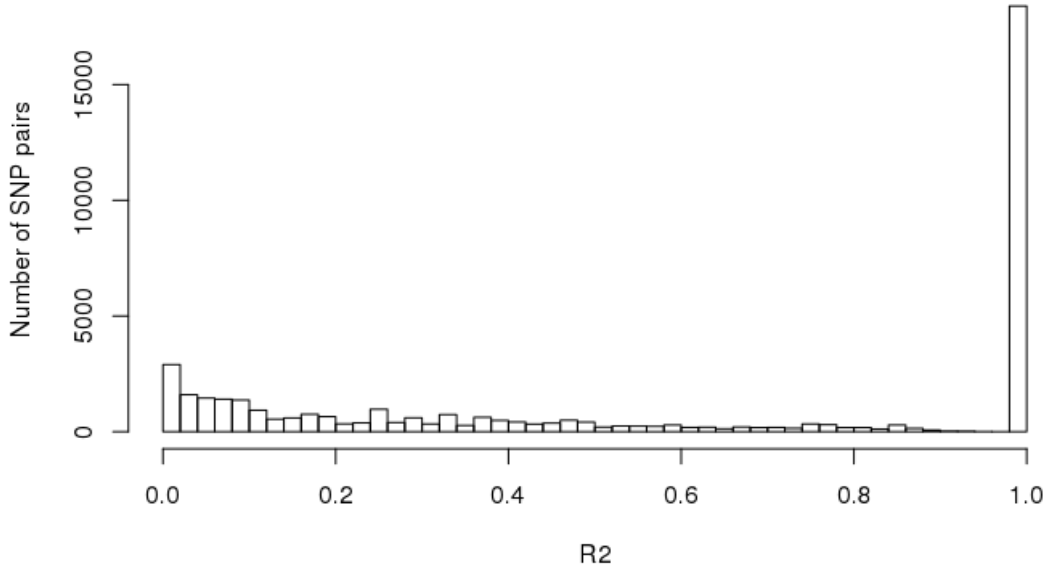

Figura 13: Histograma de r^2 de desequilibrio de ligamiento [14].

3.1.8. Determinación de las relaciones entre individuos

Las relaciones de parentesco entre individuos se obtuvieron del pedigrí y de la conversión de las identificaciones de RNA-seq (SRR1258217 – SRR1258233) con las identificaciones de los individuos (NA12877 – NA12893) [28].

Al representar en los resultados de z0, z1 y z2 (Figura 14), no se observa una diferenciación clara entre los distintos parentescos, y llama la atención las dos nubes de puntos que se producen en el primer gráfico (Figura 14a) tanto para hermanos completos como para abuelos-nietos. En los otros dos gráficos (Figura 14 b y c) vemos como existen diferentes agrupaciones de puntos pero no corresponden a las relaciones familiares establecidas.

Al representar la probabilidad que dos individuos compartan dos alelos idénticos por descendencia (z2) de las parejas identificadas como hermanos completos (Figura 15), se ve claramente que existen dos grupos, uno alrededor de 0, propio de otro tipo de relaciones de parentesco como padre-hijo, abuelo-nieto o no relacionados, y otro alrededor de 0.25, propio de hermanos completos (Tabla 1).

a) b) c)

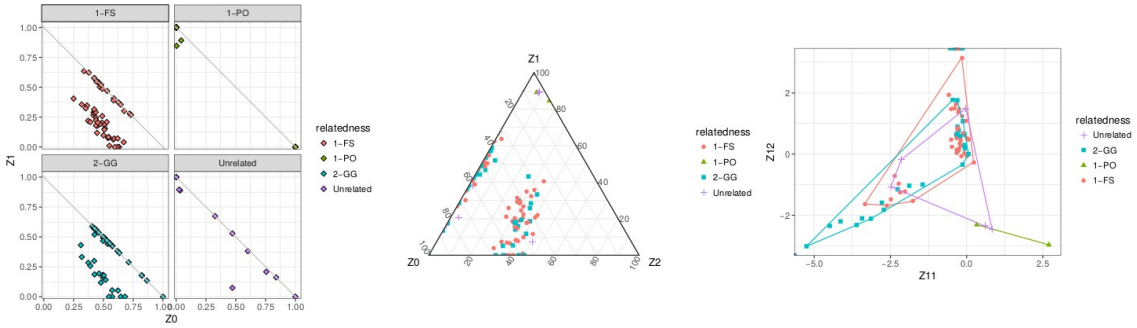

Figura 14: Gráficos de parentesco. (a) Gráfico z0-z1 para cada una de las relaciones de parentesco (1-FS: hermanos completos, 1-PO: padre-hijo, 2-GG: abuelo-nieto, Unrelated: no relacionados), (b) diagrama ternario en el que se representan los tres parámetros z0, z1 y z2, (c) Ilr-plot en el que se representa Z11 y Z12, permitiendo estimar la distancia euclidiana entre parejas de individuos [8][14]. En este caso se usa el filtro por MAF ≥ 0.4.

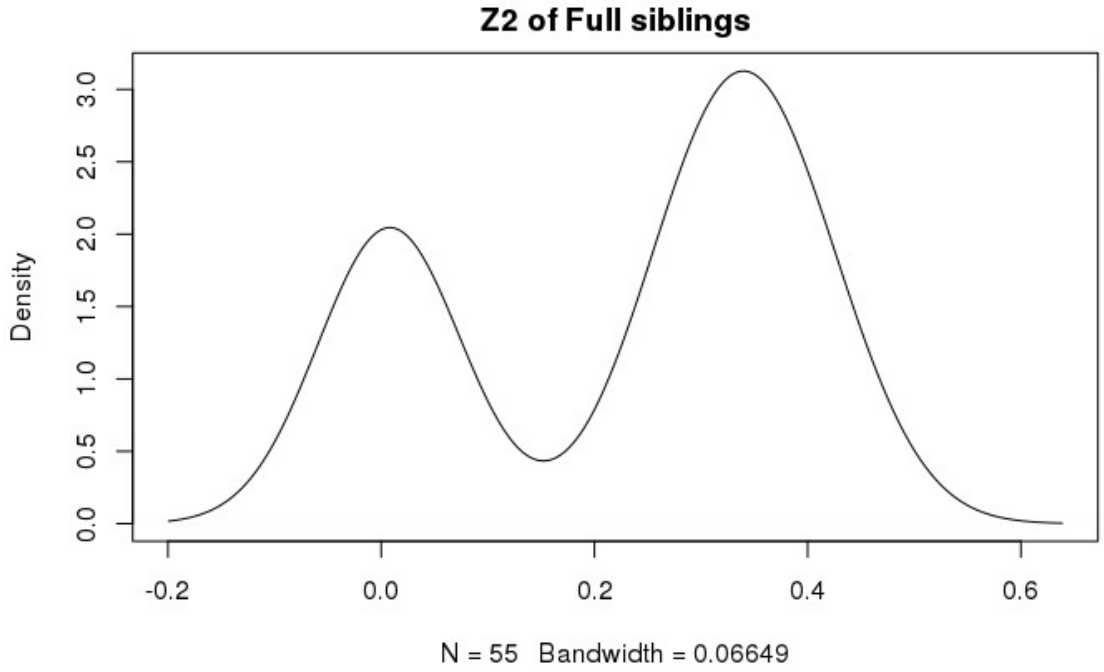

Figura 15: Gráfico de densidad de z2 para hermanos completos [14].

Al ver estos resultados se plantea la posibilidad de un mal etiquetado de las muestras o una mala conversión de las identificaciones de RNA-seq con las identificaciones de los individuos. Para comprobarlo se comparan las muestras RNA-seq con las muestras de WGS en los individuos en las que se encuentran disponibles (abuelos y padres) con un programa [29], y se detecta que algunas de las muestras no corresponden al individuo esperado. Por otro lado, al examinar el número de reads en el cromosoma Y, también se detectan incongruencias con el sexo esperado.

Con toda esta información se vuelven a establecer las relaciones de parentesco entre individuos y se realiza de nuevo el análisis de asociación

entre distintas muestras, obteniendo esta vez resultados consistentes (Figura 16). Los valores de Z11 y Z12 (Figura 16c) se obtienen mediante una fórmula y si alguno de los valores de z0, z1 o z2 es igual a 0, se obtiene un valor de ±infinito, por lo que estas parejas no aparecen en el gráfico.

a) b) c)

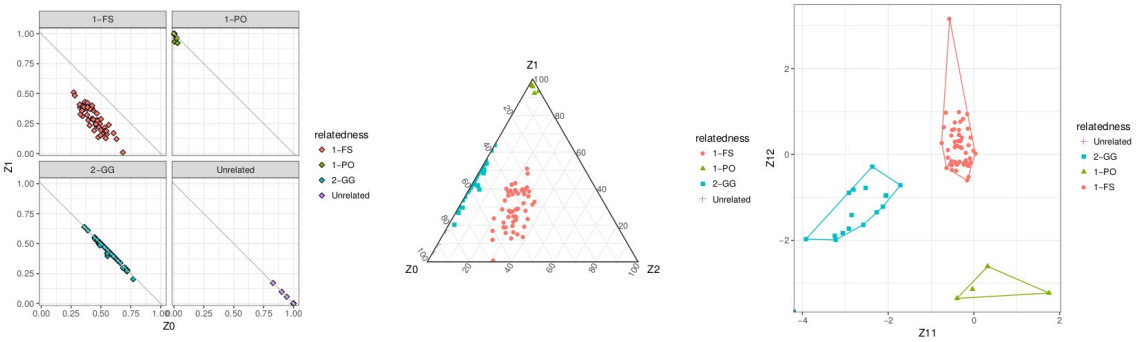

Figura 16: Gráficos de parentesco [8][14] (a) Gráfico z0-z1 para cada una de las relaciones de parentesco, (b) diagrama ternario, (c) llr-plot. Se han comparado entre 2365 y 3756 SNPs. Se muestran los resultados con filtros seleccionados (ver apartado 3.1.9.)

Se han aplicado diferentes filtros y combinaciones de estos con diferentes valores para evaluar como cambian los valores de z0, z1 y z2 y como se ajustan a los valores teóricos (Tabla 1).

Type of relative	Degree	k_0	k_1	k_2
Monozygotic twins (MZ)	0	0	0	1
Parent-offspring (PO)	1	0	1	0
Full-siblings (FS)	1	1/4	1/2	1/4
Half-siblings (HS)/avuncular (AV)/grandchild-grandparent (GG)	2	1/2	1/2	0
First cousins (FC)	3	3/4	1/4	0
Unrelated (UN)	∞	1	0	0

Tabla 1: Coeficientes de Cotterman para los diferentes tipos de relaciones familiares [8].

3.1.9. Evaluación y selección de filtros

Por lo que hace a la calidad de las bases, no se aprecian diferencias entre el trimming, la opción de mpileup para no tener en cuenta las bases con baja calidad y no aplicar ningún filtro. Por esta razón hemos decidido no aplicar ningún filtro ya que es el método más sencillo, aunque se debe tener en cuenta esta baja calidad y en caso que los resultados no fueran satisfactorios se podría aplicar alguno de los dos filtros.

Para decidir el valor de corte para la profundidad (DP) y la calidad del genotipo (GQ), tenemos en cuenta la relación lineal entre ambas (GQ = 2DP) y en el

error que representa cada valor de GQ (Tabla 2). Un GQ de 20 es suficiente, ya que la probabilidad que el genotipo sea erróneo es del 1%. Debemos tener en cuenta que si aumentamos el valor de corte de GQ aumentamos la precisión pero reducimos el número de SNPs que nos van a quedar para determinar la asociación entre individuos. Con toda esta información, decidimos utilizar el filtro DP 10 y GQ 20.

Phred Quality Score	Error	Accuracy (1 - Error)
10	1/10 = 10%	90%
20	1/100 = 1%	99%
30	1/1000 = 0.1%	99.9%
40	1/10000 = 0.01%	99.99%
50	1/100000 = 0.001%	99.999%
60	1/1000000 = 0.0001%	99.9999%

Tabla 2: Relación entre GQ y error [26].

Hemos probado diferentes valores de corte para la MAF (Figura 17) y hemos observado que con valores bajos, la separación entre grupos de parentesco no es suficiente. Para la relación de hermanos completos, la separación de otras relaciones es directamente proporcional a la MAF (Figura 17a), mientras que la separación entre abuelos-nietos e individuos no relacionados esta inversamente relacionada con la MAF (Figura 17b). Teniendo en cuenta esto, escogemos un valor de MAF intermedio (MAF ≥ 0.3), suficientemente alto como para poder diferenciar a los hermanos completos pero suficientemente bajo para poder diferenciar a los abuelos-nietos de los individuos no relacionados.

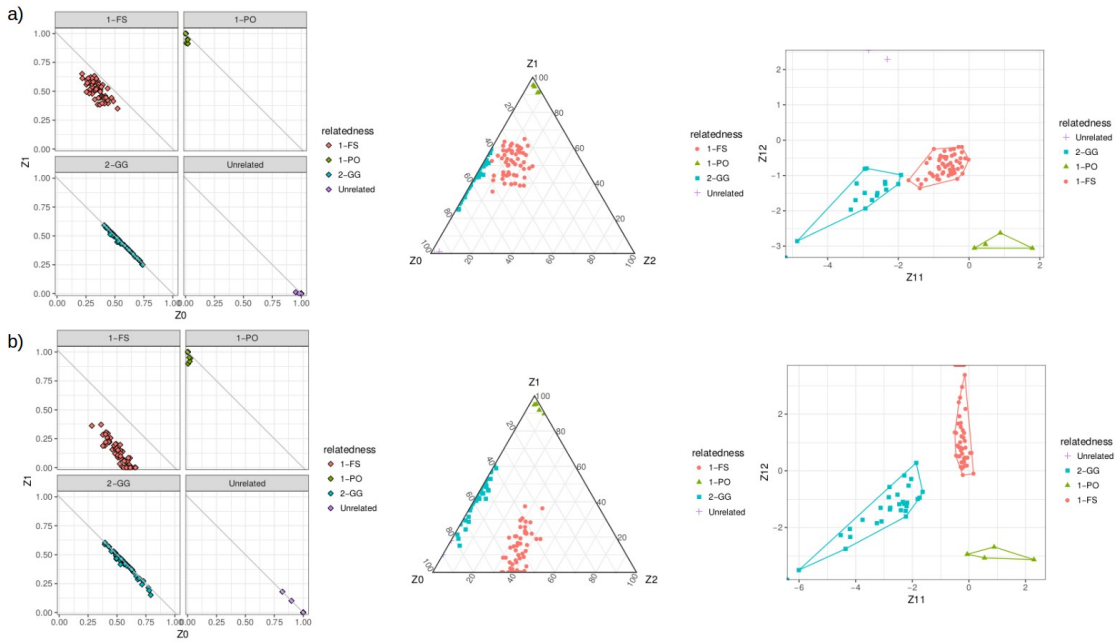

Figura 17: Gráficos de parentesco [8][14]. (a) Con filtro de MAF ≥ 0.2 y (b) con filtro de MAF ≥ 0.4.

Decidimos incorporar el filtro de desequilibrio de ligamiento para evitar los SNPs que no nos aportan información. Hemos probado diferentes parámetros para r^2 y decidimos usar el valor de corte 0.2, que nos indica que los SNPs están en equilibrio de ligamiento o muy poco ligados.

Al aplicar el filtro de genotipos faltantes juntamente con el filtro de desequilibrio de ligamiento nos quedamos con muy pocos SNPs para determinar la relación entre individuos. Por otra parte, si el filtro de MAF es superior a 0.25, nos aseguramos que no vamos a encontrar ningún SNP en el que sólo un individuo tenga un alelo menos común (si sólo dos individuos tienen el SNP, y sólo uno de ellos tiene un alelo menos común, la MAF será de 0.25 y por tanto no pasará el filtro, mientras que si sólo un individuo tiene el SNP no se usará para hacer comparaciones entre parejas de individuos). Por esta razón se decide no aplicar el filtro de genotipos faltantes.

3.1.10. Representación del pedigrí

Al representar el pedigrí se han conseguido reflejar todas las relaciones de parentesco conocidas (Figura 18). Han sido necesarios cambios en los valores por defecto de la tasa de error para llegar a determinar todos los padres, así como indicar el sexo y año de nacimiento (aproximado) de cada individuo.

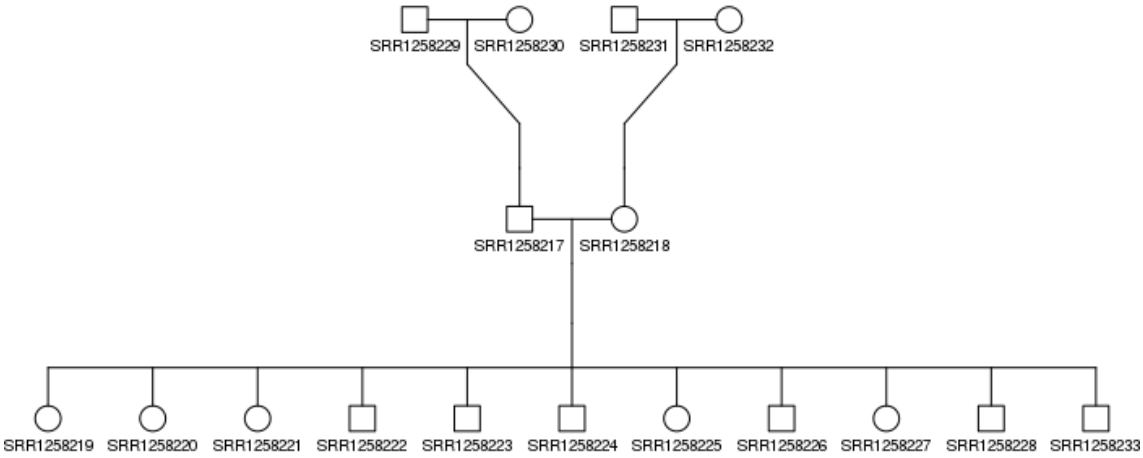

Figura 18: Representación del pedigrí en R [14] con el paquete sequoia [2] y kinship2 [25].

3.2. Resultados de los datos de familias de ratones

3.2.1. Control de calidad de los datos crudos

Para este conjunto de datos, vemos que la calidad de la secuencia por base es muy buena (Figura 19), a diferencia del anterior conjunto de datos, mientras que seguimos encontrando secuencias duplicadas (Figura 20). El resto de tests son compatibles con la normalidad o con datos de RNA-seq, aunque cabe destacar que existen algunas secuencias sobre-representadas.

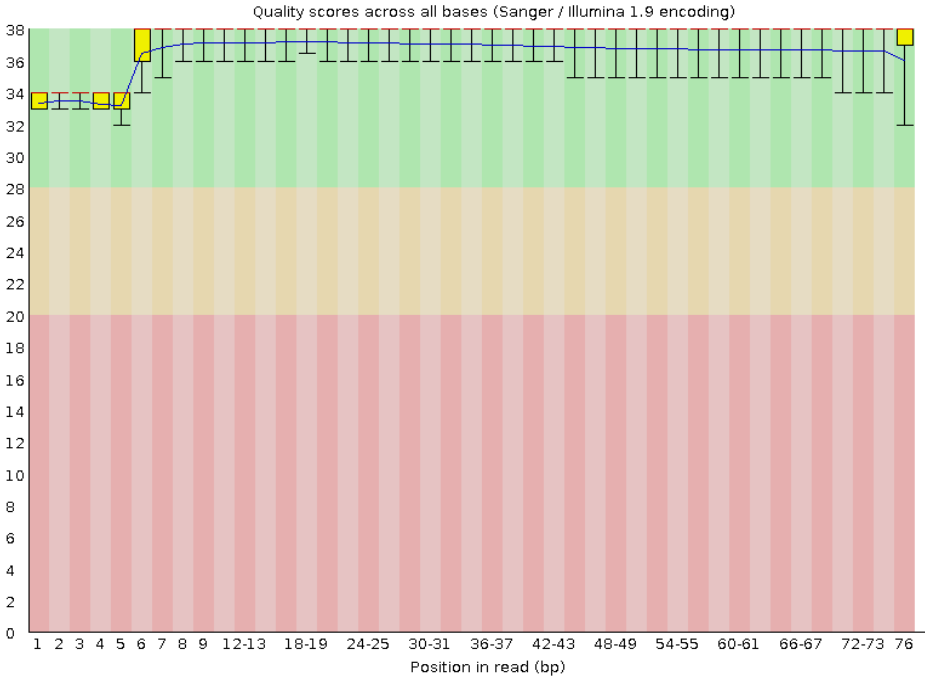

Figura 19: Calidad de las secuencias por base, en el eje x se encuentra la posición de los reads en pares de bases, y en el eje y la calidad de las bases [18].

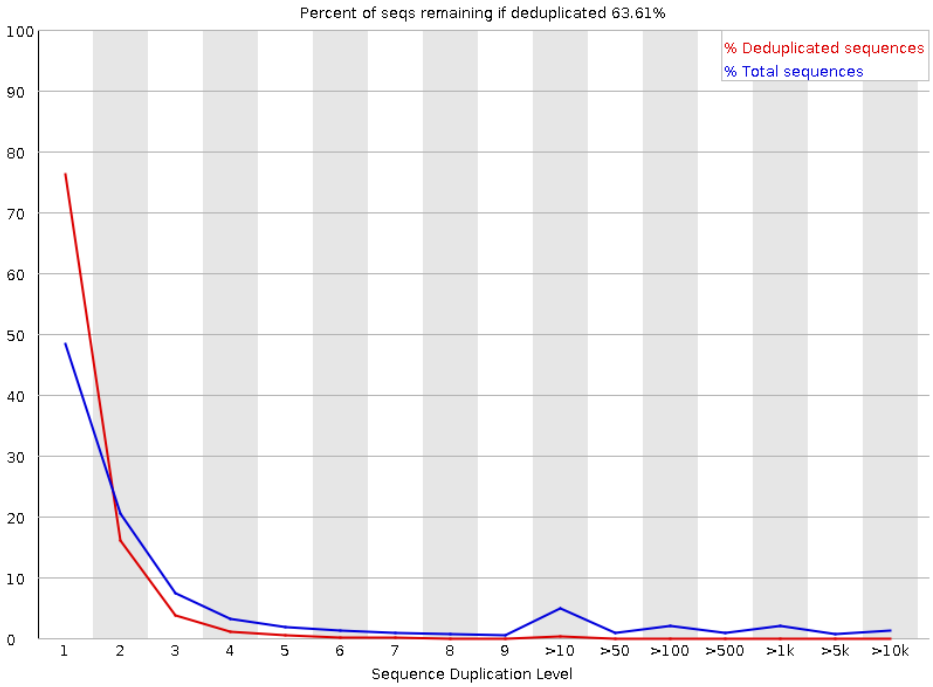

Figura 20: Nivel de secuencias duplicadas, en el eje x se encuentra el número de veces que aparece cada secuencia, y en el eje y el porcentaje de secuencias [18].

3.2.2. Mapeo de los reads en el genoma

En el proceso de mapeo con TopHat2 se eliminan entre un 14 y un 31% de los reads (Figura 21), existiendo una diferencia entre los datos de RNA-seq de testículo (14-16%) y de hígado (28-31%).

3.2.3. Eliminación de duplicados

En el proceso de eliminación de duplicados se eliminan entre un 24 y un 61% de los reads (Figura 21), encontrando diferencias entre los datos de RNA-seq de testículo (24-27%) y de hígado (52-61%).

3.2.4. Eliminación de reads con mapeo múltiple

El proceso de eliminación de reads con mapeo múltiple elimina entre un 7 y un 12% de los reads (Figura 21), encontrando otra vez diferencias entre tejidos, eliminando un 7% de los reads de testículo, y entre un 10 y un 12% de los reads de hígado.

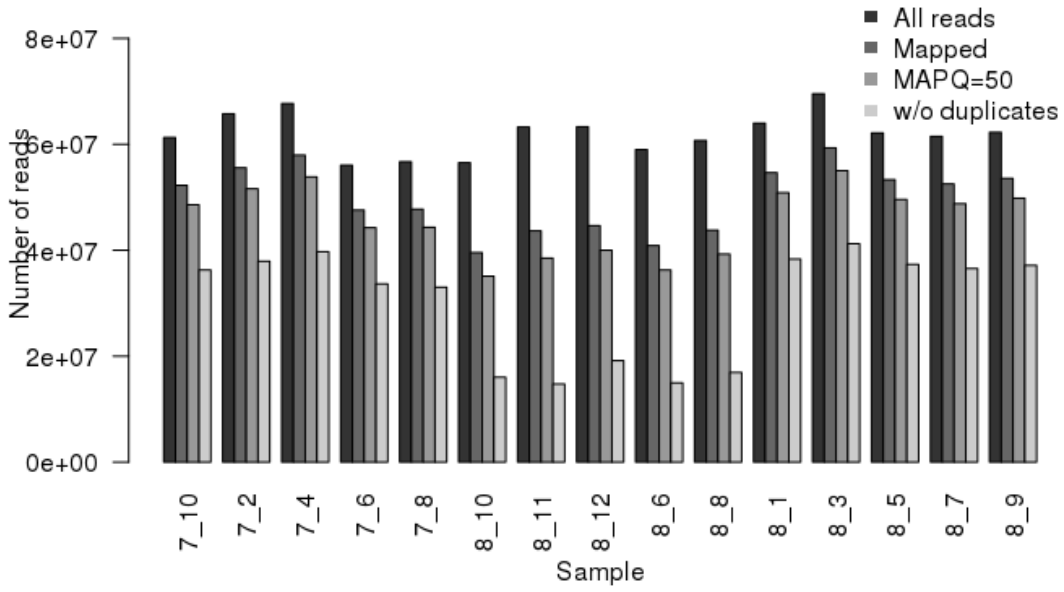

Figura 21: Número de reads antes y después de los diferentes filtros [14]. Cada filtro se aplica a los reads que han pasado el filtro anterior.

Una vez aplicados todos los filtros, se mantienen entre un 23 y un 60% de los reads iniciales. Manteniendo entre un 58 y un 60% de los reads de testículo y entre un 23 y un 30% de los reads de hígado.

3.2.5. Selección de SNPs

El fichero inicial de SNPs comunes de la anotación snp142 de mm10, contenía 8.2 millones de SNPs, de los cuales, en el proceso de selección, se eliminan 0.74 millones (Figura 22). Aunque el número de SNPs iniciales fuera menor que en humanos, tras la selección de SNPs nos quedamos aproximadamente con los mismos, ya que se han eliminado menos SNPs en repeticiones.

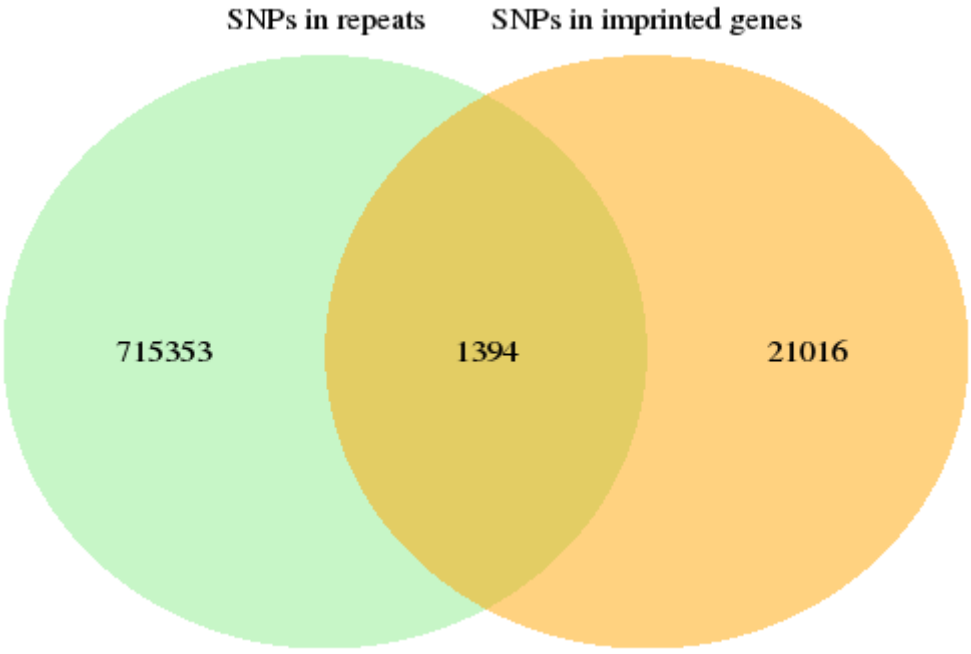

Figura 22: Diagrama de Venn de los SNPs eliminados según su localización [14].

3.2.6. Obtención de SNPs en los reads

De los 7.5 millones de SNPs que han pasado la selección, 5 millones se encuentran en los reads. La proporción es mucho más elevada que en los datos de humanos, ya que en testículo hay una mayor expresión de genes que en linfocitos B.

3.2.7. Filtro de SNPs en los reads

Aplicamos los filtros que hemos seleccionado en humanos: DP ≥ 10, GQ ≥ 20, MAF ≥ 0.3 y r^2 (desequilibrio de ligamiento) ≤ 0.2.

3.2.8. Determinación de las relaciones entre individuos

Al representar la asociación entre los diferentes individuos (Figura 23) vemos que las relaciones de padre-hijo quedan claramente separadas de las otras relaciones. Por otro lado, las relaciones abuelo-nieto y tío-sobrino son

indiferenciables, cosa que ya esperábamos de acuerdo con los valores teóricos de los coeficientes de Cotterman para estas relaciones (Tabla 1), y a su vez algunas de las parejas de individuos con estas relaciones se superponen con individuos no relacionados. Con este conjunto de datos, a diferencia del anterior, vemos que los individuos no relacionados tienen valores de z1 e incluso de z2 más elevados de los esperados, y por esta razón son más difíciles de diferenciar de otro tipo de relaciones de segundo grado.

Debido a las diferencias encontradas entre los datos de RNA-seq de testículo e hígado hemos decidido repetir el proceso utilizando sólo los individuos con datos de RNA-seq de testículo (Figura 24). En este caso no se produce una superposición entre individuos con relaciones de tío-sobrino e individuos no relacionados (sólo disponemos de una pareja con relación tío-sobrino), y también vemos una ligera disminución de los valores de z1 y z2 en los individuos no relacionados (excepto una pareja con un valor de z1 del 60%).

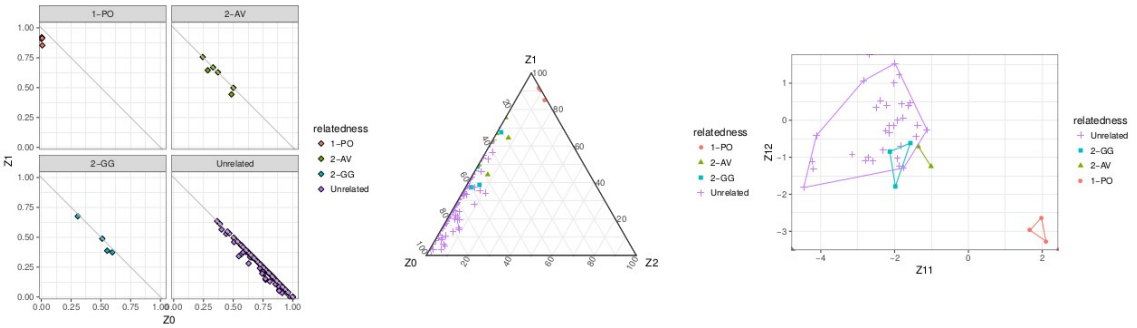

Figura 23: Gráficos de parentesco de todos los ratones [8][14]. Se han comparado entre 501 y 2059 SNPs.

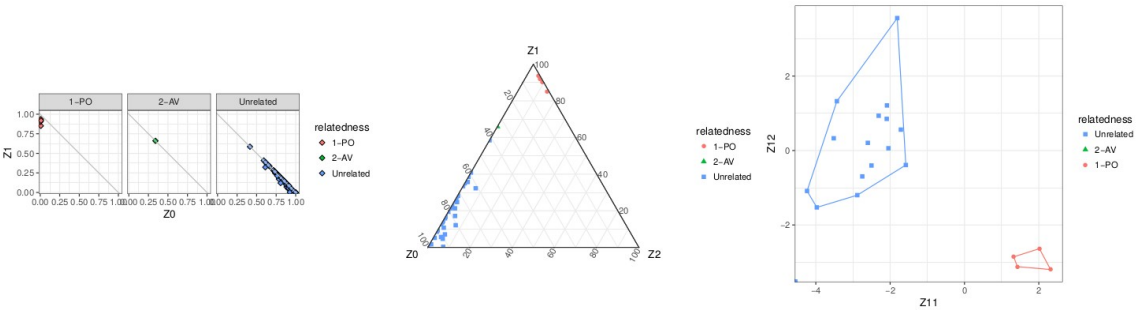

Figura 24: Gráficos de parentesco de los ratones con datos de RNA-seq de testículo [8][14]. Se han comparado entre 1724 y 2115 SNPs.

3.2.9. Representación del pedigrí

La representación del pedigrí con el método utilizado con el primer conjunto de datos no ha sido posible para este conjunto de datos. El programa sólo ha sido capaz de detectar una relación de padre-hijo y no se han podido representar las relaciones detectadas debido a la falta de madres.

4. Discusión

El hecho de no aplicar un filtro de calidad de bases no debería ser un problema, ya que el 90% de las bases que se eliminarían con una calidad de 10, estarían bien secuenciadas [26], produciendo una pérdida de información importante, y por otro lado existen otros filtros como la profundidad, la calidad del genotipo y la MAF que contrarrestan estos posibles errores de secuenciación.

Con el método utilizado se pueden diferenciar relaciones de parentesco cercanas como padre-hijo y hermanos completos de otro tipo de relaciones, y según el tipo de datos utilizados también se pueden diferenciar relaciones de segundo grado de individuos no relacionados.

Por lo que respecta a los valores obtenidos de z0, z1 y z2, se observan diferencias con los valores teóricos, y también se a observa que para cada tipo de relación existe un rango de valores obtenidos. Esto se puede deber a la propia variabilidad que existe entre individuos, existiendo hermanos completos que se parecen más entre sí que otros hermanos completos, y al filtro de MAF, observando diferencias en estos valores al aplicar diferentes filtros. Si no se utilizara este filtro, todos los SNPs con MAF = 0 provocarían una sobreestimación de las relaciones entre individuos al compartir más alelos.

La principal diferencia que se observa entre los dos conjuntos de datos, es el rango de z1 en individuos no relacionados, teniendo por un lado el factor especie y por otro lado el número de parejas de individuos con esta relación (en el primer conjunto de datos hay 10 parejas de individuos con esta relación mientras que en el segundo hay 90).

También existen diferencias entre los tejidos de los que se obtienen las muestras, siendo los filtros de los reads más restrictivos en datos de hígado que de testículo.

Otro factor que podría afectar a la eficacia del método es el número de individuos utilizado, ya que con pocos individuos el filtro de MAF podría alterar los resultados obtenidos. En este caso se podrían añadir individuos no relacionados al conjunto de datos para intentar mejorar los resultados.

Al aplicar el método con los filtros descritos se debe tener en cuenta el número de SNPs con el que se hacen las comparaciones por parejas, ya que con demasiado pocos SNPs el método puede no ser eficaz, y en este caso se deberían cambiar los valores de corte de los filtros para hacerlos menos restrictivos.

5. Conclusiones

Es posible desarrollar un método de reconstrucción de pedigríes a partir de datos de RNA-seq. Este método es capaz de detectar errores en el etiquetado de las muestras, como se ha visto en los primeros resultados.

Por lo que respecta a los objetivos planteados, se ha conseguido obtener los genotipos de los SNPs seleccionados en todos los individuos y se ha conseguido reconstruir el pedigrí (este último objetivo sólo se ha conseguido para el primer conjunto de datos ya que en el segundo sólo se tenían datos de los machos).

Se ha seguido la planificación, aunque se han incluido nuevas tareas que respondían a problemas que iban surgiendo, como el error en el etiquetado de las muestras, así como otras tareas que se han incluido para mejorar el modelo como la introducción de un nuevo conjunto de datos. Estas nuevas tareas no han impedido que se siguiera la calendarización planteada. Por lo que respecta a la metodología, se ha seguido como se había previsto y ha dado buenos resultados, aunque se probaran otras metodologías que podrían ser más eficientes.

Existen diferentes líneas de trabajo futuras:

- Probar el método con datos de small-RNA, y adecuarlo a este tipo de datos si es necesario.
- Probar otros programas que permitan acelerar el método (el proceso de mapeado y el de eliminación de duplicados son los que más tiempo consumen).
- Mejorar o simplificar el método valorando todos los filtros y su necesidad.
- Probar otros métodos para reconstruir pedigríes que permitan la representación de las relaciones de parentesco de las familias de ratones.
- Aumentar el número de SNPs mediante la imputación de alelos y comprobar si supone una mejora para el modelo.

6. Glosario

- RNA-seq: tecnología que permite obtener la secuencia del transcriptoma mediante la secuenciación de alto rendimiento [30].

- Imprinting: proceso de silenciación de genes mediante la metilación de la secuencia de ADN. El alelo inhibido está metilado mientras que el alelo activo no lo está [31].

- Read: secuencia de pares de bases obtenida de la secuenciación de un fragmento del genoma.

- SNP: acrónimo de “Single Nucleotide Polymorphism”, variaciones en la secuencia de ADN de una o unas pocas bases.

- Splicing: proceso en el que se identifican y eliminan los intrones de las secuencias de pre-mRNA y finalmente se unen los extremos de los exones [32].

- IBD: acrónimo de “Identical By Descent", se dice de dos alelos que son copias heredadas que provienen del mismo alelo ancestral [33].

- FASTQ: archivo que contiene las bases secuenciadas e información de su calidad para todos los reads que pasen el filtro [34].

- BAM: es la versión binaria comprimida de un archivo SAM (Sequence Alignment Map), en los que se representan secuencias alineadas [35].

- BCF: es la versión binaria del formato VCF (Variant Call Format), en los que se guardan los datos de variaciones [36].

- BED: acrónimo de “Browser Extensible Data”, formato que consta de una linea por entrada indicando como mínimo el cromosoma, el inicio y final de la secuencia [37].

- WGS: acrónimo de “Whole Genome Sequencing”, procedimiento que secuencia todo el genoma de un organismo en un único proceso [38].

- F0, F1 y F2: en un pedigrí se refieren a individuos de la primera, segunda y tercera generación respectivamente.

- small-RNA: secuencias de 20 a 30 nucleótidos que no se traducen en proteínas, sino que regulan diferentes procesos biológicos [39].

7. Bibliografía

[1] D. Shem-Tov y E. Halperin, «Historical Pedigree Reconstruction from Extant Populations Using PArtitioning of RElatives (PREPARE)», *PLoS Comput. Biol.*, vol. 10, n.º 6, pp. 1-13, 2014.

[2] J. Huisman, «Pedigree reconstruction from SNP data: parentage assignment, sibship clustering and beyond», *Mol. Ecol. Resour.*, vol. 17, n.º 5, pp. 1009-1024, 2017.

[3] M. Riester, P. F. Stadler, y K. Klemm, «FRANz: Reconstruction of wild multi-generation pedigrees», *Bioinformatics*, vol. 25, n.º 16, pp. 2134-2139, 2009.

[4] D. Andergassen *et al.*, «Allelome.PRO, a pipeline to define allele-specific genomic features from high-throughput sequencing data», *Nucleic Acids Res.*, vol. 43, n.º 21, 2015.

[5] X. Li *et al.*, «Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants», *Am. J. Hum. Genet.*, vol. 95, n.º 3, pp. 245-256, 2014.

[6] R. Piskol, G. Ramaswami, y J. B. Li, «Reliable identification of genomic variants from RNA-seq data», *Am. J. Hum. Genet.*, vol. 93, n.º 4, pp. 641-651, 2013.

[7] S. E. Castel, A. Levy-Moonshine, P. Mohammadi, E. Banks, y T. Lappalainen, «Tools and best practices for data processing in allelic expression analysis», *Genome Biol.*, vol. 16, n.º 1, pp. 1-12, 2015.

[8] I. Galván-Femenía, J. Graffelman, y C. Barceló-i-Vidal, «Graphics for relatedness research», *Mol. Ecol. Resour.*, vol. 17, n.º 6, pp. 1271-1282, 2017.

[9] A. Dobin y T. R. Gingeras, «TopHat2 : accurate alignment of transcriptomes in the presence of insertions , deletions and gene fusions», *Genome Biol.*, vol. 14, n.º 4, p. R36, 2013.

[10] A. Dobin *et al.*, «STAR: Ultrafast universal RNA-seq aligner», *Bioinformatics*, vol. 29, n.º 1, pp. 15-21, 2013.

[11] H. Li y R. Durbin, «Fast and accurate short read alignment with Burrows-Wheeler transform», *Bioinformatics*, vol. 25, n.º 14, pp. 1754-1760, 2009.

[12] H. Li *et al.*, «The Sequence Alignment/Map format and SAMtools», *Bioinformatics*, vol. 25, n.º 16, pp. 2078-2079, 2009.

[13] S. Purcell *et al.*, «PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses», *Am. J. Hum. Genet.*, vol. 81, n.º 3, pp. 559-575, 2007.

[14] R Development Core Team, «R: A Language and Environment for Statistical Computing», 2005.

[15] «NCBI FTP site». [En línea]. Disponible en: ftp://ftp-trace.ncbi.nih.gov. [Accedido: 27-feb-2018].

[16] National Center for Biotechnology Information (US), «Downloading SRA data using command line utilities», en *SRA Knowledge Base*, n.º Md, 2011.

[17] «fasq-dump». [En línea]. Disponible en: https://ncbi.github.io/sra-tools/fastq-dump.html. [Accedido: 20-mar-2018].

[18] Babraham Bioinformatics, «FastQC». [En línea]. Disponible en: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. [Accedido: 20-mar-2018].

[19] «FASTX-Toolkit». [En línea]. Disponible en: http://hannonlab.cshl.edu/fastx_toolkit/. [Accedido: 20-mar-2018].

[20] «Percent aligned reads for RNA-seq». [En línea]. Disponible en: http://seqanswers.com/forums/showthread.php?t=32642. [Accedido: 05-mar-2018].

[21] A. R. Quinlan y I. M. Hall, «BEDTools: A flexible suite of utilities for comparing genomic features», *Bioinformatics*, vol. 26, n.º 6, pp. 841-842, 2010.

[22] «UCSC Genome Browser». [En línea]. Disponible en: https://genome.ucsc.edu/cgi-bin/hgTables. [Accedido: 07-mar-2018].

[23] «Geneimprint». [En línea]. Disponible en: http://geneimprint.com/. [Accedido: 08-mar-2018].

[24] P. Danecek *et al.*, «The variant call format and VCFtools», *Bioinformatics*, vol. 27, n.º 15, pp. 2156-2158, 2011.

[25] J. P. Sinnwell, T. M. Therneau, y D. J. Schaid, «The kinship2 R package for pedigree data», *Hum. Hered.*, vol. 78, n.º 2, pp. 91-93, 2014.

[26] «Phred-scaled Quality Scores». [En línea]. Disponible en: https://software.broadinstitute.org/gatk/documentation/article.php?%0Aid=4260. [Accedido: 02-abr-2018].

[27] W. G. Hill y A. Robertson, «Linkage disequilibrium in finite populations», *TAG Theor. Appl. Genet.*, vol. 38, n.º 6, pp. 226-231, 1968.

[28] «Gene expression omnibus». [En línea]. Disponible en: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56961. [Accedido: 27-feb-2018].

[29] S. Lee, S. Lee, S. Ouellette, W. Y. Park, E. A. Lee, y P. J. Park, «NGSCheckMate: Software for validating sample identity in Next-generation sequencing studies within and across data types», *Nucleic Acids Res.*, vol. 45, n.º 11, p. e103, 2017.

[30] Z. Wang, M. Gerstein, y M. Snyder, «RNA-Seq: a revolutionary tool for transcriptomics», *Nat. Rev. Genet.*, vol. 10, n.º 1, pp. 57-63, 2009.

[31] E. Bajrami y M. Spiroski, «Genomic imprinting», *Open Access Maced. J. Med. Sci.*, vol. 4, n.º 1, pp. 181-184, 2016.

[32] A. Pandya-jones, «Pre-mRNA splicing during transcription in the mamamlian system», *Wiley Interdiscip. Rev. RNA*, vol. 2, n.º 5, pp. 700-717, 2012.

[33] M. Tiret y F. Hospital, «Blocks of chromosomes identical by descent in a population: Models and predictions», *PLoS One*, vol. 12, n.º 11, pp. 1-11, 2017.

[34] «FASTQ Files». [En línea]. Disponible en: http://support.illumina.com/content/dam/illumina-support/help/BaseSpaceHelp_v2/Content/Vault/Informatics/Sequencing_Analysis/BS/swSEQ_mBS_FASTQFiles.htm. [Accedido: 01-jun-2018].

[35] «BAM File Format». [En línea]. Disponible en: https://support.illumina.com/help/BS_App_ENR_OLH_15050961/Content/Source/Informatics/BAM-Format.htm. [Accedido: 01-jun-2018].

[36] «Calling SNPs/INDELs with SAMtools/BCFtools». [En línea]. Disponible en: http://samtools.sourceforge.net/mpileup.shtml. [Accedido: 01-jun-2018].

[37] «BED format». [En línea]. Disponible en: https://genome.ucsc.edu/FAQ/FAQformat.html#format1. [Accedido: 01-jun-2018].

[38] «Whole Genome Sequencing (WGS)». [En línea]. Disponible en: https://www.cdc.gov/pulsenet/pathogens/wgs.html. [Accedido: 01-jun-2018].

[39] H. Großhans y W. Filipowicz, «The expanding world of small RNAs», *Nature*, vol. 451, pp. 414–416, 2008.

8. Anexos

Descarga de los archivos SRA mediante la herramienta GNU Wget:

```
wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR125/SRR$ID/SRR$ID.sra
```

Obtención de ficheros FASTQ mediante el programa fastq-dump:

- La opción –split-3 genera el fichero *_1.fastq y *_2.fastq para reads emparejados (si hubiera algún read no emparejado, se generaría un archivo *.fastq con dichos reads).

```
fastq-dump --split-3 SRR$ID
```

Filtro de calidad de los reads con FASTQ Quality Trimmer:

- La opción -t 20 recorta las últimas bases de los reads que no llegan a una base quality de 20
- La opción -l 20 elimina aquellos reads que hayan quedado con una longitud inferior a 20 bases después del trimming

```
fastq_quality_trimmer -t 20 -l 20 -i $data/SRR$ID"_1.fastq" -o $trimmer/SRR$ID"_1_t20.fastq"
```

Mapeo de los reads en el genoma con el programa TopHat2.

- Las opciones --no-coverage-search y --no-novel-juncs se añaden porque no estamos interesados en encontrar nuevos sitios de splicing y esto permite aumentar mucho la velocidad.
- Las opciones -G y --transcriptome-index permiten añadir un archivo GTF/GFF con los tránscritos conocidos y un bowtie index del transcriptoma.

```
tophat2 -p 7 --no-coverage-search --no-novel-juncs \
        -o $path/tophat2/SRR$ID \
        -G $path2/Homo_sapiens.GRCh37.87.gtf \
        --transcriptome-index $path2/Homo_sapiens.GRCh37.87.gtf \
        $path3/Hsapiens.GRCh37_72 \
        $path/data/SRR$ID"_1.fastq" \
        $path/data/SRR$ID"_2.fastq" \

mv $path/tophat2/$ID/accepted_hits.bam $path/tophat2/$ID.bam
```

Eliminación de los reads duplicados con samtools markdup:

- Antes de poder usar la herramienta markdup se deben hacer tres pasos previos: (i) sort -n para ordenar el fichero por nombre, (ii) fixmate -m para añadir las etiquetas ms y MC que se usaran posteriormente y (iii) sort para ordenar el fichero por posición.
- La opción -r de markdup se utiliza para eliminar los duplicados.

```
samtools sort -n -o $md/namesort_$ID.bam $path/tophat2/$ID.bam
```

```
samtools fixmate -m $md/namesort_$ID.bam $md/fixmate_$ID.bam
samtools sort -o $md/positionsort_$ID.bam $md/fixmate_$ID.bam
samtools markdup -r $md/positionsort_$ID.bam $md/markdup_$ID.bam
```

Selección de los SNPs con bedtools intersect:

- El archivo de los SNPs ha sido obtenido de UCSC Genome Browser. Se han seleccionado los "Common SNPs(150)" del assembly "GRCh37/hg19".
- El archivo de los repeats ha sido obtenido de UCSC Genome Browser. Se ha seleccionado el "RepeatMasker" del assembly "GRCh37/hg19".
- Los genes improntados se han obtenido de http://geneimprint.com/. Se han seleccionado los genes humanos con "Status: Imprinted" y se han introducido en Biomart GRCh37 para obtener las regiones de los genes improntados.

```
#eliminar SNPs en repeats
bedtools intersect -v -a $path2/SNPs/hg19_snp150.bed -b
$path/repeats/hg19_repeatmasker.bed >
$path/SNPs/hg19_snp150_repeatmasker.bed

#eliminar SNPs en repeats y en imprinted genes "rmi" (repeatmasker +
imprinted genes)
bedtools intersect -v -a $path/SNPs/hg19_snp150_repeatmasker.bed -b
$path/imprinted/imprinted_genes.bed > $path/SNPs/hg19_snp150_rmi.bed
```

Filtrado de los SNPs en los reads con samtools mpileup y bcftools call:

- -A se incluye para no descartar reads con emparejamientos anómalos
- -q excluye aquellos reads con una calidad de mapeo igual o inferior al nivel de corte (se usa el número 4 para eliminar los reads de mapeo múltiple, como se ha explicado en el apartado 1.3.)
- -Q no tiene en cuenta las bases con una base quality inferior a la indicada (20)
- -t permite añadir etiquetas (AD: allele depth, DP:depth)
- -l permite incluir el fichero BED con los SNPs seleccionado, solo se tendrán en cuenta estas regiones
- -f permite añadir un fichero FASTA con el genoma de referencia
- -g indica el formato del output BCF
- -m permite identificar variantes multialélicas
- -O determina el formato de output (b: BCF)
- -f permite añadir etiquetas (GQ: genotype quality)

```
samtools mpileup -A -q 4 -t AD,DP \
        -l $path/SNPs/hg19_snp150_rmi.bed \
        -f $path1/Hsapiens.GRCh37_72.fa \
        $path/markdup/markdup_SRR12582??.bam \
        -g | bcftools call -m - --threads 7 -O b \
        -f GQ >  $path/mpileup/hg19_snp150_markdup_rmi_all.q4.bcf
```

Filtro por GQ y DP con vcftools

- --minGQ asigna como genotipos faltantes los genotipos que no lleguen a la calidad determinada
- --minDP asigna como genotipos faltantes los genotipos que no lleguen a la profundidad determinada
- --recode genera un fichero VCF
- --recode-INFO-all mantiene todas las etiquetas

```
vcftools --bcf $path/mpileup/hg19_snp150_markdup_rmi_all.q4.bcf \
    --out $path/filtered/hg19_snp150_markdup_rmi_all_q4_GQ20_DP10 \
    --minGQ 20 --minDP 10 --recode --recode-INFO-all
```

Determinar la associación entre las distintas muestras con PLINK

- --maf excluye los SNPs con una MAF inferior a la indicada
- --max-maf excluye los SNPs con una MAF superior a la indicada
- --indep-pairwise excluye los SNPs con un r^2 superior al indicado

```
plink2 --vcf $input --freq counts --maf 0.3 --make-bed --double-id --out $input.maf3
plink2 --bfile $input.maf3 --indep-pairwise 50 5 0.2 --out $input.maf3.prunning2
plink2 --bfile $input.maf3 --extract $input.maf3.prunning2.prune.in --make-bed --out $input.maf3.ld2
plink2 --bfile $input.maf3.ld2 --genome full --out $input.maf3.ld2
```

Preparar con PLINK el archivo para representar el pedigrí:

- --geno excluye los SNPs con una proporción de genotipos faltantes superior a la indicada

```
plink2 --vcf $input --freq counts --maf 0.3 --geno 0.9 --double-id --recode --out $input.maf3.sequoia
plink2 --file $input.maf3.sequoia --indep-pairwise 50 5 0.2 --out $input.maf3.prunning2.sequoia
plink2 --file $input.maf3.sequoia --extract $input.maf3.prunning2.sequoia.prune.in --recodeA --out ../sequoia/$input.maf3.ld2.sequoia
```

Representación del pedigrí en R:

```
library(sequoia)
genom <-  GenoConvert("hg19_snp150_markdup_rmi_all_q4_GQ20_
        DP10.recode.ID.vcf.maf3.ld2.sequoia.raw")
library(stringr)
rownames(genom) <- sapply(str_split(rownames(genom), "_"), "[", 2)
rownames(genom) <- substr(rownames(genom),1, nchar(rownames(genom))-4)
lifehistdata <- data.frame("ID"=rownames(genom), "Sex"= c(2, 1, 1, 1, 1, 2, 2, 2,
        1, 2, 1, 2, 2, 1, 2, 1, 2), "BY"=c(2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 1, 1, 1,
        3))
ParOUT <- sequoia(GenoM = genom, LifeHistData = lifehistdata, MaxSibIter=0,
        Err = 1e-1, quiet = T)
SeqOut <- sequoia(GenoM = genom, SeqList = ParOUT, MaxSibIter = 20, Err =
        1e-1, quiet = T)
```

```
library(kinship2)
ped <- pedigree(SeqOut$Pedigree$id, SeqOut$Pedigree$sire,
        SeqOut$Pedigree$dam, abs(lifehistdata$Sex-3))
par(xpd = T)
plot.pedigree(ped, cex = 0.5)
```