UNIVERSIDAD DE MURCIA

Entregable parte 2: Análisis de Datos Ómicos

Análisis por DNA-seq para búsqueda de variantes en pacientes con Leucemia mieloide aguda (AML) y Mieloma múltiple (MM)

Máster en Bioinformática - Facultad de Biología

Autores: **Juan Pedro López Marín** y **Francisco Javier López Carbonell**

12/04/2024

Índice

Índice de figuras

Índice de cuadros

1. Introducción

La secuenciación de nueva generación (NGS) y otras técnicas de alto rendimiento han permitido descubrir numerosos genes con mutaciones recurrentes en diferentes tipos de neoplasias mieloides, como son la leucemia mieloide aguda (AML) y el mieloma múltiple (MM).

Dentro de este marco, encontramos la técnica de secuenciación de genoma completo (WGS), que nos permite obtener la secuencia del genoma completo de una determinada especie a partir de la extracción de una muestra de ADN. Gracias a esta técnica, se han podido identificar genes y regiones intergénicas significativamente mutadas en pacientes que han sido diagnosticados con alguna de las enfermedades mencionadas.

En este caso, nosotros nos vamos a centrar en diversos genes que han sido identificados como asociados a la presencia de AML y que se sitúan en diferentes regiones cromosómicas. Realizaremos la búsqueda de variantes en una muestra con AML y a modo de comparación, en una muestra con MM.[1]

Antes de comenzar con el desarrollo del resto de experimento, aclarar que la pipeline utilizada para llegar desde los datos iniciales hasta los ficheros analizados en última instancia está representada en el Anexo I.

Dicho anexo así como el resto de ficheros generados durante el experimento se encuentran almacenados en la dirección $/home/alumno12/ADO/Documentacion\_Entrega2$ del servidor del máster *dayhoff*.

2. Objetivo

El objetivo principal de este trabajo es identificar, anotar y comparar variantes de interés asociadas con el fenotipo de pacientes con distintos tipos de neoplasias mieloides. En concreto, se trata de Leucemia Mieloide Aguda (AML) y Mieloma múltiple (MM).

3. Metodología

3.1. Obtención de los datos

Como ya se comentó en la primera entrega, los datos empleados se encuentran publicados en la base de datos del *European Bioinformatics Institute* (EBI), concretamente en el *European Nucleotide Archive* (ENA). La ubicación concreta de los datos correspondientes a cada paciente es la siguiente:

- **Paciente AML:** `https://www.ebi.ac.uk/ena/browser/view/SRR544623`
- **Paciente MM:** `https://www.ebi.ac.uk/ena/browser/view/SRR2497264`

Ambos datasets han sido obtenidos utilizando la misma herramienta de secuenciación, en este caso, el modelo *Illumina HiSeq 2000* de Illumina.

3.2. Herramientas utilizadas para DNA-seq (WGS)

La mayoría de las herramientas han sido ejecutadas desde el servidor *Galaxy* (disponible en dayhoff.inf.um.es) y *Galaxy Europe*[2]. En concreto, algunas de ellas son:

Download and Extract Reads in FASTA/Q, para la descarga e importación de datos. **Fastqc** para analizar calidad de secuencias. **Alineamiento con algoritmos BWA-backtrack y BWA-MEM**, para mapeo de lecturas en el genoma de referencia. **Integrative Genome Viewer (IGV)**, empleada como herramienta para hacer una inspección visual de nuestras lecturas mapeadas en el cromosoma de interés. **Samtools flagstat/Samtools stats**, para inspeccionar la calidad del alineamiento de forma global. **FreeBayes bayesian genetic variant detector**, para identificar variantes y la calidad de cada una de las posiciones genómicas, **Variant effect Predictor (VEP) de Ensembl**, que nos permitirá realizar una predicción del efecto de esas variantes identificadas previamente. **SnpEff**, para la anotación y filtrado de variantes. **Integrative Genomics Viewer (IGV)** para visualización del perfil de alineamiento. Por otro lado, mencionar que se han llevado ediciones de ficheros y pequeños flujos de trabajo vía terminal de comandos de bash en dicho servidor.

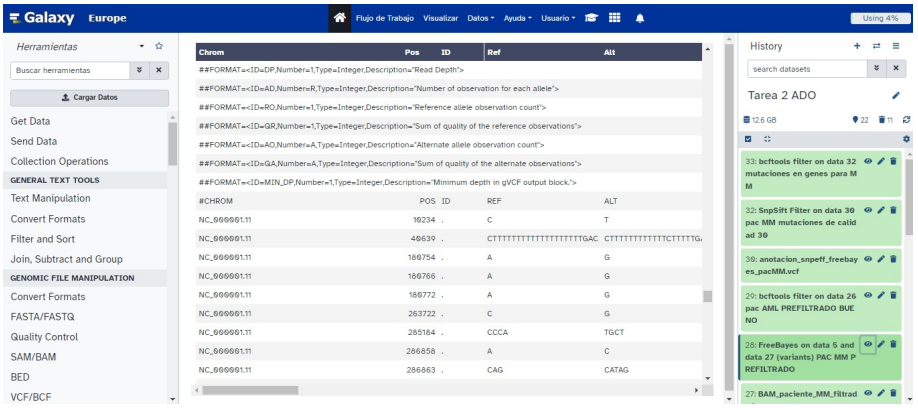

Figura 1: Entorno e historial de herramientas en Galaxy EU

4. Desarrollo del flujo de trabajo

4.1. Obtención de los ficheros FASTQ

El primer paso de todos consiste en obtener los ficheros FastQ de cada uno de los pacientes a estudiar en este caso. Para ello se ha hecho uso de la herramienta de galaxy **Faster Download and Extract Reads in FASTQ format from NCBI SRA**, una herramienta que como su propio nombre indica permite obtener datos en formato *fastq* del archivo SRA del NCBI haciendo uso de la herramienta "**fasterq-dump**"del SRA Toolkit [3].

Simplemente insertando el número de acceso del Run para cada paciente (SRR), se obtienen los ficheros fastq correspondientes a cada paciente. En este caso, se obtienen 4 ficheros fastq, 2 por cada paciente, donde uno de ellos corresponde a las lecturas en sentido **forward** y otro a las lecturas en **reverse**, de manera que a priori se va a poder trabajar con lecturas **paired-end**.

4.2. Control de calidad de las lecturas

Antes de proceder al análisis de los datos obtenidos del secuenciador es conveniente asegurarse de que estos datos son correctos y, si es necesario, eliminar los datos defectuosos.

La herramienta **FastQC** permite una exploración sencilla de los datos. Para cada fichero analizado aparecen dos archivos de salida (Webpage y RawData). El objeto Webpage contiene una serie de gráficos con las medidas de calidad mientras que, en el RawData, podemos encontrar las tablas con todos los estadísticos realizados [4]. Un resumen de los resultados del FastQ-C para cada uno de los ficheros fasta inciales se puede encontrar en el **Anexo II** adjuntado a este informe.

Si abrimos las pestañas de la página web podemos ver diferentes secciones como *estadísticas básicas* , *calidad por base de cada secuencia en promedio*, *puntuaciones de calidad por secuencia* o incluso *contenido de GC por secuencia*. Estos análisis y gráficos nos permiten comprobar la calidad de los datos y si tienen mala calidad nos indican cómo podemos modificarlos para mejorar su calidad en mayor o menor medida.

Las lecturas correspondientes al paciente con AML presentan una longitud de 100 bases. Con respecto a las lecturas en sentido **forward**, la calidad por base en todas las secuencias en promedio es bastante alta, siempre estando por encima de 30; con respecto al tipo de nucleótido por base en cada secuencia en promedio (*Per base sequence content*), la tendencia para las 4 bases se mantiene estable a lo largo de las 100 posiciones, lo que es un buen indicador de calidad.

El contenido en GC también aparenta ser correcto, y lo único que parece estar ligeramente mal es que se identifica una posible secuencia sobrerrepresentada entre las lecturas la cual puede ser un adaptador el cual haya que eliminar manualmente.

En el caso de las lecturas en sentido **reverse**, prácticamente todos los criterios de calidad son idénticos o muy parecidos a las lecturas en sentido forward, es decir, muy buenos, con la diferencia de que en este caso no parecen encontrarse secuencias sobrerrepresentadas como pasaba en el caso anterior.

Por otro lado, las lecturas correspondientes al paciente con MM en sentido **forward** presentan una longitud de 49 bases. La calidad por posición en promedio a lo largo de las secuencias vuelve a ser muy buena, estando prácticamente siempre por encima de un valor de 30. El resto de indicadores de calidad también son positivos, salvo la presencia de secuencias sobrerrepresentadas, donde se detectan 3 secuencias concretas las cuales posiblemente se deban eliminar para evitar que interfieran y contaminen los resultados en pasos posteriores del experimento.

Sin embargo, cuando se analiza el fichero fastq en sentido **reverse** para el paciente de MM, se ve como en dicho fichero las lecturas almacenadas presentan una longitud de 1 base únicamente, algo que no tiene sentido y que por lo tanto nos lleva a descartar este fichero para la realización del experimento, llevando a cabo el estudio del paciente de MM como un caso donde únicamente se tienen lecturas **single-end**.

4.3. Alineamiento

Para el proceso de alineamiento se ha decidido optar por el algoritmo **BWA** (Burrows-Wheeler Alignment), ya que al tratarse para ambos pacientes de lecturas obtenidas mediante equipamientos/herramientas de Illumina, es idóneo el uso de este algoritmo ya que está pensado para trabajar con lecturas provenientes de herramienta de Illumina.

Más concretamente, se han utilizado 2 implementaciones diferentes del algoritmo, **BWA-backtrack y BWA-MEM**. BWA-backtrack se utiliza para las lecturas del paciente con MM, ya que este algoritmo está indicado para trabajar con lecturas con una longitud menor de 100 bp; mientras que BWA-MEM se utiliza para las lecturas del paciente con AML debido a que esta adaptación del algoritmo BWA está indicada para trabajar con secuencia cuya longitud oscila entre las 70bp to 1Mbp [5].

Como genoma de referencia contra el que mapear las lecturas, se ha utilizado la versión del genoma humano **GRCh38.p14**, obtenida a partir del NCBI.

En el caso del paciente con AML, se utilizan tantos las lecturas tanto forward como reverse, sin embargo y como hemos mencionado antes, en el paciente con MM únicamente se hace uso del fichero con las lecturas en sentido forward.

4.4. Control de calidad del alineamiento

Una vez finalizado el proceso de alineamiento, se obtienen 2 ficheros ".bam", cada uno correspondiente a cada uno de los pacientes. Dichos ficheros BAM también pueden ser analizados con diferentes herramientas de Galaxy con el fin de obtener estadísticas referentes a como ha ido el proceso de alineamiento en cada caso. En este caso, se ha optado por hacer uso de 1 herramienta en concreto, **Samtools flagstat**, la cual permite ver datos estadísticos referentes a los ficheros BAM obtenidos con el fin de evaluar el alineamiento [6].

Al utilizar la herramienta **flagstat**, en el fichero BAM referente al paciente con AML se comprueba como un 95.88 % de las lecturas han sido mapeadas, lo que es un buen resultado. Además, dado que las lecturas con las que se ha hecho el alineamiento son paired-end, se obtienen estadísticos interesantes como, parejas de lecturas que han sido mapeadas en regiones diferentes. En el caso del paciente con MM, el análisis con esta herramienta indica que el 83.77 % de las lecturas han sido mapeadas.

Se ve claramente como en el caso de este último paciente el porcentaje de secuencias mapeadas es bastante menor, sin embargo, hay que tener en cuenta otro factor, y es el número de lecturas que se tiene en cada paciente. En el caso del paciente con MM se cuenta con un total de 77452621 lecturas single-end, mientras que en el caso del paciente con AML, el total de lecturas paired-end es de 8000000; por lo que quizás este peor resultado en el proceso de alineamiento en el caso del paciente con MM no sea tan significativo como si hubiese pasado en el paciente con AML.

Además, al tratarse de lecturas single-end, puede que sea más fácil que haya más problemas de alineamiento en comparación con un alineamiento donde se utilizan lecturas paired-end, dando lugar a un fichero BAM con menos lecturas mapeadas con ocurre en este caso.

4.5. Visualización del perfil de mapeo

Una vez realizado el alineamiento y comprobado que éste es de calidad, el próximo paso será ojear el perfil del mapeo de nuestras lecturas respecto al genoma de referencia de una forma mas ilustrativa.

Para ello, haremos uso de la herramienta **IGV (Integrative Genomics Viewer)**, diseñada para la exploración visual interactiva de diversos datos genómicos. Incluso para conjuntos de datos muy grandes, IGV admite la interacción en tiempo real en todas las escalas de resolución del genoma, desde el genoma completo hasta los pares de bases. [7]

La metodología en nuestro caso, se basa en subir (apartado IGV denominado *tracks*) a IGV en su formato web los dos archivos .bam con las lecturas de AML y MM junto con su índice BAM (fichero .BAI), ambos descargados del proceso de alineamiento ejecutado en Galaxy. Además, es importante seleccionar nuestra versión del genoma de referencia (apartado IGV *genome*) y que coincida con la utilizada para el mapeo de lecturas, ya mencionada anteriormente.

En nuestro caso, como este paso resulta meramente exploratorio y tenemos que acercarnos mucho a regiones concretas para mostrar algo verdaderamente informativo, hemos decidido no incluir apoyo visual en el informe. En cambio, volveremos hacer uso de IGV una vez tengamos las variantes de interés filtradas y seleccionadas, donde si mostraremos el resultado (sección 5.3 del presente informe).

4.6. Filtrado del alineamiento

Para asegurarnos de que la calidad de los alineamientos en el fichero BAM es suficiente para obtener buenos resultados en los pasos posteriores, hemos llevado a cabo un proceso de filtrado del fichero BAM mediante la herramienta **samtools view** a través de la terminal [6]. Esta herramienta permite filtrar el fichero BAM en función de la *MAPQ* o en función de diferentes *flags* o etiquetas presentes en el campo *FLAG*. El primero de los filtros aplicados fue el siguiente:

```
samtools view -b -F 0x4 -f 0x2 -q 20 BAM_paciente_AML.bam
```

Este código permite, eliminar aquellas lecturas los cuales no se han mapeado correctamente (*-F 0x4*), mantener aquellas lecturas las cuales se consideran *primarias*, es decir, lecturas las cuales han sido mapeadas en el sitio que mejor encajan a priori (*-f 0x2*); y finalmente permite filtrar y mantener únicamente aquellas lecturas cuyo valor de mapQ sea superior a 20 (*-q 20*), ya que esto es un estándar de calidad que hemos visto utilizarse en otro experimentos de análisis similares.

Una vez finalizado el proceso de filtrado, generamos un nuevo índice para el nuevo fichero BAM mediante el siguiente código:

```
samtools index BAM_paciente_AML_filtrado.bam
```

Con este nuevo índice podremos visualizar este fichero BAM y todas sus lecturas haciendo uso del portal IGV.

4.7. Detección de mutaciones

Una vez filtrados los alineamientos en función de su calidad y otros aspectos, hacemos uso de la herramienta de Galaxy **FreeBayes bayesian genetic variant detector**, la cual nos permite generar un fichero *vcf*, a partir del archivo .bam, donde se obtiene información acerca de variantes como pequeños polimorfismos [8].

Específicamente, busca SNP (polimorfismos de un solo nucleótido), indels (inserciones y eliminaciones), MNP (polimorfismos de múltiples nucleótidos) y eventos complejos (eventos compuestos de inserción y sustitución) más pequeños que el longitud de una alineación de secuenciación de lectura corta.

Esta herramienta nos permite además, obtener variantes filtradas por calidad en el mismo proceso de búsqueda de variantes, pero nosotros ya hemos filtrado previamente las lecturas mapeadas por MAPQ>20, por lo que escogemos la opción: *Simple diploid calling*.

El fichero de salida es un archivo de variantes en formato vcf, donde se obtienen un gran número de variantes en todos los cromosomas del genoma y que en etapas posteriores del análisis serán explorados más dirigidamente hacia regiones de genes concretos relacionados con la enfermedad.

4.8. Anotación de variantes

Para la anotación de variantes hemos utilizado la herramienta **SnpEff**, la cual permite llevar a cabo la anotación y filtrado de variantes a partir de un fichero VCF [9]. Para llevar a cabo el proceso de anotación hemos tenido que llevar a cabo una serie de modificaciones en el fichero VCF para que este sea compatible con la herramienta SnpEff. En primer lugar, hemos tenido que modificar el nombre de los cromosomas en el fichero VCF para que estos coincidan con los nombres de cromosomas utilizados por la base de datos propia de SnpEff. Para ello hemos seguido la siguiente pipeline, en primer lugar hemos obtenido el nombre de todos los cromosomas de nuestro fichero VCF mediante el código:

```
grep '^NC' fichero.vcf | awk '{print␣$1}' | uniq
```

Seguidamente, hemos utilizado todos los nombres de cromosomas obtenidos en dicha salida para buscar a que cromosoma hacían referencia mediante el siguiente código:

```
grep "NC_" GRCh38_latest_genomic.fna | awk '{print␣$1␣"␣"␣$5}' |
   uniq
```

Una vez obtenidas las parejas de nombres, hemos diseñado un script al cual hemos llamado *cambio_prefijos_nc.sh*, dicho script permite sustituir los nombres de los cromosomas del formato 'NC_' por su número o letra correspondiente (1,2,3,X,Y,...).

Una vez transformados los nombres de los cromosomas, hemos transformado también los nombres de los transcritos para evitar errores durante la anotación. Para ello hemos seguido una metodología similar a la anterior, en primer lugar hemos obtenido el nombre de los transcritos mediante el comando:

```
grep '^NT_\|^NW_' fichero.vcf | awk '{print␣$1}' | uniq
```

Obtenidos los nombres originales del VCF, hemos buscado sus sinónimos en el genoma de referencia mediante el comando:

```
grep 'transcrito-1\|transcrito-2\|transcrito-3\|...\|transcrito-n'
   GRCh38_latest_genomic.fna | awk '{print␣$1␣"␣"␣$NF}'
```

Ya obtenidas las parejas de nombres, otra vez hemos diseñado otro script muy similar llamado *cambio_prefijos_transcritos.sh* con el cual hemos llevado a cabo el cambio de nombre.

Finalmente, se ha llevado a cabo el proceso de anotación utilizando el comando final:

```
srun -p eck-q java -Xmx8g -jar snpEff.jar -v  -stats
   resultado_anotacion.html GRCh38.p14 fichero_vcf.vcf >
   anotacion_snpeff.vcf &
```

Una vez obtenido el fichero VCF anotado, lo siguiente que hemos hecho ha sido filtrar todas las variantes obtenidas para buscar mutaciones ubicadas en regiones determinadas, concretamente, en genes los cuales a priori podrían estar relacionados con la enfermedad AML. Para ello hemos utilizado la herramienta de Galaxy **bcftools filter** y dichos genes y sus correspondientes ubicaciones genómicas han sido obtenidas a través del portal de **OMIM** (Online Mendelian Inheritance in Man) al buscar los genes relacionados con dicha enfermedad [10]. De hecho, son varios los artículos que muestran análisis de variantes en los genes que vamos a estudiar.[11]

Las regiones y genes en concreto por los cuales se han filtrado las mutaciones son las siguientes:

Gen	Cromosoma	Coordenadas
DNMT3A	2	25,227,874-25,342,590
GATA2	3	128,479,422-128,493,201
LPP	3	188,153,021-188,890,671
CHIC2	4	54,009,789-54,091,879
KIT	4	54,657,957-54,740,715
TERT	5	1,253,167-1,295,068
NPM1	5	171,387,116-171,410,900
JAK2	9	4,984,390-5,129,948
NUP214	9	131,125,586-131,234,663
MLLT10	10	21,533,756-21,743,630
PICALM	11	85,957,175-86,069,860
ETV6	12	11,649,674-11,895,377
KRAS	12	25,205,246-25,250,929
FLT3	13	28,003,274-28,100,576
SH3GL1	19	4,360,370-4,400,547
CEBPA	19	33,299,934-33,302,534
RUNX1	21	34,787,801-35,049,302

Cuadro 1: Genes y sus posiciones genómicas

Tras llevar a cabo este filtrado, se encuentran varias mutaciones ubicadas en estos genes con diferentes efectos, tanto en el paciente con AML como en el paciente con MM, habiendo incluso mutaciones comunes entre ambos. Con el paciente de mieloma múltiple hemos llevado a cabo el mismo procedimiento, con el objetivo de **buscar mutaciones en un principio asociadas a leucemia mieloide aguda, también en el paciente con mieloma múltiple.**

La única diferencia es que en este caso, las variantes presentes en el fichero VCF resultante de utilizar FreeBayes para el paciente de MM, han sido filtradas en base a su calidad, concretamente se ha establecido un filtro para obtener únicamente aquellas variantes con un indicador de calidad superior a 30.

El por qué de esta diferencia reside en que al filtrar por regiones en el caso del paciente con AML, el VCF resultante presenta únicamente mutaciones con alta calidad, sin embargo, en el caso del paciente con MM aparecen muchas variantes con muy baja calidad, debido principalmente a que el fichero VCF obtenido con FreeBayes en el caso del paciente con MM presenta muchas mas variantes debido a que las lecturas iniciales en este paciente eran mucho mayores, como se mencionó al comienzo del experimento.

Para llevar a cabo este filtrado por calidad de las variantes se ha utilizado la herramienta **SnpSift Filter** de Galaxy [12].

Ya obtenidos los dos ficheros VCF finales para cada uno de los pacientes, hemos hecho uso de la herramienta **Variant Effect Predictor** de *Ensembl* para obtener información de una manera más gráfica y detallada acerca de las variantes presentes en cada uno de los pacientes.

5. Resultados y discusión

En cuanto a los resultados obtenidos de las diversas herramientas utilizadas, comenzaremos recordando que nuestro objetivo en última instancia es comprobar si variantes asociadas al genotipo de las enfermedades, están presentes en nuestras muestras analizadas.

5.1. Predicción de efectos por VEP

Una vez realizada la anotación y el filtrado por las regiones de nuestros genes de interés, nos llevamos el vcf resultante (tanto de AML como de MM) a VEP, una herramienta de *Ensembl* la cual permite determinar el efecto de las variantes encontradas [13]. El resultado obtenido se muestra en las siguientes figuras:

En el caso de AML, obtenemos un total de 77 variantes procesadas, el número y porcentaje de variantes novedosas es bajo frente a las variantes existentes. Por su parte, tenemos 19, 127 y 27 genes, transcripciones y características reguladoras superpuestas, respectivamente.Además, vemos que la mayor parte de las variantes son intrónicas (68 %), variante de transcrito no codificante (12 %) y degradación mediada por non-sense (7 %).

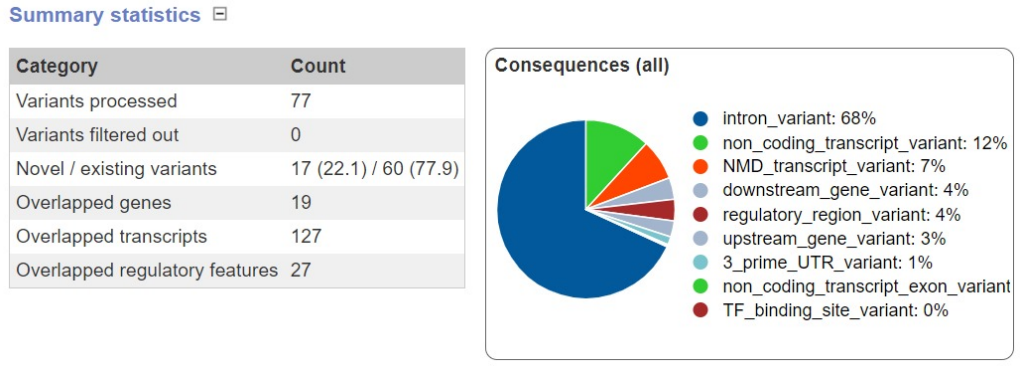

Figura 2: Resultado VEP variantes de interés AML

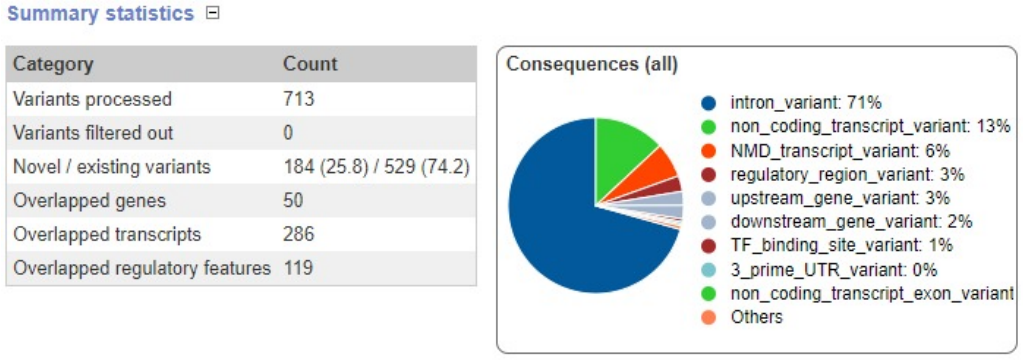

Figura 3: Resultado VEP variantes de interés MM

Por otro lado para MM, obtenemos un total de 713 variantes procesadas, el número y porcentaje de variantes novedosas de nuevo es mucho menor que el de las variantes existentes. Por su parte, tenemos 50, 286 y 119 genes, transcripciones y características reguladoras superpuestas, respectivamente. Además, vemos que también la mayor parte de las variantes son intrónicas (71 %), variante de transcrito no codificante (13 %) y degradación mediada por non-sense (6 %).

Por último, respecto a los resultados ofrecidos por VEP, obtenemos una tabla con información concreta de las variantes, como es su localización genómica, el alelo y su alelo de referencia, el symbol del gen al que corresponde la región donde se sitúa el SNP, la consecuencia o efecto de tener esa variante, el tipo biológico de la variante, el ID de la variante en RefSeq, la frecuencia de existencia de dicha variante respecto al proyecto 1000 genomas y una serie de fenotipos asociados a la variante (información accesible en la dirección del servidor de dayhoff aportada, en dos ficheros .xlsx)

5.2. Resumen de variantes comunes AML y MM

De una forma más concreta y para no alargar mucho el informe, nosotros hemos resumido variantes comunes encontradas en ambas muestras de pacientes, a priori relacionadas únicamente con fenotipo AML en la bibliografía.

Localización	Alelo	Consecuencia	Symbol	Biotipo	ID RefSeq	Ref alelo	F. Exist	Fenotipo
2:25264789-25264789	T	IV, NMDt	DNMT3A	PC, NMD	rs6546045	C	0.7602	AML
21:34921268-34921268	G	IV, NMDt	RUNX1	PC, NMD	rs4816500	A	0.9637	AML
12:11672871-11672871	T	IV, NCVT	ETV6	PC, PCCDSND	rs928949	C	0.8736	AML
2:25264789-25264789	G	IV, NCVT	LPP	PC, PCCDSND	rs13092374	A	0.8145	AML

Cuadro 2: Anotación VEP variantes presentes en AML y MM con fenotipo AML

Leyenda abreviaturas anotación VEP: *IV* (intron variant), *NCVT* (non coding transcript variant), *NMD* (nonsense mediated decay), *PC* (protein coding), *PCCDSND* (protein coding CDS not defined).

El hecho de que ambas enfermedades presenten variantes comunes puede deberse a que como vemos, su frecuencia de existencia es bastante alta y además ambas enfermedades son neoplasias mieloides, por lo que podrían ser variantes que se dieran en ambos fenotipos al ser enfermedades estrechamente ligadas.[14]

Uno de los genes donde se encuentran variantes para ambos pacientes, incluida una mutación común a ambos como se indica en la tabla 2, es el gen *DNMT3A*.

La presencia de AML frecuentemente va acompañada de mutaciones en este gen, hasta en un 20 % de los casos, y la mayoría de variaciones presentes en esta región del genoma suelen ir asociadas con cambios en la metilación del ADN, tanto de forma positiva (hipermetilación) como de forma negativa (hipo-metilación) [15].

Otra variante común para ambos pacientes se encuentra ubicada en el gen *RUNX1*, gen el cual codifica para un factor de transcripción que se expresa de forma constitutiva entre casi todas las líneas celulares y que juega un papel crítico en el desarrollo hematopoyético, regulando la expresión de genes con funciones específicas en la hematopoyesis como *IL-3* o *PF4*. La mayoría de mutaciones presentes en este gen y asociadas con este tipo de enfermedad u otras enfermedad hematopoyéticas suelen acarrerar pérdida de función [16].

ETV6, gen el cual codifica para un factor de transcripción clave en la hematopoyesis es otro de los genes en los que hemos encontrados mutaciones para ambos pacientes. En este caso, se trata de un gen cuyas mutaciones suelen estar muy frecuentemente asociadas con todo tipo de enfermedades hematopoyéticas, no solo AML, principalmente debido a que juega un papel crucial en la diferenciación mieloide y promueve la proliferación de progenitores comunes de eritrocitos y megacariocitos, entre otras cosas [17].

5.3. Perfil de mapeo de variantes comunes

Por último, mostramos el perfil de mapeo de una de las variantes en su localización concreta para ambas muestras (AML y MM). Hemos escogido la variante asociada al gen DNMT3A y la herramienta empleada ha sido IGV [7]:

Como podemos apreciar, se observa en rojo la lectura en ambos archivos de lecturas para

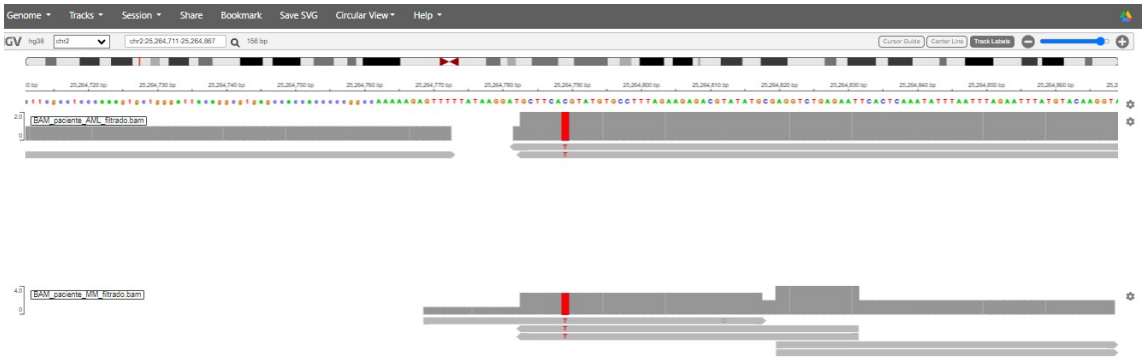

Figura 4: Lecturas variante y región específica DNMT3A

AML y MM con el alelo de ambas variantes (T) y el alelo que corresponde a esa posición en el genoma de referencia (C). Además, vemos que no hay una gran cantidad de lecturas mapeadas por lo que, la cobertura de esta variante será baja y por tanto no podemos asegurar que no se haya dado por azar.

6. Conclusiones

Como conclusión, destacar que por medio de las herramientas mencionadas a lo largo del informe, hemos sido capaces de explorar datos de secuenciación de genoma completo en base a variaciones de un solo nucleótido (SNPs) presentes en ambas muestras.

Durante dicha exploración, nos hemos centrado en encontrar variantes asociadas al fenotipo de interés (Leucemia mieloide aguda) y las hemos anotado y caracterizado en la medida de lo posible en ambos pacientes, es decir, hemos visto mutaciones asociadas a AML en un paciente diagnosticado con MM.

Esta búsqueda, ha sido dirigida por una búsqueda bibliográfica de genes donde se ha visto que mutaciones en ellos están estrechamente relacionados con los fenotipos mencionados.

Los resultados obtenidos han mostrado que, aunque existen variantes comunes en ambos fenotipos, estas no han pasado un filtro estricto de profundidad, por lo que no podemos afirmar que dichas variantes no se hayan dado de forma azarosa y tampoco podemos afirmar que estén estrechamente relacionadas con el fenotipo presente en estos pacientes.

En última instancia, destacar que para realizar un análisis más exhaustivo, se recomienda recurrir a ensayos de secuenciación más dirigidos a la zona de interés (regiones de los genes considerados en el estudio) a modo de control de calidad en la presencia de dichas variantes encontradas.

Referencias

[1] Linde A. Miles et al. "Single-cell mutation analysis of clonal evolution in myeloid malignancies". en. En: *Nature* 587.7834 (nov. de 2020). Publisher: Nature Publishing Group, págs. 477-482. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2864-x. URL: https://www.nature.com/articles/s41586-020-2864-x (visitado 18-04-2024).

[2] The Galaxy Community. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update". En: *Nucleic Acids Research* 50.W1 (abr. de 2022), W345-W351. ISSN: 0305-1048. DOI: 10.1093/nar/gkac247. eprint: https://academic.oup.com/nar/article-pdf/50/W1/W345/45189566/gkac247.pdf. URL: https://doi.org/10.1093/nar/gkac247.

[3] R. Leinonen, H. Sugawara y M. Shumway and. "The Sequence Read Archive". En: *Nucleic Acids Research* 39.Database (nov. de 2010), págs. D19-D21. DOI: 10.1093/nar/gkq1019. URL: https://doi.org/10.1093%2Fnar%2Fgkq1019.

[4] S. Andrews. "FastQC A Quality Control tool for High Throughput Sequence Data". URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[5] Heng Li y Richard Durbin. "Fast and accurate long-read alignment with Burrows–Wheeler transform". En: *Bioinformatics* 26.5 (ene. de 2010), págs. 589-595. DOI: 10.1093/bioinformatics/btp698. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbtp698.

[6] Petr Danecek et al. "Twelve years of SAMtools and BCFtools". En: *GigaScience* 10.2 (feb. de 2021). giab008. ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008. eprint: https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf. URL: https://doi.org/10.1093/gigascience/giab008.

[7] Helga Thorvaldsdóttir, James T. Robinson y Jill P. Mesirov. "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration". En: *Briefings in Bioinformatics* 14.2 (mar. de 2013), págs. 178-192. ISSN: 1467-5463. DOI: 10.1093/bib/bbs017. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3603213/ (visitado 19-04-2024).

[8] Erik Garrison y Gabor Marth. *Haplotype-based variant detection from short-read sequencing*. 2012. arXiv: 1207.3907 [q-bio.GN].

[9] Pablo Cingolani et al. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff". En: *Fly* 6.2 (abr. de 2012), págs. 80-92. DOI: 10.4161/fly.19695. URL: https://doi.org/10.4161%2Ffly.19695.

[10] Moyra Smith. *LEUKEMIA, ACUTE MYELOID; AML*. Visitada a 19/04/2024. 1/14/1997. URL: https://www.omim.org/entry/601626#creationDate.

[11] Anne Murati et al. "Myeloid malignancies: mutations, models and management". En: *BMC Cancer* 12.1 (jul. de 2012), pág. 304. ISSN: 1471-2407. DOI: 10.1186/1471-2407-12-304. URL: https://doi.org/10.1186/1471-2407-12-304 (visitado 18-04-2024).

[12] Pablo Cingolani et al. "Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift". En: *Frontiers in Gene-*

tics 3 (2012). DOI: 10.3389/fgene.2012.00035. URL: https://doi.org/10.3389%2Ffgene.2012.00035.

[13] William McLaren et al. "The Ensembl Variant Effect Predictor". En: *Genome Biology* 17.1 (6 de jun. de 2016), pág. 122. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0974-4. URL: https://doi.org/10.1186/s13059-016-0974-4 (visitado 18-04-2024).

[14] Klaus H. Metzeler et al. "Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia". eng. En: *Blood* 128.5 (ago. de 2016), págs. 686-698. ISSN: 1528-0020. DOI: 10.1182/blood-2016-01-693879.

[15] Dong Jin Park et al. "Characteristics of DNMT3A mutations in acute myeloid leukemia". En: *Blood research* 55.1 (mar. de 2020), págs. 17-26. ISSN: 2287-979X. DOI: 10.5045/br.2020.55.1.17. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7106122/ (visitado 18-04-2024).

[16] Motoshi Ichikawa et al. "A role for RUNX1 in hematopoiesis and myeloid leukemia". En: *International Journal of Hematology* 97.6 (1 de jun. de 2013), págs. 726-734. ISSN: 1865-3774. DOI: 10.1007/s12185-013-1347-3. URL: https://doi.org/10.1007/s12185-013-1347-3 (visitado 18-04-2024).

[17] Simone Feurstein y Lucy A. Godley. "Germline ETV6 mutations and predisposition to hematological malignancies". En: *International Journal of Hematology* 106.2 (1 de ago. de 2017), págs. 189-195. ISSN: 1865-3774. DOI: 10.1007/s12185-017-2259-4. URL: https://doi.org/10.1007/s12185-017-2259-4 (visitado 18-04-2024).