

Curso Genómica 2024
Práctico 15 de noviembre

FILTRADO, CLASIFICACIÓN Y VISUALIZACIÓN DE VARIANTES GÉNICAS
EN DATOS DE SECUENCIACIÓN MASIVA

Objetivos

- Visualizar variantes génicas encontradas en secuenciado de un panel de genes correspondientes a Leucemia Aguda Mieloblástica utilizando el programa IGV para la visualización de datos genómicos de gran escala.
- Filtrado y clasificación de variantes utilizando tablas generadas en QIAGEN CLC Genomics Workbench, wANNOVAR (<https://wannovar.wglab.org/>) y Ensembl Variant Effect Predictor (VEP) (<https://www.ensembl.org/info/docs/tools/vep/index.html>).

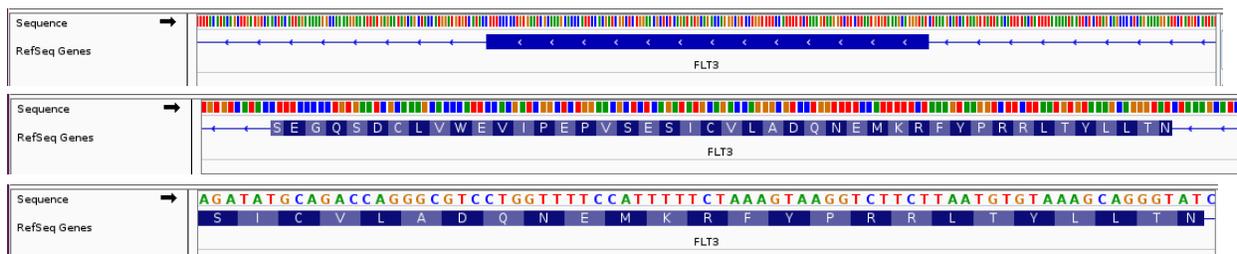
Visualización de secuencias en IGV

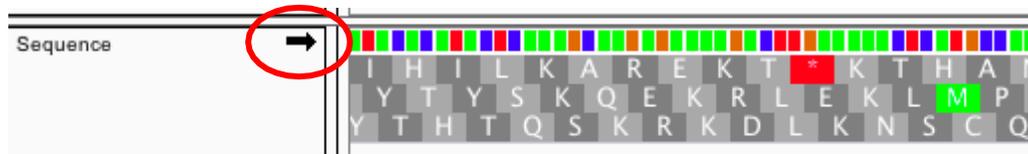
El programa Integrative Genomics Viewer (IGV) permite visualizar varios tipos de datos involucrados en cualquier análisis de NGS que use un genoma de referencia, incluyendo su mapeo, anotación y variantes génicas. En este programa se pueden analizar diferentes tipos resultados de análisis de gran escala.

Hoy vamos a visualizar un experimento de secuenciado de genes candidatos de Leucemia Mieloblástica Aguda (LMA) en una muestra de médula ósea de un paciente al momento del debut de la enfermedad.

Este archivo tiene el resultado de una secuenciación pareada de exones de 30 genes candidatos, utilizando la plataforma Illumina. Estos datos se encuentran en un archivo con formato *.bam*. Se requiere que el archivo esté ordenado e indexado por coordenadas.

1. En el programa IGV, cargar primero el genoma de referencia. Ir a Genomes > Load genome from Server... o cargar a partir del menú desplegable de la esquina superior izquierda de la ventana. Elegir el archivo de referencia Human hg19. Ver que aparecen los genes RefSeq en tracks (sección inferior de la ventana). Si se hace suficiente zoom se ve la secuencia nucleotídica. En la sección "Tracks", se pueden ver otras cosas útiles, tales como el sentido de la hebra de ADN mostrada. Una flecha hacia la derecha indica que se está mostrando la hebra positiva. También se puede mostrar la traducción en los 3 marcos de lectura de la secuencia mostrada: haciendo botón derecho en la pista RefSeq y seleccionando "Show translation".





En el Track de anotación, la opción “Expanded” permite ver mejor características solapantes, tales como múltiples transcritos para un mismo gen.

2. Cargar el archivo de datos con las secuencias alineadas. Ir a File > Load from File... y seleccionar el archivo LMA_1.bam. Estamos cargando las secuencias alineadas de los exones de 30 genes candidatos.
 Al cargar el archivo de datos se crean 2 tracks asociados: el *Track de alineamiento* para ver las secuencias y encima el de *cobertura*, con la suma de las secuencias para cada posición genómica.
 Conviene ajustar preferencias para mejor visualización:
 View > Preferences... > Alignments (Pestaña) > cambiar “Visibility range Threshold (Kb)” a 500.
3. Evaluar la secuenciación de los siguientes genes: FLT3, TET2, NPM1 y DNMT3A.
 Para ir a una región específica se puede buscar en la celda de la parte superior de la ventana. Se puede buscar por ubicación cromosómica o por nombre de gen.
 Hacer zoom para ver los alineamientos. Parándose sobre cada secuencia se ve toda su información. IGV usa marcadores de color para señalar potenciales alteraciones genéticas con respecto a la referencia. Esto incluye por ejemplo SNPs, inserciones y deleciones.
 - a. Buscar alteraciones de estos tipos en las secuencias. Para interpretar lo que estamos viendo, buscar qué representa cada color en el siguiente enlace: <http://software.broadinstitute.org/software/igv/AlignmentData#paired>
 - b. ¿La cobertura es pareja en todos los exones del mismo gen? ¿Y entre diferentes genes?
 - c. ¿Cómo discriminar entre una variante real y un error de secuenciado?
 - d. ¿El paciente tiene alteraciones genéticas en los 4 genes analizados?
4. Cargar el archivo de datos con las variantes génicas encontradas ya filtradas: SNPs_AML1.vcf. Ver las variantes encontradas en este paciente. Con la información que brinda el IGV, ¿se puede dilucidar si las variantes encontradas son somáticas o germinales?

Filtrado y clasificación de variantes

Para filtrar y clasificar las variantes es necesario obtener más información sobre ellas, esto podría hacerse de manera manual buscando una por una en bases de datos, pero existen otras herramientas que permiten agregar información de manera rápida, tales como wANNOVAR y VEP. Las mismas requieren como entrada al archivo en formato VCF generado en CLC Genomics Workbench y nos dan una salida en formato de tabla wANNOVAR y VEP, ¿qué información nos agregan estos programas?