UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Llamada y anotado de variantes genéticas de estudios de NGS utilizando Galaxy y VEP. Archivos VCF y VEP.

Apellidos, nombre	Cardona Serrate, Fernando (fcardona@btc.upv.es)
Departamento	Departamento de Biotecnología. Escuela Técnica Superior de Ingeniería Agronómica y del Medio Natural.
Centro	Universitat Politècnica de València

1 Resumen de las ideas clave

En este artículo vamos a describir paso a paso cómo analizar los datos de secuenciación de segunda generación usando el servidor Galaxy y el servidor *Variant Effect Predictor*. En concreto, estudiaremos cómo analizar los resultados que obtenemos en secuenciación de nueva generación (NGS, de *Next Generation Sequencing*). Como ejemplo se utilizan archivos obtenidos mediante la plataforma Illumina (California, Estados Unidos) de segunda generación, pero es aplicable a cualquiera de las plataformas de secuenciación NGS.

2 Objetivos

Una vez que el estudiante lea con detenimiento este documento, será capaz de:

- Obtener archivos VCF y VEP a partir de archivos BAM
- Comprender y filtrar el contenido de los archivos VCF
- Comprender y analizar el contenido de los archivos VEP
- Obtener un archivo tipo Excel con los resultados obtenidos

3 Introducción

La secuenciación de ADN ha sido revolucionaria en los campos de conocimiento de la biología molecular y la genética, ya que permite conocer la secuencia y estructura de los genes, así como las de otros elementos reguladores importantes en la regulación de la expresión génica. En el caso de humanos, es ampliamente conocido el proyecto genoma humano (1990-2001), que permitió conocer casi la totalidad de la secuencia del genoma humano. En este proyecto se utilizó el método de Sanger, conocido también primera generación de secuenciación [1].

Más recientemente (2008), las técnicas de secuenciación de segunda generación han permitido una secuenciación mucho más rápida de genomas humanos completos, en cuestión de días [2].

Por último (2010), los secuenciadores de tercera generación permiten obtener secuencias mucho más largas que los de segunda generación, facilitando el ensamblaje y análisis informático, y además no necesitan de un enriquecimiento previo de la muestra [3].

El servidor Galaxy [4] es un sistema gratuito y de código abierto para el análisis de datos, la creación de flujos de trabajo, la formación y la educación, la publicación de herramientas, la gestión de infraestructuras, entre otros, que facilita el análisis de secuencias NGS sin necesidad de tener conocimientos avanzados de bioinformática. Esta herramienta aglutina distintas herramientas bioinformáticas, de análisis NGS y de otros tipos (por ejemplo, estructura de proteínas), que procesan los datos en su propia nube y están siempre accesibles. De esta forma, es una herramienta muy adecuada para este tipo de procesos en usuarios no avanzados con recursos de hardware limitados y/o pocos conocimientos informáticos. Es posible darse de alta en el sistema de forma gratuita con una cuenta académica en https://usegalaxy.org/.

El servidor VEP (https://www.ensembl.org/info/docs/tools/vep/index.html; Variant Effect Predictor, ENSEMBL) [5], permite predecir el efecto de las variantes genéticas obtenidas por secuenciación, así como aglutinar las evidencias científicas de los efectos de los cambios o la función de los genes obtenidos en la misma.

4 Desarrollo

A la hora de realizar el análisis de unos resultados de NGS, la primera tarea a realizar es cargar los archivos de secuencia en Galaxy. Existen varias formas de hacerlo, aquí se ejemplifica una de ellas utilizando archivos BAM obtenidos a partir de 4 secuencias de la plataforma *Illumina* de un mismo trabajo disponibles en el repositorio.

4.1 Llamada de variantes desde los archivos BAM

Utilizando Galaxy, ex posible obtener, a partir de los archivos de secuencia FASTQ, archivos de mapeo de las secuencias sobre el genoma de referencia (SAM o BAM). Estos archivos, después de comprobar que tienen la calidad suficiente, pueden convertirse en archivos VCF tras hacer la "llamada de variantes" (*variant calling*). Estos archivos proporcionan un listado de variantes que puede ser utilizado para estudiar los cambios de tus secuencias respecto al genoma de referencia.

El filtrado de archivos VCF (*Variant Call Format*) se divide a grandes rasgos en dos tipos: filtrado previo y posterior al *variant calling*. El filtrado previo a la llamada es aquel en el que se decide no emitir una línea de variante al archivo VCF. El filtrado posterior a la llamada es aquel en el que se emite una variante junto con métricas auxiliares, como la calidad y la profundidad, que luego se utilizan para un filtrado posterior.

De forma previa al *variant calling*, puede ser necesario un filtrado por calidad de los archivos BAM, si estos no tienen la calidad suficiente. Esto puede hacerse en Galaxy, por ejemplo, utilizando las herramientas "Filter SAM or BAM, output SAM or BAM" o "BAM filter".

La llamada de variantes puede realizarse en Galaxy utilizando, por ejemplo, Lofreq, según se indica en la imagen 1.

Posteriormente se puede filtrar el archivo *VCF filter*, por añadiendo "*genotype filter*" con los valores por defecto, e indicando "Yes" en "*Filter entire records, not just alleles*" (imagen 2).

En la medida de lo posible, el filtrado posterior a la llamada debe ser flexible, ajustando las reglas de filtrado y producir rápidamente el archivo de variantes revisado sin perder variantes que puedan ser interesantes. Sin embargo, hay ocasiones en las que una variante no es aceptable en ningún caso, por lo que se puede especificar el número mínimo de lecturas necesarias. Este tipo de filtrado se realiza normalmente en *bcftools mpileup* o *bcftools call*, y no se tratan en este artículo docente.

Posteriormente se pueden combinar archivos VCF diferentes (por ejemplo, de genomas de diferentes pacientes de la misma familia) con *VCFcombine* (valores por defecto) y analizar los VCFs combinados, con las variantes que tienen en común.

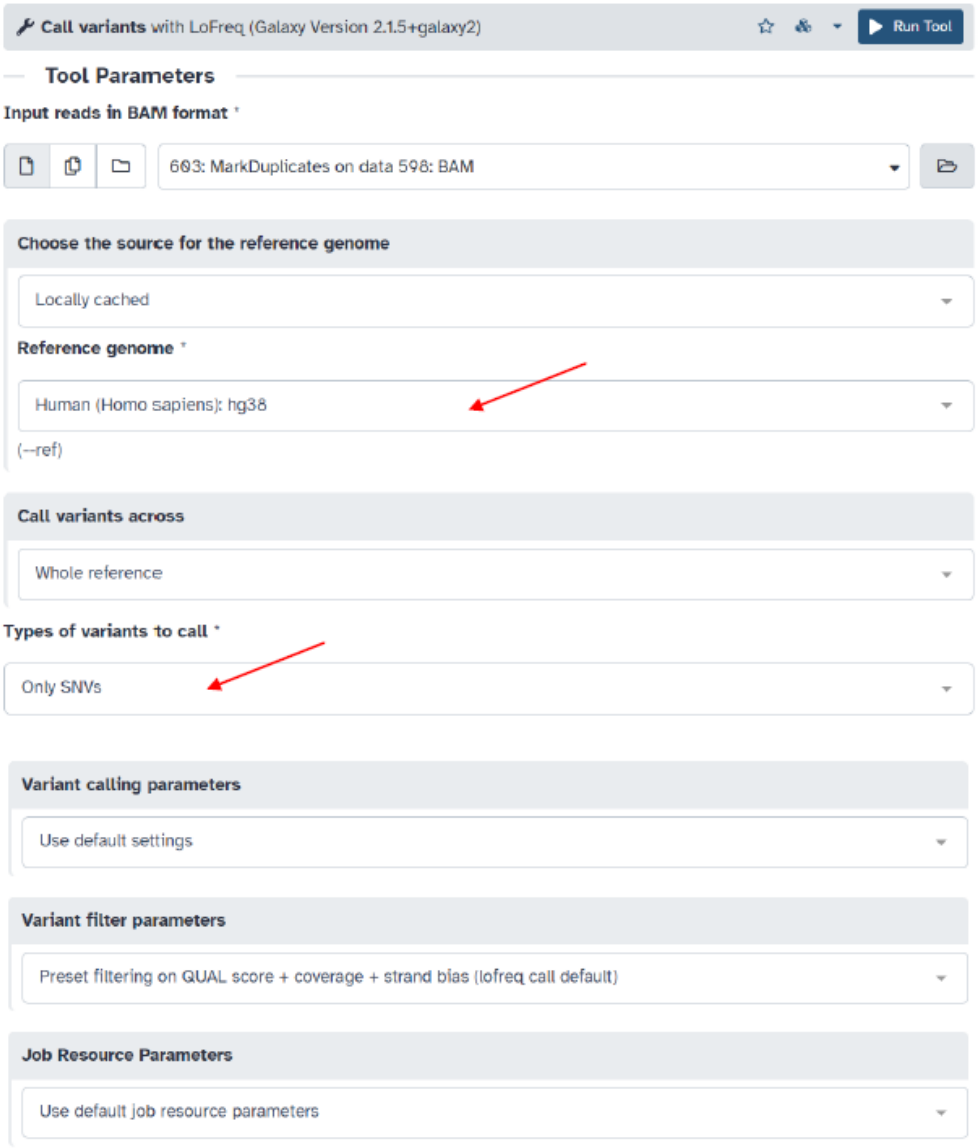
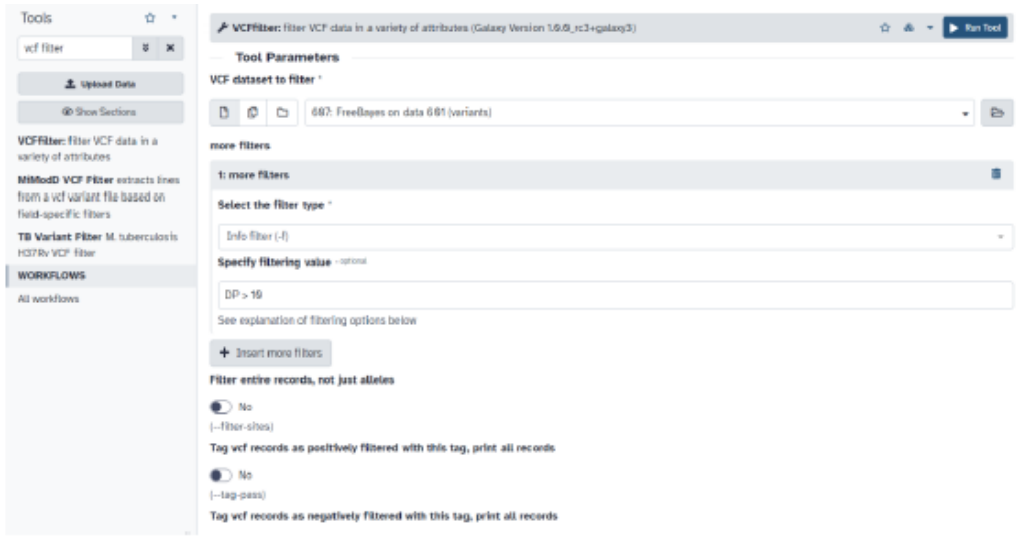


Imagen 1. Llamada de variantes utilizando Lofreq en Galaxy



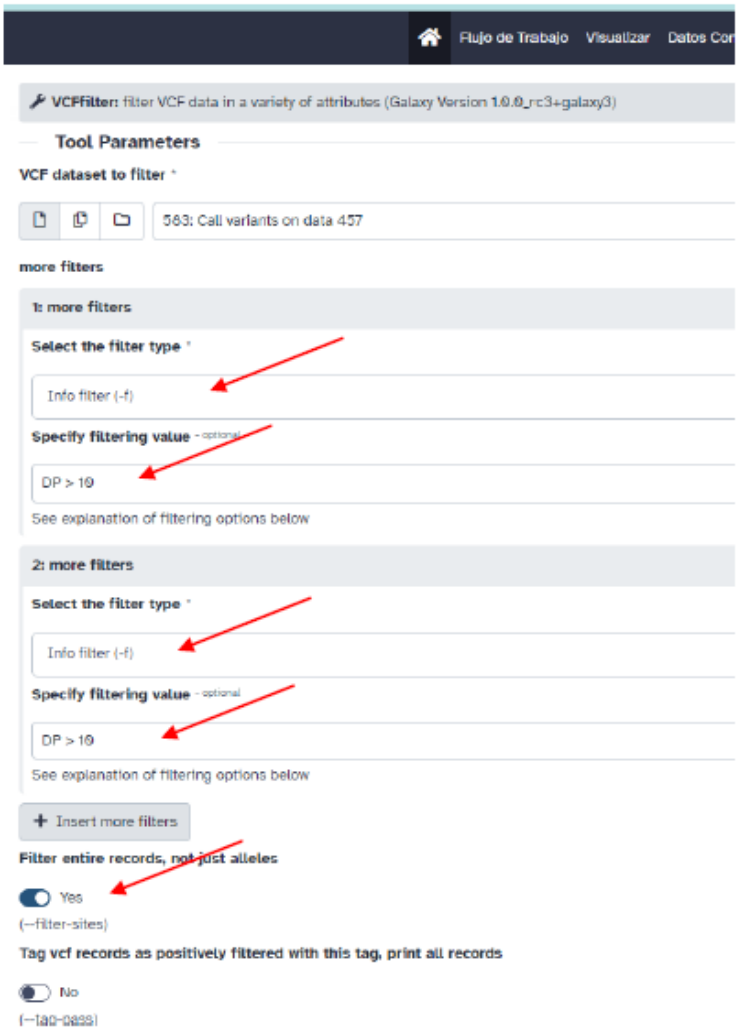


Imagen 2. Uso de VCFfilter en Galaxy para el filtrado posterior de variantes.

4.2 Análisis de los archivos VCF con *Variant Effect Predictor*.

Una vez obtenido el archivo VCF con las variantes, debe realizarse un análisis funcional de las mismas con el objetivo de anotar, seleccionar y priorizar las variantes que sean *a priori* más interesantes en nuestro estudio.

Ensembl Variant Effect Predictor (https://www.ensembl.org/vep) es un potente conjunto de herramientas para el análisis, la anotación y la priorización de variantes genómicas en regiones codificantes y no codificantes. Puede simplificar y acelerar la interpretación de variantes en una amplia gama de diseños de estudios, ya que proporciona acceso a una amplia colección de anotaciones genómicas, con una variedad de interfaces para adaptarse a diferentes necesidades, y opciones sencillas para configurar y ampliar el análisis. Es de código abierto, de uso gratuito y admite la total reproducibilidad de los resultados.

Tras darse de alta en con una cuenta institucional académica (por ejemplo, la de la Universidad), se pueden subir los archivos VCF (por separado y combinados), y obtener la anotación de variantes modificando, por ejemplo, los parámetros que aparecen por defecto como se indica en la imagen 3. Esto es únicamente un ejemplo para este caso concreto, por lo que el análisis debe ajustarse a las necesidades del trabajo en cuestión.

También existe flexibilidad en la anotación del efecto funcional de las variantes genéticas encontradas. En la imagen 4 se muestra un ejemplo de los parámetros para anotar en los diferentes transcritos el efecto funcional y los posibles fenotipos. En la imagen 5 se muestran los posibles parámetros a utilizar para la anotación del efecto funcional de las variantes genéticas encontradas en lo relativo a su patogenicidad y efectos en el *splicing*.

4.3 Análisis de los archivos VEP con *Microsoft excel*.

Los archivos generados en el apartado anterior se pueden descargar como .txt (entre otros) y abrirse en Excel. Si al intentar abrirlo no aparece la opción hay que marcar la casilla “todos los archivos” o poner * (+intro). Importar el archivo a excel indicando que contiene encabezados, dejan do el resto como aparece por defecto, según se indica en la imagen 6.

Posteriormente pueden ocultarse o eliminarse las columnas que no tienen información para facilitar el trabajo. Suele ser útil inmovilizar la fila superior (encabezados) (Vista > inmovilizar fila superior).

En este ejemplo se va a filtrar los datos según los criterios considerados. Datos > Filtro; seleccionas en el encabezado los valores que quieres. Ejemplo:

Polyphen = Probably damaging > 0.95

SIFT = deleterous + deleterous low confidence

Polyphen = probably damaging >=0.9

BayesDel pred = D (AF y no AF)

Clin Pred = D

MutationAssessor_rankscore >0.9

Provean >0.9

Con este ejemplo de priorización se llega directamente a los cambios responsables de la enfermedad de esta familia [6], pero a veces puede no ser tan claro o directo, ya que no

sabemos a priori cuáles son los cambios responsables de la enfermedad. Normalmente es necesario combinar diferentes estrategias de filtrado con búsquedas bibliográficas, otros predictores, o incluso estudios funcionales para llegar a identificar los cambios responsables. Pueden ayudar los datos obtenidos de VEP a priorizar con qué variantes comenzar el estudio (por ejemplo, genes relacionados con la enfermedad o mutaciones relacionadas con otras enfermedades, afectación de la función de la proteína, etc.). También se pueden ordenar los datos (por ejemplo, de mayor a menor valor de una columna). De esta forma dejamos visibles solo los datos utilidad, o bien los jerarquizamos para priorizarlos para su posterior análisis.

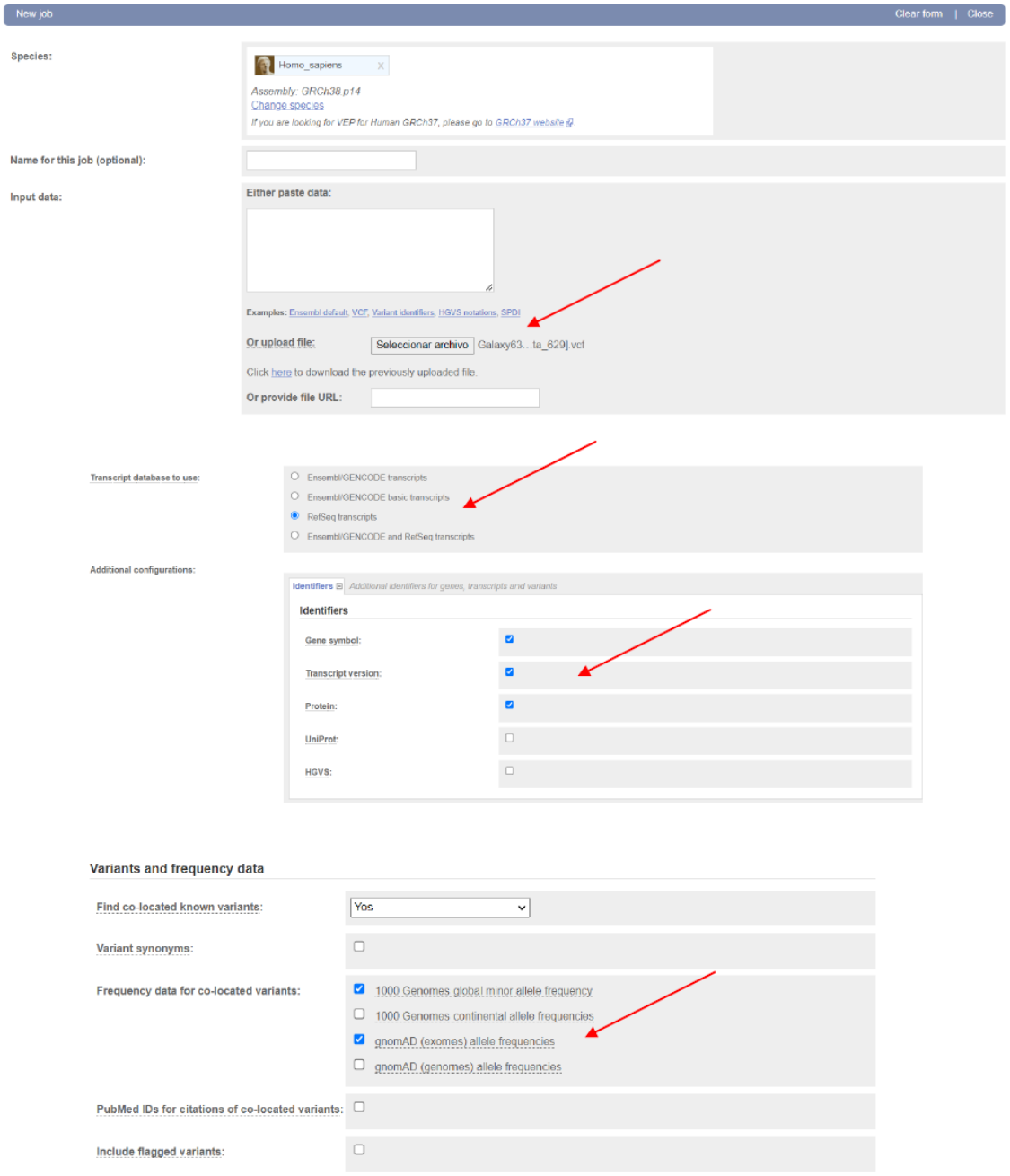


Imagen 3. Ejemplo d parámetros básicos en VEP para la anotación de variantes.

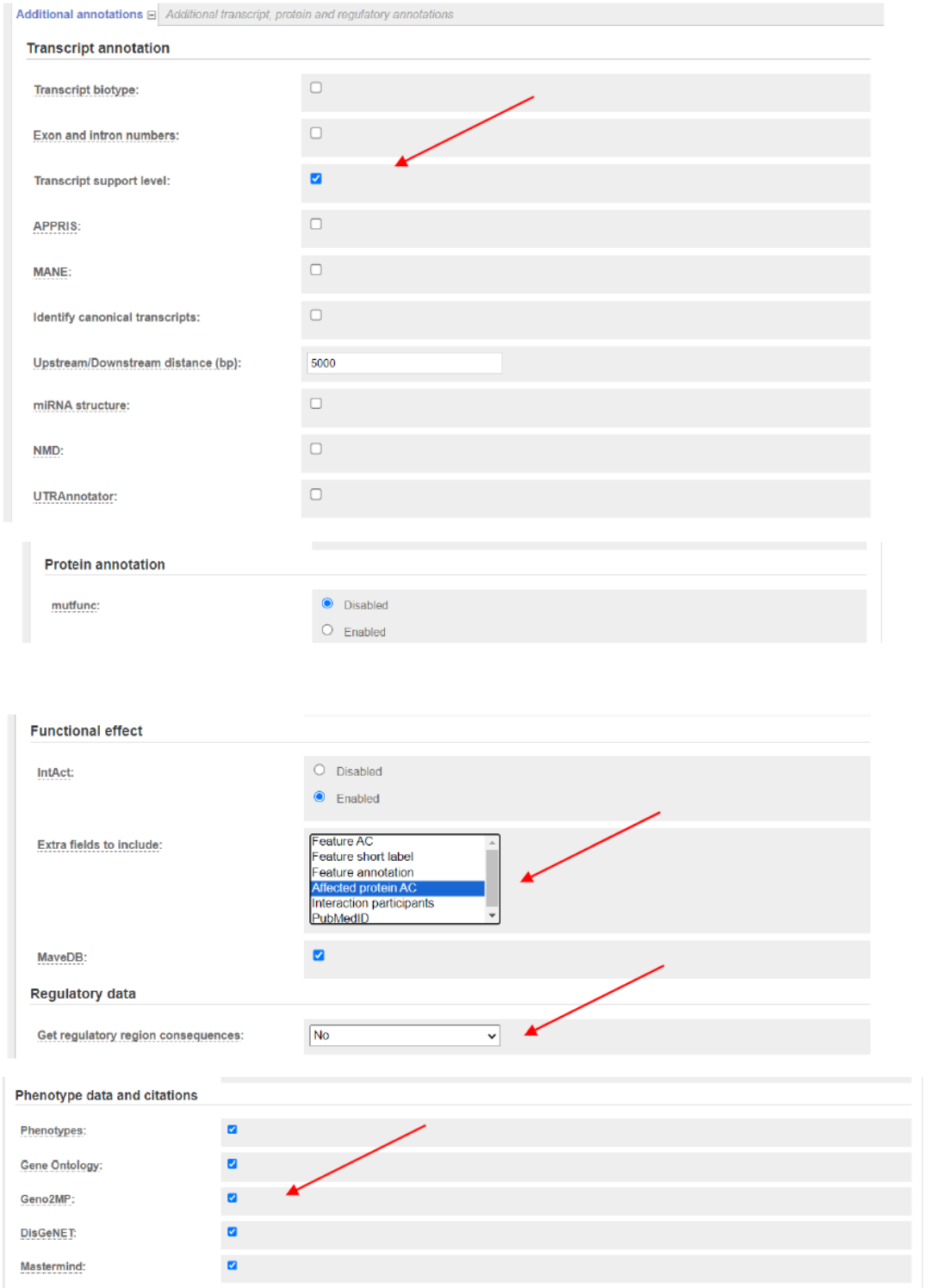


Imagen 4. Ejemplo de parámetros a utilizar en VEP para la anotación del efecto funcional de las variantes genéticas encontradas.

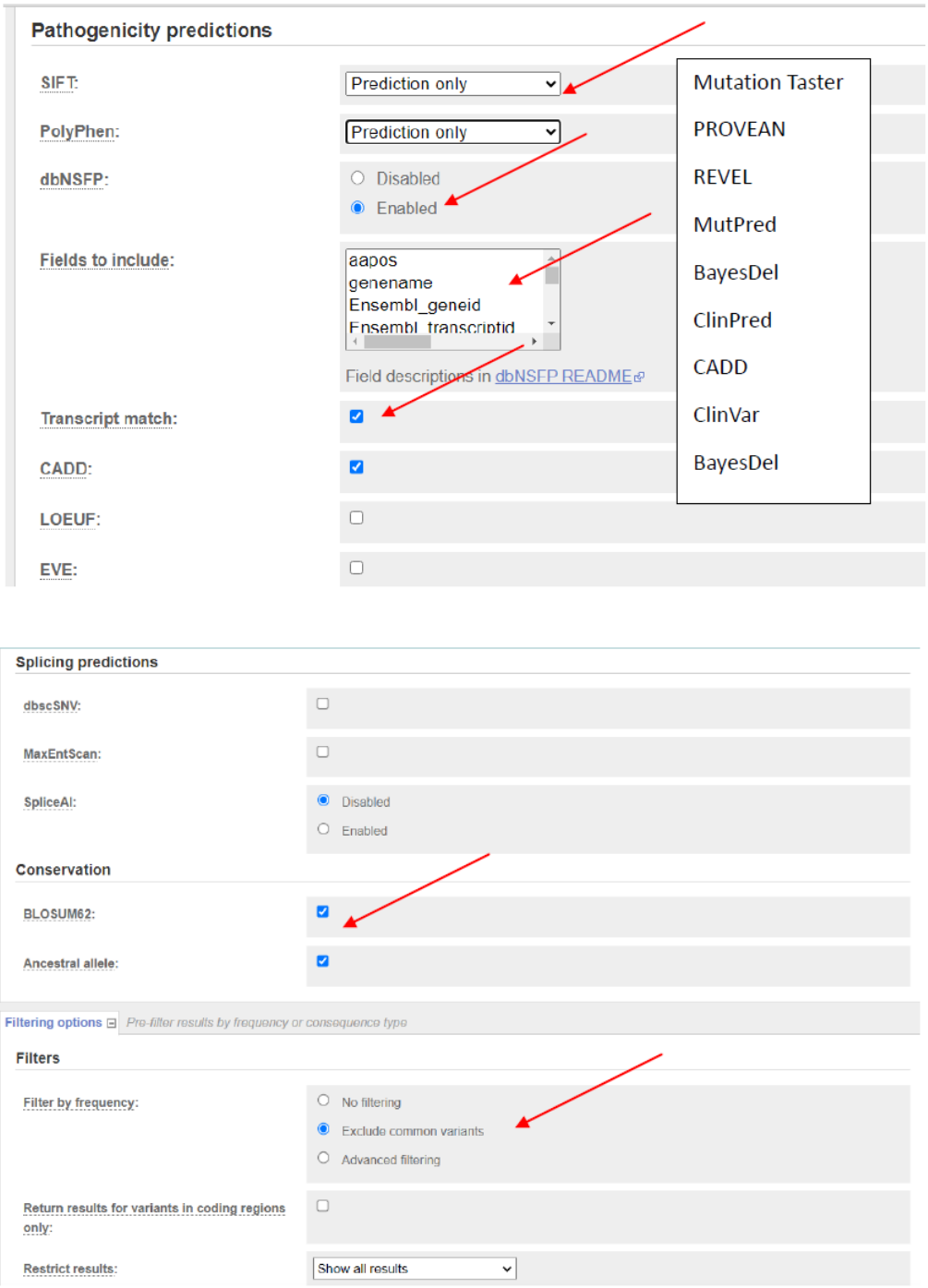


Imagen 5. Ejemplo de parámetros a utilizar en VEP para la anotación del efecto funcional de las variantes genéticas encontradas en lo relativo a su patogenicidad y efectos en el splicing.

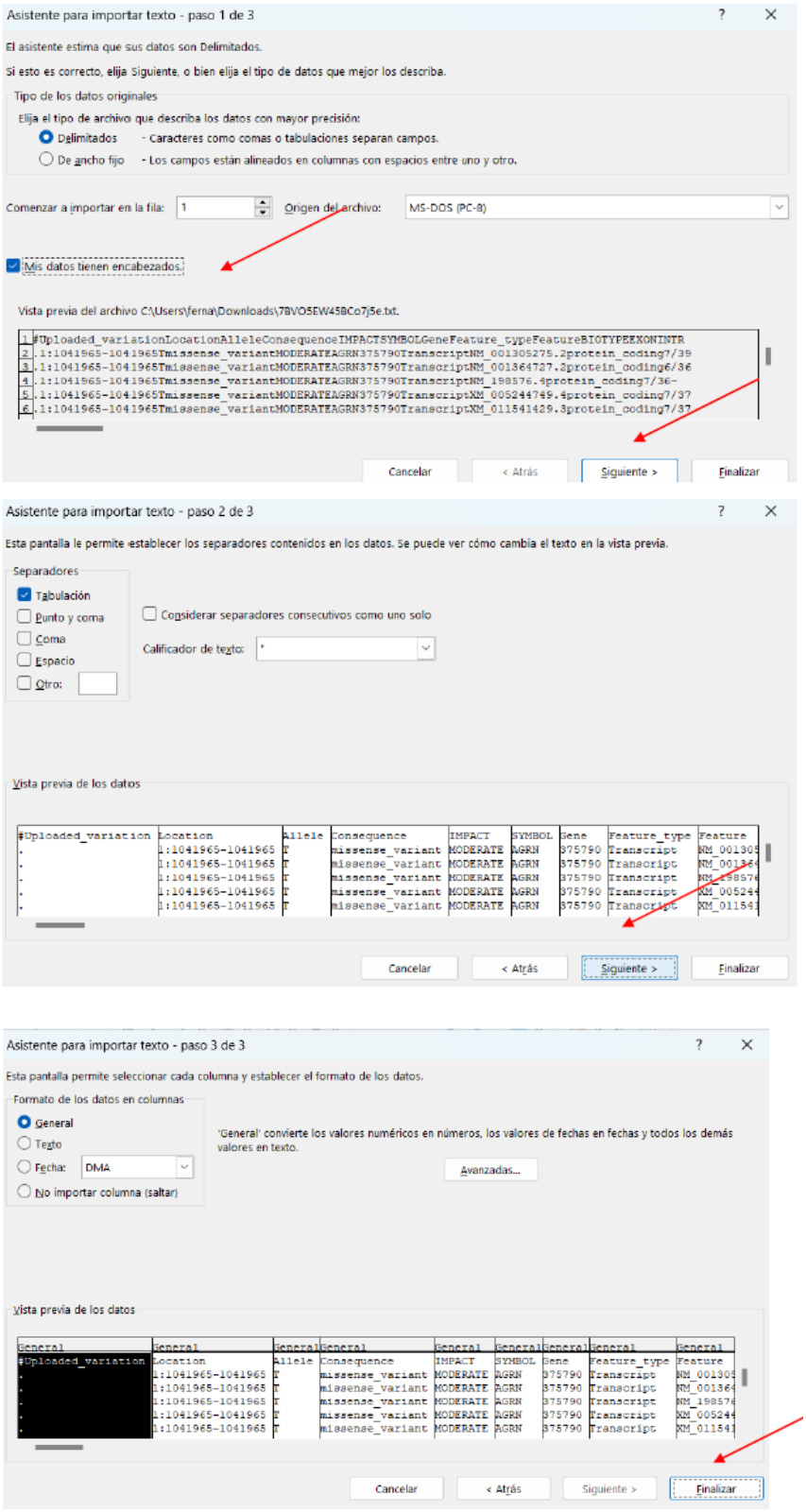


Imagen 6. Importación de los archivos obtenidos de VEP a Excel.

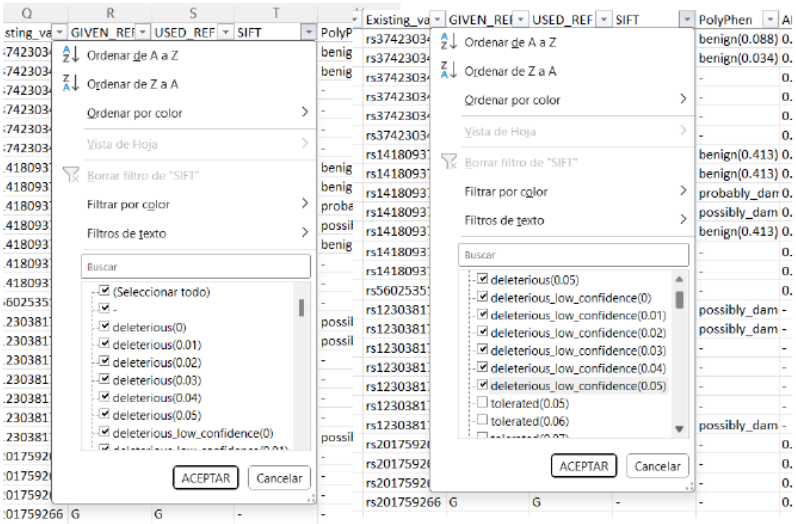


Imagen 7. Filtrado y priorización de variantes utilizando Excel.

5 Cierre

A lo largo de este objeto de aprendizaje hemos visto cómo obtener los archivos de variantes genéticas de NGS tipo .vcf. Además, hemos visto como realizar el filtrado y combinado de los mismos, todo ello utilizando herramientas disponibles en la herramienta Galaxy. También se ha estudiado un ejemplo de cómo anotar las variantes genéticas utilizando *Ensembl Variant Effect Predictor*, para posteriormente filtrar y priorizar los resultados en Excel con el objetivo de continuar el análisis con las variantes más interesantes.

6 Bibliografía

1- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature.* **409**, 860–921 (2001). https://doi.org/10.1038/35057062

2- Wadman, M. James Watson's genome sequenced at high speed. *Nature.* 452, 788 (2008). https://doi.org/10.1038/452788b

3- Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*. **7(6)**, 461-465 (2010). https://doi.org/10.1038%2Fnmeth.1459

4- The Galaxy Community. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Research*. gkae410 (2024). https://doi.org/10.1093/nar/gkae410

5- *Ensembl Variant Effect Predictor*. https://www.ensembl.org/info/docs/tools/vep/index.html

6- Yang et al. From Hyper- to Hypoinsulinemia and Diabetes: Effect of KCNH6 on Insulin Secretion, *Cell Reports* **25**, 3800–3810 (2018) https://doi.org/10.1016/j.celrep.2018.12.005