

Tenemos un archivo en el formato VCF sobre el cual podemos aplicar filtros para enfocarnos en variantes relevantes a nuestro estudio de familia monogénica. Ahora vamos a aplicar anotación adicional a nuestro archivo.

**Los siguientes ejercicios deben estar realizados sobre el archivo Example.vcf que has recibido por correo electrónico.** El archivo está truncado por diseño para que sea más manejable para este curso. (nota: la posición máxima está en el cromosoma 10). Abres el archivo en Excel o equivalente.

Q1. Variantes llamadas con un índice de confianza <Q30 seguramente son errores. Aplica el filtro en la columna requisito para ocular variantes con índice de confianza <Q30. Escribe que columna fue.

Q2. La variante que podría estar causando la enfermedad en la familia debe estar presente en el padre y la hija, pero no en la madre. Aplica el filtro en las columnas requisito para reflejar este hecho. Escribe los filtros que has usado y a que columnas se aplican.

Guarda tus resultados filtrados copiando y pegando en un nuevo archivo texto (ctrl-A, ctrl-C, ctrl-V) para evitar las comillas que aparecen con un guardado directo.

Ahora vamos a agregar anotación. Vamos a usar wAnnotator: <http://wannovar.wglab.org/>  
Llenas los campos así:

- Entrega tu correo electrónico
- Elige cualquier identificador de muestra que quieras.
- Entrega el archivo de texto que has guardado previamente.
- Duración del resultado: 1 día.
- Genoma de referencia: hg19.
- Formato de entrada: VCF.
- Definición de genes: RefSeq Gene
- Análisis individual: Todas las anotaciones.
- Modelo de enfermedad: Enfermedad Mendeliana rara dominante.

¡Listo, presiona **submit!**

En poco tiempo debería poder descargar tu archivo, descarga el archivo exoma en un formato compatible con excel (.CSV). Durante el tiempo que está esperando, verifica la **documentación Annotator**. Permitirá responder a las preguntas que siguen cuando tu archivo estará listo.

Q3. Para que una variante cause enfermedad, necesitara cambiar la estructura de la proteína que corresponde. Usando solo una columna, oculta las variantes que no van a cambiar la estructura de proteínas y también las que tienen un efecto desconocido (ing. unknown). Indica que columna usas y que palabras eliminas.

Q4. La enfermedad que afecta la familia es seguramente muy muy rara. Aplica filtros de manera que solo variantes con frecuencia <1% y variantes no presentes en bases de datos de poblaciones serán consideradas. Escribe que columnas has filtrado y con cual condiciones. (pista: no es necesario filtrar contra poblaciones específicas, solo generales).

Q5. Deberías ahora tener una lista de variantes muy manejable. ¿Puedes identificar la variante que está causando una comunicación interventricular (ing. ventricular septal defect).