Siguientes pasos

Primer filtro

Modificar el filtro para que analice todos los alineamientos de cada frame aunque con el primero ya pase el filtro y comprobar si no supone un aumento de tiempo de ejecución considerable.

Implementar en el script la generación de un informe detallado de cada secuencia que ha sido procesada por el filtro. El informe será un fichero tsv con diferentes columnas: ID frame, ID referencia, pasa filtro completo, pasa filtro parámetro 1, pasa filtro parámetro 2, … , vector de alineamientos.

Este informe es una base de datos donde se puede consultar todas las secuencias que alinean y las descartadas.

Probar con transcriptomas de los que disponemos un filtrado manual para comparar resultados.

Optimizar los parámetros para obtener el menor número de secuencias candidatas intentando que se pierda el menor número de positivos.

Segundo filtro

Después de utilizar el filtro y quedarnos con las mejores secuencias, podemos utilizar el informe para reconstruir ficheros fasta con el frame o frames que han pasado el filtro junto con las secuencias de referencias (pero sin estar visualmente alineadas con mafft, sin guiones).

Luego podemos volver a hacer un alineamiento mafft más eficaz al solo tener en cuenta un único frame (2 en muy pocos casos).

Como ocurre en TRINITY_DN4925_c0_g1_i2_Frame_R3 (seq_2 de 153A), es posible que haya frames que hayan pasado el primer filtro porque ha alineado con solo un pequeño número de secuencias de referencia pero con el resto no. Esto se debe a que en el resto ha detectado un codón de stop en mitad del alineamiento.

Sin embargo, en la secuencia de referencia no se consigue detectar una subsecuencia alineada después de la posición del codón de stop del frame porque esa secuencia es más corta o ligeramente diferente en su final. Este frame debería ser descartado ya que somos conscientes de que hay un error de lectura, ya que después del codon de stop, sigue habiendo parte del péptido.

Para solucionar esto podemos aplicar de nuevo un filtro similar al primero en el que se vuelva a analizar la condición de que la secuencia no esté interrumpida por codones de stop (para evitar errores, consideramos solo las secuencias de referencia en los que la secuencia frame alineada represente un porcentaje considerable).

Si detectamos que el alineamiento más del 50% (ajustable) de las secuencias de referencia no pasa el filtro porque después del codón de stop sigue alineando, entonces se presta atención a éstas para decidir que el frame tiene un error de lectura y descartarla.

Como con estos frames nos hemos asegurado de que la parte de secuencia bien alineada que contiene el péptido completo no esté interrumpida por ningún codón de stop, es muy fácil volver a localizar a través de un vector de alineamientos la subsecuencia real. Por lo tanto,

podemos descartar todo lo que venga después del siguiente codón de stop y todo lo anterior al codón de stop anterior a la subsecuencia bien alineada si lo hubiera.

Ahora tenemos una secuencia limpia sin codones de stop. Podemos hacer un filtro en el que se tenga en cuenta que haya una metionina al inicio de la secuencia, por ejemplo, que la encuentre en el primer 20% de la secuencia.

Además, podemos ser más estrictos forzando a que el filtro solo considere válida la metionina inicial si está alineada con la metionina inicial de la secuencia de referencia. En este caso, podemos recortar todo lo que haya antes de esa metionina para obtener el frame totalmente limpio y anotado.

En las que la metionina del frame no esté alineada con la metionina inicial de la secuencia de referencia, podemos clasificarlas como “inciertas”, irían a una carpeta distinta para que sean evaluadas a mano.

El resto de las secuencias que sí han pasado el filtro pueden procesarse automáticamente con un script para obtener un listado o base de datos en el formato deseado listo para publicar (o pendiente de añadir las “inciertas” que hayan sido validadas manualmente).