ESTUDIOS IN SILICO EN BIOMEDICINA

Examen convocatoria 1

16-01-2015

Nombre:

DNI:

Tiempo para realizar el examen 60 minutos.

Cada pregunta respondida de forma correcta suma 0.5 en un total de 10 puntos.

Indicar el nombre en todas las hojas.

---

1. ¿Cuál es el formato más extendido para intercambiar datos crudos de secuenciación masiva?
   - SFF que es el estándar de algunos secuenciadores.
   - FastQ.
   - Fasta.
   - BAM pero en formato de texto.
   - Un fichero de texto separado por tabuladores, comprimido e indexado con *tabix*.
   - Un fichero de texto separado por tabuladores, así es compatible con casi cualquier programa, incluso XLS.

2. Al hacer un estudio de expresión de microRNAs usando NGS, si mapeamos contra miRBase en lugar de contra todo el genoma de referencia ...
   - ganamos tiempo de computación pero nunca podremos descubrir nuevos microRNAs.
   - de todas formas hará falta en un último paso bajarse los genomas completos para el análisis.
   - no se podrán normalizar bien los niveles de expresión de los miRNAs.
   - usaremos menos recursos de la máquina pero harán falta bastantes réplicas para hacer un estudio de la expresión diferencial.
   - podemos reducir bastante el *coverage* y abaratar el estudio.

3. ¿Cuando conviene revisar la calidad de los datos de secuenciación?
   - Al recibir los datos crudos.
   - Después de eliminar los adaptadores
   - Al recibir los datos crudos y después de cada uno de los pasos del preprocesamiento
   - Antes de la entrega final de los resultados.
   - Al recibir los datos crudos y después de eliminar los adaptadores.

4. ¿Para que se usa principalmente el paquete de análisis de datos GATK?

   - Para hacer una detección de variantes a partir de los datos crudos de secuenciación
   - Para hacer una detección de variantes a partir de los datos mapeados frente a un genoma de referencia
   - Para hacer una detección de variantes cuando los datos tienen muchos adaptadores.
   - Para hacer un ensamblado que nos permita hacer la detección de variantes.
   - Para hacer un *calling* de SNPs pero no sirve par estudiar otro tipo de variantes como inserciones y deleciones.

5. Si ya he indexado mis ficheros utilizando *samtools* ¿Necesitare volver a indexarlos en algún momento?

   - No. Sólo con un indexado es suficiente.
   - Puede que necesite indexar ficheros indexados previamente cuando vaya a utilizar algún software nuevo para analizarlos.
   - Sólo tendré que indexar las referencias (ficheros fasta) antes de hacer el mapeo.
   - Samtools no sirve para indexar ningún tipo de datos.
   - El indexado se hace con los mapeadores como *BWA* o *Bowtie*. Luego ya es redundante e innecesario.

6. Cual es el paquete integrado se software más completo que existe por el momento para hacer análisis de datos de secuenciación?

   - Samtools.
   - GATK.
   - BWA.
   - No existe ningún paquete integrado, cada utilidad esta desarrollada por separado independientemente de las demás.
   - No existe ningún software que resuelva todo por el momento pero es previsible que se termine de desarrollar en los próximos años.

7. Si no eliminamos los adaptadores de las lecturas secuenciadas ...

   - obtendremos un mejor mapeo frente al genoma de referencia.
   - obtendremos un mejor alineamiento frente al genoma de referencia.
   - habrá que hacer un ensamblado de las lecturas antes del mapeo.
   - no habrá problema salvo que el fichero fasta ocupará mas espacio porque contendrá la parte de las secuencias de los adaptadores.
   - el mapeo sera inexacto y habrá muchas lecturas que no mapearan bien.

8. Obtener muchas lecturas exactamente iguales repetidas es ...

   - normal en algunos estudios de transcriptómica.
   - normal cuando estamos leyendo datos secuenciados directamente del DNA.
   - un indicador de baja calidad de la secuenciación en los archivos fastQ.
   - muy aburrido porque siempre sale lo mismo.
   - malo para la variabilidad estadística si es que en algún momento tenemos que hacer un test estadístico para analizar los datos.

9. Obtener muchas lecturas exactamente iguales repetidas es ...

   - bueno porque así podremos comprimir mucho los datos en el momento de almacenarlos.
   - indicativo de que la amplificación del DNA no se ha realizado bien.
   - bastante interesante en algunos estudios porque indica que hay cosas muy relevantes en ese genoma.
   - una señal de que hay amplificaciones o incrementos del número de copias de los genes.
   - una señal de que hay amplificaciones o incrementos del número de copias de algunas regiones genómicas aunque no necesariamente genes.

10. Cuando indexamos un fichero de datos de NGS ...

    - creamos los índices que son código pre compilado exclusivamente para trabajar en el ordenador en el que estamos y con el conjunto de datos indexado.
    - estamos seguramente utilizando el software BWA o Bowtie.
    - creamos ficheros adicionales que tienen que tener el mismo nombre que el original.
    - creamos ficheros adicionales que contienen la información que permite al software correspondiente acceder de forma más rápida y directa a los datos.
    - comprimimos el fichero o lo hacemos binario para que ocupe menos.

11. Se pueden almacenar mapeos secuenciados utilizando protocolos *pair ends* en un mismo fichero SAM?

    - No el almacenamiento es mejor hacerlo siempre en ficheros fasta o mejor fastQ.
    - Si y en uno de los campos del fichero tabular se indica las lecturas que están emparejadas.
    - Si pero solo si esta comprimido, es decir si es un BAM.
    - No las lecturas de las parejas se tienen que guardar en dos fichero SAM separados
    - En principio en ficheros separados pero como el formato SAM no tiene cabecera se pueden concatenar los dos ficheros en uno único escribiendo las filas de uno inmediatamente después de las del otro.

12. ¿Es preferible guardar la información de las variantes de nuestro experimento en formatos PED y MAP en lugar de en VCF?

    - Sí es mejor porque así ocupan menos espacio.
    - Sí es mejor que hacerlo en VCF porque así tenemos detallada la estructura familiar de las muestras.
    - No porque no toda la información contenida en el VCF se puede integrar en el PED y MAP.
    - Sí es mejor porque realmente el VCF no mantiene la información de los genotipos de cada individuo de la muestra.
    - No, siempre es mejor guardar los VCFs aunque sea sin cabecera.

13. Las lecturas contenidas en los ficheros SAM ¿Tienen necesariamente que estar ordenadas?

    - Sí. El orden forma parte de la especificación del formato.
    - No, es sólo un capricho.
    - No pero ayuda cuando vamos a trocear los ficheros para paralelizar algún análisis.
    - No es obligatorio que estén ordenadas pero es necesario para algunos análisis o procesos por ejemplo cuando usamos algunas opciones de *samtools*.
    - Sí el orden es necesario porque si no no se puede comprimir el fichero y no podemos generar el BAM.

14. Se pueden guardar los datos de variantes específicas de cada individuo en un fichero VCF?

    - No. Sólo se guarda información general sobre las variantes encontradas en conjunto para toda la muestra
    - Sí
    - Sólo si el número de individuos es menor que el numero de variantes almacenadas
    - Sólo si el número de individuos es mayor que el numero de variantes almacenadas
    - Depende del software que utilicemos.

15. Un fichero fastQ ...

    - ocupa mucho menos que el fichero fasta correspondiente porque está codificado en *ascii*.
    - ocupa más o menos que el doble que el fichero fasta correspondiente.
    - no tiene nada que ver con un fichero fasta. Es otro tipo de fichero aunque se llame de forma parecida.
    - tiene lecturas de mucha mas calidad que un fichero fasta.
    - es igual que el fasta pero preparado para enviarlo por mail porque todas las lineas empiezan por @.

16. Los ficheros GFF, GTF, BED de anotación ...

- sirven para indicar la posición principio y final de los genes en una referencia.
- se usan sólo para algunas características genéticas: genes, transcritos, exones y *CDS o regiones codificantes*.
- deben contener además la misma secuencia de referencia que hay en el formato fasta.
- sirven para indicar todo tipo de características asociadas a una o varias secuencias de referencia.
- contienen información funcional de los genes como por ejemplo la que se puede extraer del *Gene Ontology*.

17. Al utilizar ficheros GFF, GTF, BED ... de anotación de genomas ...

- es mejor descargarlos de Ensembl o del NCBI.
- tenemos que asegurarnos de que se corresponden con la especie que estamos estudiando.
- debemos construirlos nosotros mismos y no fiarnos de otras anotaciones anteriores.
- debemos asegurarnos de que se corresponden con el fichero fasta que contiene la secuencia de referencia que estemos utilizando.
- debemos construir el fichero fasta de forma que se corresponda con la estructura del fichero de anotación.

18. ¿Se puede plantear un experimento de expresión diferencial para una especie que no esté secuenciada?

- Bueno depende de si hay uno o varios grupos biológicos o experimentales en la comparativa
- No
- Habría que secuenciar su genoma primero y luego hacer la transcriptómica
- Habría que secuenciar su genoma primero, luego hacer una detección de la posición de los genes sobre esta secuencia y luego ya se podría hacer la transcriptómica.
- Si, pero antes de poder hacer el cálculo de la expresión de los tránscritos habrá que realizar un ensamblado precisamente para definir cuáles son esos transcritos que en principio no se conocen.

19. ¿Se puede hacer un estudio de variantes a partir de datos de transcriptómica medidos con RNAseq?

- No. Para hacer un estudio de variantes hay que tomar lecturas a partir del DNA.
- No. Para hacer un estudio de variantes hay que tomar lecturas a partir del DNA y luego si se quieres se pueden combinar con las lecturas del RNA.
- Sí. Pero sólo se podrán detecta variantes en las regiones del genoma que se transcriben...
- Sí porque en cualquier caso el protocolo de secuenciación va a convertir el RNA EN cDNA.
- No porque en el transcriptoma no se refleja las "mutaciones" o variaciones de los genes.

20. ¿Harías un estudio de variantes utilizando sólo lecturas de DNA tomadas a partir del *exoma*?

- No porque esto cubre solo una parte muy pequeña del genoma total (alrededor del 2%).
- Si. Cuando tengo evidencia de que la variante que busco está en una región codificante. Así se reduce el coste económico y también el computacional.
- Si, pero soló para especies que tienen exoma.
- Si pero solo para estudiar individuos aislados. No lo haría si quisiese hacer una comparativa caso control porque los exomas no son comparables directamente.
- No.