

# ESTUDIOS IN SILICO EN BIOMEDICINA

## Examen convocatoria 2

05-02-2015

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. ¿En que se diferencia el formato FastQ del formato Fasta?

- En nada son dos nombres que se utilizan para el mismo formato de datos.
- Incluye información de la calidad de cada nucleótido leído además de las propias lecturas.
- FastQ es un fichero complementario al Fasta y contiene solo la información de la calidad de las lecturas del fichero Fasta
- Solo se diferencian en que el FastQ utiliza el símbolo @ en lugar del > para indicar los nombres de las secuencias.
- El formato FastQ contiene la información de los alineamientos.

2. ¿Cuándo es conveniente utilizar un genoma de referencia para analizar datos de NGS?

- Cuando analizamos datos de variantes en genes.
- Cuando analizamos datos de variantes en regiones intergénicas.
- Cuando hacemos estudios de transcriptómica.
- En realidad es mejor no utilizarlo nunca aunque se haga porque lo requiere la implementación de algunas herramientas.
- En realidad siempre que esté disponible ya que contar con un genoma de referencia añade mucha información en nuestro estudio.

3. ¿Cuándo es útil hacer una secuenciación utilizando la tecnología *paired-ends*?

- Cuando queremos hacer un análisis de transcriptómica y estimar los niveles de expresión de las diferentes isoformas.
- Cuando hacemos análisis de microRNAs.
- Cuando en el estudio hay casos y controles emparejados.
- Siempre: cuando queremos medir la expresión de los genes, estudiar sus variantes o cualquier otro análisis genómico.
- Es una técnica que solo se utiliza para hacer estudios de metilación.

4. ¿Cuándo es útil hacer una secuenciación utilizando la tecnología *paired-ends*?

- Cuando queremos utilizar los datos de secuenciación para hacer un ensamblado *de novo*.
- Cuando queremos combinar datos de expresión de genes con datos de variantes genómicas en un mismo estudio.
- Solamente cuando existe un genoma de referencia.
- Solo en estudios de transcriptómica.
- Siempre.

5. ¿Cuándo es útil hacer una secuenciación utilizando la tecnología *paired-ends*?

- Cuando queremos hacer un estudio de variantes. Sobre todo si queremos ser capaces de detectar inserciones o deleciones.
- Cuando queremos analizar RNA no codificante.
- Solo cuando queremos detectar genes o transcritos que se expresan a muy bajo nivel.
- Solo cuando la secuenciación sea más barata que hacerla utilizando *single-reads*?
- Solo cuando nos interesa estudiar una cantidad muy pequeña de genes.

6. La tecnología de *paired-ends* ...

- nos permite leer el principio y el final de una molécula de DNA o RNA.
- nos permite leer el centro de una molécula de DNA o RNA pero no los extremos finales.
- nos permite leer dos segmentos de una misma molécula de DNA o RNA separados entre si por un número prefijado de bases.
- nos permite hacer estudios estadísticos de datos pareados
- nos permite hacer estudios estadísticos de datos pareados incluso cuando no hay replicas biológicas en el estudio.

7. El formato SAM se desarrollo para ...

- indicar como las lecturas secuenciadas se alinean contra una secuencia de referencia.
- poder comprimirlo en un fichero binario: BAM.
- tener un fichero más pequeño que el FastQ
- para incorporar toda la información de la anotación funcional de las lecturas.
- para contener la secuencia de referencia y luego hacer el alineamiento contra ella.

8. El formato VCF se utiliza para ...

- almacenar la información de las variantes genómicas encontradas en uno o varios individuos.
- almacenar la información de las variantes genómicas encontradas en un individuo.
- almacenar las lecturas de un experimento de secuenciación.
- almacenar los resultados estadísticos de un experimento de secuenciación
- almacenar la anotación funcional de los genes de una especie.

9. Los formatos BED, GTF, GFF contienen información ...

- que caracteriza una región de una secuencia de referencia.
- de la longitud de un genoma.
- de la longitud de las secuencias de NGS.
- de los transcritos de un genoma.
- de las variantes de un genoma.

10. ¿Por qué se debe hacer un control de calidad de las secuencias de un experimento de NGS?
- Porque en general la calidad de la secuenciación es mala
  - Para detectar y corregir posibles efectos técnicos o no biológicos en los datos.
  - Para usar el software FastQC.
  - Para usar el software FastQC o alguna de sus alternativas.
  - No siempre es necesario hacer el estudio de la calidad de los datos. Depende de quien haya hecho la secuenciación.
11. Si secuenciamos DNA capturado a partir del *exoma* (región codificante del genoma) ...
- podremos detectar variantes genéticas solamente en dicha región del genoma de los individuos.
  - podremos detectar variantes genéticas en todo el genoma de los individuos.
  - podremos hacer solo análisis de transcriptómica para los individuos del estudio.
  - no podremos hacer ninguna comparativa caso control entre los individuos del estudio.
  - no podremos enviar nuestros datos a ningún repositorio como el SRA, GEO o ArrayExpress.
12. ¿Por que es necesario normalizar los conteos medidos en un experimento de RNA-seq?
- En realidad no siempre es necesario normalizar los datos.
  - Porque la cantidad de secuencias tomadas puede variar entre las diferentes muestras biológicas con lo que la comparación directa de los conteos puede conducirnos a resultados sesgados.
  - Para que los datos sean similares a los tomados con microarrays de DNA.
  - Porque si no el software de análisis estadístico no podrá leerlos.
  - Para que el fichero este mas comprimido y ocupe menos espacio en disco.
13. ¿Por que es necesario normalizar los conteos medidos en un experimento de NGS?
- Porque la longitud de los genes influye en la probabilidad que tienen estos de ser secuenciados, influyendo en la cantidad de conteos encontrados para cada gen.
  - Para obtener así un archivo binario.
  - Para que los genes queden ordenados según su posición en el genoma. Esto es requerido por algunos software para incrementar su velocidad.
  - Para que no haya muchos decimales en las medidas y sean mas fáciles de manejar.
  - Para poder visualizarlos con el software IGV.
14. ¿Es necesario normalizar los datos de secuenciación de DNA obtenidos para un análisis de variantes?
- No. Aunque si que hay descritas metodologías de recalibrado de los datos que hacen que el *variant calling* sea mas acertado.
  - Si siempre.
  - Si siempre. Igual que cuando hacemos un estudio de RNAseq.
  - No. La normalización de los datos se hace solo si se va a aplicar algún test estadístico.
  - Si porque en cualquier experimento al fina siempre se va a hacer algún test estadístico.
15. ¿Que haremos si existen varias versiones de un mismo genoma publicadas y disponibles para la misma especie?
- Hacer un ensamblado con todas ellas y usar este para nuestro experimento.
  - Da igual porque en cada experimento hay que reconstruir o ensamblar el genoma de referencia a partir de nuestros datos experimentales.
  - Seleccionaremos la mas adecuada según criterios biológicos... que sea la mas parecida a la subespecie que estamos estudiando. También incluiremos algunos criterios más técnicos como ver cual de las versiones está mejor anotada o parece tener mejor calidad.
  - Utilizaremos siempre la referencia de ENSEMBL.
  - utilizaremos siempre la referencia del NCBI.

16. 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.
17. 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.
18. 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 se puede comprimir el fichero y no podemos generar el BAM.
19. 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.
20. ¿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 para estudiar otro tipo de variantes como inserciones y deleciones.