Seminario Análisis de secuencias de nuevas tecnologías de secuenciación en paralelo

Trabajo Práctico Final

El presente Trabajo Práctico corresponde a la instancia evaluativa final del Seminario. Para la realización del mismo, se brindarán datos reales y se requerirá la selección de herramientas para su análisis apropiado. A partir de los resultados obtenidos se elaborará un artículo científico (para ello se le brindará un documento modelo) y se realizará la presentación oral de los mismos en el congreso "Bioinformágica" a realizarse el día 8 de junio de 2023.

Título: "Análisis de la expresión génica de levadura (*Saccharomyces cerevisiae*) bajo diferentes condiciones de crecimiento"

Problema biológico:

El objetivo de este trabajo es estudiar el transcriptoma e investigar los cambios en los niveles de expresión génica de *Saccharomyces cerevisiae,* en particular de la cepa CEN.PK 113-7D bajo dos condiciones metabólicas diferentes: metabolismo respiro-fermentativo (*batch*) o totalmente respiratorio (quimiostato). En el primer caso, el crecimiento es un proceso discontinuo en el que la levadura se cultiva en un medio limitado de nutrientes y oxígeno. La levadura utiliza primero el oxígeno presente para producir energía a través de la respiración celular y, una vez agotado el oxígeno, comienza a producir energía a través de la fermentación. En este proceso, la levadura consume azúcares para producir dióxido de carbono y alcohol. Este metabolismo es útil para producir productos de fermentación como cerveza o vino. Por otro lado, el crecimiento de levadura en un metabolismo completamente respiratorio (quimiostato) es un proceso continuo en el que la levadura se cultiva en un medio con una tasa de flujo constante de nutrientes y oxígeno. En este proceso, la levadura utiliza exclusivamente el oxígeno presente para producir energía a través de la respiración celular, lo que resulta en un crecimiento más lento pero una mayor producción de biomasa. Este metabolismo es útil para producir levadura para su uso en la fabricación de pan o para la producción de proteínas recombinantes en la industria farmacéutica.

Los datos de RNA-Seq están disponibles en la base de datos SRA de NCBI, bajo el número de acceso SRS307298. En total hay 12 muestras, dos tratamientos con tres réplicas biológicas cada uno. Los datos son de tipo *paired-end*. Debido a que analizar los archivos originales llevaría mucho tiempo de procesamiento, se le brinda un *set* de datos reducido que incluye sólo las lecturas que alinean contra el cromosoma I de levadura (*S. cerevisiae*).

Completar la siguiente tabla con información (metadatos) de las muestras.

Muestra	**Condición**	**Experimento**	**Réplica**	**SRA Accession**	**Protocolo preparación biblioteca**	***Read length***	**Stranded?**
Quimiostato_1	Control	RNA-seq	1	SRR453569			No
Quimiostato_2	Control	RNA-seq	2	SRR453570			No
Quimiostato_3	Control	RNA-seq	3	SRR453571			No
Batch_1	Tratamiento	RNA-seq	1	SRR453566			No
Batch_2	Tratamiento	RNA-seq	2	SRR453567			No
Batch_3	Tratamiento	RNA-seq	3	SRR453568			No

Los pasos a seguir sugeridos son los siguientes:

1. Control de calidad y estadísticas de las secuencias.
2. Pre-procesamiento de las secuencias.
   - ➔ Las secuencias ya están pre-procesadas. Pero si lo considera necesario puede hacerles un procesamiento extra.
3. Alineamiento de las lecturas contra el genoma de referencia.
   - ➔ Se brinda el genoma de referencia que corresponde solo al cromosoma I de *S. cerevisiae* (ensamblado sacCer2) y el archivo de anotación (gtf) del genoma completo por si lo necesita emplear.
   - ➔ Se recomienda usar alguno de los programas vistos en clase del tipo "*splice-aware*".
   - ➔ Tener en cuenta el tipo de biblioteca de cada muestra.
   - ➔ Reportar los resultados/estadísticas de los alineamientos.
   - ➔ Puede ser necesario aplicar algunas operaciones con samtools a los archivos generados para usarlos en pasos posteriores.
4. Control de calidad de los mapeos.
   - ➔ Utilizar algún *software* específico para RNA-seq QC.

- ➔ Puede utilizar la herramienta bamCoverage de deepTools para generar archivos de cobertura (dada a características de los datos reducidos, aplicar un valor de *bin size* igual a 5).
- ➔ Visualizar los alineamientos y/o archivos de cobertura en un *genome browser*. Por ejemplo, en Jbrowse2 (local) puede crear tracks del tipo "*MultiWiggle track*" para visualizar archivos .bw bajo una misma escala.

5. Recuento del número de lecturas por gen (*reads counting*).

- ➔ Si utilizó STAR en el paso 2 puede incorporar la opción del conteo de reads durante el alineamiento. En caso contrario, puede emplear programas como featureCounts o HTSeq-count.
- ➔ Tener en cuenta formatear de manera apropiada las salidas de este paso (matriz de conteos *-count matrix-*) para usar esta información en el siguiente, según el programa que utilice.

6. Análisis de la expresión diferencial entre condiciones.

- ➔ Emplear los paquetes de R DESeq2 o edgeR.
- ➔ Preparar adecuadamente la matriz de conteos y la de información de las muestras, y setear correctamente el control y el tratamiento.
- ➔ Establecer valores de filtrado para obtener una lista reducida de los genes diferencialmente expresados de manera significativa.
- ➔ Realizar un análisis exploratorio de los resultados mediante diferentes visualizaciones de los datos.

7. Análisis funcionales de un conjunto de genes de interés.

- ➔ A partir del paso previo, definir un *set* de genes que muestren cambios significativos entre las condiciones analizadas y realizar un análisis funcional.
- ➔ Puede utilizar por ejemplo Enrichr específico para levadura o Panther.
- ➔ En este punto puede ser necesario convertir los ID de los genesm si no lo hizo durante el análisis del paso anterior.
- ➔ Se le proporcionó el archivo "Genes_diferencialmente_expresados_signif.tabular" el cual contiene unos 2500 genes de interés obtenidos del análisis con los datos completos. Puede extender el análisis funcional utilizando dicha lista de genes.

Material proporcionado:

- Archivo zip con los datos.
- Modelo para la escritura del artículo científico.
- Rúbrica para realizar la revisión por pares (*peer review)* de un artículo científico.

Entregables:

1. Documento Resumen con los programas y comandos ejecutados en cada una de las actividades realizadas.

2. Todos los datos procesados en archivo zip.

3. Artículo científico con un resumen de los resultados, los principales hallazgos y la metodología aplicada (se le proporciona un modelo para tal fin).
4. Presentación oral en el congreso “Bioinformágica”. Dispone de un tiempo máx. de 15 min + 5 min de espacio para preguntas.
5. Revisión del artículo científico de un compañero/a (se le brinda una rúbrica a completar).

Fechas importantes:

- Lo solicitado en los puntos 1, 2 y 3 se deberá enviar como fecha límite el día **02/06** en el recurso “Tarea” dentro de la pestaña “Trabajo Final” del aula virtual.
- La presentación en el congreso (punto 4) se realizará el día **08/06**.
- La revisión de otro artículo científico (punto 5) deberá completarse como fecha límite el **09/06**.
- En caso de recuperar este TP Final, la fecha de entrega del artículo científico revisado con los cambios sugeridos por los pares (ya sean cuestiones de redacción o del análisis de los datos) es el **16/06** (semana de recuperatorios).