Entregable: E10-E41
Valoración A16-A43

1.- Objetivos

Esta tarea ha consistido en reproducir un estudio *de novo* usando el pipeline OASES-BLAST (Schulz et al., 2012; Altschul et al., 1990) para ensamblar de novo, anotar y realizar el perfil funcional de un transcriptoma no modelo obtenido del SRA archive del NCBI. Para la anotación de transcritos, ontologías génicas y metaboloma se usaron las bases de datos de NR, Gene Ontology, KEGG.

Para más detalle a continuación se presenta una tabla con las pruebas realizadas a la aplicación DeNovoSeq en los dos modos de ejecución disponibles (step-by-step y pipeline) tanto en formato RAP como formato RCP. Este reporte de valoración de la actividad A16-A43 es parte material del entregable E10-E41.

2.- Material y métodos.

Los métodos empleados en esta prueba de concepto son los mismos usados en los artículos de investigación relacionados con este material (Pérez-Sánchez R, et al. 2021).

Para llevar a cabo el análisis *de novo* las muestras se obtuvieron de glándulas salivales de una especie de garrapata (*Ornithodoros erraticus*) a partir del SRA archive del NCBI que se detallan en tabla 1.

Concretamente se uso las siguientes librerías descargadas a partir del Bioproject accesible en esta URL https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA666995

Tabla 1: Muestras de estudio

Nombre de la librería	SRA accessions
RAO-1	SRR12763809
RAO-3	SRR12763807
RAO-5	SRR12763805

3.- Resultados de la valoración y material resultante

Todas las pruebas se realizaron tanto sobre las versiones RAP y RCP de la aplicación DeNovoSeq. Para facilitar de la realización de la prueba de concepto y dado que este material es de naturaleza big data, se ha habilitado un acceso FTP para acceder mediante el siguiente usuario anónimo y password y una carpeta con los entregables de esta prueba de concepto A16-A43 junto con otras asociadas al entregable E10-E41. Para acceder a dicho FTP recomendamos Filezilla que puede descargarse gratuitamente en https://filezilla-project.org. Las credenciales para acceder son concretamente las siguientes:

Servidor FTP: biotechvana.uv.es
Usuario: DIGITAL
Password: DiGi_19_21*

En concreto se debe acceder a la carpeta 03_valoracion_actividad_A16_A43_DeNovoSeq donde se podrá encontrar:

- Carpeta step-by-step.
- Carpeta pipeline mode

Para poder visualizar correctamente los resultados deben de descargarse al escritorio. Nótese que se ha creado una carpeta por modo de ejecución debido a que los resultados obtenidos en ambas versiones de la aplicación (RAP y RCP) son exactamente iguales y evitamos de esta forma la duplicidad de resultados.

Este material se estructura de la siguiente manera:

En la carpeta step-by-step que contiene los resultados de la ejecución del protocolo DeNovo protocols, se pueden encontrar las siguientes subcarpetas:

- **00_raw_data**: carpeta donde se depositan los archivos fastq sin procesar.
- **01_quality_analysis:** carpeta donde se depositan los resultados del análisis de calidad.
- **02_preprocessed_reads**: carpeta donde se depositan los resultados del pre-procesado
- **03_de_novo_assembly:** carpeta donde se depositan los resultados del ensamblaje *de novo.*
- **04_consenso**: carpeta donde se depositan los resultados tras realizar las secuencias consenso. Este es un paso intermedio que se ejecuta por comandos.
- **05_annotation:** carpeta donde se depositan los archivos finales con la anotación correspondiente a través del NCBI-BLAST.
- **06_functional_analysis**: carpeta donde se depositan los resultados tras realizar los análisis funcionales a través de gene ontology.

En la carpeta pipeline mode que contiene los resultados de la ejecución de dicho modo, se pueden encontrar las siguientes subcarpetas:

- **01_FASTQC:** carpeta donde se depositan los resultados del análisis de calidad.
- **02_PRINSEQ**: carpeta donde se depositan los resultados del pre-procesado
- **03_assemblyOASES:** carpeta donde se depositan los resultados del ensamblaje *de novo.*
- **04_BLAST:** carpeta donde se depositan los archivos finales con la anotación correspondiente a través del NCBI-BLAST.

4.- Análisis *de novo* y resultados

Los pasos reproducidos para realizar el análisis *de novo* fueron los siguientes:

- Modo de ejecución: STEP-BY-STEP. DeNovo protocol:

1. Quality analysis: FASTQC (Andrews 2016)
2. Preprocessing: PRINSEQ (Schmieder and Edwards 2011)
3. DeNovo Assembly: Assembly → Show input configuration file
4. DeNovo Assembly: Assembly → Transcriptomes → Oases (Schulz et al., 2012)
5. Annotation: NCBI-BLAST → Format BLAST databases (Altschul et al., 1990)
6. Annotation: NCBI-BLAST → BLAST search with fasta file query (Altschul et al., 1990)
7. Annotation: NCBI-BLAST → Process BLAST output (Altschul et al., 1990)

- Modo de ejecución: PIPELINE MODE:

    - Pipeline preprocessing and assembly:

1. Quality analysis: FASTQC (Andrews 2016)
2. Preprocessing: PRINSEQ (Schmieder and Edwards 2011)
3. DeNovo Assembly: Assembly → Transcriptomes → Oases (Schulz et al., 2012)

    - Pipeline BLAST:

4. Annotation: NCBI-BLAST→ Format BLAST databases (Altschul et al., 1990)
5. Annotation: NCBI-BLAST → BLAST search with fasta file query (Altschul et al., 1990)
6. Annotation: NCBI-BLAST → Process BLAST output (Altschul et al., 1990)

En el pipeline mode, se usan dos pipelines. Uno de calidad + pre-procesado + ensamblaje de novo y otro de Blast dado que son dos pasos independientes en el pipeline mode. Esto es debido a que los análisis *de novo* (los que realiza DeNovoSeq) son radicalmente diferentes de los análisis implementados sobre los que se implementan en RNAseq y VariantSeq dado que estas herramientas manejan datos de re-secuenciación. La dificultad de los pipelines *de novo* es que no se sabe nada sobre los datos con los que se trata porque se secuencian por primera vez. Con esto nuestra opción ha sido solamente montar los pipelines que corresponden a pasos que son automatizables, es decir pipelines de análisis de calidad, pre-procesado y ensamblaje y luego pipeline de Blast.

A continuación, se presenta la tabla 2 y tabla 3 detalladas con las pruebas realizadas a la aplicación de DeNovoSeq en el modo de ejecución step-by-step y el pipeline tanto en versión RAP como versión RCP. Por simplicidad se añade una tabla común a ambas versiones disponibles de la aplicación ya que están compuestos por las mismas herramientas.

Tabla 2. Step-by-step mode

Versión	Modo de ejecución	Herramienta	Descripción	Cumple Requisitos
RCP/RAP	STEP-BY-STEP: DeNovoProtocol	Preprocessing: Quality analysis FASTQC	Se realiza un análisis de calidad de los archivos fastq sin procesar.	Si Como resultado se obtiene un informe en el que se muestran los parámetros analizados en las muestras.
		Preprocessing: Trimming and cleaning PRINSEQ	Filtra, corta o reformatea los archivos sin procesar fastq añadiendo una serie de parámetros basados en el análisis de calidad.	Si Como resultado se obtienen nuevos archivos fastq modificados según los parámetros introducidos.
		DeNovo Assembly: Assembly Transcriptomes Oases	Ensamblador transcriptómico *de novo* diseñado para producir transcritos producidos a partir de tecnologías de secuenciación de lectura corta (Illumina, SOLiD o 454) en ausencia de ensamblaje genómico.	Si Como resultado se obtienen archivos con formato .fa y .txt
		Annotation: NCBI-BLAST Format BLAST databases	Para crear y formatear bases de datos de BLAST, DeNovoSeq implementa una interfaz que emplea el formato BLAST en el paquete de NCBI BLAST para crear una base de datos de referencia	Si Como resultado se obtiene un mensaje de confirmación, y la base de datos RDP, la base de datos nucleótidos y la base de datos proteínas en formato Blast.
		Annotation: NCBI-BLAST BLAST search with fasta file query	La implementación de NCBI BAST en DeNovoSeq permite la comparación entre archivos de secuencia. Si sólo se dispone de un archivo de secuencia, se puede comparar dicho archivo con la base de datos. El formato de este archivo es fasta. La base de datos frente a la cual se quiere hacer la comparación también debe encontrarse en formato fasta. Esto puede realizarse previamente empleando “format BLAST databases”.	Si Como resultado se obtienen una colección de ficheros con extensión .xml que son el output natural de blast (un xml por cada secuencia query)
		Annotation: NCBI-BLAST Process BLAST output	Es un script que procesa los archivos .xml generados a partir de la búsqueda BLAST y crea un archivo de anotación (formato .csv) con todos los parámetros estadísticos y anotaciones proporcionados por la búsqueda BLAST.	Si Como resultado se obtiene un archivo .csv con las anotaciones de las todas secuencias de la especie estudiada usadas como queries en el Blast contra la base de datos mediante NCBI Blast. De las queries analizadas en Blast, se obtiene un porcentaje de éxito de 45,74%.
RAP/RCP	Sistema experto	Recomendaciones y soluciones automatizadas	Según el panel de reportes el sistema experto permite dar una solución a un problema dado, en la forma de recomendación o aplicación directa de la re-ejecución del proceso.	La aplicación acierta en un 50% con la resolución de recomendación a aplicar. La aplicación de las mismas es 100% correcta si bien es un elemento prototípico que necesita ser depurado y sometido a mas entrenamiento. El que se ha aplicado aquí es básico, si bien suficiente para que podamos integrarlo en las aplicaciones de GPRO operativo y funcionando, pero destacando que es una aplicación en fase beta.

Tabla 3. Pipeline mode

Versión	Modo de ejecución		Herramienta	Descripción	Cumple Requisitos
RCP/RAP	PIPELINE	Pipeline Pre-procesado y ensamblaje	Preprocessing: Quality analysis FASTQC	Se realiza un análisis de calidad de los archivos fastq sin procesar.	Si Como resultado se obtiene un informe en el que se muestran los parámetros analizados en las muestras.
			Preprocessing: Trimming and cleaning PRINSEQ	Filtra, corta o reformatea los archivos sin procesar fastq añadiendo una serie de parámetros basados en el análisis de calidad.	Si Como resultado se obtienen nuevos archivos fastq modificados según los parámetros introducidos.
			DeNovo Assembly: Assembly Transcriptomes Oases	Ensamblador transcriptómico *de novo* diseñado para producir transcritos producidos a partir de tecnologías de secuenciación de lectura corta (Illumina, SOLiD o 454) en ausencia de ensamblaje genómico.	Si Como resultado se obtienen archivos con formato .fa y .txt
		Pipeline Blast	Annotation: NCBI-BLAST Format BLAST databases	Para crear y formatear bases de datos de BLAST, DeNovoSeq implementa una interfaz que emplea el formato BLAST en el paquete de NCBI BLAST para crear una base de datos de referencia	Si Como resultado se obtiene un mensaje de confirmación, y la base de datos RDP en formato blast
			Annotation: NCBI-BLAST BLAST search with fasta file query	La implementación de NCBI BAST en DeNovoSeq permite la comparación entre archivos de secuencia. Si sólo se dispone de un archivo de secuencia, se puede comparar dicho archivo con la base de datos. El formato de este archivo es fasta. La base de datos frente a la cual se quiere hacer la comparación también debe encontrarse en formato fasta. Esto puede realizarse previamente empleando “format BLAST databases”.	Si Como resultado se obtienen una colección de ficheros con extensión .xml que son el output natural de blast (un xml por cada secuencia query)
			Annotation: NCBI-BLAST Process BLAST output	Es un script que procesa los archivos .xml generados a partir de la búsqueda BLAST y crea un archivo de anotación (formato .csv) con todos los parámetros estadísticos y anotaciones proporcionados por la búsqueda BLAST.	Si Como resultado se obtiene un archivo .csv con las anotaciones de las todas secuencias de la especie estudiada usadas como queries en el Blast contra la base de datos mediante NCBI Blast. De las queries analizadas en Blast, se obtiene un porcentaje de éxito de 45,74%.
RAP/RCP	Sistema experto		Recomendaciones y soluciones automatizadas	Según el panel de reportes el sistema experto permite dar una solución a un problema dado, en la forma de recomendación o aplicación directa de la re-ejecución del proceso.	La aplicación acierta en un 50% con la resolución de recomendación a aplicar. La aplicación de las mismas es 100% correcta si bien es un elemento prototípico que necesita ser depurado y sometido a mas entrenamiento. El que se ha aplicado aquí es básico, si bien suficiente para que podamos integrarlo en las aplicaciones de GPRO operativo y funcionando, pero destacando que es una aplicación en fase beta.

5.- Conclusiones

Todos los análisis de reprodujeron con éxito usando el modo step-by-step y pipeline tanto en la versión RCP como la versión RAP de la aplicación DeNovoSeq. Se verifica que todas las herramientas comprobadas funcionan correctamente y la aplicación está operativa y correctamente funcionando para su uso.

6.- Bibliografía

- Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. Basic local alignment search tool. *JMolBiol* 1990;215(3): 403-410.
- Andrews, S. 2016. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc
- Pérez-Sánchez R, Carnero-Morán Á, Soriano B, Llorens C, Oleaga A. RNA-seq analysis and gene expression dynamics in the salivary glands of the argasid tick Ornithodoros erraticus along the trophogonic cycle. *Parasites Vectors* 2021; 14(170).
- Schmieder, R. and Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011;27(6):863-864.
- Schulz MH, Zerbino DR, Vingron M and Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 2012;28(8): 1086 -1092.