Entregable: E10-E41
Valoración A17-A44

1.- Objetivos

Esta tarea ha consistido en reproducir los pasos de análisis de metagenómica 16S usando el pipeline NCBI-BLAST (Altschul et al., 1990) de la herramienta DeNovoSeq en comparación con los resultados que daría el pipeline infernal-mothur aligner de la RDP Ribosome Database Project. Esto usando como caso de estudio una colección de muestras sobre una patología de base microbiana cuyas muestras fueron obtenidas a partir del SRA archive del NCBI. Para realizar la anotación de la taxonomía a los distintos taxa bacterianos identificados se usaron las bases de datos de la *RDP Project,* es decir la prueba de concepto. Se aprovechó también para testar las distintas funcionalidades estadísticas de otra de las aplicaciones de GPRO STATools creando distintas representaciones gráficas para verificar la funcionalidad de la aplicación. STATools implementa librerías de R, Python y distintas funciones de paquetes de metagenómica como Vegan y Mothur.

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. Y también las distintas pruebas realizadas con STATools también bajo los formatos RCP como RAP. Este reporte de valoración para la actividad A17-A44 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 (Herreros-Pomares A, et al. 2021) con la excepción de los ejecutados con STATools que se realizan en esta prueba de concepto para verificación de las aplicaciones.

Para llevar a cabo el análisis metagenómico 16S, las muestras se usaron las siguientes librerías descargadas a partir del Bioproject accesible en la siguiente URL https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA663424

Tabla 1: Muestras de estudio

Nombre de la librería	SRA accessions
JVB-Methyl-10	SRR12640646
JVB-Methyl-11	SRR12640658
JVB-Methyl-12	SRR12640657
JVB-Methyl-13	SRR12640656

Dichas librerías corresponden a biopsias orales de pacientes con leucoplasia verrucosa proliferativa (LVP).

3.- Resultados de la valoración y material resultante

Todas las pruebas se realizaron tanto sobre las versiones RAP y RCP de la aplicación DeNovoSeq e igualmente respecto a análisis estadístico, usando también las versiones RCP y RAP de STATools. 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 A17-A44 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
04_valoracion_actividad_A17_A44_DeNovoSeq_STATools donde se podrá encontrar:

- Carpeta 01_step-by-step
- Carpeta 02_pipeline mode
- Carpeta 03_STATools

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 01_step-by-step que contiene los resultados de la ejecución del protocolo DeNovo protocols, se pueden encontrar las siguientes subcarpetas:

Análisis realizados con DeNovoSeq:

- **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_DBs**: carpeta donde se depositan los resultados de la base de datos compilada de BLAST.
- **04_cdhit**: carpeta donde se deposita un análisis de redundancia. Este es un paso intermedio que se ejecuta por comandos.
- **05_RDP_vs_BLAST**: carpeta donde se depositan los resultados de los archivos anotados (BLAST) y clasificados (RDP project), el rendimiento de anotación de OTUs fue el parámetro a comparar.

En la carpeta 02_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_COMPARATIVE_DATA_BASES**: carpeta donde se depositan los resultados de los archivos anotados (BLAST) y clasificados (RDP project).

En la carpeta 03_STATools que contiene los resultados del análisis estadístico con el programa STATools, se pueden encontrar las siguientes subcarpetas:

Análisis realizados con STATools:

- **01_krona**: carpeta donde se depositan los resultados del análisis de taxonomía
- **02_indices_de_diversidad**: carpeta donde se depositan los resultados tras el análisis de los índices de diversidad.
- **03_heatmap**: carpeta donde se depositan los resultados de la representación gráfica del heatmap.

En el pipeline mode, se usan dos pipelines una de calidad + pre-procesado 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.

En la carpeta de pipeline mode no se reproducen los análisis estadísticos con STATools porque en este caso son lo mismo. STATools no es una aplicación de ejecución de pipelines sino de scripts estadísticos.

A continuación, se presenta la tabla 2 y la tabla 3 detalladas con las pruebas realizadas a la aplicación de DeNovoSeq en el modo de ejecución step-by-step y el pipeline mode 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.

También se sumarizan en tabla 4 los resultados de la valoración de los análisis estadísticos efectuados usando la aplicación STATools. En concreto, se han realizado un análisis KRONA, índices de diversidad y heatmaps. Los resultados de estos análisis estadísticos pueden encontrarlos en la carpeta 03_STATools con los siguientes nombres: 01_krona, 02_indices_diversidad y 03_heatmap.

Tabla 2. Step-by-step mode DeNovoSeq

Versión	Modo	Herramienta	Descripción	Cumple Requisitos
RCP /RAP	STEP-BY-STEP: DeNovoSeq	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.
		Preprocessing: Trimming and cleaning Fastx Toolkit	Convierte ficheros fastq a fasta	Si Como resultado se obtienen nuevos archivos fasta creados a partir de los fastq
		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 por cada muestra con las anotaciones de las todas secuencias de cada amplicon usadas como queries en el Blast contra la RDP mediante NCBI Blast. En comparación con el pipeline infernal aligner de la RDP el porcentaje de secuencias clasificadas por taxonomía es similar. De las queries analizadas en Blast, se obtiene un porcentaje de éxito de 40,89%. Respecto a los resultados obtenidos en la RDP se obtiene un porcentaje de éxito de 77,32%.
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 DeNovoSeq

Versión	Modo		Herramienta	Descripción	Cumple Requisitos
RCP /RAP	Pipeline Pre-procesado	Pipeline mode: DeNovoSeq	Preprocessing: Quality analysis FASTQC	Se realiza un análisis de calidad de los archivos fastq sin procesar.	Si Replica los mismos resultados que el step-by-step mode
			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.	
	Pipeline Blast	STEP-BY-STEP: DeNovoSeq	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 Replica los mismos resultados que el step-by-step mode
			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".	
			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.	
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 4. STATools pruebas realizadas sobre los datos de diversidad

Versión	Herramienta	Análisis	Descripción	Cumple Requisitos
RCP /RAP	STATools	Análisis krona	Crea una representación gráfica dinámica de la diversidad de una muestra usando el fichero de binning (el obtenido vía blast) como referencia	Si. Crea las representaciones gráficas o tabuladas de los análisis
		Análisis de diversidad	Infiere distintos índices de diversidad	
		Análisis heatmap	Crea un heatmap a partir de los datos de las muestras.	
STATools no implementa sistema experto, porque no es una herramienta gestora de flujos de trabajo				

5.- Conclusiones

Todos los análisis de reprodujeron con éxito usando el modo step-by-step y el pipeline tanto en la versión RCP como la versión RAP de ambas aplicaciones DeNovoSeq y STATools. Se verifica que todas las herramientas comprobadas funcionan correctamente y las dos aplicaciones están operativas 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
- Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res* 2005; 33:D294-D296.
- Futami R, Muñoz-Pomer A, Viu JM, Domínguez-Escribà L, Covelli L, Bernet GP, Sempere JM, Moya A, Llorens C. GPRO: the professional tool for management, functional analysis and annotation of omic sequences and databases. *Biotechvana Bioinformatics*; 2011-SOFT3.
- Hannon Lab. "FASTX Toolkit." 2016. http://hannonlab.cshl.edu/fastx_toolkit/.
- Herreros-Pomares A, Llorens C, Soriano B, Bagan L, Moreno A, Calabuig-Fariñas S, Jantus-Lewintre E, Bagan J. Differentially methylated genes in proliferative verrucous leukoplakia reveal potential malignant biomarkers for oral squamous cell carcinoma. *Oral Oncology* 2021; 116, 105191.
- Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu and Weizhong Li, CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics* 2012; 28 (23): 3150-3152.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, et al. 2009. Introducing mothur: open-source,
- Schmieder, R. and Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011; 27(6):863-864.