Entregable: E10-E41
Valoración A15-A42

1.- Objetivos

Esta tarea ha consistido en reproducir un análisis RNAseq de expresión diferencial de 9 muestras de transcriptoma humano obtenidas a partir del SRA archive del NCBI. Para la integración y anotación de las muestras se usó la aplicación Worksheet.

Este reporte de valoración de la actividad A15-A42 es parte material del entregable E10-E41.

2.- Material y métodos.

Los métodos empleados en esta prueba de concepto son los mismos usados en el artículo de investigación relacionado con este material (Pérez-Sánchez et al., 2019).

Para llevar a cabo el análisis, las muestras de RNA se obtuvieron 9 muestras de dorada (*Sparus aurata*) a partir del SRA archive del NCBI que se detallan en tabla 1.

Concretamente se usaron las siguientes librerías descargadas a partir del Bioproject accesible en esta URL https://www.ncbi.nlm.nih.gov/bioproject/PRJNA507368.

Tabla 1: Muestras y casos de estudio

SRA accession	Nombre de la librería	Tags
SRR8255970	ZFG-17-12_03_26333_S7_R1_001.fastq	BC1
SRR8255963	ZFG-17-12_06_26336_S10_R1_001.fastq	BC2
SRR8255962	ZFG-17-12_09_26339_S13_R1_001.fastq	BC3
SRR8255949	ZFG-17-12_12_26342_S16_R1_001.fastq	BC4
SRR8255945	ZFG-17-12_16_26346_S2_R1_001.fastq	BI1
SRR8255941	ZFG-17-12_20_26350_S6_R1_001.fastq	BI2
SRR8255956	ZFG-17-12_24_26354_S10_R1_001.fastq	BI3
SRR8255952	ZFG-17-12_28_26358_S14_R1_001.fastq	BI4
SRR8255939	ZFG-17-12_32_26362_S18_R1_001.fastq	BI5

*__BC__ = Control; __BI__ = Pez infectado.

También se requirió el uso del genoma y del fichero de anotaciones de *S. aurata*. Dicho genoma se puede solicitar en la siguiente URL: https://nutrigroup-iats.org/welcome/request_file. Alternativamente también se puede usar el genoma y el fichero de anotaciones de *S.aurata* disponible en el NCBI https://www.ncbi.nlm.nih.gov/assembly/GCF_900880675.1. Se ha de tener en cuenta que los resultados probablemente presenten variaciones respecto a los que facilitamos en el FTP.

Adicionalmente se usaron 4 archivos para el análisis de enriquecimiento de GOs que se pueden descargar desde los siguientes links:

1. Assayed genes:

https://ecampus.biotechvana.com/pluginfile.php/907/mod_folder/content/0/assayed_genes.txt?forcedownload=1

2. Differentially expressed genes: https://ecampus.biotechvana.com/pluginfile.php/907/mod_folder/content/0/diff_genes.csv?forcedownload=1

3. Gene size: https://ecampus.biotechvana.com/pluginfile.php/907/mod_folder/content/0/length_genes.txt?forcedownload=1

4. Go terms: https://ecampus.biotechvana.com/pluginfile.php/907/mod_folder/content/0/GO_final_saurata.txt?forcedownload=1

3.- Resultados de la valoración y material resultante

Todas las pruebas se realizaron tanto sobre las versiones RAP y RCP de la aplicación RNASeq. 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 A15-A42 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 **02_valoracion_actividad_A15_A42_RNASeq** donde se podrá encontrar:

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

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 Tophat/Hisat2 & Cufflinks, 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_refseq:** carpeta donde se depositan los siguientes ficheros correspondientes al genoma y al fichero de anotaciones de *S. aurata.*
- **04_mapping:** carpeta donde se depositan los resultados del mapeo.

- **05_transcriptome_assembly:** carpeta donde se depositan los resultados de la cuantificación y ensamblaje del transcriptoma.
- **06_differential_expression**: carpeta donde se depositan los resultados de análisis de expresión diferencial.
- **07_ go_enrichment_analysis**: carpeta donde se depositan los resultados del análisis de enriquecimiento de GOs.

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

- **01_FASTCQ:** carpeta donde se depositan los resultados del análisis de calidad.
- **02_CUTADAPT**: carpeta donde se depositan los resultados del pre-procesado procedentes de CUTADAPT
- **03_PRINSEQ**: carpeta donde se depositan los resultados del pre-procesado procedentes de PRINSEQ.
- **04_ FASTQC**: carpeta donde se depositan los resultados del segundo análisis de calidad tras el pre-procesado.
- **05_Tophat:** carpeta donde se depositan los resultados del mapeo.
- **06_Cufflinks:** carpeta donde se depositan los resultados de la cuantificación y ensamblaje del transcriptoma.
- **07_Cuffdiff:** carpeta donde se depositan los resultados de análisis de expresión diferencial.

4.- Testado de las funciones de RNASeq

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

- Modo de ejecución: STEP-BY-STEP. Tophat/Hisat2 & Cufflinks protocol:

1. Quality analysis: FASTQC (Andrews 2016)
2. Preprocessing: PRINSEQ (Schmieder and Edwards 2011)
3. Mapping: Tophat (Kim et al., 2013; Trapnell et al., 2012)
4. Transcriptome Assembly: Cufflinks (Trapnell et al., 2012)
5. Differential Expression Test: Cuffdiff (Trapnell et al., 2012; Goff et al., 2019)
6. GOseq: GOseq (Young et al., 2010)

- Modo de ejecución: PIPELINE

Muestras: Single-End

1. Quality analysis: FASTQC (Andrews 2016)

2. Preprocessing: PRINSEQ (Schmieder and Edwards 2011)

3. Mapping: Tophat (Kim et al., 2013; Trapnell et al., 2012)

4. Differential Expression: Quantification: Cuffdiff (Trapnell et al., 2012; Goff et al., 2019)

De forma adicional y aunque no forma parte de este entregable, hemos aprovechado esta prueba de concepto para crear un tutorial de uso en el análisis de expresión de genes con RNASeq. Pueden acceder al tutorial de RNASeq en el siguiente enlace https://ecampus.biotechvana.com/course/view.php?id=17

A continuación, se presentan tres tablas detalladas con las pruebas realizadas a la aplicación RNASeq en los dos modos de ejecución (step-by-step y 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: Tophat/Hisat2 & Cufflinks Protocol

Versión	Modo de ejecución	Herramienta	Descripción	Cumple Requisitos
RAP y RCP	STEP-BY-STEP: Tophat/Hisat & Cufflinks Protocol	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: Demultiplex FastqMidCleaner	Clasifica y divide las lecturas de secuenciación de los archivos fastq en archivos separados de acuerdo con identificadores moleculares predefinidos (MID).	Si Como resultado se obtienen nuevos archivos fastq.
		Preprocessing: Trimming and cleaning CUTADAPT	Encuentra y elimina secuencias de adaptadores, primers, colas poli-A y otros tipos de artefactos de secuenciación presentes en los archivos sin procesar fastq.	Si Como resultado se obtienen nuevos archivos fastq de los cuales se han eliminado las secuencias de adaptadores u otros artefactos de secuenciación.
		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 Trimommatic	Es una herramienta de recorte específica para muestras tanto pair-end como single-end obtenidas a través de secuenciación NGS de Illumina.	Si Como resultado se obtienen nuevos archivos fastq modificados según los parámetros introducidos
		Preprocessing: Trimming and cleaning Fastx Toolkit	Esta función acoge a un conjunto de herramientas destinadas al pre-procesado de las muestras fastq.	Si Según la herramienta usada se pueden obtener nuevos archivos fastq, resultados estadísticos, gráficos…
		Preprocessing: Prepseq FastqCollapser	Elimina las lecturas duplicadas en las muestras fastq. Esta herramienta se basa en el análisis de calidad según el contenido de secuencia.	Si Como resultados se obtienen nuevos archivos fastq.
		Preprocessing: Prepseq FastqIntersect	Compara la información de dos archivos pair-end que han sido pre-procesados de forma independiente y la información de ambos archivos para editarlos manteniendo solo aquellas lecturas, y en el mismo orden, que están presentes en ambos archivos	Si Como resultados se obtienen nuevos archivos fastq.
		Preprocessing: Bed to junctions	Convierte los outputs generados por Tophat: junctions.bed	Si Como resultados se obtiene un archivo llamado: junctions list
		Mapping: Tophat	Tophat alinea lecturas de RNA-seq con un genoma de referencia que identifica las uniones de empalme exón-exón.	Si Como resultado se obtienen archivos de mapeo con extensión .bam.
		Mapping: STAR	STAR es un alineador universal para mapear lecturas y transcripciones empalmadas de RNAseq	Si Como resultado se obtienen archivos de mapeo con extensión .bam.
		Mapping: Hisat2	Hisat2 es un programa altamente eficiente para alinear lecturas de experimentos de secuenciación de RNA.	Si

RAP y RCP				Como resultado se obtienen archivos de mapeo con extensión .bam.
		Transcriptome Assembly: Cufflinks	Ensambla el transcriptoma de los datos de RNAseq y cuantifica su expresión	Si Como resultado se obtiene una carpeta por muestra analizada. Dicha carpeta contiene archivos que proporcionan información del FPKM y los transcritos creados.
		Differential Expression Test: Cuffdiff	Para la expresión diferencial se usa el conjunto de herramientas del paquete de Cufflinks. Esta herramienta se encarga de buscar cambios significativos en la expresión génica, splicing y promotores.	Si Se obtiene una carpeta en la que se muestran los resultados generados entre los grupos comparados para el estudio de la expresión diferencial. Estos archivos indican si los genes están infra o sobreexpresados
		GOseq: GOseq	GOseq es una herramienta para la detección de ontologías génicas (GOs) y/o otras categorías definidas por el usuario (p.ej: mapas metabólicos) los cuales se pueden encontrar sobre/infra representadas en los datos de RNAseq analizados.	Si Como resultado se obtienen archivos en los cuales se muestran los GOs enriquecidos. En caso de haber decidido estudiar una categoría adicional, como mapas metabólicos dichos resultados se mostrarán en una carpeta diferente. Por tanto se tendrán tantas carpetas como categorías se analicen.
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. Step-by-step mode: Mapping & Counting Protocol

Versión	Modo de ejecución	Herramienta	Descripción	Cumple Requisitos
		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: Demultiplex FastqMidCleaner	Clasifica y divide las lecturas de secuenciación de los archivos fastq en archivos separados de acuerdo con identificadores moleculares predefinidos (MID).	Si Como resultado se obtienen nuevos archivos fastq.

RAP y RCP	STEP-BY-STEP: Mapping & Counting Protocol	Preprocessing: Trimming and cleaning CUTADAPT	Encuentra y elimina secuencias de adaptadores, primers, colas poli-A y otros tipos de artefactos de secuenciación presentes en los archivos sin procesar fastq.	Si Como resultado se obtienen nuevos archivos fastq de los cuales se han eliminado las secuencias de adaptadores u otros artefactos de secuenciación.
		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 Trimommatic	Es una herramienta de recorte específica para muestras tanto pair-end como single-end obtenidas a través de secuenciación NGS de Illumina.	Si Como resultado se obtienen nuevos archivos fastq modificados según los parámetros introducidos
		Preprocessing: Trimming and cleaning Fastx Toolkit	Esta función acoge a un conjunto de herramientas destinadas al pre-procesado de las muestras fastq.	Si Según la herramienta usada se pueden obtener nuevos archivos fastq, resultados estadísticos, gráficos…
		Preprocessing: Prepseq FastqCollapser	Elimina las lecturas duplicadas en las muestras fastq. Esta herramienta se basa en el análisis de calidad según el contenido de secuencia.	Si Como resultados se obtienen nuevos archivos fastq.
		Preprocessing: Prepseq FastqIntersect	Compara la información de dos archivos pair-end que han sido pre-procesados de forma independiente y la información de ambos archivos para editarlos manteniendo solo aquellas lecturas, y en el mismo orden, que están presentes en ambos archivos	Si Como resultados se obtienen nuevos archivos fastq.
	STEP-BY-STEP: Mapping & Counting Protocol	Mapping: Bowtie2	Bowtie2 es una herramienta de alineamiento de secuencias de largo tamaño (entre 50-100pb).	Si Como resultado se obtienen archivos de mapeo con extensión .bam.
		Mapping: Bwa	BWA es un paquete de software para mapear secuencias de baja divergencia contra grandes genomas de referencia	Si Como resultado se obtienen archivos de mapeo con extensión .bam.
		Mapping: Hisat2	Hisat2 es un programa altamente eficiente para alinear lecturas de experimentos de secuenciación de RNA.	Si Como resultado se obtienen archivos de mapeo con extensión .bam.
		Postprocessing: Corset	Corset usa las lecturas que han sido mapeadas en el transcriptoma para agruparlas jerárquicamente de acuerdo con la proporción de lecturas compartidas y sus patrones de expresión	Si Como resultado se obtienen dos archivos: counts.txt y clusters.txt
		Postprocessing: HTSeq	Esta herramienta cuenta aquellas lecturas que se asignan a características genómicas (exones, genes, etc)	Si Como resultado se obtiene un archivo que contiene una tabla formada por recuentos para cada característica
		Diff Expression Analysis: EdgeR	Realiza análisis diferenciales de expresión génica utilizando una serie de métodos estadísticos que se basan en distribuciones binomiales negativas, incluida la estimación empírica de Bayes, pruebas exactas, modelos lineales generalizados y pruebas de cuasi probabilidad.	Si Se obtiene una carpeta que contiene los resultados tras realizar la comparación de los grupos estudiados. Estos archivos indican si los genes están infra o sobreexpresados
		Diff Expression Analysis: DESeq	Estima las dependencias medias de varianza en la secuenciación de los datos de recuento de lectura para luego probar la expresión génica	Si

			diferencial utilizando un modelo que se basa en la distribución binomial negativa	Se obtiene una carpeta que contiene los archivos tras realizar la comparación de los grupos estudiados. Estos archivos indican si los genes están infra o sobreexpresados
		GOseq: GOseq	GOseq es una herramienta para la detección de ontologías génicas (GOs) y/o otras categorías definidas por el usuario (p.ej: mapas metabólicos) los cuales se pueden encontrar sobre/infra representados en los datos de RNAseq analizados.	Si Como resultado se obtienen archivos en los cuales se muestran los GOs enriquecidos. En caso de haber decidido estudiar una categoría adicional, como mapas metabólicos dichos resultados se mostrarán en una carpeta diferente. Por tanto se tendrán tantas carpetas como categorías se analicen.
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. Pipeline mode

Versión	Modo de ejecución	Herramienta	Descripción	Cumple Requisitos
		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 CUTADAPT	Encuentra y elimina secuencias de adaptadores, primers, colas poli-A y otros tipos de artefactos de secuenciación presentes en los archivos sin procesar fastq.	Si Como resultado se obtienen nuevos archivos fastq de los cuales se han eliminado las secuencias de adaptadores u otros artefactos de secuenciación.
		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 Trimommatic	Es una herramienta de recorte específica para muestras tanto pair-end como single-end obtenidas a través de secuenciación NGS de Illumina.	Si Como resultado se obtienen nuevos archivos fastq modificados según los parámetros introducidos
		Mapping: Tophat	Tophat alinea lecturas de RNA-seq con un genoma de referencia que identifica las uniones de empalme exón-exón.	Si

RAP y RCP	PIPELINE MODE			Como resultado se obtienen archivos de mapeo con extensión .bam.
		Mapping: STAR	STAR es un alineador universal para mapear lecturas y transcripciones empalmadas de RNAseq	Si Como resultado se obtienen archivos de mapeo con extensión .bam.
		Mapping: Bowtie2	Bowtie2 es una herramienta de alineamiento de secuencias de largo tamaño (entre 50-100pb).	Si Como resultado se obtienen archivos de mapeo con extensión .bam.
		Mapping: Bwa	BWA es un paquete de software para mapear secuencias de baja divergencia contra grandes genomas de referencia	Si Como resultado se obtienen archivos de mapeo con extensión .bam.
		Mapping: Hisat2	Hisat2 es un programa altamente eficiente para alinear lecturas de experimentos de secuenciación de RNA.	Si Como resultado se obtienen archivos de mapeo con extensión .bam.
		Postprocessing: Corset	Corset usa las lecturas que han sido mapeadas en el transcriptoma para agruparlas jerárquicamente de acuerdo con la proporción de lecturas compartidas y sus patrones de expresión	Si Como resultado se obtienen dos archivos: counts.txt y clusters.txt
		Postprocessing: HTSeq	Esta herramienta cuenta aquellas lecturas que se asignan a características genómicas (exones, genes, etc)	Si Como resultado se obtiene un archivo que contiene una tabla formada por recuentos para cada característica
		Diff Expression Analysis: EdgeR	Realiza análisis diferenciales de expresión génica utilizando una serie de métodos estadísticos que se basan en distribuciones binomiales negativas, incluida la estimación empírica de Bayes, pruebas exactas, modelos lineales generalizados y pruebas de cuasi probabilidad.	Si Se obtiene una carpeta que contiene los resultados tras realizar la comparación de los grupos estudiados. Estos archivos indican si los genes están infra o sobreexpresados
		Diff Expression Analysis: DESeq	Estima las dependencias medias de varianza en la secuenciación de los datos de recuento de lectura para luego probar la expresión génica diferencial utilizando un modelo que se basa en la distribución binomial negativa	Si Se obtiene una carpeta que contiene los archivos tras realizar la comparación de los grupos estudiados. Estos archivos indican si los genes están infra o sobreexpresados
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 tanto usando el modo step-by-step como el modo pipeline tanto en la versión RCP como la versión RAP de la aplicación RNASeq. Se verifica que todas las herramientas comprobadas funcionan correctamente y la aplicación está operativa y correctamente funcionando para su uso.

6.- Bibliografía

- Andrews, S. 2016. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc
- Goff, L., Trapnell, C. and Kelley, D. 2019. CummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. https://doi.org/doi:10.18129/B9.bioc.cummeRbund
- Kim, D*., et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 2013;14(4):R36.
- Schmieder, R. and Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011;27(6):863-864.
- Trapnell, C*., et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat.Protoc.* 2012;7(3):562-578.
- Young, M.D*., et al.* Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 2010;11(2):R14.