Entregable: E10-E41
Descripción: Pruebas de concepto y valoración de herramientas y entregables
Proyecto: TSI-100903-2019-11

Titulo proyecto: Plataforma bioinformática GPRO de biocomputación en la nube para el procesamiento masivo, integración y análisis funcional de datos ómicos.

Fecha: Valencia, 31 de diciembre de 2021
Documento presentado en sede electrónica a la apertura de plazo de justificación en marzo 2022.

1.- Objetivos y actividades asociadas.

Los objetivos del entregable E10-41 son los asociados a los objetivos de las actividades A14-A41, A15-A4, A16-A43, A17-A44, A18-A45, A19-A46. Estos fueron implementar distintos experimentos ómicos a modo de prueba de concepto en pro de testar el comportamiento de GPRO. Evaluar también la capacidad predictiva del sistema experto/asistente virtual en la toma de decisiones y entrenarlo para ajustar sus recomendaciones.

Con acuerdo a memoria, esto es dentro del paquete de trabajo PT4: "Acciones de prueba de concepto y valorización".

2.-Entregables

Este entregable incluye distintos reportes sobre los resultados de grado éxito en todas las pruebas realizadas en este paquete de trabajo y una valoración sobre la utilidad de todos los componentes en el proyecto y como cada uno de ellos impacta en el escalado de la herramienta. Más concretamente, en las tareas relativas a las actividades A14-A41, A15-A42, A16-A43 y A17-A44 lo que se ha hecho es usar material ómico de resultados conocidos para testar todas y cada una de las 6 aplicaciones de GPRO (Futami et al 2011) para que tras la re-implementación RCP/ RAP todas las funciones de las aplicaciones funcionaban correctamente y que los pipelines del server-side funcionaban correctamente tras la implementación del Docker. Hemos igualmente aprovechado las citadas actividades para testar en el contexto de las actividades A18-A45 y A19-A46 que las nuevas bases de datos de conocimiento implementadas en este proyecto funcionan correctamente y que tanto el pipeline designer, como el módulo de inteligencia artificial definido por el sistema experto y asistente virtual funcionan y son operativos.

A continuación, se desglosa una breve introducción sobre los objetivos de cada actividad asociada a este entregable y se refiere en cada sección a los documentos adjuntos justificativos de las pruebas de concepto realizadas en cada actividad.

> **2.1.- Valoración actividad A14-A41.** En esta actividad se reprodujo cuatro análisis de exoma de cáncer humano con llamada de variantes de tipo SNP/Indel usando el pipeline BWA-GATK-MUTECT y una colección de cinco muestras de cáncer obtenidas a partir del SRA archive del NCBI. Finalmente se usó la herramienta VEP de Ensembl para anotar efectos a las variantes. Todo ello usando los dos modos de ejecución de la aplicación: pipeline mode y step-by-step mode de la aplicación VariantSeq + Server-side (Hafez, et al. 2022) así como también la aplicación Worksheet para la integración de datos. Para

más detalle véase el documento denominado "01_valoracion_actividad_A14-A41" que se presenta acompañando este reporte y es parte material del entregable E10-E41.

2.2.- Valoración actividad A15-A42.- En esta actividad se reprodujo un análisis RNAseq para testar expresión diferencial usando tanto pipeline Tophat/Hisat2&Cufflinks y 9 muestras de transcriptoma de *Sparus aurata* (dorada) a obtenidas a partir del SRA archive del NCBI. Se usaron en conocimiento de distintas bases de datos Ensembl (Cunningham, et al. 2019), Gene Ontology (Huntley, et al. 2015), KEGG (Kotera, et al. 2012), Uniprot (The UniProt 2017) y otras integradas para anotar nombres y códigos de gen, ontologías de genes y rutas metabólicas. Todo ello usando el pipeline mode creado en este proyecto y el step-by-step mode de la aplicación RNAseq + Server-side (Hafez et al 2022) así como también la aplicación Worksheet (app que se introducirá en un próximo artículo en preparación) para la integración de datos. Para más detalle véase el documento denominado "02_valoracion_actividad_A15-A42" que se presenta acompañando este reporte y es parte material del entregable E10-E41.

2.3.- Valoración actividad A16-A43.- En esta actividad se reprodujo un análisis *de novo* usando el pipeline de GPRO basados en OASES y NCBI BLAST (Altschul, et al. 1997; Schulz, et al. 2012) para ensamblar *de novo*, anotar y realizar el perfil funcional de un transcriptoma no modelo obtenido SRA archive del NCBI. Se usaron las bases de datos de NR del NCBI, Gene Ontology, KEGG para anotar transcritos, ontologías y rutas metabólicas. Todo ello usando los dos modos de ejecución de la aplicación: pipeline mode y step-by-step mode de la aplicación DeNovoSeq (app que se introducirá en un próximo articulo en preparación) así como también la aplicación Worksheet para la integración de datos y la aplicación SeqEditor (Hafez, et al. 2020) para el curado de secuencias. Para más detalle véase el documento denominado "03_valoracion_actividad_A16-A43" que se presenta acompañando este reporte y es parte material del entregable E10-E41.

2.4.- Valoración actividad A17-A44.- En esta actividad se reprodujeron 4 análisis de 8 muestras pair-end (son dos fastq por muestra) de metagenómica 16S a partir de una colección de 4 muestras de microbioma bucal asociado a patología obtenidas a partir del SRA archive del NCBI. Para la prueba de concepto se usó las bases de datos de la RDP (Ribosomal Database Project) (Cole, et al. 2005) y NCBI para anotar taxonomía a los distintos taxa bacterianos identificados y el pipeline Blast implementado en la aplicación DeNovoSeq, así como también Worksheet para la integración de datos y distintas herramientas metagenómicas implementadas en la aplicación STATools (que se introducirá en un próximo articulo en preparación) a partir de paquetes de R y el paquete MOTHUR (Schloss, et al. 2009). Para más detalle véase el documento denominado "04_valoracion_actividad_A17-A44" que se presenta acompañando este reporte y es parte material del entregable E10-E41.

2.5.- Valoración actividad A18-A45.- Aquí se sometió al módulo de inteligencia artificial consistente en el sistema experto y el asistente virtual (Genie) a distintas pruebas de verificación. Todo ello para valorar la capacidad de respuesta del asistente y del sistema experto evaluándose por porcentaje de acierto y someter a ambos módulos a un proceso de entrenamiento testado y entrenamiento. Este módulo sistema de inteligencia artificial constituye un capitulo de la tesis doctoral de uno de los trabajadores de la empresa (Aya Allah Ali y su semántica comparativa será introducida

en un próximo artículo (actualmente en preparación). Este sistema de inteligencia artificial será también registrado intelectualmente en los próximos meses. Para más detalle véase el documento denominado "05_valoracion_actividad_A18-A45" que se presenta acompañando este reporte y es parte material del entregable E10-E41.

2.6.- Valoración actividad A19-A46.- Esta tarea ha consistido en valorar todos los entregables obtenidos en todos los paquetes de trabajo de este proyecto determinándose que estos han alcanzado la validación suficiente para ser integrados dentro del proyecto GPRO. Para más detalle véase el documento denominado "06_valoracion_actividad_A19-A46" que se presenta acompañando este reporte y es parte material del entregable E10-E41.

3.- Conclusiones

Todas las pruebas de concepto propuestas en el paquete de trabajo 4 han sido cumplimentadas con éxito como así se demuestra en los distintos informes de resultados presentados donde queda evidenciado que los distintos elementos desarrollados en este proyecto son estables y cumplen todos los requisitos para ser integrados en GPRO.

4.- Bibliografía

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25:3389-3402.
- Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM. 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res*. 33:D294-D296.
- Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, et al. 2019. Ensembl 2019. *Nucleic Acids Res* 47:D745-D751.
- Futami R, Muñoz-Pomer A, Viu JM, Dominguez-Escribá L, Covelli L, Bernet GP, Sempere JM, Moya A, Llorens C. 2011. GPRO: the professional tool for management, functional analysis and annotation of omic sequences and databases. *Biotechvana Bioinformatics*: 2011-SOFT3, http://biotechvana.uv.es/bioinformatics/index.php/article/35
- Hafez A, Futami R, Arastehfar A, Daneshnia F, Miguel A, Roig FJ, Soriano B, Perez-Sánchez J, Boekhout T, Gabaldón T, Llorens C. 2020. SeqEditor: an application for primer design and sequence analysis with or without GTF/GFF files. *Bioinformatics*. 37 (11):1610-1612.
- Hafez A, Futami R, Soriano B, Ceprian R, Elsayed AA, Ramos-Ruiz R, MArtinez G, Roig FJ, Torres-Font MA, Naya-Català, Calduch-Giner JA, Trilla-Fuertes L, Gamez-Pozo A, Arnau V, Perez-Sanchez J, Sempere JM, Gabaldon T, and Llorens C. 2022. RNASeq and VariantSeq applications and Server Side of GPRO suite. *ArXiv* https://arxiv.org/abs/2202.07473

- Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan C. 2015. The GOA database: gene Ontology annotation updates for 2015. *Nucleic Acids Res* 43:D1057-1063.
- Kotera M, Hirakawa M, Tokimatsu T, Goto S, Kanehisa M. 2012. The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. *Methods Mol.Biol*. 802:19-39.
- 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, platform-independent, community-supported software for describing and comparing microbial communities. *Appl.Environ.Microbiol*. 75:7537-7541.
- Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086-1092.
- The UniProt C. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 45:D158-D169.