Universidad de Navarra

Análisis e interpretación de datos de alto rendimiento (MC2)

Guía docente 2024-25

PRESENTACIÓN

Análisis e interpretación de datos de alto rendimiento

Vivimos en una época de transformación sin precedentes en el ámbito de la genómica, lo que ha propiciado avances significativos en nuestra comprensión de la biología, la medicina de precisión y la investigación de enfermedades a nivel molecular. La reducción de los costos de secuenciación masiva (NGS) y el desarrollo de la biología computacional han sido pilares fundamentales en este progreso, permitiendo el análisis de grandes conjuntos de datos genómicos de manera eficiente y precisa. En respuesta a esta creciente demanda, este curso se enfoca en proporcionar los conocimientos esenciales para el procesamiento y análisis de datos de secuenciación masiva (NGS) y tecnologías ómicas, adoptando un enfoque práctico y aplicado.

Comenzaremos explorando la evolución y los principios básicos de las tecnologías de secuenciación, así como sus diversas aplicaciones en la investigación biomédica. Además, realizaremos visitas a instalaciones de secuenciación de última generación y laboratorios de preparación de muestras, incluyendo el Centro de Secuenciación y el clúster Urederra en NASERTIC, que es uno de los nodos líderes en secuenciación y supercomputación (HPC) en Navarra.

Una vez entendido el funcionamiento técnico de la secuenciación, iniciaremos el aprendizaje del lenguaje UNIX, utilizado para acceder a grandes clústeres de computación y para el uso y creación de algoritmos de análisis de NGS. Continuaremos usando lecturas sintéticas de secuenciacion para llevar a cabo todo el proceso de análisis hasta la comparación genómica del brote epidémico de E. coli de 2011, con el objetivo de descubrir sus mecanismos de virulencia y resistencia antibiótica. Esta tecnología es puntera y actualmente se utiliza en la lucha contra la pandemia de COVID-19. Aplicaremos estos conceptos al análisis de expresión diferencial de genes, así como a los últimos avances en cuantificación y variación a nivel de transcrito. Finalmente, culminaremos con pautas sobre la nueva medicina de precisión orientada al paciente y tratamientos personalizados, equipando a los estudiantes con el conocimiento necesario para avanzar en esta área vanguardista.

La mayor parte de la computación práctica en este curso se realizará en un entorno de nube simulado, que emula un entorno HPC, donde todas las necesidades computacionales serán atendidas mediante el uso de entornos Docker y Conda. Durante el curso, también adquiriremos habilidades esenciales en el manejo del entorno UNIX, utilizando comandos de una sola línea (one-liners) para manipular y procesar datos de manera rápida y efectiva. Exploraremos el uso de entornos de gestión de paquetes como Conda y Mamba para configurar y mantener nuestras herramientas y librerías bioinformáticas. Nos familiarizaremos con la creación, configuración y despliegue de contenedores Docker, lo que nos permitirá construir ambientes de desarrollo reproducibles y portátiles. Asimismo, aprenderemos a manejar puertos y a generar salidas en formatos como HTML, facilitando la presentación y visualización de nuestros resultados. Además, aprovecharemos herramientas de inteligencia artificial como ChatGPT y GitHub Copilot para mejorar el aprendizaje y la creación de código asociado a la biología computacional.

Todo este conjunto de habilidades estará orientado a abordar grandes interrogantes biológicas mediante el análisis de datos masivos genómicos y transcriptómicos. Nos sumergiremos en el análisis de expresión génica diferencial y exploraremos los últimos avances en la cuantificación y variación de transcritos. Concluiremos el curso adquiriendo

habilidades en lenguajes de flujo de trabajo bioinformático como Snakemake y Nextflow, fundamentales para la creación de pipelines y workflows en entornos HPC.

Al finalizar este módulo, los alumnos estarán capacitados para diseñar y analizar modernas tecnologías omicas, trabajar eficientemente en entornos HPC-UNIX y preparar algoritmos para llevar a cabo todos los procesos necesarios en el análisis de datos genómicos. Esto les permitirá contribuir significativamente en la investigación biomédica y la medicina de precisión, generando conocimiento y respuestas a importantes preguntas en el campo de la biología molecular.

- **Carácter**: Optativa
- **ECTS**: 3
- **Curso y semestre**: 1º curso 2º semestre
- **Idioma**: Español. Se requieren conocimientos de inglés.
- **Título**: Máster en Métodos Computacionales en Ciencias.
- **Módulo** 4 Optativo y **materia** 4.1 Optatividad
- **Profesor responsable de la asignatura**: Igor Ruiz de los Mozos
- **Horario**:Calendario del Máster
- **Aula**: 1 edificio Los Castaños

RESULTADOS DE APRENDIZAJE (Competencias)

	Análisis e interpretación de datos de alto rendimiento
	COMPETENCIAS BÁSICAS
CB6	Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo o aplicación de ideas, a menudo en un contexto de investigación.
CB7	Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.
CB10	Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.
	COMPETENCIAS GENERALES

CG3	Conocer los principales problemas que se presentan en la adquisición y tratamiento de datos experimentales y cómo darles respuesta.
CG4	Comunicar tanto de manera oral como escrita un tema o datos de investigación en el área de las ciencias experimentales.
	COMPETENCIAS ESPECÍFICAS
CE5	Aplicar los métodos computacionales de procesamiento de datos a un problema científico particular de la disciplina de interés para el estudiante.
CE6	Diseñar un experimento científico para que sea rico en información, recogiendo gran cantidad de datos de manera estructurada que faciliten su procesamiento posterior.
CE8	Adquirir datos (bien en el laboratorio, o bien mediante minería on-line), organizarlos, filtrarlos, procesarlos, representarlos y refinarlos.
CE9	Extraer información de los datos con técnicas computacionales siguiendo un método científico.
CE10	Presentar los datos experimentales y la información científica de manera que se comuniquen de manera eficiente y fidedigna.
	COMPETENCIAS ESPECÍFICAS DE OPTATIVIDAD
CEOP5	Diseñar y analizar experimentos en el ámbito de las tecnologías ómicas, especialmente en el campo de la genómica y de la transcriptómica, con el fin de buscar nuevos biomarcadores, dianas terapéuticas o estudiar mecanismos moleculares.

Universidad de Navarra

PROGRAMA

Teoría de la secuenciación

- Evolución de las plataformas de secuenciación
- Desde las muestras hasta las librerías de secuenciación
- Desde la librería hasta las lecturas (reads)
- Estructura de los NGS (next-generation sequencing), control de calidad del pre-procesado
- Visión general de los flujos de trabajo bioinformáticos

Secuenciación práctica

- Visita al centro de secuenciación masiva de Nasertic
- Secuenciación del exoma y genoma completo (Whole Exome/Genome)
- Experimentos RNA-Seq
- Secuenciación Single-cell
- Spatial Genomics - Spatial Transcriptomics

Otras tecnologías Ómicas

- Introducción al estudio de la proteómica, metabolómica, epigenética, epigenómica y microbioma
- Técnicas de separación de proteínas (ej. electroforesis en gel bidimensional)
- Análisis de interacción proteína-proteína
- Análisis de datos metabolómicos
- Análisis de datos epigenómicos y epitranscriptómicos
- Estudio de patrones de metilación del ADN y modificaciones de histonas
- Aplicaciones de la epigenómica en la investigación de enfermedades
- Análisis de la diversidad microbiana y funciones metabólicas
- Integración de datos ómicos (multi-omics)
- Métodos bioinformáticos para la integración de datos transcriptómicos, proteómicos y metabolómicos
- Análisis de redes biológicas

Unix y clústeres de alto rendimiento

- Visita al centro de Super Computación Urederra de Nasertic
- Acceso a recursos computacionales en la nube emulando a un HPC
- Curso intensivo de UNIX/bash
- Curso intensivo en Ordenadores de Alto Rendimiento (High-Performance Computers)

Alineamiento NGS con computación avanzada

- Evaluación de calidad de datos de secuenciación
- Limpieza, descontaminación y demultiplex
- Teoría del alineamiento NGS, desde datos crudos multiplexados a lecturas alineadas.
- Mejorar los parámetros de alineamiento y estado del arte de los análisis NGS.
- Alineamiento contra referencia y de novo

Detección de diferencias NGS (variant calling) y secuenciación genómica

- Traducción de las lecturas de secuenciación para obtener el genoma completo de interés
- Identificar mutaciones de nucleótidos puntuales que pueden convertirse en diferencias fenotípicas o virulentas

Comparación computacional del genoma

- Desde la secuencia consenso del DNA hasta la exploración genómica para inferir ganancia o pérdida de funciones.
- Especialización genómica y la evolución genómica.
- Análisis genómico en un caso de pandemia.
- Diseño y selección de antibióticos susceptibles
- Evaluación de mecanismo de virulencia microbiológica

Análisis de RNA-seq

- Tecnologías transcriptómicas de secuenciación
- Cuantificación de la expresión génica
- Clusterización y caracterización de muestras, hierarchical y k-means
- Reducción de dimensiones PCA
- Análisis de expresión diferencial: teoría y practica
- Integración y anotación de datos
- Análisis funcional: Gene Set Enrichment Analysis
- Corrección de desviación por lotes
- Cuantificación a nivel de transcrito, nuevos algoritmos sin mapeo genómico
- Predicción de efectos sobre el trascriptoma

Computación de alto rendimiento con workflows en Snakemake

- Iniciación a lenguaje de *workflows* en Snakemake.
- Análisis NGS en Snakemake.
- Análisis de reproducibilidad, trazabilidad y aprovechamiento de recursos en entornos HPC.
- Descripción de Nextflow y nf-core para la creación de pipelines bioinformáticas

Aplicación de técnicas de Inteligencia Artificial para la creación de Código Bioinformático

- Iniciación IA para creación de código computacional
- Prompt engineering en bioinformática
- Auto digestión y auto-comprension de datos biológicos ayudado por AI
- Programación en Bash, Python y R guiada por AI

Documentación, trazabilidad y estabilidad del código bioinformático en GitHub

- Fundamentos de Git y GitHub para la gestión de código
- Creación y mantenimiento de repositorios
- Documentación efectiva del código y proyectos
- Estrategias para asegurar la trazabilidad del desarrollo del software
- Mantenimiento de la estabilidad del código mediante control de versiones
- Uso de GitHub Actions para automatización y CI/CD (Integración Continua /Despliegue Continuo)
- Mejores prácticas para la colaboración en proyectos bioinformáticos mediante GitHub

Universidad de Navarra

ACTIVIDADES FORMATIVAS

Cada uno de los módulos tendrá un formato diferente. El primer módulo será más teórico y se realizarán visitas guiadas a los laboratorios. El segundo y tercer módulo serán 90% prácticos. En el trabajo final se espera que se apliquen los conceptos utilizados en los tres módulos.

EVALUACIÓN

Cada uno de los módulos valdrá un 20% de la nota (con 3 módulos 60%) y el último trabajo valdrá 40%.

HORARIOS DE ATENCIÓN

No hay horario específico para ello. Concertar cita previa por e-mail y se acuerda la fecha y hora de la tutoría:

Igor Ruiz de los Mozos: igor.ruiz@unav.es

BIBLIOGRAFÍA

Bioinformatics Data Skills

by Vince Buffalo, July 2015, Publisher(s): O'Reilly Media, Inc. ISBN: 9781449367503

https://github.com/Kur1sutaru/bioinformatics-data-skills/blob/main/book-bioinformatics-data-skills.pdf

Next-generation Sequencing and Sequence Data Analysis

By: Ping Chiu, Kuo. Sharjah : Bentham Science Publishers. 2016. Localízalo en la Biblioteca [Recurso electrónico]

Bioinformatics and functional genomics

Pevsner, Jonathan. John Wiley & Sons, 2015. Localízalo en la Biblioteca

Genome Data Analysis

By: Kim, Ju Han. Singapore : Springer Singapore, 2019. XVI, 367 p. : 645 il., 236 il. col. Language: English, Base de datos: Catálogo de la Biblioteca de la Universidad de Navarra. Localízalo en la Biblioteca [Recurso electrónico]

Computational methods for next generation sequencing data analysis

Edited by Ion Măndoiu, Alexander Zelikovsky. Localízalo en la Biblioteca [Recurso electrónico]

Biostar Handbook Student Edition

Dr. István Albert, An introduction to Bioinformatics as a scientific field of study. Covers all aspects of bioinformatics. https://biostar.myshopify.com/



Universidad
de Navarra