CSIC
Consejo Superior de Investigaciones Científicas

Nota de prensa

CSIC comunicación
Tel.: 91 568 14 77
comunicacion@csic.es
www.csic.es

Valencia, jueves 11 de abril de 2024

El CSIC desarrolla una nueva herramienta informática para investigar la complejidad del genoma

- **El Instituto de Biología Integrativa de Sistemas (CSIC-UV) publica en 'Nature Methods' un 'software' propio para analizar datos obtenidos por secuenciación de lectura larga del genoma**

- **Este sistema permite descubrir nuevas moléculas de ARN y asignarles una función en la creación de tejidos. Se ahonda así en el conocimiento de la formación del organismo y sus enfermedades**

Ilustración 3D de cadena de ARN o ARNm. / iStock

La complejidad de un organismo emerge de su genoma, el libro que contiene las instrucciones de su ADN para la vida. El método para leer este libro, la secuenciación, ha evolucionado hacia la lectura de fragmentos cada vez más largos del genoma. En este campo, un grupo de investigación liderado por el Instituto de Biología Integrativa de Sistemas (I2SysBio), centro mixto del Consejo Superior de Investigaciones Científicas (CSIC) y la Universitat de València (UV), ha publicado en *Nature Methods* una mejora de

un programa informático propio capaz de descubrir nuevos tránscritos, moléculas de ARN que usan los genes para sintetizar proteínas y crear tejidos, a partir de su secuenciación con instrumentos de lectura larga, así como asignarles una función en la formación del organismo.

La secuenciación de lectura larga (long-read sequencing) es la tercera generación de métodos de secuenciación del genoma. Frente a la lectura de fragmentos cortos, que analiza unos 200 nucleótidos (las *letras* que componen los genes), los métodos de lectura larga pueden obtener lecturas 100 veces más largas, unos 20.000 nucleótidos, lo que deja menos *huecos* en la información del genoma para rellenar mediante herramientas bioinformáticas. Esta fue una de las razones para que la propia *Nature Methods* lo considerase Método del Año 2022.

Unos años antes, en 2018, la investigadora **Ana Conesa**, entonces en la Universidad de Florida, desarrolló un programa informático llamado SQANTI para analizar la información que se extraía mediante estos métodos de lectura larga. Ahora, su equipo de investigación en el I2SysBio publica en *Nature Methods* una mejora sustancial de este *software* que se puede usar libremente en los principales sistemas comerciales que emplean secuenciación de lectura larga, Pacific Biosciences (PacBio) y Oxford Nanopore Technologies (ONT).

“Las técnicas de lectura larga analizan mejor la complejidad de los tránscritos y el transcriptoma humanos”, opina Conesa. Esto identifica la porción del genoma que se lee en cada célula para dar lugar a tejidos y órganos. Así, un único gen puede dar lugar, mediante pequeños cambios en la estructura de ARN que codifica, a una gran diversidad tránscritos, y con ellos de proteínas con distintas funciones celulares. “La secuenciación de lectura corta no puede resolver este puzle. La lectura larga reconstruye mejor la complejidad funcional del transcriptoma humano, algo clave para estudiar determinadas enfermedades, sobre todo de tipo neurológico y en cáncer”, sostiene la investigadora del CSIC.

Entender mejor la complejidad del organismo y las enfermedades

La versión publicada ahora, SQANTI3, soluciona algunos problemas anteriores, derivados de la degradación del ARN o el análisis único de cada molécula, para introducir notables mejoras. El programa es capaz ahora de descubrir nuevos tránscritos que no estaban en las bases de datos del genoma que usan estos programas informáticos. Además, mediante técnicas de Inteligencia Artificial, el *software* puede asignar información funcional para el nuevo tránscrito, “algo esencial para entender la complejidad funcional del organismo y de las enfermedades”, remarca Conesa.

Para desarrollar este programa informático se ha usado el clúster de computación Garnatxa del I2SysBio, que dispone de 15 nodos de computación capaces de ofrecer 950 hilos de cómputo en paralelo. Además, el grupo Genómica de la Expresión Génica que dirige Ana Conesa en el I2SysBio participa en ELIXIR, una de las infraestructuras estratégicas para el Foro Estratégico Europeo sobre Infraestructuras de Investigación (ESFRI) que permite a laboratorios de ciencias de la vida de toda Europa compartir y almacenar sus datos.

En el desarrollo de SQANTI3 colaboraron la Universidad de Florida y Pacific Biosciences, una de las empresas que comercializa la tecnología para la secuenciación de lectura larga mediante su sistema PacBio, que recomienda el uso del software español para analizar sus datos. El uso del programa informático es libre, contando ya con "miles de usuarios en todo el mundo", según Conesa, aunque "el éxito de esta herramienta requiere también de más personal técnico para atender a las numerosas peticiones que recibimos". Así, la investigadora ha coliderado la reciente puesta en marcha de la Conexión CSIC de Biología Computacional y Bioinformática, una plataforma para conectar personas, métodos y recursos en estos ámbitos en el CSIC.

Pardo-Palacios, F.J., Arzalluz-Luque, A., Kondratova, L. et al. **SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms**. *Nature Methods*. DOI: https://doi.org/10.1038/s41592-024-02229-2

Isidoro García / CSIC Comunicación – Comunidad Valenciana

comunicacion@csic.es