CSIC comunicació
Tel.: 96 362 27 57
g.prensa@dicv.csic.es

València, 11 d'abril de 2024

El CSIC desenvolupa una nova eina informàtica per a investigar la complexitat del genoma

- **L'Institut de Biologia Integrativa de Sistemes (CSIC-UV) publica en 'Nature Methods' un 'programari' propi per a analitzar dades obtingudes per seqüenciació de lectura llarga del genoma**

- **Aquest sistema permet descobrir noves molècules d'ARN i assignar-los una funció en la creació de teixits. S'aprofundeix així en el coneixement de la formació de l'organisme i les seues malalties**

Il·lustració de renderitzat 3D de cadena d'ARN o ARNm amb espai de còpia. Crèdits: iStock.

La complexitat d'un organisme emergeix del seu genoma, el llibre que conté les instruccions del seu ADN per a la vida. El mètode per a llegir aquest llibre, la seqüenciació, ha evolucionat cap a la lectura de fragments cada vegada més llargs del genoma. En aquest camp, un grup d'investigació liderat per l'Institut de Biologia Integrativa de Sistemes (I2SysBio), centre mixt del Consell Superior d'Investigacions Científiques (CSIC) i la Universitat de València (UV), ha publicat en *Nature Methods* una millora d'un programa informàtic propi capaç de descobrir nous trànscrits, molècules d'ARN que usen els gens per a sintetitzar proteïnes i crear teixits, a partir de la seua seqüenciació amb instruments de lectura llarga, així com assignar-los una funció en la formació de l'organisme.

La seqüenciació de lectura llarga (long-read sequencing) és la tercera generació de mètodes de seqüenciació del genoma. Enfront de la lectura de fragments curts, que analitza uns 200 nucleòtids (les 'lletres' que componen els gens), els mètodes de lectura llarga poden obtindre lectures 100 vegades més llargues, uns 20.000 nucleòtids, la qual cosa deixa menys 'buits' en la informació del genoma per a emplenar mitjançant eines bioinformàtiques. Aquesta va ser una de les raons perquè la mateixa *Nature Methods* ho considerara 'Mètode de l'Any 2022'.

Uns anys abans, en 2018, la investigadora **Ana Conesa**, llavors en la Universitat de Florida, va desenvolupar un programa informàtic anomenat SQANTI per a analitzar la informació que s'extreia mitjançant aquests mètodes de lectura llarga. Ara, el seu equip d'investigació a l'I2SysBio publica en *Nature Methods* una millora substancial d'aquest programari que es pot usar lliurement en els principals sistemes comercials que empren seqüenciació de lectura llarga, Pacific Biosciences (PacBio) i Oxford Nanopore Technologies (ONT).

"Les tècniques de lectura llarga analitzen millor la complexitat dels trànscrits i el transcriptoma humans", opina Conesa. Això identifica la porció del genoma que es llig en cada cèl·lula per a donar lloc a teixits i òrgans. Així, un únic gen pot donar lloc, mitjançant xicotets canvis en l'estructura d'ARN que codifica, a una gran diversitat trànscrits, i amb ells de proteïnes amb diferents funcions cel·lulars... "La seqüenciació de lectura curta no pot resoldre aquest puzle. La lectura llarga reconstrueix millor la complexitat funcional del transcriptoma humà, una cosa clau per a estudiar determinades malalties, sobretot de tipus neurològic i en càncer", sosté la investigadora del CSIC.

Entendre millor la complexitat de l'organisme i les malalties

La versió publicada ara, SQANTI3, soluciona alguns problemes anteriors, derivats de la degradació de l'ARN o l'anàlisi única de cada molècula, per a introduir notables millores. El programa és capaç ara de descobrir nous tránscrits que no estaven en les bases de dades del genoma que usen aquests programes informàtics. A més, mitjançant tècniques d'Intel·ligència Artificial, el programari pot assignar informació funcional per al nou trànscrit, "una cosa essencial per a entendre la complexitat funcional de l'organisme i de les malalties", remarca Conesa.

Per a desenvolupar aquest programa informàtic s'ha usat el clúster de computació Garnatxa de l'I2SysBio, que disposa de 15 nodes de computació capaces d'oferir 950 fils de còmput en paral·lel. Ademés, el grup Genòmica de l'Expressió Gènica que dirigeix Ana Conesa a l'I2SysBio participa en ELIXIR, una de les infraestructures estratègiques per a Fòrum Estratègic Europeu sobre Infraestructures d'Investigació (ESFRI) que permet a laboratoris de ciències de la vida de tota Europa compartir i emmagatzemar les seues dades.

En el desenvolupament de SQANTI3 van col·laborar la Universitat de Florida i Pacific Biosciences, una de les empreses que comercialitza la tecnologia per a la seqüenciació de lectura llarga mitjançant el seu sistema PacBio, que recomana l'ús del programari

espanyol per a analitzar les seues dades. L'ús del programa informàtic és lliure, comptant ja amb “milers d'usuaris a tot el món”, segons Conesa, encara que “l'èxit d'aquesta eina requereix també de més personal tècnic per a atendre les nombroses peticions que rebem”. Així, la investigadora ha coliderat la recent posada en marxa de la Connexió CSIC de Biologia Computacional i Bioinformàtica, una plataforma per a connectar persones, mètodes i recursos en aquests àmbits en el CSIC.

Referència:

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* (2024). https://doi.org/10.1038/s41592-024-02229-2