Note: This document contains an adapted Catalan translation of the abstract of Carbonell-Sala et al. 'CapTrap-Seq: A platform-agnostic and quantitative approach for high-fidelity full-length RNA sequencing", *Nature Communications* (2024), https://doi.org/10.1038/s41467-024-49523-3, and a non-specialist summary

CapTrap-Seq: un mètode quantitatiu per a la seqüenciació completa i fidel de transcrits d'RNA

Silvia Carbonell-Sala[1], Tamara Perteghella[1,2], Julien Lagarde[1,3], Hiromi Nishiyori[4], Emilio Palumbo[1], Carme Arnan[1], Hazuki Takahashi[4], Piero Carninci[4,5], Barbara Uszczynska-Ratajczak[1,6,], Roderic Guigó[1,2,]

1. Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
2. Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
3. Flomics Biotech, SL, Carrer de Roc Boronat 31, 08005 Barcelona, Catalonia, Spain
4. Laboratory for Transcriptome Technology, RIKEN Center for Integrative Medical Sciences (IMS), Yokohama, Kanagawa, Japan
5. Human Technopole, Milan, Italy.
6. Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland.

La correspondència en català sobre aquest article es pot adreçar a RG (roderic.guigo@crg.cat) o SCS (silvia.carbonell@crg.cat).

Resum

La seqüenciació de l'RNA de lectura llarga ("long-read RNAseq") és essencial per tal de produir una anotació precisa i exhaustiva dels gens en els genomes eucariotes. Malgrat els avenços en el seu rendiment i en la seva precisió, aconseguir una identificació fiable i completa dels transcrits d'RNA continua sent un repte per als mètodes de seqüenciació de lectura llarga. Per tal de fer front a aquest repte, hem desenvolupat CapTrap-seq, un mètode de preparació de biblioteques de cDNA (DNA complementari a l'RNA), que combina l'estratègia "Cap-trapping" per seleccionar els transcrits des del seu inici amb l'estratègia "oligo(dT)" per seleccionar els transcrits des del seu final. En el nostre estudi, hem avaluat el rendiment de CapTrap-seq en teixits humans i de ratolí i l'hem comparat amb el d'altres mètodes utilitzats per a la preparació de biblioteques per a la seqüenciació d'RNA. Hem emprat dues tecnologies de seqüenciació: ONT i PacBio. Per tal d'investigar la capacitat quantitativa de CapTrap-seq a l'hora de mesurar l'abundància de transcrits i la seva precisió en la reconstrucció de molècules d'RNA, hem implementat una nova estratègia de "capping" per a seqüències d'RNA sintètic que imita la formació natural de l'estructura "cap". Les nostres avaluacions, que incorporen les dades del projecte LRGASP (Long-read RNA-seq Genome Annotation Assessment Project, és a dir, el Projecte d'Avaluació de l'Anotació del Genoma mitjançant RNAseq de Lectura Llarga), mostren que CapTrap-seq és un mètode competitiu per a la preparació de biblioteques d'ARN, capaç de generar seqüències de transcrits complets, independentment de la plataforma de seqüenciació utilitzada.

Nota: Aquest resum és una traducció adaptada de l'*abstract* de l'article Carbonell-Sala et al. 'CapTrap-Seq: A platform-agnostic and quantitative approach for high-fidelity full-length RNA sequencing", *Nature Communications* (2024), https://doi.org/10.1038/s41467-024-49523-3.

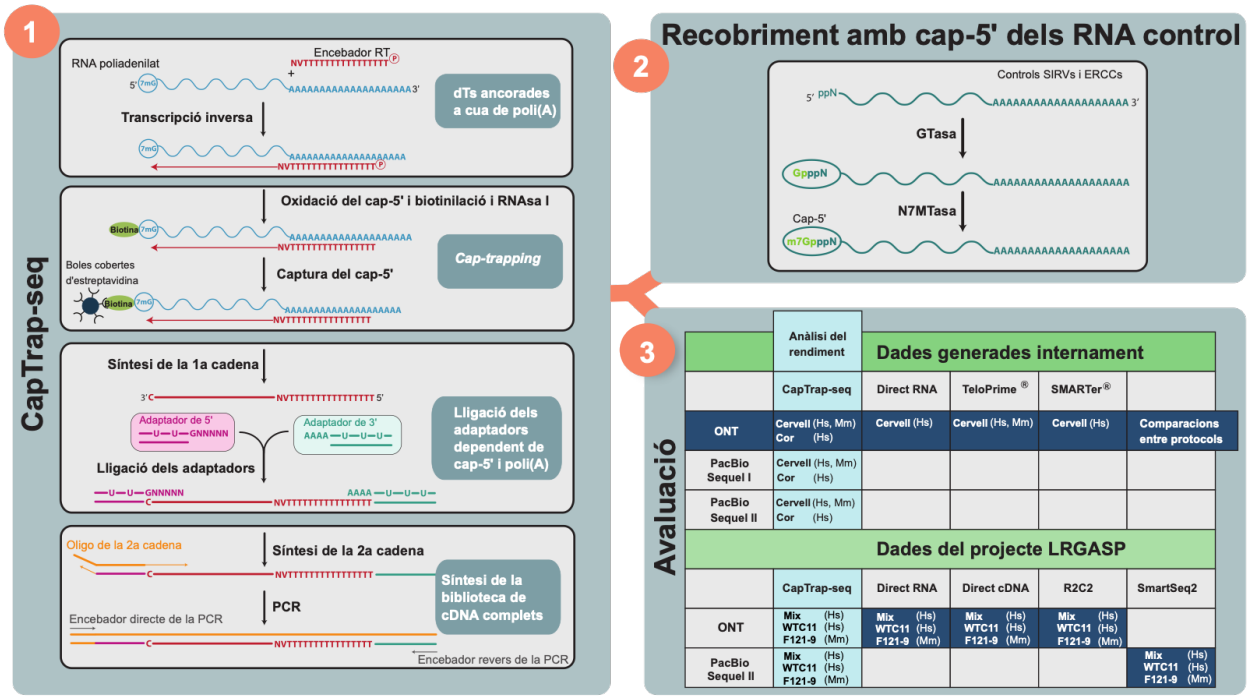
Sumari divulgatiu

El genoma -la seqüència de DNA que es troba als cromosomes dins les cèl·lules- codifica les instruccions que determinen les característiques biològiques dels éssers vius. Aquestes instruccions estan específicament localitzades en regions del genoma que anomenem gens. El genoma humà, en concret, conté uns 60.000 gens. Quan aquestes regions s'activen es transcriuen a molècules d'RNA, les quals, eventualment, donen lloc a proteïnes. Les proteïnes són els components funcionals i estructurals de les cèl·lules.

El tipus i l'estat de les cèl·lules (per exemple, en el cas de malaltia) són, doncs, determinats pel seu contingut d'RNA, és a dir, per l'abundància de les molècules d'RNA (els transcrits) que s'originen als gens (l'anomenada expressió dels gens). L'obtenció de la seqüència dels RNA que existeixen a les cèl·lules en un moment determinat (la seqüenciació de l'RNA, RNAseq, per la seva abreviació en anglès), és, en conseqüència, la millor manera de determinar l'estat cel·lular.

Fins fa poc, els mètodes d'RNAseq només permetien obtenir seqüències molt curtes, de longitud molt inferior a la dels transcrits, la qual cosa feia difícil identificar i quantificar amb precisió l'expressió dels gens. Recentment hom ha desenvolupat mètodes d'RNAseq de lectura llarga ("long-read RNAseq", en anglès) que permeten, en principi, obtenir la seqüència completa dels transcrits. Tanmateix, encara hi ha reptes importants que cal superar per tal d'obtenir aquestes seqüències. Per tal de superar-los, hem desenvolupat una nova tècnica anomenada CapTrap-seq (Figura 1). Aquesta tècnica millora la manera com s'extreu i es prepara l'RNA abans de ser seqüenciat. CapTrap-seq combina dues estratègies: la "Cap-trapping" (o atrapador de caps) per capturar els transcrits des del seu inici i l'"oligo(dT)" per capturar els transcrits des del seu final.

Hem avaluat el funcionament de CapTrap-seq en cèl·lules obtingudes de teixits humans i de ratolí i l'hem comparat amb altres mètodes de preparació d'RNA. Hem utilitzat dues plataformes de seqüenciació de lectura llarga: ONT (Oxford Nanopore Technologies) i PacBio (Pacific Biosciences) per verificar la precisió de CapTrap-seq a l'hora de mesurar l'abundància dels transcrits (és a dir, l'expressió dels gens). També hem desenvolupat una nova estratègia de "capping" (és a dir, una estratègia per afegir l'estructura química que caracteriza de manera natural l'inici dels transcrits) en seqüències d'RNA sintètic.



	Anàlisi del rendiment	Dades generades internament			
	CapTrap-seq	Direct RNA	TeloPrime®	SMARTer®	
ONT	Cervell (Hs, Mm) Cor (Hs)	Cervell (Hs)	Cervell (Hs, Mm)	Cervell (Hs)	Comparacions entre protocols
PacBio Sequel I	Cervell (Hs, Mm) Cor (Hs)				
PacBio Sequel II	Cervell (Hs, Mm) Cor (Hs)				

	Dades del projecte LRGASP				
	CapTrap-seq	Direct RNA	Direct cDNA	R2C2	SmartSeq2
ONT	Mix (Hs) WTC11 (Hs) F121-9 (Mm)	Mix (Hs) WTC11 (Hs) F121-9 (Mm)	Mix (Hs) WTC11 (Hs) F121-9 (Mm)	Mix (Hs) WTC11 (Hs) F121-9 (Mm)	
PacBio Sequel II	Mix (Hs) WTC11 (Hs) F121-9 (Mm)				Mix (Hs) WTC11 (Hs) F121-9 (Mm)

Figura 1. (1) **Flux de treball del CapTrap-seq** per enriquir transcrits de longitud completa des de l'extrem 5' fins a l'extrem 3'. (2) **Recobriment amb cap-5' dels RNA control:** Esquema de la nova estratègia dissenyada per afegir un recobriment a les seqüències dels RNA sintètics (controls), (3) **Avaluació de CapTrap-seq:** protocols alternatius amb els quals hem comparat CapTrap-seq en dues plataformes diferents de seqüenciació: PacBio (Sequel I i II) i ONT, en dades produïdes a partir de teixits d'origen humà (Homo sapiens, Hs) i de ratolí (Mus musculus, Mm). Algunes d'aquestes dades han estan generades internament i d'altres han estat generades com a part del projecte LRGASP (https://www.gencodegenes.org/pages/LRGASP). Mix és una barreja de cèl·lules mare embrionàries i cèl·lules d'origen endodèrmic. WTC11 són cèl·lules mare pluripotents. F121-9 són cèl·lules mare embrionàries.

Els resultats del nostre estudi demostren que CapTrap-seq és un mètode competitiu per a la preparació de l'ARN, i que pot ser utilitzat indistintament en diferents plataformes de seqüenciació, la qual cosa el fa flexible i adaptable a les necessitats dels investigadors. Hom està utilitzant CapTrap-seq dins el projecte GENCODE, el consorci internacional que produeix el mapa detallat dels gens i transcrits en els genomes humà i de ratolí.

Tot i que CapTrap-seq ofereix molts avantatges, també té certes limitacions. En primer lloc, requereix una quantitat més gran d'RNA que altres protocols. En segon lloc, tendeix a produir seqüències més curtes. En tercer lloc, té com a objectiu principal la identificació d'un tipus determinat de molècules d'RNA, les quals constitueixen només una fracció (biològicament molt important, això sí) de totes les molècules d'RNA que es troben a les cèl·lules.

En conclusió, la seqüenciació d'RNA de lectura llarga, juntament amb les eines d'anàlisi de dades adequades, té el potencial de produir per primera vegada mapes de transcrits acurats dels genomes dels éssers vius i estimacions precises de l'expressió gènica en tipus cel·lulars diferents. Això és especialment important perquè hi ha grans projectes internacionals en marxa per tal de seqüenciar els genomes de totes les espècies de la Terra (l'anomenat "Earth BioGenome Project") i per tal de identificar-ne tots els tipus cel·lulars (l'anomenat "Biodiversity Cell Atlas"). En tots dos projectes, el nostre país té un paper rellevant.