UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN BIOINGNERIA

Sanità digitale e ingegneria clinica

Rilevazione qualitativa e quantitativa dell'espressione di RNA circolari in campioni longitudinali

Relatore: Prof. Enrico Lavezzo

Laureanda: Mariachiara Vardeu

ANNO ACCADEMICO 2022 – 2023

Data di Laurea: 18/10/2023

Sommario

Abstract

La biologia molecolare ha compiuto notevoli progressi nel corso degli anni, portando alla scoperta di una vasta gamma di RNA non codificanti coinvolti nella regolazione e nell'espressione dei geni. Tra questi, gli RNA circolari (circRNA) hanno attirato particolare attenzione negli ultimi anni per il loro ruolo potenziale come marcatori diagnostici e prognostici in diverse malattie umane.

I circRNA sono una classe di RNA che si differenzia dai classici RNA lineari per la loro struttura a forma di anello, che li rende resistenti all'azione delle esonucleasi e conferisce loro una maggiore stabilità rispetto agli RNA lineari.

Questa caratteristica è il risultato di una reazione di "backsplicing" durante la trascrizione del DNA: mentre gli RNA lineari vengono generati attraverso lo splicing classico, in cui si ha un legame covalente tra gli esoni in sequenza, negli RNA circolari le estremità 3' e 5' sono unite covalentemente formando una struttura circolare.

I circRNA esercitano un ruolo chiave nella regolazione dell'espressione genica attraverso l'interazione con i microRNA e le proteine e si ipotizza che l'alterazione dell'espressione di specifici circRNA possa avere un ruolo importante nelle malattie e/o nei disturbi umani. Nonostante il loro coinvolgimento in diversi processi biologici, i circRNA sono però molto poco abbondanti nel trascrittoma cellulare, e la loro individuazione rimane complessa.

L'obiettivo principale di questa tesi è stato lo sviluppo di una pipeline innovativa per l'identificazione e la quantificazione dei circRNA in dataset di RNA-seq simulati e reali. Particolare attenzione è stata dedicata all'analisi di campioni longitudinali, consentendo l'osservazione delle variazioni dei circRNA nel tempo. Sono stati utilizzati tre diversi software, già pubblicati in letteratura per questo preciso scopo, per i quali sono state calcolate e comparate le misure di *precision* e *recall*

I risultati evidenziano che il tool CIRI2 è il più affidabile nell'identificazione e quantificazione dei circRNAs nei diversi dataset.

Questo lavoro di ricerca fa parte di un progetto più ampio che mira a valutare il ruolo dei circRNA nello sviluppo neuronale fetale in presenza di infezioni congenite da virus patogeni umani. Nell'ambito di questo progetto sono in corso di produzione dei dataset sperimentali di trascrittomica, arricchiti della componente di circRNA, provenienti da cellule staminali neurali infettate con diversi virus umani patogeni e responsabili di malattie congenite. Su questi campioni sarà possibile studiare i profili di espressione dei circRNA nelle diverse fasi del processo di infezione e di identificare potenziali candidati per una successiva caratterizzazione funzionale. La pipeline sviluppata in questa tesi sarà impiegata per l'analisi

di tali dataset, fornendo un tassello fondamentale per la buona riuscita di questo ambizioso progetto.

Introduzione

La storia della scoperta degli RNA circolari risale al 1976, quando il noto chimico Heinz L. Sänger e il suo team incapparono casualmente in un RNA circolare durante uno studio sui viroidi, agenti patogeni delle piante.[1] Allo stesso tempo, l'uso della microscopia elettronica permise l'osservazione della presenza di RNA circolari anche nel virus di Sendai.[2] In questo contesto, si osservò che gli RNA circolari derivano da processi autocatalitici a carico di introni negli eucarioti unicellulari, o da introni dell'RNA ribosomiale negli archei.[3,4] Tuttavia, la scoperta di RNA circolari endogeni espressi negli eucarioti superiori avvenne solo negli anni '90.

La comprensione degli RNA circolari ha richiesto anni di studio e tecnologie avanzate per poterli isolare e caratterizzare.

I vari studi hanno rivelato che i circRNA sono molto più diffusi di quanto si pensasse inizialmente e che potrebbero svolgere ruoli importanti nelle cellule, ma, nonostante i progressi, il campo degli RNA circolari rimane largamente inesplorato e presenta sfide stimolanti.

Il meccanismo di generazione e di regolazione dei circRNA, così come il loro coinvolgimento in malattie umane e processi fisiologici, rappresentano aree di ricerca in continua espansione.[4,5]

Processo di formazione degli RNA circolari: back-splicing

Il processo di splicing è un fondamentale meccanismo biologico che si verifica nel nucleo delle cellule eucariotiche, svolgendo un ruolo cruciale nella regolazione dell'espressione genica e nella produzione di proteine funzionali. Il termine "splicing" ha origine dal verbo inglese "to splice" che indica l'atto di unire o giungere. In questo contesto, lo splicing si riferisce alla giunzione di diverse parti di un RNA pre-messaggero (pre-mRNA) al fine di formare un RNA messagero (mRNA) maturo, pronto per la traduzione in proteine.

Lo splicing è mediato dallo spliceosoma, un complesso ribonucleoproteico costituito da molteplici proteine leganti l'RNA, comunemente note come fattori di splicing. Questo processo è soggetto a una stretta regolazione, e lo spliceosoma riveste un ruolo di notevole rilevanza, in quanto si configura come una struttura dinamica, in cui i fattori di splicing possono modulare l'attività del complesso, potenziandola o reprimendola a seconda delle condizioni cellulari.[5,6]

Inizialmente, il sistema di splicing identifica una sequenza contenente una guanina e un uracile (GU) all'estremità 5' dell'introne, e allo stesso tempo riconoscendo una sequenza chiamata "branch point sequence", contenente una adenina all'estremità 3' dell'introne stesso.

Mediante una serie di reazioni altamente coordinate, gli scRNP (small cytoplasmic ribonucleoprotein), complessi molecolari costituiti da RNA ribosomiale e proteine presenti nello spliceosoma, avvicinano con precisione il punto di diramazione verso la regione di giunzione a monte dell'introne, contribuendo alla formazione del complesso di splicing.
Un passaggio cruciale avviene quando una reazione chimica porta all'ancoraggio dell'estremità 5' dell'introne con l'esone situato a monte, creando una struttura circolare chiamata lariat. La risoluzione di questa struttura richiede l'intervento dell'estremità 3' dell'esone a monte, che reagisce con una sequenza nucleotidica contenente adenina e guanina localizzata all'estremità 5' dell'esone a valle. Questo complesso meccanismo di reazioni culmina nell'unione dei due esoni e nell'eliminazione dell'introne, un processo noto come giunzione (fig.1).[7]

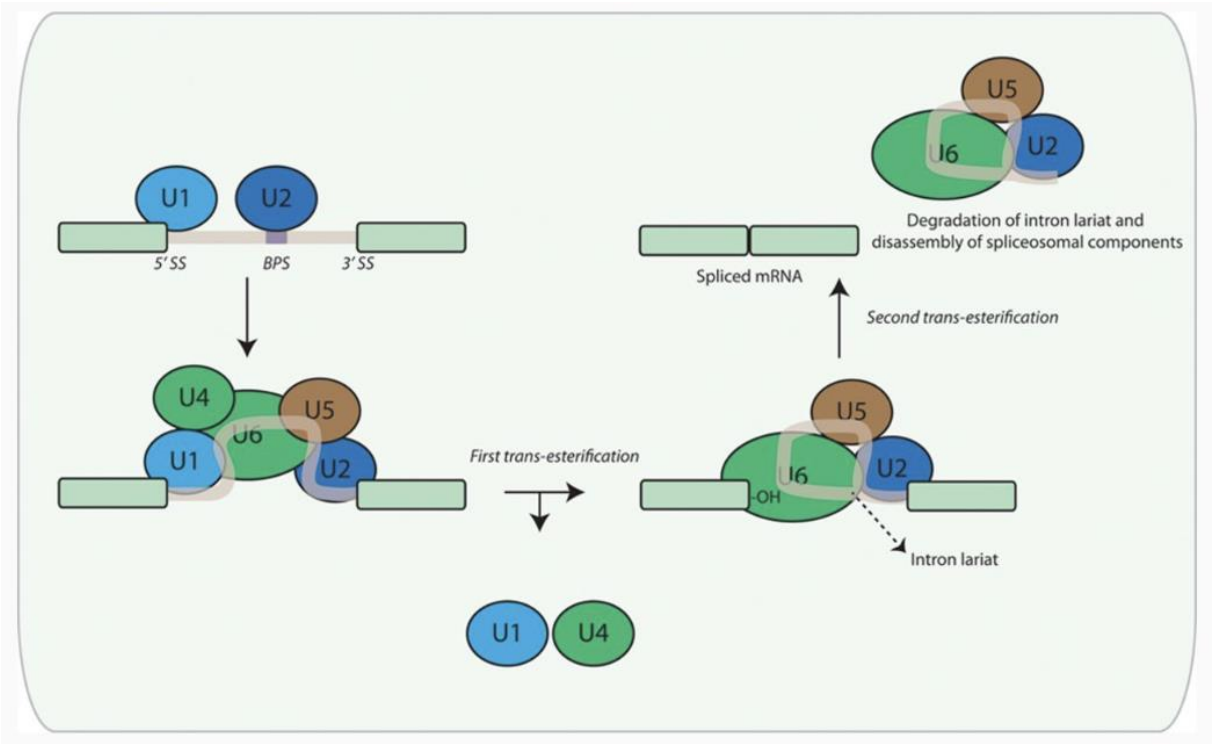


Figura 1: meccanismo di splicing[8]

Nella maggior parte dei casi negli organismi eucarioti, il processo di splicing del pre-mRNA avviene contemporaneamente al processo di trascrizione del DNA in RNA, in una modalità nota come cotrascrizionale. Tuttavia, ci sono alcuni casi significativi in cui lo splicing avviene successivamente, a livello post-trascrizionale. In tali circostanze, lo splicing si verifica dopo la completa trascrizione dell'RNA a partire dal DNA. Questo evento può influenzare la struttura e la funzione dell'mRNA maturo prima che venga tradotto in proteina.[6]
Lo splicing maggiore e minore rappresentano due meccanismi distinti di rimozione degli introni durante la maturazione dell'mRNA (fig.2).

Lo splicing maggiore è il tipo di splicing più comune ed è coinvolto nella rimozione degli introni e nell'unione degli esoni per formare una sequenza di mRNA matura. Lo splicing minore effettua lo splicing di un piccolo sottoinsieme di introni chiamati introni minori, che presentano sequenze che differiscono significativamente dalle sequenze tipiche o consensuali degli introni maggiori. Gli spliceosomi responsabili dei due processi condividono la maggior parte delle proteine, tranne che lo spliceosoma minore contiene sette proteine uniche che fanno parte dell'U11/U12 di-snRNP.[9]

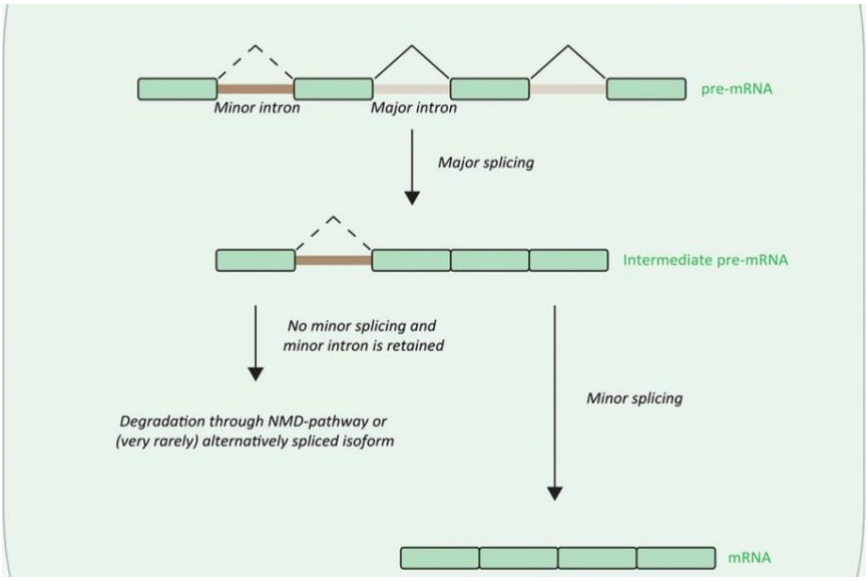


Figura 2: meccanismo di splicing maggiore e minore[8]

Gli RNA circolari (circRNA) rappresentano una classe particolare di RNA non codificanti che emerge dal complesso meccanismo di back-splicing, che comporta la fusione di esoni, introni, o entrambi, formando anelli covalentemente chiusi.

Nonostante il primo circRNA sia stato identificato più di 40 anni fa, questi trascritti sono stati a lungo considerati rari nelle cellule e spesso considerati come rumore di fondo in esperimenti di sequenziamento.

Questa percezione è stata in parte influenzata dal dogma prevalente che la maggior parte degli eventi di splicing avvenga co-trascrizionalmente, cioè poco dopo che un introne è stato completamente trascritto. Di conseguenza, l'introne insieme al suo accettore dello splicing associato (sito di splicing 3') dovrebbe essere rimosso prima che siano stati trascritti i siti di donazione dello splicing a valle (sito di splicing 5'), rendendo la maggior parte delle reazioni di back-splicing impossibili.

Oggi si riconosce che alcuni introni vengono rimossi lentamente o post-trascrizionalmente, consentendo il verificarsi delle reazioni di back-splicing.

Le reazioni di back-splicing sono spesso facilitate da sequenze introniche ripetute che si appaiano tra loro e avvicinano siti di splicing intermedi. È stato infatti osservata in numerosi studi la presenza di ripetizioni introniche complementari. Oltre agli elementi Alu, che sono specifici per i primati, una gamma diversificata di sequenze complementari può essere localizzata attorno agli RNA circolari, inclusi segmenti non ripetitivi.[4]

I circRNA derivano dai precursori di RNA che danno origine agli mRNA codificanti proteine mediante lo splicing canonico. La loro struttura caratteristica è costituita da loop chiusi in cui le estremità 3' e 5' sono unite covalentemente.

Durante la trascrizione di un gene codificante proteine in condizioni fisiologiche, il precursore dell'mRNA subisce uno splicing canonico, un processo altamente regolato, in cui gli introni sono rimossi e gli esoni vengono giunti per produrre un mRNA maturo funzionale.

Tuttavia, in determinate circostanze, il back-splicing del pre-mRNA può portare a una mescolanza di esoni. In questa particolare modalità di splicing l'estremità donatrice a valle del prodotto dello splicing si lega covalentemente al sito accettore a monte, dando origine a un circRNA.[10]

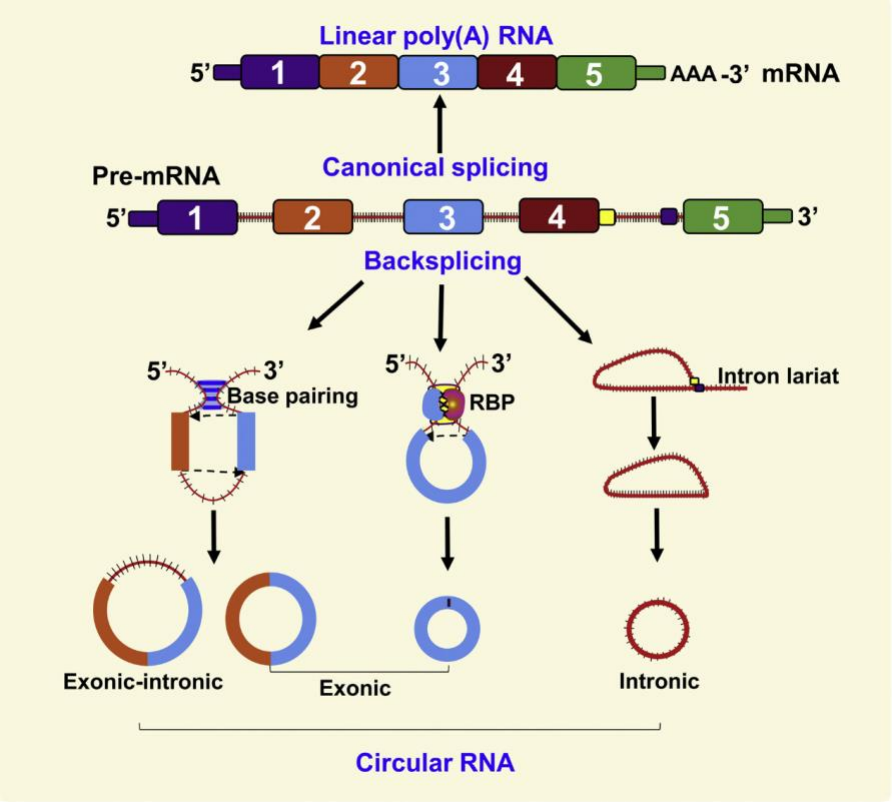


Figura 3: Biogenesi dei circRNA. I circRNA si formano da esoni, introni o da una combinazione di entrambi gli esoni-introni mediante eventi di back-splicing e l'intervento del macchinario di splicing[10]

Nell'immagine sopra (fig. 3) sono rappresentati i diversi meccanismi che guidano la formazione dei circRNA, tra cui il back-splicing convenzionale, la circolarizzazione attraverso l'accoppiamento degli introni e la circolarizzazione tramite lariat.

Il back splicing convenzionale prevede che la circolarizzazione sia mediata da sequenze complementari in cui la formazione di circRNA avviene attraverso l'appaiamento diretto tra sequenze complementari inverse nelle regioni di flanking[1] dell'RNA, facilitando la giunzione delle estremità e la creazione della struttura circolare tramite un legame fosfodiesterico 3',5'.[6]
Il secondo meccanismo evidenziato nella fig. 3 raffigura il processo di circolarizzazione guidato dalle proteine leganti l'RNA (RBPs), che coinvolge l'interazione tra gli introni di flanking mediata da queste proteine. Questa interazione contribuisce a stabilizzare la vicinanza tra il recettore di splicing e il donatore di splicing, agevolando così la formazione degli RNA circolari.[4,11]
L'interessante configurazione ad anello chiuso, nota come lariat, sorge dalla congiunzione covalente del sito donatore di splicing nell'esone a valle del precursore dell'mRNA con il sito accettore di splicing nell'esone a monte. Questo complesso processo di back splicing dà luogo alla formazione sia di circRNA contenenti esoni e introni, sia di circRNA esonici, ampliando la gamma di molecole circolari generabili.
In aggiunta a ciò, ulteriori eventi di back splicing possono condurre alla formazione di lariat circRNA intronici, attraverso il legame covalente degli estremi 3' e 5', mentre l'eliminazione dell'esone può portare alla comparsa di lariat circRNA misti.
Il processo di circolarizzazione guidato dai lariat coinvolge l'interazione di elementi regolatori cis- attivi, come le ripetizioni Alu, presenti sia negli esoni che negli introni vicini. Inoltre, l'intervento di fattori trans- attivi, come le proteine leganti l'RNA, contribuisce all'efficacia del processo di circolarizzazione.
L'azione di lariat intronici, che sfuggono alla consueta deramificazione e degradazione intronici, amplia ulteriormente le vie attraverso le quali si formano i circRNA, in questo caso infatti il lariat mantiene una forma circolare con un legame fosfodiesterico 3',5' tra il donatore di splicing e il punto di diramazione.[10]
I circRNA esonici-intronici (EIciRNAs) e gli intronici (ciRNAs) sono principalmente localizzati nel nucleo cellulare, suggerendo un coinvolgimento significativo nella regolazione dell'espressione genica a livello trascrizionale.
Al contrario, i circRNA esonici costituiscono la maggioranza dei circRNA presenti nel citoplasma, indicando la loro probabile partecipazione a processi regolatori a livello post-trascrizionale.

[1] Le regioni di flanking dell'RNA sono le sequenze di nucleotidi situate immediatamente ai lati di una particolare regione di interesse nell'RNA. Nel contesto dei circRNA e della circolarizzazione mediata da sequenze complementari, le regioni di flanking si riferiscono alle sequenze di nucleotidi che circondano l'area in cui si verifica il back-splicing.

Tutti gli esoni interni hanno segnali di splicing alle loro estremità 5' e 3', rendendo teoricamente possibile la circolarizzazione. Tuttavia, osserviamo solo un limitato sottoinsieme di eventi di backsplicing nelle cellule, principalmente a causa dell'efficienza estremamente bassa di queste reazioni.[4]

La biogenesi dei circRNA è stata proposta come competizione con il processo di splicing pre-mRNA ma diverse ricerche riportano risultati contrastanti perché le forma lineari e circolari dello stesso gene ospite hanno lo stesso modello di espressione e sono regolate in modo simile in diversi contesti cellulari.[12]

Caratteristiche e funzioni degli RNA circolari

Nel paragrafo precedente, è stato esaminato il processo di back splicing. È importante notare che gli RNA circolari possono derivare da diversi processi di back splicing e nonostante abbiano in generale bassi livelli di espressione, i circRNA presentano diversi profili di espressione tra tipi cellulari e tessuti nei mammiferi.

Circa il 75% del genoma umano può essere trascritto in RNA, tuttavia soltanto l'1,5% di questa trascrizione corrisponde a regioni codificanti.[13,14] Nel cervello umano, il 20% dei geni produce circRNA, mentre nel cuore solo il 9% dei geni espressi produce circRNA.[15]

L'abbondanza dei circRNA manifesta specificità cellulare, evidenziando una maggiore espressione nelle cellule a bassa proliferazione, come i cardiomiociti, in contrasto alle cellule ad alta proliferazione come quelle del fegato. L'osservazione di aumentati livelli di circRNA nei tessuti in via di sviluppo cardiaci, polmonari e cerebrali pare derivare principalmente da un processo di accumulo, ed inoltre, da vari studi, pare si verifichi un incremento di circRNA legato all'età.[16]

La prima volta che è stato osservato l'accumulo dei circRNA durante l'invecchiamento è stato durante il processo di invecchiamento in un'analisi di RNA-seq condotta su Drosophila melanogaster[17]. Studi successivi hanno confermato che questo fenomeno di accumulazione si verifica anche in altre specie animali, tra cui l'uomo.[18–21] È quindi plausibile che i circRNA possano essere utilizzati come biomarcatori dell'invecchiamento umano. [21]

La maggior parte dei circRNA si origina dai pre-mRNA che danno anche luogo a forme lineari di RNA, e i loro profili di espressione risultano in sintonia con quelli dei corrispondenti mRNA ospiti. Pur essendo abbondanti, i circRNA tendono ad essere espressi generalmente a livelli inferiori rispetto agli mRNA. Ciò nonostante, alcune ricerche hanno indicato che l'espressione di un RNA circolare può discostarsi dall'espressione del suo mRNA lineare corrispondente; in taluni contesti, i livelli di espressione dei circRNA possono risultare notevolmente superiori rispetto ai loro corrispettivi lineari.[15]

Dalle ricerche condotte, è stato osservato che durante la differenziazione e lo sviluppo neuronale, ad esempio, i circRNA manifestano un aumento dell'espressione e un notevole arricchimento nelle sinapsi.[22]

La mancata presenza di code di poliadenilazione 3' ha inizialmente ostacolato la rilevazione della maggioranza dei circRNA all'interno dei dataset convenzionali di sequenziamento dell'RNA, i quali prevalentemente contengono RNA poliadenilati. Tuttavia, recenti analisi del trascrittoma non poliadenilato e trascrittoma trattato con l'enzima RNasi R hanno svelato l'ampia diffusione dell'espressione dei circRNA in varie specie.

Le analisi più recenti, basate sull'impiego di algoritmi computazionali per l'identificazione dei siti di giunzione back-splice a partire da letture RNA-seq o attraverso la ricostruzione completa dei circRNA delle long *reads* RNA-seq, hanno portato all'individuazione di più di 100.000 circRNA nei trascrittomi umani.[6,23]

Poichè i circRNA sono anelli chiusi in modo covalente che non presentano le tipiche strutture all'estremità 5' e code di poli-A all'estremità 3', sono resistenti all'azione dell'enzima RNasi, che agisce in senso esonucleasico da 3' a 5' e degrada efficacemente quasi tutte le specie di RNA lineare. Questa caratteristica fa si che i circRNA siano incredibilmente stabili, con una durata di emivita media all'interno delle cellule superiore alle 48 ore, mentre gli RNA lineari durano mediamente solo 10 ore.

Questa caratteristica distintiva rende i circRNA un'opzione ideale come biomarcatori per diverse patologie. Tuttavia, i circRNA potrebbero comunque essere sensibili ad altre RNasi, come l'RNasi A, l'RNasi T1 e l'RNasi T2, che potrebbero costituire una via cruciale per la degradazione dei circRNA. La stabilità dei circRNA nel siero, ad esempio, è di circa 15 secondi, e la ragione potrebbe essere la presenza di endonucleasi circolanti.[24] Oltre alla loro struttura circolare, è possibile che altri fattori e meccanismi concorrano alla stabilità dei circRNA, sebbene tali elementi siano ancora ampiamente inesplorati.[11,25]

Numerosi circRNA mostrano specificità per determinati tessuti e periodi di sviluppo. In particolare, i circRNA sono elementi evolutivamente conservati attraverso diverse specie e sono presenti nella maggior parte degli organismi.

Il sequenziamento dell'RNA nei tessuti umani, sia adulti che fetali, ha rivelato che fino al 50% dei circRNA presenti manifesta una specificità tissutale, e sia la numerosità che i livelli espressione dei diversi circRNA nei tessuti fetali sono superiori rispetto a quelli riscontrati nei tessuti adulti. [26]

A lungo considerati sottoprodotti trascurabili della trascrizione, i circRNA sono ora riconosciuti come elementi chiave nella regolazione genica e nell'interazione con diversi componenti cellulari.

Una delle funzioni più studiate è la capacità dei circRNA di agire come “spugne” per i microRNA (miRNA), competendo per i siti di legame e riducendo così l’effetto delle attività regolatorie mediate da essi. [27]

I miRNA sono caratterizzati da una lunghezza compresa tra 18 e 25 nucleotidi, fanno parte del gruppo degli RNA non codificanti e giocano un ruolo cruciale nella regolazione di una diversificata gamma di funzioni biologiche fondamentali, tra cui la proliferazione cellulare, lo sviluppo, l'apoptosi e l'insorgenza di patologie. Le molecole di miRNA hanno la capacità di silenziare un gene attraverso il riconoscimento di sequenze complementari, e tramite l’aiuto del complesso di proteine (Argonauta) che fanno parte del RNA-induced silencing complex (RISC). I meccanismi sottostanti la regolazione dei miRNA risultano ancora ampiamente incompleti ma sembra che gli RNA circolari contribuiscano significativamente alla rete di RNA endogeni competitivi.[16]

Una grande quantità di studi ha confermato la relazione tra circRNA e miRNA, tuttavia, i circRNA solitamente assorbono più di un miRNA per esercitare la loro funzione. Pertanto, per gli RNA circolari con un singolo miRNA target o un unico sito di assorbimento, questa funzione rimane controversa.[28] L’ipotesi dell’assorbimento circRNA-miRNA deve essere analizzata con un approccio critico; infatti, la presenza di specifici circRNA è generalmente scarsa. Questo è in contrasto con le possibili teorie di assorbimento, perché il miRNA potrebbe essere più abbondante rispetto all’RNA totale con cui il miRNA può legarsi e competere. Tuttavia, è ancora possibile che alcuni circRNA possano agire cataliticamente, cioè mobilitando, inattivando e/o degradando i miRNA.[29]

Uno degli esempi più significativi della relazione tra RNA circolari e microRNA è CDR1as (fig.4), un circRNA identificato come regolatore di processi cellulari contenente oltre 70 siti di legame convenzionali per miR-7 e abbondantemente espresso nel cervello dei mammiferi. CDR1as agisce come regolatore negativo di miR-7, determinando una variazione dell’espressione di numerosi geni chiave. La ridotta espressione di CDR1as nelle linee cellulari umane si è tradotta in una diminuzione dei livelli di mRNA che presentano siti di legame per miR-7. Inoltre, l’interazione tra CDR1as e miR-7 è influenzata anche da un lungo RNA non codificante chiamato Cyrano. Cyrano lega il miR-7 e favorisce la degradazione mirata dei miRNA bersaglio, attivata attraverso il taglio dell’estremità 3’ di miR-7. Questo meccanismo contribuisce alla regolazione dell'asse Cdr1as-miR-7.[6,30]

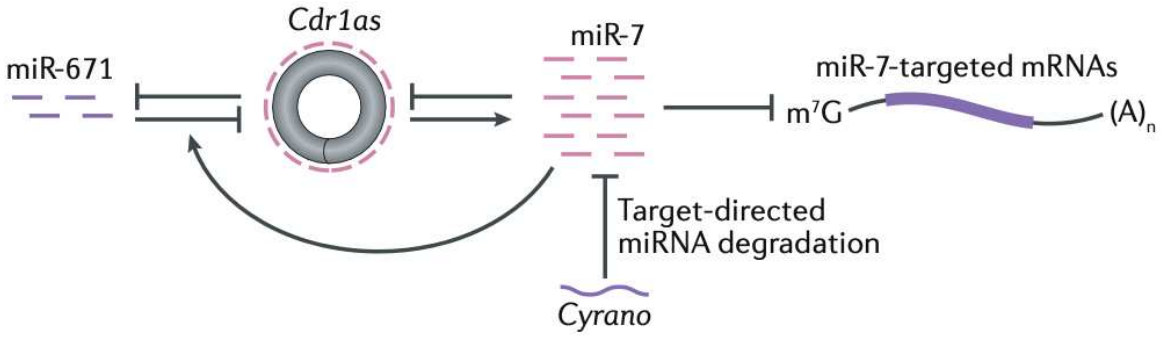


Figura 4: asse Cdr1as-miR-7[6]

Alla luce della generalmente bassa abbondanza dei circRNA e del numero limitato di siti di legame, non è ancora chiaro come i circRNA esercitino effetti biologici e ne rende difficile anche la rilevazione. L'analisi bioinformatica ha infatti dimostrato che non tutti i circRNA contengono un numero sufficiente di siti di legame per i miRNA e la funzione tipica di "spugna" di miRNA sembra essere unica solo per alcuni circRNA.[10]

La maggior parte degli RNA circolari caratterizzati dimostra una predominante localizzazione nel citoplasma; tuttavia, stanno emergendo esempi di RNA circolari che svolgono funzioni nel nucleo. Gli RNA circolari esone-introne sono sottosposti a splicing incompleto e hanno un introne trattenuto che consente loro di interagire con alcuni snRNA[2] promuovendo la trascrizione del loro gene ospite.[4]

La seconda funzione più importante attribuita agli RNA circolari è esercitata tramite interazioni circRNA-proteina. Le proteine che maggiormente interagiscono con le molecole di RNA sono conosciute come RNA-binding protein (RBPs). Le RBPs costituiscono una categoria di proteine coinvolte nel processo metabolico degli RNA, regolando la loro maturazione, il trasporto, la localizzazione e la traduzione. Dagli studi è emerso che molti circRNA interagiscono con le RBPs attraverso siti di legame specifici. È importante notare che gli effetti delle RBPs sulla regolazione della formazione dei circRNA variano a seconda dei tipi di circRNA, del tessuto o della cellula e delle circostanze biologiche. La presenza di una singola proteina legante l'RNA può avere effetti duali sui circRNA, poiché differenti elementi di legame che circondano i circRNA possono interagire con diversi domini funzionali della RBP. Inoltre, l'espressione delle RBPs è altamente specifica sia in termini di localizzazione nello spazio che di momento temporale, contribuendo così alla variazione dell'espressione dei circRNA in differenti tipi cellulari e sotto varie circostanze patofisiologiche. Di conseguenza, le RBPs possono agire sia come attivatori che come

[2] snRNA è l'acronimo di small nuclear RNA. Si tratta di piccole molecole di RNA nucleare, tendenzialmente ricche in uridina, e con una funzione fondamentale nel processo di splicing. Il ruolo degli snRNA nello splicing si esercita attraverso la formazione di complessi RNA-RNA o RNA-proteine e questi complessi interagiscono base contro base con l'mRNA che deve essere tagliato.

inibitori nella formazione dei circRNA, determinando regolazioni differenziate dei livelli di espressione dei circRNA tramite molteplici meccanismi.
Ricerche recenti hanno inoltre rivelato che le interazioni tra RNA e RBPs sono notevolmente influenzate dalla struttura terziaria delle molecole di RNA. Pertanto, la particolare struttura terziaria dei circRNA potrebbe influenzare la loro capacità di legare le proteine.
È stato dimostrato che le interazioni RNA-proteina influenzano l'espressione e la funzione delle proteine, oltre a regolare la sintesi e la degradazione dei circRNA stessi.
Ci sono tre modi principali tramite il quale i circRNA interagiscono con le proteine: possono stabilizzare l'interazione tra due proteine, migliorando l'affinità o l'adesione reciproca; possono legarsi ad una proteina A e influenzare l'interazione con un'altra proteina B che non è direttamente legata al circRNA; oppure possono legarsi ad entrambe le proteine che normalmente interagiscono, causandone la dissociazione. In tutti e tre i casi si forma un complesso ternario, ma gli effetti risultanti sono differenti.

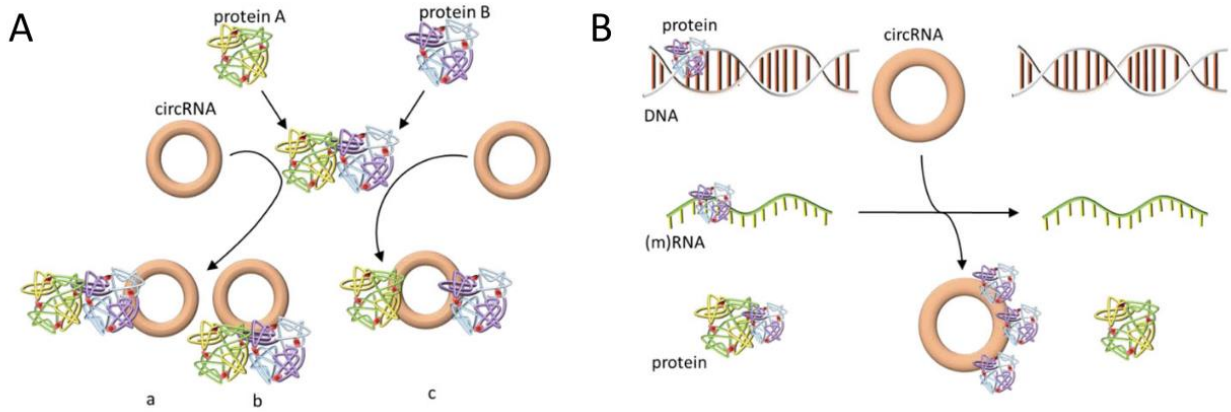


Figura 5: interazioni circRNA-proteina[31]

Nella fig. 5 sono rappresentati alcuni processi di interazione circRNA-proteina. In particolare, nella figura a sinistra (A) possiamo vedere 3 processi differenti di interazione: nella prima rappresentazione a sinistra (a) il circRNA si lega ad entrambe le proteine, e questa interazione sinergica rafforza il legame tra le proteine stesse. Il circRNA in questo caso agisce da mediatore, favorendo una comunicazione più efficiente e intensa tra le proteine coinvolte, che possono così coordinare e potenziare le loro funzioni biologiche.
Nella seconda modalità (b), il circRNA stabilisce un legame specifico con una proteina A, e questo legame a sua volta amplifica l'interazione tra la proteina A e un'altra proteina B che normalmente non avrebbe un legame diretto con l'RNA circolare. In questo modo il circRNA agisce da "ponte molecolare" facilitando l'interazione tra proteine che altrimenti potrebbero avere un coinvolgimento limitato. La terza modalità (c) vede il circRNA interagire con due

proteine che, in precedenza, interagivano tra loro. Tuttavia, una volta che il circRNA si lega ad entrambe le proteine, interrompe questa interazione precedentemente stabilita.
Nell'immagine in alto a destra (B) viene invece rappresentato il circRNA che agisce come un blocco che impedisce alle proteine di interagire con il DNA, l'RNA o altre proteine. In questo modo, i circRNA limitano l'accesso delle proteine a specifiche molecole biologiche, compromettendo le loro funzioni originali.

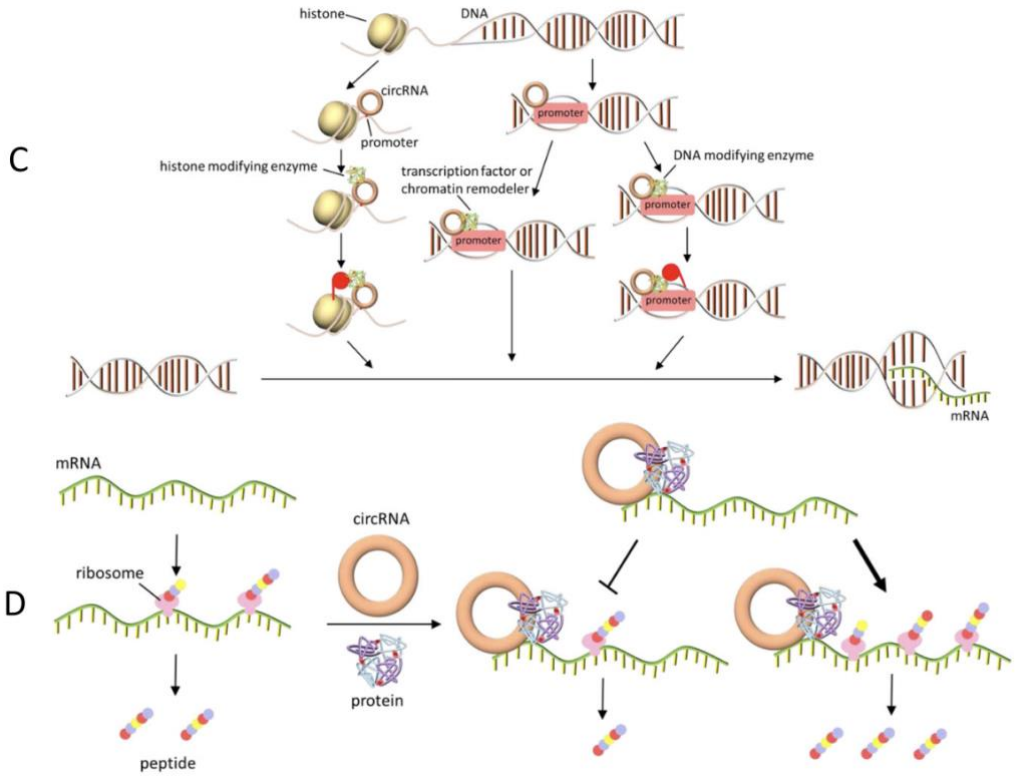


Figura 6: interazioni circRNA-proteina[31]

Altre due tipologie di interazioni possibili tra gli RNA circolari e le proteine sono raffigurate nella figura 6. La prima in alto (C) mostra come i circRNA possono reclutare i fattori di trascrizione, i rimodellatori della cromatina e gli enzimi che modificano il DNA. Questo processo complesso conduce a modificazione dell'attività trascrizionale, che può comprendere sia attivazioni che inibizioni. L'immagine D invece mostra come i circRNA possono aiutare le RBP a combinarsi con l'mRNA e stabilizzare l'mRNA (promuovendo indirettamente la traduzione) o regolare direttamente la traduzione.[24,25,31]

Ruolo degli RNA circolari nelle varie condizioni patologiche

I circRNAs sono stati collegati ad una serie di malattie, svolgendo ruoli chiave in processi patologici. Ad esempio, si è scoperto che alcuni circRNA sono sovraespressi o sottoespressi in diverse neoplasie e questa deregolazione dei livelli di circRNA è stata associata alla

promozione della proliferazione cellulare, alla soppressione dell'apoptosi, all'invasione e alla metastasi tumorale.[32]

Inoltre, alcuni circRNA sono implicati nelle malattie neurodegenerative, come l'Alzheimer e il morbo di Parkinson. Questi circRNAs possono interagire con proteine coinvolte nella formazione di placche amiloidi, contribuendo così alla progressione di queste patologie.[33]

Oltre alle malattie oncologiche e neurodegenerative, i circRNA sono stati collegati anche ad altre condizioni patologiche, come le malattie cardiovascolari[34], le malattie infiammatorie e le malattie metaboliche.[35] Ad esempio, alcuni studi hanno suggerito che circRNA possono influenzare la risposta infiammatoria delle cellule e contribuire alla regolazione della funzione delle cellule del sistema immunitario.[11]

La ricerca sulla relazione tra i circRNA e il cancro ha riscontrato che spesso i circRNA sono deregolati nei tumori, il che consente di distinguere il tessuto tumorale dal tessuto normale adiacente. Sono stati identificati diversi circRNA funzionali associati al cancro, con espressione differenziale, in diversi tipi di tumori. Questi circRNA agiscono come soppressori o promotori tumorali e influenzano i fenotipi cancerosi in modi diversi.[36,37]

Studi recenti hanno documentato un incremento frequente dei livelli di RNA circolari quando le cellule si differenziano e smettono di proliferare ed è stato dimostrato che i circRNA possono influenzare la patogenesi e la progressione del cancro (fig.7) regolando geni bersaglio agendo, ad esempio, come spugne per miRNA.[25]

L'interazione tra circRNA e miRNA riveste un ruolo chiave nella comprensione dei meccanismi di sviluppo dei tumori.[32]

Le traslocazioni cromosomiche associate al cancro possono dare luogo alla produzione anomala di RNA circolari di fusione, dove sequenze introniche adiacenti al punto di rottura si appaiano reciprocamente. Questi RNA circolari di fusione possono esercitare un'azione favorevole sulla trasformazione cellulare, sulla sopravvivenza e sulla capacità di resistere alle terapie, fornendo così nuovi e promettenti bersagli terapeutici. [4]

Sono stati trovati legami tra gli RNA circolari e vari tipologie di tumori, tra cui il tumore al pancreas, il tumore della cervice uterina e il tumore al cervello.

Ad esempio, il silenziamento di circ_0030235 tramite un siRNA[3] sopprime principalmente la proliferazione, la migrazione e/o l'invasione del cancro al pancreas.[11] L'espressione di Hsa_circ_002059 è significativamente più bassa nel cancro gastrico.[32] Nel caso del tumore della cervice uterina studi recenti hanno dimostrato che 45 diversi circRNA erano espressi

[3] I siRNA, ovvero small interfering RNA, è una classe di molecole di RNA a doppio filamento, lunghe tra i 19 e i 21 nucleotidi. I siRNA sono coinvolti nell'interferenza dell'espressione di specifici geni con sequenze nucleotidiche complementari, degradando l'mRNA dopo la trascrizione, per non far avvenire la traduzione

maggiormente nei tessuti tumorali rispetto a quelli sani nel controllo. Tra questi l'eliminazione del circRNA maggiormente espresso comportava un'inibizione della proliferazione e dell'invasione delle cellule tumorali.[27,38] Nel tumore al cervello, differenti studi[39–41] hanno dimostrato che oltre a promuovere l'angiogenesi e la proliferazione delle cellule di glioma, diversi circRNA sembrano avere un potenziale ruolo regolatorio nella vitalità, migrazione e invasione delle cellule tumorali.

Ad esempio nello studio "Circular RNA circBCBM1 promotes breast cancer brain metastasis by modulating miR-125a/BRD4 axis"[41] è stato dimostrato che funzionalmente, circBCBM1 promuove la proliferazione e la migrazione delle cellule 231-BR[4] in vitro e la crescita e la metastasi cerebrale in vivo.

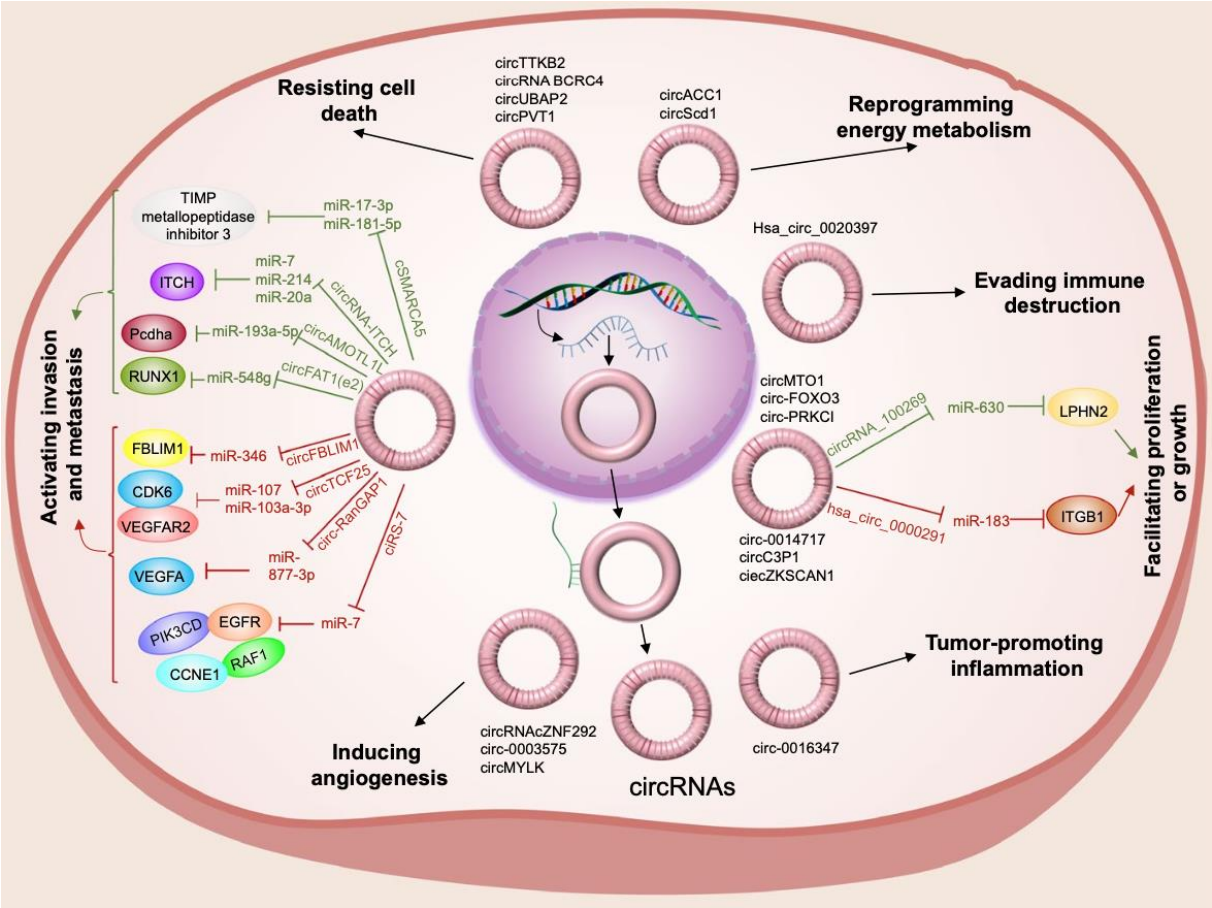


Figura 7:Illustrazione del Circular RNA nel cancro. I Circular RNA (circRNA) sono stati identificati (principalmente nel citoplasma) come attori chiave in vari aspetti della progressione tumorale. Ad esempio, alcuni circRNA potrebbero contribuire alla proliferazione, alla crescita, all'invasione e alla metastasi. I circRNA soppressori dei tumori sono rappresentati da linee verdi, mentre i circRNA promotori dei tumori sono indicati da linee rosse.[36]

La presenza stabile degli RNA circolari negli esosomi[42] e nel plasma rappresenta una modalità più agevole per l'individuazione diagnostica dei fenomeni oncologici. [43]

[4] Le cellule 231-BR fanno parte della linea cellulare metastatica al cervello del cancro al seno, questa linea cellulare deriva dalle cellule MDA-MB-231 che sono conosciute per essere altamente metastatiche e vengono utilizzte come modello per studiare le metastasi del cancro al seno in laboratorio[117]

In ambito clinico, i marcatori molecolari attualmente utilizzati sono spesso costituiti da proteine che hanno una scarsa specificità, quindi potrebbero non essere molto utili per distinguere con precisione la fonte o l'origine di una condizione patologica; l'utilizzo degli RNA circolari come potenziali biomarcatori per l'identificazione dei tumori potrebbe contribuire ad affrontare la limitata specificità dei marcatori esistenti, grazie alla loro espressione mirata e differenziata.[44,45]

Tuttavia, va sottolineato che l'impiego degli RNA circolari nell'ambito diagnostico non è esente da alcune criticità. In primo luogo, alcuni RNA circolari richiedono prelievi di campioni tissutali dai pazienti per la diagnosi, comportando un certo livello di invasività. In secondo luogo, la rilevazione degli RNA circolari nei tessuti o negli esosomi risulta essere un processo più oneroso rispetto alle metodologie diagnostiche esistenti, aspetto che potrebbe limitarne l'ampia adozione come biomarcatori. In terzo luogo, l'affidabilità dell'impiego degli RNA circolari per la diagnosi necessita ancora di ulteriori validazioni.[23]

Come accennato in precedenza, i circRNA manifestano un notevole arricchimento all'interno dei tessuti neuronali.

Ciò trova fondamento nell'osservazione che molti dei geni che danno origine ai circRNA sono esclusivamente espressi nel tessuto cerebrale. Inoltre, è stato riscontrato che nei casi in cui il gene ospitante sia espresso anche in altri tessuti, nel cervello la percentuale di trascritti che generano circRNA è notevolmente maggiore. Questo suggerisce l'esistenza di una regolazione specifica dei circRNA nell'ambito della produzione neuronale.[46,47]

D'altra parte, i geni che ospitano i circRNA risultano essere arricchiti in termini di processi biologici strettamente legati alla funzione sinaptica, come lo sviluppo del sistema nervoso, la neurogenesi e la differenziazione neuronale. Questi geni includono componenti cellulari di rilevanza sinaptica, come la sinapsi stessa, la zona attiva presinaptica, la membrana presinaptica e la densità postsinaptica.

Nel corso del confronto tra differenti regioni cerebrali, la maggioranza dei circRNA dimostra un profilo di espressione che concorda con i livelli di espressione dei geni ospiti lineari. Tuttavia, emergono casi di circRNA specifici che presentano un'ampia espressione in specifiche aree cerebrali, indipendentemente dai livelli di espressione dei corrispondenti geni lineari.

Emerge, inoltre, una marcata abbondanza di circRNA all'interno del cervelletto, una regione cerebrale molto densa di spine neuronali, sinapsi e neuroni in generale.[10,47]

Diversi studi hanno collegato i circRNA alla degenerazione e alle malattie cerebrali e l'identificazione dell'espressione specifica di alcuni RNA circolari nel cervello li eleva al ruolo di potenziali biomarcatori per le malattie neurodegenerative.

Ad esempio, sono stati individuati diversi circRNA con espressione differenziale nei tessuti cerebrali e nel plasma dei pazienti affetti da malattia di Alzheimer (AD). È stato dimostrato, ad esempio, che hsa_circRNA_001481 e hsa_circRNA_000479 sono significativamente sovraespressi nei campioni di sangue in diverse fasi dei pazienti affetti da AD. I profili di espressione dei circRNA potrebbero contribuire ad una più approfondita comprensione dei dettagli molecolari e dei possibili meccanismi sottostanti per lo sviluppo di indicatori disgnostici e approcci terapeutici.[48]

Ulteriori ricerche hanno individuato circRNA associati a malattie neurodegenerative (fig.8) come l'atrofia multisistemica, e sono stati segnalati studi che implicano i circRNA nella malattia di Parkinson e nella sclerosi laterale amiotrofica (SLA).[49,50] Questi circRNA sono coinvolti in rilevanti vie di trasduzione del segnale che regolano diverse attività neurali.

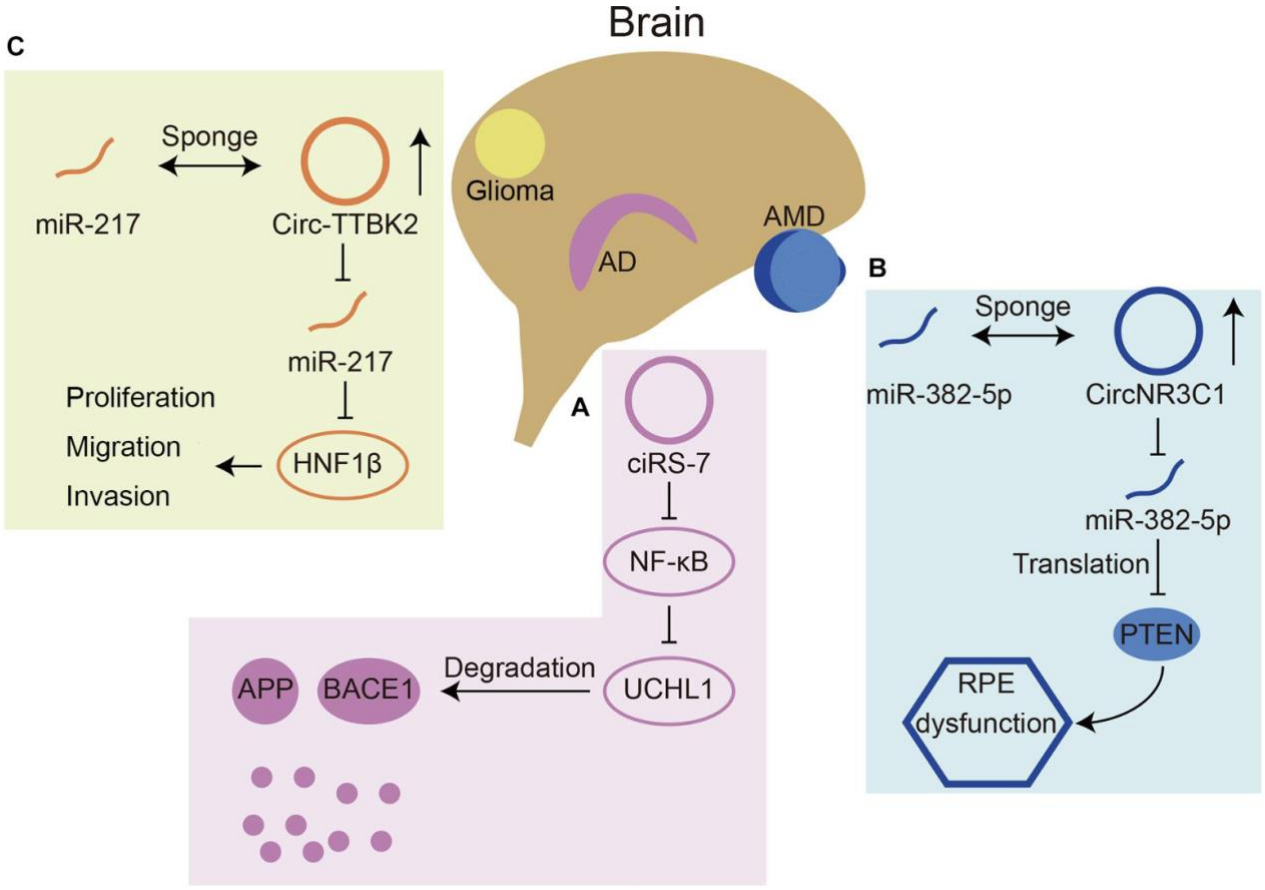


Figura 8: Esempio di 3 RNA circolari associati a malattie neurologiche: (A) il circular RNA CiRS-7 è in grado di rallentare la produzione della proteina NF-kB. Questo porta a un aumento della quantità di UCHL1, la quale favorisce la riduzione delle proteine APP e BACE1, che sono coinvolte nella malattia di Alzheimer. (B) L'RNA circolare CircNR3C1 agisce come una sorta di spugna per una piccola molecola chiamata miR-382-5p. Questa interazione impedisce che miR-382-5p inibisca la produzione di PTEN, una proteina legata alla degenerazione maculare. (C) L'RNA circolare Circ-TTBK2 può agire come spugna per la molecola miR-217, impedendo a quest'ultima di inibire la produzione di una proteina chiamata HNF1β che può promuovere la crescita aggressiva dei tumori cerebrali chiamati gliomi.[50]

Molti studi hanno dimostrato che le lesioni acute del sistema nervoso centrale (SNC) comportano significative alterazioni nei profili di espressione dei circRNA.[51] Ad esempio, l'ipossia-ischemia neonatale (HI) è un problema frequente che si verifica quando il flusso di sangue al feto viene interrotto a partire dalla settimana gestazionale 36 o in seguito. Questo

evento dà origine all'encefalopatia ipossico-ischemica, una condizione caratterizzata da disabilità motorie, sensoriali e cognitive a lungo termine. Uno studio recente ha dimostrato che in un modello di ratto affetto da HI, si verifica una significativa alterazione dell'espressione di molti circRNA. L'analisi bioinformatica delle reti circRNA/mRNA ha suggerito che questi RNA circolari potrebbero essere coinvolti sia nel danno cerebrale che nella degenerazione neuronale.[52] La significativa abbondanza dei circRNA nel cervello e negli esosomi potrebbe renderli preziosi biomarcatori per i disturbi del sistema nervoso centrale (CNS).[47]

Per quanto riguarda il sistema immunitario, recenti studi hanno evidenziato un ruolo significativo e una varietà di funzioni biologiche dei circRNA nelle infezioni virali, nonché nella regolazione delle risposte immunitarie innate.[53]

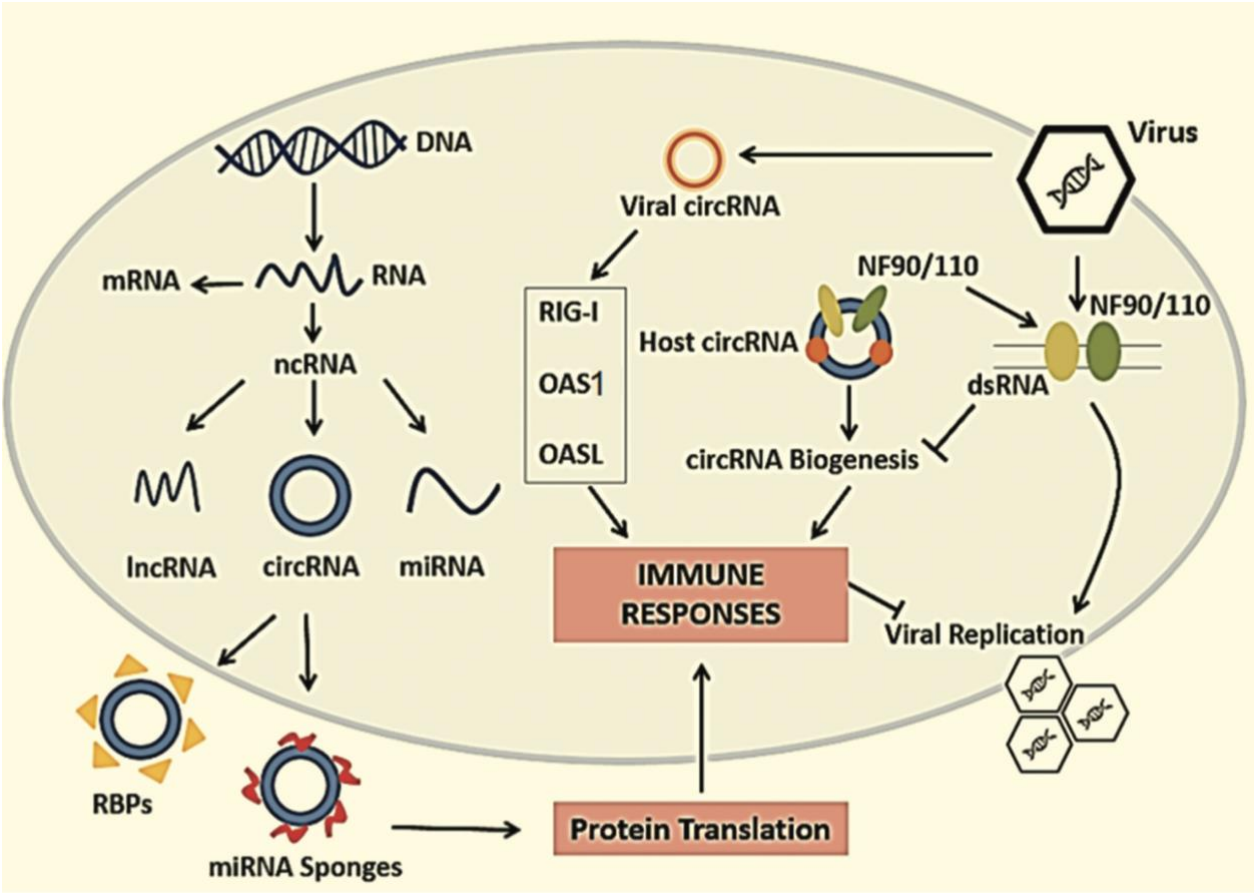


Figura 9: illustrazione schematica del significato funzionale dei circRNA nelle risposte immunitarie antivirali[54]

L'associazione tra i circRNA endogeni e le risposte dell'immunità innata è emersa attraverso una ricerca dei fattori di trascrizione che possono influenzare la produzione di circRNA e ad oggi sono note più di 100 proteine che possono specificamente modulare il back-splicing.

Nel sistema immunitario innato antivirale, le proteine che legano il dsRNA svolgono un ruolo chiave ed essenziale inducendo varie modifiche nei processi dell'RNA cellulare e virale per reprimere la replicazione virale. I circRNA esogeni possono innescare una risposta

immunitaria innata che conferisce protezione contro l'infezione virale. È stato inoltre riportato che i circRNA inducono in modo potente l'espressione di diversi geni regolatori del sistema immunitario innato. Dai vari studi è emerso che la produzione di circRNA deve in qualche modo essere limitata dalle cellule ospiti in caso di infezione virale, mentre i livelli aumentati di circRNA possono facilitare la propagazione virale.[55] Sono stati osservati circRNA derivati dai virus e/o espressi in modo diverso in seguito a varie infezioni virali.[56] Il Citomegalovirus (CMV), ad esempio, può influenzare l'espressione dei circRNA nelle cellule ospiti; inoltre i circRNA potrebbero essere coinvolti nella regolazione dell'infiammazione e nella risposta immunitaria dell'ospite all'infezione da CMV.[57,58]
Un altro esempio è il virus Zika è un flavivirus trasmesso principalmente dalle zanzare e la cui infezione durante la gravidanza è associata a gravi difetti congeniti. Anceh per questo virus è emerso che durante l'infezione si modifica l'espressione dei circRNA nell'encefalo dei feti umani. Alcuni di questi circRNA mostrano livelli di espressione alterati in risposta all'infezione e sono stati collegati alla regolazione dei geni coinvolti nello sviluppo cerebrale.[59]

Protocollo di sequencing degli RNA circolari

Nelle prime analisi di sequenziamento dell'RNA i circRNA sono rimasti inosservati. Ciò è principalmente attribuibile a due motivi: l'ampio impego di passaggi di selezione della catena poli(A) nei protocolli di preparazione delle librerie, che hanno portato alla perdita degli RNA circolari e di altri trascritti privi di coda di poli(A), e l'utilizzo di algoritmi computazionali che richiedevano che le letture dell'RNA-seq si allineassero linearmente al genoma, scartando quindi tutte le letture corrispondenti alle giunzioni backspliced.
È stato solo quando Salzman e i suoi colleghi hanno utilizzato l'RNA-seq per identificare trascritti specifici, associati a riarrangiamenti cromosomici in cellule cancerose, che hanno casualmente scoperto la presenza di migliaia di RNA circolari nelle cellule.[60] La maggior parte di questi trascritti è stata poi sorprendentemente riscontrata anche nelle cellule normali, suggerendo che questi RNA insoliti non erano causati da riarrangiamenti strutturali del DNA genomico, ma piuttosto da processi di splicing attivi in tutte le cellule.
I circRNA sono stati inoltre arricchiti nelle frazioni prive di poli(A) e risultavano resistenti all'azione dell'RNase R, un enzima che degrada quasi tutti gli RNA lineari.[4] È quindi possibile ottenere delle librerie arricchite per i circRNA mediante raccolte di RNA non poliadenilati e mediante trattamento preliminare con l'enzima RNasi R che riduce la rilevazione di falsi positivi.[24,29] Tuttavia, l'eliminazione degli RNA lineari rende complesso stabilire se le variazioni nell'espressione dei circRNA sono autonome rispetto ai loro geni ospiti lineari.[61]

L’esonucleasi RNasi R 3’-5’ ha la capacità di degradare preferenzialmente gli RNA lineari, mantenendo per lo più intatti i circRNA. Questa efficacia può essere compromessa quando ci si trova di fronte a RNA altamente strutturati, come quelli che formano G-quadruplex, per cui la resistenza all'esonucleasi non può stabilire inequivocabilmente la circolarità. Per superare questa sfida, è possibile aggiungere ioni Li+ al buffer di reazione dell'RNase R, poiché questo aiuta a destabilizzare la struttura secondaria dell'RNA e a migliorare l'eliminazione dell'RNA lineare. Oltre al rischio di falsi positivi dovuti alla presenza di RNA lineari non degradati, possono verificarsi falsi negativi se i circRNA vengono danneggiati durante la purificazione. Questo sembra essere particolarmente rilevante per circRNA più grandi e durante tempi di incubazione prolungati. Un altro approccio per aumentare la quantità di circRNA in un campione è eseguire una controselezione delle poli(A) utilizzando primer oligo(dT) immobilizzati, che consentono di eliminare gli mRNA lineari poliadenilati dal campione di interesse. Tuttavia, è importante notare che questo metodo può influenzare i circRNA che contengono tratti A interni, come nel caso di CDR1as32.[61]

Un approccio alternativo, in cui è possibile evitare l’utilizzo della RNasi R, prevede l’analisi di cattura su gel, in cui i circRNA vengono intrappolati nei pozzetti di un gel di agarosio e possono essere estratti, purificati e sottoposti a sequenziamento per confermare la loro presenza. Inoltre, l’uso dell’elettroforesi bidimensionale su gel di poliacrilammide non denaturante può separare i circRNA dalle molecole lineari in base alla loro migrazione elettroforetica più lenta.

Un ulteriore metodo alternativo prevede l’utilizzo dell’RNasi H e di una sonda di DNA complementare a una regione specifica del circRNA. Una volta che la sonda di DNA si lega al circRNA, l’enzima RNasi H può intervenire causando una singola rottura nella regione legata. Questo evento porta alla linearizzazione del circRNA che può quindi essere visualizzato come una singola banda su un gel di agarosio. D’altro canto, se il circRNA è assente e al suo posto c’è RNA lineare, possono formarsi diverse bande nel gel di agarosio. Questo metodo sfrutta quindi la capacità dell’RNase H di degradare l’ibrido RNA-DNA e fornisce un modo per distinguere tra circRNA e RNA lineare in base alla presenza di una singola o doppia banda di elettroforesi. [24]

L’analisi Nothern blot con sonde lunghe che coprono l’intero circRNA o con sonde corte che fiancheggiano i siti di giunzione di splice permette quindi di individuare i circRNA in quanto un circRNA si sposta con una velocità inferiore rispetto ad un RNA lineare della stessa lunghezza. Questo metodo è però limitato dalla bassa sensibilità, dalla bassa capacità di throughput e da fasi molto lunghe che richiedono molto tempo.[62]

L’utilizzo della fluorescenza durante l’ibridazione dell’RNA, abbinata alla microscopia ad alta risoluzione che utilizza sonde progettate per affiancare i siti di giunzione, è un metodo efficace per individuare la distribuzione e l’abbondanza dei circRNA.[10]

Diversi gruppi di ricerca hanno lavorato per potenziare la sensibilità e abbreviare il tempo necessario per le analisi Northern Blot sui circRNA. A titolo d'esempio, Xiaolin Wang e Ge Shan hanno introdotto l'utilizzo del sistema di marcatura con digossigenina (DIG) per la rilevazione dei circRNA. Questo approccio offre diversi vantaggi, tra cui una sensibilità elevata, tempi di esposizione più brevi, una maggiore durata e una sicurezza superiore rispetto al tradizionale sistema di marcatura isotopica. Altre strategie di progettazione delle sonde, come l'impiego di sonde oligonucleotidiche con acidi nucleici a blocco (LNA) al posto delle sonde tradizionali a base di DNA, hanno dimostrato un'efficienza almeno dieci volte superiore per la rilevazione degli RNA più piccoli.[62]

La PCR con primer convergenti e divergenti è un altro metodo ampiamente utilizzato per l’identificazione dei circRNA. Problemi come la bassa abbondanza possono però impedire l’accuratezza della quantificazione dei circRNA. Questi ostacoli possono essere superati con la PCR ad alto rendimento, ad esempio tramite la Droplet Digital PCR (ddPCR), durante la quale il campione viene suddiviso in microgoccioline indipendenti. L’amplificazione della PCR avviene in ogni gocciolina e i compartimenti che finiscono per contenere la molecola bersaglio completano la PCR e si leggono come ‘positivi’ mentre quelli che non contengono DNA si registrano come negativi. La misura dei livelli assoluti di acidi nucleici a bassa abbondanza può essere fatta in modo preciso utilizzando il rapporto tra positivi e negativi.[10]

La somiglianza tra i circRNA e i corrispettivi RNA lineari rende cruciale la convalida sperimentale dell’espressione dei circRNA. Tra i metodi più popolari ci sono la PCR quantitativa con trascrizione inversa (RT-qPCR) e la RT-PCR seguita dall’elettroforesi su gel o il sequenziamento di Sanger. L'analisi tramite RT-qPCR rappresenta una metodologia che richiede quantità minime di RNA iniziale, permettendo la determinazione del rapporto tra circRNA e RNA lineare. In alternativa, l'approccio della RT-PCR semiquantitativa può essere impiegato per amplificare direttamente sia le isoforme circolari che lineari in modo indipendente all'interno di una singola reazione, come comunemente utilizzato nell'ambito dello studio dello splicing alternativo. Questo metodo richiede la progettazione di tre primer: una coppia di primer per amplificare la giunzione di back-splicing (BSJ) e un ulteriore primer situato in uno degli esoni adiacenti, il quale servirà per amplificare la forma lineare.[61]

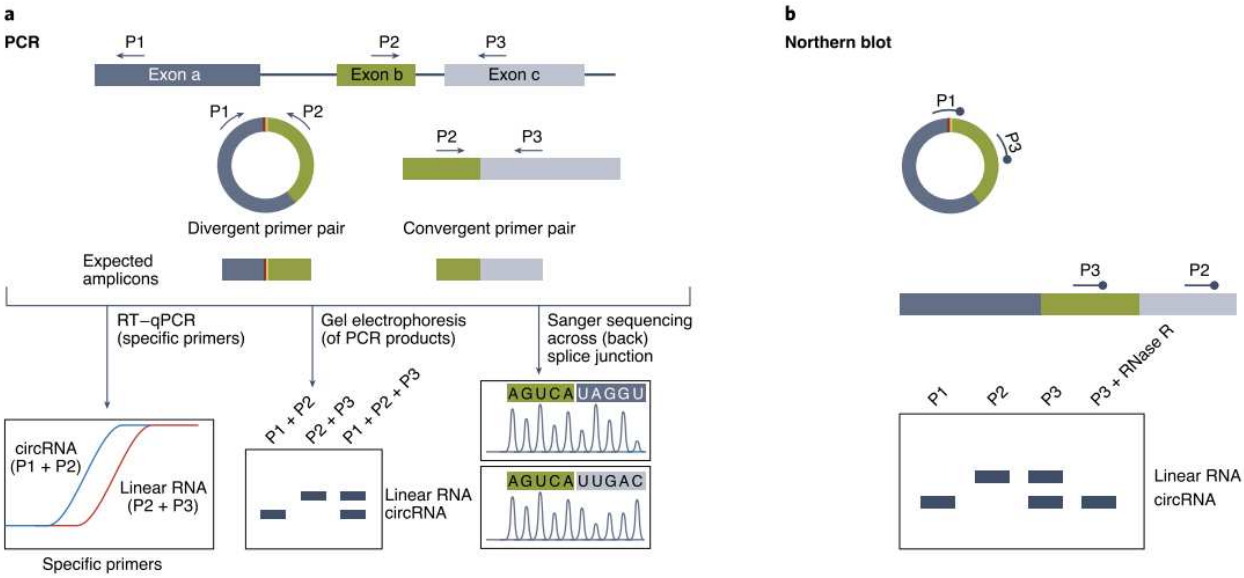


Figura 10: metodi di detection dei circRNA[61]

Rilevazione e quantificazione dei circRNA tramite sequenziamento

Molti degli studi più recenti per il rilevamento e la quantificazione dei circRNA sono basati sul sequenziamento, in particolare grazie a piattaforme di next generation sequencing (NGS). Il NSG offre notevoli vantaggi rispetto ai metodi di sequenziamento tradizionali, tra cui alto throughput, elevata sensibilità, rapido tempo di esecuzione e costi inferiori.

Ci sono metodi di sequenziamento basati su corte lunghezze di lettura e altri che invece producono letture anche molto lunghe.

Il sequenziamento delle *reads* corte viene eseguito solitamente per sintesi o ligazione. Ciascuna strategia utilizza rispettivamente enzimi DNA polimerasi o ligasi per estendere numerosi filamenti di DNA in parallelo. I nucleotidi possono essere forniti uno alla volta o possono essere modificati con tag identificativi. Il sequenziamento di un genoma altamente complesso e ripetitivo, come quello umano, può essere difficoltoso utilizzando queste tecnologie proprio a casua della ridotta lunghezza delle letture che crea problemi in fase di assemblaggio delle sequenze.

Uno dei metodi di sequenziamento *short reads* maggiormente conosciuto è Illumina, un metodo ad alta capacità di sequenziamento, che consente di generare grandi quantità di dati in un breve periodo di tempo.

Le tecnologie di sequenziamento *long-reads* sono in grado di leggere porzioni di DNA comprese tra 5000 e 30000 paia di basi sequenziando una singola molecola. Questo permette di eliminare il bias di amplificazione e di identificare la sovrapposizione tra le letture, consentendo un migliore assemblaggio delle sequenze.

Una delle tecnologie di sequenziamento *long-reads* è il nanopore-sequencing. Il sequenziamento a nano-pori migliora l'efficienza di rilevamento delle letture circolari (con un aumento di 20 volte) e fornisce un aumento di cinque volte nell'identificazione di eventi di circolarizzazione alternativi rispetto ad Illumina, indicando la sua maggiore sensibilità per l'identificazione di isoforme circolari. Inoltre, il sequenziamento con nanopori è in grado di riconoscere circRNA con un'abbondanza relativamente bassa e di fornire informazioni non solo sulla giunzione di splicing, ma anche relative all'intera sequenza di un circRNA.[28]
Una limitazione delle tecnologie di sequenziamento *long-reads* è la minore accuratezza delle singole letture rispetto alle tecnologie di sequenziamento *short-reads*. Un ulteriore problema riguarda il tempo necessario per l'elaborazione dei dati, che aumenta significativamente con la dimensione del genoma dell'organismo in esame.
CIRI-long è un innovativo algoritmo progettato per condurre un'analisi dei circRNA tramite l'utilizzo della tecnologia di sequenziamento a nanopori. Questa tecnologia ha utilizzato la trascrizione circolare inversa per amplificare i circRNA producendo lunghe molecole di DNA complementare.[63,64] Questo sistema di amplificazione utilizza un solo innesco di DNA complementare e una sonda molecolare per la rilevazione del segnale. L'innesco di DNA si lega al sito di giunzione sulla circRNA bersaglio, avviando il processo di amplificazione, e successivamente la sonda molecolare si lega a un sito sul filamento di DNA di amplificazione inversa, con conseguente emissione del segnale fluorescente. A differenza della tradizionale qRT-PCR, questo metodo offre vantaggi significativi, tra cui la semplicità operativa, il costo ridotto e l'assenza della necessità di cicli termici altamente precisi o passaggi di separazione aggiuntivi. [62] Successivamente all'amplificazione dei circRNA, è stato utilizzato un approccio a nanopori per sequenziare direttamente le sequenze di circRNA a lunghezza completa ed è stato applicato un algoritmo specifico per quantificare l'espressione dei circRNA e riconoscere le sequenze di trascritti mutanti a lunghezza completa. I dati ottenuti evidenziano il notevole vantaggio del sequenziamento a nanopori rispetto alla tecnica Illumina RNA-seq.

Quantificazione dell'espressione dei circRNA

Per la quantificazione dei circRNA si possono utilizzare varie tecniche. Gli approcci più diffusi sono quelli basati sulla PCR e RNA-seq che prevedono l'allineamento delle *reads* e il conteggio delle *reads* allineate per ogni gene/trascritto, e quelli basati sulla fluorescenza.
Nel 2018, ad esempio, Li et al. hanno sviluppato un metodo per misurare quantitativamente i circRNA sfruttando la loro capacità di agire come "spugne" per i miRNA e l'amplificazione attraverso una nucleasi specifica per il duplex (DSN).[62]

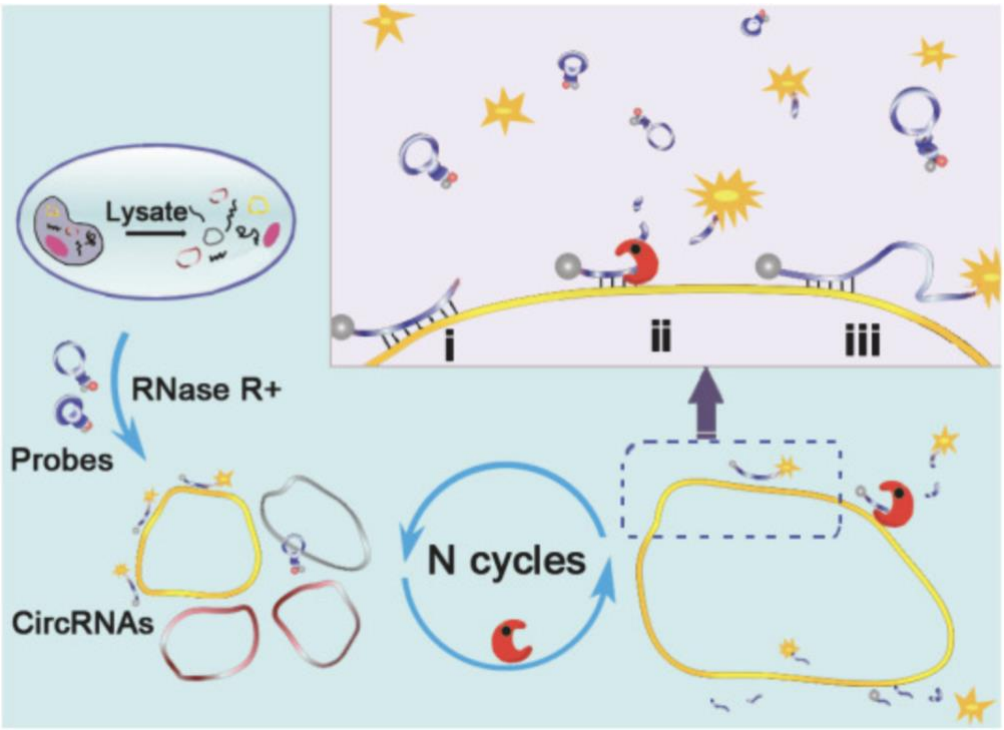


Figura 11:rappresentazione schematica del rilevamento dei circRNA tramite DSN

La DSN è un tipo di enzima nucleasi che degrada il DNA in duplex a doppio filamento DNA o ibridi DNA:RNA, ma non è in grado di tagliare il DNA a singolo filamento, l'RNA a singolo filamento o il RNA a doppio filamento. Nello specifico del loro studio, sono state progettate delle sonde a "molecular beacon" in grado di legarsi a sette siti specifici su un circRNA target. Il legame tra il "molecular beacon" e il suo target induce una modifica conformazionale nella struttura a doppio loop della sonda, scatenando il rilascio di un fluoroforo-quencher e consentendo la rilevazione della fluorescenza della sonda. Successivamente, la DSN degrada il filamento di DNA dell'ibrido DNA/circRNA, rilasciando sia il frammento della sonda fluorescente che il circRNA stesso, il quale può quindi ibridarsi con una nuova sonda "molecular beacon". Questo processo ciclico di degradazione delle sonde da parte della DSN amplifica in modo significativo il segnale fluorescente quando è presente il circRNA target.

Li et al. hanno creato un test che semplifica ulteriormente il rilevamento dei circRNA. Questo test utilizza una sonda che si lega al circRNA target in un punto specifico. Successivamente il DNS taglia il DNA solo nelle regioni lunghe del complesso DNA/RNA (20 bp) ma rimane inattiva nelle regioni più corte (15bp). Solo le sonde legate al circRNA e non all'RNA lineare possono essere tagliate dalla DSN, innescando un processo di amplificazione. Questo porta alla rimozione dei marcatori dalla superficie degli elettrodi, causando una diminuzione della corrente elettrica misurata.

Questo test è in grado di rilevare direttamente i circRNA in campioni biologici complessi con un limite di rilevamento (LOD) bassissimo, fino a 3,5 fM, senza richiedere un pretrattamento

con l'RNase R. Di conseguenza, questa piattaforma elettrochimica offre un approccio semplice per il profilo dei circRNA senza l'uso di apparecchiature costose o procedure complesse.
Tuttavia, gli studi basati sulla DSN, pur amplificando il segnale in presenza di un circRNA target, mostrano un aumento di fluorescenza solo 35 volte superiore rispetto a un controllo senza DSN, non potendo quindi competere con l'efficienza di amplificazione della PCR. Trovare nuovi enzimi con maggiore efficienza di amplificazione o ingegnerizzarli, oppure combinare la DSN con altre amplificazioni isotermiche, agevolerà l'applicazione pratica dei metodi di amplificazione assistita dalla nucleasi nella rilevazione dei circRNA. [62]
Nel caso dei circRNA la quantificazione tramite RNA seq prevede l'allineamento e il conteggio delle *reads* che mappano sulle BSJ. Il numero di *reads* che mappano su una certa BSJ è proporzionale al livello di espressione, ma si deve prestare attenzione ai falsi positivi. I metodi di quantificazione basati su RNA-seq e PCR includono una fase di trascrizione inversa e durante la preparazione delle librerie per il sequenziamento possono verificarsi artefatti come il template-switching, che hanno il potenziale di generare falsi positivi nelle BSJ. Pertanto, la convalida mediante tecniche complementari è essenziale, specialmente quando si trattano nuovi candidati circRNA.
Per ottenere risultati significativi, è inoltre importante eseguire un'analisi statistica dei dati RNA-seq per identificare i geni o le isoforme che sono significativamente diversi tra campioni o condizioni sperimentali.
Un approccio consigliato è eseguire l'RNA-seq sia con che senza il trattamento dell'RNase R e poi cercare un aumento nelle letture delle BSJ in rapporto al numero totale di letture. Inoltre, è importante notare che, a causa della relativa bassa abbondanza di molti circRNA, la quantificazione basata sui dati del RNA-seq può diventare meno affidabile quando si lavora con campioni di dimensioni ridotte o molto diluiti. Questo diventa particolarmente critico quando si confronta l'espressione dei circRNA tra diversi campioni, poiché un rapporto segnale-rumore ridotto e la variabilità nelle popolazioni cellulari possono influenzare notevolmente l'identificazione dei circRNA differenzialmente espressi.
È fondamentale eseguire un sequenziamento con una copertura profonda. Inoltre, l'utilizzo di letture più lunghe (oltre 100 nucleotidi) contribuisce ad incrementare il numero complessivo di letture che mappano sulle BSJ, migliorando la sensibilità dell'analisi.[61]

Metodologie bioinformatiche per la rilevazione e l'analisi dei circRNA in dati di sequenziamento

Dal 2012 è stato sviluppato un gran numero di strumenti di bioinformatica per lo studio delle circRNA. Sebbene le funzioni degli strumenti per i circRNA siano diverse, possono essere

classificati in tre categorie principali che comprendono strumenti per l'identificazione dei circRNA, database di annotazione delle circRNA e altri strumenti.[65]
Nella fig.12 è rappresentata la cronologia storica della ricerca sui circRNA che mostra l'evoluzione della conoscenza e degli strumenti sperimentali e computazionali legati a queste molecole. I segni blu, verdi e rossi rappresentano rispettivamente le scoperte biologiche, gli approcci sperimentali e gli strumenti di bioinformatica rappresentativi.

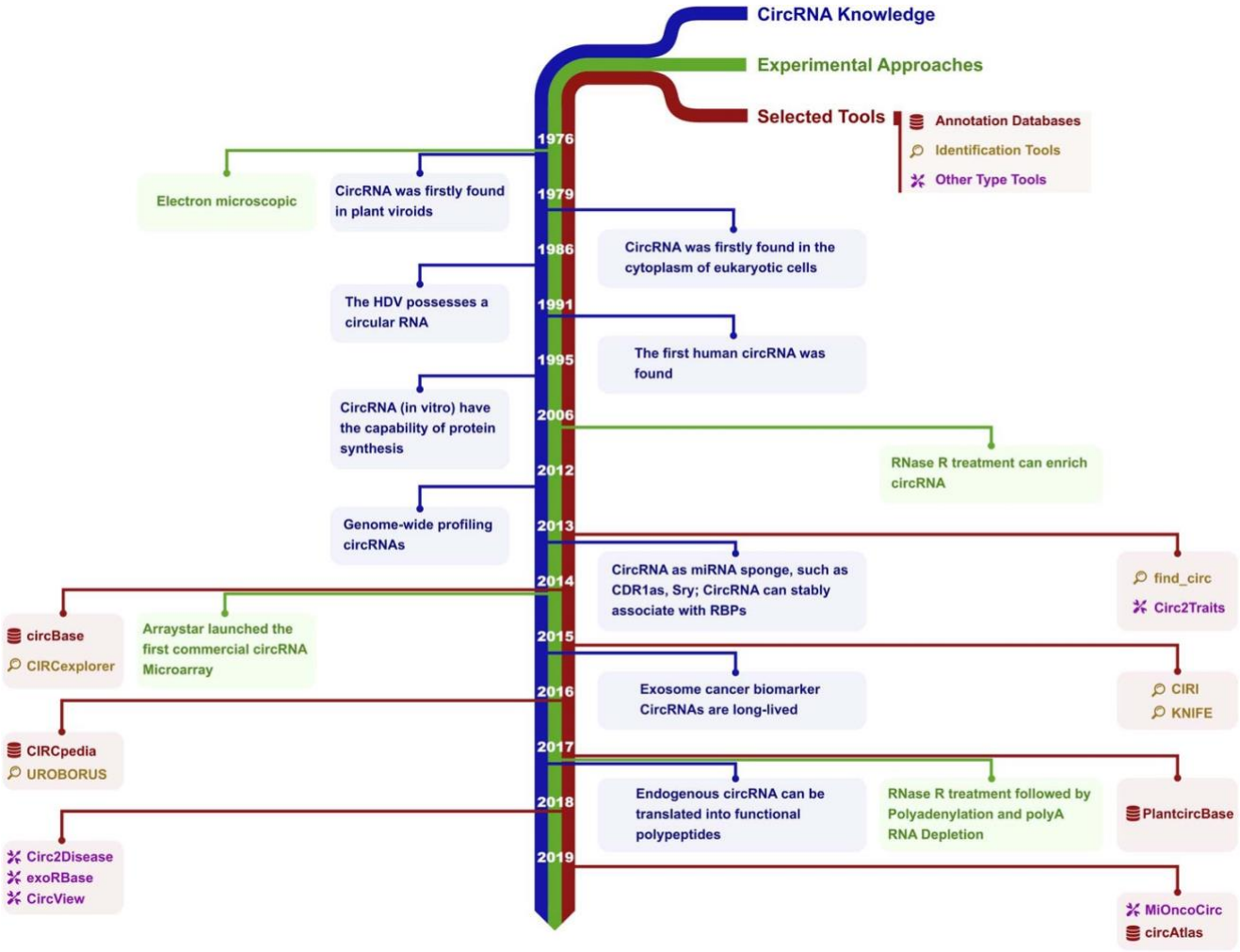


Figura 12: schema sulla storia dei tool utilizzati per la detection degli RNA circolari.[65]

Alcune delle pipeline computazionali ampiamente note e spesso utilizzate per individuare le circRNA includono CIRI2[66,67], Circexplorer2[68], circRNA_finder[69], KNIFE[70], Mapsplice[71] e Segemehl[72].
Ogni circRNA rappresenta un singolo trascritto e le Backsplice Junction *Reads* (BJR) si originano esclusivamente dal sito specifico della sequenza dei circRNA in cui termina la giunzione stessa. Le BJR sono notoriamente più difficili da identificare tramite approcci computazionali rispetto alle letture non splittate e a quelle linearmente splittate, in quanto richiedono allineamenti non collineari e necessitano di ulteriori elaborazioni per eliminare

corrispondenze spurie. Per questo motivo la maggior parte degli strumenti di detection dei circRNA tendono a soffrire di bassi tassi di rilevamento.
La combinazione di queste caratteristiche biologiche dei circRNA e delle sfide computazionali associate alla stima della loro espressione può condurre alla creazione di dataset in cui una notevole frazione dei conteggi risulta essere molto bassa. Questa caratteristica è stata confermata in uno studio che ha coinvolto 34 set di RNA-seq arricchiti per circRNA, provenienti da 17 differenti tessuti umani.[73] Suddividendo i dati in 4 gruppi di allineamenti delle *reads* che rappresentano il segnale di espressione disponibile per la stima dell'espressione genica, lo studio dello splicing alternativo, il confronto tra l'abbondanza di trascritti circolari e lineari espressi da un gene e la stima dell'abbondanza dei circRNA, e confrontando l'entità dei segnali di espressione è stato osservato che indipendentemente dall'arricchimento dei circRNA nella libreria, il segnale più alto è stato ottenuto per le stime dell'espressione genica, seguito dalle letture linearmente splittate. Le BJR hanno mostrato i valori più bassi, anche nei campioni arricchiti di circRNA.
I circRNA meno espressi potrebbero non essere rilevati in alcuni campioni a causa di un bias di campionamento che potrebbe portare i conteggi a zero. La maggioranza dei circRNA (86,6%) viene rilevata solitamente con un conteggio delle BSJ inferiore a 5, con solo il 46,1% dei circRNA rilevati che presenta almeno 2 conteggi delle BSJ. Al fine di aumentare l'attendibilità dei risultati, alcuni strumenti, come circRNA_finder e segemehl, applicano un filtro per riportare solo i circRNA con un conteggio delle BSJ di almeno 5, mentre CirComPara2 e KNIFE effettuano una selezione basata su un conteggio delle BSJ di almeno 2. Circtools filtra i circRNA con almeno 2 conteggi in almeno 2 campioni.[74] Nello stesso studio, per determinare che i risultati non dipendessero da qualche artefatto dell'algoritmo di stima dell'espressione dei circRNA, sono stati utilizzati altri sei ulteriori metodi di quantificazione dei circRNA ed è stato osservato che la distribuzione dei conteggi delle BJR era comparabile tra i metodi di quantificazione. [73] Questi strumenti verificano la presenza di almeno un evento di backsplice nei dati di sequenziamento di nuova generazione (NGS). Naturalmente, ciascuno di questi strumenti ha i propri vantaggi e limitazioni. In generale, l'identificazione dei circRNA a livello genomico da qualsiasi organismo potrebbe effettivamente sottostimare il numero totale di circRNA rilevate, oltre a trascurare i circRNA scarsamente espressi. [29]

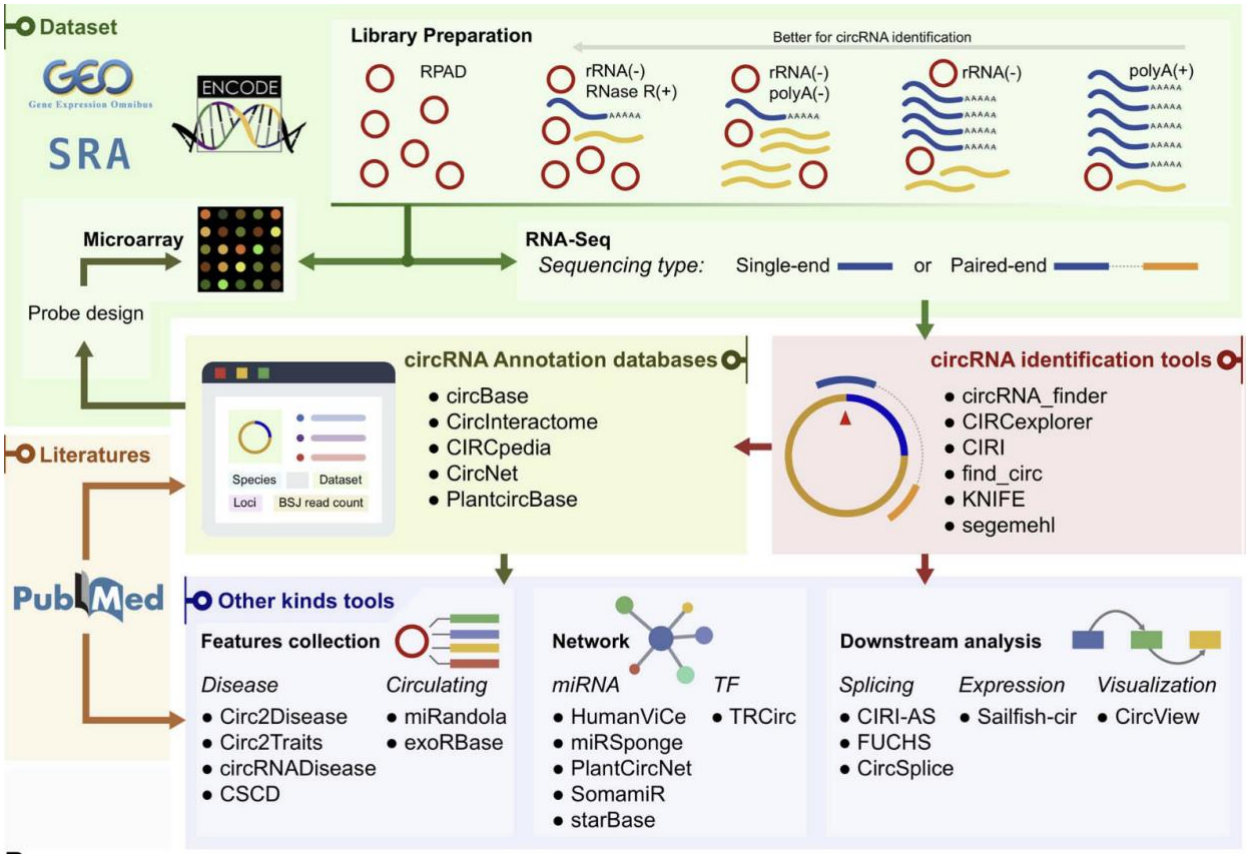


Figura 13: Schema degli strumenti bioinformatici per i circRNA[65].

I circRNA possono essere identificati dai dati di sequenziamento dell'RNA attraverso due principali approcci: l'approccio basato su pseudoriferimento, noto anche come approccio basato su candidati, e l'approccio basato su frammenti, anche chiamato approccio basato su lettura segmentata. Nel primo approccio, viene utilizzata una lista di riferimento contenente potenziali sequenze di giunzione di backsplice (BSJ), spesso derivata da tutte le possibili combinazioni di esoni annotati all'interno di un gene. Questo metodo è limitato alle specie con genomi annotati e ai geni precedentemente annotati, rilevando solamente i circRNA che utilizzano gli stessi siti di splicing dei trascritti lineari, viene utilizzato per avere risultati più affidabili ed è il metodo più veloce da implementare nelle librerie prive di RNA ribosomiale.[62]

Nel secondo approccio, le librerie sono prive di rRNA e trattate con trattamento RNase R prima del sequenziamento high throughput e le letture di sequenziamento non mappate vengono suddivise in sequenze più brevi e poi riallineate sul genoma di riferimento. Questo secondo approccio evita il bias introdotto dall'utilizzo dei modelli esistenti per creare un elenco di candidati e consente l'identificazione di nuovi circRNA attraverso la rilevazione di BSJ non annotate. Tuttavia, è meno accurato, meno sensibile e richiede una maggiore quantità di RNA totale rispetto all'approccio basato su candidati. Inoltre, è più incline a essere influenzato dalla contaminazione da endonucleasi. [62]

In aggiunta alcuni strumenti integrativi, come CirComPara2, combinano i risultati ottenuti da più strumenti.[74]

Ad eccezione di segemehl, gli strumenti che si basano su k-mer e quelli basati su machine learning richiedono l'uso di allineatori esterni, con Bowtie e BWA-MEM (Bowtie-Watson Alignment with the Maximal Exact Matches) come opzioni comuni.[65,75]

Nell'algoritmo di allineamento Bowtie si hanno quattro fasi principali, inizialmente prende una *read* del sequenziamento e crea delle sottostringhe più piccole, queste vengono allineate al genoma di riferimento senza considerare delezioni o inserzioni nella *read*. A questo punto vengono valutati gli allineamenti di queste mini-sequenze e ordinati in base a dei criteri specifici. A questo punto i "seed" identificati vengono estesi in allineamenti completi delle letture. Questo approccio permette di ottenere un allineamento rapido e preciso delle letture di sequenziamento rispetto a un genoma di riferimento noto.[76]

L'algoritmo BWA-MEM è in grado di adattarsi in modo automatico tra allineamenti locali ed end-to-end, supporta l'allineamento di *reads* in modo paired-end e gestisce con precisione gli allineamenti chimerici. Viene utilizzato spesso come primo passaggio negli algoritmi di detection dei circRNA in quanto la sua robustezza nei confronti degli errori di sequenziamento lo rende adatto per una vasta gamma di sequenze.[77]

La maggior parte degli strumenti per l'identificazione dei circRNA rientra nella categoria "stand-alone", ovvero possono funzionare in maniera indipendente da altri oggetti o software. La stragrande maggioranza di questi strumenti mette a disposizione il proprio codice sorgente su GitHub, e il linguaggio di programmazione più utilizzato è Python, anche se R e Perl sono anch'essi linguaggi comuni, mentre la Shell è una scelta frequente per la creazione di flussi di lavoro. Questi strumenti sono in gran parte progettati per essere eseguiti su sistemi Linux o su piattaforme simili a UNIX. Alcuni richiedono una compilazione a partire dal codice sorgente, mentre altri necessitano di una procedura di installazione manuale con le relative dipendenze. Attualmente, l'installazione su Linux è spesso la più agevole, grazie all'ampia disponibilità di metodi di installazione semplici come Conda (bioconda), Docker, l'indice dei pacchetti Python (PyPI) e BiocManager (Bioconductor).[65]

Strumenti bioinformatici di identificazione dei circRNA basati sulle back-spliced Junction

Nel contesto dell'identificazione dei circRNA basata su dati di RNA-Seq, è interessante notare che il pioniere in questo campo è stato il software denominato "Find_circ". Questo strumento ha inaugurato l'approccio di utilizzare le letture di sequenziamento back-spliced (BSJ) per la predizione dei circRNA.[65,78] Tuttavia, successivamente è stato sviluppato uno strumento più sofisticato e robusto denominato "CIRI". CIRI è notevole perché, anziché limitarsi a

identificare le letture di giunzione BSJ, effettua una scansione completa dei dati di sequenza per individuare inizialmente queste letture di giunzione, seguita dall'implementazione di diverse strategie di filtraggio per ridurre al minimo i falsi positivi nell'identificazione delle circRNA.[65,66]

Un altro strumento degno di menzione è "CIRCexplorer2", che identifica le letture di giunzione BSJ originate da esoni back-spliced e lariat intronici. Questo software ha recentemente ricevuto un aggiornamento significativo, che ha ampliato la sua funzionalità per includere l'analisi dell'alternative splicing dei circRNA e la capacità di eseguire l'assemblaggio de novo dei trascritti di RNA circolare.[65,79]

D'altra parte, esistono strumenti come "DCC" e "CircTest" che sfruttano l'output generato dal lettore di sequenze STAR per l'identificazione delle letture di BSJ. Inoltre, va sottolineato che il campo dell'identificazione dei circRNA continua a evolversi, con strumenti più recenti come "KNIFE" che migliorano ulteriormente la sensibilità e la specificità attraverso l'implementazione di un modello lineare generalizzato logistico (GLM). [65,80,81]

Nel panorama degli strumenti basati su BSJ, è possibile individuare alcune applicazioni più specifiche. Ad esempio, "Ularcirc" e "UROBORUS" si distinguono per la loro capacità di rilevare circRNA con bassi livelli di espressione in set di dati di RNA-seq senza la necessità di un trattamento con RNase R. Infine, "circRNA_finder" rappresenta uno strumento di notevole interesse in quanto consente la predizione de novo dei circRNA senza dipendere da annotazioni genetiche o dalla struttura esone-introne. Ciò rende quest'ultimo particolarmente utile per l'identificazione di circRNA con siti di splicing sconosciuti.[62,65,82]

Nonostante la rilevazione delle BSJ sia essenziale, alcuni strumenti si basano su altri parametri per poi ricondursi all'identificazione dei circRNA.

Un primo esempio è rappresentato da "PTESFinder", che si basa sull'identificazione della struttura di exon shuffling post-trascrizionale e successivamente sull'associazione a un modello di sequenza. Al contrario, "CircMarker" adotta un approccio basato su k-mer, mentre "CircDBG" si basa su grafi di De Bruijn per l'identificazione dei circRNA.

Un'altra categoria di strumenti, come "NCLcomparator", è progettata per rilevare trascritti non lineari, inclusi quelli circolari, di splicing trasversale o di fusione. Questi strumenti eseguono un post-screening dei trascritti non lineari precedentemente individuati da altri metodi. Inoltre, alcuni strumenti, come "ACValidator", utilizzano un approccio di convalida in silico basato sull'assemblaggio.[81,83–85]

Tuttavia, una sfida significativa nell'identificazione dei circRNA è rappresentata dalla presenza di eventi di fusione. In questo contesto, "ACFS" si distingue per l'uso di dati di RNA-seq single-end e paired-end per rilevare i circRNA di fusione derivati da eventi di

translocazione cromosomica. In modo simile, "ROP" è in grado di profilare ripetizioni, circRNA e fusioni geniche, mentre "STARChip" supporta la quantificazione dell'abbondanza dei circRNA, l'annotazione e la predizione delle fusioni genomiche.
Inoltre, è importante considerare che la rilevazione dello splicing alternativo rappresenta un'ulteriore sfida, soprattutto in contesti con differenze tra specie. Per affrontare questa sfida, sono stati sviluppati strumenti specifici per le diverse specie, come "ANNOgesic", uno strumento di annotazione del genoma per batteri ed archei che può anche rilevare circRNA e altre caratteristiche genomiche. "AutoCirc" è invece un'applicazione adatta a tutte le specie con sequenza genomica disponibile.
Un'altra strategia per affrontare le sfide dell'identificazione dei circRNA è fornire un punteggio statistico per i circRNA previsti. Questo approccio è stato adottato da "CircRNAFisher", che può eseguire una predizione sistematica de novo dei circRNA basata su valori derivanti dall'analisi delle letture sovrapposte BSJ e delle letture BSJ discordanti.
Infine, sono stati sviluppati alcuni strumenti multifunzionali per l'identificazione dei circRNA. Ad esempio, "hppRNA" è in grado di eseguire il mappaggio delle *reads*, rilevare variazioni nella sequenza e identificare i geni di fusione, oltre a individuare circRNA. In modo simile, "miARma" può identificare mRNAs, miRNAs e circRNA in qualsiasi organismo sequenziato, fornendo un approccio completo per lo studio dei diversi tipi di molecole nucleotidiche. "CIRCexplorer2", invece, può essere considerato multifunzionale in quanto contiene un modulo dedicato all'analisi dello splicing alternativo dei circRNA. "Circtools" offre una vasta gamma di funzionalità, dalla rilevazione e ricostruzione dei circRNA alla progettazione di primer specifici per circRNA, passando per lo screening delle proteine di legame all'RNA (RBP) e l'analisi dell'uso differenziale degli esoni.
I circRNA possono essere individuati anche sfruttando dati provenienti da varie tecniche di sequenziamento, tra cui CLIP-Seq (Cross-linking and immunoprecipitation sequencing), ISO-Seq (Isoform sequencing), Ribo-Seq (Ribosome profiling) e miRNA-Seq (microRNa sequencing), e per questo scopo esistono numerosi strumenti specifici.
Un'applicazione di base per identificare i circRNA basandosi sui dati di CLIP-Seq è rappresentata da "CircScan". Questo strumento riconosce le letture di giunzione back-spliced (BSJ) e le utilizza per individuare i circRNA. Per un approccio più completo, vi è "CircTools" (starBase), che costituisce una pipeline composta da tre distinti software (circSeeker, circAnno e clipSearch) progettati per individuare e annotare i circRNA e per studiare le loro interazioni con i miRNA, facendo uso dei dati provenienti da CLIP-Seq. [83,86,87] Un'alternativa completa è fornita da "PRAPI", una pipeline dedicata all'analisi dei dati ISO-Seq, che include la rilevazione delle iniziazioni di trascrizione alternative, lo splicing alternativo e

l'identificazione dei circRNA. Uno strumento specializzato che può identificare circRNA con potenziale di codifica proteica e letture di giunzione è Ribo-Seq CircPro.
"CircularRNAPipeline", invece, è una pipeline versatile in grado di individuare circRNA direttamente dai dati grezzi in formato fastq.
L'aggiunta di "CIRCexplorer2" alla pipeline "CircularRNAPipeline", ha permesso di individuare i circRNA anche nei dati di RNA-Seq a singola cellula, ampliando così le possibilità di studio dei circRNA. Queste pipeline rappresentano un importante strumento per la ricerca, consentendo ai ricercatori di esplorare i circRNA utilizzando dataset più ampi e di sfruttare dati preesistenti analizzati per altri scopi.[65,88]

Strumenti integrati di identificazione dei circRNA

È stato dimostrato che l'integrazione di diversi strumenti di identificazione dei circRNA può ridurre il tasso di falsi positivi. Va notato che alcuni circRNA sono sensibili al trattamento con RNase R, il che significa che possono essere notevolmente ridotti o assenti dopo questo trattamento, come nel caso di circ_CDR1as. Pertanto, affidarsi esclusivamente a strumenti basati su librerie preparate con trattamento RNase R può essere problematico per identificare questi circRNA sensibili a RNase R. Di conseguenza, è importante utilizzare diversi strumenti di identificazione e condurre l'identificazione dei circRNA su diversi set di dati che possono variare nei trattamenti applicati. Questo approccio migliora notevolmente l'affidabilità nell'identificazione dei circRNA.
Per l'identificazione dei circRNA sono state sviluppate pipeline integrate. Ad esempio, CirComPara è una pipeline completa che comprende l'identificazione, la quantificazione dell'abbondanza e l'annotazione dei circRNA. Questa pipeline integra diversi software, tra cui CIRCexplorer, CIRI, find_circ e segemehl, consentendo agli utenti di sfruttare le capacità di diversi algoritmi di identificazione. Inoltre, è possibile confrontare o combinare diverse predizioni di circRNA ottenute da algoritmi come circRNA_finder, CIRCexplorer, CIRI, find_circ, MapSplice, ACSF, DCC, KNIFE e UROBORUS per migliorare ulteriormente la sensibilità e la specificità dell'identificazione dei circRNA mediante l'utilizzo di circ_battle.[81,85,86,89]
Un altro esempio è rappresentato da RAISE, un'ulteriore pipeline che non solo misura l'abbondanza dei circRNA, ma è in grado anche di prevedere la loro struttura. RAISE integra vari strumenti, tra cui MapSplice, find_circ, ACFS e circRNA_finder, fornendo un approccio completo per l'analisi dei circRNA.[65,84,90] Un ulteriore passo avanti è rappresentato da CircRNAwrap, una pipeline estremamente completa che combina l'identificazione dei circRNA utilizzando diversi strumenti come KNIFE, find_circ, CIRI, CIRCexplorer, MapSplice, ACFS, circRNA_finder e DCC. Inoltre, questa pipeline consente di prevedere i

trascritti circRNA incorporando RAISE e CIRI-as e di stimare l'abbondanza utilizzando sailfish-cir. Questo approccio integrato offre una comprensiva panoramica dell'identificazione, predizione e quantificazione dei circRNA.

In sintesi, le pipeline integrate rappresentano un passo avanti significativo nell'identificazione dei circRNA, consentendo agli studiosi di sfruttare al meglio le risorse di diversi algoritmi e migliorare l'affidabilità delle analisi dei circRNA.[65,91]

Strumenti di identificazione basati sul machine learning

Per la predizione dei circRNA c'è inoltre una categoria di strumenti che si avvale delle tecniche di machine learning. Questi strumenti utilizzano la conoscenza acquisita sui circRNA esistenti e le caratteristiche distintive di tali molecole per addestrare modelli di classificazione. Gli algoritmi di machine learning consentono a questi modelli di apprendere da un insieme di circRNA precedentemente identificate, migliorando così la loro capacità predittiva.

Man mano che si ha una maggiore comprensione dei circRNA, si ampliano anche le caratteristiche utilizzate come input per questi modelli di machine learning. Tra le caratteristiche comuni utilizzate figurano le ripetizioni di sequenze ALU, motivi strutturali e motivi sequenziali. L'obiettivo principale di tali strumenti è distinguere i circRNA da altre categorie di RNA non codificanti lunghi (lncRNA) o molecole RNA simili.

Esempi di questi approcci includono PredcircRNA, che si basa su un algoritmo di apprendimento a kernel multiplo per identificare i circRNA, e WebCircRNA, che viene addestrato su dati provenienti da cellule staminali e consente di prevedere circRNA specifiche per questo tipo cellulare. Altri strumenti, come PredicircRNATool, sfruttano proprietà conformazionali e termodinamiche nelle regioni circostanti per effettuare le predizioni, mentre DeepCirCode si avvale di una rete neurale convoluzionale (CNN) per predire i circRNA umane. Quest'ultimo rappresenta un notevole avanzamento in quanto costituisce il primo modello di deep learning progettato specificamente per la predizione dei circRNA.[65,92,93]

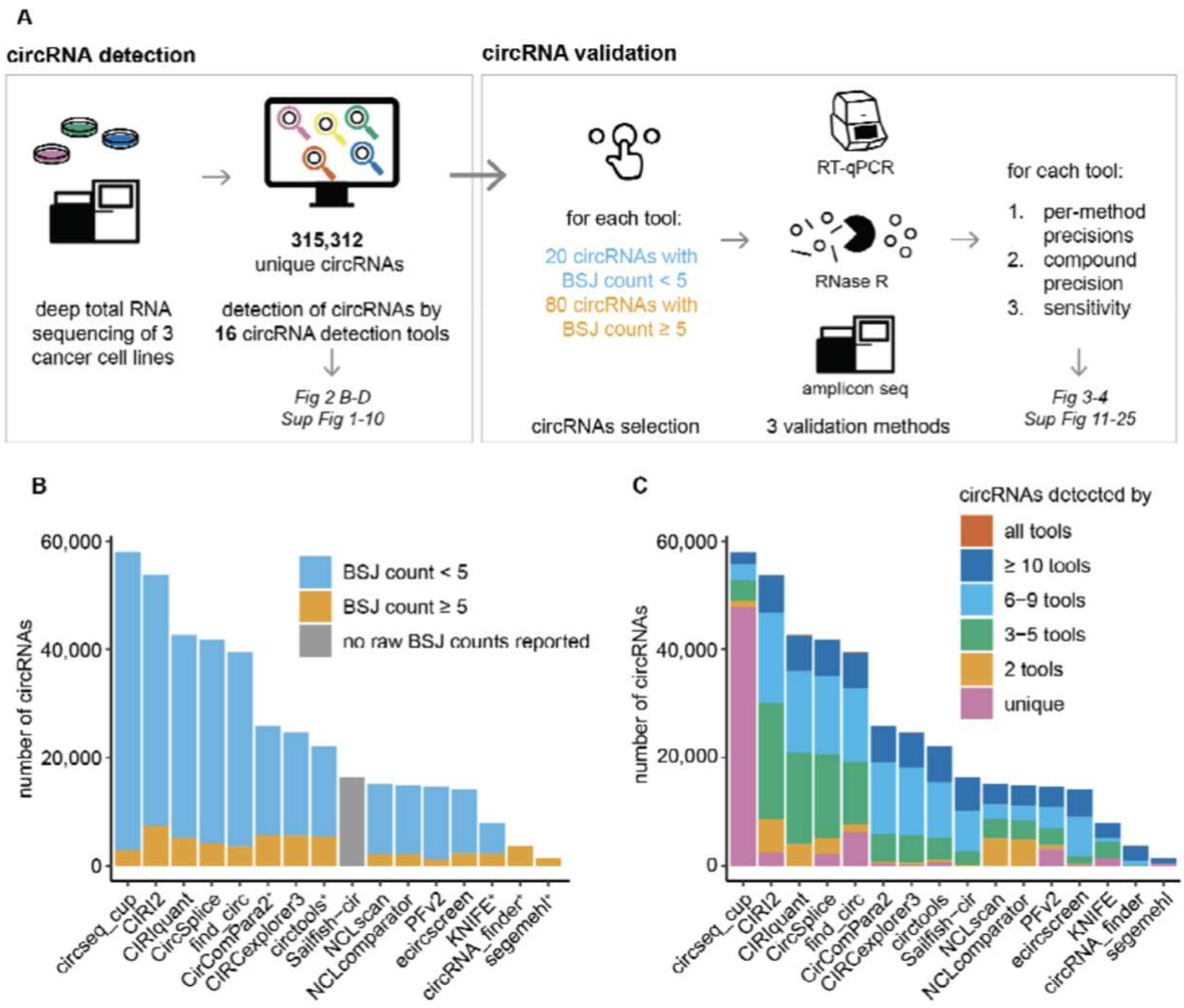


Figura 14: confronto tra i vari tool di detection[74]

Database circRNA

Molti dei database circRNA attualmente disponibili raccolgono circRNA da letteratura scientifica utilizzando una varietà di strumenti di identificazione e set di dati NGS specifici. Altri, invece, adottano una pipeline unificata per elaborare i dati RNA-seq e conservare i risultati delle predizioni dei circRNA. Poiché lo sviluppo di questi database è in corso, è di fondamentale importanza continuare a espanderne e migliorarne il contenuto e la qualità. Alcuni di questi database sono progettati appositamente per i circRNA e contengono informazioni dettagliate sulle specie di circRNA e i dati relativi al numero di letture di back-splicing junction (BSJ). Un esempio di questo tipo è circBase, che ospita circRNA di animali, fornendo sequenze e coordinate genomiche. L'ultima versione di circBase annota i circRNA basandosi su dati provenienti da nove pubblicazioni diverse. [94]Un altro esempio è CircFunBase, che contiene circRNA annotati manualmente.

Il database CIRCpedia (http://www.picb.ac.cn/rnomics/circpedia) è una raccolta di tutti gli eventi di back-splicing alternativo e splicing alternativo identificati nei circRNA, insieme a

esoni di nuova identificazione. Questo strumento online consente la ricerca, l'analisi e il download agevole di molteplici circRNA generati dal medesimo locus genico in differenti linee cellulari. Al momento, il database comprende retro-splicing e splicing alternativo di circRNA provenienti da 13 linee cellulari umane. Si prevede di arricchire il repertorio includendo dati provenienti da una vasta gamma di campioni, quali linee cellulari, tessuti e specie, man mano che saranno disponibili ulteriori dataset di RNA-seq di alta qualità. Specificando la posizione genomica, è possibile recuperare tutti i circRNA identificati in quel locus genico, insieme ai relativi eventi di retro-splicing e splicing alternativo. Gli utenti hanno inoltre la possibilità di restringere la ricerca a un particolare tipo di back-splicing o splicing alternativo o a specifiche linee cellulari tramite diverse opzioni di configurazione. CircRNADb, invece, si concentra sui circRNA umani, includendo circRNA esonici annotati e circRNA con potenziale codifica proteica, basandosi su dati estratti dalla letteratura.[95]
Alcuni database cercano di raccogliere informazioni sui circRNA all'interno di database più ampi che includono anche altri ncRNA (non-coding RNA). Questi database offrono dati sulle interazioni dei circRNA con altri ncRNA e informazioni sull'espressione. Ad esempio, CircInteractome fornisce strumenti per recuperare informazioni sui siti di legame delle proteine leganti l'RNA (RBP) e dei microRNA (miRNA) sui circRNA umane, oltre a strumenti per la progettazione di siRNA per il silenziamento delle circRNA. DeepBase è un altro database ampio che annota e scopre piccoli RNA, lncRNA e circRNA dai dati di sequenziamento di nuova generazione, utilizzando dati estratti da circBase e dalla letteratura.
Un problema attuale nei database circRNA è la nomenclatura. Non esiste attualmente una nomenclatura unificata per i circRNA, e gli identificatori utilizzati tra i diversi database non sono standardizzati. L'adozione di una nomenclatura unificata faciliterebbe notevolmente l'integrazione dei dati provenienti da diversi database circRNA. [65,95–97] CircBase, ad esempio, utilizza un sistema di identificazione basato sulla specie, il numero circ, o il simbolo del gene ospite, o un nome di convenzione. Alcuni database utilizzano anche il simbolo del gene ospite con numeri di accesso o le locazioni BSJ come identificatori. Unificare la nomenclatura rimane una sfida significativa per il campo della ricerca sui circRNA.
Molti database raccolgono informazioni sui circRNA che vanno al di là delle sequenze stesse dei circRNA, enfatizzando aspetti specifici come le regioni regolatorie associate ai Single Nucleotide Polymorphism (SNP), ovvero varianti genetiche comuni in cui un singolo nucleotide (A, T, C o G) all'interno del DNA è sostituito con un altro nucleotide nella sequenza genica di un individuo. Un esempio di ciò è rSNPBase, un database che raccoglie elementi regolatori associati ai SNP, tra cui quelli nelle regioni circRNA. Altre caratteristiche dei circRNA possono essere associate a specifiche malattie.

I circRNA si prestano bene come biomarcatori grazie al loro ciclo di vita prolungato, all'abbondanza in cellule specifiche e alla loro rilevabilità in vari fluidi corporei e per sfruttare appieno questi vantaggi, sono stati creati database volti a conservare le relazioni tra circRNA e le malattie umane. Tra i database di interesse figurano Circ2Disease, circRNADisease e CircR2Disease, che sono database curati manualmente e convalidati sperimentalmente che catalogano le associazioni tra circRNA e malattie.
In modo analogo, Circ2Traits raccoglie le associazioni potenziali tra circRNA e malattie umane. Per quanto riguarda il cancro, CSCD e MiOncoCirc sono database specifici che mirano a agevolare lo studio funzionale dei circRNA legate al cancro, mentre HDncRNA e LncRNADisease catalogano ncRNA associati a malattie, tra cui i circRNA. Ulteriori database rilevanti per la scoperta di biomarcatori includono exoR-Base, che contiene RNA associati agli esosomi del sangue umano (mRNA, lncRNA e circRNA) e BBBomics, che conserva dati omici relativi alla barriera emato-encefalica umana, tra cui miRNA, lncRNA e circRNA.
È importante sottolineare che alcune delle evidenze sull'associazione tra circRNA e malattie provenienti da diversi database risultano essere in conflitto. Questa discrepanza potrebbe riflettere l'uso di criteri diversi per stabilire l'associazione con le malattie o differenze nei campioni di popolazione studiati. Circ2Disease, CircR2Disease e circRNADisease sembrano essere i database principali per registrare e cercare circRNA correlati alle malattie, poiché raccolgono le associazioni con una vasta gamma di malattie basate su una revisione manuale della letteratura scientifica. [65,98]

Contesto e obiettivi di studio

Partendo dall'analisi bioinformatica, che svolge un ruolo cruciale nella detection dei circRNA, questa tesi ha come obiettivo primario lo sviluppo di una pipeline per l'identificazione e la quantificazione dei circRNA presenti all'interno di dataset di RNAseq. A tal fine, sono stati impiegati e valutati algoritmi e strategie diverse per garantire un rilevamento affidabile dei circRNAs, e la determinazione dei loro livelli di espressione anche in campioni longitudinali raccolti nel corso del tempo. Il prodotto sviluppato con questo lavoro di tesi è stato testato su dataset scaricati da repository pubblici, e sarà impiegato successivamente nel contesto di un progetto di ricerca per la valutazione del ruolo dei circRNAs nello sviluppo neuronale del feto esposto a diverse infezioni congenite da parte di virus patogeni umani.

Materiali e Metodi

Nel presente capitolo, verranno presentati i metodi utilizzati nella ricerca dei circRNA. La scoperta e l'analisi degli RNA circolari richiedono l'utilizzo di strumenti e approcci specifici. Questo capitolo si propone di fornire una panoramica dettagliata dei materiali e dei metodi adottati nel corso del lavoro di tesi per identificare e caratterizzare i circRNA in campioni biologici. In particolare, sono stati utilizzati tre tool per la detection: CIRI2 e CIRCexplorer2 e CIRIquant, e CIRI2 e CIRIquant sono stati utilizzati anche per la quantificazione degli RNA circolari. Circall simulator è stato invece utilizzato in un primo momento per la simulazione di *reads* contenenti RNA circolari.

CIRI2

CIRI è un tool bioinformatico per la rilevazione dei circRNA che è stato implementato nel 2015[66]. Questo algoritmo, a differenza di quelli sviluppati precedentemente che dipendono da annotazioni preesistenti o richiedono un passaggio di arricchimento specifico per i circRNA, si basa su un algoritmo innovativo che sfrutta la rilevazione delle BJS. [99] Questo algoritmo è implementato nel tool di allineamento (SAM) di BWA-MEM e accompagna una procedura di filtrazione sistematica finalizzata all'eliminazione dei falsi positivi.
Durante la prima scansione dell'allineamento SAM, CIRI rileva le BSJ che riflettono un candidato circRNA. Viene implementata una filtrazione preliminare utilizzando il mapping in paired-end (PEM) e i segnali di splicing GT-AG per le giunzioni. Dopo aver raggruppato le letture di giunzione e registrato ciascun candidato circRNA, CIRI esegue nuovamente la scansione dell'allineamento SAM per rilevare ulteriori letture di giunzione e nel frattempo esegue ulteriori filtraggi per eliminare candidati falsi positivi risultanti da letture mappate erroneamente di geni omologhi o sequenze ripetitive. Infine, i circRNA identificati vengono restituiti con le informazioni di annotazione.[66]
CIRI rappresenta un approccio indipendente dalle banche dati dei circRNA per il loro rilevamento tramite l'uso di un algoritmo de novo.
Questo approccio può individuare nuovi candidati circRNA per la convalida sperimentale e la formulazione di nuove ipotesi. In particolare, CIRI presenta i seguenti vantaggi fondamentali rispetto agli algoritmi basati sull'annotazione: è in grado di rilevare circRNA trascritti da regioni genomiche introniche o intergeniche ed è applicabile a dati di sequenziamento di eucarioti che non sono ben annotati o addirittura privi di annotazione.
CIRI raccoglie e confronta le informazioni di allineamento grezze per tutte le divisioni di allineamento di una lettura al fine di individuare BSJ, anziché suddividere artificialmente una

lettura parzialmente mappata in due parti o allineare tutte le letture a un database personalizzato basato su presupposti a priori.

In base alle prestazioni di CIRI su dati simulati [66] è stato osservato che le *reads* single end tendono maggiormente a produrre falsi positivi rispetto alle letture paired end perché manca l'informazione PEM, uno dei filtri utilizzati nei parametri predefiniti. Tuttavia, CIRI mette a disposizione dei parametri in modo da poter impostare delle soglie che garantiscano una qualità di mappatura minima, anche se a costo di una minore sensibilità. Inoltre, per le letture di breve lunghezza, che possono comportare una minore sensibilità, è possibile migliorare le prestazioni di CIRI modificando alcuni parametri di BWA-MEM in modo che consentano l'allineamento anche per punteggi di mappatura bassi.[66]

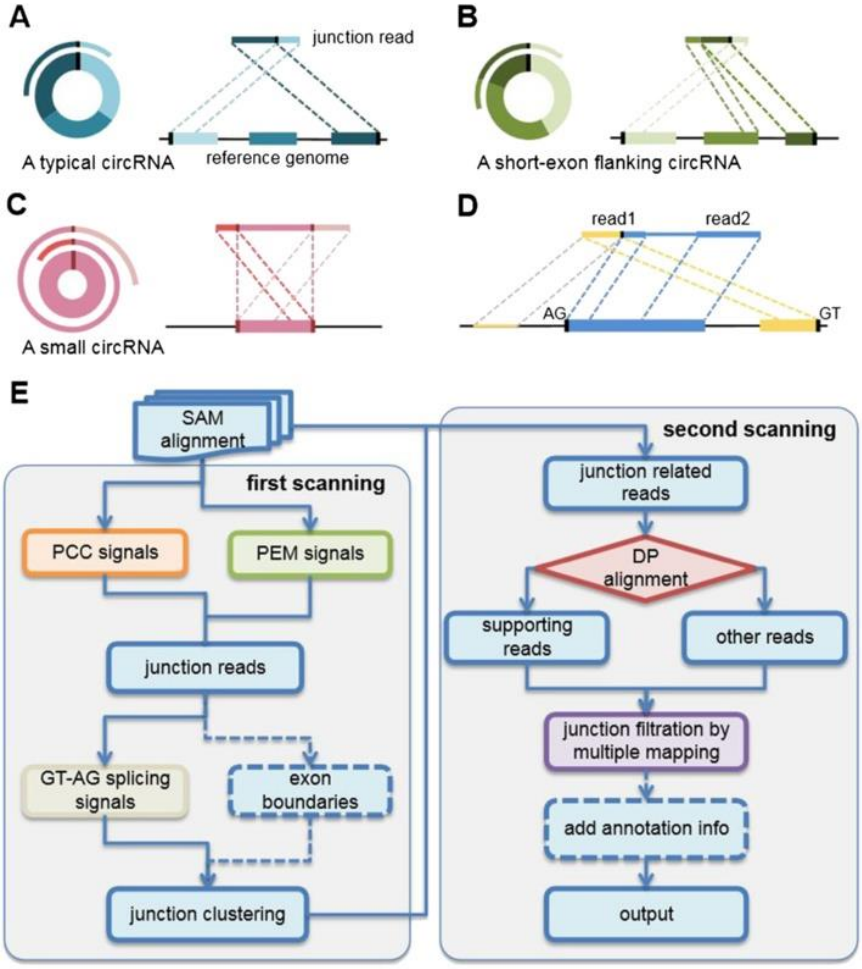


Figura 15:pipeline di identificazione dei circRNA in CIRI[66]

CIRI2 rappresenta un passo avanti nella rilevazione dei circRNA, ed è la naturale evoluzione dell'algoritmo CIRI, che ha consentito di superare alcune delle sfide precedenti e di migliorare la precisione e l'efficienza nell'identificazione dei circRNA. CIRI2 utilizza una strategia più avanzata basata su una stima efficiente della massima verosimiglianza (MLE) per differenziare in modo più preciso le letture di giunzione retro-spliced (BSJ) dalle letture non-

BSJ. Questo significa che CIRI2 è in grado di identificare i circRNA con una maggiore precisione rispetto a CIRI.
Inoltre, CIRI2 è stato progettato per funzionare più rapidamente ed efficientemente, grazie all'implementazione del multithreading e a un'ottimizzazione complessiva delle prestazioni che gli consentono di gestire grandi quantità di dati in modo più efficiente rispetto a CIRI.[67]
CIRI2 richiede gli stessi tipi di dati in input di cui ha bisogno anche CIRI, tra cui sequenze di riferimento formattate in formato FASTA e l'allineamento SAM generato da BWA-MEM. Questi dati sono fondamentali per la rilevazione de novo dei circRNA basata sulle BSJ (Back-Spliced Junction). Inoltre, CIRI2 offre la possibilità di includere un input GTF opzionale, che può essere utilizzato per una rilevazione supplementare basata sull'annotazione, prendendo come riferimento i noti limiti degli esoni degli RNA lineari. Questo input GTF può anche essere utilizzato per ottenere un'annotazione dettagliata di tutti i loci dei circRNA.
CIRI2 è stato soggetto a un'ottimizzazione approfondita, in particolare per le fasi chiave del processo di rilevazione dei circRNA. Queste ottimizzazioni includono la capacità di inferire la regione originale per i segmenti di lettura del sequenziamento basandosi sulla corrispondenza dei *seed,* ovvero piccoli segmenti di una lettura di sequenziamento che vengono utilizzati per cercare corrispondenze nella sequenza di riferimento genomica, e la capacità di distinguere le letture di BSJ dalle letture di giunzione forward-spliced (FSJ) attraverso un adattamento del calcolo di massima verosimiglianza (MLE).
Un altro punto importante da notare è che CIRI2 è in grado di gestire dati di sequenziamento con diverse lunghezze di lettura.
Esistono situazioni in cui un software di allineamento non può determinare la posizione di mappatura di un segmento in una lettura di sequenziamento. Tuttavia, è spesso essenziale per uno strumento di rilevazione dei circRNA poter dedurre se il segmento provenga da una specifica regione genomica. In CIRI, durante la seconda scansione, vengono individuate alcune letture di BSJ con un breve segmento che affianca il BSJ. Questi segmenti brevi, chiamati "letture di giunzione sbilanciate", non possono essere allineati con precisione alla sequenza di riferimento dai programmi di allineamento delle letture corte. Per risolvere questa sfida, vengono impiegati allineamenti basati su programmazione dinamica utilizzando un gruppo di segmenti provenienti da altre letture con posizioni di mappatura affidabili. Questi allineamenti basati su programmazione dinamica richiedono spesso un considerevole calcolo computazionale.
Al contrario, CIRI2 è in grado di affrontare rapidamente questa fase chiave attraverso il matching di *seeds.* Poiché i segmenti di lettura più corti sono più suscettibili a causare corrispondenze spurie, l'inferenza basata su un singolo match spesso produce risultati casuali.

Una soluzione notevolmente più robusta è quella di applicare il matching di multipli seed per determinare la posizione di un segmento. In questo scenario, la maggior parte dei seed dovrebbe corrispondere perfettamente alla loro posizione genomica originale, mentre una piccola frazione potrebbe non farlo a causa di errori di sequenziamento o di interruzioni causate dagli introni.

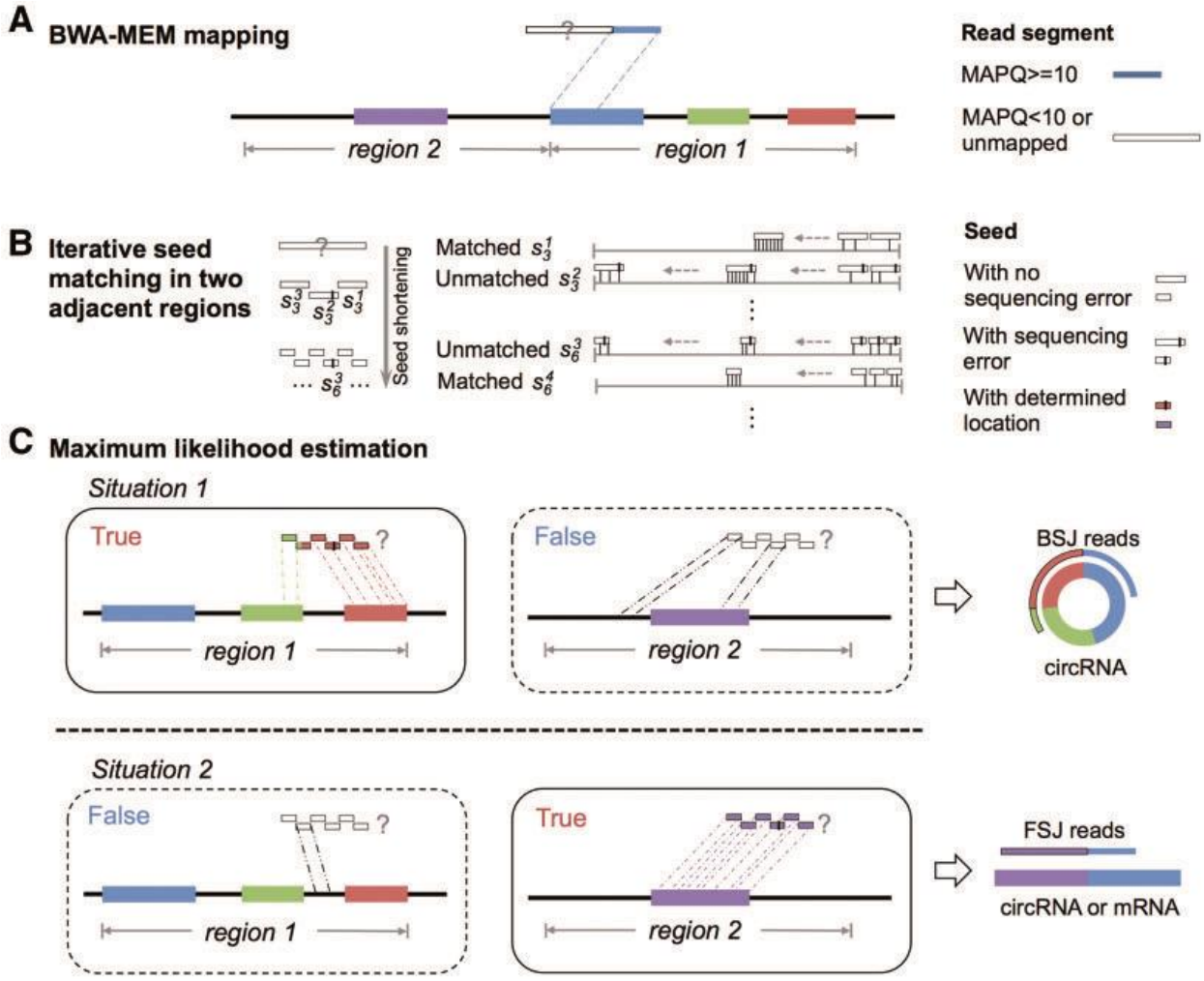


Figura 16: Rilevamento delle letture BSJ basato su MLE[67]

L'algoritmo MLE in CIRI2 è utilizzato per identificare le regioni genomiche più probabili da cui provengono i segmenti chiave nei *reads* BSJ e per distinguere tra regioni di back-splice e forward splice.

Inizialmente, CIRI2 seleziona due regioni genomiche candidate, Region 1 (downstream) e Region 2 (upstream), da cui potrebbero provenire i segmenti chiave nei read BSJ. Queste regioni sono basate su considerazioni di back-splice e forward splice. Per ciascuna delle regioni candidate, CIRI2 esegue un allineamento dei "seed" delle *reads* e calcola la probabilità di allineare con successo un seed con o senza errori di sequenziamento alla regione genomica candidata. Queste probabilità sono calcolate utilizzando una formula basata sulla probabilità binomiale e tengono conto del numero di seed corrispondenti e del tasso di errore di sequenziamento. L'algoritmo MLE sceglie la regione genomica candidata che ha la

probabilità massima di generare i read BSJ osservati. Questo viene fatto utilizzando le probabilità calcolate durante l'allineamento dei seed. Inoltre, stima il numero di seed che corrispondono alla regione genomica ottimale, tramite le probabilità calcolate durante l'allineamento dei seed e sulla distribuzione binomiale.
Al fine di agevolare l'analisi di vasti insiemi di dati, CIRI2 implementa il multithreading mediante il modulo Perl 'threads'. Quando l'utente specifica più di un thread tramite il parametro T, CIRI2 procede innanzitutto alla suddivisione del file SAM in un numero corrispondente di parti equivalenti. È da notare che l'uso del modulo 'threads' comporta un aumento dell'utilizzo della RAM per ciascun thread assegnato. Poiché tale suddivisione può causare la separazione dei record di allineamento dalle rispettive *reads* nei punti di divisione, CIRI2 registra tali *read* e, successivamente, estrae i rimanenti allineamenti dalla parte successiva per riunirli, garantendo così che ogni porzione suddivisa del file SAM contenga il completo registro di allineamento per le coppie di *read*. Nel corso della doppia scansione dei record SAM, CIRI2 assegna thread a ciascuna delle parti suddivise e identifica i *read* BSJ mediante l'utilizzo del MLE come precedentemente descritto. Al fine di limitare l'uso di memoria RAM, CIRI2 archivia i risultati intermedi (ossia i BSJ candidati e i *read* corrispondenti) in un file temporaneo durante la scansione dell'allineamento SAM. Al termine del lavoro di tutti i thread, tali file temporanei sono processati da un singolo thread per ottenere le previsioni definitive dei circRNA.
Nel seguente lavoro di tesi è stato utilizzato CIRI2 per l'identificazione dei circRNA in quanto rappresenta uno degli approcci maggiormente utilizzati e più efficienti come dimostrato da vari studi.[100,101]

Implementazione pratica: comandi e istruzioni

Per l'utilizzo efficace di CIRI2 è fondamentale disporre della versione 5.8 o superiore di Perl, in quanto CIRI2 fa uso di moduli Perl[5], e il sistema operativo deve essere MacOs o Linux. Prima di utilizzare CIRI2, è necessario eseguire una fase di pre-elaborazione dei dati di sequenziamento. La prima fase, sebbene opzionale, comprende la procedura di "trimming" delle *read*, ovvero la rimozione di segmenti indesiderati o di bassa qualità dei dati di sequenziamento. Questo passaggio è consigliato se si desidera migliorare la qualità dei dati in vista dell'analisi successiva. Successivamente al "trimming", le *read* vengono allineate al genoma di riferimento tramite l'utilizzo del tool BWA-MEM.

[5] Un modulo in Perl è un insieme di subroutine e variabili correlate che eseguono un insieme di compiti di programmazione.

Per quanto riguarda i requisiti di input di BWA-MEM, è necessario fornire il genoma di riferimento in formato FASTA e le *read* che si intendono allineare. In questo lavoro di tesi sono stati utilizzati i genomi di riferimento GRCh38 e hg19, scaricati tramite il comando "wget" che consente di recuperare contenuti e file dai server Web tramite il terminale.
Le sequenze destinate all'allineamento con il genoma di riferimento sono reperibili nel database SRA (Sequence Read Archive), accessibile tramite il seguente link: https://www.ncbi.nlm.nih.gov/sra.
Per estrarre i dati in formato FASTQ e FASTA dal database SRA è stato utilizzato il comando fasterq-dump che è parte del toolkit SRA.
Tale database è principalmente focalizzato sulla conservazione dei dati di sequenziamento di nuova generazione (NGS) generati da una varietà di piattaforme, tra cui Illumina, Ion Torrent, PacBio e altre.
Per estrarre i dati in formato FASTQ e FASTA dal database SRA, è stato impiegato il comando "fasterq-dump," che costituisce una componente del toolkit SRA.
n questa fase è stato utilizzato il comando integrato fornito da CIRI2. È stato dato in input il file SAM, ottenuto attraverso l'allineamento al genoma di riferimento, insieme al genoma stesso in formato FASTA e in formato GTF. Il risultato è stato un file di output in formato .txt con le colonne delimitate da tabulazioni ("\t" in ambiente shell e Perl). Ciascuna colonna ha fornito dettagliate informazioni su un circRNA previsto, inclusi il cromosoma di riferimento in cui è stata individuata la giunzione back-splice, le posizioni di inizio e fine della BSJ su quel cromosoma e gli ID delle *read* in cui sono state individuate le giunzioni circolari (separati da virgole).
Per aumentare l'efficienza complessiva del processo di rilevamento dei circRNA, è stato sviluppato uno script personalizzato usando il linguaggio bash che ha consentito di eseguire tutte queste operazioni in un singolo comando, semplificando notevolmente il flusso di lavoro e accelerando la rilevazione dei circRNA.

Figura 17: script CIRI2

La fig. 17 rappresenta l'analisi CIRI2 eseguita sulle sequenze SRR7062764- SRR7062769, che rappresentano il trascrittoma di cellule staminali neurali umane infette con il virus Zika[6].[102]

Gli algortimi bioinformatici per la quantificazione dei circRNA sono inclini a fornire risultati falsi positivi. Utilizzare l'intersezione tra più di un algoritmo può fortemente ridurre il numero di errori. Per questo motivo, nel seguente lavoro di tesi è stato utilizzato anche il tool CIRCexplorer2.[61,64,67]

CIRCexplorer2

La prima versione del tool CIRCexplorer è stata presentata nel 2014 (Zhang et. al) come uno strumento tramite il quale identificare le BSJ. Questa metodologia coinvolge un processo di allineamento delle *reads* di RNA-seq al genoma di riferimento utilizzando l'algoritmo TopHat[103]. Le letture che non riescono ad allinearsi in modo appropriato vengono sottoposte a un secondo allineamento univoco al genoma utilizzando il tool TopHat-Fusion[7].[104]Le *reads* che soddisfano il criterio di allineamento con TopHat-Fusion sullo stesso cromosoma in un ordine non colineare, ma che non corrispondono all'allineamento con TopHat, vengono selezionate come potenziali candidati per le giunzioni back-splice (BSJ). Successivamente, queste letture vengono ulteriormente allineate rispetto all'annotazione genica esistente per determinare con precisione le posizioni di inizio e fine delle BSJ.[105]

La figura 18 illustra il flusso di lavoro per l'identificazione e l'annotazione delle letture contenenti BSJ utilizzando CIRCexplorer. Questo approccio è stato successivamente implementato e ampliato in CIRCexplorer2, che offre la flessibilità di supportare altri strumenti di allineamento come STAR e MapSplice. Ciò consente di adattare l'analisi alle diverse esigenze di allineamento per i dati di RNA circolari e per il data mining.[68]

[6] Il virus Zika è un virus a RNA del genere Flavivirus, trasmesso principalmente attraverso le punture di zanzara. o. Nei casi più comuni, l'infezione da ZIKV provoca sintomi lievi o addirittura può essere asintomatica. Tuttavia, l'infezione durante la gravidanza può portare a gravi complicanze nello sviluppo fetale, come la microcefalia.

[7] TopHat Fusion è una versione migliorata di TopHat con la possibilità di allineare le letture attraverso i punti di fusione, che derivano dalla rottura e dalla ricongiunzione di due cromosomi diversi o da riarrangiamenti all'interno di un cromosoma.

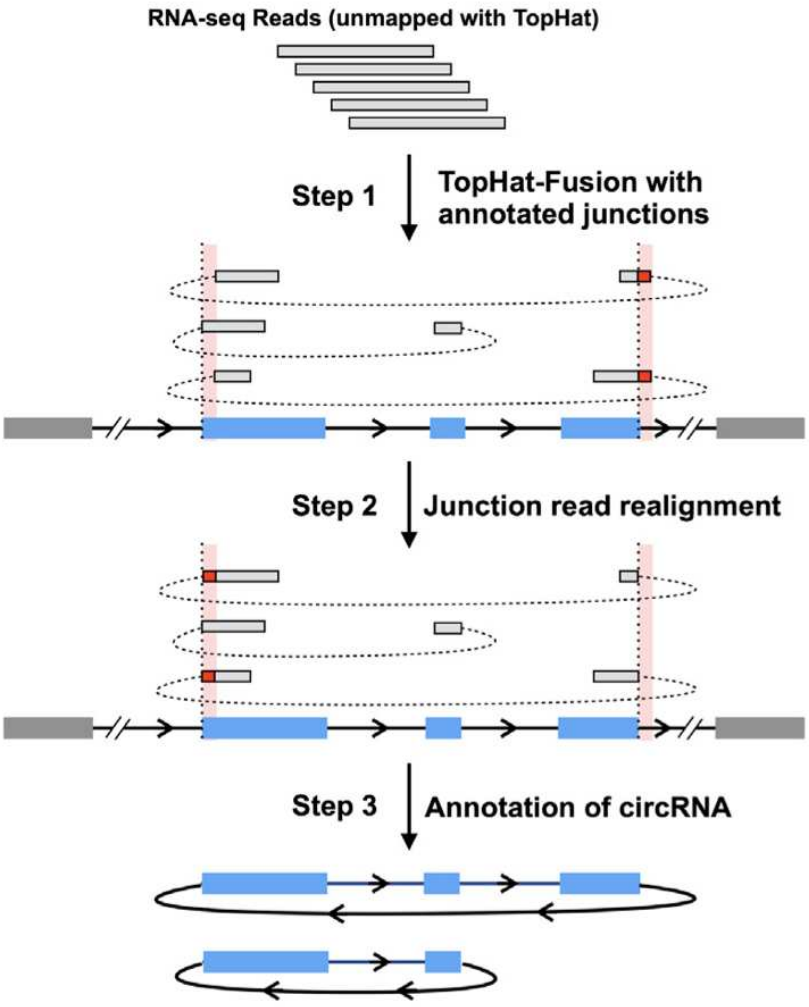


Figura 18: Pipeline dell' algoritmo CIRCexplorer[105]

Gli strumenti di allineamento che sono stati integrati in CIRCexplorer2 permettono alla pipeline si essere più flessibile. La combinazione di diversi strumenti di allineamento potrebbe inoltre fornire una migliore predizione del back-splicing.

Durante la fase di validazione di CIRCexplorer2 sono stati utilizzati data di RNA-seq derivanti da cellule staminali embrionali H9 R e cellule carcinoma ovarico con e senza trattamento con RNase. Queste *reads* sono state mappate inizialmente sul genoma di riferimento umano hg19 e le letture non mappate sono state quindi estratte e allineate nuovamente sul genoma hg19 (con i parametri --fusion-search --keep-fasta-order --bowtie1 --no-coverage-search).

Per gli altri aligner, gli stessi dataset di RNA-seq sono stati allineati utilizzando diversi aligner con parametri specifici.

Nella fig.19 sono mostrati il numero di circRNAs trovati con i diversi tipi di tool di allineamento.

Aligner	Algorithm	Compatible with Cufflinks	Memory consumption	# of circular RNAs with mapped fusion reads ≥ 1	# of circular RNAs with mapped fusion reads ≥ 2
STAR	suffix arrays	partial	~28G	circRNAs: 11,155	4,199
				ciRNAs: 1,521	334
segemehl	suffix arrays	poor	~70G	circRNAs: 12,871	5,030
				ciRNAs: 3,263	824
MapSplice	FM index	poor	~5G	circRNAs: 4,609	4,609
				ciRNAs: 86	86
TopHat 2 & TopHat-Fusion	FM index	perfect	~3G	circRNAs: 9,957	5,082
				ciRNAs: 983	327

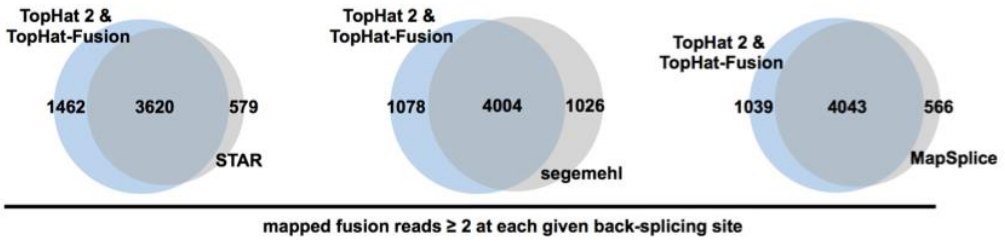


Figura 19: confronto tra i differenti tool di allineamento utilizzati da CIRCexplorer2[68]

Nella nuova versione dell'algoritmo, le *reads* di RNA-seq che hanno allineamenti al genoma e alle giunzioni esoniche collineari non sono semplicemente scartate, ma vengono ulteriormente assemblate in modo de novo. Questo passo di assemblaggio ha consentito l'identificazione di nuovi esoni ed eventi di splicing. Infine, le letture che non si allineano correttamente con TopHat ma che hanno un allineamento corretto con TopHat-Fusion vengono riallineate utilizzando sia le annotazioni esistenti che quelle nuove. Questo ha permesso di identificare le giunzioni di back-splicing da esoni che erano già annotati e da quelli nuovi nell'annotazione.

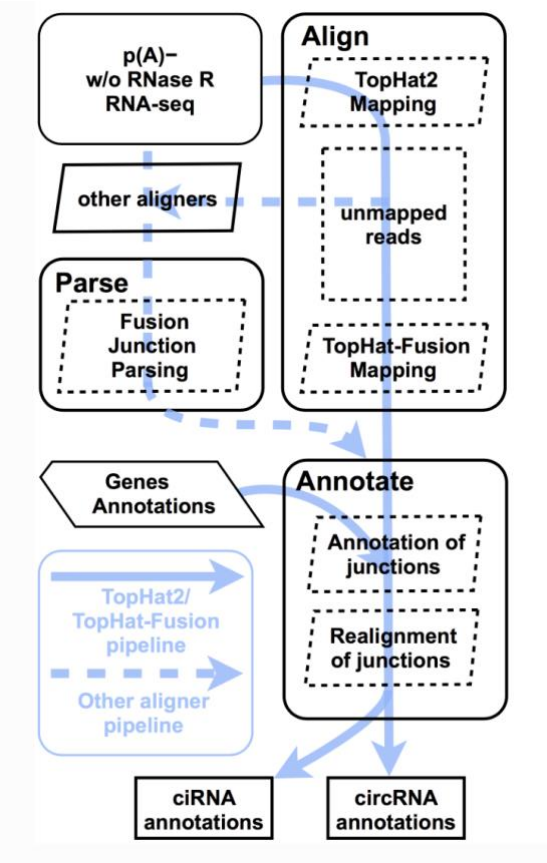


Figura 20: schema dell'algoritmo CIRCexplorer2

Implementazione pratica: comandi e istruzioni

CIRCexplorer2 è costituito da cinque moduli, ognuno dei quali funge da entità indipendente con un ruolo specifico e ben definito. Tuttavia, questi moduli interagiscono tra loro e le diverse pipeline di analisi dell'RNA circolare sono composte da varie combinazioni di tali moduli.

Il primo modulo, noto come “align”, gestisce l'allineamento delle *read* al genoma di riferimento. Inizialmente, viene utilizzata la pipeline combinata TopHat2/TopHat-Fusion per effettuare l'allineamento e recuperare le *read* delle giunzioni di fusione non lineari.

Nel contesto di questo lavoro di tesi, è stata effettuata un'operazione di allineamento manuale per garantire la coerenza con l'uso precedente dello strumento CIRI2. Per le *read* single-end, sono stati utilizzati i file .sam generati dall'algoritmo BWA MEM, mentre per le *read* paired-end è stato impiegato l'algoritmo STAR. Per l’allineamento con STAR, è necessario fornire gli indici del genoma di riferimento e le *read* da allineare. La generazione degli indici di riferimento è stata eseguita seguendo le istruzioni fornite nel manuale di STAR[106] . Durante questa fase, vengono fornite le sequenze di riferimento del genoma e le annotazioni (file GTF), e l'algoritmo crea gli indici di riferimento, che vengono salvati su disco. Gli indici devono essere generati una sola volta per ogni combinazione di genoma e annotazione.

Una volta che sono stati forniti gli input richiesti, STAR effettua l'allineamento delle *read* al genoma e produce diversi file di output. Nella pipeline di CIRCexplorer2, il risultato dell'allineamento è un file Chimeric.out.junction quando si utilizza STAR o un file .sam quando si utilizza BWA. Questi file costituiranno rispettivamente gli input per il modulo successivo della pipeline di CIRCexplorer2.
Il secondo modulo di CIRCexplorer2, denominato "parse," analizza le informazioni sulle giunzioni a partire dai risultati ottenuti con i diversi strumenti di allineamento, al fine di preparare i file necessari per le analisi successive. Durante questa fase, viene creato automaticamente un file denominato back_spliced_junction.bed che è necessario per le analisi successive.
Nel terzo step, noto come “annotate”, si effettua un confronto tra il file back_spliced_junction.bed e il file di annotazione dei geni (hg19_ref_all.txt) al fine di determinare con precisione i confini degli RNA circolari. Inoltre, vengono eseguiti riallineamenti per correggere alcuni allineamenti errati. Il modulo CIRCexplorer2 "annotate" crea un file di output denominato circularRNA_known.txt contenente informazioni sugli RNA circolari noti.
Ci sono due ulteriori moduli, denominati “assemble” e “denovo”. In particolare, CIRCexplorer2 assemble utilizza Cufflinks per effettuare l'assemblaggio de novo dei trascritti di RNA circolari e caratterizza lo splicing alternativo in base ai risultati dell'assemblaggio. È quindi il passo fondamentale prima di analizzare il panorama del back-splicing alternativo e dello splicing alternativo degli RNA circolari. CIRCexplorer2 denovo analizza invece i risultati dell'assemblaggio de novo di RNA circolari per identificare nuovi circRNA e caratterizzare vari eventi di splicing alternativo.
Anche in questo caso ho utilizzato degli script bash (Figura 21-22) che mi permettono di effettuare tutti i passaggi con una sola riga di codice.

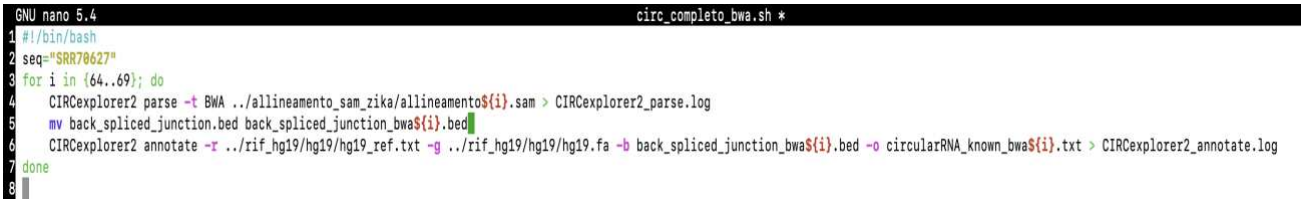


Figura 21: script algoritmo CIRCexplorer2 automatizzato che permette di scorrere tra le reads e effettuare tutti I passaggi dell’algoritmo in un solo run, questa pipeline si può utilizzare quando si utilizza come tool di allineamento BWA-MEM(reads single-end)

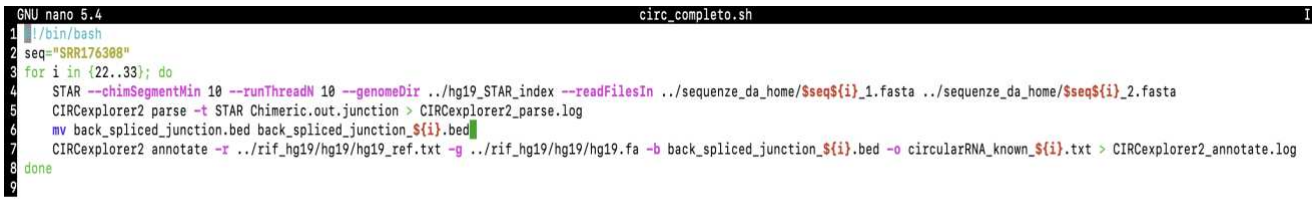


Figura 22:script algoritmo CIRCexplorer2 automatizzato che permette di scorrere tra le reads e effettuare tutti i passaggi dell'algoritmo in un solo run, questa pipeline si può utilizzare quando si utilizza come tool di allineamento STAR (reads single-end)

Circall – Circall Simulator

Circall è uno strumento innovativo per identificare i circRNA dai dati di RNA-seq.

L'algoritmo considera attentamente la probabilità di ottenere risultati falsi positivi in diverse parti o regioni dei dati di RNA-seq, e prende in considerazione molteplici aspetti dei dati per identificare i circRNA in modo più accurato.

Il processo di identificazione dei circRNA mediante Circall si compone di due fasi principali: la prima fase riguarda la scoperta dei potenziali candidati circRNA, mentre la seconda fase consiste in una valutazione statistica.

Nella prima fase, tutte le letture di input vengono mappate sul trascrittoma annotato per eliminare le letture provenienti da trascritti lineari ed estrarre quelle non mappate. Successivamente, queste ultime vengono mappate su un database di riferimento BSJ pre-costruito a partire dal trascrittoma annotato, al fine di individuare letture a supporto di BSJ e potenziali candidati circRNA. Infine, queste letture a supporto di BSJ vengono mappate su pseudo-sequenze di circRNA[8] e possibili RNA tandem[9], per escludere letture FP e generare un elenco di candidati circRNA. Questi vengono quindi valutati e classificati statisticamente in base ai loro tassi di falsi positivi locali bidimensionali.

La procedura di individuazione dei candidati circRNA si basa sull'identificazione delle BSJ. Circall rileva i circRNA attraverso un approccio basato su un riferimento. In questo processo, le sequenze di tutti i BSJ di tutti i potenziali circRNA esonici vengono generate a partire dall'annotazione dei geni e utilizzate come riferimento per l'allineamento delle letture. Circall fa uso del velocissimo strumento di quasi-mappatura RapMap[107] per eseguire l'allineamento delle letture.

[8] La pseudo-sequenza di un circRNA viene generata aggiungendo le ultime L - 1 basi della sequenza del circRNA all'inizio della sequenza stessa del circRNA, dove L è la lunghezza delle letture dei dati di RNA-seq. Per i candidati circRNA con più di 2 esoni, si utilizza l'informazione dello splicing alternativo dai trascritti lineari. In caso contrario, le pseudo-sequenze dei circRNA e delle RNA tandem vengono raccolte dalle sequenze di tutti gli esoni costituenti.

[9] L'RNA tandem è una sequenza simulata che contiene sequenze di circRNA affiancate o sovrapposte alle sequenze di RNA lineari.

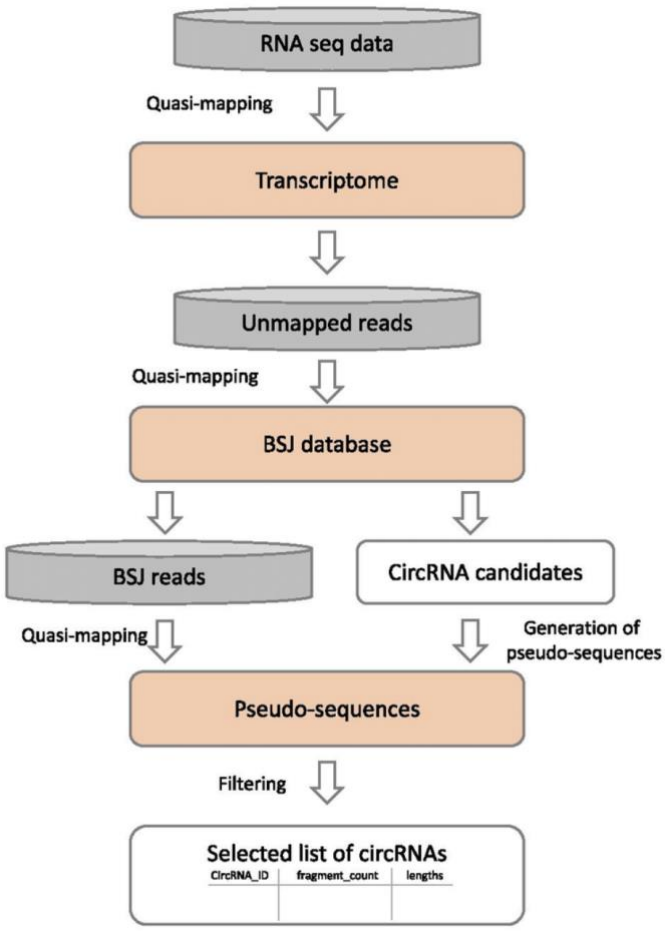


Figura 23: schema algortimo Circall

Nel presente lavoro di tesi è stato utilizzato Circall simulator, una pipeline creata per la simulazione di dati RNA-seq contenenti giunzioni back-spliced, RNA tandem e RNA lineari. Lo script Circall simulator può essere eseguito sulla console R e richiede diversi input tra cui un elenco di candidati circRNA e i conteggi desiderati delle letture dei circRNA.
In aggiunta, è necessaria una completa annotazione genomica, che comprenda un riferimento genomico, un riferimento del trascrittoma e un file di annotazione dei geni in formato Sqlite, al fine di acquisire i modelli genici e le sequenze degli esoni. È importante sottolineare che gli eventi di splicing alternativo rappresentano un fenomeno ubiquitario nella formazione dei circRNA. In situazioni in cui più di un trascritto lineare contribuisce al supporto della BSJ di un circRNA, si adotta una selezione casuale di uno dei trascritti lineari, utilizzando i suoi esoni per la procedura di simulazione.
Il pacchetto di simulazione dà inizio alla suddivisione dell'elenco dei circRNA in due liste separate, le quali verranno sfruttate per simulare i dati relativi ai circRNA e ai tandem RNA in accordo con il tasso di tandem preimpostato dall'utente. Le liste dei candidati vengono quindi sottoposte a due sotto-processi all'interno del pacchetto al fine di simulare i dati di sequenziamento relativi ai circRNA e ai tandem RNA. Ciascun sotto-processo presenta due fasi chiave nel processo di simulazione: l'ottenimento delle sequenze dei trascritti e la

simulazione di coppie di letture sintetiche mediante l'utilizzo di Polyester. Polyester è uno strumento solitamente impiegato per la generazione di RNA lineari di tipo wild-type. Attraverso questo pacchetto, il simulatore di Circall offre agli utenti numerose opzioni per gli esperimenti di simulazione, tra cui il tasso di errore di sequenziamento, la lunghezza delle letture e la distribuzione delle lunghezze dei frammenti.

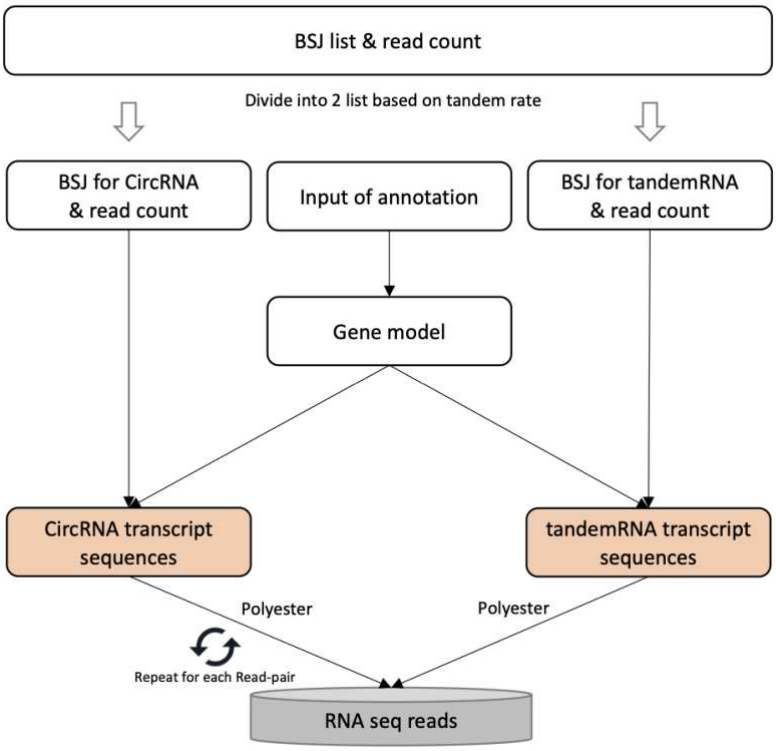


Figura 24: schema di simulazione dei dati circRNA e tandem RNA seq

A causa della forma circolare dei circRNA viene adottata una procedura per emulare il processo di rottura dei circRNA in forme lineari. Le sequenze degli esoni raccolte vengono preliminarmente concatenate in modo sequenziale, assumendo una forma lineare e successivamente vengono divise casualmente in due sequenze separate. Queste sequenze vengono quindi nuovamente concatenate nell'ordine inverso al fine di ottenere le sequenze dei circRNA.

Infine, queste sequenze vengono impiegate per generare artificialmente una coppia di letture di RNA sintetico attraverso l'utilizzo del pacchetto Polyester. Questa procedura viene applicata per ciascuna coppia di letture associate al circRNA target.

Per quanto riguarda i tandem RNA, il procedimento è più agevole. Basandosi sulle informazioni relative alle BSJ, le sequenze degli esoni raccolte vengono suddivise in due categorie: la prima corrisponde alla regione del circRNA, mentre la seconda rappresenta la regione non circRNA. Le sequenze dei trascritti vengono ottenute duplicando gli esoni nella regione del circRNA e, successivamente, concatenando tutte le parti in modo congiunto.

Dal momento che i tandem RNA sono già presenti in forma lineare, è possibile simulare direttamente le letture di sequenziamento dell'RNA sintetico mediante l'utilizzo del pacchetto Polyester.

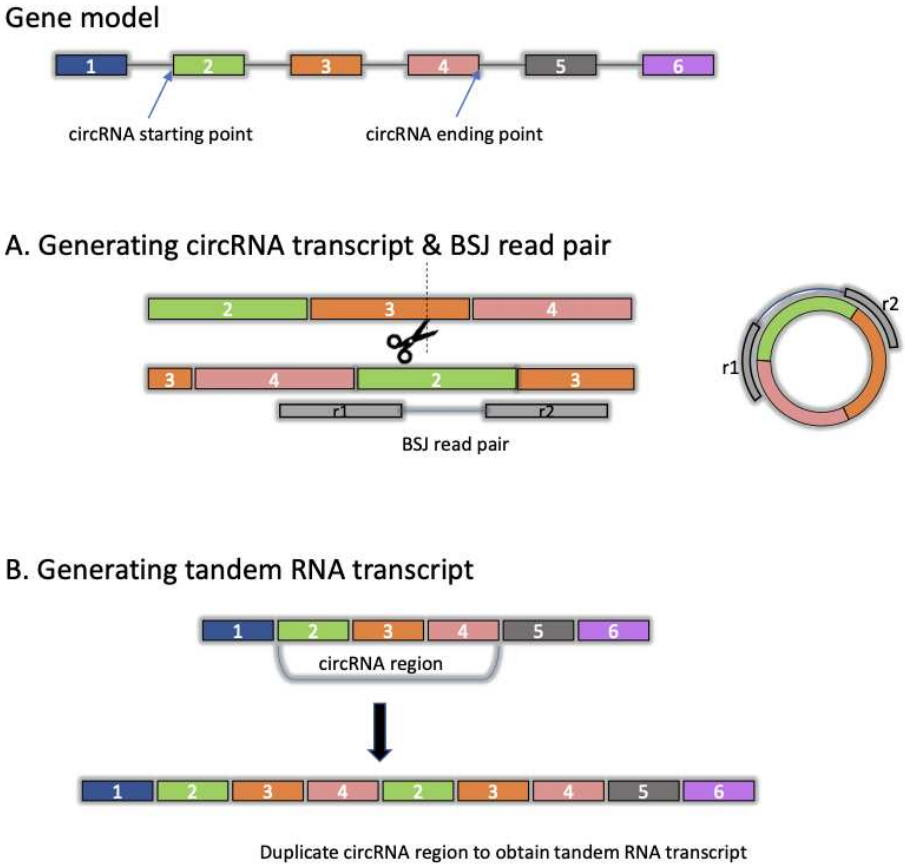


Figura 25: illustrazione della generazione delle pseudosequenze si circRNA e tandem RNA

Implementazione pratica: comandi e istruzioni

Lo script di Circall è implementato mediante l'utilizzo di una funzione R denominata "Circall_simulator", che richiama al suo interno ulteriori funzioni. Questa funzione prevede l'utilizzo di diverse librerie R, tra cui GenomicFeatures, Biobase, BiocManager e Biostrings, che fanno parte della repository open-source Bioconductor.[108,109]

Bioconductor attualmente comprende una vasta gamma di strumenti per l'analisi e l'annotazione di altri dati ad alto rendimento, tra cui quelli derivanti dal sequenziamento di nuova generazione (NGS). [78]

La funzione principale richiede in input un data frame con 6 colonne che contengono informazioni sui circRNA, tra cui il cromosoma di riferimento, le posizioni di inizio e fine della BSJ su quel cromosoma, l'ID del gene contenente il circRNA, il numero di coppie di letture che si desidera generare per il circRNA target e l'FPKM (Fragments Per Kilobase of transcript per Million) dei circRNA target.

L'utente ha anche la possibilità di definire il tasso di errore di sequenziamento (di default pari a 0.05). Inoltre, è necessario fornire il percorso del genoma di riferimento in formato FASTA, il percorso del file FASTA della trascrizione (cDNA) e il percorso della cartella di output. L'esecuzione della funzione principale produce in uscita un file contenente le pseudo-sequenze che rappresentano i circRNA. Questo sistema di simulazione consente di validare i sistemi di rilevamento CIRI2 e CIRCexplorer2 precedentemente descritti.

CIRIquant

Il software CIRI quant è stato sviluppato con l'obiettivo di migliorare la quantificazione accurata dei circRNA e condurre un'analisi diferenziale dell'espressione in maniera più precisa. Questo risultato è ottenuto attraverso l'implementazione di una serie di approcci avanzati.
Innanzitutto, CIRIquant costruisce un riferimento pseudo-circolare che viene utilizzato per riallineare le letture provenienti dai dati di RNA-seq. Questa strategia consente una rappresentazione più accurata dei circRNA, mitigando gli errori di allineamento.
Inoltre, per affrontare i possibili bias introdotti durante il trattamento con l'RNase R, CIRIquant si avvale di modelli statistici sofisticati che contribuiscono a ottenere valori di espressione dei circRNA con una riduzione significativa del tasso di falsi positivi.
Un altro punto di forza di CIRIquant è l'implementazione di una pipeline completa per l'analisi dell'espressione differenziale. Questa pipeline integra due misure indipendenti, che forniscono una valutazione più completa della regolazione dello splicing competitivo tra i circRNA e le loro controparti lineari. Tale approccio offre una visione dettagliata dei cambiamenti nell'espressione dei circRNA in risposta a specifiche condizioni sperimentali.
CIRIquant sfrutta gli strumenti comunemente utilizzati (come CIRI2, CIRCexplorer2, find_circ2, ecc.) per identificare i circRNA. Attraverso l'utilizzo dei risultati di allineamento delle letture rispetto al genoma di riferimento e ai trascritti pseudo-circolari, CIRIquant raggiunge una migliore identificazione delle letture di BSJ in termini di accuratezza e sensibilità, ma consente anche una quantificazione delle BSJ. Inoltre, CIRIquant è dotato di una funzione per eseguire un'analisi dell'espressione differenziale dei circRNA.[110]

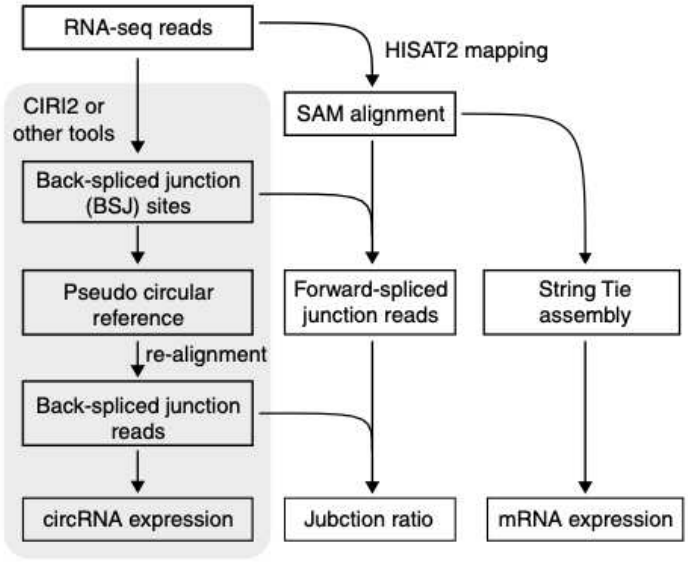


Figura 26: schema quantificazione dei circRNA

Come si può vedere dalla fig.26, le *reads* vengono allineate al genoma di riferimento utilizzando l'algoritmo HISAT2[111] . Parallelamente vengono utilizzati strumenti di rilevazione dei circRNA come CIRI2, per identificare i circRNA di interesse.

Al fine ottenere una valutazione accurata dei livelli di espressione dei circRNA e per minimizzare la presenza di falsi positivi nelle letture BSJ, l'algoritmo implementa una procedura che coinvolge la generazione di una sequenza di riferimento pseudo-circolare. Tale sequenza pseudo-circolare si ottiene concatenando due sequenze complete corrispondenti alla regione di BSJ. Successivamente, le *reads* circolari candidate vengono sottoposte a un nuovo allineamento utilizzando la sequenza pseudo circolare come riferimento. Le letture di BSJ vengono confermate solo se possono essere allineate in modo completo e lineare alla regione di BSJ.

Per determinare la percentuale di letture di giunzione relative allo splicing circolare, i risultati dell'allineamento rispetto al genoma di riferimento vengono combinati con quelli ottenuti utilizzando le sequenze pseudo di riferimento.

Nel caso di dati di RNA-seq trattati con RNase R, i valori di espressione delle BSJ dei circRNA non possono essere utilizzati direttamente per analisi comparative a causa dell'efficienza non uniforme del trattamento con RNase R in diversi studi. Pertanto, è stato implementato un modello gaussiano per adattare la distribuzione dell'efficienza del trattamento con RNase R, e successivamente è stato utilizzato il modello adattato come distribuzione posteriore per la correzione dei coefficienti di RNase R.

CIRIquant permette di valutare sia l'espressione differenziale (DE) che lo splicing differenziale (DS) dei circRNA nei campioni di caso e controllo.

Quando non sono disponibili repliche biologiche si effettua una misura che tiene conto sia delle differenze medie nell'espressione delle circRNA tra due condizioni o gruppi di campioni, sia della varianza dei dati all'interno di ciascun gruppo.
Quando sono disponibili repliche biologiche, si può eseguire un test statistico per valutare la significatività del cambiamento nei valori di espressione dei circRNA.
CIRIquant effettua una normalizzazione che assicura che le differenze siano dovute ai circRNA e non ad altri fattori come il numero di letture o piccole variazioni nei campioni. Applicando inoltre dei metodi matematici, CIRIquant ci permette di capire con un livello abbastanza elevato di accuratezza che le differenze siano reali e significative.

Implementazione pratica: comandi e istruzioni

CIRI-quant richiede diversi prerequisiti e può essere eseguito esclusivamente attraverso il linguaggio di programmazione Python2.
Tuttavia, poiché la versione più recente di Python non è compatibile, è stato necessario utilizzare un ambiente virtuale Conda. Ho proceduto alla creazione di un file di configurazione nel formato YAML, all'interno del quale sono stati specificati i percorsi per i tool impiegati dall'algoritmo e i percorsi relativi al genoma di riferimento utilizzato.
In termini di input, CIRIquant richiede il suddetto file di configurazione e le sequenze di lettura. L'output principale consiste in un file GTF, che offre una dettagliata annotazione riguardo alle giunzioni BSJ e FSJ dei circRNA, nonché l'annotazione delle regioni circolari back-spliced nelle colonne degli attributi.
Qualora si desideri quantificare i circRNA ottenuti tramite altri strumenti, è possibile aggiungere un file nel formato BED attraverso l'opzione "--bed". Allo stesso modo, è possibile fornire l'output dell'altro strumento che si intende quantificare utilizzando l'opzione "--circ" seguita dal nome dello strumento dopo l'opzione "--tool". È da notare che CIRIquant è in grado di elaborare i risultati di vari strumenti, tra cui CIRI2 e CIRCexplorer2, che sono stati impiegati nell'ambito di questa ricerca.

Genomi di riferimento e dataset utilizzati

L'obiettivo primario del Progetto del Genoma Umano consiste nella creazione di un genoma di riferimento rappresentativo, che sia accurato e completo, al fine di riflettere la vasta diversità genetica all'interno della popolazione umana.
Tuttavia, raggiungere tale obiettivo è un'impresa complessa a causa della notevole variabilità genetica presente in specifiche regioni del genoma umano e dal fatto che tali genomi di riferimento vengono costruiti utilizzando sequenze di DNA provenienti da diversi donatori individuali.

Un esempio illustrativo di questa complessità è rappresentato dal genoma di riferimento GRCh38/hg38, il quale è stato ottenuto mediante l'utilizzo di oltre 60 librerie di cloni genomici.
I genomi di riferimento sono disponibili per numerose specie e solitamente vengono utilizzati come punto di partenza per la costruzione di nuovi genomi.
Nel contesto del presente studio di ricerca, sono stati utilizzati due genomi di riferimento umani distinti, ossia HG19 e HG38, sviluppati dal Genome Reference Consortium.
Questi genomi di riferimento svolgono un ruolo cruciale nell'ambito della mappatura dei geni umani e forniscono un fondamento essenziale per l'esecuzione di analisi bioinformatiche e filogenetiche approfondite.
Va sottolineato che esistono notevoli differenze tra questi due tipi di genomi di riferimento, in quanto il genoma HG38 incorpora una maggiore quantità di varianti genomiche inclusi SNP ed è sottoposto ad aggiornamenti continui, il che implica che i risultati ottenuti mediante l'utilizzo di un genoma di riferimento (come ad esempio HG38) non sono direttamente comparabili con quelli ottenuti mediante l'utilizzo dell'altro genoma (come ad esempio HG19). Questa variazione genetica può avere un impatto significativo sui risultati delle analisi condotte.[112]
Nei vari strumenti utilizzati per la rilevazione dei circular RNAs, i file del genoma di riferimento sono richiesti come input e sono generalmente utilizzati per l'allineamento iniziale dei dati, nei formati gtf, fasta e txt.
É possibile accedere ai genomi di riferimento online attraverso vari siti web, utilizzando appositi browser come Ensemble o UCSC Genome Browser, è inoltre possibile eseguire il download al link (https://www.ncbi.nlm.nih.gov/datasets/genome) o scaricarli da terminale tramite uno script contenuto nella pipeline CIRCexplorer2 che può scaricare e formattare il file di annotazione dei geni e il file della sequenza del genoma di riferimento per la specie umana e murina[113].
Le variazioni non sono solo di tipo genetico ma anche nella struttura dei file; per questo motivo, durante l'impiego di CIRCexplorer2, è stato necessario apportare una modifica specifica al file che rappresenta il genoma HG38, scaricato con un altro link rispetto a quello suggerito dalla pipeline.
Si verifica infatti una disparità nella configurazione della prima colonna tra i due riferimenti, il che genera una incompatibilità nell'impiego di CIRCexplorer2.
Questa divergenza nella struttura della colonna di input provoca una mancata produzione di risultati da parte del software, poiché CIRCexplorer2 richiede un formato specifico che non è soddisfatto da Hg38.

```python
#!/usr/bin/env python3
def main():
    chr = "chr"

    with open('../Homo_sapiens.GRCh38.dna_sm.primary_assembly.fa', 'r') as input_file, open('output.fa', 'w') as output_file:
        for line in input_file:
            if line.startswith('>'):
                line_parts = line.split(' ')
                number = line_parts[0][1:]
                modified_line = '>' + chr + number + '\n'
                output_file.write(modified_line)
            else:
                output_file.write(line)

if __name__ == '__main__':
    main()
```

Figura 27: script che modifica il genoma di riferimento Hg38 per renderlo utilizzabile nell'algoritmo CIRCexplorer2

Per l'analisi e la detection dei circRNA sono stati utilizzate delle sequenze appartenenti a 3 diversi dataset.

Il primo gruppo di *reads* RNA-seq è relativo al trascrittoma di cellule staminali neurali (hNSC) umane infettate con il virus Zika (ZIKV). Le hNSC sono state infettate con ZIKV MR766 o Paraiba per 3 giorni a un MOI di 1. L'RNA totale è stato analizzato mediante RNA-seq o miRNA-seq per identificare reti di geni regolati da miRNA implicati nella patogenesi di ZIKV.

L'RNA totale è stato isolato utilizzando il Qiagen Rneasy Minikit[10] e le librerie di RNA sono state preparate per il sequenziamento seguendo i protocolli standard Illumina.[114]

Il secondo gruppo di *reads* RNA-seq riguarda il trascrittoma di organoidi cerebrali derivati da cellule staminali indotte pluripotenti (iPSC) in seguito all'infezione da citomegalovirus. Nello studio di riferimento gli organoidi infettati sono stati separati in base all'espressione di GFP 14 giorni dopo l'infezione, ottenendo 3 popolazioni distinte. Dopo la separazione, le cellule sono state centrifugate e da ciascuna popolazione è stato isolato l'RNA. Infine, dopo l'isolamento e la quantificazione dell'RNA, è stato utilizzato il kit NEBNext Poly(A) mRNA Magnetic Isolation Module per generare librerie di cDNA.

L'ultimo campione rappresenta una libreria arrichita per i circRNA, eliminando gli RNA lineari in ogni campione. L'RNA totale è stato trattato con l'enzima RNasi R ed è stato poi purificato tramite il kit RNeasy. Tutte le librerie sono poi state sottoposte a sequenziamento e le letture grezze ottenute sono state ulteriormente filtrate con il software fastp[115] per ottenere *reads* di alta qualità.

[10] I kit RNeasy rappresentano la metodologia di riferimento per l'estrazione di RNA totale. Garantiscono una veloce purificazione di RNA di elevata qualità da varie quantità di cellule e tessuti.[118]

Risultati

Il seguente capitolo presenta in dettaglio la pipeline che è stata sviluppata durante il lavoro di tesi, per simulare dataset di RNA-seq e poter testare su questi dati gli algoritmi di detection e quantificazione dei circRNA prima dell'applicazione su dati sperimentali.

La pipeline è stata sviluppata per simulare dati di RNA seq con un focus specifico sulle *read* di circRNA. La simulazione comprende la generazione sia di *read* circRNA che di *read* randomiche di trascrittoma, al fine di rappresentare realisticamente la complessità dei dati sperimentali.

Nello script utilizzato sono presenti, inoltre, i comandi di tutti i tool utilizzati (CIRI2, CIRCexplorer2 e CIRIquant) e un confronto automatizzato sugli output di questi tool per valutare l'affidabilità della detection.

Simulatore

È stato sviluppato un nuovo simulatore per le *reads* contenenti circRNA a causa delle limitazioni riscontrate nell'utilizzo di Circall, il simulatore precedentemente menzionato nei Materiali e metodi. Circall operava in un ambiente locale basato su R e, quando impiegato con dataset di notevole dimensione, risultava essere soggette a un malfunzionamento dovuto all'eccessivo utilizzo di memoria.

La pipeline che è stata sviluppata consente una rappresentazione accurata e controllata dei dati di RNA-seq simulati, considerando diversi parametri configurabili.

Per simulare i circRNA, è necessario selezionare un numero specifico di circRNA da una banca dati interna precedentemente creata. Questa banca dati è stata generata "in-house" a partire dai file in formato GFF, dai quali sono state estratte le coordinate degli esoni per ciascun trascritto. Da queste coordinate, sono state generate tutte le possibili combinazioni di esoni contigui. Tuttavia, le combinazioni relative a esoni singoli, sebbene possibili dal punto di vista teorico, sono state escluse dalla simulazione, in quanto nella pratica risultano solitamente brevi. Se ad esempio si ha un gene con 4 esoni, le combinazioni risultati saranno 1-2, 2-3, 3-4,1-2-3,2-3-4,1-2-3-4.

In seguito, queste combinazioni di esoni sono state divise in due parti e unite in una configurazione "testa-coda" al fine di creare sequenze con una BSJ al centro, necessarie al simulatore a seguire.

Per ciascun circRNA selezionato dalla banca dati, viene scelto casualmente un valore all'interno di un intervallo specificato dall'utente che determina il numero di *reads* da generare per ciascun circRNA. Viene salvato, per scopi di confronto e valutazione successiva, un file contenente identificatori univoci per i circRNA e le rispettive quantità.

La BSJ viene assegnata casualmente ad ogni *read* simulata, con circa il 50% di probabilità di finire nella parte forward o reverse della *read paired-end*.
La posizione in cui inizia la *read* varia in modo da garantire una diversificazione delle sequenze generate. La posizione di partenza viene selezionata casualmente all'interno di un intervallo compreso tra un terzo e due terzi della lunghezza della *read* stessa, garantendo in questo modo che entrambi i lati della giunzione siano rappresentati in modo adeguato, occupando almeno un terzo della lunghezza totale della *read.*
La *mate read* associata viene determinata in base alla lunghezza media dei frammenti da simulare, un parametro controllato dall'utente. Ad esempio, se la lunghezza del frammento è pari a 500 nucleotidi e quella delle *reads* 150 nucleotidi, l'inizio della *mate read* risulta 350 nucleotidi a monte della prima *read* e termina 200 nucleotidi prima della fine della prima *read.* Per ciascuna sequenza generata, viene assegnato uno score di qualità predefinito (punteggio Phred 30).
In aggiunta alla generazione delle *reads* contenenti BSJ per simulare circRNA, è stato implementato un processo per simulare *reads* che non presentano BSJ. Questo passaggio serve per creare un rumore di fondo all'interno del dataset simulato di RNA-seq. La simulazione di queste *reads* si basa sul trascrittoma di riferimento.
Si effettua il campionamento di una parte dei trascritti dal trascrittoma di riferimento. La selezione dei trascritti è guidata dalla lunghezza dei frammenti determinata precedentemente, rispettando un fattore moltiplicativo specificato rispetto all'intervallo predefinito per le *reads* circolari. Questo processo di campionamento mira a riflettere la diversità dei trascritti presenti nel trascrittoma di riferimento e ad assicurare la presenza di una varietà di lunghezze di frammenti nei dati simulati.
In modo analogo a quanto fatto per le *reads* contenenti BSJ, vengono generate coppie di *reads* forward e reverse alle estremità dei frammenti generati. Queste coppie di *reads* vengono poi integrate nel dataset simulato in formato fastq per le *reads paired-end*.
L'inclusione di queste *reads* senza BSJ permette di simulare il contesto completo di un dataset di RNA-seq, contribuendo a rappresentare realisticamente i dati biologici sperimentali in cui non tutti i trascritti saranno circolari. La combinazione di *reads* contenenti BSJ e *reads* senza BSJ offre una visione completa dell'espressione genica all'interno del campione simulato e consente una valutazione accurata degli algoritmi e delle analisi applicate ai dati di RNAseq simulati.

Generazione dei circRNA simulati e successiva rilevazione

Nel presente lavoro di tesi, sono state simulate le *reads* con parametri differenti e sono stati utilizzati vari tool di detection che sono stati confrontati tra loro per valutarne l'affidabilità. In particolare, per la simulazione delle *reads*, sono state variate le seguenti condizioni: il numero di circRNA tra 100, 1000 e 10000, la lunghezza delle letture è stata impostata a 75 e 150 basi. Per quanto riguarda il numero di *reads* per ciascun circRNA, sono stati considerati i range 1-10, 11-100 e 101-1000. La lunghezza dei frammenti è stata mantenuta costante a 500 basi, e il valore moltiplicativo per il rumore di fondo è stato impostato pari a 10. Per ogni run della pipeline si hanno 4 timepoint che vanno a simulare l'andamento dell'espressione dei circRNA in quattro istanti differenti.

È stato effettuato un confronto tra i circRNA trovati da ogni tool considerando le posizioni di inizio e fine dei circRNA che sono stati mappati nel genoma di partenza e sono stati ricondotti agli esoni tra cui avviene la circolarizzazione. La pipeline tiene nota degli ID dei circRNA simulati. Gli ID sono formati a partire dal codice identificativo del trascritto e contengono anche l'informazione sugli esoni di inizio e fine dei circRNA. Mentre nell'output di CIRI2 e CIRIquant è presente l'informazione sul trascritto contenente la BSJ, CIRCexplorer2 fornisce in output solo le posizioni genomiche, rendendo necessario un controllo su tali indici per poter effettuare una comparazione tra i tool.

1	circID	# simulated circRNAs reads	#circRNAs reads identified by CIRI2	circRNAs identified by CIRCexplorer	#circRNAs reads identified by CIRIquant
2	NM_000014.6:2-21	334	334	NF	1
3	NM_000023.4:1-3	738	738	NF	738
4	NM_000036.3:5-8	970	970	NF	970
5	NM_000037.4:32-38	586	586	NF	586
6	NM_000038.6:1-13	695	695	NF	1
7	NM_000051.4:39-62	294	294	NF	294
8	NM_000057.4:1-3	506	506	NF	506
9	NM_000059.4:1-16	478	478	NF	NA
10	NM_000065.5:4-18	694	694	NF	NA
11	NM_000068.4:9-29	159	159	NF	1
12	NM_000069.3:22-33	118	118	NF	118
13	NM_000070.3:5-22	884	884	NF	1
14	NM_000081.4:13-28	579	579	NF	1
15	NM_000084.5:2-10	501	501	NF	1
16	NM_000088.4:11-33	954	954	NF	954
17	NM_000089.4:28-37	903	903	NF	629
18	NM_000090.4:21-41	431	431	NF	431
19	NM_000091.5:10-47	678	678	NF	678
20	NM_000092.5:22-28	751	751	NF	1
21	NM_000093.5:46-52	695	695	NF	695
22	NM_000094.4:7-19	765	765	NF	1
23	NM_000095.3:4-16	493	493	NF	1
24	NM_000097.7:2-7	194	194	NF	191
25	NM_000109.4:27-78	587	NA	NF	NA
26	NM_000110.4:2-22	860	NA	NF	NA
27	NM_000111.3:7-9	962	962	NF	962
28	NM_000128.4:5-13	357	357	NF	1
29	NM_000129.4:2-8	972	972	NF	972
30	NM_000132.4:1-9	221	221	NF	221
31	NM_000135.4:15-21	801	801	NF	1
32	NM_000136.3:1-11	645	NA	NF	NA
33	NM_000137.4:5-9	470	470	NF	470
34	NM_000138.5:36-38	738	738	NF	738

Tabella 1: output di confronto tra i tool

La tabella 1 mostra una parte dell'output della pipeline su dati simulati con i seguenti parametri: numero di circRNAs pari a 10000, lunghezza delle *reads* pari a 150 e il numero di *read* per circRNA in un range tra 101 e 1000.
Nella prima colonna ci sono gli ID dei circRNAs che sono stati simulati. Gli ultimi due numeri dell'ID, separati da un trattino, indicano i due esoni che circolarizzano (la BSJ sarà formata dall'inizio del primo esone e la fine del secondo). Nella seconda colonna troviamo le informazioni fornite dalla pipeline riguardo la quantità di *reads* contenenti BSJ generate per ciascuno dei circRNA. Le ultime tre colonne rappresentano invece gli output dei tre tool utilizzati.
Nelle colonne corrispondenti agli output di CIRI2 e CIRIquant è presente la quantificazione per ogni circRNA, mentre l'etichetta "NA" in queste colonne rappresenta i circRNA presenti nelle *reads* ma non identificati dai tool. CIRCexplorer2 si concentra solo sulla rilevazione dei circRNA ma non sulla loro quantificazione, per cui in questo caso, "NF" rappresenta l'individuazione dei circRNA, mentre "NA" indica che quel circRNA non è stato riconosciuto dal tool.
Nella tabella 2 sono riportati i risultati della detection dei vari tool nelle simulazioni. In particolare, è presente il numero di circRNAs trovati da ogni tool, il numero dei veri positivi, il numero dei falsi positivi, quello dei falsi negativi e i corrispondenti valori di precisione e *recall* su sulle *read* simulate con i diversi parametri.
Il numero dei falsi negativi corrisponde al numero degli NA, il numero dei veri positivi è calcolato invece contando quanti dei circRNAs trovati dai tool sono quelli effettivamente simulati. Il numero dei falsi positivi si ottiene sottraendo al totale dei circRNAs identificati da ciascun tool il numero di veri positivi.
Per il calcolo della precisione e *recall* sono state utilizzate le seguenti formule: (TP/TP+FP) e (TP/TP+FN), rispettivamente.
L'etichetta dell'esperimento da informazione sui diversi parametri di configurazione: il primo numero rappresenta il numero di circRNAs simulati (10000, 1000 o 100), il secondo la lunghezza delle reads e il terzo numero rappresenta il numero di *read* per circRNA (es. quando è pari a 10 il range è 1-10).

Exp 10000-150-1000	#circRNAs	TP	FP	FN	Precision	Recall
CIRI2	10644	9445	1199	555	0,887354378053363	0,9445
CIRIquant	8446	8507	0	1493	1	0,8507
CIRCexplorer	10628	9419	1209	581	0,886243884079789	0,9419
Exp 10000-150-100						
CIRI2	10217	9405	812	595	0,920524615836351	0,9405
CIRIquant	8393	8447	0	1553	1	0,8447
CIRCexplorer	9929	9356	573	644	0,94229026085205	0,9356
Exp 3 10000-150-10						
CIRI2	8365	8307	58	1693	0,993066347878063	0,8307
CIRIquant	8016	8092	0	1908	1	0,8092
CIRCexplorer	9363	9266	97	734	0,989640072626295	0,9266
Exp 1000-150-1000						
CIRI2	1123	952	171	48	0,847729296527159	0,952
CIRIquant	879	867	12	133	0,986348122866894	0,867
CIRCexplorer	1143	940	203	60	0,822397200349956	0,94
Exp 1000-150-100						
CIRI2	1015	930	85	70	0,916256157635468	0,93
CIRIquant	867	851	16	149	0,981545559400231	0,851
CIRCexplorer	1083	927	156	73	0,85595567867036	0,927
Exp 1000-150-10						
CIRI2	817	796	21	204	0,974296205630355	0,796
CIRIquant	839	827	12	173	0,98569725864124	0,827
CIRCexplorer	995	912	83	88	0,916582914572864	0,912
Exp 100-150-1000						
CIRI2	110	91	19	9	0,827272727272727	0,91
CIRIquant	92	84	8	16	0,91304347826087	0,84
CIRCexplorer	170	95	75	5	0,558823529411765	0,95
Exp 100-150-100						
CIRI2	126	90	36	10	0,714285714285714	0,9
CIRIquant	88	76	12	24	0,863636363636364	0,76
CIRCexplorer	181	89	92	11	0,49171270718232	0,89
Exp 100-150-10						
CIRI2	89	84	5	16	0,943820224719101	0,84
CIRIquant	89	80	9	20	0,898876404494382	0,8
CIRCexplorer	150	88	62	12	0,586666666666667	0,88

Exp 10000-75-1000						
CIRI2	10719	9444	1275	556	0,881052336971732	0,9444
CIRIquant	8462	8523	0	1477	1	0,8523
CIRCexplorer	10738	9400	1338	600	0,875395790650028	0,94
Exp 10000-75-100						
CIRI2	10281	9462	819	538	0,920338488473884	0,9462
CIRIquant	8514	8595	0	1405	1	0,8595
CIRCexplorer	9945	9442	503	558	0,949421820010055	0,9442
Exp 10000-75-10						
CIRI2	8422	8345	77	1655	0,990857278556162	0,8345
CIRIquant	8308	8370	0	1630	1	0,837
CIRCexplorer	9341	9236	105	764	0,988759233486779	0,9236
Exp 1000-75-1000						
CIRI2	1078	940	138	60	0,871985157699443	0,94
CIRIquant	887	877	10	123	0,988726042841037	0,877
CIRCexplorer	1386	947	439	53	0,683261183261183	0,947
Exp 1000-75-100						
CIRI2	1024	931	93	69	0,9091796875	0,931
CIRIquant	867	860	7	140	0,991926182237601	0,86
CIRCexplorer	1075	948	127	52	0,881860465116279	0,948
Exp 1000-75-10						
CIRI2	851	816	35	184	0,958871915393655	0,816
CIRIquant	871	862	9	138	0,989667049368542	0,862
CIRCexplorer	1008	943	65	57	0,935515873015873	0,943
Exp 100-75-1000						
CIRI2	111	97	14	3	0,873873873873874	0,97
CIRIquant	93	86	7	14	0,924731182795699	0,86
CIRCexplorer	125	90	35	10	0,72	0,9
Exp 100-75-100						
CIRI2	98	89	9	11	0,908163265306122	0,89
CIRIquant	86	78	8	22	0,906976744186047	0,78
CIRCexplorer	124	93	31	7	0,75	0,93
Exp 100-75-10						
CIRI2	81	78	3	22	0,962962962962963	0,78
CIRIquant	94	87	7	13	0,925531914893617	0,87
CIRCexplorer	118	93	25	7	0,788135593220339	0,93

Tabella 2: risultati della detection sulle reads simulate usando diverse combinazioni di parametri di simulazione

Nella fig. 28 sono rappresentati i risultati, per i tre tool, della detection nelle *read* simulate con un numero di circRNAs pari a 10000, lunghezze delle *reads* pari a 150 e numero di *read* per circRNA nel range 101-1000. La simulazione è stata ripetuta quattro volte a partire dagli stessi circRNAs, in modo da simulare quattro punti temporali relativi allo stesso dataset.

Esperimento 10000-150-1000

Timepoint	CIRI2	CIRCexplor	CIRIquant
1	10644	10628	8446
2	10626	10486	8447
3	10649	10495	8436
4	10687	10717	8447

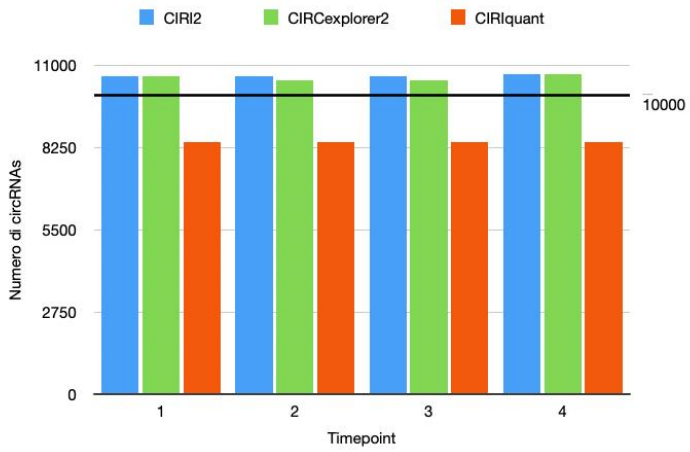


Figura 28: rappresentazione del numero di circRNA trovati dai 3 tool nei 4 timepoint simulati

Nella fig. 29 sono invece presenti le informazioni sulla precisione e la *recall* dei tool relative agli stessi parametri di simulazione.

Precisione e Recall

	FP	FN	TP	Recall	Precision
CIRI2	1199	555	9445	0,9445	0,887
CIRCexplorer	1209	581	9419	0,9419	0,886
CIRIquant	0	1209	8507	0,8507	1

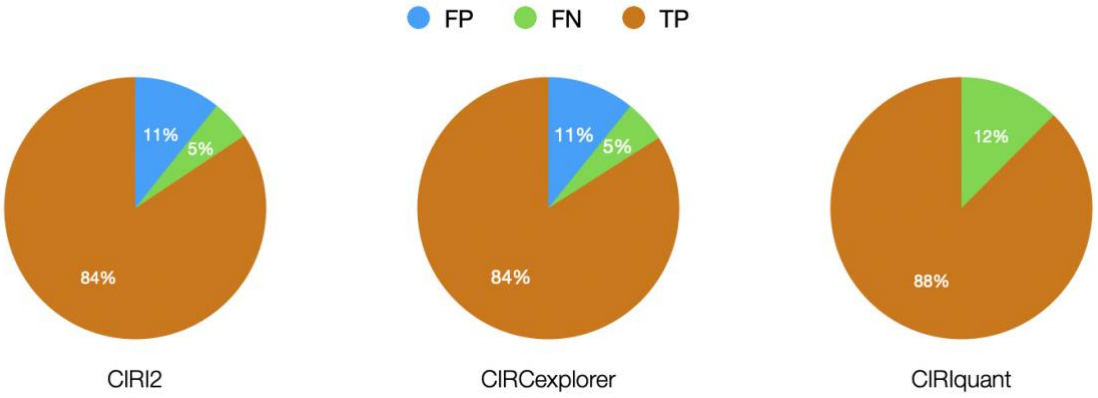


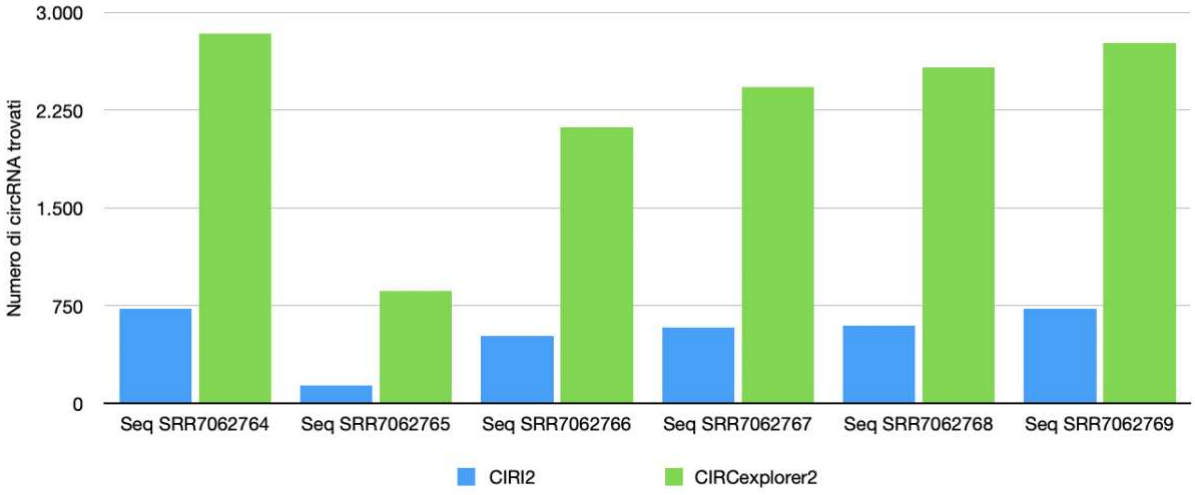
Figura 29:affidabilità dei tool

Detection dei circRNAs sui dati reali

I tool descritti precedentemente sono stati utilizzati per la detection anche sui dati reali. In questo paragrafo saranno mostrati i risultati relativi a due dataset, uno con *read single-end* e uno con *read paired-end.*

Sul dataset di sequenze SRR7062764- SRR7062769, che rappresentano il trascrittoma di cellule staminali neurali umane infette con il virus Zika, è stato possibile utilizzare solo CIRI2 e CIRCexplorer2 in quanto le *reads* sono *single-end* e CIRIquant lavora solo su *reads paired end.*

Nella fig. 30 è riportato il numero dei circRNA trovato da ciascun tool.



Reads dataset Zika infected neural stem cells

	CIRI2	CIRCexplorer2
Seq SRR7062764	728	2838
Seq SRR7062765	137	861
Seq SRR7062766	519	2119
Seq SRR7062767	586	2428
Seq SRR7062768	600	2582
Seq SRR7062769	725	2767

Figura 30: detection dei circRNA sul trascrittoma di cellule staminali neuronali infettate con virus Zika

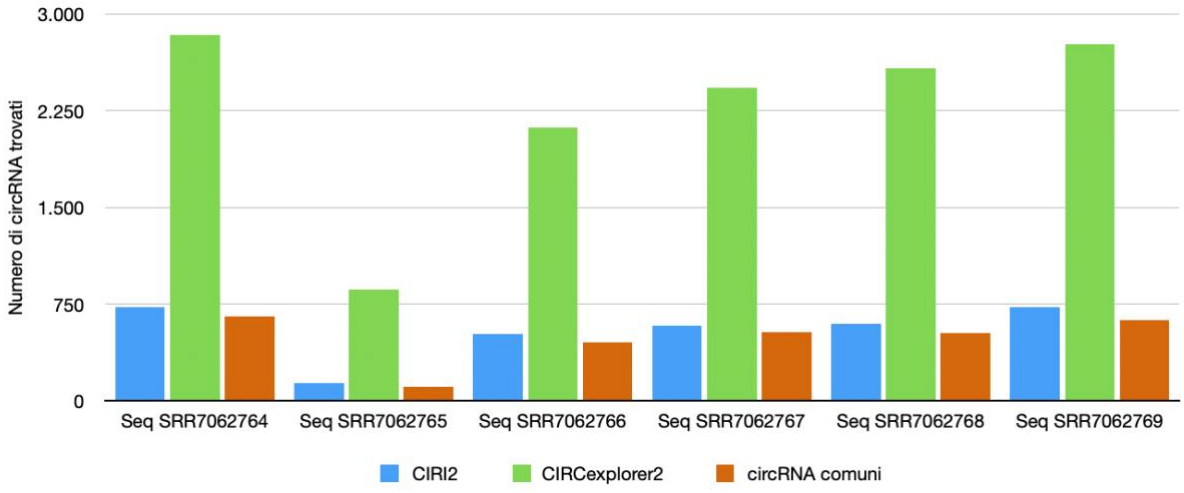
A questo punto è stato fatto un confronto tra gli output dei due tool per verificare quanti circRNA sono comuni a entrambi.

È stato usato lo script Python in figura 31, il quale prende in ingresso gli output dei due tool, crea una lista per ciascun output in cui sono presenti la posizione di inizio e fine di ciascuna BSJ, e il relativo cromosoma di appartenenza del circRNA. A questo punto vengono confrontati gli indici e se c'è uguaglianza tra questi (CIRCexplorer2 ha la posizione di inizio della BSJ traslata di uno rispetto all'output di CIRI2) si crea un ulteriore lista in cui sono riportati i circRNA comuni.

```
 GNU nano 5.4                                          confronto.py *
#!/usr/bin/env python3

def main():
    lista1 = []
    lista2 = []

    with open('CIRCexplorer2/circularRNA_known_bwa69.txt', 'r') as out:
        out.readline()
        for i in out:
            campi = i.split('\t')
            lista1.append([int(campi[1]), int(campi[2]),campi[0]])

    with open('outfile69_hg19.txt', 'r') as fj:
        fj.readline()
        for j in fj:
            p = j.split('\t')
            lista2.append([int(p[2]), int(p[3]),p[3]])

    with open('equal69.txt', 'w') as conf:
        all_uguale = []
        for i in lista1:
           for j in lista2:
               if i[0]==(j[0]-1) and i[1]==j[1]:
                   conf.write('\t'.join(map(str, i)) + '\n')

if __name__ == '__main__':
    main()
```

Figura 31: script per il confronto degli output dei tool CIRI2 e CIRCexplorer2

Nella fig.32 è possibile vedere il risultato di questo confronto:

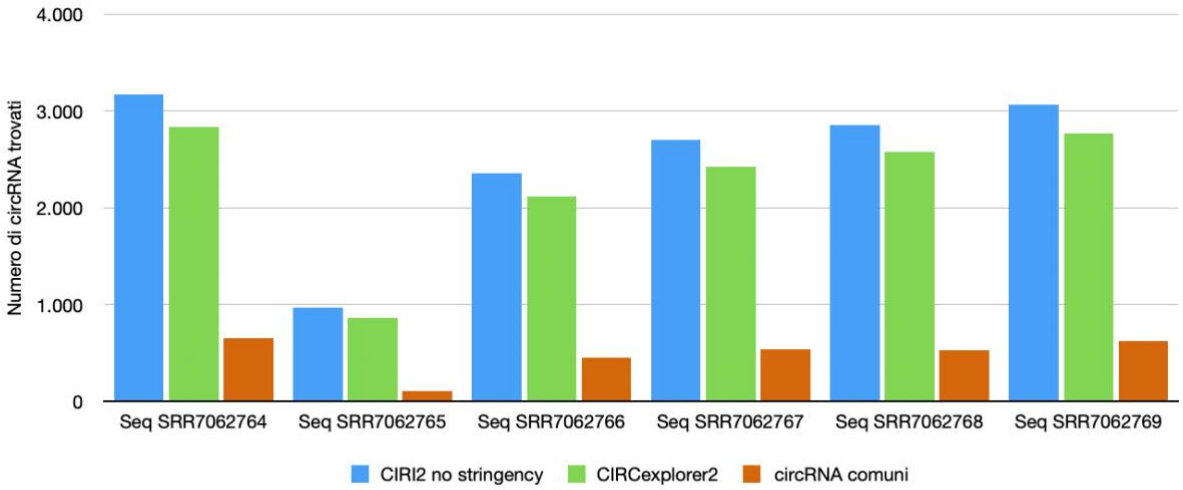


Reads dataset Zika infected neural stem cells

	CIRI2	CIRCexplorer2	circRNA comuni
Seq SRR7062764	728	2838	652
Seq SRR7062765	137	861	112
Seq SRR7062766	519	2119	456
Seq SRR7062767	586	2428	535
Seq SRR7062768	600	2582	528
Seq SRR7062769	725	2767	630

Figura 32: circRNA comuni ai due tool

Per fare un confronto coerente tra i risultati trovati è stato eseguito CIRI2 con il parametro -0 (no-stringency) ed in questo caso, come ci aspettavamo, è stato trovato un numero molto maggiore di circRNA.

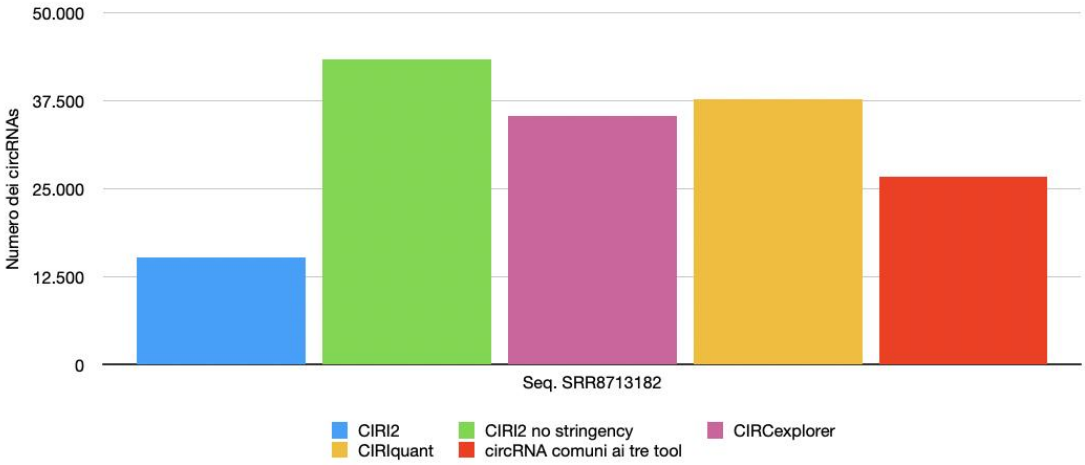


Reads dataset Zika infected neural stem cells

	CIRI2 no stringency	CIRCexplorer2	circRNA comuni
Seq SRR7062764	3.174	2838	652
Seq SRR7062765	974	861	112
Seq SRR7062766	2.357	2119	456
Seq SRR7062767	2.699	2428	535
Seq SRR7062768	2.856	2582	528
Seq SRR7062769	3.067	2767	630

Figura 33: circRNA comuni ai due tool

La detection dei circRNA è stata fatta anche su *reads-paired end* RNA-seq appartenenti a una libreria arricchita per i circRNA di organoidi cerebrali derivati da cellule staminali pluripotenti indotte (iPSC) in seguito all'infezione da citomegalovirus (seq. SRR8713182_1.fastq -SRR8713182_2.fastq).
Nella fig. 34 è rappresentato il numero di circRNAs trovato da ogni tool e i circRNA comuni tra i tool.



Numero di circRNAs rilevati

	CIRI2	CIRI2 no stringency	CIRCexplorer	CIRIquant	circRNA comuni ai tre tool
SRR8713182	15.226	43449	35.337	37.712	26.769

Figura 34: rappresentazione del numero di circRNA trovati da ciascun tool in un dataset di RNA-seq paired-end reale

Quantificazione dei circRNAs sui dati simulati

Per la quantificazione dei circRNA sono stati utilizzati i tool CIRI2 e CIRIquant.

La fig.35 mostra quanto la quantificazione ottenuta con i due strumenti è notevolmente differente.

Quantificazione	Numero circRNAs quantificati correttamente 10000-150-1000
CIRI2	9292
CIRIquant	1851

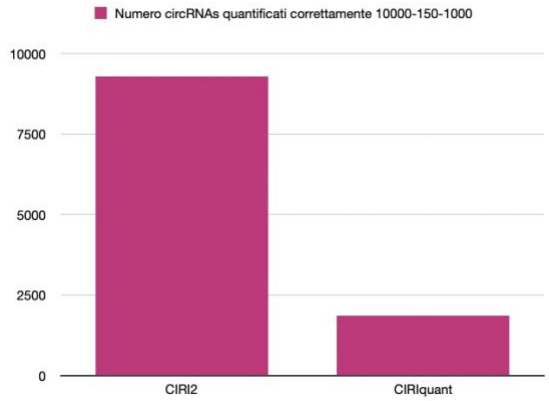


Figura 35: divergenza nella quantificazione dei circRNA da parte di CIRI2 e CIRIquant

Ad esempio nell'esperimento 10000-150-1000, se si va a controllare la quantificazione del circular con ID NM_000097.7:2-7, si hanno i risultati mostrati in fig.36:

CIRIquant tende a sottostimare il numero di *reads* con BSJ trovate.

Timepoint	#BSJ simulated	CIRI2 quantification	CIRIquant quantification
1	194	194	191
2	455	455	450
3	361	361	356
4	364	364	362

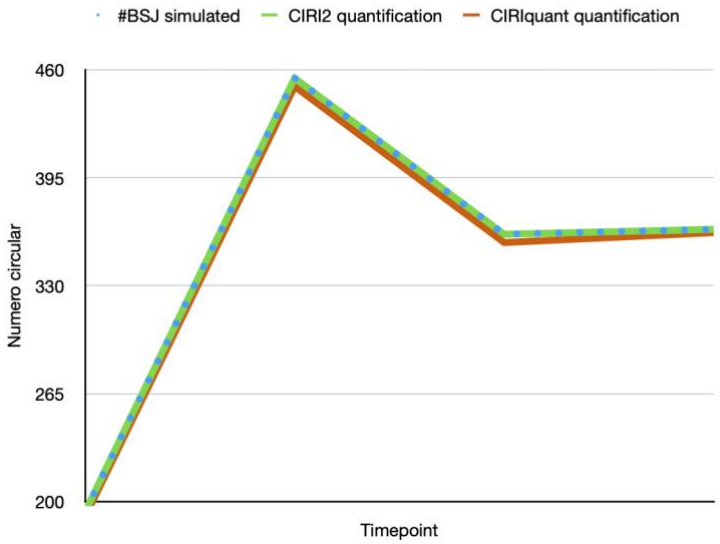


Figura 36: quantificazione delle reads che contengono BSJ per il circRNA NM_000097.7:2-7

Un esempio ancora più evidente è mostrato dalla fig. 37 in cui CIRIquant fa la detection del circRNA ma la quantificazione dello stesso è pari ad 1 nonostante il numero di *reads* associate a quel circRNA sia molto più elevato.

Timepoint	#BSJ simulated	CIRI2 quantification	CIRIquant quantification
1	676	676	1
2	357	357	1
3	962	962	1
4	995	995	1

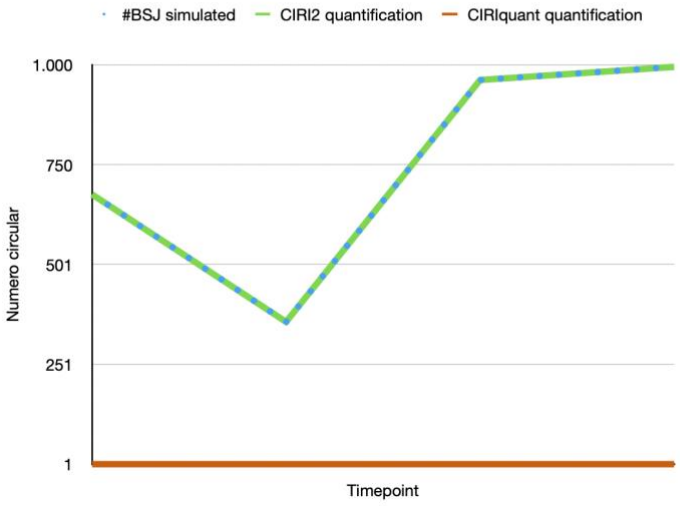


Figura 37: quantificazione delle reads che contengono BSJ per il circRNA NM_000128.4:5-13

La figura 38 e la figura 39 mostrano invece la correlazione tra due tool relativa alla quantificazione dei diversi circRNAs al variare del numero di *read* per circRNA.

Quantificazione	Numero circRNAs quantificati correttamente 1000-150-10	Numero circRNAs quantificati correttamente 1000-150-100	Numero circRNAs quantificati correttamente 1000-150-1000
CIRI2	784	912	925
CIRIquant	321	244	235

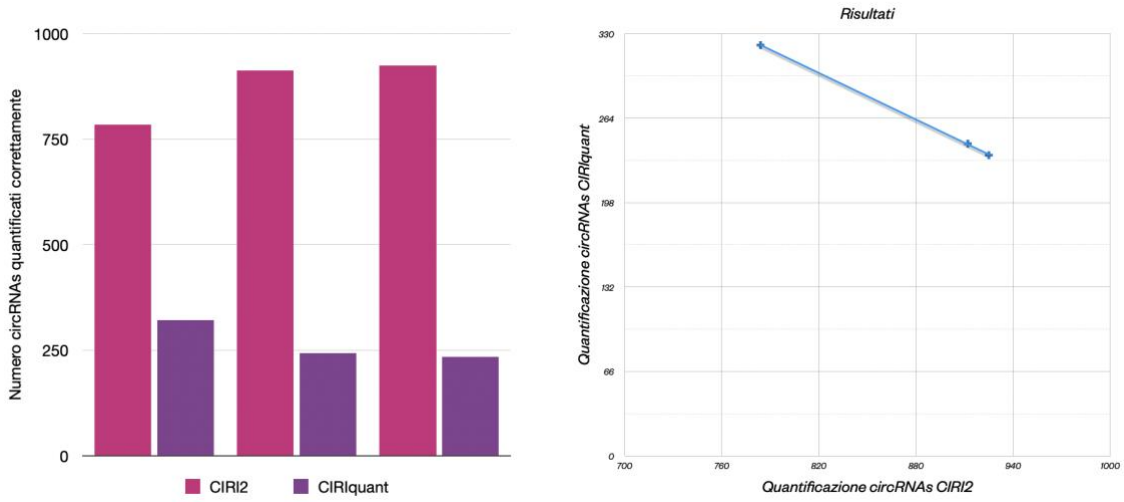


Figura 38: correlazione nella quantificazione da parte di CIRI2 e CIRIquant al variare del numero di read per circRNA

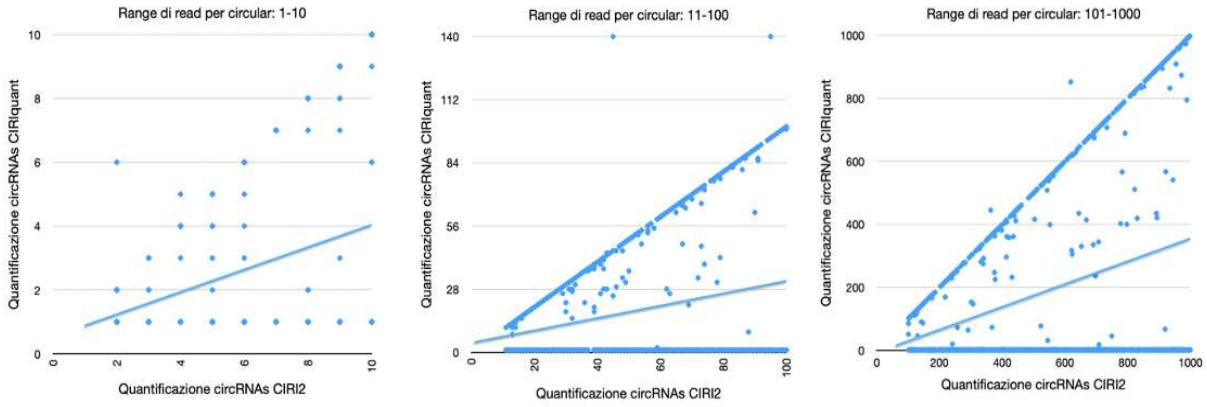


Figura 39:grafico a dispersione della quantificazione dei circRNA da parte di CIRI2 e CIRIquant al variare del numero di read per circRNA

Nella fig.39 si vede il grande numero di circRNAs quantificati in modo errato da CIRIquant (molti punti in corrispondenza di 1), le diagonali dei grafici rappresentano invece le quantificazioni corrette comuni ad entrambi i tool.

Quantificazione dei circRNA sui dati reali

Sulle stesse *reads-paired end* utilizzate per la detection, è stata fatta anche la quantificazione utilizzando CIRI2 e CIRIquant. Una volta ottenuti gli output dei due tool è stato utilizzato lo script in fig. 40 per avere come output un file di comparazione in cui sono presenti i circRNAs identificati da entrambi i tool e il numero delle BSJ identificate da ognuno di essi:

```
 GNU nano 5.4                                    quant_confronto.py
#!/usr/bin/env python3
def main():
    lista1 = []
    lista2 = []

    with open('CIRIquant-master/dati_reali/SRR8713182.gtf', 'r') as out:
        for _ in range(5):
            out.readline()
        for i in out:
            campi = i.split('\t')
            pos = campi[8]
            e = pos.split(';')
            lista1.append([int(campi[3]), int(campi[4]), e[2]])

    with open('out_82_hg19_nostringency.txt', 'r') as fj:
        fj.readline()
        for j in fj:
            p = j.split('\t')
            lista2.append([int(p[2]), int(p[3]), p[4]])

    with open('real_quant.txt', 'w') as conf:
        all_uguale = []
        for i in lista1:
            for j in lista2:
                if i[0] == j[0] and i[1] == j[1]:
                    conf.write('\t'.join(map(str, i)) + '\t' + j[2] + '\n')

if __name__ == '__main__':
    main()
```

Figura 40: script che compara gli output di CIRI2 e CIRIquant generando un file in cui sono presenti gli indici iniziale e finale delle BSJ e la loro quantificazione

La fig.41 rappresenta il risultato della quantificazione tramite correlazione:

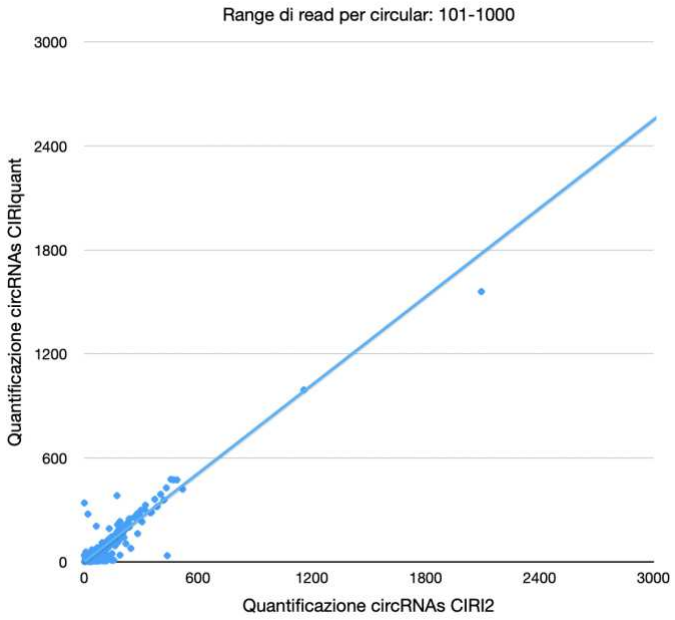


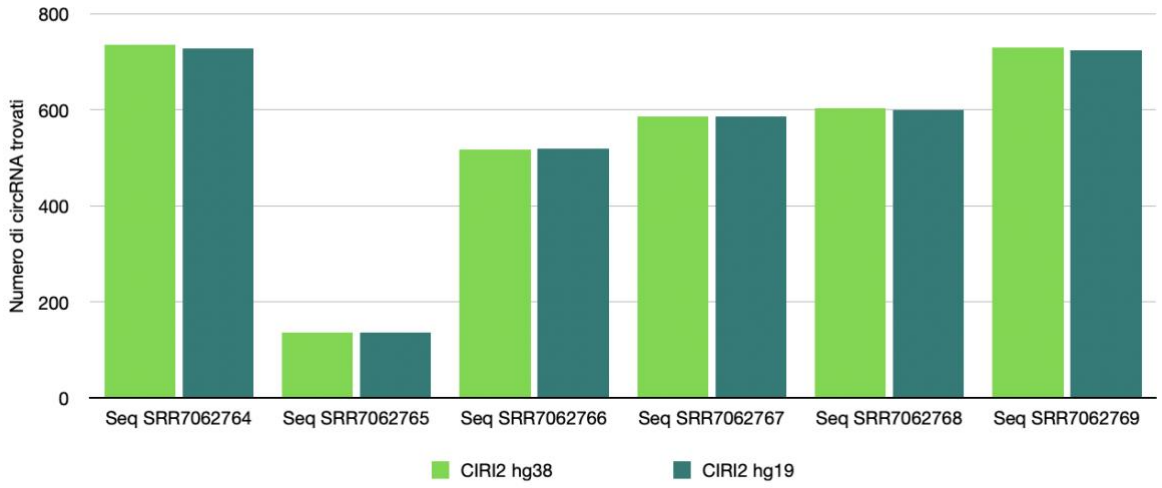
Figura 41:correlazione della quantificazione di BSJ ottenuta da CIRI2 e CIRIquant sui dati reali

Importanza del genoma di riferimento

Un aspetto importante da prendere in considerazione consiste nella scelta del genoma di riferimento utilizzato nella ricerca dei circRNA.

Per le sequenze *single-end* sono stati utilizzati CIRI2 e CIRCexplorer2 su entrambe le versioni del genoma di riferimento (hg19 e hg38) ed è stato ottenuto un numero di circRNAs leggermente differente, come si può vedere in fig.42.

Va sottolineato che non è possibile effettuare un confronto diretto tra gli output generati da questi strumenti a causa delle differenze nei valori degli indici associati all'inizio e alla fine delle BSJ. Al fine di condurre un confronto accurato, sarebbe necessario valutare l'uguaglianza tra le differenze riscontrate tra le posizioni di fine e di inizio delle BSJ, imponendo come ulteriore condizione che tali giunzioni siano localizzate sul medesimo cromosoma.



Reads dataset Zika infected neural stem cells

	CIRI2 hg38	CIRI2 hg19
Seq SRR7062764	736	728
Seq SRR7062765	136	137
Seq SRR7062766	518	519
Seq SRR7062767	587	586
Seq SRR7062768	603	600
Seq SRR7062769	729	725

Figura 42: differenza nella rilevazione dei circRNAs quando si utilizzano dei genomi di riferimento differenti

Discussione

Nel percorso di tesi, ci siamo concentrati sull'obiettivo di sviluppare una pipeline innovativa per l'identificazione e la quantificazione dei circRNAs all'interno di dataset di RNA-seq. Questa ricerca ha richiesto l'impiego e la valutazione di algoritmi e strategie diverse al fine di garantire un rilevamento affidabile dei circRNAs, nonché la determinazione dei loro livelli di espressione, anche in campioni longitudinali raccolti nel corso del tempo.
È stata fatta una valutazione dei tool utilizzando sia *reads* simulate che dati provenienti da repository pubblici. Questo risultato sarà in seguito utilizzato come parte integrante di un progetto di ricerca dedicato all'analisi del contributo dei circRNAs nel processo di sviluppo neurale fetale in presenza di diverse infezioni congenite causate da virus patogeni.
Un aspetto rilevante di questa ricerca è stato lo sviluppo di un simulatore in grado di generare circRNA artificiali a partire da una vasta banca dati, con la possibilità di variare parametri e configurazioni. Lo sviluppo di un nuovo simulatore ci ha permesso di creare dei dataset di grandi dimensioni senza che vengano salvati sulla memoria locale, al contrario di altri simulatori già disponibili dalla letteratura e che abbiamo testato, come Circall Simulator.
In particolare, la pipeline è stata eseguita 18 volte, ogni volta con combinazioni differenti dei parametri: numero di circRNAs, lunghezza delle *reads* e numero di *read* per ogni circRNA.
Infine è stato effettuato un confronto dettagliato tra tre differenti strumenti di analisi su quattro punti temporali differenti per ogni dataset simulato.

Detection dei circRNA sui dati simulati

In tutte le simulazioni, CIRI2 e CIRCexplorer2 hanno trovato un numero di circRNA maggiore rispetto a CIRIquant (tabella 2, fig.28). Questo è un risultato che ci aspettavamo dalla letteratura in quanto CIRIquant filtra i circRNA con lo scopo di eliminare potenziali falsi positivi, ma impiegando evidentemente un approccio troppo aggressivo. [116]
Per quanto riguarda i valori di precisione, il tool CIRIquant risulta più preciso, in quanto rileva meno falsi positivi, ma a discapito del valore di *recall*, in quanto il numero di falsi negativi è molto maggiore rispetto a CIRI2 e CIRCexplorer2. Dalla tabella 2 si nota inoltre che, per gli esperimenti con 10000 circRNAs simulati, il valore dei falsi positivi per il tool CIRIquant è pari a zero. Questo è dovuto al fatto che nella pipeline sviluppata ci si riconduce ai circRNAs utilizzando le posizioni di inizio e fine della BSJ e ciò può portare alla rilevazione di più circRNAs con gli stessi indici quando appartengono a varianti trascrizionali differenti.

Questo fenomeno non rappresenta un errore della pipeline, ma piuttosto una conseguenza delle limitazioni dei tool di detection, che elaborando delle *short reads*, non riescono a darci informazioni accurate sull'intero trascritto.
Le *short reads, infatti,* potrebbero non coprire totalmente il circRNA e, di conseguenza, risulta impossibile riconoscere le varianti trascrizionali.
Una possibile soluzione sarebbe l'utilizzo delle *long reads*, in maniera da poter distinguere le isoforme trascrizionali e identificare quindi i circRNAs in maniera più accurata.
CIRI2 e CIRCexplorer2 sono abbastanza concordanti sui valori di precisione e *recall* e concordi con i risultati presenti in letteratura.[91,116] La precisione aumenta con il diminuire del numero di *reads* per circRNA, la *recall* cresce invece in maniera proporzionale al numero di *reads* per circRNA.
È importante notare che la precisione di CIRCexplorer2 diminuisce notevolmente quando si ha un numero di circRNA simulati più basso. Di conseguenza, in un dataset reale non arricchito per i circRNA ci aspettiamo che CIRCexplorer2 performi in maniera peggiore rispetto agli altri due tool.

Detection circRNA sui dati reali

La rilevazione dei circRNA sui dati reali mostrata in fig.30 è discordante con i risultati sui dati simulati. Questo è dovuto al fatto che CIRI2 sui dati simulati è stato utilizzato con il parametro -0 (*no-stringency*) poiché la qualità dei dati simulati era molto alta. Nei dati reali, invece, CIRI2 è stato utilizzato nella sua configurazione di default, perché non sappiamo qual è la qualità dei dati, per cui c'è un filtro sul numero di BSJ che devono essere trovate per poter chiamare un circRNA.
Dalla fig.33 si può vedere come, nel momento in cui si utilizza lo stesso parametro -0 sui dati reali, si hanno risultati coerenti con i dati simulati.
CIRIquant è stato utilizzato per i dati reali solo su dataset con *reads paired-end* in quanto non lavora su *read single-end.*
Sulle *reads paired-end* appartenenti ad un dataset arricchito per i circRNAs, CIRIquant trova più circRNAs rispetto a CIRCexplorer, al contrario di ciò che avevamo riscontrato sui dati simulati. Questo potrebbe essere attribuito al fatto che ci sia del rumore di fondo che influisce sulla rilevazione dei circRNA da parte di CIRCexplorer.

Quantificazione dei circRNA

Per quanto riguarda la quantificazione, dalla tabella 1 e dalla fig.35 emerge chiaramente che CIRI2 fornisce una quantificazione più corretta rispetto a CIRIquant. È stato fatto un confronto tra il numero di *reads* contenenti BSJ reale e quello trovato da entrambi i tool nella simulazione con parametri 10000-150-1000, ed è stato riscontrato che CIRI2 ha quantificato

in maniera corretta il 94,4% dei circRNA. CIRIquant invece, ha quantificato correttamente solo il 21,76%, mentre alla restante parte dei circRNAs è stato erroneamente associato il conteggio di una sola *read.*

Questa informazione si rispecchia nelle figure 35-37, in cui si nota che anche con un numero di *read* per lo stesso circRNA molto elevato, nella maggior parte dei casi CIRIquant non riesce a fornire una quantificazione accurata.

All'aumentare del numero di *read* per circRNA, CIRI2 migliora la quantificazione, mentre CIRIquant quantifica in maniera migliore quando ci sono meno *read* per ciascun circRNA (fig38).

Nei dati sperimentali ci aspettiamo di avere poche *read* per ciascun circRNA. Dalla fig.41, si nota che sui dati reali CIRIquant esegue una quantificazione molto simile a quella di CIRI2, confermando la migliore prestazione quando il numero di *read* per circRNA diminuisce.

Conclusioni

Sulla base di questi risultati possiamo concludere che CIRI2 è risultato essere il tool più performante in vari scenari, dimostrandosi leggermente più restrittivo ma estremamente affidabile nella quantificazione dei circRNAs.

Le prospettive future per questa ricerca includono il potenziamento della pipeline, con l'obiettivo di utilizzare la versione più aggiornata del genoma di riferimento per la costruzione della banca dati, in quanto essa contiene una maggiore quantità di informazioni. Inoltre, stiamo lavorando per perfezionare il simulatore, in modo da evitare che ci sia il bias dovuto alle varianti trascrizionali, che attualmente possono portare all'assegnazione delle stesse posizioni a più circRNAs.

La pipeline sviluppata durante il progetto di tesi contribuirà a valutare il ruolo dei circRNA nello sviluppo neuronale fetale. In questo ambito, sono in corso di produzione dei data set sperimentali di trascrittomica, arricchiti per i circRNAs, provenienti da cellule staminali neurali infettate con diversi virus umani patogeni, sui quali sarà possibile studiare i profili di espressione dei circRNA attraverso l'utilizzo della pipeline. L'obiettivo è quello di riuscire ad identificare potenziali candidati per una successiva caratterizzazione funzionale.

I passi avanti nella metodologia contribuiranno alla precisione e all'affidabilità della nostra analisi dei circRNA, fornendo così un contributo significativo al successo di questo progetto di ricerca.

BIBLIOGRAFIA

1. Sanger, H. L. *et al. Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures (electron microscopy/endgroup analysis/ultracentrifugation/thermal denaturation)*. vol. 73 (1976).
2. Kolakofsky, D. Isolation and characterization of Sendai virus DI-RNAs. *Cell* **8**, 547–555 (1976).
3. Kjems, J. & Garrett, R. A. Novel splicing mechanism for the ribosomal RNA intron in the archaebacterium desulfurococcus mobilis. *Cell* **54**, 693–703 (1988).
4. Wilusz, J. E. A 360° view of circular RNAs: From biogenesis to functions. *Wiley Interdisciplinary Reviews: RNA* vol. 9 Preprint at https://doi.org/10.1002/wrna.1478 (2018).
5. Giaretta, A., Ghusinga, K. R. & Elston, T. C. A Stochastic model for RNA splicing. in *2022 European Control Conference (ECC)* 1164–1169 (2022). doi:10.23919/ECC55457.2022.9838423.
6. Chen, L. L. The expanding regulatory mechanisms and cellular functions of circular RNAs. *Nature Reviews Molecular Cell Biology* vol. 21 475–490 Preprint at https://doi.org/10.1038/s41580-020-0243-y (2020).
7. Clancy, S. RNA Splicing: Introns, Exons and Spliceosome. *Nature Education* (2008).
8. Van Den Hoogenhof, M. M. G., Pinto, Y. M. & Creemers, E. E. RNA Splicing regulation and dysregulation in the heart. *Circulation Research* vol. 118 454–468 Preprint at https://doi.org/10.1161/CIRCRESAHA.115.307872 (2016).
9. Olthof, A. M. *et al.* The minor and major spliceosome interact to regulate alternative splicing around minor introns. *bioRxiv* 2020.05.18.101246 (2020) doi:10.1101/2020.05.18.101246.
10. Mehta, S. L., Dempsey, R. J. & Vemuganti, R. Role of circular RNAs in brain development and CNS diseases. *Progress in Neurobiology* vol. 186 Preprint at https://doi.org/10.1016/j.pneurobio.2020.101746 (2020).
11. Ikeda, Y. *et al.* CircRNAs and RNA-Binding Proteins Involved in the Pathogenesis of Cancers or Central Nervous System Disorders. *Non-coding RNA* vol. 9 Preprint at https://doi.org/10.3390/ncrna9020023 (2023).
12. Lin, Z. *et al.* Functions and mechanisms of circular RNAs in regulating stem cell differentiation. *RNA Biol* **18**, 2136–2149 (2021).
13. Park, E. G. *et al.* Genomic Analyses of Non-Coding RNAs Overlapping Transposable Elements and Its Implication to Human Diseases. *International Journal of Molecular Sciences* vol. 23 Preprint at https://doi.org/10.3390/ijms23168950 (2022).
14. *Evaluate both RNA and protein targets in single cells PrimeFlow RNA assay for detecting RNA targets by flow cytometry*. (2017).
15. Huang, A., Zheng, H., Wu, Z., Chen, M. & Huang, Y. Circular RNA-protein interactions: Functions, mechanisms, and identification. *Theranostics* vol. 10 3506–3517 Preprint at https://doi.org/10.7150/thno.42174 (2020).
16. Bachmayr-Heyda, A. *et al.* Correlation of circular RNA abundance with proliferation - Exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and normal human tissues. *Sci Rep* **5**, 8057 (2015).
17. Westholm, J. O. *et al.* Genome-wide Analysis of Drosophila Circular RNAs Reveals Their Structural and Sequence Properties and Age-Dependent Neural Accumulation. *Cell Rep* **9**, 1966–1980 (2014).
18. Cortés-López, M. *et al.* Global accumulation of circRNAs during aging in Caenorhabditis elegans. *BMC Genomics* **19**, 8 (2018).
19. Gruner, H., Cortés-López, M., Cooper, D. A., Bauer, M. & Miura, P. CircRNA accumulation in the aging mouse brain. *Sci Rep* **6**, 38907 (2016).

20. Mahmoudi, E. & Cairns, M. J. Circular RNAs are temporospatially regulated throughout development and ageing in the rat. *Sci Rep* **9**, 2564 (2019).
21. Knupp, D. & Miura, P. CircRNA accumulation: A new hallmark of aging? *Mechanisms of Ageing and Development* vol. 173 71–79 Preprint at https://doi.org/10.1016/j.mad.2018.05.001 (2018).
22. Rybak-Wolf, A. *et al.* Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Mol Cell* **58**, 870–885 (2015).
23. Meng, S. *et al.* CircRNA: Functions and properties of a novel potential biomarker for cancer. *Molecular Cancer* vol. 16 Preprint at https://doi.org/10.1186/s12943-017-0663-2 (2017).
24. Guria, A., Sharma, P., Natesan, S. & Pandi, G. Circular RNAs—The Road Less Traveled. *Frontiers in Molecular Biosciences* vol. 6 Preprint at https://doi.org/10.3389/fmolb.2019.00146 (2020).
25. Geng, X. *et al.* Circular RNA: Biogenesis, degradation, functions and potential roles in mediating resistance to anticarcinogens. *Epigenomics* vol. 12 267–283 Preprint at https://doi.org/10.2217/epi-2019-0295 (2020).
26. Xu, T., Wu, J., Han, P., Zhao, Z. & Song, X. Circular RNA expression profiles and features in human tissues: A study using RNA-seq data. *BMC Genomics* **18**, (2017).
27. Ma, Y. *et al.* A Comprehensive Overview of circRNAs: Emerging Biomarkers and Potential Therapeutics in Gynecological Cancers. *Frontiers in Cell and Developmental Biology* vol. 9 Preprint at https://doi.org/10.3389/fcell.2021.709512 (2021).
28. Shen, H. *et al.* Circular RNAs: characteristics, biogenesis, mechanisms and functions in liver cancer. *Journal of Hematology and Oncology* vol. 14 Preprint at https://doi.org/10.1186/s13045-021-01145-8 (2021).
29. Patop, I. L., Wüst, S. & Kadener, S. Past, present, and future of circ RNA s . *EMBO J* **38**, (2019).
30. Zou, Y. *et al.* The role of circular RNA CDR1as/cirs-7 in regulating tumor microenvironment: A pan-cancer analysis. *Biomolecules* **9**, (2019).
31. Zhou, W. Y. *et al.* Circular RNA: metabolism, functions and interactions with proteins. *Molecular Cancer* vol. 19 Preprint at https://doi.org/10.1186/s12943-020-01286-3 (2020).
32. Zhao, Z. J. & Shen, J. Circular RNA participates in the carcinogenesis and the malignant behavior of cancer. *RNA Biology* vol. 14 514–521 Preprint at https://doi.org/10.1080/15476286.2015.1122162 (2017).
33. Zhang, M. & Bian, Z. The Emerging Role of Circular RNAs in Alzheimer's Disease and Parkinson's Disease. *Frontiers in Aging Neuroscience* vol. 13 Preprint at https://doi.org/10.3389/fnagi.2021.691512 (2021).
34. Kishore, R., Garikipati, V. N. S. & Gonzalez, C. Role of Circular RNAs in Cardiovascular Disease. *Journal of cardiovascular pharmacology* vol. 76 128–137 Preprint at https://doi.org/10.1097/FJC.0000000000000841 (2020).
35. Wu, Y. L., Li, H. F., Chen, H. H. & Lin, H. Emergent Roles of Circular RNAs in Metabolism and Metabolic Disorders. *International Journal of Molecular Sciences* vol. 23 Preprint at https://doi.org/10.3390/ijms23031032 (2022).
36. Li, W., Liu, J. Q., Chen, M., Xu, J. & Zhu, D. Circular RNA in cancer development and immune regulation. *Journal of Cellular and Molecular Medicine* vol. 26 1785–1798 Preprint at https://doi.org/10.1111/jcmm.16102 (2022).
37. Lei, M., Zheng, G., Ning, Q., Zheng, J. & Dong, D. Translation and functional roles of circular RNAs in human cancer. *Molecular Cancer* vol. 19 Preprint at https://doi.org/10.1186/s12943-020-1135-7 (2020).
38. Liu, K.-S., Pan, F., Mao, X.-D., Liu, C. & Chen, Y.-J. *Biological functions of Circular RNAs and their roles in occurrence of reproduction and gynecological diseases*. *Am J Transl Res* vol. 11 www.ajtr.org (2019).

39. Gao, X. *et al.* Circular RNA-encoded oncogenic E-cadherin variant promotes glioblastoma tumorigenicity through activation of EGFR–STAT3 signalling. *Nat Cell Biol* **23**, 278–291 (2021).
40. Chen, J. *et al.* CircPTN sponges miR-145-5p/miR-330-5p to promote proliferation and stemness in glioma. *Journal of Experimental and Clinical Cancer Research* **38**, (2019).
41. Fu, B. *et al.* Circular rna circbcbm1 promotes breast cancer brain metastasis by modulating mir-125a/brd4 axis. *Int J Biol Sci* **17**, 3104–3117 (2021).
42. Li, Y. *et al.* Circular RNA is enriched and stable in exosomes: A promising biomarker for cancer diagnosis. *Cell Research* vol. 25 981–984 Preprint at https://doi.org/10.1038/cr.2015.82 (2015).
43. Zhang, Z., Yang, T. & Xiao, J. Circular RNAs: Promising Biomarkers for Human Diseases. *EBioMedicine* vol. 34 267–274 Preprint at https://doi.org/10.1016/j.ebiom.2018.07.036 (2018).
44. Wei, X. *et al.* Underlying metastasis mechanism and clinical application of exosomal circular RNA in tumors (Review). *International Journal of Oncology* vol. 58 289–297 Preprint at https://doi.org/10.3892/ijo.2021.5179 (2021).
45. Tang, Q. & Hann, S. S. Biological roles and mechanisms of circular RNA in human cancers. *OncoTargets and Therapy* vol. 13 2067–2092 Preprint at https://doi.org/10.2147/OTT.S233672 (2020).
46. Dong, X. *et al.* Circular RNAs in the human brain are tailored to neuron identity and neuropsychiatric disease. *Nat Commun* **14**, 5327 (2023).
47. Hanan, M., Soreq, H. & Kadener, S. CircRNAs in the brain. *RNA Biology* vol. 14 1028–1034 Preprint at https://doi.org/10.1080/15476286.2016.1255398 (2017).
48. Zheng, D. *et al.* Screening of Human Circular RNAs as Biomarkers for Early Onset Detection of Alzheimer's Disease. *Front Neurosci* **16**, (2022).
49. Aquilina-Reid, C. *et al.* Circular RNA Expression and Interaction Patterns Are Perturbed in Amyotrophic Lateral Sclerosis. *Int J Mol Sci* **23**, (2022).
50. Li, M. L., Wang, W. & Jin, Z. B. Circular RNAs in the Central Nervous System. *Frontiers in Molecular Biosciences* vol. 8 Preprint at https://doi.org/10.3389/fmolb.2021.629593 (2021).
51. Zhang, L., Li, Z., Mao, L. & Wang, H. Circular RNA in Acute Central Nervous System Injuries: A New Target for Therapeutic Intervention. *Frontiers in Molecular Neuroscience* vol. 15 Preprint at https://doi.org/10.3389/fnmol.2022.816182 (2022).
52. Yu, X. *et al.* Circular RNAs: New players involved in the regulation of cognition and cognitive diseases. *Frontiers in Neuroscience* vol. 17 Preprint at https://doi.org/10.3389/fnins.2023.1097878 (2023).
53. Li, Z. *et al.* The emerging landscape of circular RNAs in immunity: Breakthroughs and challenges. *Biomarker Research* vol. 8 Preprint at https://doi.org/10.1186/s40364-020-00204-5 (2020).
54. Awan, F. M. *et al.* The emerging role and significance of circular RNAs in viral infections and antiviral immune responses: possible implication as theranostic agents. *RNA Biology* Preprint at https://doi.org/10.1080/15476286.2020.1790198 (2020).
55. Xie, R., Zhang, Y., Zhang, J., Li, J. & Zhou, X. The Role of Circular RNAs in Immune-Related Diseases. *Frontiers in Immunology* vol. 11 Preprint at https://doi.org/10.3389/fimmu.2020.00545 (2020).
56. Xie, H. *et al.* The role of circular RNAs in viral infection and related diseases. *Virus Research* vol. 291 Preprint at https://doi.org/10.1016/j.virusres.2020.198205 (2021).
57. Deng, J. *et al.* Human Cytomegalovirus Influences Host circRNA Transcriptions during Productive Infection. *Virol Sin* (2021) doi:10.1007/s12250-020-00275-6.
58. Yang, S. *et al.* Circular RNAs Represent a Novel Class of Human Cytomegalovirus Transcripts. *Microbiol Spectr* **10**, (2022).

59. Cai, Z. *et al.* VirusCircBase: A database of virus circular RNAs. *Brief Bioinform* **22**, 2182–2190 (2021).
60. Salzman, J., Gawad, C., Wang, P. L., Lacayo, N. & Brown, P. O. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One* **7**, (2012).
61. Nielsen, A. F. *et al.* Best practice standards for circular RNA research. *Nature Methods* vol. 19 1208–1220 Preprint at https://doi.org/10.1038/s41592-022-01487-2 (2022).
62. Mi, Z. *et al.* Circular RNA detection methods: A minireview. *Talanta* **238**, 123066 (2022).
63. Hou, L., Zhang, J. & Zhao, F. Full-length circular RNA profiling by nanopore sequencing with CIRI-long. *Nat Protoc* **18**, 1795–1813 (2023).
64. Zhang, J. *et al.* Comprehensive profiling of circular RNAs with nanopore sequencing and CIRI-long. *Nat Biotechnol* **39**, 836–845 (2021).
65. Chen, L. *et al.* The bioinformatics toolbox for circRNA discovery and analysis. *Briefings in Bioinformatics* vol. 22 1706–1728 Preprint at https://doi.org/10.1093/bib/bbaa001 (2021).
66. Gao, Y., Wang, J. & Zhao, F. CIRI: An efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol* **16**, (2015).
67. Gao, Y., Zhang, J. & Zhao, F. Circular RNA identification based on multiple seed matching. *Brief Bioinform* **19**, 803–810 (2018).
68. Zhang, X. O. *et al.* Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res* **26**, 1277–1287 (2016).
69. circRNAFinder_BiCoB-2014.
70. Szabo, L. *et al.* Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol* **16**, (2015).
71. Wang, K. *et al.* MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* **38**, (2010).
72. Th, M. E. *et al. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection*. *Genome Biology* vol. 15 http://genomebiology.com/2014/15/2/R34 (2014).
73. Buratin, A., Bortoluzzi, S. & Gaffo, E. Systematic benchmarking of statistical methods to assess differential expression of circular RNAs. *Brief Bioinform* **24**, (2023).
74. Vromman, M. *et al.* Large-scale benchmarking of circRNA detection tools reveals large differences in sensitivity but not in precision. doi:10.1101/2022.12.06.519083.
75. Li, X. & Wu, Y. Detecting circular RNA from high-throughput sequence data with de Bruijn graph. *BMC Genomics* **21**, (2020).
76. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
77. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
78. Chen, L., Huang, C., Wang, X. & Shan, G. *Send Orders for Reprints to reprints@benthamscience.ae Circular RNAs in Eukaryotic Cells*. *Current Genomics* vol. 16 (2015).
79. Ma, X.-K., Xue, W., Chen, L.-L. & Yang, L. CIRCexplorer pipelines for circRNA annotation and quantification from non-polyadenylated RNA-seq datasets. *Methods* **196**, 3–10 (2021).
80. Cheng, J., Metge, F. & Dieterich, C. Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* **32**, 1094–1096 (2016).
81. Hansen, T. B. Improved circRNA identification by combining prediction algorithms. *Front Cell Dev Biol* **6**, (2018).

82. Fu, X. & Liu, R. *Circrnafinder: A tool for identifying circular RNAs using RNA-Seq data*. *Proceedings of the 6th International Conference on Bioinformatics and Computational Biology, BICOB 2014* (2014).
83. Chen, C. Y. & Chuang, T. J. NCLcomparator: Systematically post-screening non-co-linear transcripts (circular, trans-spliced, or fusion RNAs) identified from various detectors. *BMC Bioinformatics* **20**, (2019).
84. Li, X. & Wu, Y. Detecting circular RNA from high-throughput sequence data with de Bruijn graph. *BMC Genomics* **21**, (2020).
85. Nguyen, M. H., Nguyen, H. N. & Vu, T. N. Evaluation of methods to detect circular RNAs from single-end RNA-sequencing data. *BMC Genomics* **23**, (2022).
86. Jakobi, T., Uvarovskii, A. & Dieterich, C. Circtools—a one-stop software solution for circular RNA research. *Bioinformatics* **35**, 2326–2328 (2019).
87. Jia, G. yi *et al.* CircRNAFisher: a systematic computational approach for de novo circular RNA identification. *Acta Pharmacol Sin* **40**, 55–63 (2019).
88. Nisar, S. *et al.* Insights Into the Role of CircRNAs: Biogenesis, Characterization, Functional, and Clinical Impact in Human Malignancies. *Frontiers in Cell and Developmental Biology* vol. 9 Preprint at https://doi.org/10.3389/fcell.2021.617281 (2021).
89. Rebolledo, C., Silva, J. P., Saavedra, N. & Maracaja-Coutinho, V. Computational approaches for circRNAs prediction and in silico characterization. *Brief Bioinform* **24**, bbad154 (2023).
90. Szabo, L. *et al.* Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol* **16**, 126 (2015).
91. Gaffo, E., Buratin, A., Dal Molin, A. & Bortoluzzi, S. Sensitive, reliable and robust circRNA detection from RNA-seq with CirComPara2. *Brief Bioinform* **23**, bbab418 (2022).
92. Pan, X. *et al.* WebCircRNA: Classifying the Circular RNA Potential of Coding and Noncoding RNA. *Genes (Basel)* **9**, 536 (2018).
93. Chaabane, M., Williams, R. M., Stephens, A. T. & Park, J. W. circDeep: deep learning approach for circular RNA classification from other long non-coding RNA. *Bioinformatics* **36**, 73–80 (2020).
94. Glažar, P., Papavasileiou, P. & Rajewsky, N. CircBase: A database for circular RNAs. *RNA* **20**, 1666–1670 (2014).
95. Dong, R., Ma, X. K., Li, G. W. & Yang, L. CIRCpedia v2: An Updated Database for Comprehensive Circular RNA Annotation and Expression Comparison. *Genomics Proteomics Bioinformatics* **16**, 226–233 (2018).
96. Glažar, P., Papavasileiou, P. & Rajewsky, N. CircBase: A database for circular RNAs. *RNA* **20**, 1666–1670 (2014).
97. Chen, X. *et al.* CircRNADb: A comprehensive database for human circular RNAs with protein-coding annotations. *Sci Rep* **6**, (2016).
98. Fan, C. *et al.* CircR2Disease v2.0: An Updated Web Server for Experimentally Validated circRNA–disease Associations and Its Application. *Genomics Proteomics Bioinformatics* **20**, 435–445 (2022).
99. Wang, Y. *et al.* A machine learning framework for accurately recognizing circular RNAs for clinical decision-supporting. *BMC Med Inform Decis Mak* **20**, (2020).
100. Zeng, X., Lin, W., Guo, M. & Zou, Q. A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Computational Biology* vol. 13 Preprint at https://doi.org/10.1371/journal.pcbi.1005420 (2017).
101. Hansen, T. B., Venø, M. T., Damgaard, C. K. & Kjems, J. Comparison of circular RNA prediction tools. *Nucleic Acids Res* **44**, (2015).

102. Dang, J. W., Tiwari, S. K., Qin, Y. & Rana, T. M. Genome-wide Integrative Analysis of Zika-Virus-Infected Neuronal Stem Cells Reveals Roles for MicroRNAs in Cell Cycle and Stemness. *Cell Rep* **27**, 3618-3628.e5 (2019).
103. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
104. Kim, D. & Salzberg, S. L. TopHat-Fusion: An algorithm for discovery of novel fusion transcripts. *Genome Biol* **12**, (2011).
105. Zhang, X.-O. *et al.* Complementary Sequence-Mediated Exon Circularization. *Cell* **159**, 134–147 (2014).
106. Dobin, A. *STAR manual 2.7.11a*. (2023).
107. Srivastava, A., Sarkar, H., Gupta, N. & Patro, R. RapMap: A rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics* **32**, i192–i200 (2016).
108. Bioconductor. http://bioconductor.unipi.it/packages/3.15/bioc/html/GenomicFeatures.html.
109. Sepulveda, J. Using R and Bioconductor in Clinical Genomics and Transcriptomics. *The Journal of Molecular Diagnostics* **22**, (2019).
110. Zhang, J., Chen, S., Yang, J. & Zhao, F. Accurate quantification of circular RNAs identifies extensive circular isoform switching events. *Nat Commun* **11**, (2020).
111. Zhang, Y., Park, C., Bennett, C., Thornton, M. & Kim, D. Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N. *Genome Res* **31**, 1290–1295 (2021).
112. Pan, B. *et al.* Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinformatics* **20**, (2019).
113. Xiao-Ou Zhang & Li Yang. CIRCexplorer: Installation and Setup . https://circexplorer2.readthedocs.io/en/latest/tutorial/setup/.
114. Dang, J. W., Tiwari, S. K., Qin, Y. & Rana, T. M. Genome-wide Integrative Analysis of Zika-Virus-Infected Neuronal Stem Cells Reveals Roles for MicroRNAs in Cell Cycle and Stemness. *Cell Rep* **27**, 3618-3628.e5 (2019).
115. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
116. Mathanakumara Ealam Selvan, Kai Shen Lim, Chee How Teo & Yat-Yuen Lim. In Silico Identification and Characterization of circRNAs During Host-Pathogen Interactions.
117. Dun, M. D. *et al.* Proteotranscriptomic profiling of 231-BR breast cancer cells: Identification of potential biomarkers and therapeutic targets for brain metastasis. *Molecular and Cellular Proteomics* **14**, 2316–2330 (2015).
118. RNeasy Plus Kits. https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/rna-purification/total-rna/rneasy-plus-kits.