



UNIVERSITÀ DEGLI STUDI DI MILANO - BICOCCA

Scuola di Scienze

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di laurea in Informatica

Confronto di strumenti software per rilevare pattern di splicing alternativo

Relatore: Prof. Gianluca Della Vedova

Co-relatore: Prof. Raffaella Rizzi

Relazione della prova finale di:
Mattia Previtali
Matricola 807564

Anno Accademico 2017-2018

Sommario

Questa tesi ha il compito di mettere a confronto fra loro tre diversi strumenti software, che hanno lo scopo di individuare eventi di splicing alternativo. I software considerati sono ASGAL, SplAdder ed rMATS. Per fare questo si è ritenuto importante capire quali fossero gli eventi riconosciuti da un tool rispetto all'altro a partire da un input comune. Per riuscire ad armonizzare i dati in possesso, al fine di avere un input comune, si è provveduto a generare dei sample di read di RNA con il simulatore noto con il nome di RNASeqReadSimulator. È stato inoltre necessario fornire un genoma di riferimento e l'annotazione dei geni interessati ad ogni tool analizzato. Queste tre informazioni (genoma, annotazione, sample di read di RNA) sono poi state manipolate in modo tale da poter essere usate come input per i tre strumenti software. In primo luogo si è voluto analizzare eventi semplici, molto frequenti nell'espressione genica, per poi passare ad eventi più complessi e quindi più rari. Lo studio è stato in particolare effettuato sul cromosoma Y, avendo come genoma di riferimento quello umano. Si è giunti alla conclusione che, oltre alle differenze implementative, ci sono delle differenze fra l'utilizzo di un software rispetto all'altro in termini di output e quindi di eventi rilevati e no. Ad ogni modo è importante sottolineare che ad influenzare i risultati è stato anche l'utilizzo di uno specifico software di allineamento delle read simulate. In questo caso si è usato il tool STAR, non prendendo in considerazione altri software di allineamento.

Indice

Sommario	I
1 Introduzione	1
1.1 Importanza dello studio dello splicing alternativo	1
2 Contesto biologico	3
2.1 Espressione genica	3
2.2 Splicing alternativo	6
3 Approccio informatico	8
3.1 Approccio informatico allo splicing alternativo	8
3.1.1 Allineamento di sequenze	8
3.1.2 Contenuto e conservazione di sequenze	10
3.1.3 Microarray	11
3.1.4 Sequenziamento ad alta capacità produttiva	12
3.2 Conoscenze apprese durante lo stage	13
3.2.1 Nomenclatura per eventi di splicing	13
3.2.2 Pipeline generica per utilizzo dei tool messi a confronto	14
3.3 Eventi di splicing considerati	16
3.4 Descrizione dei tool utilizzati	16
3.4.1 AStalavista	16
3.4.2 RNASeqReadSimulator	18

3.4.3	STAR	19
3.4.4	ASGAL	21
3.4.5	SplAdder	22
3.4.6	rMATS	23
4	Risultati	26
4.1	Pipeline	26
4.1.1	Pipeline per utilizzo di ASGAL	26
4.1.2	Pipeline per utilizzo di SplAdder	27
4.1.3	Pipeline per utilizzo di rMATS	28
4.1.4	Riproducibilità	29
4.2	Confronto di eventi semplici	30
4.3	Confronto di eventi complessi	33
5	Conclusioni	36
	Riferimenti bibliografici	37

Elenco delle figure

2.1	Codoni mRNA codificanti per traduzione	6
2.2	Principali tipi di splicing alternativo	7
3.1	Splice graph	9
3.2	Punteggio di conservazione flanking regions	11
3.3	Tipologie di microarray	12
3.4	Esempi di utilizzo codice AS di AStalavista	14
4.1	Esempio confronto output dei tre tool utilizzati per evento semplice	31
4.2	Utilizzo di IGV su evento semplice di exon skipping	31
4.3	Sashimi plot di IGV su evento semplice di exon skipping	32
4.4	Esempio confronto output dei tre tool utilizzati per evento complesso	34

Elenco delle tabelle

3.1	Eventi considerati	17
4.1	Eventi AS semplici analizzati	30
4.2	Comportamento di ASGAL, SplAdder e rMATS su eventi semplici	32
4.3	Eventi AS complessi analizzati	33
4.4	Comportamento di ASGAL, SplAdder e rMATS su eventi complessi	35

Capitolo 1

Introduzione

1.1 Importanza dello studio dello splicing alternativo

Lo splicing alternativo è un processo che abilita l'RNA messaggero (mRNA) alla codifica di trascritti diversi per la sintesi di differenti proteine, con funzioni e proprietà diversificate. Si verifica attraverso la ricombinazione di introni ed esoni di un particolare gene presente sulla catena di DNA in questione [1]. Numerosi studi hanno rivendicato il fondamentale e critico ruolo dello splicing alternativo per quanto riguarda i sistemi biologici. Lo splicing alternativo è responsabile di diversi processi biologici lungo l'intero arco di vita degli organismi. Esso svolge un significativo ruolo funzionale nella differenziazione delle specie e nell'evoluzione del genoma, nonché nello sviluppo di tessuti funzionalmente semplici o complessi con diversi tipi di cellule, come per il cervello, i testicoli e il sistema immunitario. Lo splicing alternativo partecipa anche all'elaborazione stessa dell'RNA per gli eventi pre e post trascrizionali. Nel complesso lo splicing alternativo è un elemento centrale nell'espressione genica. [2]

Gli eventi di splicing alternativo sono osservati frequentemente quando si parla di cancro ed assumono un ruolo importante sia durante l'avanzamento del tumore sia sul fronte terapeutico. Ad ogni modo il loro impatto funzionale e la loro rilevanza inerente alla tumorigenesi rimangono per lo più sconosciuti. [3] Al momento sono state riportate ancora poche sperimentazioni cliniche basate sul controllo terapeutico dello splicing alternativo.

Nonostante questo si ritiene che lo splicing alternativo abbia un potenziale terapeutico in patologie come il cancro, le malattie degli occhi, le malattie ereditarie, come l'atrofia spinale muscolare, il colangiocarcinoma e le infezioni virali. Lo studio relativo allo splicing alternativo e al suo obiettivo terapeutico non può limitarsi a queste patologie dal momento che fa riferimento ad aree divergenti come i processi fisiologici, le patologie, le conseguenze avverse all'utilizzo dei medicinali. È importante sottolineare che lo splicing alternativo, in opposizione allo splicing costitutivo, potrebbe essere un approccio terapeutico non sfruttato in settori quali la nocicezione e/o il dolore, e la neuroprotezione. [4]

È bene sottolineare che ci sono stati alcuni tentativi nel riconoscere su larga scala i geni dei quali gli eventi di splicing alternativo fossero affetti da vari tipi di cancro. Si è anche cercato di riprodurre le condizioni per lo sviluppo di tali eventi al fine di una diagnosi preventiva delle relative malattie. Per fare degli esempi in questo senso, sono stati condotti studi per episodi di cancro alle ovaie e al seno, nei quali il profilo di 600 geni correlati al cancro provenienti da 26 differenti tipi di tessuti cancerogeni sono stati analizzati. Queste analisi hanno portato ad una validazione di 41 eventi di splicing alternativo, costantemente presenti in tutti i tessuti cancerogeni presi in considerazione, da poter utilizzare come marcatori molecolari. In uno studio simile, eventi differenti di splicing sono stati attribuiti a 29 persone a cui è stato diagnosticato un cancro al polmone. In questo caso, dopo aver analizzato 5183 eventi di splicing si è concluso che 4 di questi erano presenti nella maggior parte dei pazienti. In un altro tipo di studi invece è stato messo sotto analisi il legame fra l'effetto di numerosi farmaci impiegati in chemioterapia e la regolazione di eventi di splicing di geni coinvolti nel cancro. Nonostante i risultati promettenti, c'è ancora molto da approfondire soprattutto relativamente ai meccanismi coinvolti nei diversi tipi di tumore. [5]

Per i motivi sopra citati lo splicing alternativo si crede abbia un ruolo chiave in medicina, e per questo si necessita di ulteriori studi e approfondimenti al riguardo.

Capitolo 2

Contesto biologico

2.1 Espressione genica

Il DNA (Acido Desossiribonucleico) è essenzialmente una molecola di stoccaggio. Contiene tutte le istruzioni di cui una cellula ha bisogno per sopravvivere. Queste istruzioni si trovano all'interno dei geni, che sono sezioni di DNA costituite da sequenze specifiche di nucleotidi. Per poter essere seguite, le istruzioni contenute nei geni, devono essere espresse in una forma che può essere utilizzata dalle cellule per produrre le proteine di cui hanno bisogno per sostenere la vita. Le istruzioni memorizzate nel DNA vengono lette e processate da una cellula in due passi: trascrizione e traduzione.

Durante la trascrizione, una parte del DNA serve come modello per la creazione di una molecola di RNA (Acido Ribonucleico). In alcuni casi la nuova molecola di RNA è pronta per essere utilizzata così com'è, svolgendo una funzione importante all'interno della cellula. In altri casi questa molecola ha il compito di portare dei messaggi ad alcune aree della cellula. Quando si verifica tale situazione lo specifico tipo di RNA viene chiamato mRNA (RNA messaggero). La trascrizione ha inizio quando un enzima, detto RNA polimerasi, si ancora alla catena di DNA che viene usata come modello ed inizia a produrre una nuova catena complementare di RNA. Questo enzima si muove lungo la catena di DNA seguendo la sua direzione 3' - 5' e producendo le basi complementari della catena di RNA in direzione 5' - 3'. In questa fase iniziale l'RNA polimerasi si lega ad una specifica

area del DNA, detta regione promotrice, che spesso include una sequenza nucleotidica specializzata: TATAAA. Sul DNA si trovano quattro diversi tipi di basi azotate: Adenina (A), Citosina (C), Guanina (G), Timina (T). Queste basi costituiscono anche l'RNA ad eccezione della Timina, che viene rimpiazzata dall'Uracile (U). Durante il processo di trascrizione vengono utilizzate le seguenti regole di complementarità: $A \rightarrow U$, $C \rightarrow G$, $G \rightarrow C$, $U \rightarrow A$. Quindi se sulla catena di DNA si trova una base azotata di Adenina (A), sulla catena di RNA in costruzione verrà posizionata una base azotata di Uracile (U). Questa sostituzione di basi non è l'unica differenza fra il DNA e l'RNA, infatti l'RNA è costituito da una sola catena non elicoidale al contrario del DNA che ha una struttura a doppia elica. C'è anche da sottolineare che l'RNA contiene molecole di ribosio, a differenza di quelle di desossiribosio del DNA. La molecola di RNA messaggero appena creata (mRNA) ha il compito di portare informazioni dal nucleo della cellula ai ribosomi presenti nel citoplasma, dove verranno assemblate le proteine. Tuttavia, prima di poter compiere questa funzione, la catena di mRNA deve essere separata dalla catena di DNA utilizzata come modello e, in alcuni casi, deve essere sottoposta ad un processo di ordinamento. Questo processo, che mette termine alla trascrizione negli eucarioti, può avvenire in modi differenti. In alcuni casi è proprio la polimerasi che interrompe la trascrizione una volta trovata una specifica sequenza di basi nucleotidiche sul DNA, conosciuta come sequenza di terminazione. In altri casi è la presenza di una speciale proteina, nota come fattore di terminazione, che ha il ruolo di porre fine al processo di trascrizione. A questo punto, una volta che la molecola di mRNA si è staccata dalla catena di DNA usata come modello, viene sottoposta ad un processo, noto come splicing, durante il quale le parti di mRNA non codificanti per la costruzione delle future proteine, chiamate introni, vengono rimosse. Dopodiché viene aggiunta una sequenza di nucleotidi di adenina (A) alla fine (3') della molecola di mRNA. Questa sequenza comunica alla cellula che la molecola di mRNA è pronta per lasciare il nucleo ed entrare nel citoplasma. [6]

Il secondo passo per l'interpretazione delle informazioni memorizzate all'interno del DNA è la traduzione. La lettura della molecola di mRNA viene eseguita da una struttura proteico-sintetizzante nota come ribosoma. Prima di descrivere come avviene il processo

di traduzione è utile illustrare cosa sia il codice genetico. Il codice genetico è l'insieme di regole che una cellula usa per interpretare la sequenza di nucleotidi all'interno dell'mRNA. Questa sequenza viene suddivisa in una serie di triplette di nucleotidi conosciute come codoni. Dal momento che i codoni sono costituiti da tre nucleotidi, questo significa che le quattro basi azotate che si trovano nell'mRNA (A, C, G, U) possono produrre un totale di 64 differenti combinazioni. Di questi 64 codoni, 61 rappresentano gli aminoacidi e i rimanenti tre rappresentano segnali di stop, che segnalano la fine della sintesi di proteine. Dal momento che ci sono 20 differenti aminoacidi, ma 64 possibili codoni, la maggior parte degli aminoacidi corrispondono a più di un codone. È da sottolineare inoltre che ogni codone rappresenta un solo aminoacido o segnale di stop. Questo fenomeno di ridondanza è importante perché minimizza gli effetti dannosi che nucleotidi posizionati erroneamente possono avere sulla sintesi proteica. Un altro importante fattore è che all'interno del codice genetico non si verifica il fenomeno di sovrapposizione e quindi i nucleotidi di un particolare codone fanno parte esclusivamente di quel codone e di nessun altro dei codoni adiacenti. La traduzione inizia quando il ribosoma addetto a tale funzione si attacca alla catena di mRNA nel punto in cui trova un codone che rappresenta il segnale di inizio. Solitamente si tratta del codone AUG, rappresentante l'aminoacido metionina. I codoni responsabili dei segnali di fine traduzione invece sono così composti: UAA, UAG, UGA. Tutti gli altri codoni, rappresentanti gli aminoacidi, sono illustrati in [Figura 2.1].

Grazie al processo di traduzione si è quindi costituita una catena di aminoacidi. L'insieme ordinato di questi aminoacidi costituisce una proteina. Per molte proteine il processo di traduzione è solo il primo step del loro ciclo di vita. Alcune di esse infatti necessitano di ulteriori modifiche prima di essere considerate proteine complete. Una volta che una proteina è considerata completa, è pronta per portare a termine il suo compito. Alcune proteine sono enzimi che fanno da catalizzatori di reazioni biochimiche. Altre proteine hanno un ruolo nella replicazione e nella trascrizione del DNA. Altre ancora provvedono a fornire un sostegno strutturale per la cellula, a creare canali attraverso le membrane della cellula oppure a svolgere una delle tante altre funzioni di supporto alla cellula stessa. [7]

		Second nucleotide					
		U	C	A	G		
First nucleotide	U	UUU Phe UUC UUA Leu UUG	UCU UCC Ser UCA UCG	UAU Tyr UAC UAA STOP UAG STOP	UGU Cys UGC UGA STOP UGG Trp	U C A G	
	C	CUU CUC Leu CUA CUG	CCU CCC Pro CCA CCG	CAU His CAC CAA Gln CAG	CGU CGC Arg CGA CGG	U C A G	
	A	AUU Ile AUC AUA AUG Met	ACU ACC Thr ACA ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G	
	G	GUU GUC Val GUA GUG	GCU GCC Ala GCA GCG	GAU Asp GAC GAA Glu GAG	GGU GGC Gly GGA GGG	U C A G	
		Third nucleotide					

Figura 2.1: Gli aminoacidi codificati da ogni codone di mRNA. Più codoni possono codificare per lo stesso aminoacido.

2.2 Splicing alternativo

Il processo di splicing, noto come splicing costitutivo, si verifica nel momento in cui una molecola di mRNA si stacca dal modello di DNA utilizzato per la trascrizione. In questo frangente, prima che la catena di RNA messaggero venga sottoposta alla traduzione in aminoacidi, questa viene spezzata e ricomposta a seconda delle porzioni codificanti, esoni, e non codificanti, introni, presenti al suo interno. È importante evidenziare che il prodotto di questa selezione, a partire dalla stessa catena di mRNA, non è univoco. Infatti è possibile che durante il processo di splicing vengano prese in considerazione porzioni di mRNA che non sarebbero codificanti oppure scartate sezioni che avrebbero ricoperto tale ruolo, gli esoni. Questa alternativa allo splicing costitutivo è meglio nota con il nome di splicing alternativo.

Questo processo è mediato da un grosso complesso enzimatico, lo spliceosoma. Oltre allo splicing costitutivo [Figura 2.2.A] si possono classificare cinque tipi principali di splicing alternativo [Figura 2.2]. Il pattern più ricorrente (ca. 30%) per vertebrati e invertebrati è quello noto come exon skipping (salto dell'esone) [Figura 2.2.C]. Quando si parla di

exon skipping, un esone non viene considerato nella fase di splicing. Altro tipo di splicing alternativo è quello noto come intron retention (retenzione di un introne) [Figura 2.2.F], mediante il quale un introne non viene preso in considerazione. Questo tipo di evento nei trascritti umani si trova principalmente nelle regioni non tradotte (UTR) ed è associato ai siti di giunzione più deboli, ad una lunghezza ridotta dell'introne e alla regolazione degli elementi cis-regolatori. Questi ultimi sono regioni di DNA non codificante che regolano la trascrizione dei geni vicini. Altra tipologia ricorrente (ca. 25%) è l'alternative selection (selezione alternativa) dei siti donatori o accettori. In questo caso viene considerato solo un suffisso di uno specifico esone se si tratta di alternative acceptor site [Figura 2.2.E] oppure un suo prefisso se si sta parlando di alternative donor site [Figura 2.2.D]. Altro tipo di pattern noto è l'alternanza mutualmente esclusiva di esoni [Figura 2.2.B]. In questa situazione, dati due esone A e B, si verificano eventi in cui viene preso in considerazione un esone A e non un esone B e altri eventi in cui l'esone B sarà sezione codificante al contrario dell'esone A. [2]

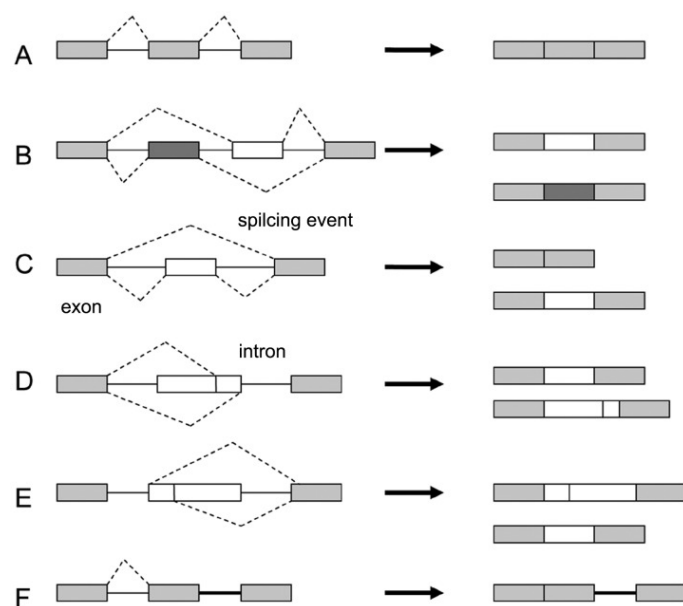


Figura 2.2: Principali tipi di splicing: costitutivo (A), alternanza mutualmente esclusiva di esoni (B), exon skipping (C), alternative donor site (D), alternative acceptor site (E), intron retention (F).

Capitolo 3

Approccio informatico

3.1 Approccio informatico allo splicing alternativo

Dal punto di vista informatico esistono metodi diversi per poter individuare eventi di splicing alternativo. Qui di seguito ne vengono riportati alcuni. [8]

3.1.1 Allineamento di sequenze

Un metodo utilizzato è quello che si basa sull'allineamento di sequenze EST con sequenze genomiche e di mRNA. Le sequenze EST sono piccoli frammenti delle sequenze trascritte di cDNA, solitamente con una lunghezza di 300-400 basi. Il cDNA è un DNA a doppia elica sintetizzato a partire da un campione di RNA messaggero maturo. Le sequenze EST sono prodotte dal sequenziamento di entrambi i filamenti di un clone di mRNA. Nel database pubblico dbEST sono state depositate circa 61 milioni di sequenze EST (aprile 2009). Sono stati sviluppati programmi per allineare le sequenze EST alle sequenze complete di genoma in modo efficiente. Una volta in possesso di tali allineamenti, si è in grado di contrassegnare le posizioni di esoni e introni. I confronti delle strutture esone-introne distinguono ulteriormente gli eventi di splicing alternativo. In alcuni casi una sequenza EST può essere mappata su più di una posizione del genoma di riferimento con un punteggio di allineamento elevato. Questi allineamenti possono essere poi corretti considerando i siti di giunzione (splice). Sebbene l'approccio per l'allineamento di sequenze ha portato ad un

importante progresso nella rilevazione di eventi di splicing alternativo, rimangono problemi nel trattare le giunzioni non canoniche, nel rilevare piccoli esoni, nel gestire gli errori di sequenziamento delle EST e gli errori di polarizzazione relativi alla preparazione delle EST, e altro ancora. Altre limitazioni riguardano l'insufficiente copertura di sequenze per alcuni trascritti e il campionamento distorto di un limitato numero di tipi di cellule e tessuti. Dopo aver individuato i singoli eventi di splicing alternativo, un compito più complesso è quello della costruzione degli interi trascritti corrispondenti. Per questo motivo è stato introdotto il concetto dello Splice Graph. Questo tipo di grafo rappresenta un gene con un grafo aciclico all'interno del quale gli esoni sono i vertici e ogni giunzione corrisponde ad un arco orientato fra due esoni [Figura 3.1].

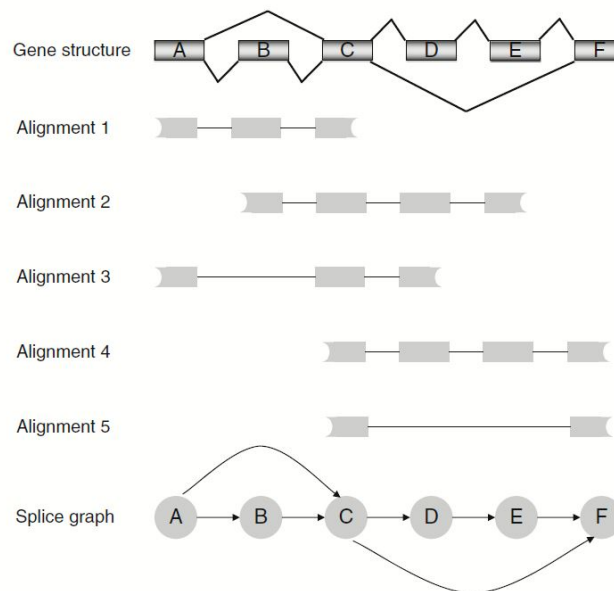


Figura 3.1: Splice graph costruito a partire da allineamenti EST di un genoma di riferimento. Sono mostrati anche la struttura del gene e gli allineamenti considerati.

Le diverse varianti di splicing possono essere dedotte grazie ad algoritmi specifici che operano sui grafi, partendo da un vertice senza archi entranti e arrivando ad un vertice senza archi uscenti. Grazie ad uno Splice Graph può essere elencato un vasto numero di potenziali varianti, ma molte di queste potrebbero essere dovute a costruzioni artificiali, senza quindi avere un corrispettivo significato biologico. Gli esoni infatti non sono uniti

fra di loro in modo casuale al fine di produrre tutte le possibili combinazioni di trascritti isomorfi. Sono stati proposti diversi metodi per selezionare o prioritizzare i trascritti che avessero la più alta probabilità di essere reali, date le sequenze osservate. Per esempio esistono software come AIR, che assegnano un punteggio differente alle diverse varianti di trascritti sulla base di caratteristiche quali la qualità della mappatura, la lunghezza dell'allineamento, l'accuratezza dei segnali di giunzione e il livello di frammentazione degli allineamenti di prova. Dopodiché le varianti con un alto punteggio vengono selezionate per l'annotazione. L'algoritmo ECgene invece valuta ogni possibile variante in base alla qualità della sequenza e al numero di allineamenti di cDNA. Un'altra tipologia di software applica l'algoritmo Expectation-Maximization sulla base del numero di allineamenti osservati lungo il gene considerato. Le prestazioni di questi metodi sono limitate dalla contaminazione delle sequenze EST con frammenti genomici, errori di allineamento, e altro ancora.

3.1.2 Contenuto e conservazione di sequenze

Dal momento che lo splicing alternativo su mRNA è un processo altamente regolamentato, il confronto fra sequenze genomiche potrebbe fornire indizi sull'esistenza di un esone alternativo nei siti con un alta pressione di selezione, ovvero quei siti coinvolti da mutazioni dovute al fenomeno della selezione evolutiva. Sono stati quindi proposti metodi per la predizione di esoni alternativi basati su algoritmi di apprendimento automatico che pongono la loro attenzione su caratteristiche quali il contenuto della sequenza analizzata e la conservazione della stessa. Una sequenza è conservativa se ha un basso numero di mutazioni sulle basi azotate che la costituiscono. In questo contesto vengono utilizzati modelli di Markov e classificatori di Bayes per identificare esoni a partire da una sequenza di introni. Grazie ad alcuni studi effettuati su esoni di umani e topi si è riusciti a comprendere come gli esoni di splicing alternativo tendano ad essere più piccoli, con una lunghezza multipla di tre, rispetto a quelli coinvolti nello splicing costitutivo. Inoltre si è osservato che hanno una più alta conservazione nelle regioni vicine a quelle introniche, note con il nome di flanking regions. In altri studi si è adottato l'algoritmo delle foreste casuali (Random Forests). Le foreste casuali sono costituite da molti alberi di decisione. Ogni albero è un insieme di

funzioni booleane relative a caratteristiche specifiche. Le congiunzioni fra le caratteristiche partizionano i sample di allenamento in gruppi con etichette di classi omogenee. Grazie all'utilizzo di tale algoritmo si sono potute osservare le enormi differenze fra il punteggio di conservazione delle flanking regions di esoni alternativi e di esoni costitutivi [Figura 3.2]. Inoltre si è scoperto come le flanking regions siano coinvolte nella regolazione dello splicing. Sono necessari però ulteriori studi per distinguere la pressione di selezione di esoni alternativi, esoni costitutivi e vincoli su aminoacidi (vincoli spaziali che riguardano il ripiegamento delle proteine).

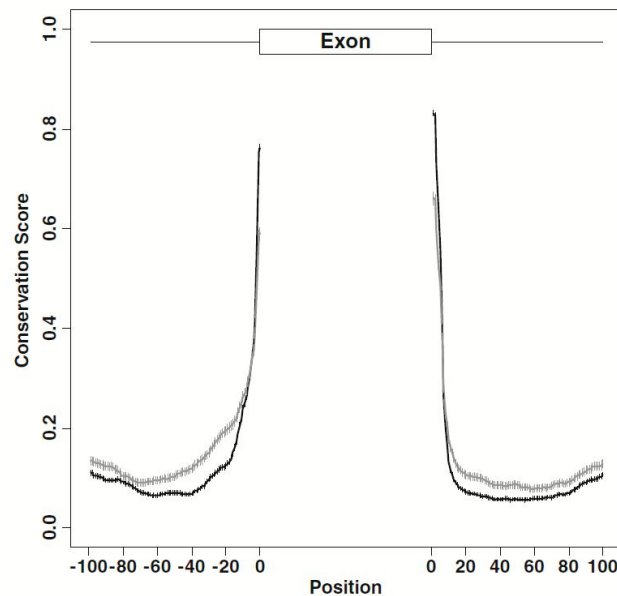


Figura 3.2: Punteggio di conservazione per flanking intronic regions di esoni costitutivi (nero) e esoni alternativi (grigio). Sull'asse delle Y si vede il punteggio di conservazione medio per ogni posizione. Gli intervalli di confidenza rappresentano l'errore standard rispetto alla media.

3.1.3 Microarray

Sebbene gli approcci basati sull'allineamento di sequenze e sul confronto del contenuto e della conservazione di sequenze abbiano dato origine ad un forte progresso nella predizione di eventi, essi forniscono solo un punto di vista qualitativo piuttosto che una visione

quantitativa dello splicing alternativo. Essi forniscono solo le prove dell'esistenza di un evento di splicing alternativo, ma non aggiungono informazioni sulla regolazione spaziale e temporale e nemmeno sul grado di splicing alternativo. La natura altamente parallela dei micorarray rende possibile identificare e quantificare tutti gli eventi di splicing alternativo per uno specifico tessuto oppure le condizioni di malattia rispetto alle condizioni normali di una cellula. Esistono diversi tipi di microarray come si può vedere in [Figura 3.3].

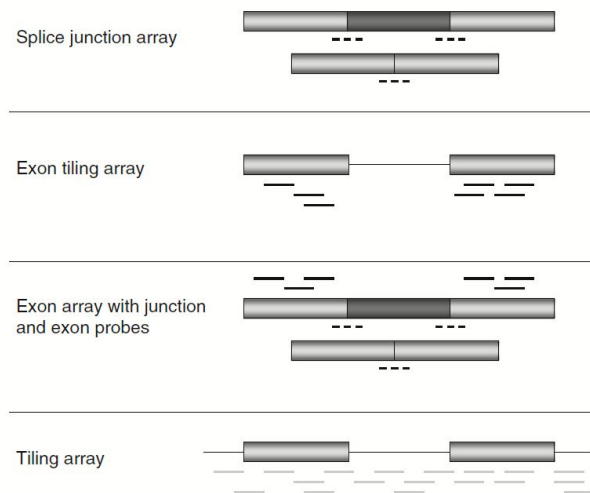


Figura 3.3: Microarray di splicing alternativo. Le linee nere tratteggiate rappresentano le sonde di giunzione. Le linee nere continue rappresentano le sonde per gli esoni. Per i tiling array, le sonde sono distribuite lungo il genoma indipendentemente dalla struttura del gene (linee grigie).

In linea di principio, tutti i tool di analisi dei dati, sviluppati per i microarray genetici standard, possono essere usati per l'analisi di microarray di splicing alternativo. La vera sfida è quella di riuscire a distinguere il segnale di splicing da quello di trascrizione. Alcuni metodi presenti attualmente sono i seguenti: indice di splicing, ANOSVA, FIRMA, DECONV, SPACE, GenASAP.

3.1.4 Sequenziamento ad alta capacità produttiva

Sono stati sviluppati anche approcci basati sul sequenziamento ad alta capacità produttiva (RNA-Seq) per mappare e quantificare i trascrittomi. Le sequenze di mRNA (comprese le

sezioni poliadeniniche) vengono isolate dalle cellule e frammentate in sequenze di dimensione minore (per es. circa 200 basi). Dopodiché vengono convertite in cDNA e sequenziate attraverso le tecniche di sequenziamento ad alta capacità produttiva (High Throughput Sequencing). A questo punto le read provenienti dalle macchine di sequenziamento vengono mappate sui geni e usate come una misura quantitativa del livello di espressione. Le varie tecniche per l'identificazione di eventi di splicing alternativo basate su sequenziamento ad alta capacità produttiva hanno un concetto comune: viene introdotta una variabile latente per poter effettuare una diretta selezione statistica dei differenti trascritti isomorfi espressi.

3.2 Conoscenze apprese durante lo stage

3.2.1 Nomenclatura per eventi di splicing

Durante l'esperienza di stage è stata appresa la nomenclatura adottata dal software AStalavista per identificare i diversi eventi di splicing alternativo analizzati. Solitamente gli eventi vengono categorizzati con il loro nome e vengono illustrati con alcuni semplici schemi. Gli autori di AStalavista hanno però ritenuto importante adottare un codice univoco per riferirsi ai singoli eventi in modo tale da non avere ambiguità dal momento che in alcuni casi si può usare più di una classificazione per rappresentare lo stesso evento. Inoltre, aggiungono gli autori, la nomenclatura tradizionale, si limita agli eventi menzionati più di frequente in letteratura. In questo modo quindi gli eventi che presentano un'elevata complessità non possono essere descritti esclusivamente con termini comuni. Viene inoltre sottolineato che, a seconda dell'annotazione che viene adottata, circa tra un quinto e un terzo delle diverse varianti di splicing vengono trascurate. Nella nomenclatura utilizzata da AStalavista ad ogni evento è assegnato un codice AS, sulla base delle relative posizioni dei siti di splice coinvolti nella variante analizzata. Ad ogni sito di splice viene assegnato un numero incrementale a partire da 1 in accordo con la relativa posizione nell'evento analizzato e un simbolo che dipende dal tipo di sito. Per denotare un sito donatore viene utilizzato il simbolo “^” (raffigurante l'introne a valle del sito donatore), mentre viene

utilizzato il simbolo “-” per caratterizzare un sito accettore. Per esempio un sito di splice denotato da “3-” indica che il terzo sito dell’evento considerato è un sito accettore. Il codice AS è progettato in modo tale che i siti di splice di un trascritto siano separati da una virgola da quelli di un altro trascritto. Uno 0 viene utilizzato nel caso in cui non siano coinvolti siti di splice, come per esempio in un exon skipping. In [Figura 3.4] sono illustrati alcuni esempi di eventi con il relativo codice AS. [9]

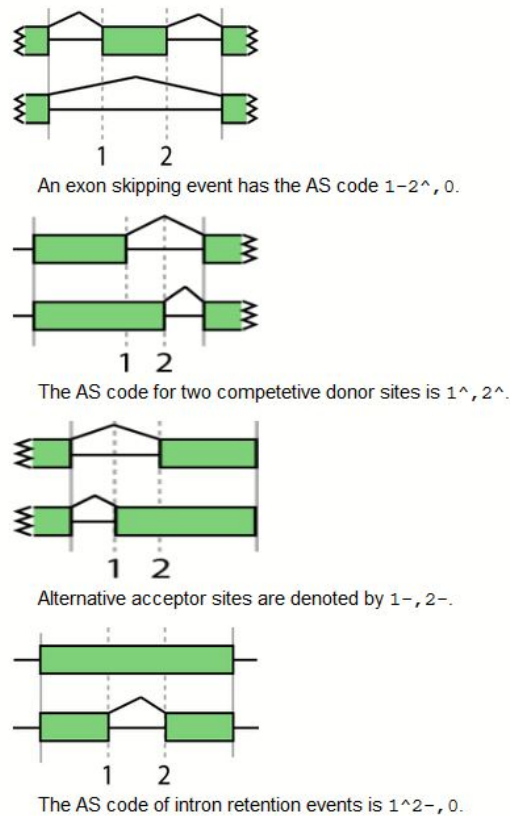


Figura 3.4: Questi sono alcuni esempi di utilizzo del codice AS. In ordine si trovano un evento di exon skipping, un evento di alternative donor site, un evento di alternative acceptor site ed un evento di intron retention.

3.2.2 Pipeline generica per utilizzo dei tool messi a confronto

Per poter mettere a confronto i tre tool citati, che si occupano di rilevare eventi di splicing alternativo, si è dovuta seguire una pipeline operativa specifica. In linea generale si sono

trovati gli eventi di splicing relativi ad un genoma di riferimento e alla sua annotazione, con il software AStalavista. Dopo questo passo si è proceduto ad individuare coppie di trascritti citati negli eventi individuati da AStalavista e si è provato a verificare che tali eventi fossero riconosciuti dai tre software sotto analisi. Entrando più nello specifico, dal momento che si è deciso di fare analisi sul cromosoma Y del genoma umano, si è provveduto a cercare la sequenza genomica di riferimento (file FASTA) dalla banca dati di ENSEMBL [10] e la relativa annotazione genetica (file GTF). Un file FASTA è un file di testo utilizzato per rappresentare sequenze di nucleotidi o peptidi nel quale ogni nucleotide o amminoacido viene codificato mediante un codice a singola lettera [11]. Nel caso in questione le lettere usate sono proprio quelle rappresentative delle basi azotate: A, C, G, T. Viene usata la lettera N per un acido nucleico generico. Un file GTF invece è un file di testo che raccoglie le annotazioni di geni di una particolare sequenza genomica di riferimento. Ogni riga, nota con il nome di feature, può rappresentare le informazioni e le posizioni sulla sequenza genomica, di un esone, di una cds (coding DNA sequence), di un gene, o altro ancora. Viene quindi utilizzato AStalavista, che necessita del file GTF delle annotazioni genetiche, in questo caso relativo al cromosoma Y. Tra gli output di AStalavista si trova una lista di tutti gli eventi di splicing individuati ordinata in base alla frequenza nella sequenza genomica di riferimento, corredata di codice AS e dello schema della struttura dei trascritti corrispondenti. A questo punto si seleziona un evento che dovrà poi essere individuato dai tre differenti tool e si osserva quale gene e quali trascritti sono interessati da tale evento. È poi necessario generare i file GTF, partendo da quello originale del cromosoma Y, che contengano esclusivamente le informazioni dei geni e dei corrispettivi trascritti che si è deciso di considerare. Dal momento che si sta procedendo con una logica inversa rispetto a quanto viene realmente fatto quando vengono utilizzati i software di rilevamento di eventi di splicing alternativo, occorre simulare sample di sequenze di RNA. Per fare questo si è usato il software RNASeqReadSimulator. I sample di RNA in questo caso sono scritti secondo il formato FASTA. Poi, a seconda del tool che viene utilizzato, potrebbe essere opportuno procedere all'allineamento di questi sample. Infine viene utilizzato il tool sotto analisi (ASGAL, SplAdder, rMATS) per verificare l'identificazione dell'evento selezionato

dalla lista inizialmente prodotta da AStalavista. Questo è il processo di lavoro generico. In base al tool di rilevamento di eventi che viene utilizzato si utilizzano input diversi. Questi saranno poi spiegati in dettaglio nel *Capitolo 4 - Risultati* di questa tesi.

3.3 Eventi di splicing considerati

Durante l'analisi dei tre software per l'identificazione di eventi di splicing alternativo, come sopracitato, si è deciso di far riferimento al cromosoma Y del genoma umano, per semplicità strutturale. Sono stati considerati eventi semplici, quindi più frequenti all'interno del cromosoma in questione, e poi eventi più complessi, quindi più rari. In [Tabella 3.1] vengono riportati tali eventi con il codice identificativo dei trascritti coinvolti, l'identificativo del gene di appartenenza, la percentuale di frequenza di tale evento all'interno del cromosoma Y, la nomenclatura AS e lo schema della struttura dei trascritti. Nonostante ci si sia focalizzati sul cromosoma Y si è osservato che le frequenze degli eventi rispettano quelle attese, citate nella parte iniziale di questa tesi.

3.4 Descrizione dei tool utilizzati

Per poter mettere a confronto i tre tool di rilevamento di eventi di splicing alternativo sono stati utilizzati anche software di allineamento di sequenze, come STAR, e di enumerazione di eventi, come AStalavista. Di seguito vengono descritti tutti i software utilizzati durante l'esperienza di stage.

3.4.1 AStalavista

AStalavista (Alternative Splicing Transcriptional Landscape VISualization Tool) è un software che implementa un sistema di notazione intuitivo per identificare eventi di splicing alternativo, forniti in input il genoma di riferimento e l'annotazione corrispondente. Questo strumento rileva anche gli eventi, considerati complessi, sulla base delle posizioni relative ai siti di splice, riportate nell'annotazione fornita come input. In progetti antecedenti a questo software si tendeva invece ad utilizzare matrici di bit per identificare segmenti di










RANK	FREQUENCY (%)	TRANSCRIPT ID		GENE ID	NOMENCLATURE	STRUCTURE
		TRANSCRIPT 1	TRANSCRIPT 2			
1,1	30,0	ENST00000383036	ENST00000215479	ENSG00000099721	1-2^,0	
2,1	11,6	ENST00000640033	ENST00000426950	ENSG00000233803	1^2-,0	
4,1	9,6	ENST00000383008	ENST00000426950	ENSG00000233803	1-,2-	
7,1	3,2	ENST00000382392	ENST00000602549	ENSG00000183795	1^,2^	
12,5	0,8	ENST00000447585	ENST00000545582	ENSG00000240438	1-2^3-4^5-,6-	
12,6	0,8	ENST00000421205	ENST00000635343	ENSG00000224989	1^4-,2^3-	
13,23	0,4	ENST00000317961	ENST00000469599	ENSG00000012817	1^3-4^,2^	
13,24	0,4	ENST00000441139	ENST00000513194	ENSG00000215580	1^3-6^8-,2^4-5^7-	
13,5	0,4	ENST00000624098	ENST00000538878	ENSG00000183878	1^,2^3-4^	

Tabella 3.1: In tabella sono riportati gli eventi considerati in questo studio con la percentuale di frequenza di tale evento all'interno del cromosoma Y, il codice identificativo dei trascritti coinvolti, l'identificativo del gene di appartenenza, la nomenclatura AS e lo schema della struttura dei trascritti.

esoni o introni e descrivere quindi gli eventi di splicing alternativo. AStalavista adotta una definizione generica degli eventi di splicing alternativo e un sistema di notazione flessibile usando il codice AS presentato precedentemente.

In modo particolare AStalavista è stato usato nella fase iniziale di questo studio per poter scegliere gli eventi semplici e complessi da poter analizzare attraverso i tre tool sotto analisi e di conseguenza capire i trascritti coinvolti.

Algoritmo

L'algoritmo implementato da AStalavista consiste, in una parte iniziale, nel considerare tutti i confronti di coppie di trascritti sovrapposti. A questo punto viene rilevata una

variante della struttura di splicing se alcuni siti di splice non sono presenti in entrambi i trascritti. Dopodiché viene utilizzato il relativo ordine dei siti di splice inclusi nelle varianti, secondo le coordinate genomiche, per costruire un codice rappresentativo dell'evento di splicing alternativo corrispondente. Questo approccio, secondo gli autori, permette di non avere i limiti dei metodi focalizzati esclusivamente su eventi semplici e aggira il problema di dover scegliere un trascritto di riferimento, che sarebbe usato come modello per tutti i confronti.

Durante questo studio si è provveduto a considerare il genoma umano (human hg18) e l'annotazione del cromosoma Y, recuperata dall'archivio di ENSEMBL [10].

Output

Dopo l'elaborazione dei dati forniti in input, AStalavista web server mostra una pagina dove ogni tipo di evento identificato è rappresentato dal codice AS e da uno schema visuale. La lista delle tipologie di evento è ordinata in base alla frequenza con la quale tale tipologia è stata rilevata. Inoltre nella parte alta della pagina si può osservare un diagramma a torta che mostra la distribuzione delle tipologie stesse. Da ogni raggruppamento, in base al tipo di evento, si accede poi alla corrispondente lista di trascritti coinvolti. [12]

3.4.2 RNASeqReadSimulator

RNASeqReadSimulator è un insieme di script preposti alla generazione di read RNA-Seq simulate. Questo software consente di assegnare in modo casuale livelli di espressione di trascritti e generare read single-end o paired-end di RNA-Seq. Consente anche di simulare randomicamente errori di read delle piattaforme di sequenziamento.

Input

Per questo studio specifico si è utilizzato il seguente comando con le relative opzioni: *RNASeqSim [genome] [gtf] [fasta_output] [nreads]*. Dove [genome] corrisponde al genoma

di riferimento in formato FASTA, [gtf] è l'annotazione genetica in formato GTF, [fasta_output] corrisponde al nome da assegnare al file FASTA di output e [nreads] sono il numero di read che si desidera generare con RNASeqReadSimulator.

Output

Questo simulatore genera sample di read di RNA-Seq in formato FASTA. [13]

3.4.3 STAR

STAR (Spliced Transcripts Alignment to a Reference) è uno strumento software adepto all'allineamento di sequenze genomiche come le read di RNA. Il suo flusso operativo è costituito da due step. Il primo step consiste nella generazione di indici genomici che vengono poi utilizzati nel secondo step. In questo step è necessario fornire le sequenze genomiche di riferimento, attraverso file FASTA, e le annotazioni, mediante file GTF. Gli indici possono essere generati una volta sola quando si usa la stessa combinazione di genoma e annotazione. Il secondo step invece utilizza gli indici appena creati per mappare le read in possesso dall'utente (in questo caso specifico generate da RNASeqReadSimulator) sul genoma fornito come riferimento. In questa fase, oltre ai file in formato SAM di read allineate, STAR scrive anche diversi file che riepilogano informazioni statistiche, read senza un mapping, e altro ancora.

Opzioni comandi

Le opzioni principali per generare gli indici genomici sono le seguenti:

- `-runThreadN` : <numero di thread utilizzati>
- `-runMode` : `genomeGenerate`
- `-genomeDir` : <percorso della directory dove memorizzare gli indici>
- `-genomeFastaFiles` : <percorso della directory con i file FASTA da analizzare>
- `-sjdbGTFfile` : <percorso della directory con i file GTF da analizzare>

- `-sjdbOverhang` : (lunghezza delle read - 1), solitamente 100 è un buon valore

Per quanto riguarda lo step di mapping invece le opzioni principali sono:

- `-runThreadN` : <numero di thread utilizzati>
- `-genomeDir` : <percorso della directory dove sono stati memorizzati gli indici>
- `-readFilesIn` : <percorso dei file contenenti le sequenze da mappare>

[\[14\]](#)

Algoritmo

L'algoritmo utilizzato da STAR è stato progettato per allineare sequenze non contigue direttamente al genoma di riferimento. È costituito da due parti principali: lo step di seed searching e quello di clustering/stitching/scoring.

L'idea centrale della fase di seed searching è la ricerca sequenziale di un Maximal Mappable Prefix (Massimo Prefisso Mappabile), noto come MMP. Un MMP è simile al concetto di Maximal Exact (Unique) Match (Massimo Match Esatto) usato da tool come Mummer e MAUVE su larga scala. Data una sequenza di read R , una posizione i di read e un genoma di riferimento G , l' $\text{MMP}(R, i, G)$ è definito come la più lunga sottostringa $(R_i, R_{i+1}, \dots, R_{i+\text{MML}-1})$ che ha un match esatto con una o più sottostringhe di G , dove MML è la massima lunghezza mappabile. Questo approccio rappresenta un modo naturale per trovare precise locazioni di giunzioni di splice in una sequenza (read) ed è più vantaggioso rispetto ad una suddivisione arbitraria usata in metodi split-read. Inoltre è bene sottolineare che le giunzioni di splice vengono individuate senza una conoscenza pregressa di proprietà e loci delle stesse, e senza un allineamento preliminare contiguo. STAR implementa la ricerca di un MMP attraverso suffix array non compressi. Trovare un MMP è molto simile alla classica ricerca di una stringa binaria all'interno di suffix array non compressi. Tutto questo permette tempi di esecuzione di scala logaritmica in correlazione alla lunghezza del genoma di riferimento, permettendo quindi ricerche veloci anche con genomi di grandi dimensioni.

Nella seconda fase dell'algoritmo invece, STAR costruisce allineamenti dell'intera sequenza di read assemblando tutti i seed allineati al genoma di riferimento nella prima fase. In un primo momento i seed vengono raggruppati in base alla loro prossimità rispetto a un insieme di seed ancora (anchor-seed). In un secondo momento questi vengono assemblati attraverso un punteggio di allineamento locale. Questo si basa sui match, i mismatch, le inserzioni, le cancellazioni e i gap delle giunzioni. La combinazione di seed con il punteggio di assemblamento maggiore viene scelta come il migliore allineamento di una read. [15]

3.4.4 ASGAL

ASGAL (Alternative Splicing Graph ALigner) è un tool per l'identificazione di eventi di splicing alternativo espressi in un sample di RNA-Seq, non necessariamente allineate, con la rispettiva annotazione dei geni. L'approccio di questo strumento si costituisce di tre step. Il primo step consiste nella costruzione di un grafo di splice a partire dall'annotazione dei geni, che rappresenta la struttura genetica dei trascritti di input. Il secondo step tratta l'allineamento detto splice-aware. In questa fase ASGAL allinea le read di RNA-Seq al grafo di splice precedentemente creato. L'ultimo step invece è la vera e propria individuazione degli eventi di splicing alternativo. ASGAL può riportare come output finale tutti gli eventi trovati oppure solo quelli che non erano presenti nell'annotazione fornita in input, i cosiddetti novel events [16]. I singoli step possono essere eseguiti separatamente grazie ai relativi script. Ad ogni modo viene messa a disposizione dell'utente una pipeline già pronta per poter portare a termine l'elaborazione di tutte le tre fasi con un solo comando [17]. Attualmente ASGAL supporta le seguenti tipologie di evento: exon skipping, alternative acceptor site, alternative donor site, intron retention [16].

Input

ASGAL richiede in input un cromosoma di riferimento in formato FASTA, l'annotazione genetica in formato GTF e un sample di read RNA-Seq in formato FASTA o FASTQ (o compresso con gzip). Inoltre supporta anche un'analisi genomica più ampia di un singolo

cromosoma, ma in questo caso bisogna far riferimento all'apposita pipeline sviluppata dagli autori.

Output

Gli eventi di splicing alternativo individuati vengono scritti all'interno di file a valori separati da spazi dove ogni riga rappresenta un singolo evento. Questi file di output sono così costituiti:

1. tipo di evento (ES,A3,A5,IR);
2. posizioni genomiche di inizio e fine dell'introne che ha portato all'evento di splicing in questione;
3. numero di read che confermano tale evento;
4. lista dei trascritti annotati rispetto ai quali si è espresso l'evento

[17]

3.4.5 SplAdder

SplAdder (Splicing Adder) è un software atto all'analisi di splicing alternativo sulla base di read RNA-Seq allineate. In breve, SplAdder trasforma l'annotazione fornita in input in un grafo di splice e arricchisce tale grafo con le informazioni estratte dalle read RNA-Seq. In questa fase gli elementi aggiunti al grafo sono le intron retention, i cassette exon (per exon skipping) e gli archi dei nuovi introni. Dopodiché estrae gli eventi di splicing alternativo dal grafo e quantifica tali eventi sulla base dei dati di allineamento. Gli eventi di splicing che sono supportati sono: exon skipping, alternative acceptor site, alternative donor site, intron retention, multiple exon skip e mutually exclusive exons. Esistono sia l'implementazione in Matlab che quella in Python. Per questo studio è stata adottata la versione in Python che, a detta degli autori, sarà quella che verrà mantenuta aggiornata in futuro. [18]

Input

SplAdder necessita dei file di input di annotazione genomica e del sample di read RNA-Seq già allineate. Durante questo studio si sono allineate le read con il software STAR, per avere i dati allineati in formato BAM.

Output

SplAdder produce diversi file in output. Di seguito riportiamo solo quelli che l'utente può leggere per comprendere gli eventi di splicing individuati e non i file intermedi che SplAdder stesso utilizza per il processo di elaborazione. I file in formato GFF3 contengono gli eventi che sono stati individuati da SplAdder utilizzando la formattazione tipica dei file GFF3. In questo contesto ogni evento viene mostrato come un mini gene costituito da due differenti trascritti isomorfi. I nomi assegnati a questi file seguono il pattern generico *merge_graphs_<event_type>_C<confidence_level>.confirmed.gff3*. I file in formato HDF5 seguono una memorizzazione gerarchica che consente un'efficiente interrogazione sui dati e una buona interoperabilità fra piattaforme differenti. I file in formato txt contengono essenzialmente le stesse informazioni dei file HDF5 secondo una formattazione a colonne delimitate da tab. [19]

3.4.6 rMATS

rMATS (replicate Multivariate Analysis of Transcript Splicing) è un tool computazionale per identificare differenti eventi di splicing alternativo a partire da dati RNA-seq. Il modello statistico di rMATS calcola il P-value e altre statistiche. rMATS supporta la maggior parte delle tipologie di splicing alternativo come: exon skipping, alternative acceptor site, alternative donor site, intron retention e mutually exclusive exons [20]. L'utilizzo di rMATS ha tre vantaggi essenziali. Innanzitutto rMATS fornisce un rigoroso framework statistico per i biologi al fine di identificare variazioni nello splicing alternativo di ogni tipo di grandezza. In secondo luogo rMATS garantisce robustezza contro il rumore che è presente in dati RNA-Seq. Infine è da sottolineare che la flessibile formulazione di ipotesi permette di testare anche altri tipi di comportamento di splicing alternativo come per esempio il

pattern *switch-like*, secondo il quale un esone è prevalentemente incluso in un trascritto in determinate condizioni ed è invece parte di un altro trascritto in presenza di altre condizioni. [21]

Algoritmo

L'algoritmo utilizzato da rMATS si suddivide in quattro fasi principali. Nel primo step rMATS stima il livello di inclusione di un esone nei due sample forniti in input contando le read RNA-Seq mappate sulle giunzioni esone-esone. In un secondo momento i livelli calcolati nella prima fase vengono usati per costruire un modello preventivo multivariato uniforme che rappresenta la similarità complessiva, a livello di splicing alternativo, fra i due sample utilizzati. Nella terza fase rMATS usa un metodo MCMC (Markov chain Monte Carlo) per calcolare la probabilità Bayesiana per le differenze di splicing. Questo viene fatto basandosi sul modello preventivo multivariato uniforme e su un altro modello di probabilità per i conteggi delle read RNA-Seq. Infine rMATS calcola un P-value per ogni esone comparando la probabilità posteriore osservata, calcolata nel terzo step, con un insieme di probabilità posteriori simulate a partire dalle ipotesi nulle. [21]

Input

In questo studio sono stati utilizzati sample di RNA-Seq in formato BAM generati da STAR. Di conseguenza i dati di input richiesti da rMATS sono i seguenti:

1. file BAM per il primo sample di RNA-Seq
2. file BAM per il secondo sample di RNA-Seq
3. tipo di read utilizzate (paired-end o single-end)
4. lunghezza di ogni read
5. file dell'annotazione genica in formato GTF
6. percorso della directory contenente gli indici STAR utilizzati per l'allineamento delle read dei sample

Output

Fra gli output si trovano i seguenti file:

- `AS_Event.MATS.JC.txt` = valuta lo splicing solamente con le read che riguardano le giunzioni dello splicing
- `AS_Event.MATS.JCEC.txt` = valuta lo splicing con le read che riguardano le giunzioni dello splicing e con le target read
- `fromGTF.AS_Event.txt` = tutti i possibili eventi di splicing alternativo derivatni dall'annotazione GTF e dall'RNA considerato
- `JC.raw.input.AS_Event.txt` = valuta lo splicing solamente con le read che riguardano le giunzioni dello splicing
- `JCEC.raw.input.AS_Event.txt` = valuta lo splicing con le read che riguardano le giunzioni dello splicing e con le target read

In tutti questi diversi tipi di file la parola *AS_Events* è sostituita con la tipologia di evento di splicing alternativo considerata. [22]

Capitolo 4

Risultati

4.1 Pipeline

La pipeline generica adottata per l'utilizzo dei tre diversi tool di identificazione di eventi di splicing alternativo è stata descritta precedentemente. Ora viene spiegato in modo più dettagliato il flusso di lavoro adottato, che poi è stato automatizzato utilizzando uno script in Python, per ognuno di questi software.

4.1.1 Pipeline per utilizzo di ASGAL

Utilizzando ASGAL, dopo aver individuato gli eventi di interesse con AStalavista, occorre testare le diverse combinazioni di file GTF e sample di read RNA-Seq. A tale proposito sono stati generati tre file GTF per l'annotazione dei geni. Il primo file contiene esclusivamente le informazioni del gene e degli esoni coinvolti dal primo trascritto considerato. Il secondo file GTF per l'annotazione dei geni contiene le stesse informazioni relative invece al secondo trascritto coinvolto. Il terzo file GTF è costituito dalle informazioni di entrambi i trascritti, e relativi geni, che hanno preso parte all'evento di splicing alternativo che si sta considerando. Per fare queste operazioni si è utilizzato il comando GREP messo a disposizione dalla BASH di Linux. Sono quindi necessari anche tre diversi sample di read RNA-Seq. Utilizzando il software RNASEqReadSimulator si è generato un sample relativo esclusivamente al primo trascritto coinvolto, un sample relativo al secondo trascritto

coinvolto e un terzo sample che comprendesse le read per entrambi i trascritti considerati dall'evento di splicing alternativo in questione. Avendo quindi tre file GTF e tre sample di read RNA-Seq, si è eseguito ASGAL per nove run differenti. Questo ha permesso di garantire la massima copertura da parte di ASGAL per poter trovare l'evento di splicing alternativo sotto analisi.

Tutto questo processo, dopo essere stato provato manualmente, è stato automatizzato ricorrendo ad un script Python che provvedesse a creare i file GTF, i sample FASTA e gli output delle nove run di ASGAL, utilizzando come input:

- codice id del primo trascritto
- codice id del secondo trascritto
- file GTF dell'intera annotazione del cromosoma Y
- file FASTA del genoma di riferimento per il cromosoma Y

Per ASGAL è importante sottolineare che non c'è bisogno di effettuare allineamenti dei sample preventivamente, dal momento che si tratta di un'operazione già compresa nella pipeline propria di questo software.

4.1.2 Pipeline per utilizzo di SplAdder

La pipeline riguardante SplAdder segue anch'essa le linee generali della pipeline generica ma si differenzia da quella di ASGAL per alcuni motivi. In questo caso infatti è sufficiente avere un unico file GTF per le annotazioni di entrambi i trascritti coinvolti nell'evento di splicing alternativo sotto analisi. Anche il sample di read RNA-Seq da generare con RNASEqReadSimulator è comprensivo sia delle informazioni del primo trascritto che del secondo trascritto. Con SplAdder è necessario però fornire il sample di read dopo aver già eseguito su di esso un processo di allineamento. Per questo motivo lo script Python utilizzato per automatizzare questa serie di processi comprende anche la creazione di indici STAR, nel caso non fossero già esistenti, e l'allineamento vero e proprio fatto con STAR. Inoltre, per utilizzare un sample allineato, SplAdder richiede un indice BAI che

viene costruito sulla base del sample allineato in formato BAM. Per fare questo si è ricorsi all'utilizzo del software SAMTOOLS [23]. La pipeline Python richiede in definitiva i seguenti input:

- codice id del primo trascritto
- codice id del secondo trascritto
- file GTF dell'intera annotazione del cromosoma Y
- file FASTA del genoma di riferimento per il cromosoma Y
- una flag ('y' o 'n') per indicare l'esistenza di un eventuale indice STAR già esistente
- directory dell'eventuale indice STAR già esistente

4.1.3 Pipeline per utilizzo di rMATS

Anche la pipeline adottata per l'utilizzo di rMATS segue la pipeline generica precedentemente spiegata ed è molto simile a quella di SplAdder. In questo caso però rMATS necessita dei sample di read RNA-Seq dei due trascritti coinvolti nell'evento di splicing alternativo considerato, in modo distinto. Per questo motivo, occorre generare il file GTF con l'annotazione per il primo trascritto e il file GTF con l'annotazione per il secondo trascritto. rMATS non richiede il formato BAI dei due sample generati e, oltre al file GTF costituito dalle informazioni di entrambi i trascritti presi in analisi, richiede solamente i file in formato BAM dei sample già allineati dal software STAR. In breve la pipeline Python scritta per utilizzare rMATS ha i seguenti input:

- codice id del primo trascritto
- codice id del secondo trascritto
- file GTF dell'intera annotazione del cromosoma Y
- file FASTA del genoma di riferimento per il cromosoma Y
- una flag ('y' o 'n') per indicare l'esistenza di un eventuale indice STAR già esistente

- directory dell’eventuale indice STAR già esistente

4.1.4 Riproducibilità

Per quanto riguarda la riproducibilità dei test che sono stati eseguiti si può far riferimento all’archivio disponibile su GitHub [24]. È importante sottolineare che gli script Python che sono stati scritti per poter eseguire le tre differenti pipeline necessitano di tutte le dipendenze richieste dai singoli software per poter funzionare. Per sapere di quali dipendenze i tool abbiano bisogno occorre consultare i corrispettivi manuali disponibili online. Inoltre all’inizio di ognuno di questi script vi è una sezione nominata *CONFIGURATION SETTINGS*, all’interno della quale sono stati specificati i percorsi relativi dei software coinvolti. Per poter ripetere quanto fatto occorrerà modificare tali path a seconda delle posizioni, dei singoli software, sul proprio computer. Si è deciso di non parametrizzare anche questi valori sia per non eccedere con gli argomenti richiesti in input sia perché una volta settati, vanno bene per tutte le esecuzioni successive. Durante questo caso specifico si è deciso di generare 9000 read quando viene utilizzato RNASeqReadSimulator e di prendere in considerazione il cromosoma Y, come precedentemente accennato. Anche queste impostazioni possono essere modificate nella stessa sezione degli script. Per questo studio si è utilizzata una macchina Linux (Xubuntu 17.10 - 64bit) con 4GB di RAM e si è ricorsi all’utilizzo della shell BASH per eseguire i diversi comandi e software citati.

Archivio GitHub

L’archivio GitHub comprende:

- tre script Python3 per l’esecuzione delle tre pipeline descritte
- tutti i file di output generati dalle tre pipeline, per ogni evento di splicing alternativo considerato
- nove file in cui sono stati riuniti tutti gli output rilevanti al fine di confrontare i tre software sotto analisi
- il genoma di riferimento, in formato FASTA, del cromosoma Y

- l’annotazione, in formato GTF, del cromosoma Y

4.2 Confronto di eventi semplici

In [Tabella 4.1] vengono riportati gli eventi semplici di splicing alternativo che sono stati considerati.





RANK	FREQUENCY (%)	TRANSCRIPT ID		GENE ID	NOMENCLATURE	STRUCTURE
		TRANSCRIPT 1	TRANSCRIPT 2			
1,1	30,0	ENST00000383036	ENST00000215479	ENSG00000099721	1-2^,0	
2,1	11,6	ENST00000640033	ENST00000426950	ENSG00000233803	1^2-,0	
4,1	9,6	ENST00000383008	ENST00000426950	ENSG00000233803	1-,2-	
7,1	3,2	ENST00000382392	ENST00000602549	ENSG00000183795	1^,2^	

Tabella 4.1: In tabella sono riportati gli eventi semplici di splicing alternativo considerati per una prima analisi dei tre tool utilizzati (ASGAL, SplAdder, rMATS).

Per ognuno degli eventi considerati si sono ottenuti, grazie all’esecuzione delle tre pipeline, diversi file di testo da poter confrontare. Dopodiché si è provveduto a fare un merge di tali output. Viene mostrato un esempio in [Figura 4.1], per l’evento che ha coinvolto i trascritti ENST00000383036 e ENST00000215479.

In questo specifico caso l’evento di exon skipping è stato individuato da tutti i tre tool sotto analisi. ASGAL ha messo in evidenza le posizioni di inizio e fine dell’introne che ha portato all’evento. SplAdder e rMATS invece hanno riportato le posizioni delle sezioni codificanti, esoni, del trascritto ENST00000215479. Per la posizione di inizio dell’esone non considerato dal trascritto ENST00000383036, rMATS adotta una numerazione che parte da zero. Per un’ulteriore verifica si è utilizzato il software IGV. IGV (Integrative Genomics Viewer) è un tool di visualizzazione per l’esplorazione di dataset genomici. Supporta una vasta varietà di tipi di dato, inclusi sequenze array-based e next-generation, e annotazioni

Trascritti: ENST00000383036 e ENST00000215479

ASGAL:

Type,	Start,	End,	Support,	Transcripts
ES,	6868777,	6870005,	617,	ENST00000383036
Type,	Start,	End,	Support,	Transcripts
ES,	6868777,	6870005,	335,	ENST00000383036

SPLADDER:

contig	strand	event_id	gene_name	exon_pre_start	exon_pre_end	exon_start	exon_end	exon_aft_start	exon_aft_end
exon_aft_end	Aligned.sortedByCoord.out:exon_pre_cov	Aligned.sortedByCoord.out:exon_cov							
Aligned.sortedByCoord.out:intron_pre_conf	Aligned.sortedByCoord.out:intron_aft_conf								
Aligned.sortedByCoord.out:intron_skip_conf	Aligned.sortedByCoord.out:psi								
Y	-	exon_skip_1	ENSG00000099721	6868732	6868776	6868868	6868909	6870006	6870053
1021.2	606.5	849.2	591	546	433	0.56			

RMATS:

ID	GeneID	geneSymbol	chr	strand	exonStart_0base	exonEnd	upstreamES	upstreamEE	downstreamES	downstreamEE
0	"ENSG00000099721"	"AMELY"	chrY	-	6868867	6868909	6868731	6868776	6870005	6870053

Figura 4.1: Questo è un esempio delle informazioni che sono state messe a confronto per capire se i tre tool utilizzati siano stati in grado di riconoscere un evento di splicing alternativo oppure no. Il caso specifico riguarda i trascritti ENST00000383036 e ENST00000215479.

genomiche [25]. La [Figura 4.2] e la [Figura 4.3] mostrano visualmente quanto è stato individuato dai tre software di rilevamento di eventi.

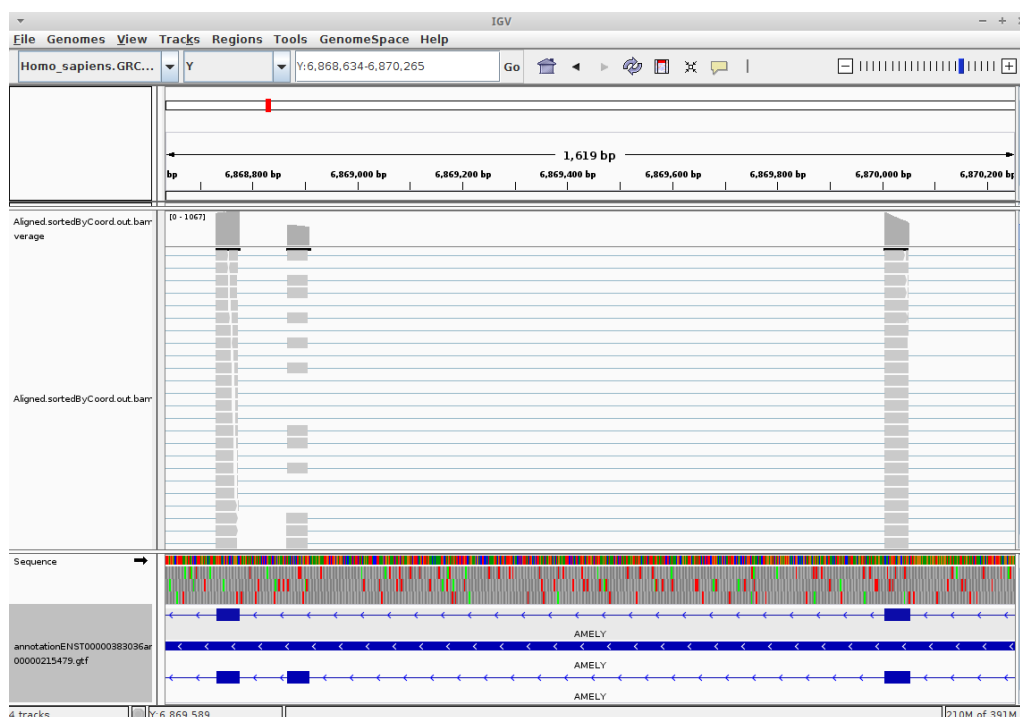


Figura 4.2: In figura viene mostrata la track delle read RNA-Seq allineate relative alla sezione dei due trascritti coinvolta nell'exon skipping. La track in basso mostra invece l'annotazione GTF dei due trascritti coinvolti.

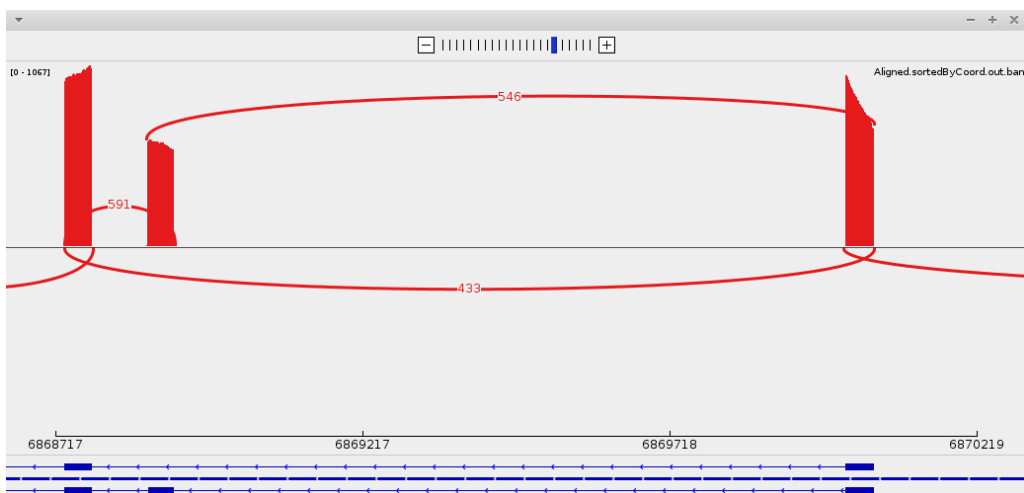


Figura 4.3: La figura mostra il *sashimi plot* generato da IGV sulla parte di exon skipping dei trascritti coinvolti ENST00000383036 e ENST00000215479.

La tabella [Tabella 4.2] riporta gli eventi semplici che sono stati individuati correttamente da ASGAL, SplAdder e rMATS oppure no.

TRANSCRIPT ID		GENE ID	NOMENCLATURE	ASGAL	SPLADDER	RMATS
TRANSCRIPT 1	TRANSCRIPT 2					
ENST00000383036	ENST00000215479	ENSG00000099721	1-2^,0	Found	Found	Found
ENST00000640033	ENST00000426950	ENSG00000233803	1^2-,0	Found	Found	
ENST00000383008	ENST00000426950	ENSG00000233803	1-,2-	Found	Found	Found
ENST00000382392	ENST00000602549	ENSG00000183795	1^,2^	Found	Found	Inaccurate

Tabella 4.2: In tabella sono riportati gli eventi semplici di splicing alternativo che sono stati identificati oppure no da ASGAL, SplAdder e rMATS. *Found* sta per evento trovato mentre la cella vuota sta per evento non trovato. *Inaccurate* significa che sono stati identificati alcuni eventi ma non quelli attesi.

Questi eventi semplici (exon-skipping, intron retention, alternative acceptor site, alternative donor site) sono stati individuati da quasi tutti i tool sotto analisi. Solo un evento

di intron retention e un evento di alternative donor site non sono stati individuati da rMATS].

4.3 Confronto di eventi complessi

In [Tabella 4.3] vengono riportati gli eventi complessi di splicing alternativo che sono stati considerati.






RANK	FREQUENCY (%)	TRANSCRIPT ID		GENE ID	NOMENCLATURE	STRUCTURE
		TRANSCRIPT 1	TRANSCRIPT 2			
12,5	0,8	ENST00000447585	ENST00000545582	ENSG00000240438	1-2^3-4^5-,6-	
12,6	0,8	ENST00000421205	ENST00000635343	ENSG00000224989	1^4-,2^3-	
13,23	0,4	ENST00000317961	ENST00000469599	ENSG00000012817	1^3-4^-,2^	
13,24	0,4	ENST00000441139	ENST00000513194	ENSG00000215580	1^3-6^8-,2^4-5^7-	
13,5	0,4	ENST00000624098	ENST00000538878	ENSG00000183878	1^-,2^3-4^	

Tabella 4.3: In tabella sono riportati gli eventi complessi di splicing alternativo considerati per un'analisi più specifica dei tre tool utilizzati (ASGAL, SplAdder, rMATS).

Come per gli eventi più semplici si è provveduto ad eseguire le tre pipeline per ottenere file output confrontabili. Ora viene mostrato nel dettaglio l'evento di splicing alternativo che ha coinvolto i trascritti ENST00000624098 e ENST00000538878.

ENST00000624098 vs ENST00000538878

In [Figura 4.4] si vede il merge delle informazioni utili alla comprensione dell'evento analizzato. In questo caso specifico l'evento complesso può essere considerato come la combinazione di un alternative donor site ed un exon skipping. ASGAL è stato in grado di riconoscere l'evento di donor site in questione ed altri eventi di exon skipping presenti sui

trascritti. Spladder ed rMATS hanno identificato solo eventi di exon skipping. Nessun tool ha quindi identificato l'evento complesso nella sua interezza.

```

Trascritti: ENST00000624098 e ENST00000538878

ASGAL:

Type,Start,End,Support,Transcripts
ES,13249883,13260277,60,ENST00000538878
A5,13336336,13355374,99,ENST00000538878

Type,Start,End,Support,Transcripts
ES,13249883,13260277,29,ENST00000538878
A5,13336336,13355374,48,ENST00000538878

Type,Start,End,Support,Transcripts
ES,13357970,13359766,66,ENST00000624098

Type,Start,End,Support,Transcripts
ES,13357970,13359766,30,ENST00000624098

SPLADDER:

contig strand event_id gene_name exon_pre_start exon_pre_end exon_start exon_end exon_aft_start
exon_aft_end Aligned.sortedByCoord.out:exon_pre_cov Aligned.sortedByCoord.out:exon_cov Aligned.sortedByCoord.out:exon_aft_cov
Aligned.sortedByCoord.out:intron_pre_conf Aligned.sortedByCoord.out:intron_aft_conf
Aligned.sortedByCoord.out:intron_skip_conf Aligned.sortedByCoord.out:psi
Y - exon_skip_1 ENSG00000183878 13357877 13357969 13359088 13359222 13359767
13359986 121.1 73.6 119.1 86 72 39 0.65

RMATS:

ID GeneID geneSymbol chr strand exonStart_Obase exonEnd upstreamES upstreamEE downstreamES downstreamEE
0 "ENSG00000183878" "UTY" chrY - 13251016 13251187 13248378 13249882 13260277
13260404
1 "ENSG00000183878" "UTY" chrY - 13359087 13359222 13357876 13357969 13359766
13359986

```

Figura 4.4: Questo è un esempio delle informazioni messe a confronto per capire se i tre tool utilizzati siano stati in grado di riconoscere un evento di splicing alternativo complesso oppure no. Il caso specifico riguarda i trascritti ENST00000624098 e ENST00000538878.

Questo tipo di analisi è stata effettuata anche sugli altri eventi complessi precedentemente citati e in [Tabella 4.4] vengono mostrati i risultati.

ENST00000447585 vs ENST00000545582

In questo caso solo ASGAL è riuscito a trovare qualcosa. Sono stati trovati un evento di exon skipping e un alternative donor site, ma non quelli che avrebbero potuto comporre l'evento complesso in questione.

ENST00000421205 vs ENST00000635343

Nessuno tool analizzato ha trovato un evento di splicing alternativo.

TRANSCRIPT ID		GENE ID	NOMENCLATURE	ASGAL	SPLADDER	RMATS
TRANSCRIPT 1	TRANSCRIPT 2					
ENST00000447585	ENST00000545582	ENSG00000240438	1-2^3-4^5-,6-	Incorrect		
ENST00000421205	ENST00000635343	ENSG00000224989	1^4-,2^3-			
ENST00000317961	ENST00000469599	ENSG00000012817	1^3-4^-,2^	Inaccurate	Inaccurate	Inaccurate
ENST00000441139	ENST00000513194	ENSG00000215580	1^3-6^8-,2^4-5^7-	Inaccurate	Inaccurate	
ENST00000624098	ENST00000538878	ENSG00000183878	1^-,2^3-4^	Inaccurate	Inaccurate	Inaccurate

Tabella 4.4: In tabella sono riportati gli eventi complessi di splicing alternativo che sono stati identificati oppure no da ASGAL, SplAdder e rMATS. *Inaccurate* significa che sono stati identificati alcuni eventi, ma non la combinazione esatta di eventi semplici che possa descrivere correttamente l'evento complesso. *Incorrect* significa che è stato identificato un evento in modo inesatto. La presenza di una cella vuota rappresenta un evento complesso non trovato.

ENST00000317961 vs ENST00000469599

I tre tool hanno individuato eventi di alternative donor site e intron retention. Per comporre l'evento più complesso occorre la composizione di un evento di alternative donor site e di un exon skipping.

ENST00000441139 vs ENST00000513194

ASGAL ha evidenziato un evento di alternative donor site, SplAdder un evento di intron retention e rMATS non ha identificato nessun evento. In questo caso l'evento complesso avrebbe potuto essere costituito dalla composizione di più eventi di alternative donor site e alternative acceptor site.

Per gli output specifici di ogni singolo evento complesso, per ogni tool analizzato, si rimanda ai file contenuti nell'archivio GitHub [24].

Capitolo 5

Conclusioni

Lo scopo principale di questa tesi, e quindi dell'esperienza svolta, era quello di mettere a confronto tre diversi software che si occupano di identificare eventi di splicing alternativo: ASGAL, SplAdder e rMATS. È importante sottolineare che ad eccezione di ASGAL, che implementa un tool interno per l'allineamento di read RNA-Seq, durante l'esecuzione delle diverse pipeline si è sempre utilizzato il software STAR. Questo significa che utilizzare un software di allineamento differente potrebbe portare a risultati differenti. Ad ogni modo si è constatato come i tre tool sotto analisi supportino, per quasi la totalità dei casi, gli eventi di splicing semplici che dichiarano di supportare. Per quanto riguarda gli eventi più complessi si è potuto osservare che i software riescono spesso a identificare alcuni eventi più semplici, ma non sono in grado di evidenziare la combinazione degli stessi fino a ricostruire l'evento complesso nella sua interezza. Questo vale esclusivamente per i casi specifici che sono stati provati con il cromosoma Y, ma dà l'idea di una situazione che può essere estesa ad altri geni e cromosomi. Da questo punto di vista sarebbe quindi interessante poter eseguire un'analisi più ampia che tenga in considerazione molti più eventi e geni, compresi tool differenti di allineamento. Data l'importanza e il ruolo chiave dello splicing alternativo per la salute delle persone, si crede che i tre software considerati avranno sviluppi futuri. Questo a dire che gli eventi complessi che in questo studio non sono stati identificati correttamente potranno riscontrare esiti migliori in versioni successive.

Bibliografia

- [1] Nature. *Definizione di splicing alternatvio*. URL: <https://www.nature.com/subjects/alternative-splicing> (visitato il 08/07/2018).
- [2] Wang Yan et al. «Mechanism of alternative splicing and its regulation». In: *Biomed Rep.* (17 Dic. 2014), pp. 152–158. DOI: [10.3892/br.2014.407](https://doi.org/10.3892/br.2014.407). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4360811/> (visitato il 08/07/2018).
- [3] Climente-González Héctor et al. «The Functional Impact of Alternative Splicing in Cancer». In: *Cell Reports* 20 (9 29 ago. 2017), pp. 2215–2226. DOI: [10.1016/j.celrep.2017.08.012](https://doi.org/10.1016/j.celrep.2017.08.012). (Visitato il 08/07/2018).
- [4] Donaldson Lucy F. e Beazley-Long Nicholas. «Alternative RNA splicing: contribution to pain and potential therapeutic strategy». In: *Drug Discovery Today* 21 (11 18 giu. 2016), pp. 1787–1798. DOI: [10.1016/j.drudis.2016.06.017](https://doi.org/10.1016/j.drudis.2016.06.017). URL: <https://www.sciencedirect.com/science/article/pii/S1359644616302355?via%3Dihub> (visitato il 08/07/2018).
- [5] Martínez-Montiel Nancy, Rosas-Murrieta Nora e Martínez-Contreras Rebeca. «Alternative splicing regulation: Implications in cancer diagnosis and treatment». In: *Medicina Clínica* 144 (7 8 apr. 2015), pp. 317–323. DOI: [10.1016/j.medcle.2015.11.005](https://doi.org/10.1016/j.medcle.2015.11.005). URL: <https://www.sciencedirect.com/science/article/pii/S2387020615001709?via%3Dihub> (visitato il 08/07/2018).
- [6] Nature. *The Information in DNA Is Decoded by Transcription*. URL: <https://www.nature.com/scitable/topicpage/the-information-in-dna-is-decoded-by-6524808> (visitato il 08/07/2018).

- [7] Nature. *The Information in DNA Determines Cellular Function via Translation*. URL: <https://www.nature.com/scitable/topicpage/the-information-in-dna-determines-cellular-function-6523228> (visitato il 08/07/2018).
- [8] Chen Liang. *Statistical and Computational Studies on Alternative Splicing*. In: *Handbook of Statistical Bioinformatics*. A cura di Springer. 2011, p. 630. ISBN: 978-3-642-16344-9. URL: <https://goo.gl/arv9zL> (visitato il 08/07/2018).
- [9] Genome BioInformatics Research Lab. *The AS code and the underlying method*. URL: <http://genome.crg.es/astalavista/FAQ.html> (visitato il 08/07/2018).
- [10] *Ensembl datasets*. URL: <https://www.ensembl.org/info/data/ftp/index.html> (visitato il 08/07/2018).
- [11] Wikipedia. *FASTA format*. URL: https://en.wikipedia.org/wiki/FASTA_format (visitato il 08/07/2018).
- [12] Foissac Sylvain e Sammeth Michael. «ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets». In: *Nucleic Acids Research* 35 (1 lug. 2007), pp. 297–299. DOI: 10.1093/nar/gkm311. URL: https://academic.oup.com/nar/article/35/suppl_2/W297/2922496 (visitato il 08/07/2018).
- [13] Li Wei. *RNASeqReadSimulator*. 1 Feb. 2013. URL: <https://github.com/davidliwei/RNASeqReadSimulator/blob/master/README> (visitato il 08/07/2018).
- [14] Dobin Alexander. *STAR manual*. 24 Apr. 2018. URL: <https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf> (visitato il 08/07/2018).
- [15] Dobin Alexander et al. «STAR: ultrafast universal RNA-seq aligner». In: *Bioinformatics* 29 (1 25 ott. 2012), pp. 15–21. DOI: 10.1093/bioinformatics/bts635. URL: <https://academic.oup.com/bioinformatics/article/29/1/15/272537> (visitato il 08/07/2018).
- [16] Denti Luca et al. *ASGAL: Aligning RNA-Seq Data to a Splicing Graph to Detect Novel Alternative Splicing Events*. URL: <http://asgal.algolab.eu/> (visitato il 08/07/2018).

- [17] Denti Luca et al. *ASGAL documentation*. URL: <http://asgal.algolab.eu/documentation> (visitato il 08/07/2018).
- [18] Ong Cheng Soon, Raetsch Gunnar e Kahles Andre. *SplAdder*. URL: <https://github.com/ratschlab/spladder/blob/master/README.md> (visitato il 08/07/2018).
- [19] Ong Cheng Soon, Raetsch Gunnar e Kahles Andre. *File Format Descriptions*. 2 Set. 2015. URL: <https://github.com/ratschlab/spladder/wiki/File-Format-Descriptions> (visitato il 08/07/2018).
- [20] Xing Lab. *Multivariate Analysis of Transcript Splicing (MATS)*. URL: <http://rnaseq-mats.sourceforge.net/> (visitato il 08/07/2018).
- [21] Shen Shihao et al. «MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data». In: *Nucleic Acids Research* 40 (8 20 gen. 2012), p. 61. DOI: [10.1093/nar/gkr1291](https://doi.org/10.1093/nar/gkr1291). URL: <https://academic.oup.com/nar/article/40/8/e61/2411737> (visitato il 08/07/2018).
- [22] Xing Lab. *rMATS v4.0.2 (turbo)*. URL: http://rnaseq-mats.sourceforge.net/user_guide.htm (visitato il 08/07/2018).
- [23] Li Heng et al. *Samtools repository*. URL: <https://github.com/samtools/samtools> (visitato il 10/07/2018).
- [24] Previtali Mattia. *GitHub repository*. 8 Lug. 2018. URL: <https://github.com/MattiaPrevitali/comparisonASSW/> (visitato il 08/07/2018).
- [25] Broad Institute e University of California. *Integrative Genomics Viewer*. URL: <https://software.broadinstitute.org/software/igv/> (visitato il 08/07/2018).