Metodi e analisi statistiche 2024 ISBN: 978-88-6629-053-7
DIPARTIMENTO DI ECONOMIA E FINANZA p. 143-154

L'impatto delle mutazioni del DNA sul processo di *splicing*: un'analisi statistica nel contesto del gene BRCA1

Barbara Nocca[1], Nunziata Ribecco[1], Alessandro Stella[2]

[1] Dipartimento di Economia e Finanza, Università degli Studi di Bari Aldo Moro,
[2] Dipartimento di Scienze biomediche e oncologia umana, Università degli Studi di Bari Aldo Moro

Riassunto: In questo lavoro ci si propone di analizzare l'effetto delle mutazioni del DNA sul processo di splicing, con particolare attenzione al gene BRCA1 sito sul cromosoma 17. Utilizzando un campione di 105 soggetti suddivisi tra i soggetti portatori di mutazioni note per influenzare lo splicing (positivi) e i soggetti portatori di mutazioni che non lo influenzano (negativi), a seguito di una prima analisi esplorativa dei dati, è stata condotta l'analisi delle corrispondenze multiple. Essa ha rivelato che specifiche mutazioni negli esoni 16, 19, 20 e 21 del gene BRCA1, localizzato sul cromosoma 17, sono significativamente associate a variazioni nel processo di splicing. Al contrario, mutazioni negli esoni 3, 4, 17 e 18 non hanno mostrato un impatto significativo sullo splicing. Questi risultati forniscono nuove informazioni sulle mutazioni genetiche del gene BRCA1 con importanti conseguenze sul processo di splicing, contribuendo alla comprensione della patogenesi delle malattie che ne derivano e suggerendo potenziali direzioni per future ricerche.

Keywords: Mutazione, Processo di Splicing, Corrispondenze multiple.

1. Introduzione

La trascrizione è il processo attraverso il quale una cellula produce una copia complementare di una sequenza di DNA in forma di RNA messaggero (mRNA). Tale trascritto è quindi utilizzato per la sintesi di proteine, non prima di aver subito un processo di maturazione detto "splicing" durante il quale dal pre-mRNA sono rimosse dal trascritto primario le sequenze non codificanti dette *introni*. Le rimanenti porzioni codificanti, dette *esoni*, saranno unite tra loro per ottenere l'mRNA maturo

che sarà tradotto in proteine.

Ogni individuo è portatore di centinaia di variazioni nucleotidiche rispetto alla sequenza di riferimento (normale) del DNA che sono responsabili della naturale diversità tra esseri umani. Tuttavia, alcune di queste variazioni possono essere causa di malattia. Il processo di splicing è un processo fondamentale perché l'inserimento erroneo di sequenze *introniche*, o l'omissione di sequenze *esoniche*, porterebbe alla traduzione di proteine anomali con perdita della normale funzione biologica. Tuttavia, non è facile prevedere quali sostituzioni nucleotidiche possano avere delle conseguenze sull'esecuzione corretta del processo di splicing.

2. Materiali e metodi

Il campione analizzato era costituito da 105 individui a sua volta distinto in due sottocampioni:

- il primo sotto-campione, composto da 50 individui, è relativo alla popolazione dei **"positivi"** ovvero coloro in cui la mutazione ha effetto sullo splicing;
- il secondo sotto-campione, composto da 55 individui, è relativo alla popolazione dei **"negativi"** ovvero coloro in cui la mutazione non ha effetto sullo splicing;

Il dataset include le seguenti variabili rilevanti per l'analisi degli errori nella trascrizione associate a un effetto sullo splicing:

- Transcript ID: rappresenta l'Id del transcript in banca dati (www.ensembl.org);
- Exon: rappresenta l'esone interessato dalla mutazione;
- Chrom: è il cromosoma dove è localizzato il gene in cui è presente la mutazione;
- HG19 pos: sono le coordinate nel genoma umano (versione 19) di ogni singola mutazione;
- Ref: è il nucleotide di riferimento; ovvero il nucleotide presente nella sequenza normale;
- Alt: è il nucleotide alternativo, ovvero il nucleotide presente a seguito della mutazione;
- Splicing Site: è la tipologia di sito di splicing: SD è localizzato all'inizio dell'introne dove avviene il primo taglio ed è definito sito donatore, SA è alla fine dell'introne dove avviene il secondo taglio ed è definito sito accettatore;

- Distance from SJ: rappresenta la distanza dalla più vicina giunzione di splicing espressa in numero di nucleotidi;
- Gene: è il gene su cui è presente la mutazione.

Si è passati dapprima a selezionare le variabili più rilevanti del dataset. In entrambi i due sotto-campioni analizzati, il gene che si presenta più frequentemente mutato è il gene BRCA1 questo perché in generale, nel campione di mutazioni analizzate, quelle che possono essere responsabili di tumore della mammella ereditario sono tra le più studiate e il gene BRCA1 è quello più comunemente coinvolto.

Pertanto, l'analisi si è focalizzata sulle mutazioni che riguardavano il gene BRCA1 sito sul Cromosoma 17.

I dati sono stati studiati mediante l'analisi esplorativa (EDA, Exploratory Data Analysis), una fase importante del processo di analisi dei dati, in quanto mira a comprendere la struttura e le caratteristiche principali dei dati. Questo approccio è fondamentale per identificare modelli, anomalie, relazioni e per formulare ipotesi iniziali che guideranno ulteriori analisi. Sulla base dei risultati delle analisi esplorative sono state effettuate delle ipotesi, poi confermate dall'analisi delle corrispondenze multiple.

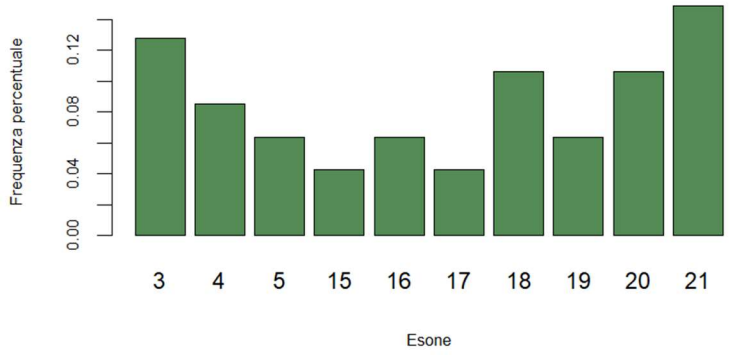
L'analisi delle corrispondenze multiple (ACM), o Multiple Correspondence Analysis (MCA), è una tecnica statistica utilizzata per esplorare e visualizzare le relazioni tra più variabili categoriali. Questa tecnica mira a identificare la struttura dell'associazione all'interno di un dataset, rappresentando graficamente le modalità dei caratteri in uno spazio di dimensioni minime, generalmente rappresentato su un piano cartesiano. Sulla base della vicinanza tra le varie mutabili nel piano è possibile ricavare un'associazione tra le variabili permettendo di dedurre quali relazioni caratterizzano il dataset.

3. Risultati

3.1 Analisi della distribuzione delle mutazioni negli esoni

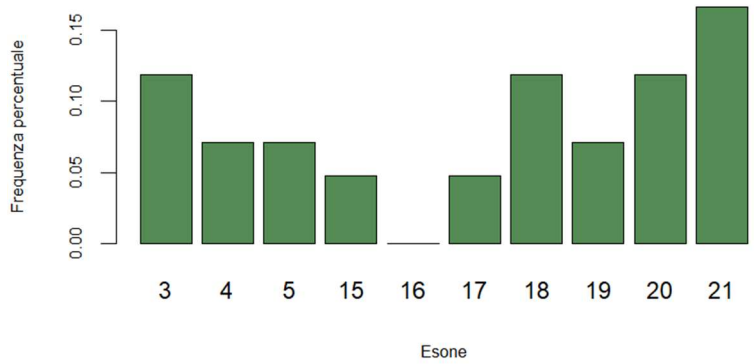
Analizzando la distribuzione delle mutazioni con putativo effetto sullo splicing nei diversi esoni del gene BRCA1 si ottengono i seguenti risultati:

Figura 1. Distribuzione degli esoni interessati nella mutazione riguardante il cromosoma 17 nella popolazione dei positivi



La distribuzione degli esoni interessati da mutazioni nel gene BRCA1 nella popolazione dei positivi è rappresentata nella Figura 1. Si osserva che l'esone 21 presenta la frequenza percentuale maggiore. Inoltre, osservando il grafico, è possibile affermare che gli esoni maggiormente colpiti da mutazioni sono l'esone 21, l'esone 3, gli esoni 18 e 20 e infine l'esone 4.

Figura 2. *Distribuzione degli esoni interessati nella mutazione riguardante il cromosoma 17 nella popolazione dei negativi*



La distribuzione degli esoni nella popolazione dei negativi è rappresentata nella Figura 2. La distribuzione presenta anche qui un punto di massimo in corrispondenza dell'esone 21, conseguentemente la distribuzione si presenta asimmetrica. È possibile, quindi, affermare che nella popolazione dei negativi gli esoni con maggiore frequenza di alterazioni prive di effetto sullo splicing sono l'esone 21, l'esone 3 e gli esoni 18 e 20.

Confrontando ora i due grafici, si denota come tra la popolazione dei Positivi e la popolazione dei Negativi la differenza principale risiede nella frequenza di alterazioni patogenetiche nell'esone 16. Esso, infatti, è presente nella popolazione dei Positivi, ma risulta essere totalmente assente nella popolazione dei Negativi.

Si denota poi, che la frequenza relativa degli esoni 3 e 4 è maggiore nella popolazione dei positivi, mentre la frequenza relativa degli esoni 5, 15, 17, 18, 19, 20 e 21 è maggiore nei negativi.

In conclusione, si osserva che le due distribuzioni sono molto simili. BRCA1 è, composto da 27 esoni, pertanto nei due grafici gli esoni più frequentemente presenti corrispondono a esoni con una regolazione dello Splicing più complicata e più sensibile ai cambiamenti nel DNA. Tuttavia, la maggiore frequenza di alterazioni con effetto sullo splicing nell'esone 16 di BRCA1 fa sospettare che in questo esone le sostituzioni nucleotidiche abbiano una più alta probabilità di provocare anomalie nel processo di splicing.

Nella Tabella 1 è riportata la distribuzione degli esoni nella popolazione dei positivi e in quella dei negativi allo scopo di analizzare se esiste una relazione di dipendenza fra le due popolazioni. Essendoci caselle vuote, si è reso necessario considerare come unico gruppo gli esoni che nella popolazione dei Positivi, presentavano una maggiore frequenza rispetto alla popolazione dei Negativi.

Tabella 1. *Distribuzione delle popolazioni (positivi/negativi) per presenza di esoni*

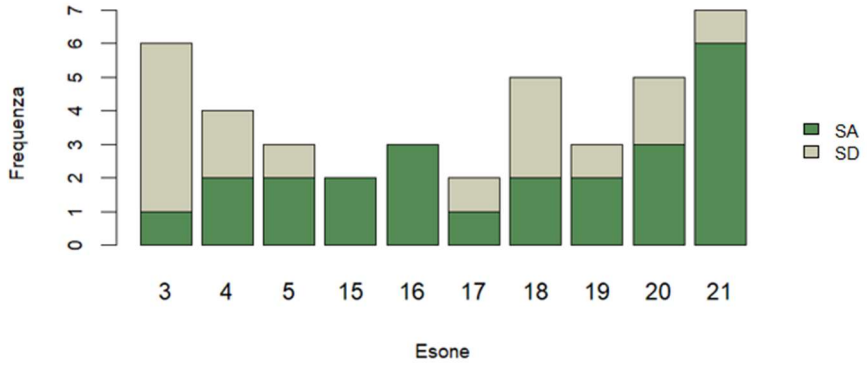
	Esone	
Popolazione	Esone 3-4-16	Altri
Positivi	19	21
Negativi	0	39
Totale	20	59

Avendo verificato che nessuna frequenza teorica è più piccola di cinque, al fine di verificare che non ci fosse alcuna relazione fra le due variabili osservate, è stato utilizzato il test Chi-quadro. I risultati del test portano a rifiutare l'ipotesi di indipendenza fra le distribuzioni in quanto il p-value (2.933e-06) è inferiore al livello di significatività α posto pari a 0,05.

3.2 Analisi del sito di splicing nei diversi esoni

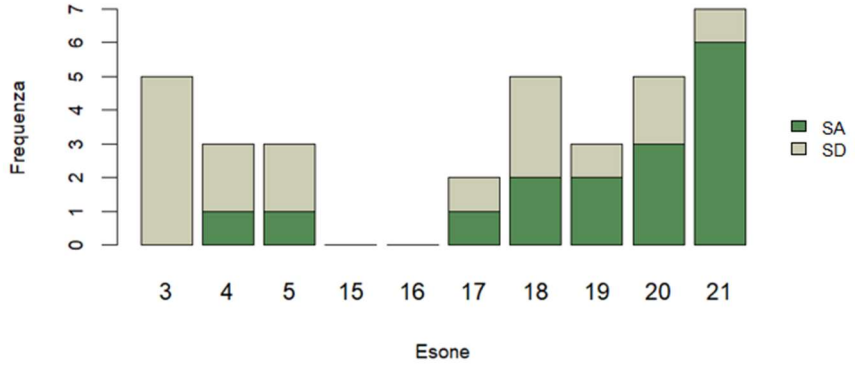
Analizzando la tipologia di sito di Splicing (SD vs SA) interessato dalle mutazioni nelle due sottopopolazioni nei diversi esoni si osserva che:

Figura 3. *Distribuzione degli esoni interessati nella mutazione riguardante il cromosoma 17 nella popolazione dei positivi e il sito di splicing da essi posseduto*



La distribuzione del tipo di sito di splicing oggetto di mutazione nella popolazione dei positivi e nei diversi esoni del gene BRCA1 è rappresentata nella Figura 3, da cui si evince che l'esone 3 è caratterizzato prevalentemente da sostituzioni nel sito di splicing di tipo SD, mentre l'esone 16, assente nella popolazione dei negativi, è caratterizzato esclusivamente da sostituzioni nel sito di splicing di tipo SA.

Figura 4. *Distribuzione degli esoni interessati nella mutazione riguardante il cromosoma 17 nella popolazione dei negativi e il sito di splicing da essi posseduto*



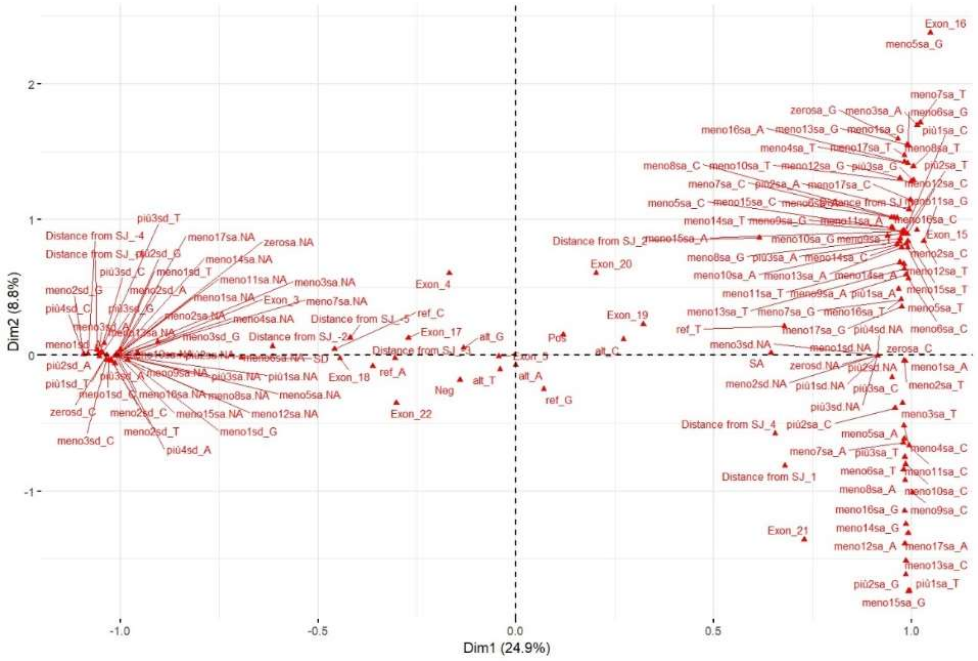
La distribuzione degli esoni in relazione al sito di splicing nella popolazione dei negativi è rappresentata nella Figura 4. Dal grafico si osserva che l'esone 3 è caratterizzato esclusivamente da sostituzioni nel sito di splicing di tipo SD, non si possiedono osservazioni sui siti di splicing dell'esone 15, mentre nella popolazione dei positivi due sostituzioni interessano un sito di splicing di tipo SA.

Inoltre, per l'esone 3 si nota come il sito di splicing di tipo SA è presente solo nella popolazione dei positivi. Analogamente per l'esone 4 si nota che il sito di splicing di tipo SA è maggiormente presente nei positivi.

3.3 Analisi delle corrispondenze multiple

Al fine di studiare l'associazione fra le variabili è stata effettuata l'analisi delle corrispondenze, considerando la totalità delle variabili e successivamente solo le 4 variabili principali. Considerando il dataset nella sua totalità è possibile osservare il seguente plot delle corrispondenze:

Figura 5. *Primo plot delle corrispondenze sull'intero dataset.*



Dal grafico è possibile notare che sia l'asse delle ordinate sia quello delle ascisse risultano essere molto discriminanti; in particolare, la popolazione dei positivi è situata nel 1° quadrante, mentre la popolazione dei negativi è situata nel 3° quadrante. Si nota che nel semiasse negativo delle ascisse le variabili sono più raggruppate, segno di una maggiore associazione.

Il sito di Splicing donatore (SD) giace in corrispondenza del semiasse negativo delle ascisse ed è quindi verosimilmente associato alla popolazione dei negativi e risulta vicino agli esoni 4, 17, 18 e 22. Nel 2° e nel 3° quadrante, inoltre, associati alla popolazione dei negativi si ritrova l'esone 3.

In corrispondenza della popolazione dei negativi si ritrovano inoltre le modalità della variabile "Distance from Sj" da -1 a -5, anche se la modalità -3 e la modalità -5 sono più prossime alla popolazione negativa.

Per quanto concerne le altre variabili è possibile notare che sono associati alla popolazione dei negativi i nucleotidi A, T e G in posizione -1 nel sito di splicing SD, tutti e 4 i nucleotidi in posizione -2 nel sito di splicing SD, i nucleotidi A e G in posizione -3 nel sito di splicing SD, il nucleotide C in posizione 0 nel sito di splicing SD, il nucleotide T in posizione +1 nel sito di splicing SD, i nucleotidi A e G in posizione +2 nel sito di splicing SD, tutti e 4 i nucleotidi in posizione +3 nel sito di splicing SD, i nucleotidi A e C in posizione più 4 nel sito di splicing SD.

I suddetti risultati sono riassunti nella seguente tabella:

Tabella 2. *Nucleotidi presenti nelle varie posizioni con sito di splicing accettore associato alla popolazione dei positivi*

-3	-2	-1	0	1	2	3	4
A, G	A, T, C, G	A, T, G	C	T	A, G	A, T, C, G	A, C

Il sito di Splicing SA giace, invece, in corrispondenza del semiasse positivo delle *x* ed è quindi verosimilmente associato alla popolazione dei positivi. In corrispondenza del semiasse positivo delle *x* si ritrovano anche gli esoni 5, 19, 20. Inoltre, l'esone 5 si trova in posizione intermedia tra le due popolazioni. Infine, nel 1° e nel 4° quadrante, associati alla popolazione dei negativi si notano gli esoni 16 e 21.

In corrispondenza della popolazione dei positivi si osservano le modalità della variabile "Distance from Sj" da 1 a 4, anche se la modalità 2, la modalità 3 e la modalità 4 sono più prossime, in termini di distanza, alla popolazione dei positivi e quindi maggiormente associate.

Per quanto concerne le altre variabili si nota che sono associate alla popolazione dei positivi tutti e 4 i nucleotidi in posizione -17 nel sito di splicing SA, i nucleotidi C, T e G in posizione -16 nel sito di splicing SA, tutti e 4 i nucleotidi in posizione -15 nel sito di splicing SA, i nucleotidi C, T e G in posizione -14 nel sito di splicing

SA, i nucleotidi C, A e G in posizione -13 nel sito di splicing SA, i nucleotidi C e A in posizione -12 nel sito di splicing SA, i nucleotidi C, A e G in posizione -11 nel sito di splicing SA, i nucleotidi C, A e G in posizione -10 nel sito di splicing SA, i nucleotidi C, T e G in posizione -9 nel sito di splicing SA, i nucleotidi C, A e G in posizione -8 nel sito di splicing SA, tutti e 4 i nucleotidi in posizione -7 nel sito di splicing SA, tutti e 4 i nucleotidi in posizione -6 nel sito di splicing SA, tutti e 4 i nucleotidi in posizione -5 nel sito di splicing SA, il nucleotide C in posizione -4 nel sito di splicing SA, i nucleotidi C, A e T in posizione -3 nel sito di splicing SA, i nucleotidi C e T in posizione -2 nel sito di splicing SA, i nucleotidi A e G in posizione -1 nel sito di splicing SA, i nucleotidi C e G in posizione zero nel sito di splicing SA, i nucleotidi A e T in posizione +1 nel sito di splicing SA, tutti e 4 i nucleotidi in posizione +2 nel sito di splicing SA, i nucleotidi T, A e G in posizione +3 nel sito di splicing SA.

I suddetti risultati sono riassunti nella Tabella 3 seguente:

Tabella 3. *Nucleotidi presenti nelle varie posizioni nel sito di splicing donatore associato alla popolazione dei negativi.*

-17	-16	-15	-14	-13	-12	-11
A, T, C, G	T, C, G	A, T, C, G	T, C, G	A, C, G	A, C	A, C, G
-10	-9	-8	-7	-6	-5	-4
A, C, G	T, C, G	A, C, G	A, T, C, G	A, T, C, G	A, T, C, G	C
-3	-2	-1	0	1	2	3
A, T, C	T, C	A, G	C, G	A, T	A, T, C, G	A, T, C

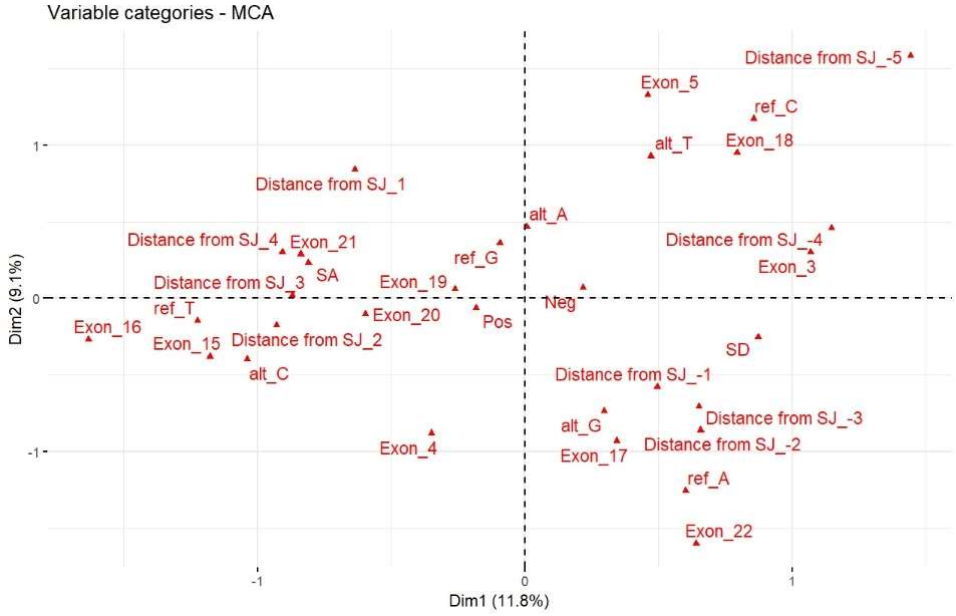
Si considerano ora solo le variabili di seguito elencate:

- Exon: esone interessato nella mutazione
- Distance from SJ: distanza dal sito di splicing espressa in nume-ro di nucleotidi

- Ref: nucleotide presente nella mutazione
- Alt: nucleotide che si trova normalmente

Il grafico dell'analisi effettuata risulta:

Figura 6. *Secondo plot delle corrispondenze.*



Considerando la Figura 6 è possibile notare come l'asse delle *y* risulta essere discriminante per la variabile popolazione; in particolare, la popolazione dei negativi è situata a destra tra il 1° e il 4° quadrante, mentre la popolazione dei positivi è situata a sinistra nel 3° quadrante, entrambi a ridosso dell'asse delle ascisse. Si osserva che nel semiasse negativo delle ascisse le variabili sono più concentrate, segno di una maggiore associazione.

Il sito di Splicing SA giace in corrispondenza del semiasse negativo delle ascisse ed è quindi verosimilmente associato alla popolazione dei positivi e risulta vicino agli esoni 19, 20 e 21; i 3 esoni, infatti, possedevano con frequenza maggiore mutazioni in un sito di splicing di tipo SA. Nel 3° quadrante, inoltre, associati alla popolazione dei positivi si ritrovano gli esoni 4, 15 e 16 che in questa popolazione erano quelli più frequentemente mutati.

Il sito di Splicing SD giace, invece, in corrispondenza del semiasse positivo delle ascisse ed è quindi verosimilmente associato alla popolazione dei negativi. Sempre

con riferimento all'asse delle ascisse, in corrispondenza del semiasse positivo si notano gli esoni 3, 5, 17 e 18.

In particolare, si osserva la formazione di due gruppi nel semiasse positivo delle ascisse: nel primo quadrante si posizionano gli esoni 3, 5 e 18 associati a una distanza dal sito di splicing donatore pari a -4 o -5 dove presentano tipicamente un nucleotide C, sostituito da un nucleotide T; nel quarto quadrante si posizionano gli esoni 17 e 22 associati a una distanza dal sito di splicing (donatore) pari a -1, -2, -3 dove presentano tipicamente un nucleotide A, sostituito da un nucleotide G.

4. Conclusioni

In conclusione, lo studio condotto ha portato a significativi risultati.

Dall'analisi esplorativa, si era dedotto che gli esoni maggiormente soggetti a errori nello splicing determinanti nello sviluppo del carcinoma mammario ereditario legato a mutazioni del gene BRCA1 erano rappresentati dagli esoni 3, 4, 16.

Dall'analisi delle corrispondenze si osserva che una mutazione che colpisce gli esoni 16, 19, 20, 21 a una distanza dal sito di splicing accettore in termini di nucleotidi pari a 1, 2, 3, 4 che presentavano come nucleotide di riferimento il nucleotide G, sostituito in conseguenza della mutazione da un nucleotide A o C, risulta essere associato a una probabile influenza sullo splicing.

Si denota inoltre, che in corrispondenza delle posizioni più prossime al sito di splicing il genoma risulta essere meno flessibile in quanto in corrispondenza della posizione 0 si posiziona esclusivamente il nucleotide C e in posizione 1 esclusivamente il nucleotide T. La presenza di nucleotidi differenti in corrispondenza delle suddette posizioni quasi certamente porta ad un'alterazione del processo di splicing.

Al contrario una mutazione negli esoni 3, 4, 17 18 e 22 a una distanza dal sito di splicing donatore in termini di nucleotidi pari a -1, -2, -3, -4 e -5 che presentavano come nucleotide di riferimento il nucleotide A o C, sostituiti in conseguenza della mutazione da un nucleotide G o T risulta essere probabilmente NON associato a una possibile influenza sullo splicing.

In merito all'esone 5 che presentava in posizione -4 con sito di Splicing di tipo SD un nucleotide di riferimento G, trascritto erroneamente come A o C e in posizione 2 con sito di Splicing SA un nucleotide di riferimento C trascritto erroneamente come A non è possibile ritenere che non sia associato a una probabile influenza sullo splicing in quanto risulta essere vicino a entrambe le due popolazioni.

Si denota inoltre che in corrispondenza delle posizioni più prossime al sito di splicing il genoma risulta essere meno flessibile in quanto in corrispondenza della posizione -4 ritroviamo il nucleotide C, in posizione -2 ritroviamo esclusivamente il nucleotide T o C, in posizione -1 il nucleotide A o G, in posizione 0 il nucleotide C o G e in posizione 1 il nucleotide A o T.

5. Bibliografia

Atkins, P. W., Jones, L., & Laverman, L. (2005). *Principi di chimica*. Zanichelli.

Becker, W. M., Kleinsmith, L. J., Hardin, J., & Raasch, J. (2002). *Il mondo della cellula*. EdiSES.

Chan, B. K. (2015). Biostatistics for epidemiology and public health using R. Springer Publishing Company.

Elston, RC, Olson, JM e Palmer, L. (a cura di). (2002). *Genetica biostatistica ed epidemiologia genetica (Vol. 1).* John Wiley & Figli.

Pagano, M., Gauvreau, K., & Mattie, H. (2022). *Principi di biostatistica*. Chapman e Hall/CRC.