UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN BIOINGEGNERIA

"VIMENTINA E G-QUADRUPLEX NELLE REGIONI ACCESSIBILI DELLA CROMATINA: UN'ANALISI GENOMICA PER NUOVI TARGET TERAPEUTICI NELLA PROGRESSIONE TUMORALE"

Relatore: Prof. Enrico Lavezzo

Laureanda: Valentina Ceriani

Correlatore: Prof. Stefano Toppo

ANNO ACCADEMICO 2023 – 2024

Data di laurea 27/11/2024

INDICE

ABSTRACT

La modulazione dell'accessibilità cromatinica e dell'espressione genica è cruciale nella regolazione delle vie metaboliche e di segnalazione delle cellule. L'alterazione di questi delicati meccanismi può portare a diverse patologie, fino alla progressione tumorale. Lo studio di questi processi è, quindi, una via promettente per lo sviluppo di nuove terapie mirate.

Questa tesi indaga la relazione tra le regioni accessibili della cromatina (ATAC) e la localizzazione di vimentina, una proteina strutturale presente anche nel nucleo della cellula.

Lo studio è stato condotto su cellule HGC-27, una linea cellulare derivata da un carcinoma gastrico umano indifferenziato, ampiamente utilizzata nella ricerca oncologica. Le HGC-27 offrono un modello sperimentale ideale per analizzare le dinamiche molecolari in contesti tumorali grazie alla loro elevata capacità proliferativa.

L'obiettivo della ricerca è identificare motivi ricorrenti nelle regioni ATAC che influenzano il legame di vimentina, specialmente in presenza di particolari strutture non canoniche del DNA chiamate G-quadruplex.

Studi recenti hanno dimostrato che vimentina lega preferenzialmente regioni del genoma caratterizzate da due G-quadruplex molto vicini, che interagiscono tra loro. Questa interazione sembra essere funzionale alla capacità di vimentina di promuovere la regolazione genica e forse la migrazione delle cellule tumorali. L'estensione di tale interazione verrà saggiata a livello genomico, al fine di identificare potenziali strutture vimentina-G-quadruplex coinvolte nella regolazione dell'espressione di geni potenzialmente implicati nella progressione tumorale.

L'analisi si avvale di tecniche avanzate di sequenziamento ATAC-seq e strumenti bioinformatici per il trattamento dei dati, con un workflow rigoroso che include FastQC, Trim Galore, Bowtie2 e altre piattaforme per garantire l'accuratezza del preprocessing e dell'analisi.

Questo studio mira ad aprire nuove prospettive nella comprensione della regolazione epigenetica della vimentina e a contribuire allo sviluppo di strategie terapeutiche innovative basate su interventi epigenetici, con potenziali implicazioni nel trattamento dei tumori.

CAPITOLO 1: INTRODUZIONE

1.1 IL DNA

Il DNA, o acido desossiribonucleico, costituisce il genoma e nell'uomo conta circa 3 miliardi di paia di basi contenute in 46 cromosomi all'interno del nucleo della cellula. Esso è identico in tutte le cellule dell'organismo e contiene le informazioni necessarie per la loro sopravvivenza e il loro funzionamento. Inoltre, è alla base dell'ereditarietà perché responsabile della trasmissione delle caratteristiche che rendono un individuo simile ai propri genitori.

1.1.1 Struttura

Dal punto di vista chimico, il DNA è una macromolecola costituita dalla concatenazione di unità ripetute, i nucleotidi. Ciascuno di essi è formato da uno zucchero a cinque atomi di carbonio, il desossiribosio, un gruppo fosfato e una base azotata, uniti da legami specifici che garantiscono la stabilità e la continuità della struttura del DNA.
Il desossiribosio si lega al gruppo fosfato tramite un legame fosfodiesterico, che coinvolge uno degli ossigeni del gruppo fosfato e il carbonio in posizione 5' dello zucchero.
A sua volta, la base azotata si lega allo zucchero tramite un legame glicosidico, che si stabilisce tra il carbonio in posizione 1' del desossiribosio e un atomo di azoto presente nella base.

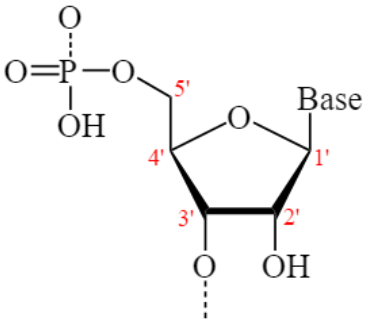

Figura 1: Struttura di un nucleotide composto da un gruppo fosfato, uno zucchero a cinque atomi di carbonio (desossiribosio) e una base azotata. Il gruppo fosfato si lega al carbonio 5' del desossiribosio, mentre la base azotata si unisce al carbonio 1' dello zucchero.

Nella catena del DNA, i nucleotidi successivi si uniscono attraverso legami fosfodiesterici, che collegano il gruppo fosfato di un nucleotide al carbonio 3' del desossiribosio del nucleotide seguente. Questa serie di legami crea la spina dorsale zucchero-fosfato che mantiene uniti i nucleotidi. Due filamenti di DNA vengono tenuti insieme dall'interazione tra basi azotate complementari. Le basi azotate si dividono in due categorie: le purine, caratterizzate da una struttura a doppio anello, e le pirimidine, con una struttura a singolo anello. Le purine, adenina e guanina, e le pirimidine, citosina e timina, si legano tra loro tramite legami a idrogeno, stabilizzando la doppia elica del DNA: citosina e guanina sono legate da tre legami a idrogeno, mentre adenina e timina da due.

Quindi, il DNA si presenta come una coppia di filamenti antiparalleli aventi un'orientazione 5'-3' che si associano a formare una doppia elica.

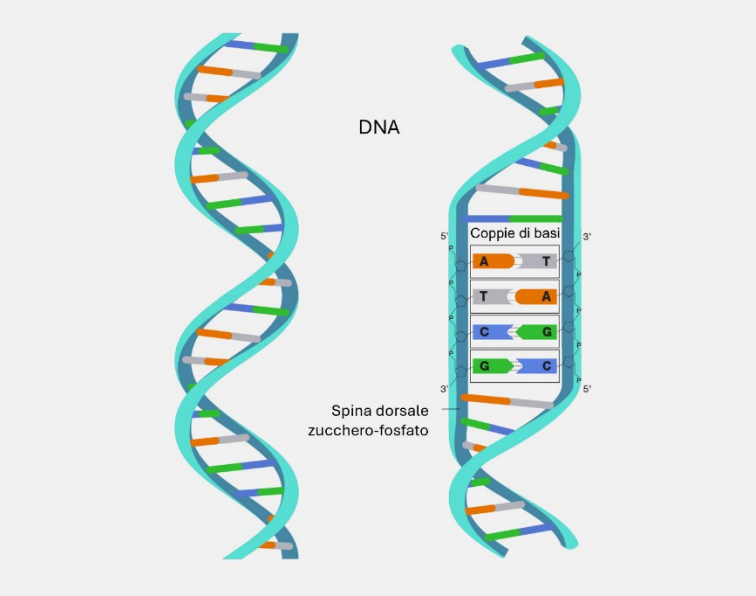

Figura 2: Struttura della doppia elica del DNA. I due filamenti antiparalleli e complementari sono legati tra loro tramite basi azotate appaiate (adenina (A), guanina (G), citosina (C), timina (T)) e sono tenuti insieme dallo scheletro zucchero-fosfato.

La struttura del DNA fu annunciata negli anni ’50, in un articolo pubblicato dai due scienziati James Watson e Francis Crick. La scoperta rivoluzionò il campo della medicina, in quanto fece emergere le due caratteristiche principali che svolgono un ruolo fondamentale nella funzione biologica del DNA: la complementarietà delle sequenze di basi sui due filamenti e la natura a doppia elica del polimero (1). La prima suggerisce che l’informazione possa essere replicata in due copie identiche, mentre la seconda conferisce proprietà fisico-chimiche cruciali nel definire la funzione biologica del polimero (1).

Questo fu l’inizio degli immensi progressi nella comprensione e nella manipolazione genetica avvenuti negli ultimi 70 anni (1).

1.1.2 Funzioni

Il DNA ha tre funzioni principali date dalla sinergia tra l'informazione contenuta nella sequenza di basi e dinamicità della struttura a doppia elica (1).

La prima funzione è quella di conservare l'informazione genetica contenuta nella sequenza delle basi e di trasmetterla di generazione in generazione attraverso la replicazione. Questo processo, cruciale per la divisione cellulare, crea copie identiche del DNA che vengono trasferite alla cellula figlia. La replicazione sfrutta la complementarità delle basi, che consente di duplicare le informazioni contenute nella doppia elica a partire da entrambi i filamenti. L'enzima elicasi apre la doppia elica separando i filamenti attraverso il taglio dei legami idrogeno tra le basi azotate. Successivamente, l'enzima DNA polimerasi sintetizza un filamento complementare su ciascuno dei filamenti originari, garantendo così la conservazione di tutto il patrimonio genetico.

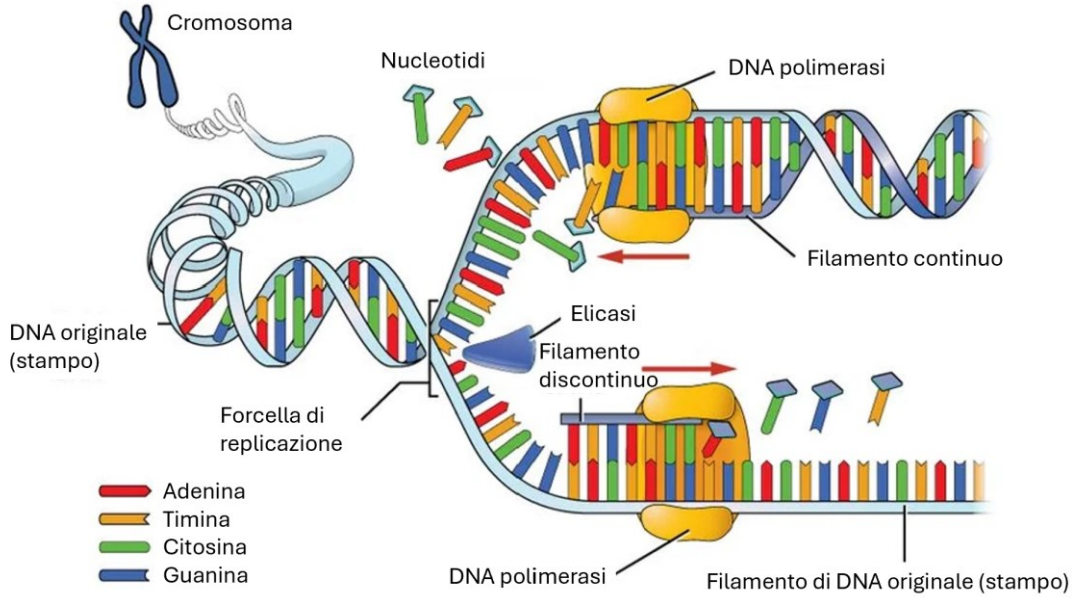

Figura 3: Processo di replicazione del DNA. L'enzima elicasi separa i due filamenti e l'enzima DNA polimerasi aggiunge i nucleotidi complementari. La replicazione avviene in direzione opposta nei due filamenti, risultando in una sintesi continua su un filamento e discontinua sull'altro.

La seconda funzione riguarda la sintesi proteica. Le informazioni codificate nel DNA vengono utilizzate per produrre le proteine tramite l'intervento di una molecola che fa da mediatore, l'RNA. Il processo avviene in due passaggi fondamentali: la trascrizione e la traduzione.

L'RNA serve per trascrivere l'informazione contenuta nel DNA in modo che possa essere tradotta in proteina. L'RNA, come il DNA, è costituito da nucleotidi, ma ha una struttura a singolo filamento, lo zucchero è un ribosio e la timina è sostituita dall'uracile.

Durante la trascrizione, viene sintetizzata una molecola di RNA che ha una sequenza di basi complementare a quella del filamento stampo del DNA. L'enzima RNA polimerasi si lega al sito di inizio della sequenza di DNA, chiamato promotore, separa i filamenti e sintetizza la molecola di RNA che può essere utilizzata nel processo di traduzione per sintetizzare le proteine.

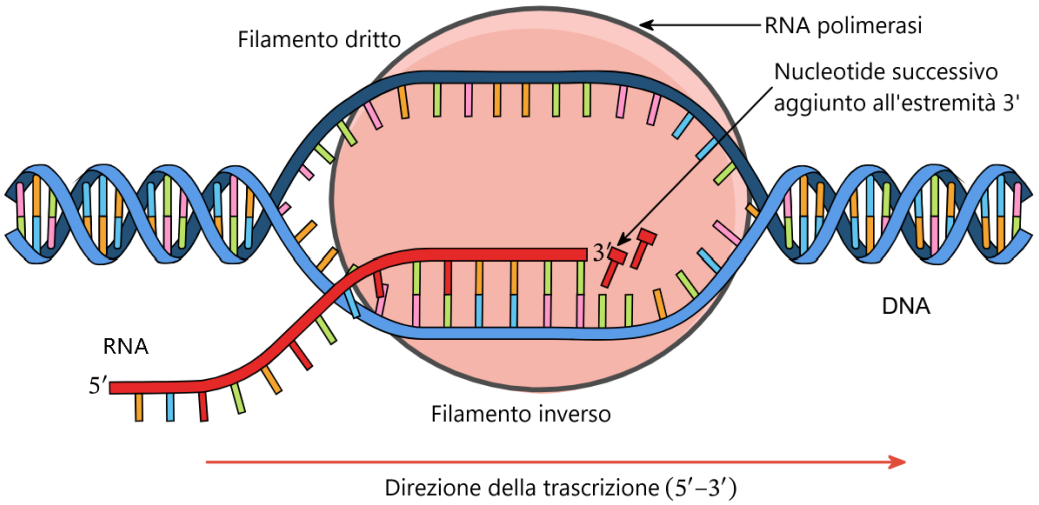

Figura 4: Processo di trascrizione del DNA in RNA. L'RNA polimerasi separa i filamenti di DNA e utilizza uno di essi come stampo per sintetizzare una molecola di RNA complementare. Durante questo processo, il DNA si riunisce dietro l'enzima, mentre l'RNA appena formato si stacca dal filamento stampo.

La traduzione avviene quando l'RNA si lega ai ribosomi nel citoplasma, dove viene letto per costruire una proteina, formata da più amminoacidi legati tra loro. La corrispondenza tra sequenze di basi e amminoacidi è scritta nel codice genetico: ogni tripletta di basi, detta codone, è associata a un particolare amminoacido. Poiché esistono 64 triplette possibili e solo 20 amminoacidi, il codice genetico è detto ridondante: alcuni amminoacidi possono infatti essere codificati da più triplette diverse, ma a ogni tripletta corrisponderà un solo amminoacido in modo da eliminare ambiguità.

La terza funzione è quella di regolazione genica: il DNA contiene regioni che regolano quando e dove determinati geni vengono espressi. Un gene è una sequenza di basi nella molecola di DNA che contiene le informazioni in grado di influire sulle caratteristiche del fenotipo dell'organismo. La regolazione dell'espressione genica è cruciale per il corretto sviluppo e la risposta alle condizioni ambientali; questo processo avviene tramite una serie di meccanismi epigenetici e strutturali.

Alla base dell'espressione genica agisce una rete di regolazione: un complesso sistema di interazioni tra i vari elementi molecolari, come fattori di trascrizione, proteine regolatorie, RNA e sequenze di DNA, che collaborano per controllare l'attivazione o la repressione dei geni.

1.1.3 La cromatina come modello di compattamento e regolazione genica

Il DNA di una cellula umana è lungo circa 2 metri, ma deve adattarsi al volume ridotto del nucleo cellulare, che ha un diametro di appena 5-10 micrometri. Per fare ciò, esso si organizza secondo precise regole gerarchiche di impacchettamento avvolgendosi attorno a proteine, principalmente gli istoni. Questa configurazione compatta e ordinata, chiamata cromatina, permette di regolare l'accessibilità delle informazioni genetiche (1).

La cromatina è, a sua volta, organizzata in modo gerarchico. Il primo livello di compattamento del genoma è rappresentato dal nucleosoma, l'unità fondamentale della cromatina, che consiste in un segmento di DNA avvolto intorno a un ottamero di istoni. I nucleosomi vengono riarrangiati per formare strutture secondarie stabili e si dispongono in domini e sotto-domini topologicamente associati (TADs), che consentono una regolazione più fine dell'espressione genica separando porzioni del genoma con funzioni diverse. Nei sotto-domini la cromatina forma dei loop che avvicinano regioni distanti della struttura lineare del DNA, permettendo l'interazione di elementi regolatori che giocano un ruolo chiave nell'espressione genica, come *promoter* (promotori), *enhancer* (potenziatori), *silencer* (silenziatori) e *insulator* (isolatori).

Le regioni regolatorie, come gli *enhancer,* spesso si trovano a migliaia di basi di distanza dai geni che devono regolare. Tuttavia, grazie alle interazioni a lungo raggio, possono entrare in contatto con il *promoter* attivato sullo stesso cromosoma. Questo porta alla formazione di un anello di DNA che, se osservato linearmente, separerebbe l'*enhancer* dal promotore.
Gli *insulator* fanno da mediatore per consentire o impedire questi contatti in modo da prevenire l'interferenza funzionale tra elementi inappropriati (3).
A un livello più alto di organizzazione, i domini sono ordinati in compartimenti e la loro unione costituisce i cromosomi, ripiegati per occupare territori e posizioni distinte nello spazio nucleare. I cromosomi, costituiti da cromatina condensata, consentono la corretta distribuzione del materiale genetico tra le cellule figlie.

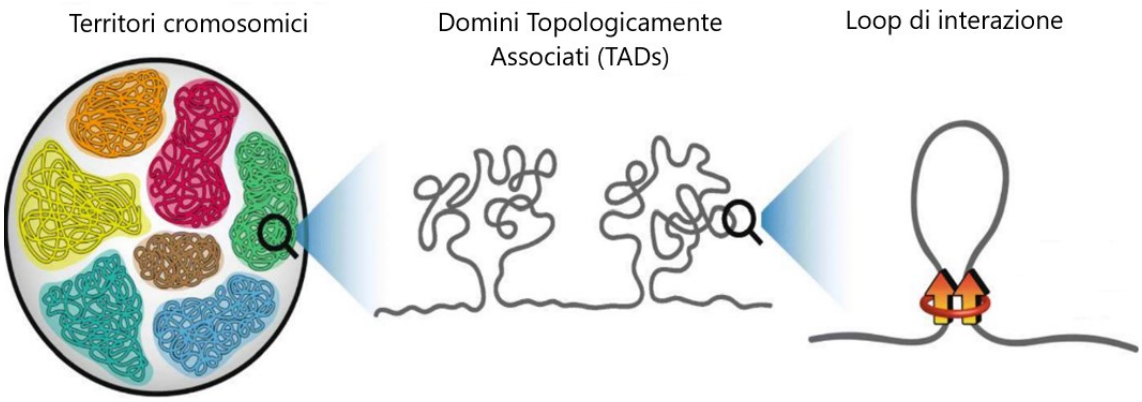

Figura 5: L'immagine illustra la disposizione della cromatina all'interno del nucleo cellulare. I domini topologicamente associati (TADs), nei territori cromosomici, rappresentano regioni del genoma che sono organizzate spazialmente in modo da permettere interazioni specifiche tra sequenze distanti del DNA. I loop collegano in modo dinamico queste regioni consentendo l'interazione di elementi regolatori come i promotori e gli enhancer.

Per regolare l'attivazione o l'inibizione dei geni, possono avvenire modifiche chimiche reversibili che non alterano la sequenza di basi, ma che agiscono direttamente sul DNA o sugli istoni. Queste modifiche, come l'acetilazione, la metilazione e la fosforilazione, coinvolgono specifici amminoacidi e controllano l'accessibilità del DNA fungendo anche da substrati di legame che possono reclutare o escludere complessi proteici. Ad esempio, per ridurre l'attività trascrizionale a livello del DNA, può avvenire la metilazione della citosina, che comporta l'aggiunta di un gruppo metilico alla base azotata.

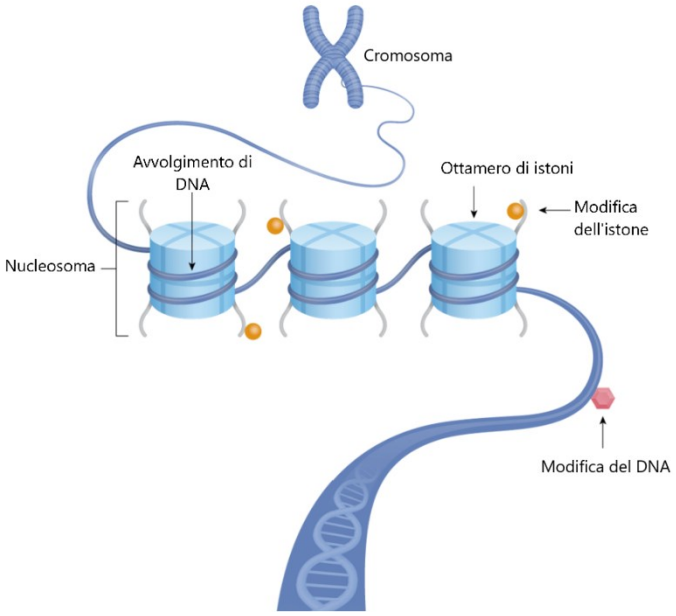

Figura 6: L'immagine illustra l'impacchettamento della cromatina. Le modifiche epigenetiche influenzano l'accessibilità del DNA: possono favorire la condensazione della cromatina, rendendo il DNA meno accessibile, o l'attivazione di geni specifici per la trascrizione.

La cromatina, quindi, può esistere in diversi stati, definiti da combinazioni di modifiche epigenetiche e associati a specifici schemi di regolazione genica: il DNA è accessibile quando è in forma di eucromatina, un tipo di cromatina meno compatta e solitamente attiva nella trascrizione, mentre diventa inaccessibile in forma di eterocromatina, ovvero uno stato cromosomico condensato e inaccessibile ai fattori di trascrizione.

L'accessibilità della cromatina è un meccanismo chiave nella regolazione dell'espressione genica e assume particolare importanza nella progressione tumorale. Cambiamenti nelle sequenze regolatorie o nella struttura della cromatina possono, infatti, attivare oncogeni o reprimere geni oncosoppressori: i primi favoriscono la crescita dei tumori, mentre i secondi la rallentano. In generale, i prodotti proteici di questi geni sono diventati i bersagli di numerose terapie mirate (4).

1.1.4 G-Quadruplex: struttura, localizzazione e funzione biologica

Il DNA è strutturalmente dinamico e può assumere una vasta gamma di conformazioni alternative biologicamente importanti, oltre alla struttura ampiamente conosciuta della doppia elica. In specifiche regioni ricche di guanina, le sequenze di DNA possono ripiegarsi in strutture secondarie non canoniche a quattro filamenti chiamate G-quadruplex (G4) (5). Questi G-quadruplex si formano quando quattro guanine sono collegate tra loro da legami di Hoogsteen per formare una tetrade. Tetradi successive si impilano per formare una struttura compatta, stabilizzata da cationi monovalenti, come potassio o sodio, che si posizionano tra di esse (6,7). Le guanine impilate formano quattro filamenti, che sono connessi da brevi segmenti di nucleotidi chiamati loop. Le diverse combinazioni tra direzione dei filamenti e lunghezza e composizione dei loop permettono al G4 di assumere una vasta gamma di topologie (5). La forma canonica del G4 è costituita da una singola catena contenente quattro tratti di almeno due guanine e tre loop (7). Quando due o più G-quadruplex sono vicini, risultano collegati tra loro da una sequenza che d'ora in poi definiremo con il nome di linker.

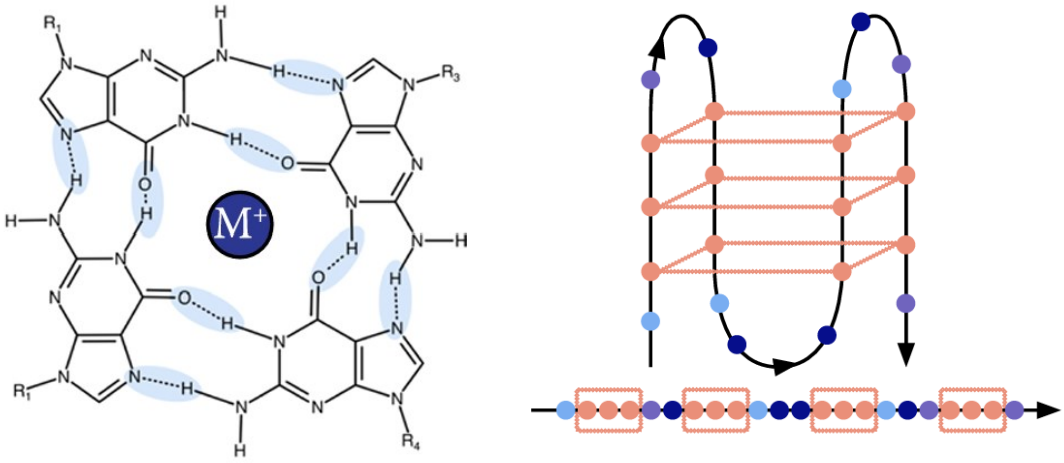

Figura 7: Struttura del G-quadruplex. A sinistra, una tetrade planare costituita da quattro guanine unite da legami di Hoogsteen e stabilizzata da uno ione metallico. A destra, la struttura canonica del G-quadruplex, una delle molteplici topologie che può assumere, formata dall'impilamento di più tetradi di guanine.

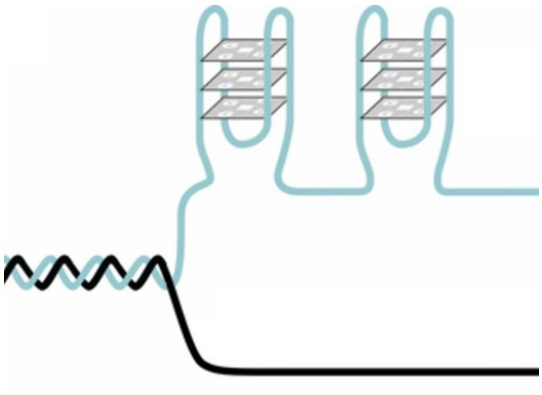

Figura 8: Struttura di un doppio G-quadruplex.

I G-quadruplex sono localizzati in regioni genomiche di grande rilevanza funzionale, in particolare nei telomeri e nei promotori genici, dove agiscono come elementi regolatori.
I telomeri, situati alle estremità dei cromosomi e composti da sequenze altamente ripetute, proteggono il materiale genetico durante la replicazione, impedendo la perdita di informazioni. Qui, i G-quadruplex svolgono un ruolo protettivo, ostacolando l'azione degradativa della telomerasi e contribuendo alla stabilità cromosomica.
Nei promotori genici, che sono sequenze di DNA all'inizio dei geni e fungono da segnali di avvio per la trascrizione, i G-quadruplex possono interferire con l'espressione di determinati geni, in particolare quelli legati alla proliferazione e alla crescita cellulare.
Durante la trascrizione, la separazione del doppio filamento di DNA espone le sequenze ricche di guanina presenti nel filamento singolo che possono ripiegarsi temporaneamente in strutture G-quadruplex. In questo modo si crea una barriera fisica che può ostacolare l'azione della RNA polimerasi, arrivando anche a bloccare la trascrizione. Questo fenomeno è particolarmente importante nei promotori di oncogeni, dove la formazione dei G4 aiuta a modularne l'espressione. Inoltre, i G4 sono in grado di interagire con specifiche proteine regolatorie che influenzano il loro comportamento, in quanto capaci di favorirne lo scioglimento o di stabilizzare la loro struttura.

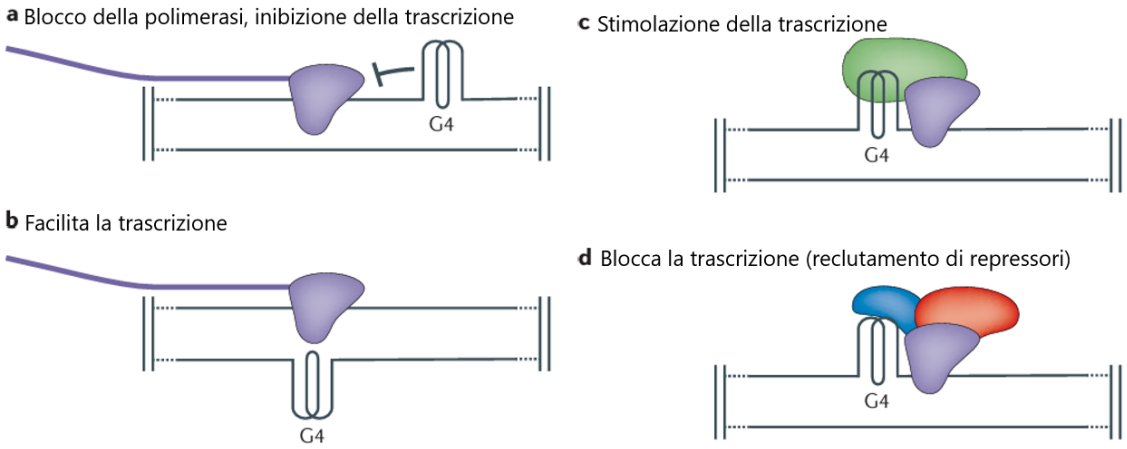

Figura 9: Meccanismi attraverso cui i G-quadruplex influenzano la trascrizione. (a) La formazione di G-quadruplex può ostacolare l'attività dell'RNA polimerasi, impedendo il normale proseguimento della trascrizione. (b) I G-quadruplex possono contribuire a mantenere separati i filamenti di DNA, facilitando la trascrizione. (c) I G-quadruplex reclutano proteine che stimolano l'attività trascrizionale. (d) Al contrario, i G-quadruplex possono legare proteine che inibiscono la trascrizione.

I G-quadruplex sono quindi strettamente legati a numerose funzioni biologiche e, in particolare, hanno un ruolo nel determinare l'identità della cellula attraverso la modulazione dell'espressione genica. Data l'entità del loro coinvolgimento in questi meccanismi, risultano essere bersagli di interesse per lo sviluppo di trattamenti terapeutici specifici (8). Ad esempio, è stata osservata una maggiore presenza di strutture G4 in stati tumorali rispetto a quelli normali, dove svolgono un ruolo nella crescita e progressione di essi (5).

1.2 LE REGIONI ATAC E LA VIMENTINA

1.2.1 La tecnica ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing): mappatura dell'accessibilità cromatinica a scopo genomico

Ogni evento che si verifica all'interno della cellula è la diretta conseguenza di variazioni nell'espressione genica. L'espressione genica è finemente controllata e l'insieme delle componenti che costituiscono la rete di regolazione è coordinato dall'attività dei fattori di trascrizione (TF), che interpretano e modificano lo stato epigenetico della cromatina (9). I fattori di trascrizione sono proteine che regolano l'espressione genica legandosi a specifiche sequenze di DNA nei promotori o in altre regioni regolatorie dei geni. Attraverso questo legame, controllano l'attivazione o la repressione della trascrizione del gene.

Sono stati sviluppati vari saggi, ognuno con specificità diverse, per analizzare e comprendere il paesaggio epigenetico e le sue variazioni in contesti particolari o in seguito a perturbazioni.

Il metodo ATAC-seq è una tecnica avanzata per la mappatura dell'accessibilità della cromatina a livello genomico. Questa tecnologia offre dei vantaggi rispetto ad altre tecniche: riduce la quantità di materiale biologico richiesto a livelli compatibili con l'uso clinico, aumenta la capacità di questi saggi di distinguere tra diversi domini regolatori putativi e non necessita di conoscere preventivamente il meccanismo atteso.

Questo approccio sfrutta la trasposasi Tn5 ingegnerizzata, un enzima caratterizzato da un'elevata attività e dalla capacità di inserirsi nelle regioni prive di nucleosomi. La trasposasi Tn5 catalizza il movimento dei trasposoni da una posizione del genoma a un'altra, si sposta nel genoma grazie al riconoscimento di specifiche sequenze e agisce in modo controllato attraverso un meccanismo detto 'taglia e incolla', in cui taglia il DNA e facilita l'inserimento del trasposone in una nuova posizione.

Il legame dei fattori di trascrizione al DNA è associato alla creazione di regioni estese prive di nucleosomi, dove la trasposasi Tn5 può insinuarsi legandosi direttamente al DNA. Queste aree si trovano vicino ai siti di inizio della trascrizione e in corrispondenza di *enhancer*, *insulator* o *silencer*, offrendo informazioni utili per comprendere come le cellule modulano l'espressione genica (9). Quindi, l'enzima frammenta selettivamente il DNA in regioni di cromatina accessibile, creando un'incisione sfalsata di 9 bp (ovvero una rottura asimmetrica su ciascun filamento di DNA) e integra adattatori per il sequenziamento alle estremità tagliate. Gli adattatori, brevi sequenze di DNA, rendono il DNA compatibile con il sequenziamento, in

modo che possa essere letto e identificato dai sistemi di sequenziamento ad alta capacità. Durante questo processo, l'incisione viene riparata dalla cellula, generando una duplicazione di 9 bp nelle estremità tagliate, una 'firma' caratteristica che permette di identificare le regioni di DNA accessibile. Poiché ciascun estremo di un frammento ATAC-seq corrisponde a un evento di trasposizione unico, viene eseguito il sequenziamento paired-end, che legge entrambe le estremità di ogni frammento di DNA e fornisce informazioni più precise sulla posizione e sull'allineamento dei frammenti. Questa modalità facilita tassi di allineamento unici e più elevati nelle regioni aperte della cromatina. Grazie a questa azione combinata, l'ATAC-seq permette di identificare con rapidità e precisione le aree genomiche prive di nucleosomi, consentendo così l'analisi approfondita dell'architettura della cromatina e delle regolazioni epigenetiche associate.

Dopo il sequenziamento, è necessario eseguire un'analisi dei dati che comprende passaggi di controllo qualità, allineamento, chiamata dei picchi e ulteriori analisi avanzate.

L'allineamento dei frammenti generati con il genoma permette di identificare i 'picchi', che segnalano gli eventi di trasposizione corrispondenti alle regioni di cromatina accessibile e con una possibile funzione regolatoria, come promotori, *enhancer* o silenziatori.

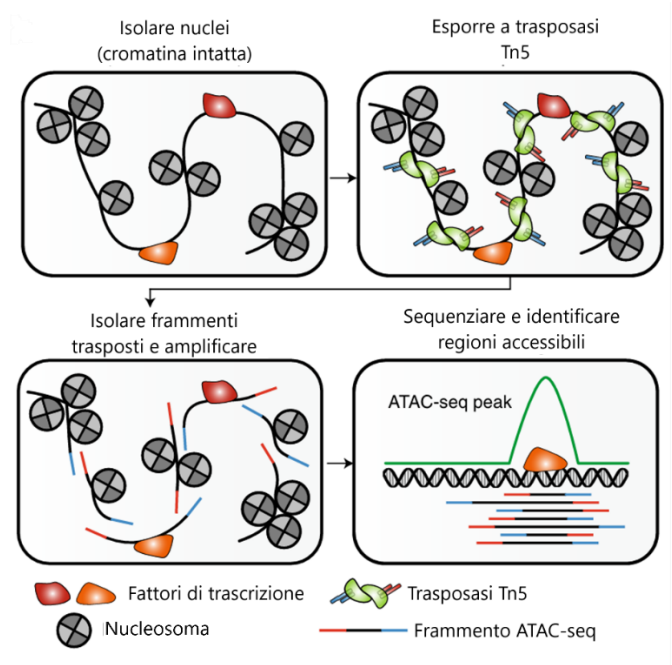

Figura 10: Panoramica dei diversi passaggi della tecnica ATAC-seq. La trasposasi Tn5 agisce sulla cromatina intatta frammentandola e inserendo gli adattatori nelle regioni prive di nucleosomi. Successivamente, i frammenti ottenuti correttamente vengono amplificati e sequenziati. I dati raccolti permettono di identificare i picchi di accessibilità della cromatina.

1.2.2 Il ruolo della vimentina

Il citoscheletro è un componente cruciale per la vita cellulare. Si estende attraverso il citoplasma ed è responsabile della morfologia e della funzionalità delle cellule in vari contesti biologici: è essenziale per il mantenimento della struttura e della forma cellulare, interviene nella mitosi e nella meiosi, facilita la mobilità della cellula e il trasporto intracellulare e contribuisce alla risposta della cellula a segnali esterni.

Gli elementi fondamentali del citoscheletro sono i microtubuli, i microfilamenti e i filamenti intermedi. All'interno della famiglia dei filamenti intermedi si trova la proteina vimentina. La sua struttura è costituita da un dominio centrale ad alfa elica ed è caratterizzata da un'elevata variabilità ai domini N- e C- terminali. Queste differenze strutturali conferiscono alla proteina una diversità funzionale (12). Vimentina svolge un ruolo essenziale nella stabilizzazione della struttura intracellulare e contribuisce alla plasticità e all'assorbimento dello stress meccanico (13). Partecipando all'organizzazione di proteine coinvolte nell'adesione, nella migrazione e nella segnalazione cellulare (14), regola la meccanica cellulare ed è fondamentale per la coordinazione dei processi di meccanosensibilità, trasduzione del segnale e risposte infiammatorie. Inoltre, funziona come piattaforma per il reclutamento di componenti cellulari responsabili della segnalazione, esercitando così un controllo sulle reti di regolazione genica (13). Questa proteina svolge ruoli significativi anche all'interno del nucleo cellulare: fornisce supporto meccanico al nucleo e aiuta a mantenere l'integrità della membrana nucleare durante eventi di migrazione cellulare o stress meccanico (15), oltre ad aiutare a proteggere il DNA da potenziali danni.

Le interazioni di vimentina con altre proteine e con i fattori di trascrizione influenzano i processi cellulari, modulando la condensazione della cromatina e l'attività di numerosi fattori implicati in fenomeni come l'infiammazione e la progressione tumorale (16). L'associazione tra vimentina e DNA, soprattutto in corrispondenza di telomeri e centromeri, suggerisce che vimentina possa mediare modificazioni epigenetiche (17).

Un ruolo fondamentale di vimentina è quello che svolge durante l'evento di transizione epiteliale-mesenchimale (EMT), dove partecipa alla differenziazione e alla migrazione delle cellule. La sua capacità di promuovere l'EMT è collegata a cambiamenti nell'espressione del DNA e alla predisposizione a danni al DNA, influenzando negativamente la stabilità genomica. L'EMT è un processo biologico durante il quale le cellule epiteliali acquisiscono il fenotipo delle cellule mesenchimali per migrare verso i tessuti circostanti e mantenere la capacità di differenziarsi.

Esistono diverse tipologie di EMT legate a condizioni specifiche. L'EMT di tipo 1 è un evento fisiologico che avviene durante lo sviluppo embrionale e l'organogenesi. L'EMT di tipo 2 si attiva in risposta a danni ai tessuti e durante le infiammazioni, favorendo la cicatrizzazione; in condizioni croniche, come nella fibrosi, le cellule non tornano allo stato epiteliale, ma portano alla formazione di tessuti fibrotici. Infine, l'EMT di tipo 3 è associata alla progressione delle metastasi tumorali.

In quest'ultima condizione, vimentina risulta essenziale per favorire la migrazione delle cellule e mantenere la loro capacità staminale. Inoltre, promuovendo l'espressione di geni coinvolti nell'auto-rinnovamento delle cellule tumorali (13,18), essa contribuisce alla formazione di metastasi (19). Per via della sua sovraespressione nei tessuti tumorali, vimentina è utilizzata come marcatore per identificare l'EMT e rappresenta un potenziale bersaglio per la terapia contro il cancro (16,19).

1.2.3 Relazione tra vimentina e strutture G-quadruplex nelle regioni ATAC: implicazioni nella progressione tumorale

In uno studio recente condotto su cellule HGC-27 (17) (cellule derivate da un carcinoma gastrico umano indifferenziato, ampiamente utilizzate nella ricerca), vimentina ha mostrato un'affinità specifica ed elevata per sequenze di DNA costituite da almeno due unità G-quadruplex, note come ripetizioni di G-quadruplex, presenti nei telomeri, nei centromeri e nei promotori genici. Questa scoperta suggerisce che vimentina possa avere un ruolo chiave nel mantenimento dell'integrità del genoma e nella regolazione dell'espressione genica.

Poiché le strutture G4 si formano preferenzialmente nelle regioni ATAC, l'interazione di vimentina con specifiche regioni accessibili di DNA è guidata dalla struttura piuttosto che dalla sequenza. Vimentina, infatti, mostra una predilezione particolare per le ripetizioni G4, indipendentemente dalla loro composizione nucleotidica o dalla loro topologia (17). Questo suggerisce che le ripetizioni G4 possano essere elementi strutturali importanti per l'organizzazione del genoma e potrebbero partecipare alla regolazione dell'espressione di geni chiave durante la migrazione cellulare e lo sviluppo.

L'assenza di interazioni tra vimentina e singoli G4 potrebbe indicare un meccanismo selettivo, che favorisce il legame con regioni genomiche specifiche come telomeri, centromeri e un sottoinsieme selezionato di promotori genici. I geni contenenti ripetizioni G4 su cui si concentra vimentina mostrano una correlazione significativa con geni associati a processi di

comunicazione cellula-cellula e di risposta a stimoli esterni. Questi risultati indicano che il legame della vimentina ai G4 può influenzare l'espressione genica e contribuire a cambiamenti strutturali nel DNA (17). In particolare, il legame vimentina-DNA è funzionale a vimentina per supportare la crescita e la motilità delle cellule tumorali.

Inoltre, le ripetizioni G4 sono presenti nei promotori di geni implicati nella neurogenesi. Un aumento dei livelli di vimentina è stato osservato anche nell'assone in seguito a lesioni neuronali, suggerendo un suo potenziale ruolo nella rigenerazione nervosa (17). Ciò implica che vimentina potrebbe non solo incidere sulla struttura genomica, ma anche influenzare lo sviluppo e la maturazione del sistema nervoso.

In conclusione, la selettività di vimentina per le ripetizioni G4 evidenzia un meccanismo attraverso il quale questa proteina può modulare l'espressione genica in contesti fisiologici e patologici. Questi risultati sottolineano come l'inibizione del legame tra vimentina e G4 possa diventare un potenziale ambito di studio per lo sviluppo di un nuovo approccio terapeutico, volto a contenere la formazione di metastasi nelle cellule tumorali.

1.3 CONTESTO E IMPORTANZA DELLA RICERCA

Il cancro è una delle principali sfide per la salute pubblica e rappresenta la seconda causa di morte a livello mondiale (20). Si tratta di una patologia complessa, caratterizzata da una crescita cellulare irregolare e incontrollata che può colpire vari tipi di cellule e tessuti (21) e, nelle fasi avanzate, può diffondersi in altre parti del corpo attraverso le metastasi (22).

Il cancro è una malattia genetica, scatenata dall'accumulo di mutazioni che alterano i geni coinvolti nella regolazione della crescita e della morte cellulare (22). Le cellule tumorali, infatti, perdono la capacità di regolare la propria proliferazione e di entrare in apoptosi, cioè in morte cellulare programmata. Le mutazioni possono essere causate da errori durante la replicazione del DNA o da danni dovuti a fattori ambientali.

Non sono solo le mutazioni a dare inizio al processo tumorale, anche le alterazioni dei sistemi di regolazione epigenetica giocano un ruolo fondamentale. Cambiamenti epigenetici anomali possono infatti indurre espressioni geniche inappropriate, favorendo l'insorgenza e lo sviluppo del tumore (23). Le alterazioni sia nei meccanismi genetici sia nella regolazione epigenetica sono dunque essenziali per la progressione tumorale. L'iperespressione degli oncogeni e il

silenziamento dei geni oncosoppressori sono fattori chiave per l'aggressività del cancro. Inoltre, le interazioni tra proteine regolatorie e DNA modulano l'accessibilità della cromatina, rendendo alcune aree del genoma più vulnerabili a modifiche epigenetiche che influenzano percorsi cellulari vitali per la sopravvivenza del tumore.

Negli ultimi anni, la ricerca si è concentrata su nuove strategie per bloccare la crescita tumorale e migliorare l'efficacia dei trattamenti. Le tecnologie di sequenziamento di nuova generazione (NGS) hanno rivoluzionato la comprensione del genoma e dell'epigenoma, consentendo l'identificazione di mutazioni e alterazioni epigenetiche specifiche, fondamentali per lo sviluppo di terapie personalizzate. Queste informazioni permettono di intervenire su anomalie molecolari precise delle cellule tumorali, riducendo gli effetti collaterali rispetto alle terapie tradizionali.

1.3.1 Il ruolo della regolazione genica nello sviluppo di tumori

Oltre alle mutazioni, i cambiamenti epigenetici contribuiscono significativamente alla progressione del cancro, intervenendo nei processi di espressione genica e favorendo la perdita o il guadagno di funzioni cellulari (24). I meccanismi epigenetici, come la metilazione del DNA e le modificazioni degli istoni, risultano specifici per tipo di tumore e presentano un'interconnessione con altri segni epigenetici. Queste alterazioni si verificano nelle regioni regolatorie dei geni e in sequenze ripetitive del genoma, influenzando stabilità trascrizionale e regolazione della cromatina (24).

Una caratteristica distintiva delle alterazioni epigenetiche è la loro reversibilità, che le rende bersagli promettenti per nuove terapie oncologiche. A differenza delle mutazioni genetiche permanenti, i cambiamenti epigenetici possono essere modulati rapidamente con composti che mirano agli enzimi epigenetici che risultano utili per controllare l'espressione genica nelle cellule tumorali (23).

La formazione di strutture secondarie del DNA, come i G-quadruplex, è un altro aspetto rilevante per la regolazione genica e la biologia del cancro. Queste strutture possono promuovere o inibire l'espressione dei geni a seconda della loro posizione e dell'interazione con fattori di trascrizione. I G-quadruplex sono particolarmente abbondanti nelle regioni promotrici, contribuendo all'organizzazione genomica e all'accessibilità della cromatina, con impatti diretti sull'inizio della trascrizione (24).

L'influenza epigenetica sui tumori non si limita a singoli geni: essa modella il panorama epigenetico e genomico del cancro, contribuendo alla complessità e alla variabilità della

malattia. L'accumulo di mutazioni genetiche e di alterazioni epigenetiche determina caratteristiche chiave dei tumori, come proliferazione incontrollata, resistenza all'apoptosi, invasione, metastasi e instabilità genomica (25).
Tali anomalie possono essere corrette farmacologicamente, pur senza modificare direttamente la sequenza genetica, rallentando la progressione tumorale o riducendo la necessità di dosi elevate di chemioterapici (24).
In sintesi, le alterazioni epigenetiche giocano un ruolo determinante nel cancro, operando in sinergia con le mutazioni genetiche per mantenere il fenotipo maligno. La comprensione approfondita di questi meccanismi potrebbe favorire lo sviluppo di nuove strategie terapeutiche mirate, applicabili a un'ampia gamma di patologie (24).

1.3.2 Obbiettivo della tesi: identificare pattern genomici che colleghino le regioni ATAC e vimentina come potenziali target terapeutici

L'ipotesi che la vimentina leghi preferenzialmente ripetizioni di G-quadruplex apre nuove possibilità per sviluppare approcci terapeutici. In questo contesto si colloca questa ricerca, il cui obiettivo principale è identificare sequenze di G-quadruplex nelle regioni ATAC che possano fungere da bersagli terapeutici per inibire il legame con vimentina e la progressione tumorale che ne potrebbe derivare.
Analizzando le regioni ATAC, sarà possibile mappare le aree genomiche accessibili e attive nella regolazione dell'espressione genica. Successivamente, l'individuazione delle sequenze potenzialmente capaci di formare strutture G4 permetterà di esaminare le regioni in cui si localizza vimentina per identificare motivi specifici che caratterizzano il legame G4-vimentina. Questi dati contribuiranno a proporre le ripetizioni di G4 come potenziali target terapeutici e l'integrazione delle informazioni ottenute consentirà di migliorare la comprensione dei meccanismi regolatori alla base della patologia tumorale.

CAPITOLO 2: MATERIALI E METODI

2.1 ANALISI BIOINFORMATICA DEI DATI ATAC-SEQ: WORKFLOW E STRUMENTI UTILIZZATI

L'analisi dei dati ATAC-seq si svolge attraverso tre fasi principali: una pre-analisi, l'identificazione dei picchi e un'analisi avanzata (9,10). La pre-analisi è necessaria per preparare l'allineamento al genoma dei file FASTQ ottenuti dal sequenziamento. Successivamente, le sequenze vengono utilizzate per identificare i picchi, che corrispondono agli eventi di trasposizione e quindi identificano le regioni di accessibilità della cromatina. Infine, l'analisi a valle si concentra sull'estrazione di informazioni biologiche significative dai dati elaborati.

2.1.1 Acquisizione e preprocessing dei dati ATAC-seq

Attraverso la piattaforma NCBI Gene Expression Omnibus è possibile ricavare i dati generati dal sequenziamento utilizzando un codice. Questo codice permette di accedere a una pagina contenente un altro codice, disponibile in Bioproject, da inserire in ENA Browser. Sempre in NCBI è necessario aprire SRA run selector, che fornisce informazioni sulla tecnica utilizzata per ricavare le sequenze e sul tipo di sequenziamento. Dopo avere ottenuto i codici per le sequenze relative ad ATAC-seq, due per ogni libreria ottenuta con sequenziamento paired-end, è possibile identificare in ENA Browser le sequenze di interesse. Per scaricare i file direttamente nell'ambiente di calcolo si utilizza il comando *wget* con il riferimento al file desiderato.

```
wget -nc -P ${cartella1} ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR148/089/SRR14879789/SRR148
79789_2.fastq.gz

wget -nc -P ${cartella1} ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR148/089/SRR14879789/SRR148
79789_1.fastq.gz
```

Codice 1: Esempio di linea di comando per scaricare i dati.

Si ottengono così file in formato FASTQ, contenenti quattro elementi informativi diversi disposti su righe: una stringa che identifica la sequenza, la sequenza di basi, uno spaziatore e un valore che indica la qualità della sequenza.

Una volta ottenuti i dati, si può procedere con i vari passaggi previsti da questo tipo di analisi, ciascuno caratterizzato dall'utilizzo di uno strumento bioinformatico diverso.

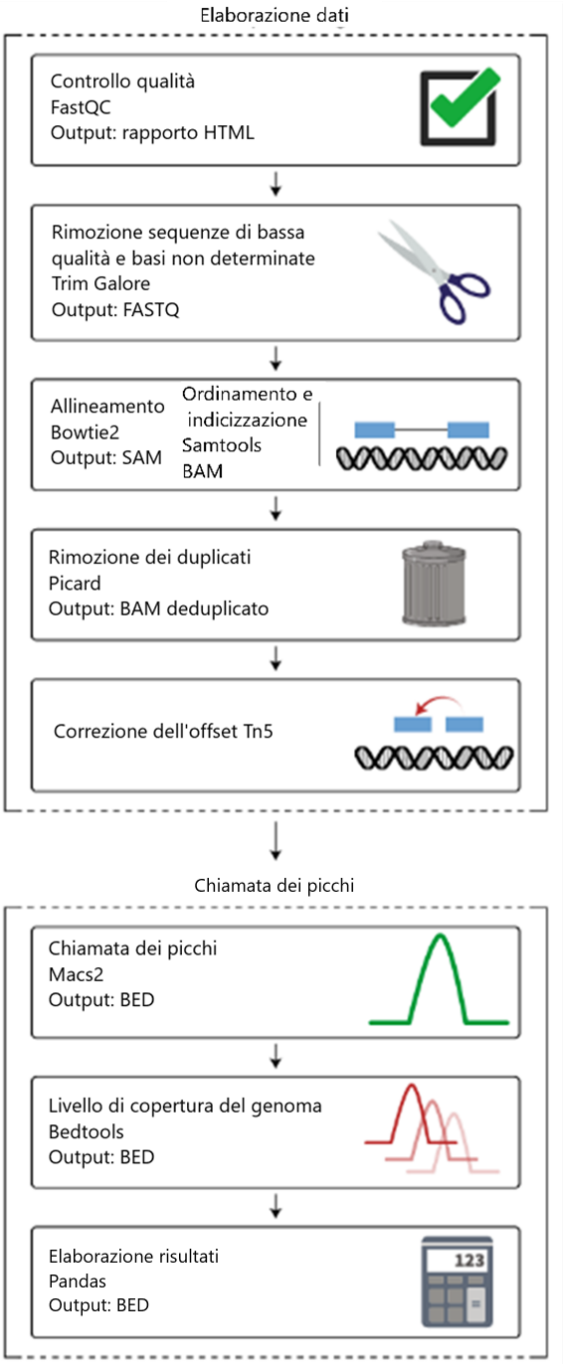

Figura 11: Workflow dell'analisi ATAC-seq.

2.1.2 FastQC

Il primo passaggio dell'analisi consiste in semplici controlli di qualità per identificare possibili problemi nei dati grezzi e per valutare se sono necessari passaggi di filtraggio o pre-elaborazione prima di proseguire con l'analisi. FastQC fornisce una misura della precisione con cui le basi di una sequenza di DNA sono state lette durante il processo di sequenziamento.
A questo scopo, FastQC analizza ogni sequenza ed elabora una serie di statistiche e grafici che forniscono una panoramica della qualità complessiva del set di dati. Infine, fornisce un rapporto in formato HTML visualizzabile nel browser. Ogni sezione del report viene accompagnata da un segnale di qualità che varia tra *pass*, *warn* e *fail*, questo consente di individuare rapidamente le aree problematiche che potrebbero richiedere correzioni.

```
fastqc -o ${cartella1} -t 2 ${cartella0}"$3_1.fastq.gz" ${cartella0}"$3_2.fastq.gz"
```

Codice 2: Esempio di linea di comando per eseguire FastQC.

Opzioni utilizzate:

- -o: specifica la cartella di output.
- -t: indica il numero di file FASTQ da analizzare.

Il report generato da FastQC include vari moduli di analisi della qualità, tra cui:

- 'Statistiche di base': una tabella con informazioni semplici sul file FASTQ di input, inclusi nome, tipo di codifica del punteggio di qualità, numero totale di reads, lunghezza di read e contenuto di GC.

- ‘Qualità della sequenza per base’: mostra la qualità media di ciascun nucleotide in tutte le sequenze, visualizzando le deviazioni standard. Il punteggio di qualità medio tende a diminuire progressivamente lungo la lunghezza della sequenza. Le statistiche aggregate dei punteggi di qualità in ciascuna posizione lungo tutte le reads nel file vengono mostrate con un grafico BoxWhiskers.

  La Figura 12 mostra la differenza nella qualità della sequenza per base di dati aventi un buon punteggio di qualità e dati di bassa qualità. Nel grafico a sinistra il punteggio di qualità è elevato e costante, questo indica che le reads sono di alta qualità, senza significativi cali di qualità nelle posizioni esaminate; il grafico a destra evidenzia fluttuazioni e cali significativi nel punteggio di qualità, ciò suggerisce la presenza di errori o sequenze di bassa qualità in alcune posizioni delle reads.

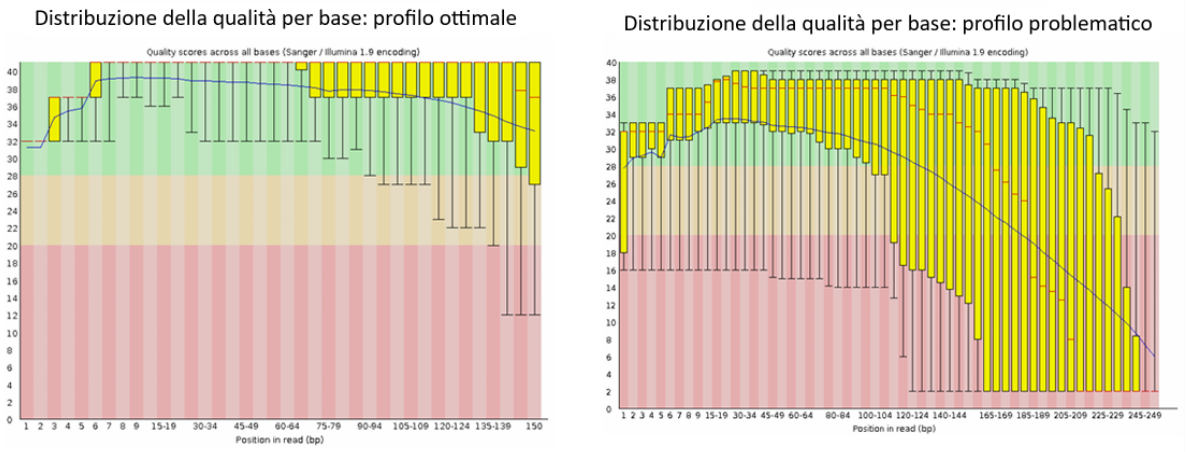

Figura 12: Confronto tra due grafici BoxWhiskers ottenuti con FastQC per valutare la qualità della sequenza per base. L’asse delle ascisse rappresenta la posizione della base lungo la lunghezza delle reads di sequenziamento. L’asse delle ordinate rappresenta il punteggio di qualità per ciascuna posizione della base.

- ‘Punteggi di qualità per sequenza’: mostra la distribuzione dei punteggi di qualità di tutte le sequenze, fornendo un’indicazione generale della qualità del set di dati. La distribuzione della qualità media delle reads dovrebbe essere piuttosto uniforme nella parte alta del grafico.

  In Figura 13 è possibile osservare il confronto tra due grafici che rappresentano il punteggio di qualità per sequenza. Il grafico a sinistra mostra una distribuzione dei punteggi di qualità considerata ‘buona’, con la maggior parte delle sequenze che presenta punteggi elevati e costanti. Il grafico a destra illustra una distribuzione ‘cattiva’, caratterizzata da punteggi di qualità variabili e inferiori, indicativa di potenziali problemi nella qualità dei dati di sequenziamento.

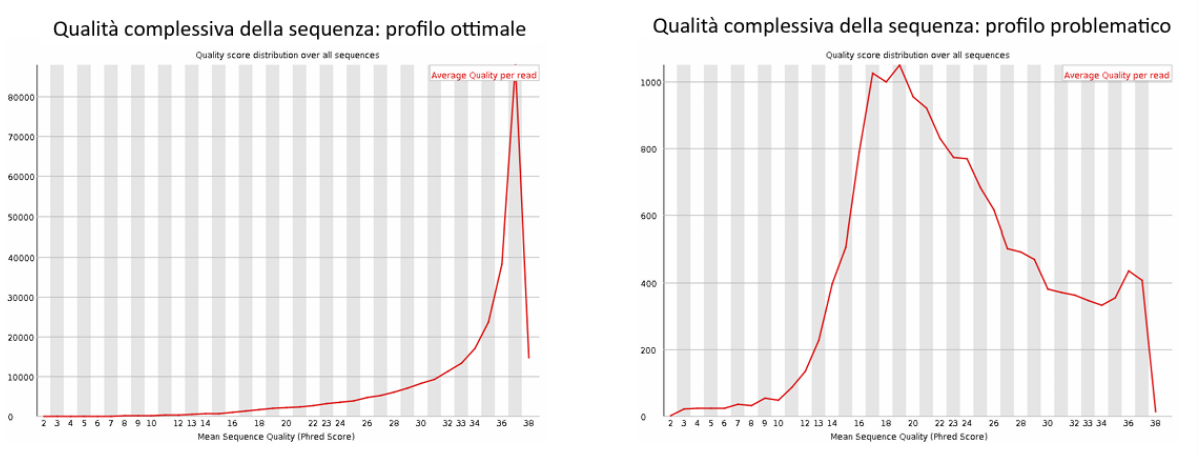

Figura 13: Grafici comparativi dei punteggi di qualità per sequenza. L’asse delle ascisse indica la qualità media della sequenza per ciascun intervallo di reads. L’asse delle ordinate rappresenta la quantità di reads che rientrano in ciascun intervallo di qualità.

- ‘Contenuto della sequenza per base’: indica la proporzione dei quattro nucleotidi (A, T, C, G) in ciascuna posizione all'interno delle reads. La proporzione di ciascuna delle quattro basi dovrebbe rimanere relativamente costante lungo la lunghezza della read ed eventuali squilibri possono indicare un bias di sequenziamento. Tuttavia, il contenuto di sequenza discordante all'inizio delle reads è un fenomeno comune per le reads ottenute in seguito agli eventi di trasposizione con trasposasi Tn5 che induce incisioni sfalsate di 9 bp.

  In Figura 14, il grafico a sinistra mostra una distribuzione costante delle proporzioni delle basi nel campione, indicativa di una buona qualità della sequenza. Al contrario, il grafico a destra presenta un contenuto discordante delle basi, segnalando potenziali problemi di qualità e uniformità nelle reads sequenziate.

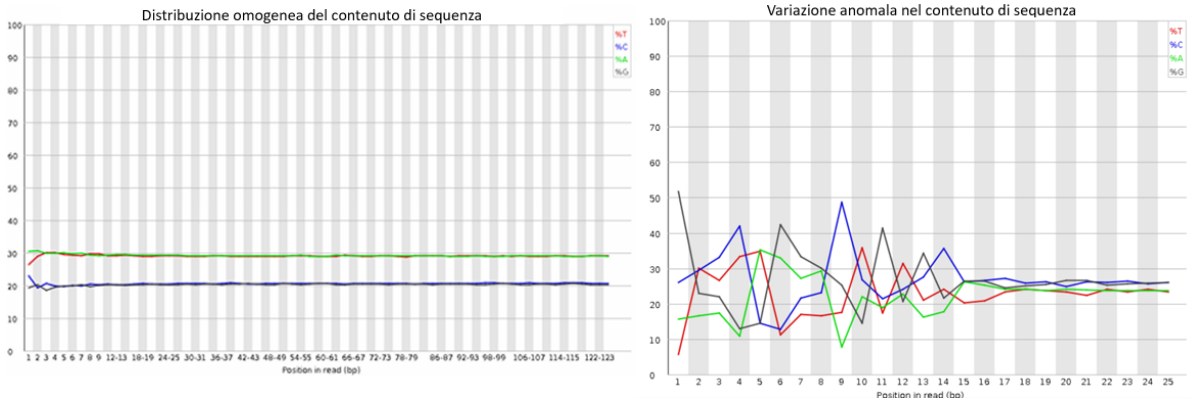

Figura 14: Contenuto della sequenza per base secondo FastQC. L’asse delle ascisse rappresenta la posizione delle basi lungo le reads sequenziate. L’asse delle ordinate indica la percentuale di ciascuna delle basi (A, T, C, G) presenti a ogni posizione della sequenza.

- ‘Contenuto GC per sequenza’: mostra la distribuzione delle percentuali di GC in tutte le sequenze, aiutando a identificare contaminazioni come RNA ribosomale o frammenti di altri organismi. Ciò che ci si aspetta è che il contenuto di GC di tutte le reads formi una distribuzione normale, con il picco della curva al contenuto medio di GC.

  La Figura 15 mostra il picco al contenuto medio di GC atteso che indica una buona qualità dei dati. Mentre una deviazione significativa dalla distribuzione normale attesa potrebbe indicare problemi nella qualità del sequenziamento.

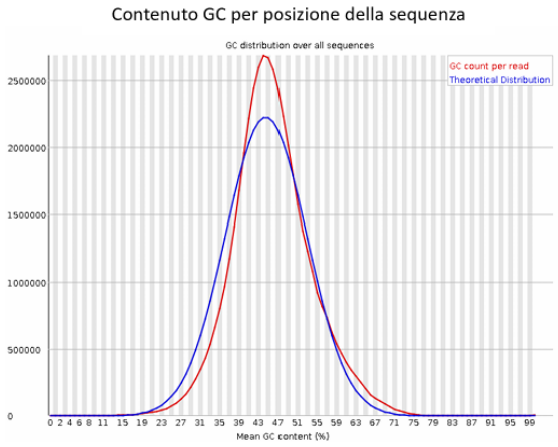

Figura 15: Grafico del contenuto di GC delle reads, che mostra la distribuzione osservata (in rosso) rispetto alla distribuzione teorica prevista (in blu). L’asse delle ascisse rappresenta la posizione delle basi lungo la sequenza. L’asse delle ordinate indica la percentuale di contenuto GC per ogni posizione della sequenza.

- ‘Contenuto di "N" per base’: indica la percentuale di basi ‘N’, cioè senza assegnazione di base, in ogni posizione delle reads. Se la curva si innalza in modo evidente sopra lo zero indica un problema verificatosi durante il ciclo di sequenziamento.

  La Figura 16 mostra la curva relativa a un incremento del numero di basi non determinate (N), indicando una potenziale problematica durante il sequenziamento.

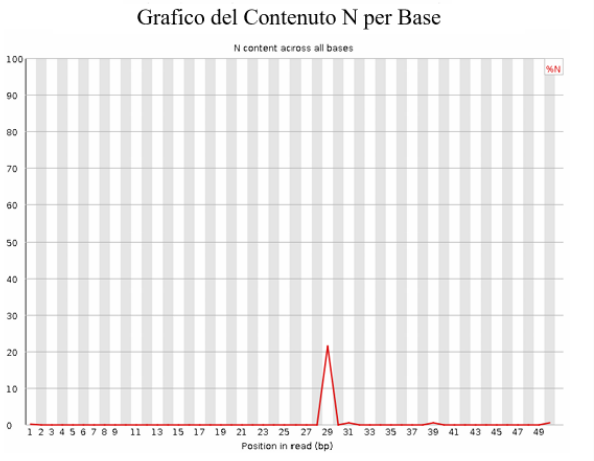

Figura 16: Grafico del contenuto di 'N' per base di FastQC. L’asse delle ascisse indica la posizione delle basi lungo la sequenza. L’asse delle ordinate indica la percentuale di basi ‘N’ per ogni posizione della sequenza.

- ‘Distribuzione della lunghezza delle sequenze’: visualizza la lunghezza delle reads nel set di dati. In generale, ci si aspetta frammenti di lunghezza uniforme, ma possono essere presenti anche reads di lunghezze molto variabili.

  In Figura 17, il grafico evidenzia un picco significativo all'inizio, indicando che la maggior parte dei frammenti di sequenza analizzati ha una lunghezza uniforme. Questo è tipico per librerie di sequenziamento di alta qualità, suggerendo un'adeguata preparazione della libreria e una bassa presenza di contaminazioni o frammenti di lunghezza variabile.

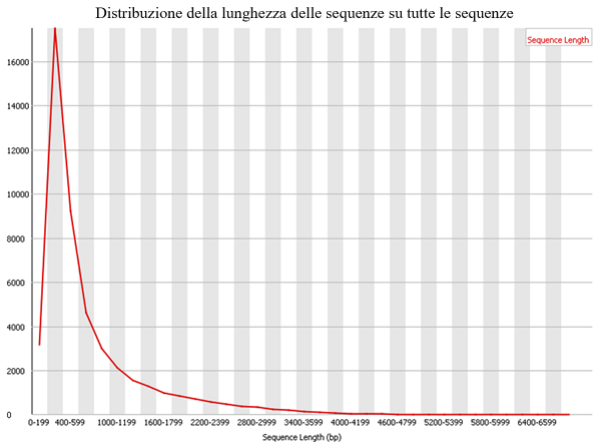

Figura 17: Distribuzione della lunghezza delle sequenze. L’asse delle ascisse rappresenta la lunghezza delle sequenze, cioè il numero di basi di ciascuna. L’asse delle ordinate indica il numero di sequenze che corrispondono a ciascuna lunghezza specifica.

- ‘Livelli di duplicazione delle sequenze’: calcola la percentuale di sequenze duplicate nel file. In una libreria altamente diversificata e non sovra-sequenziata quasi il 100% delle reads è unico. Un’elevata percentuale di duplicazione può indicare che i dati provengono da una sequenza specifica sovrarappresentata o che ci sono problemi di sequenziamento.

  In Figura 18, il grafico a sinistra mostra un'alta percentuale di reads uniche, ciò indica una libreria ben diversificata e non sovrasequenziata, suggerendo una buona qualità del campione. Nel grafico a destra, si osservano dei picchi che indicano una significativa presenza di duplicazioni, il che potrebbe compromettere l'affidabilità dei risultati dell'analisi.

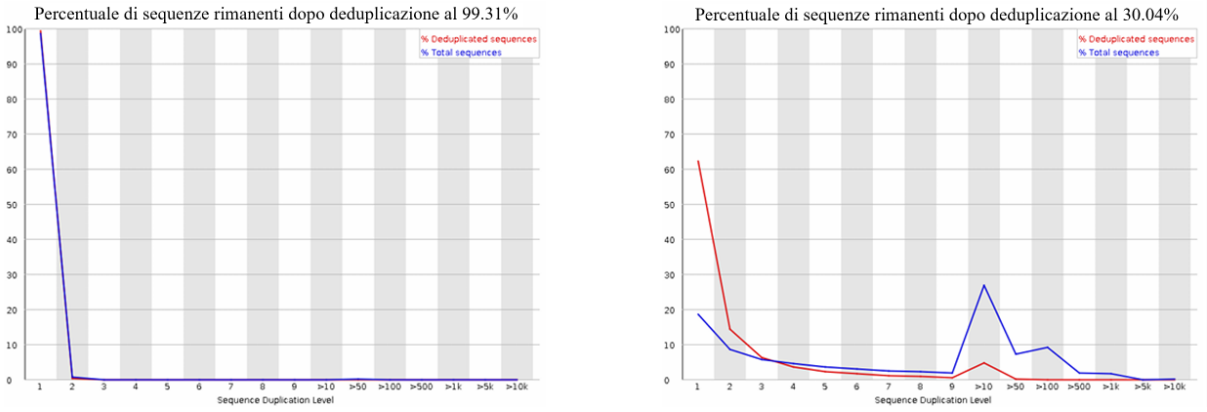

Figura 18: Grafico della distribuzione dei livelli di duplicazione delle sequenze. L’asse delle ascisse indica la percentuale di sequenze deduplicate. L’asse delle ordinate indica la percentuale di sequenze rimanenti dopo il processo di deduplicazione.

- ‘Sequenze sovra-rappresentate’: elenca le sequenze che compaiono più frequentemente del previsto nel file. Nei dati di sequenziamento del DNA, non dovrebbe esserci alcuna singola sequenza presente a una frequenza sufficientemente alta da essere elencata (≥ 0,1% delle reads totali), anche se non è insolito vedere una piccola percentuale di reads di adattatori.

- ‘Contenuto di adattatori’: mostra la presenza di adattatori o altre sequenze indesiderate. Un elevato contenuto di adattatori può richiedere la rimozione prima dell’analisi successiva.

  In Figura 19, la curva mostra un aumento del contenuto degli adattatori verso la fine delle reads, suggerendo che alcune delle sequenze di DNA presenti nella libreria siano più corte della lunghezza della read.

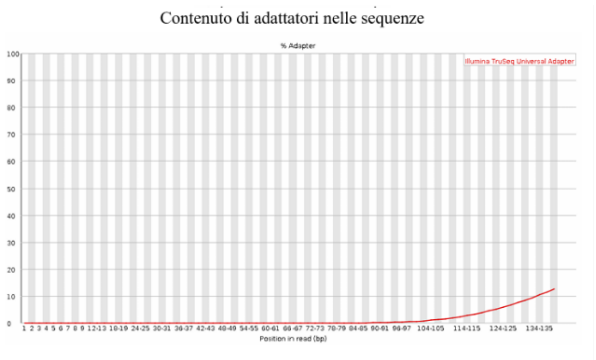

Figura 19: Grafico della frazione di reads in cui è stata identificata la sequenza di un adattatore della libreria alla posizione di base indicata. L’asse delle ascisse rappresenta le posizioni all’interno della sequenza. L’asse delle ordinate indica la quantità di nucleotidi che corrispondono a sequenze di adattatori in quella posizione specifica della sequenza.

2.1.3 Trim Galore

Trim Galore è uno strumento di pre-elaborazione utilizzato per il trimming delle reads di sequenziamento, ovvero per rimuovere le sequenze di adattatori e le basi di bassa qualità dai file FASTQ. Consente di ottimizzare la qualità delle reads prima dell'analisi successiva, migliorando così l'accuratezza dei risultati in esperimenti di sequenziamento di nuova generazione.

```
singularity exec --bind /data:/data ${TRIM_G} trim_galore -q $4 --fastqc --length $5 --tr
im-n --max_n $6 --output_dir ${cartella0}"trim" --paired --cores 4 ${cartella0}"$3_1.fast
q.gz" ${cartella0}"$3_2.fastq.gz"
```

Codice 3: Esempio di riga di comando per eseguire Trim Galore.

Singularity è una piattaforma che permette di eseguire applicazioni in ambienti scientifici, senza bisogno di installazioni locali. Quindi, si utilizza *singularity* per eseguire *trim_galore* in modo da rimuovere le sequenze di bassa qualità e le basi non determinate N e per generare un report FastQC gestendo le reads secondo specifiche opzioni di trimming.

Opzioni utilizzate:

- -q: imposta una soglia minima di qualità per il trimming delle basi.
- --fastqc: genera un report di qualità FastQC sui file output, permettendo di verificare la qualità delle sequenze dopo il trimming.
- --length: imposta la lunghezza minima della read dopo il trimming; tutte le sequenze con una lunghezza inferiore a questo valore verranno scartate.
- --trim-n: rimuove le eventuali basi N all'inizio e alla fine delle reads.
- --max_n: specifica la percentuale massima di basi N consentite in una read. Se una read contiene più di questo numero di N, verrà scartata (in paired-end vengono rimosse entrambe).
- --output_dir: specifica la directory di output, se non esiste la crea.
- --paired: indica che i file di input sono reads paired-end. Serve per assicurare che le coppie di reads rimangano allineate anche dopo il trimming, rimuovendo intere coppie se una delle due sequenze è più corta della soglia.
- --cores: numero di core del processore che verranno utilizzati durante l'esecuzione del programma.
- file FASTQ in input.

2.1.4 Bowtie2

Bowtie2 è un software di allineamento utilizzato per mappare rapidamente sequenze di DNA corte su un genoma di riferimento. La fonte del genoma di riferimento è T2T-CHM13v2.0, che rappresenta la versione più recente derivante dal sequenziamento completo del genoma umano ed è disponibile su NCBI.

```
singularity exec --bind /data:/data ${Bowtie2} bowtie2 -x "/data/valentina.ceriani/
reference/$7" -1 ${cartella0}"trim/"*"_val_1"*".gz" -2 ${cartella0}"trim/"*"_val_2"
*".gz" -X $8 --no-unal --no-mixed --threads $9 -S ${cartella2}"$3" &> ${cartella2}"
log.txt"
```

Codice 4: Esempio di riga di comando per Bowtie2.

Usando *singularity*, viene eseguito *bowtie2*, che esegue un allineamento tra le sequenze di read in formato FASTQ e una sequenza di riferimento, creando un file di output in formato SAM (Sequence Alignment Map) che contiene i risultati degli allineamenti.

Opzioni utilizzate:

- -x: specifica l'indice del genoma di riferimento su cui valutare l'allineamento.
- -1: specifico per reads paired-end, indica i file contententi le reads forward.
- -2: specifico per reads paired-end, indica i file contententi le reads reverse.
- -X: imposta la distanza massima (in numero di basi) per le sequenze paired-end.
- --no-unal: esclude dall'output le sequenze che non sono state allineate.
- --no-mixed: assicura che vengano allineate solo le raeds paired-end complete, impedendo di cercare allineamenti singoli.
- --threads: specifica il numero di thread da utilizzare per ricerca parallela degli allineamenti in modo da aumentare la velocità dell'allineamento.
- -S: indica il nome del file di output in formato SAM che conterrà gli allineamenti.
- &>: reindirizza l'output in un file di log, utile per monitorare il processo e per eventuali diagnostiche.

2.1.5 Samtools

Samtools è un insieme di utilità per la manipolazione degli allineamenti nel formato BAM. Consente di convertire un file SAM in un file BAM, di eseguire ordinamenti, unioni e indicizzazione e di recuperare le reads in qualsiasi regione in modo rapido.

```
singularity exec --bind /data:/data ${Samtools} samtools view -bSh --threads $9 ${c
artella2}"$3.sam" > ${cartella3}"$3.bam"
```

Codice 5: Esempio di linea di comando per eseguire samtools view.

Usando *singularity*, viene eseguito il comando *view* di *samtools*, che permette di convertire un file SAM in un file BAM (i file BAM sono semplicemente file SAM compressi).

Opzioni utilizzate:

- -bSh: è una combinazione di diverse opzioni. -b indica che l'output deve essere in formato BAM, -S specifica che l'input è in formato SAM, -h include l'intestazione del file SAM nel file di output.
- --threads: utilizza il numero di thread specificato per migliorare le prestazioni, permettendo una conversione più veloce.
- file.sam > file.bam: specifica il percorso e il nome del file di input in formato SAM e reindirizza l'output al file di output in formato BAM.

```
singularity exec --bind /data:/data ${Samtools} samtools sort ${cartella3}"$3.bam"
--threads $9 -o ${cartella3}"$3_sorted.bam"
```

Codice 6:Esempio di linea di comando per eseguire samtools sort.

Usando *singularity*, viene eseguito il comando *sort* di *samtools*, che ordina le sequenze nel file BAM in base alle coordinate genomiche. L'ordinamento è essenziale per molte analisi downstream, come la marcatura dei duplicati e la visualizzazione dei dati. Un file BAM ordinato per coordinate genomiche facilita infatti sia l'identificazione dei duplicati sia la navigazione tra diverse regioni genomiche.
Ordinare il BAM quindi assicura che i dati siano pronti per questi e altri tipi di analisi, migliorando l'efficienza e l'accuratezza degli strumenti che sfruttano l'ordine delle coordinate per analizzare le sequenze genomiche.

Opzioni utilizzate:

- file.bam: specifica il file di input in formato BAM da ordinare.
- --threads: utilizza il numero di thread specificato per accelerare il processo di ordinamento.
- -o: specifica il file di output in formato BAM ordinato.

```
singularity exec --bind /data:/data ${Samtools} samtools index ${cartella3}"$3_sorted.bam
"
```

Codice 7: Esempio di linea di comando per eseguire samtools index.

Usando *singularity*, viene eseguito il comando *index* di *samtools*, che viene utilizzato per creare un file indice per un file BAM. Questo file indice è fondamentale per le analisi bioinformatiche, in quanto consente di accedere rapidamente a specifiche posizioni nel file BAM senza dover caricare l'intero file in memoria.

Opzioni utilizzate:

- file.bam: specifica il file BAM ordinato per il quale si desidera creare l'indice.

L'output di questo comando è un file BAM index (.bai), cioè un file di indice associato a un file BAM che contiene informazioni sulle posizioni delle reads mappate nel file BAM, facilitando l'accesso rapido ai dati.

2.1.6 Picard

Picard permette di manipolare e analizzare file di allineamento nei formati BAM e SAM.

```
singularity exec --bind /data:/data ${Picard} picard MarkDuplicates \
--INPUT ${cartella3}"$3_sorted.bam" \
--OUTPUT ${cartella4}"$3_deduplicates.bam" \
--METRICS_FILE ${cartella4}"$3_metrics.txt" \
--REMOVE_DUPLICATES true
```

Codice 8: Esempio di linea di comando per eseguire Picard.

Usando *singularity*, viene eseguito *picard* con il comando *MarkDuplicates* per marcare e rimuovere reads duplicate da un file BAM già ordinato, registrando le metriche relative

all'operazione in un file di output specificato. Questo è un passo cruciale nei flussi di lavoro di analisi genomica, poiché le reads duplicate possono introdurre errori nelle analisi successive.

Opzioni utilizzate:

- --INPUT: specifica il file di input che è un file BAM già ordinato.
- --OUTPUT: specifica il file BAM di output dove scrivere le sequenze deduplicate.
- --METRICS_FILE: specifica il file in cui registrare le metriche relative alle reads duplicate.
- --REMOVE_DUPLICATES: indica che le reads duplicate devono essere rimosse. Impostando true, Picard non solo marca le raeds duplicate, ma le rimuove anche dal file di output.

Prima di procedere con la chiamata dei picchi, è necessario correggere le posizioni di inizio di ciascuna read. Infatti, la riparazione delle incisioni create dalla trasposasi Tn5 genera una duplicazione di 9 bp che potrebbe portare a risultati errati. La regolazione dell'offset di Tn5 avviene aggiungendo 4 bp per il filamento positivo e togliendo 5 bp per il filamento negativo.

2.1.7 Macs2

MACS2 è uno strumento progettato per identificare i picchi, ovvero aree del genoma arricchite con reads allineate. Queste regioni suggeriscono che molte cellule nel campione abbiano una cromatina accessibile al Tn5 in quei siti specifici. Questo processo è fondamentale per studiare la regolazione genica e le interazioni tra proteine e DNA.

```
#NarrowPeaks

singularity exec --bind /data:/data ${Macs2} macs2 callpeak -t ${cartella4}"$3_dedu
plicates.bam" -n macs2_01 -f BAMPE -g hs --keep-dup all -q 0.01 --outdir ${cartella
0}"macs_narrow_01"

#BroadPeaks

singularity exec --bind /data:/data ${Macs2} macs2 callpeak -t ${cartella4}"$3_dedu
plicates.bam" -n macs2_01 -f BAMPE -g hs --keep-dup all --broad -q 0.01 --outdir ${
cartella0}"macs_broad_01"
```

Codice 9: Esempio di linea di comando per Macs2.

Usando *singularity*, viene eseguito *macs2* col comando *callpeak* per l'identificazione di picchi su dati di sequenziamento contenuti nel file BAM, dove l'allineamento è già stato effettuato.

Ogni comando ha una configurazione leggermente diversa in termini di tipo di picco (*narrow* o *broad*) e soglia di significatività (q-value), il che permette di ottenere risultati variabili per analisi comparative.

Opzioni utilizzate:

- -t: specifica il file BAM di input.
- -n: specifica il nome del file di output.
- -f BAMPE: indica un file di input con allineamento paired-end.
- -g: specifica il genoma di riferimento, hs sta per Homo sapiens (genoma umano).
- --keep-dup all: conserva tutti i duplicati.
- --broad: per picchi Broad, indentifica picchi ampi per legami di proteine su regioni più estese.
- -q: imposta una soglia di q-value per la chiamata dei picchi. Si utilizzano diversi valori per evidenziare come ognuno influisce sulla scoperta dei picchi. Un q-value di 0.01 può essere scelto per massimizzare la specificità, riducendo i falsi positivi. Un q-value di 0.05 per chiamate di picchi meno stringenti. Un q-value di 0.00001 può aiutare a catturare picchi di bassa abbondanza, aumentando il numero totale di picchi identificati.

  I q-value sono i p-value aggiustati che si ottengono utilizzando un approccio ottimizzato per il False Discovery Rate (FDR). Mentre un p-value misura la probabilità di ottenere risultati significativi per caso, il q-value tiene conto del numero totale di test condotti e della distribuzione dei valori p, permettendo di stimare la proporzione di falsi positivi tra i risultati significativi. Quindi, i q-value offrono un modo per interpretare i risultati in modo più affidabile, migliorando l'accuratezza delle conclusioni tratte da analisi di dati complessi.
- --outdir: definisce la directory per i file di output. L'output di base è automatico e comprende diversi file che forniscono informazioni sui picchi identificati e sull'analisi eseguita. Il file BED (Browser Extensible Data) è costituito da un insieme di righe e ognuna di esse rappresenta un intervallo genomico. Le informazioni di maggiore interesse si trovano nelle prime tre colonne del file e sono il cromosoma su cui si trova l'intervallo, la posizione iniziale (start) e la posizione finale (end) dell'intervallo sul cromosoma. Il file Excel riporta i picchi identificati insieme alle loro coordinate e alle statistiche associate. Il file file NarrowPeak o BroadPeak è simile al BED file, ma contiene solo i picchi ritenuti, rispettivamente, 'stretti' o 'larghi' dal processo di chiamata dei picchi

  I file NarrowPeak permettono di ottener una mappatura più precisa delle regioni di accessibilità della cromatina, ciò può essere utile per identificare specifiche regioni

regolatorie e comprendere meglio la struttura della cromatina in specifiche regioni del genoma. Mentre i file BroadPeak identificano regioni di accessibilità della cromatina più estese e diffuse e sono più adatti per comprendere la struttura e la dinamica della cromatina su larga scala.

2.1.8 Bedtools

Bedtools è uno strumento che consente di combinare, filtrare e manipolare file che contengono informazioni su posizioni genomiche per creare un set di picchi di consenso attraverso tutti i campioni e un file tabulare per facilitare il filtraggio dei dati. Questo è utile per analizzare la quantità di reads che coprono specifiche regioni del genoma in un esperimento di sequenziamento.

```
singularity exec --bind /data:/data ${Bedtools} bedtools bamtobed -i ${cartella4}"$
3_deduplicates.bam" > ${cartella5}"$3.bed"
```

Codice 10: Esempio di linea di comando per bedtools bamtobed.

Il comando *bamtobed* crea un file BED contenente le informazioni di allineamento dalle reads nel file BAM. Il formato BED è facile da interpretare e può essere utilizzato nei passaggi successivi dell'analisi.

Opzioni utilizzate:

- -i: specifica il file di input BAM
- >: reindirizza l'output in un file BED

```
singularity exec --bind /data:/data ${Bedtools} bedtools genomecov -bg -g ${home}"C
hr_length.txt" -i ${cartella5}"$3.bed" > ${cartella5}"$3.bedgraph"
```

Codice 11: Esempio di linea di comando per bedtools genomecov con input BED.

Il comando *genomecov* genera un file bedGraph che mostra la copertura delle reads lungo il genoma, evidenziando le diverse regioni genomiche e i loro livelli di copertura. Questo file è utile per visualizzare l'accessibilità o l'espressione delle regioni genomiche nel dataset di sequenziamento.

Opzioni utilizzate:

- -bg: imposta l'output in formato bedGraph, che mostra i livelli di copertura lungo il genoma.
- -g: specifica un file TXT di lunghezza dei cromosomi (Chr_length.txt) necessario per determinare l'intervallo genomico completo.
- -i: usa il file BED generato nel primo comando come input per calcolare la copertura.
- >: salva l'output in un file bedGraph.

Per creare il file TXT con la dimensione dei cromosomi da utilizzare in *bedtools genomecov*:

- sulla piattaforma NCBI, inserire il codice del genoma di riferimento, lo stesso utilizzato in Bowtie2;
- selezionare il genoma e andare nella sezione Chromosome;
- copiare in un file TXT i valori RefSeq (o GenBank) e Size (bp) con una tabulazione a separarli.

In alternativa, utilizzando l'opzione -ibam è possibile dare in input un file BAM e in questo caso non è necessario specificare un file di lunghezza dei cromosomi.

```
singularity exec --bind /data:/data ${Bedtools} bedtools genomecov -bg -ibam ${cart
ella4}"$3_deduplicates.bam" > ${cartella6}"$3_ibam.bedgraph"
```

Codice 12: Esempio di linea di comando per bedtools genomecov con input BAM.

Opzioni utilizzate:

- -bg: indica che il formato di output deve essere bedGraph.
- -ibam: specifica che il file di input è in formato BAM.
- >: reindirizza l'output del comando a un file.

2.1.9 IGV (Integrative Genomics Viewer)

IGV è uno strumento interattivo progettato per la visualizzazione e l'analisi di dati genomici. I dati vengono visualizzati come tracciati (tracks) che consentono di esplorare visivamente sequenze genomiche e allineamenti permettendo così di individuare differenze di espressione, mutazioni e varianti strutturali in geni specifici o interi cromosomi.

Una volta caricati i file NarrowPeak e BroadPeak ottenuti dall'analisi di ATAC-seq su IGV, è possibile ispezione i risultati della chiamata dei picchi. Questi file contengono informazioni sulle regioni di picco identificate nell'analisi di ATAC-seq. Le regioni sono indicate dalle coordinate genomiche (asse x), dall'altezza del picco (asse y) e dalla larghezza del picco. Per valutare l'affidabilità e l'importanza dei picchi identificati possono essere utilizzati i valori di punteggio del picco e la significatività statistica del picco, già contenuti nel file.

2.2 ANALISI BIOINFORMATICA DEI DATI CUT&TAG (CLEAVAGE UNDER TARGETS AND TAGMENTATION) PER VIMENTINA

CUT&Tag è una tecnica di sequenziamento che utilizza anticorpi per individuare proteine specifiche legate alla cromatina. Ciò la rende particolarmente utile per studiare la distribuzione genomica di proteine, come vimentina, o per osservare modificazioni epigenetiche (36).
Il processo inizia con il legame di un anticorpo specifico per la proteina di interesse associata alla cromatina. Successivamente, un secondo anticorpo lega il primo per reclutare l'enzima per il taglio, ottenuto dall'unione della trasposasi Tn5 e una proteina A/G (A/G-Tn5). L'A/G-Tn5, provvista di adattatori, frammenta il DNA nei siti legati dagli anticorpi producendo brevi sequenze di DNA per il sequenziamento (36).

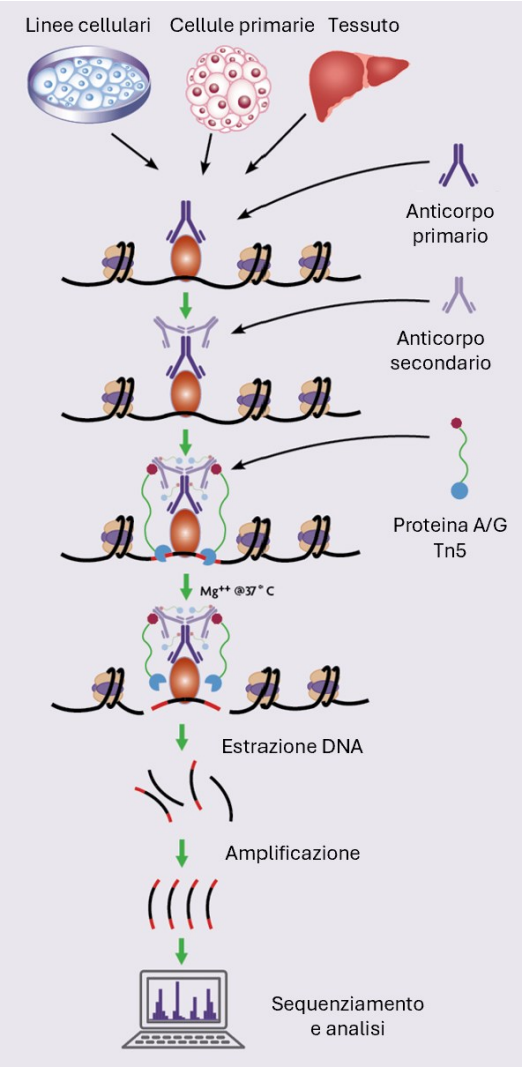

Figura 20: Schema della tecnica CUT&Tag: si aggiunge un primo anticorpo specifico per la proteina, si lega un secondo anticorpo dotato di affinità per il primo che facilita il legame del complesso pA-Tn5. L'attivazione della Tn5 mediante l'aggiunta di ioni magnesio consente l'integrazione degli adattatori nel DNA vicino. I frammenti di DNA così ottenuti vengono poi amplificati tramite PCR per creare librerie destinate al sequenziamento.

Dal punto di vista dell'analisi bioinformatica, CUT&Tag segue passaggi molto simili a quelli di ATAC-seq, tra cui la pulizia dei dati, l'allineamento e l'identificazione dei picchi (37). Tuttavia, poiché CUT&Tag non implica la stessa modalità di accessibilità della cromatina di ATAC-seq, non è necessario applicare la correzione dell'offset Tn5 che precede la chiamata dei picchi. Questa correzione, tipica di ATAC-seq, viene utilizzata per compensare la leggera distorsione di posizione dovuta al taglio non diretto della transposasi, mentre in CUT&Tag, grazie alla precisione del taglio specifico vicino alla proteina bersaglio, questo passaggio non è richiesto.

2.3 ANALISI AVANZATA: RICERCA DI G-QUADRUPLEX

2.3.1 Download dei file FASTA dei cromosomi

I file FASTA sono un formato di file comunemente utilizzato per memorizzare sequenze biologiche. Ogni file inizia con una riga di intestazione caratterizzata da un simbolo '>' e contiene informazioni descrittive sulla sequenza, come il nome della sequenza o altre annotazioni. Mentre le righe successive contengono la sequenza, che in questo caso è di un intero cromosoma.
Per scaricare i file fasta di tutti i cromosomi è sufficiente andare sulla piattaforma NCBI, nella pagina riguardante il genoma di riferimento utilizzato in precedenza, selezionare la tendina '*send to*', scegliere le opzioni '*complete record'* e '*file'* e assegnare il formato FASTA.

2.3.2 QParse

QPARSE è uno strumento bioinformatico utilizzato principalmente per identificare e analizzare i motivi G-quadruplex nelle sequenze genomiche. QPARSE consente di modellare tutti i possibili pattern considerando tutte le combinazioni di isole, cioè i raggruppamenti di guanine, che vengono rilevati in base ai parametri specificati.
L'output tipico di QPARSE è organizzato in blocchi, ognuno associato a una sequenza fornita come input. I quadruplex individuati includono i dettagli principali in un formato tabellare: ogni riga contiene la sequenza, il punteggio, gli indici di inizio e fine e la lunghezza delle isole. Il punteggio assegnato al quadruplex è basato sulla stabilità e sulla probabilità di formazione, considerando la lunghezza delle isole di guanina e dei loop.

```
$QPARSE -i ${home}"fasta/sequence_chr16.fasta" -o ${home}"risultati/patternG_chr16.
txt" -p 8 -n 8 -b G
```

Codice 13: Esempio di linea di comando per QPARSE.

Opzioni utilizzate:

- -i: specifica il file di input in formato FASTA.
- -o: specifica il file di output dove verranno salvati i risultati.

- -p: indica il numero minimo di isole perfette.
- -n: indica il numero di isole consecutive.
- -b: specifica la base di interesse.

QPARSE_parser è un programma associato a QPARSE, progettato per elaborare i risultati ottenuti da quest'ultimo. Attraverso il suo output elaborato, QPARSE_parser facilita la comprensione delle strutture identificati da QPARSE e delle loro caratteristiche, rendendo più semplice il lavoro di ricerca.

```
$QPARSER -i ${home}"risultati/patternG_chr16.txt" -o ${home}"risultati/patternG_chr16_par
ser.txt" -x
```

Codice 14: Esempio di linea di comando per QPARSER.

Opzioni utilizzate:

- -i: specifica il file di input, che è quello generato da QPARSE.
- -o: specifica il file di output per i risultati del parser.
- -x: restituisce solo il numero massimo di potenziali quadruplex non sovrapposti.

2.3.3 Pandas

Per i prossimi passaggi dell'analisi è necessario utilizzare Pandas, una potente libreria Python progettata per la manipolazione e l'analisi dei dati. Pandas offre strumenti versatili per gestire dataset di grandi dimensioni e per facilitare operazioni di filtraggio, raggruppamento e aggregazione, semplificando l'analisi dei dati generati nelle fasi precedenti della pipeline.
Prima di tutto si ricavano le aree in cui è presente una corrispondenza tra ATAC e vimentina. Le sovrapposizioni si ottengono dai dati relativi alle regioni ATAC e alle regioni di vimentina contenuti nei file BED ottenuti dalla pipeline.

```
singularity exec --bind /data:/data ${Bedtools} bedtools multiinter \
-header \
-i ${DATA}"ATAC_regioni_fuse_finali.bed" ${DATA}"VIM_regioni_fuse_finali.bed" \
-names ATAC VIM \
> ${CON}"intersect_atac_vim_con_tc.bed"
```

Codice 15: Esempio di linea di comando per bedtools multiintersect.

Il comando *multiinter* di *bedtools* identifica le regioni sovrapposte tra più file BED.

Opzioni utilizzate:

- -header: aggiunge una riga di intestazione all'output, utile per sapere a quale file e condizione si riferiscono le colonne riportate.
- -i: specifica i file di input contenenti le regioni genomiche di interesse.
- -names: assegna nomi ai file di input, facilitando l'identificazione dei file corrispondenti nelle colonne di output.
- >: reindirizza l'output.

Il file di output contiene le regioni genomiche sovrapposte tra i due file, indicando in quali file specifici è presente l'intersezione. Questo output viene utilizzato per creare tre file: uno per le intersezioni con solo regioni ATAC (atac), uno per le intersezioni con solo regioni di vimentina (vim) e uno con le regioni di sovrapposizione (atac-vim). Le informazioni salvate riguardano le coordinate di inizio (start) e fine (end) della regione e il cromosoma di appartenenza.
I dati ottenuti in precedenza con QPARSE sono molto ridondanti e risultano utili solo per una ricerca iniziale dei G-quadruplex, ma non sono adatti a portare avanti una ricerca di sequenze per regione e con dimensioni di isole e loop specifiche. Perciò, per ogni cromosoma, si esegue la ricerca di potenziali G-quadruplex considerando gli start ed end di ciascuna regione e scartando le sequenze che non hanno le seguenti caratteristiche: almeno otto isole consecutive intervallate da loop di lunghezza compresa tra 1 e 7 (*trenoG_argparse_nuovo.py**).
La sequenza del cromosoma ottenuta da NCBI è relativa al filamento a polarità positiva della catena di DNA. Per poter ripetere la ricerca dei G4 sul filamento inverso viene creata una funzione che, data in input la sequenza, restituisca in output la sequenza inversa complementare (*funz_conversione.py**).
Alla fine del processo, utilizzando il comando *cat*, si riuniscono i dati in tre file, uno per regione, contenenti le coordinate e le sequenze dei potenziali G4 per ciascun filamento di ogni cromosoma.
In atac, vim e atac-vim sono state incluse le sequenze relative ai telomeri e ai centromeri. Per osservare un risultato che non risulti alterato dalle sequenze altamente ripetute presenti in queste regioni, si esegue nuovamente *bedtools multiinter*: le coordinate dei telomeri e dei centromeri vengono aggiunte in input e poi le regioni di ATAC, di vimentina e le sovrapposizioni tra esse vengono nuovamente selezionate. Infine, si ripete la ricerca dei G4.
Per poter osservare come la presenza di sequenze telomeriche e centromeriche influenzi il risultato, si ricavano le lunghezze dei loop sia per i dati che includono i telomeri e centromeri

sia per i dati senza. Queste informazioni vengono utilizzate per creare dei grafici di distribuzione delle lunghezze tramite la libreria *matplotlib* di Python.

Il passo successivo consiste nel filtrare i G4 in modo da ottenere sequenze che abbiano isole 'regolari' (*lunghezze_isole_regolari.py**). Si tratta di un passaggio fondamentale per isolare sequenze che rappresentano potenziali siti di interazione funzionale con vimentina. Poiché vimentina mostra una preferenza per strutture G4 specifiche, la definizione di un set filtrato e regolare consente di concentrare l'analisi su regioni genomiche che hanno maggiore probabilità di essere biologicamente rilevanti, minimizzando il rumore e ottimizzando l'interpretazione dei dati. Questa selezione viene fatta utilizzando una *sliding window* che si sposta di un'isola alla volta e applica una condizione di lunghezza delle isole per ogni coppia di G4 che trova. In questo modo si considerano tutte le ripetizioni di due G4 che potrebbero formarsi. La condizione di lunghezza viene applicata ai G4 singoli e prevede che le isole centrali debbano avere una lunghezza di due o tre guanine e che le isole laterali siano di lunghezza uguale o aumentata di uno rispetto a quella delle isole centrali.

Di nuovo, si ricavano le lunghezze dei loop per ottenere un grafico di distribuzione delle sequenze dei G4 regolari, sia delle regioni con telomeri e centromeri sia di quelle senza.

Un'ulteriore ricerca, fatta per verificare la presenza di corrispondenze tra i possibili G4 presenti nelle regioni ATAC, nelle regioni di vimentina e in quelle di sovrapposizione, è di cercare caratteristiche comuni nei loop in termini di sequenza e lunghezza (*loop_dipendenti.py**). Considerando i G4 singoli, si osserva la distribuzione dei loop in gruppi da tre, indagando la presenza, tra i vari possibili G4, di ripetizioni di lunghezze e sequenze quando i loop vengono considerati come gruppo ordinato.

*https://github.com/valeceri/Pipeline-ATAC-VIM-G-quadruplex-

CAPITOLO 3: RISULTATI

3.1 CONFRONTO DEI REPORT FASTQC PRIMA E DOPO IL TRIMMING DELLE SEQUENZE

L'analisi di qualità delle sequenze è un passaggio essenziale nel preprocessing dei dati di sequenziamento. Di seguito viene presentato il report di FastQC per le sequenze prima e dopo il trimming, permettendo un confronto diretto degli indicatori di qualità migliorati grazie a questo processo. L'analisi si concentra su due aspetti principali: la qualità della sequenza per base (Figura 20) e il contenuto degli adattatori (Figura 21). Questi parametri sono stati scelti in quanto mostrano differenze significative dopo la rimozione delle sequenze di bassa qualità, mentre le altre sezioni del report non presentano variazioni rilevanti, dato il trimming limitato applicato ai dati. Questo confronto evidenzia come il processo di trimming migliori specifici parametri di qualità, garantendo dati più affidabili per le successive analisi bioinformatiche.
I report di FastQC completi sono presenti su github (https://github.com/valeceri/Pipeline-ATAC-VIM-G-quadruplex-).

In Figura 21, nei grafici FastQC si osserva un miglioramento della qualità complessiva delle basi nelle sequenze dopo il trimming, specialmente nelle ultime posizioni, dove inizialmente era presente un calo di qualità. Questo miglioramento è dovuto alla rimozione delle sezioni di bassa qualità nelle estremità delle letture.

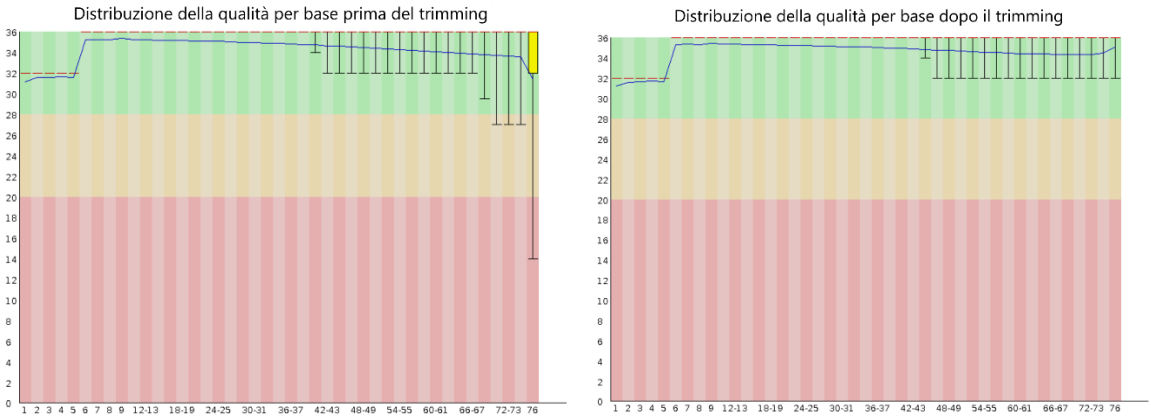

Figura 21: Confronto della distribuzione della qualità per base nelle sequenze prima e dopo trimming. L'asse delle ascisse rappresenta la posizione della base lungo la lunghezza delle reads di sequenziamento. L'asse delle ordinate rappresenta il punteggio di qualità per ciascuna posizione della base.

In Figura 22, nel grafico che rappresenta il contenuto di adattatori prima del trimming, è evidente un accumulo di sequenze adattatrici nelle ultime posizioni, che indica la necessità di un processo di pulizia. Mentre nel grafico che mostra il contenuto degli adattatori dopo il trimming si osserva un'assenza di picchi, indicando che gli adattatori sono stati efficacemente rimossi. Questo confronto evidenzia l'efficacia del trimming nel rimuovere gli adattatori indesiderati, migliorando la qualità complessiva delle sequenze.

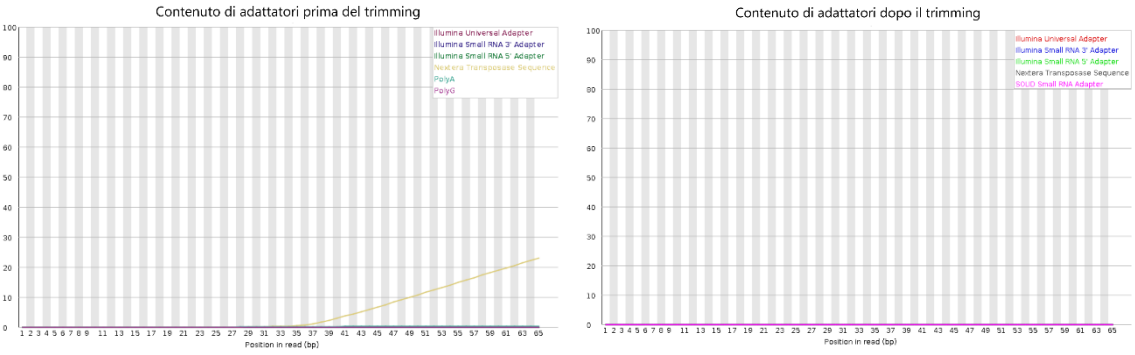

Figura 22: Confronto del contenuto di adattatori nelle sequenze prima e dopo il trimming. L'asse delle ascisse rappresenta le posizioni all'interno della sequenza. L'asse delle ordinate indica la quantità di nucleotidi che corrispondono a sequenze di adattatori in quella posizione specifica della sequenza.

3.2 IDENTIFICAZIONE DELLE SOVRAPPOSIZIONI GENOMICHE TRA ATAC E VIMENTINA

Le posizioni relative alle regioni accessibili della cromatina e le regioni di vimentina sono contenute nei file BED ottenuti rispettivamente tramite le tecniche ATAC-seq e CUT&Tag. Oltre ad essere particolarmente utilizzato per visualizzare dati su browser genomici (come IGV), questo formato è anche compatto e versatile, adatto per descrivere regioni specifiche del genoma, come i picchi di ATAC-seq e CUT&Tag.

Nel contesto di questa analisi, l'obiettivo principale è individuare le regioni di sovrapposizione tra i dati di accessibilità cromatinica e le aree di legame della proteina. Queste intersezioni rappresentano i punti di contatto in cui, tramite i G-quadruplex, vimentina potrebbe influenzare l'espressione genica e le strutture cromatiniche locali.

Per identificare con precisione queste sovrapposizioni, è stata utilizzata la funzione *multiinter* di Bedtools. Questo strumento consente di identificare le regioni genomiche condivise tra più dataset, rendendolo particolarmente utile per individuare sovrapposizioni tra diverse tipologie di sequenze regolatorie o di legame proteico.

Inizialmente, viene effettuata un'intersezione tra i dati di accessibilità della cromatina e di vimentina per delineare i siti condivisi. Successivamente, l'analisi è stata approfondita aggiungendo le informazioni relative ai telomeri e ai centromeri, con l'intento di escludere queste regioni dalla sovrapposizione finale. Questo approccio permette di selezionare in modo preciso le aree di sovrapposizione tra ATAC e vimentina, distinguendo le regioni ATAC, le regioni di vimentina, e le regioni ATAC-vimentina che siano prive di telomeri e centromeri. Questa separazione delle intersezioni permette di avere una visione più pulita e mirata delle aree genomiche di reale interesse, poiché l'assenza di telomeri e centromeri minimizza l'interferenza di strutture cromosomiche particolari che potrebbero alterare i risultati, consentendo un'analisi più specifica sulle funzioni potenziali di vimentina nelle regioni accessibili del genoma.

Il file generato con Bedtools riporta tutte le regioni comuni e non comuni tra i file BED specificati, consentendo un'analisi dettagliata delle diverse combinazioni di intersezione (Figura 23). Il conteggio delle sovrapposizioni facilita la selezione delle aree di interesse, permettendo di isolare le aree ATAC, vimentina e ATAC-vimentina che non presentano sovrapposizioni con telomeri o centromeri.

chrom	start	end	num	list	ATAC	VIM	centromeres	telomeres
chr1	1	820	1	telomeres	0	0	0	1
chr1	820	1158	2	VIM,telomeres	0	1	0	1
chr1	1158	3449	1	telomeres	0	0	0	1
chr1	3449	4042	2	ATAC,telomeres	1	0	0	1
chr1	4042	10000	1	telomeres	0	0	0	1
chr1	12689	12876	1	ATAC	1	0	0	0
chr1	18581	18703	1	ATAC	1	0	0	0
chr1	22615	23351	1	VIM	0	1	0	0
chr1	32062	32621	1	ATAC	1	0	0	0
chr1	36314	37193	1	ATAC	1	0	0	0
chr1	47356	47713	1	VIM	0	1	0	0
chr1	57997	58239	1	ATAC	1	0	0	0
chr1	61725	62014	1	VIM	0	1	0	0
chr1	89315	95207	1	VIM	0	1	0	0
chr1	102015	102665	1	VIM	0	1	0	0
chr1	125279	125603	1	VIM	0	1	0	0
chr1	141631	141729	1	ATAC	1	0	0	0
chr1	148900	149047	1	ATAC	1	0	0	0
chr1	153871	154168	1	ATAC	1	0	0	0
chr1	155549	155893	1	ATAC	1	0	0	0
chr1	159956	160279	1	ATAC	1	0	0	0
chr1	163369	163830	1	ATAC	1	0	0	0
chr1	166780	167087	1	VIM	0	1	0	0
chr1	167087	167168	2	ATAC,VIM	1	1	0	0
chr1	167168	167208	1	ATAC	1	0	0	0
chr1	177986	178129	1	ATAC	1	0	0	0
chr1	198586	199700	1	VIM	0	1	0	0

Figura 23: Esempio di file BED di output generato da multiinter di Bedtools, che mostra le regioni di intersezione rilevate tra dataset di ATAC e vimentina. Questo formato evidenzia le coordinate genomiche e i conteggi delle sovrapposizioni tra diverse regioni.

3.3 IDENTIFICAZIONE DEI G-QUADRUPLEX E ANALISI DELLE LUNGHEZZE DEI LOOP

Per la ricerca dei G-quadruplex, sono stati selezionati entrambi i gruppi di regioni genomiche: quello che include telomeri e centromeri e quello da cui queste regioni sono state escluse.
La ricerca si è concentrata sull'individuazione di strutture G-quadruplex all'interno di entrambe le categorie, considerando il filamento diretto e il filamento inverso complementare per garantire una copertura completa dell'intero genoma. L'analisi separata tra regioni con e senza telomeri e centromeri è utile per evidenziare eventuali differenze nella distribuzione di G-quadruplex tra queste aree genomiche.
I G-quadruplex da identificare sono stati definiti come sequenze contenenti almeno 8 isole, intervallate da loop di lunghezza variabile da 1 a 7 nucleotidi. La scelta di considerare sequenze con più di 8 isole è motivata dalla necessità di garantire una ricerca esaustiva di tutti i possibili G-quadruplex. Poiché la posizione di formazione del G4 dipende dalla distribuzione delle isole di guanine, è fondamentale non limitarsi al primo quadruplex doppio individuato. L'uso della *sliding window* consente di catturare isole e loop successivi lungo la sequenza, evitando il bias che potrebbe derivare dal considerare solo una porzione iniziale della sequenza stessa e garantendo che tutte le possibili strutture G4 vengano identificate.

Le Figure 24, 25 e 26 mostrano degli esempi di file risultanti dall'analisi dei G-quadruplex trovati nelle regioni genomiche selezionate. Questi file forniscono una visione approfondita della distribuzione e della struttura dei G-quadruplex in sequenze genomiche sia con sia senza telomeri e centromeri.

chr	start	end	quadruplex
chr1	36338	36432	GGCCACTAGGGAACCTCAGGAGAGGCCGCCGCGGTGGGGTGGGCGG
chr1	334780	334846	GGAGGAGGCGGCGGCCACCCGGGCCTGGCGGTGGAGGGGG
chr1	346388	346604	GGGATGGAGGCGGGAGAGCCAGGAGGCCTGAGGCAACCCTGGGACGGTTCCTGGATCGAGGAGAGCAGGGGGGTGATGAGGG
chr1	353402	353473	GGCCGACTCGGGGCATCCGGCGGCGGCAGCGGGACAGGGCTGCGGCCCCCGG
chr1	357622	357676	GGAAGGAGGCCAGGGCTGGCTGGGAATGGAGCCCAGG
chr1	369032	369071	GGCCGGCTCCGCGGCGGGGCGGGGGGGGGGCGGCGGCGCGG
chr1	388089	388155	GGCTGGGCGTGGCGCGGCGGGGCGGGGAAAGGGGCGGGGCCTGGGCGGCGG
chr1	395445	395560	GGGGACAAGGCAAGGGGTCTCGTGGCTTGGCCCCGGAGCGGGTGGGGG
chr1	449660	449712	GGGGCTCCGGGCGGCGCGGAGCCGGGCGGCAGGGCTCGAAGGCCGCCGGGAGGGG
chr1	449731	449839	GGGGCCGGACACCGAGGCTCGGCCGGGCCGGACCCCCCGGGAGCTGCGGCGCGGAGGCGGGTCCCGGTGCCGGGGCGGG
chr1	455040	455103	GGCCTAGGCAGCTGGCCAGGCAGGGAGGAAGGGAGGG

Figura 24: Esempio di file che riporta le coordinate genomiche e la sequenza completa del G-quadruplex identificato.

chr	start	end	pos_isola	num_isole	len_isole	isole
chr1	36338	36432	3	8	2-3-2-2-2-4-3-2	GG-GGG-GG-GG-GG-GGGG-GGG-GG
chr1	334780	334846	22	10	2-2-2-2-2-3-2-2-2-5	GG-GG-GG-GG-GG-GGG-GG-GG-GG-GGGGG
chr1	346388	346604	24	13	3-2-2-3-2-2-2-3-2-2-2-6-3	GGG-GG-GG-GGG-GG-GG-GG-GGG-GG-GG-GG-GGGGGG-GGG
chr1	353402	353473	26	9	2-4-2-2-2-3-3-2-2	GG-GGGG-GG-GG-GG-GGG-GGG-GG-GG
chr1	357622	357676	29	8	2-2-2-3-2-3-2-2	GG-GG-GG-GGG-GG-GGG-GG-GG
chr1	369032	369071	33	8	2-2-2-4-10-2-2-2	GG-GG-GG-GGGG-GGGGGGGGGG-GG-GG-GG
chr1	388089	388155	38	11	2-3-2-2-4-4-4-4-3-2-2	GG-GGG-GG-GG-GGGG-GGGG-GGGG-GGGG-GGG-GG-GG
chr1	395445	395560	40	8	4-2-4-2-2-2-3-5	GGGG-GG-GGGG-GG-GG-GG-GGG-GGGGG
chr1	449660	449712	48	10	4-3-2-2-3-2-3-2-3-4	GGGG-GGG-GG-GG-GGG-GG-GGG-GG-GGG-GGGG
chr1	449731	449839	48	14	4-2-2-2-3-2-3-2-2-2-3-2-4-3	GGGG-GG-GG-GG-GGG-GG-GGG-GG-GG-GG-GGG-GG-GGGG-GGG
chr1	455040	455103	50	8	2-2-2-2-3-2-3-3	GG-GG-GG-GG-GGG-GG-GGG-GGG

Figura 25: Esempio di file contenente informazioni specifiche sulle isole di guanine, come il numero totale di isole e la lunghezza di ciascuna.

chr	start	end	pos_trenoG	num_loop	len_loop	loop
chr1	36338	36432	3	7	6-7-3-7-1-1-1	CCACTA-AACCTCA-AGA-CCGCCGC-T-T-C
chr1	334780	334846	22	9	1-1-1-1-6-3-1-1-1	A-A-C-C-CCACCC-CCT-C-T-A
chr1	346388	346604	24	12	2-1-1-7-1-5-7-2-5-5-6-6	AT-A-C-AGAGCCA-A-CCTGA-CAACCCT-AC-TTCCT-ATCGA-AGAGCA-TGATGA
chr1	353402	353473	26	8	7-5-1-1-4-3-4-5	CCGACTC-CATCC-C-C-CAGC-ACA-CTGC-CCCCC
chr1	357622	357676	29	7	2-1-3-2-2-3-6	AA-A-CCA-CT-CT-AAT-AGCCCA
chr1	369032	369071	33	7	2-6-1-1-1-1-3	CC-CTCCGC-C-C-C-C-CGC
chr1	388089	388155	38	10	2-3-3-1-1-3-1-3-1-1	CT-CGT-CGC-C-C-AAA-C-CCT-C-C
chr1	395445	395560	40	7	4-3-6-3-4-3-1	ACAA-CAA-TCTCGT-CTT-CCCC-AGC-T
chr1	449660	449712	48	9	4-1-3-4-1-2-6-5-1	CTCC-C-CGC-AGCC-C-CA-CTCGAA-CCGCC-A
chr1	449731	449839	48	13	2-7-3-2-2-7-6-3-1-1-4-4-1	CC-ACACCGA-CTC-CC-CC-ACCCCCC-AGCTGC-CGC-A-C-TCCC-TGCC-C
chr1	455040	455103	50	7	4-5-3-2-1-2-1	CCTA-CAGCT-CCA-CA-A-AA-A

Figura 26: Esempio di file che mostra dettagli sui loop, come il numero di loop e la loro lunghezza.

La ricerca è proseguita con un'analisi dettagliata delle caratteristiche strutturali di queste sequenze, focalizzandosi inizialmente sulle lunghezze dei loop. La ricerca è stata condotta in parallelo sui due gruppi distinti di regioni genomiche per capire se la presenza di telomeri e centromeri influisse sulle caratteristiche strutturali dei G-quadruplex, suggerendo eventuali variazioni nella loro funzione o stabilità a seconda del contesto genomico. Questa analisi comparativa mirava a individuare se esistessero lunghezze dei loop ricorrenti che potessero suggerire una disposizione strutturale favorita o un pattern preferenziale per i G-quadruplex.
Dall'analisi delle lunghezze dei loop dei G-quadruplex doppi nelle regioni prive di telomeri e centromeri la distribuzione delle lunghezze risulta uniforme con un lieve picco osservato alla lunghezza di 1 (Figura 27).

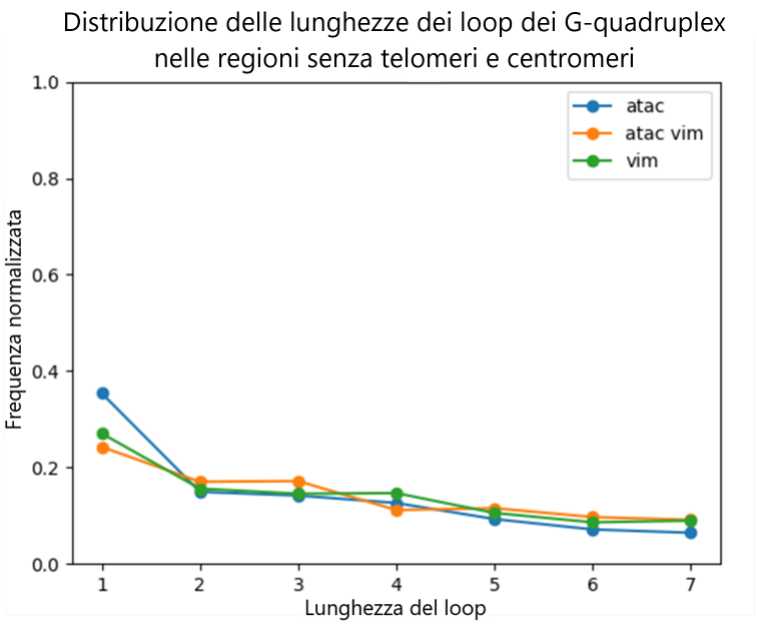

Figura 27: Grafico delle lunghezze dei loop dei G-quadruplex normalizzate nelle regioni prive di telomeri e centromeri. La curva orizzontale mostra una distribuzione uniforme delle lunghezze, indicando l'assenza di una frequenza predominante per specifiche lunghezze dei loop.

Diversamente, l'analisi delle lunghezze dei loop nei G-quadruplex doppi delle aree che includono telomeri e centromeri ha rivelato un picco evidente in vimentina in corrispondenza dei loop di lunghezza 3 (Figura 28). Questo risultato suggerisce una possibile preferenza strutturale per i loop di tale lunghezza in questi contesti specifici. Tuttavia, il fenomeno è probabilmente legato alla significativa presenza di sequenze ripetitive tipiche delle zone telomeriche e centromeriche, che potrebbero giustificare la formazione di loop di questa lunghezza. L'analisi delle triplette di basi, che ha permesso di identificare le combinazioni di basi più frequenti nei loop di lunghezza 3, ha confermato questa ipotesi, rivelando una predominanza di sequenze ripetitive che sono comuni nelle regioni telomeriche.

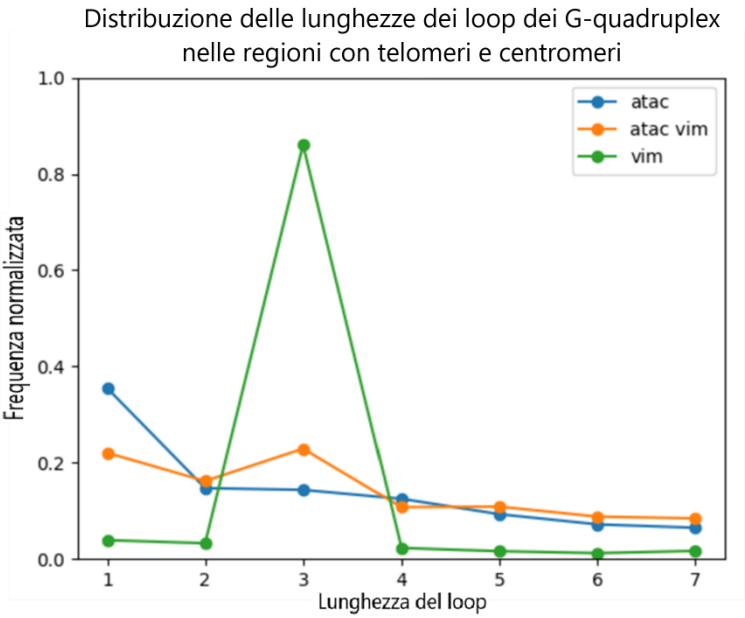

Figura 28: Grafico delle lunghezze dei loop dei G-quadruplex normalizzate nelle regioni con telomeri e centromeri. Il grafico evidenzia un picco a 3, suggerendo una predominanza di loop di lunghezza 3 nelle regioni che includono telomeri e centromeri.

3.4 SELEZIONE E ANALISI DEI G-QUADRUPLEX PIÙ ADATTI PER L'INTERAZIONE CON VIMENTINA

Fino ad ora sono state considerate tutte le sequenze possibili che rispettano i criteri di almeno 8 isole intervallate da loop di lunghezza compresa tra 1 e 7. Tuttavia, per raffinare ulteriormente l'analisi, si è deciso di applicare un filtro per isolare i G-quadruplex più regolari, ossia quelli con una maggiore probabilità di interagire con vimentina.

In particolare, il nuovo criterio selettivo per i G-quadruplex regolari prevede che le isole centrali abbiano una lunghezza di 2 o 3 basi, mentre le isole laterali devono essere della stessa lunghezza o di una base più lunghe rispetto alle isole centrali. Questo modello di coerenza nelle lunghezze delle isole dovrebbe favorire la formazione di strutture più stabili e quindi più suscettibili all'interazione con vimentina.

I risultati ottenuti nell'analisi delle regioni prive di telomeri e centromeri mostrano un piccolo picco alla lunghezza 5 nelle regioni ATAC-vimentina, oltre al picco a 1 già osservato in

precedenza (Figura 29). Nonostante ATAC e vimentina presentino entrambi un leggero picco a 4, in ATAC-vimentina non si osserva: mentre ATAC potrebbe rendere disponibile il sito per il legame, vimentina non sembra effettivamente legare in queste regioni.

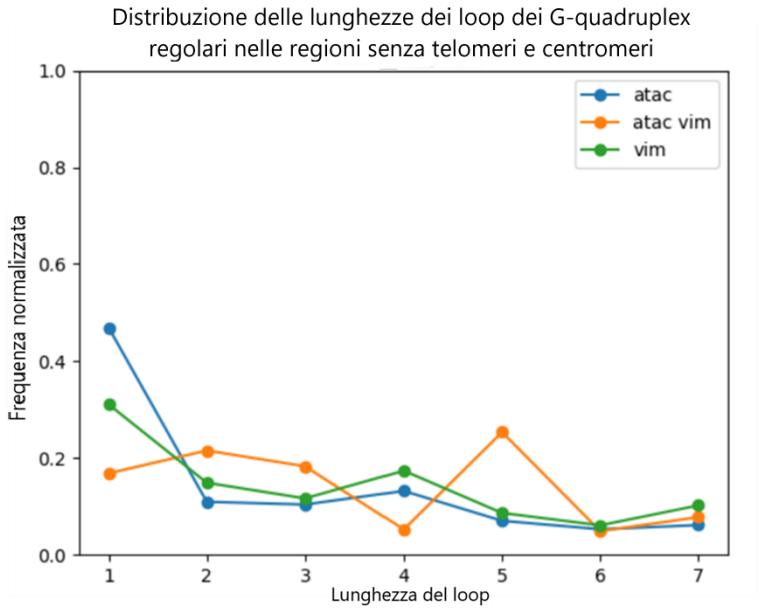

Figura 29: Grafico che mostra la distribuzione delle lunghezze dei loop nei G-quadruplex nelle regioni senza telomeri e centromeri. Si osserva il picco a 1, già stato riscontrato nelle analisi precedenti. Oltre a questo, il grafico mostra un piccolo picco a 5 e un minimo a 4 nelle regioni ATAC-vimentina.

Per le regioni con telomeri e centromeri, l'analisi ha evidenziato un marcato picco a una lunghezza di 3 nelle regioni di vimentina, che si riflette anche nelle regioni di intersezione ATAC-vimentina (Figura 30). Come per i G4 non regolari osservati in precedenza, questo picco significativo è correlato alla presenza di ripetizioni all'interno delle zone telomeriche e centromeriche. Similmente a quanto visto nelle regioni senza telomeri e centromeri, si osserva un piccolo picco alla lunghezza di 5, che potrebbe suggerire una dinamica di interazione differente tra ATAC e vimentina rispetto alla semplice presenza di telomeri e centromeri.

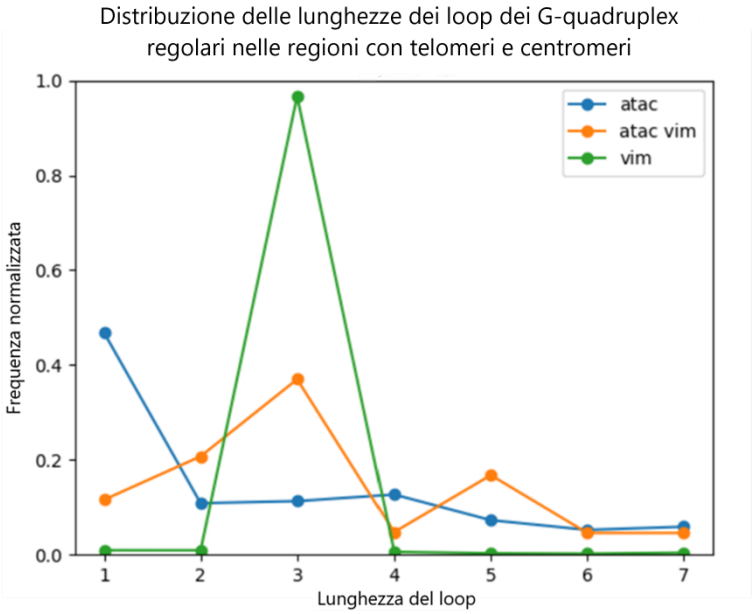

Figura 30: Grafico che mostra l'analisi delle lunghezze dei loop dei G-quadruplex nelle regioni con telomeri e centromeri, con evidenza di un marcato picco a lunghezza 3 nelle regioni legate a vimentina. Inoltre, si osserva un piccolo picco a lunghezza 5 nelle intersezioni ATAC-Vimentina.

3.5 ANALISI GENERALE DEI PATTERN NEI LOOP DEI G-QUADRUPLEX

Poiché dalle analisi precedenti non sono emerse caratteristiche predominanti nei loop dei G-quadruplex, è stato deciso di adottare un approccio più generale per identificare potenziali pattern. In questa fase, i loop sono stati esaminati come gruppi ordinati di ogni possibile struttura G4 individuale. Questo tipo di analisi mira a evidenziare se esistono pattern di ripetizione nelle lunghezze e nelle sequenze dei loop, permettendo di cogliere configurazioni stabili e ricorrenze strutturali comuni, che potrebbero essere meno visibili con un'analisi focalizzata su singoli parametri. Dall'analisi dei pattern di lunghezze dei loop è emersa una predominanza di specifiche configurazioni in base alla categoria di regione analizzata.

Nelle regioni senza telomeri e centromeri, il pattern di lunghezza più comune per i loop in ATAC e in vimentina è '1-1-1', mentre in ATAC-vimentina prevale la configurazione '5-5-5' (Figura 31).

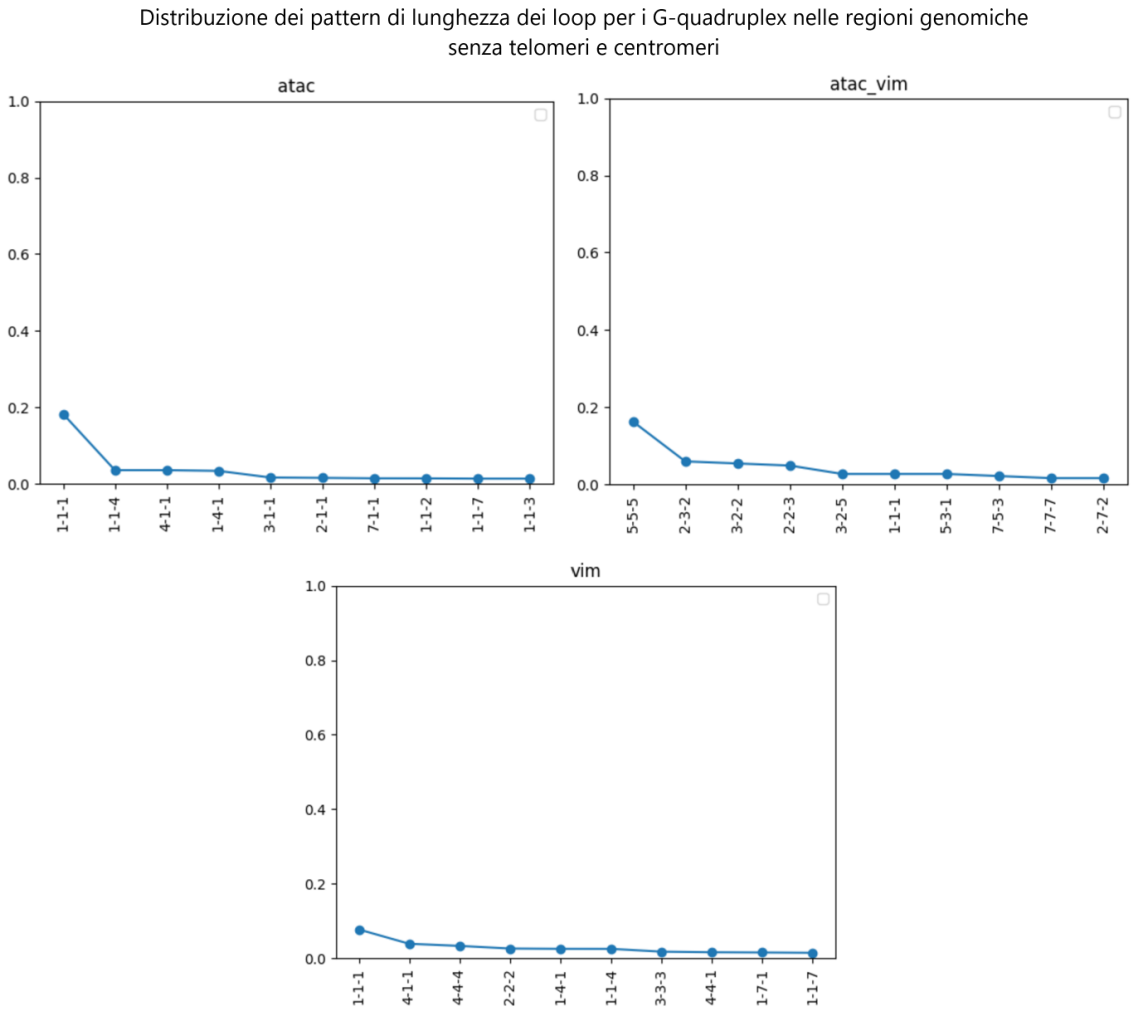

Figura 31: Distribuzione dei pattern di lunghezza dei loop per i G-quadruplex nelle regioni genomiche senza telomeri e centromeri. L'asse delle ascisse indica il pattern di lunghezza dei loop. L'asse delle ordinate indica la frequenza relativa. I grafici mostrano i dieci pattern più frequenti nelle categorie di sequenze ATAC, ATAC-vimentina e vimentina.

Il pattern di lunghezza '5-5-5' è stato oggetto di una verifica approfondita per individuare se le sequenze di questi loop mostrano somiglianze significative. L'analisi ha rivelato che i gruppi di loop non sono necessariamente una caratteristica distintiva per il legame tra ATAC e vimentina, come inizialmente ipotizzato. In realtà, si è scoperto che le sequenze '5-5-5' sono prevalentemente riconducibili a una sequenza molto ripetitiva, in cui tutti i loop hanno una lunghezza di 5 (Figura 32). Oltre alla lunghezza, anche la composizione nucleotidica è simile e le differenze riscontrate sono probabilmente il risultato di mutazioni puntiformi che si sono

verificate durante un processo di duplicazione. Inoltre, il metodo della *sliding window*, utilizzato per identificare i G-quadruplex regolari, ha contribuito a rinforzare questo fenomeno di ripetizione (Figura 33), fornendo anche una spiegazione per il piccolo picco osservato a 5 nei grafici di distribuzione delle lunghezze dei quadruplex regolari. Quando le ripetizioni '5-5-5' vengono eliminate, il pattern di lunghezza 5 non emerge più come una tendenza significativa. L'analisi dei pattern successivi più abbondanti ha portato alla conclusione che non è più utile proseguire con la ricerca di altri pattern di lunghezza tra i loop, poiché i risultati sembrano essere dominati dalla ripetitività di queste sequenze particolari.

Le Figure 32 e 33 mostrano i file rappresentativi della sequenza ripetitiva che provoca la prevalenza del pattern '5-5-5'. In queste due immagini è possibile osservare come la tecnica della *sliding window* abbia prodotto diverse configurazioni di G-quadruplex, accentuando le ripetizioni di lunghezza 5 come pattern dominante rispetto agli altri.

chr	start	end	pos_trenoG	num_loop	len_loop	loop
chr5	270432	270480	3257	7	4-2-7-6-4-2-4	CAGA-CA-CTGTGCC-CACTGT-CTGT-AT-TGCT
chr5	270634	270687	3257	8	3-4-2-6-7-4-2-4	CCA-CAGA-CA-CTGTGA-CACTGTA-CTGT-AT-AGCT
chr5	36713890		36713931	3447	7	1-2-6-2-3-6-4 A-CT-AAGCCT-TT-TGT-CAAAGT-CAGT
chr5	59755751		59755851	3493	15	1-1-1-4-2-1-1-1-3-1-1-3-1-1-1 T-C-C-AGCA-AA-A-C-C-AGC-C-C-AGC-C-C-C
chr5	81135871		81135919	3530	8	3-4-6-6-1-3-3-3 TTT-TTAA-TTGTTT-TAGTAA-T-AAT-TTT-CTA
chr5	110426943		110426987	3582	8	5-1-1-3-3-2-1-4 AGCAC-T-A-AAA-TGT-CT-A-CTGT
chr5	113947765		113947809	3593	7	5-5-4-6-2-2-6 CCTGA-CACTT-ACTT-AATCCT-AA-CT-CCAGTA
chr5	131538518		131538557	3629	7	1-3-2-7-1-5-4 T-TGA-AA-ACCATTA-T-CGACA-ACTA
chr5	133046174		133046213	3634	8	3-3-3-4-6-1-1-2 TGT-AGT-TCT-AGAA-AGTAGA-T-A-AT
chr5	150697939		150697973	3688	7	1-2-6-2-1-1-3 A-AT-ACGAAA-CA-A-A-CAT
chr5	168824961		168825008	3735	7	2-7-5-5-7-6-1 CT-TGTCCCT-ACTCT-TGCCT-TATTTAT-AATGAA-A
chr5	176590342		176590386	3770	7	5-1-7-5-1-6-5 TCCCA-T-CTGTGTA-TCCCA-T-ATATCA-CATCA
chr5	181215246		181215438	3787	24	5-5-5-5-5-5-5-5-5-5-5-5-5-5-5-5-5-5-5-5-5-5-5-5 ACAGA-ACAGC-ACAGA-ACAGA-A
chr5	181584547		181584603	3788	10	3-1-6-4-2-3-2-3-1-4 TCT-A-TTTATA-CACA-TT-CGT-CA-CCA-T-TCTT
chr5	181601402		181601441	3789	8	4-5-2-2-1-1-3-2 AAGT-ACGAA-TT-TT-T-T-TGA-CA

Figura 32: Estratto parziale del file che riporta la sequenza originale con la struttura del quadruplex.

chr	pos_isola	len_isole	sequenza_isole	len_loop	sequenza_loop
5	3634	2-2-2-2-2-2-2-2	GG-GG-GG-GG-GG-GG-GG-GG	3-3-3-4-6-1-1	TGT-AGT-TCT-AGAA-AGTAGA-T-A
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGA-ACAGC-ACAGA-ACAGA-ACAGC-ACAGA-ACAGC
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGC-ACAGA-ACAGA-ACAGC-ACAGA-ACAGC-ACAGA
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGA-ACAGA-ACAGC-ACAGA-ACAGC-ACAGA-ACAGA
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGA-ACAGC-ACAGA-ACAGA-ACAGA-ACAGA-ACAGA
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGC-ACAGA-ACAGC-ACAGA-ACAGA-ACAGA-ACAGA
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGA-ACAGC-ACAGA-ACAGA-ACAGA-ACAGA-ACAGC
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGC-ACAGA-ACAGA-ACAGA-ACAGA-ACAGC-ACAGA
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGA-ACAGA-ACAGA-ACAGA-ACAGC-ACAGA-ACAGC
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGA-ACAGA-ACAGA-ACAGC-ACAGA-ACAGC-ACAGA
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGA-ACAGC-ACAGC-ACAGA-ACAGA-ACAGC-ACAGA
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGA-ACAGC-ACAGA-ACAGC-ACAGA-ACAGA-ACAGC
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGC-ACAGA-ACAGC-ACAGA-ACAGA-ACAGC-ACAGA
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGA-ACAGC-ACAGA-ACAGA-ACAGC-ACAGA-ACAGA
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGC-ACAGA-ACAGC-ACAGA-ACAGA-ACAGC-ACAGA
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGA-ACAGA-ACAGC-ACAGA-ACAGA-ACAGC-ACAGA
5	3787	3-3-3-3-3-3-3-3	GGG-GGG-GGG-GGG-GGG-GGG-GGG-GGG	5-5-5-5-5-5-5	ACAGC-ACAGA-ACAGA-ACAGC-ACAGA-ACAGA-ACAGA

Figura 33: Estratto parziale del file che riporta le ripetizioni derivate dall'uso della sliding window.

Per quanto riguarda le regioni che includono telomeri e centromeri, il pattern più frequente in ATAC rimane ‘1-1-1’, mentre sia in ATAC-vimentina sia in vimentina si osserva una predominanza del pattern ‘3-3-3’ (Figura 34).

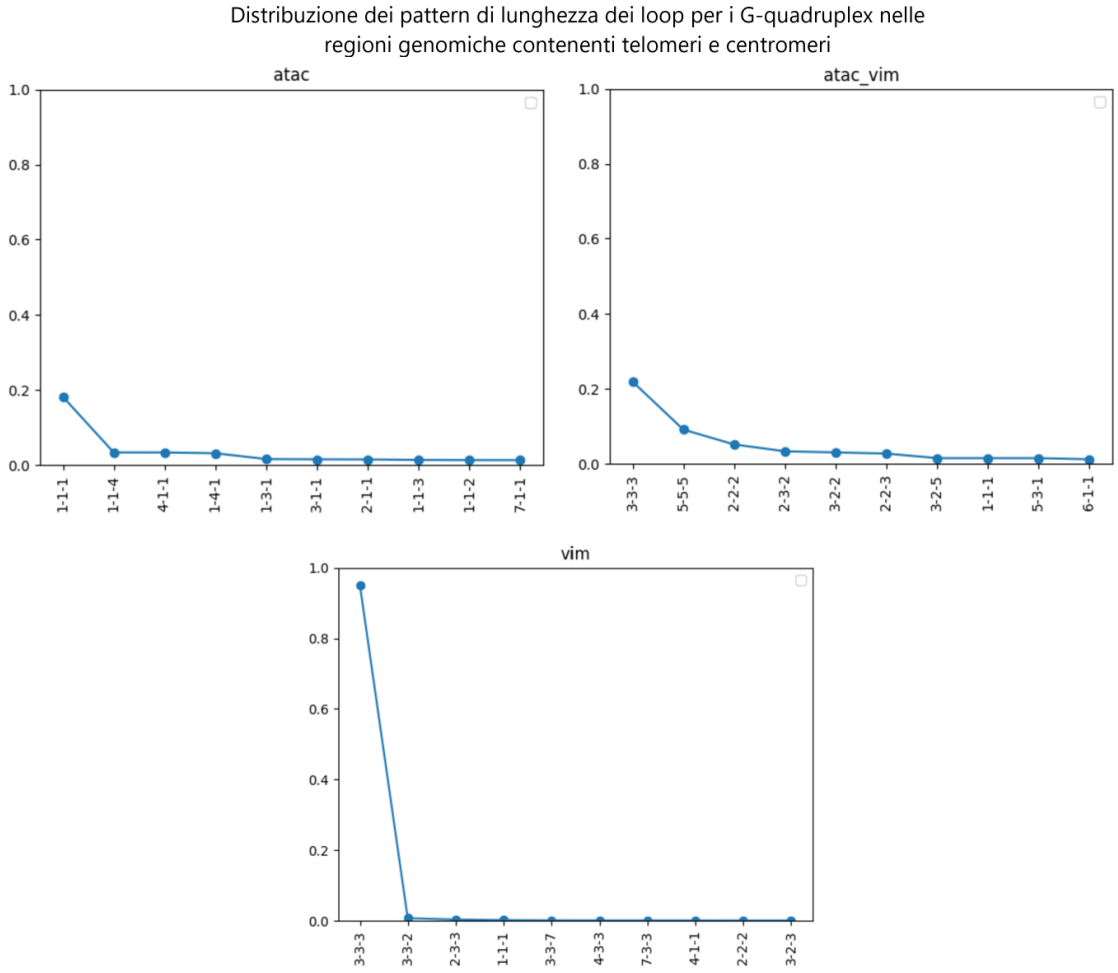

Figura 34: Distribuzione dei pattern di lunghezza dei loop per i G-quadruplex nelle regioni genomiche contenenti telomeri e centromeri. L’asse delle ascisse indica il pattern di lunghezza dei loop. L’asse delle ordinate indica la frequenza relativa. I grafici illustrano i dieci pattern più frequenti nelle categorie di sequenze ATAC, ATAC-vimentina e vimentina, consentendo di evidenziare le configurazioni ricorrenti in queste aree.

CAPITOLO 4: DISCUSSIONE

4.1 METODOLOGIE E APPROCCI NELL'ANALISI DEI G-QUADRUPLEX: ATAC-SEQ E CUT&TAG

Le scelte metodologiche compiute in questa ricerca riflettono la necessità di coniugare precisione e rappresentatività per produrre risultati significativi e comparabili con studi preesistenti, con l'obiettivo di approfondire le caratteristiche strutturali dei G-quadruplex nelle regioni genomiche studiate.

L'analisi delle regioni accessibili della cromatina tramite ATAC-seq e delle interazioni tra DNA e proteine con CUT&Tag si è dimostrata particolarmente efficace: ATAC-seq consente di valutare l'accessibilità della cromatina, una misura fondamentale per comprendere la distribuzione dei G-quadruplex nelle regioni di interesse, mentre CUT&Tag fornisce informazioni sulla localizzazione di vimentina.

Diversi studi recenti hanno valutato l'efficacia delle pipeline di ATAC-seq e CUT&Tag nell'analisi bioinformatica, mostrando come queste tecniche siano diventate un metodo standard consolidato per l'identificazione delle regioni accessibili della cromatina (40). Tale standardizzazione permette di seguire una pipeline semplice e funzionale, con passaggi condivisi che garantiscono risultati riproducibili.

Nonostante l'efficacia di ATAC-seq, vari fattori possono influire sulla rilevazione precisa dei picchi di accessibilità della cromatina. La qualità dei campioni e la gestione degli artefatti rappresentano sfide note che possono compromettere l'identificazione delle regioni aperte. Il preprocessing dei dati ATAC-seq è quindi essenziale per ottenere profili cromatinici accurati e privi di bias. Passaggi come il controllo qualità, la rimozione degli adattatori e il filtraggio dei duplicati sono fondamentali per migliorare la precisione della rilevazione dei picchi di accessibilità. In quest'ottica, la scelta di strumenti come Bowtie2, Samtools, Macs2 e Bedtools, supportata da una vasta comunità scientifica, risulta ottimale per garantire che i picchi identificati rappresentino realmente regioni biologicamente rilevanti.

Per quanto riguarda i G-quadruplex, è stato dimostrato che hanno un ruolo significativo nel controllo trascrizionale nelle regioni promotrici, soprattutto in contesti oncologici. Diversi studi sfruttano le tecniche CUT&Tag e ATAC-seq per mappare i G4 a livello genomico, fornendo

un quadro dettagliato sulla distribuzione e il ruolo dei G4 nel genoma umano, in particolare nelle regioni promotrici (41,42). I risultati dimostrano che i G4, abbondanti nei promotori genici, possono esercitare un controllo importante sull'espressione dei geni, stabilizzando o inibendo la formazione di strutture che influenzano la trascrizione, soprattutto nei processi di sviluppo e in contesti patologici come il cancro (41).

4.2 ANALISI DEI G-QUADRUPLEX NELLE REGIONI GENOMICHE

L'analisi sperimentale condotta in questa ricerca ha permesso di identificare G-quadruplex distribuiti nelle varie regioni genomiche. Nelle regioni prive di telomeri e centromeri, le lunghezze dei loop non hanno mostrato tendenze strutturali significative (Figura 27), suggerendo che in tali regioni non vi siano particolari configurazioni favorite per i G4. Sebbene sia stato osservato un lieve picco alla lunghezza di 1, tale valore non viene considerato rilevante per diversi motivi, legati principalmente alla stabilità e alla configurazione delle strutture G4. Poiché la stabilità di un G4 richiede una certa lunghezza e un'organizzazione specifica dei loop, loop troppo brevi possono risultare instabili e più difficili da associare a funzioni biologiche particolari, come l'interazione con proteine o il legame con altre strutture genomiche.
Al contrario, nelle regioni contenenti telomeri e centromeri è emerso un picco significativo per i loop di lunghezza 3 (Figura 28), che può essere interpretato come una manifestazione delle caratteristiche intrinseche di queste aree, dove sequenze ripetitive influenzano la formazione dei G-quadruplex.
La decisione di analizzare separatamente le regioni telomeriche e centromeriche si basa sull'importanza regolativa e strutturale attribuita a queste aree. Nei telomeri, notoriamente ricchi di sequenze che favoriscono la formazione di G-quadruplex, si osservano profili di stabilità specifici che richiedono un'analisi distinta per evitare che i loro effetti dominino i risultati delle aree genomiche più eterogenee. Studi precedenti hanno dimostrato come i G4 nei telomeri siano fondamentali per proteggere le estremità cromosomiche, inibendo l'attività della telomerasi e prevenendo l'instabilità genomica (42). I dati confermano queste osservazioni, rivelando un pattern ripetitivo nei telomeri coerente con la loro funzione stabilizzante.

Nei promotori genici, invece, i G4 sono coinvolti nell'attività trascrizionale, modulando l'accessibilità ai fattori regolatori (42). Qui, l'analisi ha mostrato una variabilità nella lunghezza dei loop, suggerendo una maggiore flessibilità compatibile con il ruolo dinamico delle regioni promotrici nella regolazione genica. Di conseguenza, anche se i G4 telomerici sono noti per la loro stabilità e funzione protettiva, questa ricerca si focalizza sui G4 nelle regioni promotrici e sulla loro interazione con vimentina, per comprendere come tali strutture possano influenzare la regolazione genica e contribuire a processi tumorigenici.

Un aspetto importante da notare è che il picco a loop di lunghezza 3 non implica necessariamente un legame diretto di vimentina con il DNA o con la formazione di G-quadruplex. La spiegazione più plausibile per questo picco è una convergenza di segnali, probabilmente derivante dall'interazione di vimentina con altre proteine. Queste interazioni potrebbero portare a segnali erroneamente associati ai G-quadruplex, creando una sovrapposizione di dati che suggerisce una connessione che, in realtà, non esiste dal punto di vista della formazione di strutture G4. Questa interpretazione è supportata dal fatto che vimentina è una proteina versatile che, pur non legandosi direttamente al DNA in modo stabile, è coinvolta in una vasta gamma di funzioni cellulari, tra cui il trasporto nucleare, la risposta allo stress e l'adesione cellulare, tutte attività che la mettono in contatto con molte altre proteine, ma senza un legame diretto con la formazione di G-quadruplex nei telomeri o nei centromeri.

Un ulteriore affinamento dell'analisi ha incluso criteri più selettivi nella definizione dei G-quadruplex per isolare strutture più regolari e ipoteticamente più stabili, che potrebbero avere maggiori probabilità di interagire con vimentina. Ripetendo l'analisi su questa selezione più restrittiva, si mirava a individuare pattern o tendenze che potessero rivelare una correlazione più forte con la presenza di vimentina nelle regioni di interesse, rispetto ai G-quadruplex più eterogenei precedentemente considerati. Ciò che emerge da questa analisi è una prevalenza di loop di lunghezza 5 nelle regioni di intersezione tra ATAC e vimentina sia in presenza sia in assenza di telomeri e centromeri (Figura 29 e 30). Questo risultato potrebbe riflettere una preferenza strutturale specifica nelle regioni regolatorie, suggerendo un possibile ruolo nel controllo genico. Tuttavia, la predominanza di configurazioni '5-5-5' emersa in Figura 31 ha mostrato che la preferenza per i loop di lunghezza 5 è data dall'effetto della ripetitività nelle sequenze e non per una specifica funzione biologica. Questo evidenzia una criticità metodologica legata all'uso della *sliding window*, che può generare tendenze apparenti a causa del ripetersi di sequenze già ridondanti, piuttosto che da un'effettiva configurazione strutturale predisposta a funzioni specifiche.

Invece, la prevalenza del pattern ‘3-3-3’ nei telomeri e centromeri (Figura 34) sembra confermare le caratteristiche strutturali di queste regioni, supportando l’ipotesi che le sequenze ripetitive ricche di guanine nei telomeri facilitino la formazione di G4 altamente stabili (42).
Altri studi, in linea con i dati ottenuti, esaminano la formazione dei G4 nelle regioni promotrici degli oncogeni e la loro interazione con vimentina (43), rafforzando l’idea che usare i G4 in tali aree come bersagli potrebbe rappresentare una strategia utile per modulare l’espressione dei geni coinvolti nella proliferazione cellulare (43). Tuttavia, la variabilità di queste strutture rende difficile riconoscere caratteristiche strutturali uniche che potrebbero renderli selettivamente riconoscibili da agenti terapeutici, rimanendo un ostacolo importante nello sviluppo di trattamenti mirati (43).

CAPITOLO 5: CONCLUSIONE

Sebbene l'obiettivo principale di questa tesi, ossia individuare pattern nei G-quadruplex in corrispondenza del legame di vimentina con ATAC, non è stato completamente raggiunto, il lavoro svolto ha comunque contribuito a una comprensione più profonda della complessità di queste strutture genomiche. L'analisi ha rivelato tendenze interessanti nelle lunghezze dei loop dei G-quadruplex, che potrebbero suggerire nuovi spunti per l'esplorazione delle interazioni tra proteine e sequenze ripetitive del DNA.

Utilizzando la tecnica ATAC-seq, è stato possibile mappare in modo approfondito le aree di cromatina accessibile, evidenziando la presenza di G4 in prossimità di regioni promotoriali.

Dal punto di vista metodologico, la scelta di utilizzare criteri rigorosi per la selezione dei G-quadruplex e un'analisi basata sullo *sliding window* ha permesso di evidenziare la frequenza di pattern ricorrenti. Questo ha ridotto eventuali bias, sebbene alcune lunghezze dei loop siano risultate sovrarappresentate. La metodologia applicata, pur presentando queste limitazioni, ha quindi contribuito a una caratterizzazione più accurata delle regioni di interesse.

La selezione dei G-quadruplex secondo criteri definiti e l'adozione di approcci specifici per le aree con telomeri e centromeri rappresentano scelte mirate a ottenere un'analisi equilibrata, che tenga conto delle peculiarità strutturali di queste regioni senza trascurare l'eterogeneità delle altre zone genomiche.

I risultati ottenuti non solo dimostrano le potenzialità dell'ATAC-seq nel rilevare aree di cromatina accessibile e siti G4 con potenziale regolatorio, ma offrono anche una base per futuri studi mirati a esaminare in dettaglio le interazioni dirette tra G4 e vimentina.

Da un punto di vista teorico, questi risultati ampliano la comprensione del ruolo dei G-quadruplex nelle regioni non telomeriche e non centromeriche, suggerendo che la ripetitività dei loop possa influenzare la stabilità e la funzionalità delle strutture quadruplex in contesti genomici particolari. Questo contributo rappresenta un avanzamento nell'identificazione dei pattern preferenziali dei G-quadruplex che potrebbero essere coinvolti nella regolazione della cromatina e nell'interazione con proteine strutturali come la vimentina.

Dal punto di vista pratico, la caratterizzazione di queste strutture può avere un impatto su future ricerche che mirano a sviluppare farmaci o strategie terapeutiche basate sull'alterazione di questi motivi strutturali nei G-quadruplex. La comprensione approfondita delle condizioni che favoriscono la formazione di questi quadruplex in regioni specifiche del genoma potrebbe infatti supportare la progettazione di molecole mirate a stabilizzare o destabilizzare determinate

strutture quadruplex, che potrebbero avere applicazioni nel trattamento di malattie genetiche e di alcune forme di cancro.

In conclusione, pur non avendo completamente raggiunto l'obiettivo iniziale, questi risultati potranno essere ulteriormente approfonditi con studi futuri che includano tecniche complementari, per confermare e ampliare le osservazioni fatte. Un aspetto cruciale che merita attenzione è la possibilità di indagare più a fondo le implicazioni funzionali delle diverse configurazioni di G-quadruplex.

BIBLIOGRAFIA

(1) Andrew Travers ,Georgi Muskhelishvili. *DNA structure and function*. FEBS Journal 282: 2279–2295, 2015.

(2) Aaron Klug. *The discovery of the DNA double helix*. J. Mol. Biol 335: 3–26, 2004.

(3) Marek Bartkuhn, Rainer Renkawitz. *Long range chromatin interactions involved in gene regulation*. Biochimica et Biophysica Acta 1783: 2161–2166, 2008.

(5) Jochen Spiegel, Santosh Adhikari, Shankar Balasubramanian. *The structure and function of DNA G-quadruplexes*. Trends in Chemistry, Vol. 2, No. 2, February 2020.

(6) Daniela Rhodes, Hans J. Lipps. *G-quadruplexes and their regulatory roles in biology.* Nucleic Acids Research, Vol. 43, No. 18 8627–8637, 2015

(7) Marc Farag, Liliane Mouawad. *Comprehensive analysis of intramolecular G-quadruplex structures: furthering the understanding of their formalism*. Nucleic Acids Research, 52 , 3522–3546, 2024.

(9) Fiorella C. Grandi, Hailey Modi, Lucas Kampman, M. Ryan Corces. *Chromatin accessibility profiling by ATAC-seq*. Nature Protocols Vol 17: 1518–1552, Giugno 2022.

(10) Feng Yan, David R. Powell, David J. Curtis, Nicholas C. Wong. *From reads to insight: a hitchhiker's guide to ATAC-seq data analysis*. Genome Biology 21:22, 2020

(11) Sandy L. Klemm, Zohar Shipony, William J. Greenleaf. *Chromatin accessibility and the regulatory epigenome*, Nature Reviews, Genetics volume 20, Aprile 2019.

(12) Frida Danielsson, McKenzie Kirsten Peterson, Helena Caldeira Araújo, Franziska Lautenschläger, Annica Karin Britt Gad. *Vimentin diversity in health and disease*. Cells 7, 147, 2018.

(13) Denise Paulin, Alain Lilienbaum, Sareen Kardjian, Onnik Agbulut, Zhenlin Li. *Vimentin: Regulation and pathogenesis*. Biochimie 197: 96-112, 2022.

(14) Johanna Ivaskaa, Hanna-Mari Pallarib, Jonna Nevoa, John E. Erikssonb. *Novel functions of vimentin in cell adhesion, migration, and signaling*. Elsevier, 2007.

(15) Sarthak Gupta, Alison E. Patteson, J. M. Schwarz. *The role of vimentin–nuclear interactions in persistent cell motility through confined spaces*. New Journal of Physics 23, 2021.

(16) Michael C. Keeling, Luis R. Flores, Asad H. Dodhy, Elizabeth R. Murray, Núria Gavara. *Actomyosin and vimentin cytoskeletal networks regulate nuclear shape, mechanics and chromatin organization.* Scientific Reports 7: 5219, 2017.

(17) Silvia Ceschi, Michele Berselli, Marta Cozzaglio, Mery Giantin, Stefano Toppo, Barbara Spolaore, Claudia Sissi. *Vimentin binds to G-quadruplex repeats found at telomeres and gene promoters*. Nucleic Acids Research, Vol. 50, No. 3: 1370–138, 2022.

(18) Saima Usman, Naushin H. Waseem, Thuan Khanh NgocNguyen, Sahar Mohsin ,Ahmad Jamal, Muy-Teck Teh, Ahmad Waseem1. *Vimentin is at the heart of epithelial mesenchymal transition (EMT) mediated metastasis*. Cancers 13, 4985, 2021.

(19) Katerina Strouhalova, Magdalena Prechová, Aneta Gandalovicová, Martin Gregor, Daniel Rosel. *Vimentin intermediate filaments as potential target for cancer treatment*. Cancers 12, 184, 2020.

(20) Rebecca L. Siegel, Kimberly D. Miller, Nikita Sandeep Wagle, Ahmedin Jemal. *Cancer statistics, 2023*. CA Cancer J Clin; 73:17–48, 2023.

(21) Esam Bashir Yahya, Amaal Mohammed Alqadhi. *Recent trends in cancer therapy: A review on the current state of gene delivery*. Life Sciences 269 119087, 2021.

(23) Yuanjun Lu, Yau-Tuen Chan, Hor-Yue Tan, Sha Li, Ning Wang, Yibin Feng. *Epigenetic regulation in human cancer: the potential role of epi-drug in cancer therapy*. Molecular Cancer 19:79, 2020.

(24) Ilias Georgakopoulos-Soares, Candace S. Y. Chan, Nadav Ahituv, Martin Hemberg. *High-throughput techniques enable advances in the roles of DNA and RNA secondary structures in transcriptional and post-transcriptional gene regulation*. Genome Biology 23:159, 2022.

(25) Luciane T. Kagohara, Genevieve L. Stein-O'Brien, Dylan Kelley, Emily Flam, Heather C. Wick, Ludmila V. Danilova, Hariharan Easwaran, Alexander V. Favorov, Jiang Qian, Daria A. Gaykalova, Elana J. Fertig. *Epigenetic regulation of gene expression in cancer: techniques, resources and analysis.* Briefings in Functional Genomics, 17(1), 49–63, 2018.

(36) Zhijun Fu, Sanjie Jiang, Yiwen Sun, Shanqiao Zheng, Liang Zong, Peipei Li. *Cut&tag: a powerful epigenetic tool for chromatin profiling.* Epigenetics, Vol. 19, No. 1, 2293411, 2024.

(40) Siyuan Cheng, Benpeng Miao, Tiandao Li, Guoyan Zhao, Bo Zhang. *Review and evaluate the bioinformatics analysis strategies of ATAC-seq and CUT&Tag data.* Genomics, Proteomics & Bioinformatics, 22(3), 2024.

(41) Jing Lyu, Rui Shao, Philip Yuk Kwong Yung, Simon J. Elsasser. *Genome-wide mapping of G-quadruplex structures with CUT&Tag.* Nucleic Acids Research, Vol. 50, No. 3 e13, 2022.

(42) Jun Tan, Li Lan. *The DNA secondary structures at telomeres and genome instability.* Cell & Bioscience, 10:47, 2020.

(43) Marta Cozzaglio, Silvia Ceschi, Elisabetta Groaz, Mattia Sturlese, Claudia Sissi. *G-quadruplexes formation within the promoter of TEAD4 oncogene and their interaction with Vimentin*. Frontiers in Chemistry, 2022.

SITOGRAFIA

(26) https://www.ncbi.nlm.nih.gov/geo/

(27) https://www.ebi.ac.uk/ena/browser/home

(28) https://home.cc.umanitoba.ca/~psgendb/doc/fastqc.help

(29) https://github.com/FelixKrueger/TrimGalore/blob/master/Docs/Trim_Galore_User_Guide.md

(30) https://open.bioqueue.org/home/knowledge/showKnowledge/sig/bowtie2

(31) https://www.htslib.org/doc/1.9/samtools.html

(32) https://broadinstitute.github.io/picard/

(33) https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html

(34) https://bedtools.readthedocs.io/en/latest/content/overview.html

(35) https://igv.org/doc/webapp/#

(37) https://yezhengstat.github.io/CUTTag_tutorial/

(38) https://github.com/B3rse/qparse

(39) https://pandas.pydata.org/docs/user_guide/index.html

RINGRAZIAMENTI

Desidero esprimere la mia sincera gratitudine a tutte le persone che mi hanno supportato e aiutato durante il percorso che ha portato alla realizzazione di questa tesi.
In primo luogo, desidero ringraziare il mio relatore, il professore Enrico Lavezzo, per la sua disponibilità e la sua guida. Ringrazio anche il mio correlatore, il professore Stefano Toppo, per aver condiviso con me il suo sapere e le sue intuizioni.
Ringrazio sinceramente Marco per avermi seguita durante tutto il percorso e aver creato un ambiente stimolante e ricco di opportunità. La sua esperienza è stata fondamentale nel mio percorso e nella realizzazione di questo lavoro.
Un sentito grazie alle mie colleghe, Federica, Matilde, Ilaria, Gloria e Chiara, che con il loro sostegno, motivazione e compagnia, hanno reso questo percorso più piacevole e ricco di spunti.
Un grazie speciale va alle mie amiche, che conosco da quando ero piccola, per il loro supporto incondizionato e per aver condiviso risate e momenti indimenticabili lungo tutto questo percorso. Grazie di cuore Alessia, Alice, Giulia e Agnese.
Desidero ringraziare la mia famiglia per l'amore, la pazienza e la comprensione dimostrata in ogni momento. Senza il loro supporto incondizionato, questo lavoro non sarebbe stato possibile.
Non posso non menzionare i miei gatti, la loro compagnia è stata per me fonte di leggerezza nei momenti di fatica e stress.
Infine, un ringraziamento speciale va a Matteo per la sua pazienza e la sua capacità di farmi vedere le cose da una prospettiva migliore.
Grazie a tutti voi, che avete reso possibile questo traguardo. Senza di voi, non sarebbe stato lo stesso.