



UNIVERSITÀ
DI TRENTO

Dipartimento di Ingegneria e Scienza dell'Informazione

Corso di Laurea in
Informatica

ELABORATO FINALE

ANALYSIS OF RNA-SEQ TRANSCRIPTOMIC
DATA FROM TOTAL AND POLYSOMAL
mRNA FRACTIONS FROM AN EPITHELIAL
CANCER CELL LINE

Supervisore

.....

Laureando

Giacomo Fantoni

Anno accademico 2020/2021

Ringraziamenti

...thanks to...

Indice

Sommario	3
1 Introduzione	3
1.1 Controllo traduzionale nel cancro	3
1.1.1 Panoramica dell'iniziazione della traduzione	3
1.1.2 Meccanismi di traduzione deregolati e selettivi nel cancro	4
1.1.3 Vantaggi oncogenici selettivi di una traduzione deregolata	7
1.2 Polimorfismi a singolo nucleotide	8
1.2.1 Espressione genica allelo-specifica	8
1.3 TransSNP	8
1.3.1 Profilamento polisomico	9
1.3.2 Sequenziamento	9
2 Linea cellulare e dati di partenza	10
2.1 HCT116	10
2.1.1 Controllo traduzionale da parte di p53	10
2.2 Culture cellulari	11
2.3 Apoptosi indotta da Nutlin causata da un processo traduzionale regolato da PCBP2 e DHX30	11
2.3.1 Gli esiti diversi del trattamento nelle due linee sono riflessi da cambi nel traduttoma	12
2.3.2 Il potenziamento traduzionale dipendente dalla linea cellulare è associato con la presenza di un elemento nella 3'-UTR	12
2.3.3 Le proteine PCBP2 e DHX30 legano il motivo CGPD in maniera dipendente dalla linea cellulare	12
2.4 Dati di partenza	13
2.4.1 Dati di RNA-seq	13
2.4.2 Dati WES	13
3 Processamento dei dati	14
3.1 Pre-processamento e allineamento dei dati RNA-seq	15
3.1.1 Trimmomatic	15
3.1.2 Star	16
3.1.3 Samtools	17
3.1.4 Deduplicazione e ricalibrazione	17
3.2 Ottenimento degli SNP di interesse	20
3.3 Calcolo dei dati di sbilanciamento allelico	21
4 Analisi dei dati	23
4.1 Conta degli SNP trovati con ASEQ	23
4.1.1 Distribuzione degli SNP	24
4.2 Qualità dei campioni	24
4.3 Considerazioni sulla recalibrazione	24
4.4 Ottenere i dati per gli SNP di interesse	25
4.4.1 Boxplot	26

4.5 Caratterizzazione degli SNP di interesse	26
4.6 Conclusioni	26
Bibliografia	26
A Titolo primo allegato	29
A.1 Titolo	29
A.1.1 Sottotitolo	29
B Titolo secondo allegato	30
B.1 Titolo	30
B.1.1 Sottotitolo	30

Sommario

Sommario è un breve riassunto del lavoro svolto dove si descrive l'obiettivo, l'oggetto della tesi, le metodologie e le tecniche usate, i dati elaborati e la spiegazione delle conclusioni alle quali siete arrivati.

Il sommario dell'elaborato consiste al massimo di 3 pagine e deve contenere le seguenti informazioni:

- contesto e motivazioni
- breve riassunto del problema affrontato
- tecniche utilizzate e/o sviluppate
- risultati raggiunti, sottolineando il contributo personale del laureando/a

1 Introduzione

Il progetto di ricerca presentato in questo elaborato tenta di replicare il processo presentato in [1] su un'altra linea cellulare: *HCT116*¹. Questo viene fatto in modo da eliminare la possibilità che il fenomeno osservato sia specifico a *MCF7*. Sempre per determinare se l'evento è estensibile al di fuori dell'essere umano il progetto è stato replicato in parallelo sulla linea cellulare murina *B16-F10*². In particolare si tenta di determinare una nuova classe di mutazioni dette *transSNP* che potrebbero essere in grado di avere effetto sui livelli di espressione di proteine controllandone la traduzione e perciò essere una fonte potenziale di variazione inter-individuale nel rischio di cancro³.

1.1 Controllo traduzionale nel cancro

La traduzione di mRNA in proteine è un evento chiave nella regolazione dell'espressione genica. Questo è specialmente vero nel contesto del cancro in quanto molti oncogeni ed eventi di trasformazione sono regolati a questo livello. In [2] vengono esplorati i diversi modi in cui le cellule del cancro deregolano e riprogrammano la traduzione e il loro impatto oncogenico. Risorse considerevoli sono dedicate alla traduzione di mRNA in cellule normali. Fino al 20% dell'energia cellulare viene impiegata nella sintesi delle proteine e anche la maggior parte della trascrizione è volta alla produzione di RNA ribosomiale e di mRNA codificante proteine ribosomiali. La proliferazione di cancri maligni richiede una sintesi continua di proteine e un aumento del contenuto di ribosomi. Molte cellule tumorali subiscono stress fisiologico come ipossia o mancanza di nutrienti. In queste condizioni l'efficienza della traduzione si riduce, ma i meccanismi di regolazione vengono disaccoppiati dal processo nelle cellule tumorali come conseguenza del processo di trasformazione, aumentando lo stress della cellula. Si nota come le cellule tumorali possono sfruttare il meccanismo di traduzione per sostenere la loro proliferazione, sopravvivenza e metastasi. Questo avviene cambiando l'attività e l'espressione di fattori di traduzione che conferiscono alla cellula una capacità di traduzione di mRNA specifica per il cancro.

1.1.1 Panoramica dell'iniziazione della traduzione

L'iniziazione della traduzione ha un ruolo fondamentale nella regolazione della proliferazione cellulare, differenziazione e apoptosis, come descritto in [3]. L'iniziazione è regolata dall'assemblaggio dei complessi ternario e eIF4F. Il complesso ternario è formato da eIF3 e dal Met-tRNAs e recluta la subunità

¹Riferimento a linea cellulare

²Ce lo metto? Devo citare in qualche modo?

³Magari qua ci va un termine tecnico?

ribosomiale 40S in modo da formare il complesso di pre-iniziazione 43S. Questo si lega al cap di mRNA insieme ad altri fattori di traduzione e inizia a scansionare la 5'-UTR fino al codone iniziatore AUG, dove si unisce la subunità 60S e si forma il ribosoma 80S.

Un punto importante di regolazione è lo scambio di eIF2 di guanosina difosfato e guanosina trifosfato catalizzato da eIF2B. Questo scambio viene inibito quando la subunità eIF2 α viene fosforilata, riducendo così il tasso di iniziazione della traduzione. Questo avviene in quanto si lega con alta affinità al fattore di scambio del nucleotide guanosina eIF2B inibendo la sua funzione. Questo riduce nel complesso l'espressione di proteine oncogeniche e aumenta l'espressione di proteine soppressori dei tumori e pro-apoptotiche.

Nonostante la maggior parte degli mRNA siano reclutati al ribosoma attraverso il riconoscimento del loro cap a 5', un sottoinsieme può essere tradotto utilizzando un punto di ingresso del ribosoma interno IRES. Oltre al complesso ternario è coinvolto nell'iniziazione il complesso eIF4F, composto da una RNA elicasi, la subunità legante il cap e una proteina strutturale. Il complesso è si lega alla struttura del cap, svolge la struttura secondaria del 5' del mRNA e recluta il complesso di pre-iniziazione 43S. Stimola pertanto il reclutamento del ribosoma al mRNA, si è notato come questo avviene sia per mRNA con cap che senza cap in vitro.

Un altro passaggio di regolazione coinvolge le proteine leganti eIF4E: 4E-BP, una famiglia di represori di eIF4F. Una loro ipo-fosforilazione causa il legame con eIF4E e impedisce il reclutamento del macchinario di traduzione al mRNA. Una sua fosforilazione invece distrugge il legame e favorisce la traduzione.

1.1.2 Meccanismi di traduzione deregolati e selettivi nel cancro

Dati i tassi di allungamento del ribosoma sul RNA costanti e i fattori di iniziazione di traduzione limitati, la traduzione di mRNA specifici per cui il reclutamento del ribosoma è inefficiente viene disproporzionalmente affetta da cambi nell'attività dei fattori di iniziazione. Da questo si nota come le cellule tumorali possiedono una varietà di complesse alterazioni molecolari che aumentano la traduzione selettiva di mRNA mal tradotti⁴. In particolare aumenta l'espressione o la disponibilità di fattori di iniziazione della traduzione specifici e l'attività dei pathways di segnalazione che li regolano. L'espressione aberrante dei fattori di iniziazione di traduzione è il primo meccanismo scoperto attraverso cui le cellule di cancro deregolano la traduzione. È stato dimostrato originariamente dall'abilità di eIF4E sovra-espresso di trasformare le cellule NIH 3T3. Ulteriori fattori di iniziazione sono stati scoperti in tumori umani.

Formazione del complesso eIF4F

I ribosomi sono reclutati alla terminazione 5' del mRNA attraverso il complesso eIF4F, composto di tre subunità, tra cui eIF4A che fornisce l'attività elicasica necessaria per svolgere le strutture secondarie presenti nella 5'-UTR. Questo processo è aiutato dalle altre due subunità eIF4G e eIF4E. Considerando che tipicamente tale regione degli mRNA oncogenici è lunga e stabile questi sono molto sensibili all'attività di eIF4A e alla formazione di eIF4F. Tutte e tre le subunità possono essere deregolate nelle cellule cancerogene: i loci genomici sono amplificati in molti tumori umani e sono tutti obiettivi dell'oncoproteina MYC. La sovra-espressione di eIF4E e eIF4G, che agiscono come classici oncogeni, risulta nella trasformazione delle cellule in vitro e in vivo.

La regolazione dell'iniziazione della traduzione nel cancro può essere anche regolata alla fosforilazione del complesso eIF4F. La fosforilazione di eIF4E promuove lo sviluppo di tumori e la loro disseminazione ed è elevata in tumori umani di polmoni, prostata e seno. Un sito di fosforilazione di eIF4G quando legato promuove la fosforilazione di eIF4E, coinvolta nella riprogrammazione traduzionale che porta alla resistenza al tamoxifene nel cancro al seno. Questo meccanismo è poco compreso, ma si pensa coinvolga il riciclo dei fattori di iniziazione.

Un ulteriore meccanismo di regolazione coinvolge il sequestro dei fattori di iniziazione per impedire la formazione del complesso eIF4F. Questo avviene da parte di PDCD4 (tumor suppressor programmed cell death 4), la cui perdita è associata con un invasione della cellula tumorale e un abbassamento della probabilità di sopravvivenza per alcuni tumori. Un meccanismo simile coinvolge 4E-BP, che competo-

⁴Il paper intendeva trascritti?

no con eIF4G per il legame con eIF4E, inibendo la traduzione dipendente dal cap. La loro espressione può essere persa o la loro funzione inibita attraverso fosforilazione. 4E-BP possono contrastare la metastasi, ma facendolo promuovono lo sviluppo di grandi tumori locali.

Formazione del complesso ternario

Il complesso ternario o *TC* è composto di eIF2, GTP e dal tRNA con metionina iniziatore. La formazione deregolata del TC è un meccanismo complesso, in particolare sono stati trovati risultati contrastanti riguardo il ruolo della fosforilazione di eIF2 α . Un aumento in questa modifica garantisce alle cellule tumorali un aumento nella loro abilità di rispondere a condizioni di stress promuovendo la traduzione di open reading frame a valle o *uORF* contenenti mRNA dedicati alla risposta a tale stress. Inoltre una sovra-espressione di eIF2 α o di una delle sue chinasi promuove in alcuni contesti la trasformazione. Una fosforilazione di eIF2 α a lungo termine, invece, promuove apoptosi e ha permesso lo sviluppo di terapie che promuovono l'attività delle sue chinasi o un'inibizione delle fosfatasi. Da questi risultati si osserva come il risultato della fosforilazione di eif2 α è specifica al contesto e potrebbe cambiare con il tempo.

Altri meccanismi per modulare l'attività del TC includono la sovra-espressione di eIF5 o delle sue proteine simili *MP* che, quando presenti in eccesso, si legano a eIF2 sequestrandolo dal ribosoma 40S. In modo simile alla fosforilazione di eif2 α il sequestro riduce la sintesi proteica globale, ma aumenta la traduzione di mRNA contenenti uORF. Questo meccanismo sembra essere rilevante per le proprietà maligne di alcuni tipi di cancro.

eIF3, connessione di eIF4F e il complesso di pre-iniziazione

eIF3 è un complesso che si lega direttamente a eIF4G unendolo al complesso di pre-iniziazione. In questo modo si connettono gli mRNA con la subunità 40S del ribosoma. Un aumento dei livelli di eIF3 dovrebbe promuovere l'unione dei due elementi, aumentando il tasso di iniziazione della traduzione. Questo aumenta la traduzione di trascritti oncogenici. Diversi studi hanno notato come quando sovra-espresso alcune subunità di eIF3 mostrano proprietà oncogene, altre agiscono come soppressori. Questo avviene a causa dei ruoli non traduzionali di alcune subunità, come eIF3a che si lega a componenti del citoscheletro e eIF3f e eIF3i, che regolano pathway di trasduzione del segnale. Sono stati trovati anche altri ruoli regolatori di eIF3 nella traduzione che includono il legame a strutture del mRNA nella 5'-UTR di trascritti rilevanti per il cancro.

Allungamento e terminazione della traduzione

Il processo di traduzione viene regolato anche durante l'allungamento e la terminazione e nuovi studi hanno mostrato cambi oncogenici in questi due processi. Un ruolo dominante è stato mostrato per la perita della regolazione inibitoria dell'allungamento attraverso la fosforilazione di eEF2. Inoltre l'aumento della disponibilità di specie di tRNA iso-acettanti nelle cellule tumorali sembra avere un ruolo nella tumorigenesi. In quanto la velocità di incorporazione degli amminoacidi durante la fase di allungamento è dipendente dalla disponibilità dei tRNA carichi corrispondenti, diversi studi hanno mostrato come nelle cellule del cancro il repertorio di tRNA disponibili viene riprogrammato in modo che le specie richieste per la traduzione di mRNA oncogenici siano presenti a livelli sufficienti. Oltre a questo l'allungamento può essere deregolato attraverso un cambiamento programmato del frame di lettura a -1. Lo scivoltamento indietro di una base porta alla creazione di codoni di stop prematuri e a decadimento del mRNA mediato dal non-senso. Questo meccanismo spiega il ruolo oncogenico di mutazioni silenti che inducono frameshift nei soppressori dei tumori.

La terminazione aberrante o alterata può portare alla fine della traduzione a codoni di stop prematuri come risultato di mutazioni somatiche e porta al cancro se queste si trovano su geni soppressori del tumore.

Oltre a questo si trovano due fattori di iniziazione con multipli ruoli nella traduzione del mRNA, che sono associati con regolazione alterata nelle cellule del cancro. eIF6 è un fattore anti-associazione ribosomale che impedisce interazioni aberranti tra le subunità 40S e 60S. Deve pertanto essere spostato

dal ribosoma nel passo finale della sintesi del ribosoma 60S nel nucleo e può promuovere il disassemblaggio del ribosoma 80S nel citosol impedendo la riassociazione dei ribosomi 60S post-terminazione. Questo impedisce ulteriori passaggi di iniziazione con un sequestro prolungato. Un'espressione aberrante di eIF6 causa un suo accumulo nel nucleo, con un ruolo in diversi tipi di cancro. Livelli ridotti della sua espressione invece impediscono trasformazioni indotte da oncogeni.

Il secondo fattore con multiple attività oncogene è eIF5A che oltre ad essere un fattore di iniziazione importante per la formazione del primo legame peptidico, ha un ruolo durante l'allungamento di regioni tripeptidiche mal tradotte. Si è notato come entrambe le sue isoforme siano sovra-espresso in molti tumori e sono state associate con la capacità metastatica delle cellule tumorali.

Cambi nelle regioni UTR nelle cellule di cancro

Sequenze e motivi strutturali presenti negli mRNA determinano la loro efficienza traduzionale e la loro abilità di essere regolati da fattori agenti in trans come microRNA, proteine leganti RNA e fattori di iniziazione. Questi elementi si trovano nelle regioni 5' e 3' UTR e tendono ad essere sovra-rappresentati in mRNA oncogenici garantendo una loro precisa regolazione. È stato mostrato come mutazioni come gli SNP in questi motivi non codificanti possono modulare in maniera significativa l'espressione di proto-oncogeni.

L'aumento di strutture secondarie nel 5' UTR ha effetto sul tasso di iniziazione della traduzione di mRNA cap-dipendente. In particolare si nota come mRNA oncogenici possiedono strutture stabili nella 5' UTR e hanno una maggiore dipendenza da *eIF4F*. Altri elementi nelle due UTR possono regolare l'efficienza della traduzione: una maggiore dipendenza da *eIF4E* ma non da *eIF4A* è stata mostrata per l'elemento iniziatore di traduzione del 5' UTR corto di alcuni mRNA.

In contrasto mRNA contenenti siti di ingresso di ribosomi interni o *IRES* sono altamente dipendenti da *eIF4G* e *eIF4A*. Inoltre gli mRNA possono contenere codoni di iniziazione alternativi e open reading frames *ORF* inibitori a valle del codone di inizio canonico che possono severamente contrastare la normale identificazione del normale sito di inizio di traduzione. Questi elementi di sequenza sono arricchiti di trascritti oncogenici e in condizioni di stress oncogenico alcuni di questi mRNA mostrano una traduzione aumentata.

Oltre a questo motivi di sequenza o strutturali in alcune delle 5' UTR mediane il reclutamento di proteine leganti RNA che modulano la sua traduzione. Un esempio ben caratterizzato di questo è l'elemento ditraduzione attivata dal transforming growth factor β *TGF- β* che regola la traduzione di certi mRNA coinvolti nella transizione da cellula epiteliale a mesenchimale promuovendo la migrazione cellulare.

Infine un altro elemento da tenere in conto sono i siti di legame per i microRNA, motivi particolarmente comuni con un effetto sulla traduzione e sulla stabilità del mRNA. La maggior parte di questi elementi riduce l'efficienza di traduzione del mRNA e si trovano in varie combinazioni in mRNA oncogenici attenuandone la traduzione in modo da impedire la trasformazione della cellula causata da una loro sovra-espressione. Nonostante tutto questo le cellule di cancro riescono a trovare meccanismi in grado di superare questi controlli.

Segnalazione oncogenica

La maggior parte dei segnali fisiologici, tra cui stimolazione dei fattori di crescita e di stress, funzioni metaboliche e fattori endocrini sono integrati attraverso il macchinario traduzionale. In particolare il target mammifero della rapamicina (*mTOR*) ha un ruolo fondamentale nella segnalazione regolatoria della traduzione: fosforila 4E-BP permettendo la formazione del complesso eIF4F e la proteina chinasi ribosomiale S6. S6K regola altri processi che alleviano l'inibizione della traduzione. Molti dei geni più comunemente mutati in molti tipi di cancro codificano proteine chiave che regolano pathway di segnalazione riguardanti la traduzione. Questo è fondamentale per iniziare e mantenere il fenotipo trasformato.

Ribosoma tumorale

È stato a lungo discusso se esistano modifiche specifiche al cancro nei ribosomi che potrebbero promuovere il riprogrammamento della traduzione di mRNA. Questa ipotesi è supportata dalla scoperta di ribosomopatie, una famiglia di sindromi causate da mutazioni ereditate nei geni codificanti proteine ribosomiali e i loro regolatori. Sono caratterizzate da difetti iniziali nell'ematopoiesi seguita da un'aumento alla suscettibilità al cancro. Nonostante questo l'effetto oncogenico di questi meccanismi rimane non chiaro.

Si nota inoltre come la stechiometria delle proteine ribosomiali e le modifiche al rRNA varia nelle cellule tumorali, suggerendo che ribosomi individuali potrebbero possedere modifiche uniche che alterano la loro abilità a tradurre certi mRNA. Non è chiaro però se questi ribosomi particolari siano in grado di restringere la sintesi dei soppressori dei tumori o di aumentare la traduzione di mRNA oncogenici. In supporto a questo si è trovato che la distruzione della discherina o di piccoli RNA nucleici che la guidano ai siti di rRNA sono comuni in molti tumori e possono impedire la traduzione di mRNA codificanti soppressori dei tumori come p53.

Infine il colligamento tra proteine ribosomiali, ribosomopatie e cancro è stato attribuito al ruolo non traduzionale dei componenti del macchinario di traduzione, principalmente la stabilizzazione di p53. Pertanto un complesso sub-ribosomiale composto del rRNA 5S, da RPL5 e da RPL11 si lega e sequestra MDM2, risultando in una stabilizzazione di p53 e nell'arresto del ciclo cellulare. Questo complesso si forma quando una biogenesi deregolata dei ribosomi porta a uno sbilanciamento delle componenti ribosomiali. Una perdita somatica di p53 permette alle cellule ematopoietiche di sfuggire all'arresto del ciclo causato dalla biosintesi difettiva del ribosoma, causando la predisposizione al cancro associata con le ribosomopatie.

1.1.3 Vantaggi oncogenici selettivi di una traduzione deregolata

Un gran numero di ricerche ha determinato la regolazione traduzionale di fattori ati apoptotici, cicline e chinasi dipendenti da cicline. Questi fattori sono fondamentali per la proliferazione e l'apoptosi delle cellule e una loro deregolazione può determinare la trasformazione della cellula in una di cancro.

Angiogenesi

L'angiogenesi dei tumori è un processo di continuo rimodellamento in modo da permettere la crescita del tumore ed è promosso da una varietà di meccanismi traduzionali. Gli mRNA codificanti i due principali regolatori id angiogenesi, VEGFA e HIF1 α sono tradotti grazie a una varietà di meccanismi che garantiscono la capacità della cellula tumorale di adattarsi all'ipossia. Pertanto la traduzione di questi due mRNA può essere promossa da meccanismi sia dipendenti dal cap che indipendenti da esso attraverso l'utilizzo di IRES e uORF, oltre a ulteriori meccanismi regolatori non canonici. Nonostante la loro traduzione sia associata con l'aumento dell'espressione di eIF4E nei tumori umani, una loro regolazione traduzionale complessa permette di mantenere la loro traduzione anche in condizioni di profonda ipossia e depravazione di nutrienti. Si nota come HIF1 α si lega al promotore di eIF4E promuovendo la sua trascrizione, suggerendo che in risposta all'ipossia la cellula potrebbe passare da un meccanismo cap-dipendente a uno cap-indipendente di traduzione.

Risposta allo stress

Oltre all'ipossia le cellule tumorali devono modulare la traduzione in modo da rispondere a una varietà di altri tipi di stress. Si nota come la risposta a diversi agenti di stress condivida meccanismi regolatori comuni: tutti gli mRNA tradotti in condizioni di stress sono regolati dalla fosforilazione di eIF2 α e includono uORF. Questi mRNA codificano proteine coinvolte in pathway che permettono l'adattamento delle cellule tumorali al loro ambiente. Meccanismi di iniziazione della traduzione non canonici, IRES e metilazione di mRNA mantengono la sintesi delle proteine nonostante lo stress che inibisce il processo cap-dipendente. Non è chiaro come gli mRNA siano tradotti selettivamente in risposta ad ogni tipo di stress. Inoltre la fosforilazione prolungata di eIF2 α porta alla morte cellulare, evento a cui le cellule tumorali potrebbero riuscire a sfuggire in quanto promuove la traduzione di fattori che permettono la sua defosforilazione portando a un feedback loop inibitorio.

Vantaggi oncogenici della traduzione deregolata emergenti

Considerando che la maggior parte delle morti causate dal cancro sono dovute alla disseminazione metastatica, un concetto chiave sotto studio è l'abilità delle cellule tumorali di deregolare la traduzione di fattori prometastatici.

Un altro aspetto sotto studio è l'importanza della traduzione nell'aspetto del mantenimento del bilanciamento energetico e come lo stato energetico e la sintesi proteica sono regolate reciprocamente per raggiungere un equilibrio.

Inoltre si sta analizzando il rapporto tra la traduzione e le specie reattive con l'ossigeno *ROS*: componenti del macchinario di traduzione sono particolarmente sensibili all'ossidazione della cisteina da parte di tali elementi. Gli mRNA codificanti proteine antiossidanti posseggono un motivo che conferisce una regolazione tradizionale in risposta all'aumento dei livelli di espressione di eIF4E.

Infine la traduzione deregolata può promuovere l'espressione di proteine coinvolte nella riparazione del DNA permettendo la fuga dalla senescenza indotta dagli oncogeni e resistenza agli agenti danneggianti il DNA.

Si nota come la sintesi delle proteine fornisce alle cellule tumorali un modo cruciale per distruggere una varietà di processi importanti per tutti i passaggi nella biologia del cancro.

1.2 Polimorfismi a singolo nucleotide

I polimorfismi a singolo nucleotide, da qui in avanti denominati *SNP* sono delle mutazioni nel genoma causate dal cambio di un singolo nucleotide nella molecola di DNA presenti in almeno 1% della popolazione. Sono una delle classi più grandi di variabilità genetica che possono sottostare o essere responsabili di variazioni inter-individuali in fenotipi complessi di malattie. Grazie allo sviluppo tecnologico nelle tecnologie di sequenziamento e alla nascita del next-generation sequencing sono state rese disponibili informazioni a livello della singola base sul genoma umano e sul trascrittoma permettendo l'esplorazione di questioni biologiche prima insondabili. Infatti diversi strumenti sono stati implementati per studiare i dati di espressione genica basati su RNA-sequencing in modo da identificare istanze di espressione genica allelo-specifica.

1.2.1 Espressione genica allelo-specifica

Si intende per espressione genica allelo-specifica o *ASE* una condizione per cui alleli diversi di un gene, per lo scopo di questo progetto uno contenente uno SNP e uno no, mostrano un'attività trascrizionale considerevolmente diversa. Un evento di ASE si osserva pertanto nelle cellule umane dove la trascrizione si origina principalmente da un allele. Questi fenomeni sono dovuti principalmente a geni imprinted, condizioni fisiologiche come l'inattivazione del cromosoma *X* e contribuiscono alla variabilità fenotipica umana. Ulteriori meccanismi comprendono degradazione dei trascritti da parte di miRNA, distruzione monoallellica di regioni regolatorie, pattern di splicing alternativi o fenomeni epigenetici.

1.3 TransSNP

Una frazione di SNP identificati nella popolazione umana sono locati nelle regioni codificanti o negli *UTR*. In questo caso studi guidati da meccanismo e da associazione hanno tentato di studiare SNP funzionali che possono modificare aspetti di regolazione genica post-trascrizionale. Non sono però stati ancora esplorati SNP associati con alterazioni nel potenziale di traduzione del mRNA, estendendo il concetto di ASE dall'aspetto trascrizionale all'aspetto tradizionale. Lo scopo del progetto è pertanto identificare questi ultimi, SNP in grado di cambiare l'efficienza della traduzione del mRNA che li contiene, e una loro eventuale correlazione con il cancro, denominati transSNP. In particolare si tenta di determinare mutazioni in grado di andare a inficiare l'integrità dei motivi presenti nelle UTR, come quelli descritti in §1.1.2 Si nota infatti come la regolazione tradizionale governa la produzione di proteine in risposta a un gran numero di situazioni fisiologiche e patologiche: circa metà della variazione della concentrazione di una proteina è dovuta a questo tipo di controllo. Per farlo si utilizza un'analisi comparativa di sbilanciamento allelico tra frazioni di mRNA totali e polisomiali estratti dallo stesso campione cellulare in modo da superare il rumore causato dalla chiamata degli

SNP e dal coverage derivato dai dati di RNA-seq. Questo tipo di analisi viene svolto unicamente su SNP in eterozigosi nella cellula: in questo modo si ottiene la percentuale di allele presente in una frazione rispetto all’altro⁵. In questo modo si crea un catalogo di SNP codificanti e negli UTR associati e cause potenziali con alterazioni nel potenziale di traduzione degli mRNAs.

1.3.1 Profilamento polisomico

Il profilamento polisomico è il metodo con cui le frazioni di mRNA totali e polisomali vengono separate in un campione e il suo protocollo viene descritto in [4]. Permette di determinare il sottoinsieme di mRNA attivamente coinvolti nella traduzione o traduttoma, ritornando una visione funzionale del genoma in un dato momento in una data cellula. Questo metodo offre diversi vantaggi rispetto ad altri, per esempio, a differenza del profilamento a ribosomi questa tecnica dà accesso all’intera lunghezza degli mRNA, comprese le UTR, le regioni che questo progetto vuole analizzare. La separazione delle due frazioni si basa su una centrifugazione con gradiente: i ribosomi hanno un coefficiente di sedimentazione molto maggiore rispetto alle molecole di mRNA e pertanto si troveranno ad altezze diverse della colonna. Pertanto le cellule vengono lisate e i lisati citoplasmatici vengono caricati su un gradiente di saccarosio lineare 10 – 50%, ultra-centrifugate e frazionate attraverso un collettore automatico di frazioni che tiene conto dell’assorbanza a 254nm. Tutte le parti più leggere contenenti frazioni subpolisomali presenti dalla cima fino alla frazione corrispondente al monosoma 80S sono assunte non attivamente coinvolte nel processo di traduzione e raccolte in una provetta. Le frazioni più pesanti sono quelle attivamente tradotte e sono raccolte in una seconda provetta⁶. Successivamente le molecole di RNA sono purificate e sospese in acqua sterile. In questo modo la seconda provetta contiene la frazione polisomale di interesse per il progetto.

La frazione totale viene ottenuta attraverso un’estrazione con TRIzol (ThermoFisher) di una popolazione cellulare separata preparata in parallelo⁷.

In questo modo sono state ottenute tutte le due frazioni di mRNA, prima la polisomale e poi la totale che verranno sequenziate in modo poi da poter studiare lo sbilanciamento allelico al loro interno. Valori diversi di ASE tra le due frazioni indicano un cambio nel potenziale di traduzione causato dallo SNP considerato.

1.3.2 Sequenziamento

Le frazioni ottenute attraverso il profilamento polisomico vengono poi sequenziate attraverso *HiSeq 2500* di Illumina. Le molecole di RNA vengono frammentate e convertite in cDNA a cui viene aggiunta una sequenza adattatrice. Successivamente avviene un’amplificazione con *PCR* i cui risultati sono caricati in una *flow cell* dove i frammenti sono catturati da oligonucleotidi legati alla superficie complementari agli adattatori. Si formano in questo modo dei cluster di frammenti che presentano la stessa sequenza adattatrice. Successivamente i reagenti di sequenziamento, che includono nucleotidi etichettati con un nucleotide fluorescente sono aggiunti in modo da incorporare la prima base. La flow cell è letta dalla macchina che registra la lunghezza d’onda emessa dai cluster. La particolare lunghezza d’onda permette di identificare il nucleotide. Il ciclo viene poi ripetuto *n* volte in modo da creare una read lunga *n* basi. In questo modo si ottengono per ogni frazione i file *fastq* poi utilizzati per l’analisi dell’espressione allelo-specifica.

⁵Non riesco a spiegare bene il concetto

⁶Provetta è il termine giusto? Nel paper si parla di “tube”

⁷Sto prendendo dal paper, magari per la mia linea cellulare queste cose sono state fatte in modo diverso

2 Linea cellulare e dati di partenza

2.1 HCT116

La linea cellulare soggetto di analisi è HCT116. È una linea cellulare di un carcinoma colon-rettale umano presente nel pannello originale delle 60 linee caratterizzate a fondo dall'iniziativa del national cancer institute statunitense (NCI-60). Come descritto in [5] la linea è perfettamente diploide e mostra un'abilità tumorigenica intermedia: se vengono iniettate in una popolazione di topi nudi atimici $5 \cdot 10^6$ cellule il 50% di questi sviluppano un tumore dopo un periodo di latenza di 16 giorni. Oltre a queste caratteristiche la linea cellulare è wild-type per p53, una proteina soppressore dei tumori in grado di regolare la sintesi proteica in modo da inibire la crescita cellulare.

2.1.1 Controllo traduzionale da parte di p53

La proteina p53 soppressore dei tumori è il fattore di trascrizione mammifero meglio caratterizzato che media diversi processi anti-proliferativi. Ha un effetto importante sull'iniziazione della traduzione e una sua caratterizzazione è fondamentale per comprendere il ruolo che sue mutazioni o deregolazioni hanno nella biologia del cancro. p53, oltre a regolare la trascrizione, controlla la biogenesi dei ribosomi e dei fattori di iniziazione eucarioti. Si analizza pertanto il suo ruolo come regolatore dei complessi ternario e eIF4E e nella biogenesi dei ribosomi.

p53 limita la biogenesi dei ribosomi

I ribosomi sono responsabili del trasferimento dell'informazione contenuta negli mRNA in proteine. La loro biogenesi ha luogo nel nucleolo, in cui il DNA ribosomiale è organizzato. La subunità 60S è una molecola complessa composta di 3 RNA ribosomalni e 47 prtoeine. Questo complesso è responsabile per la formazion edel legame peptidico e del controllo della qualità del peptide nascente. La subunità 40S invece è responsabile per lo svolgimento e la scansione del mRNA ed è composta da 1 rRNA e 33 prtoeine. Disturbi nel processo di biosintesi delle componenti del ribosoma ha un ruolo centrale nella tumorigenesi.

p53 limita questo processo: regola infatti la RNA polimerasi I, che sintetizza gli RNA ribosomalni, inibendola. Questo processo coinvolge l'interferenza di p53 con un insieme di proteine richieste per l'assemblaggio e l'iniziazione del macchinario trascrizionale sul promotore del gene del rRNA. p53 Si lega alla proteina legante la TATA-box e ai suoi fattori associati impedendo la loro interazione con fattori di legame a monte e reprimendo la trascrizione del RNA polimerasi I.

Inoltre p53 inibisce l'attività della RNA polimerasi III, diminuendo la produzione di tRNA, del rRNA 5S e di altri piccoli RNA coinvolti nel trasporto e nel processamento del RNA impedendo l'attacco della polimerasi al DNA.

p53 regola la trascrizione dei geni RP

Le proteine ribosomalni o RP di nuova sintesi sono importate nel nucleolo dal citosol. In risposta a stress nucleolare diverse RP traslocano nel nucleoplasma legandosi e inibendo l'attività di MDM2 e causando un arresto del ciclo cellulare e apoptosi mediati da p53. Inoltre in risposta a danni al DNA la proteina ribosomalna RPL26 si lega alle terminazioni del mRNA di p53 aumentando la sua traduzione e portando all'arresto del ciclo cellulare. Infine durante una condizione di stress genotossico p53 induce l'espressione di una proteina ribosomalna che aumenta l'espressione di p21, che media a sua volta l'arresto del ciclo cellulare.

Oltre a regolare la trascrizione di rRNA p53 controlla il processamento dei pre-rRNA inibendo i livelli di espressione di FBL (fibrillarin), una proteina nucleolare vitale per la metilazione e il processamento dei

pre-rRNA. L'inibizione causa un'alta infedeltà della traduzione e aumenta l'iniziazione di traduzione dipendente da *IRES*.

p53 regola l'assemblaggio del complesso ternario e di eIF4F

p53 è in grado di attenuare la sintesi globale di proteine grazie all'inibizione della proteina ribosomiale S6 chinasi, una chinasi a monte di 4E-BP1. Inoltre un gene obiettivo di p53, TRIM22 inibisce il legame di eIF4E a eIF4G. Inoltre p53 causa defosforilazione e rottura del fattore di iniziazione eIF4GI e di 4E-BP1. Si è dimostrato come gli effetti combinati dell'assenza delle 4E-BP e di p53 aumentano sinergisticamente la proliferazione cellulare e la tumorigenesi. p53 pertanto inibisce la sintesi proteica attraverso l'inibizione di eIF4E e l'assemblaggio del complesso eIF4F aumentando la de-fosforilazione di 4E-BP1. Non interagisce con il complesso ternario ma unicamente con diverse componenti del complesso eIF4F.

p53 inibisce anche la segnalazione con mTOR e l'attività della proteina chinasi CK2, in grado di regolare a sua volta eIF2 α e quindi il complesso ternario. CK2 è in grado di fosforilare p53 a un residuo altamente conservato, causandone la traslocazione nel mitocondrio e apoptosi indipendente dalla trascrizione dopo l'esposizione della cellula ad agenti genotossici. Infine la subunità regolatoria β di CK2 è in grado di interagire con p53 riducendone la funzione transattivatrice e l'affinità con il DNA.

2.2 Culture cellulari

I dati di RNA-seq sono stati ottenuti per colture di HCT116 in diverse condizioni. Oltre a un ambiente di controllo in presenza di *DMSO* è stata utilizzata anche *Nutlin*, una piccola molecola in grado di attivare p53 e pertanto arresto del ciclo cellulare e apoptosi. Anche lo stato genetico è stato modificato dalla condizione di controllo denominata *scr* è stato svolto il knockout di due geni: *PCBP2* e *DHX30*, coinvolti nel processo traduzionale attivato da p53 che porta all'apoptosi. Si ottengono infine 6 condizioni diverse su cui viene svolto il profilamento polisomico e da cui si ottengono i dati di RNA-seq:

Denominazione	Ambiente	Stato genetico
scr_DMSO	DMSO	Normale ¹
scr_NUTLIN	NUTLIN	Normale
shDHX30_DMSO	DMSO	knockout di DHX30
shDHX30_NUTLIN	NUTLIN	knockout di DHX30
shPCBP2_DMSO	DMSO	knockout di PCBP2
shPCBP2_NUTLIN	NUTLIN	knockout di PCBP2

Tabella 2.1: Condizioni di coltura

2.3 Apoptosi indotta da Nutlin causata da un processo traduzionale regolato da PCBP2 e DHX30

Come descritto in §2.1.1 p53 è una proteina strettamente controllata e altamente pleiotropica tipicamente disattivata nelle cellule tumorali umane. Tra le varie funzioni svolte da essa quella considerata più rilevante nel contesto del cancro è il controllo della morte cellulare programmata. In quanto una funzione di p53 non controllata produrrebbe una morte cellulare massiva esiste un MDM2, una proteina agente come una ubiquitina ligasi E3 che ne inibisce l'attività. Questa proteina inibitrice è tipicamente sovra-espressa nella frazione di tumori che mantengono p53 wild-type, come succede per HCT116. Un trattamento possibile per questi tipi di tumori va ad attaccare questo meccanismo, in particolare Nutlin inibisce l'interazione tra p53 e MDM2, attivando la prima. I risultati di tale trattamento variano in una combinazione di arresto del ciclo cellulare, senescenza e apoptosi, in proporzioni di difficile previsione. Lo studio [6] tenta di determinare le cause dietro le diversità fenotipiche a seguito del trattamento con Nutlin. Utilizza come modello due linee cellulari: HCT116 e SJSA1, un

osteosarcoma maligno. Se la prima a seguito del trattamento con Nutlin subisce un arresto del ciclo cellulare, la seconda subisce un'apoptosi massiva.

2.3.1 Gli esiti diversi del trattamento nelle due linee sono riflessi da cambi nel traduttoma

Se i cambi trascrizionali causati dal trattamento con Nutlin contengono invariantemente geni coinvolti in multipli pathway, si analizza il traduttoma in modo da determinare differenze in esso tra le due linee cellulari. Dopo il trattamento per ognuna delle linee cellulari si sono analizzati i dati di RNA-seq frazionati con profilamento polisomico e sono stati identificati diversi geni espressi in maniera diversa rispetto alla condizione di controllo. Questi sono indicati come *DEG* e sono stati divisi in 3 classi:

- Accoppati: DEG in cui cambi nella frazione polisomiale sono accoppiati con cambi nella frazione totale.
- Regolati traduzionalmente: DEG che esibiscono cambi unicamente nella frazione polisomiale.
- Invariati nella traduzione: DEG che esibiscono cambi unicamente nella frazione sub-polisomiale.

Tra questi sono stati presi in analisi i DEG regolati traduzionalmente e si è notato come i geni associati con la segnalazione apoptotica sono traduzionalmente potenziati unicamente nelle cellule SJS1, spiegando la loro abilità di avviare l'apoptosi a seguito del trattamento. Questo mostra come anche se l'attivazione di p53 da parte di Nutlin attivi programmi trascrizionali nelle due linee cellulari, questi causano programmi traduzionali diversi.

2.3.2 Il potenziamento traduzionale dipendente dalla linea cellulare è associato con la presenza di un elemento nella 3'-UTR

Data la specificità degli obiettivi regolati traduzionalmente in ogni linea cellulare è stato cercato un motivo de novo. Questo ha permesso l'identificazione di un elemento cis-regolatorio specificatamente arricchito nella 3'-UTR di geni potenziati traduzionalmente solo nelle cellule SJS1. Il motivo, definito come motivo-CGPD corrisponde al consenso 5'-*CCCC(A/C)(T/G)GGCCCT*-3'. Circa il 65% dei geni potenziati traduzionalmente nelle cellule SJS1 dopo il trattamento presentano almeno una copia del motivo-CGPD. In HCT116 invece l'espressione dei geni contenenti il motivo o geni-CGPD non è attivata da p53.

Ulteriori analisi hanno mostrato come l'effetto maggiore si ottiene quando il motivo-CGPD è presente in due copie nel mRNA. Si è notato come il motivo-CGPD è sufficiente per potenziare la traduzione di un mRNA reporter in cellule SJS1 e conferisce una tendenza alla repressione traduzionale nelle cellule HCT116.

2.3.3 Le proteine PCBP2 e DHX30 legano il motivo CGPD in maniera dipendente dalla linea cellulare

Per scoprire il meccanismo sottostante il risultato diversificato del trattamento con Nutlin nelle due linee cellulari si analizzano proteine leganti il RNA capaci di riconoscere il motivo-CGPD in maniera dipendente dal ciclo cellulare. Ci si aspetta che queste proteine abbiano modelli di espressioni opposti tra le due linee cellulari. Questo ha portato all'identificazione di DHX30 tra gli estratti di HCT116 che ha mostrato un cambio di espressione tra il mRNA polisomiale e tra i livelli di proteina. Inoltre si è analizzata PCBP2, facente parte della famiglia di proteine PCBP, capaci di creare forti legami con lunghezze poli-C come quella presente nel motivo. Quest'ultima presenta maggiori livelli di espressione a livello proteico in HCT116 rispetto a SJS1, suggerendo un suo ruolo nella regolazione del motivo-CGPD in HCT116. DHX30 e PCBP2 sono state sottoposte ad ulteriori analisi e si è notato come gli mRNA che presentano un aumento del potenziale di traduzione in SJS1 sono obiettivi di queste due proteine. Questo aumento non è così significativo invece in HCT116. Si è anche notata una sovrapposizione tra i siti di legame di DHX30, PCBP2 e il motivo-CGPD. Questo suggerisce PCBP2 e DHX30 come probabili candidati capaci di legare il motivo-CGPD in vivo. Inoltre le due proteine sono espresse a diversi livelli nelle due linee cellulari, suggerendo che potrebbero contribuire alla repressione traduzionale di mRNA contenenti il motivo nelle cellule HCT116. Ulteriori anali hanno dimostrato come il legame di DHX30 al motivo-CGPD avviene principalmente nelle cellule HCT116, mentre

PCBP2 in entrambe. Viene suggerito come PCBP2 agisce come fattore di vincolo al motivo-CGPD in quanto il suo legame è dipendente dalla sequenza ma indipendente dalla linea cellulare. L'espressione dipendente dalla linea cellulare di DHX30 potrebbe agire come repressore dell'efficienza di traduzione di mRNA contenenti il motivo in HCT116. Non si può però escludere che DHX30 sia in grado di legare il motivo indipendentemente da PCBP2. Si può osservare come il legame di queste due proteine è potenzialmente coinvolto nel controllo traduzionale di mRNA contenenti il motivo-CGPD.

2.4 Dati di partenza

I dati di partenza necessari per l'analisi di sbilanciamento allelico comprendono:

- I dati di RNA-seq per ogni condizione della linea cellulare divisi tra frazione allelica e totale.
- I dati di whole exome sequencing contenenti informazioni riguardo agli SNP presenti nella linea cellulare, divisi tra file GTF e VCF.

2.4.1 Dati di RNA-seq

Dove sono stati ottenuti i dati di RNA-seq.

2.4.2 Dati WES

Dove sono stati ottenuti i dati WES.

3 Processamento dei dati

Definito in §1 l’obiettivo di questo progetto si deve implementare una pipeline che, da dati di sequenziamento di RNA e da una lista di SNP esonici sia in grado di fornire le istanze di espressione allelo-specifica. Questa pipeline viene definita basandosi su quella presentata in [7] e si compone pertanto di tre fasi principali:

1. Pre-processamento e allineamento dei dati di RNA-seq, con eventuale deduplicazione e ricalibrazione.
2. Analisi di dati *WES* in modo da ottenere una lista di possibili SNP esonici da considerare.
3. Calcolo dei dati di sbilanciamento allelico.

La figura 3.1 è una visualizzazione della pipeline e dei tool utilizzati.

Essendo che questo lavoro prevede il processamento di un gran numero di file, unito al fatto che l’implementazione dei tool utilizzati permette lo sfruttamento delle pipe di unix e la possibilità di un’esecuzione in multi-threading è stato di fondamentale importanza l’utility parallel [8]. Questo tool permette di trasformare l’output di un programma in parametri per un successivo e consente di decidere quante istanze del secondo possono essere eseguite contemporaneamente.

In particolare nel caso di star (§3.1.2) è stata notata una diminuzione lineare delle performance per thread del tool. Parallel permette di mitigare questo problema in modo da, invece di eseguire un’unica istanza di star con molti thread, di averne diverse in contemporanea con meno thread. In questo modo aumenta il tempo di esecuzione dell’istanza singola, ma il tempo di esecuzione globale diminuisce.

Si nota pertanto che parallel permette non solo un’elegante implementazione della pipeline, ma fornisce un livello di controllo tale da sfruttare completamente la potenza computazionale disponibile, riducendo in questo modo il tempo globale di esecuzione.

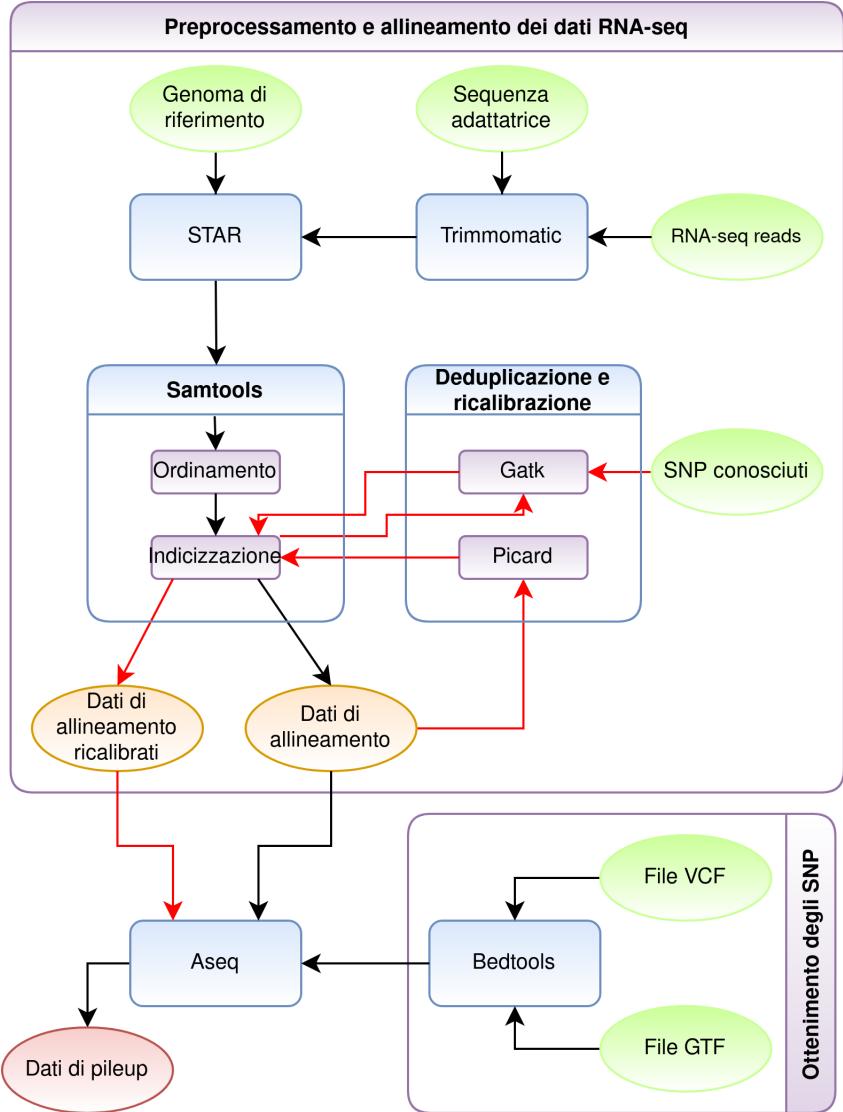


Figura 3.1: Pipeline per l'ottenimento dei dati di allelic imbalance

3.1 Pre-processamento e allineamento dei dati RNA-seq

Il primo passo della pipeline appena definita è il pre-processamento e l'allineamento dei dati RNA-seq. L'input della pipeline sono file di testo in formato fastq che contengono i dati di sequenziamento raccolti in laboratorio. Questi file subiscono un processo di preparazione in modo da ottenere dei file di allineamento utilizzabili per ottenere i dati di sbilanciamento allelico. Opzionalmente i file di allineamento possono subire ulteriori modifiche prima di passare alla prossima fase in modo da tentare di migliorare i risultati ottenuti. Tali modifiche sono in particolare deduplicazione e ricalibrazione.

3.1.1 Trimmomatic

Trimmomatic, presentato in [9], è il primo tool ad essere utilizzato. Il suo obiettivo è quello di eliminare dalle read la sequenza adattatrice, un artificio aggiunto al RNA per rendere possibile il sequenziamento, o suoi frammenti. Essendo le RNA-seq fornite single-ends¹ viene utilizzata la simple mode del tool. Questa scansione ogni read dalla terminazione 5' alla 3' per determinare la presenza della sequenza adattatrice. La scansione avviene attraverso il metodo "seed and extend" per trovare corrispondenze iniziali, anche non perfette, tra la read e la sequenza adattatrice. Successivamente svolge un allineamento locale a cui assegna uno score. Se lo score è maggiore di una soglia predefinita l'allineamento e la porzione che lo segue sono rimossi. Questa modalità permette di identificare ogni sequenza adattatrice in ogni luogo della read a patto che l'allineamento sia abbastanza lungo e la read

¹Punta a spiegazione nel capitolo 1

abbastanza accurata. Si nota però come nelle regioni dove solo una corta corrispondenza parziale è possibile, come alle estremità della read, i contaminanti non possono essere identificati attendibilmente. Oltre alla rimozione delle sequenze adattatrici Trimmomatic tronca un'estremità secondo un algoritmo di filtraggio secondo qualità. Tra i metodi forniti dallo strumento è stato utilizzato quello del “sliding window quality filtering”: scansiona la read dal 5' e rimuove la terminazione 3' quando la qualità media di un gruppo di basi scende sotto una soglia specificata. Il risultato di questo passaggio sarà un altro file fastq con le sequenze adattatrici rimosse.

Parametro	Valore
ILLUMINACLIP:	TruSeq3-SE.fa:2:30:10
phred	33
LEADING:	3
TRAILING:	3
SLIDINGWINDOW:	4 : 15
MINLEN:	36

Tabella 3.1: Parametri utilizzati per trimmomatic

3.1.2 Star

Star (spliced transcript alignment to a reference), presentato in [10], è il secondo tool della pipeline. Prende come input il file fastq generato da Trimmomatic e un genoma di riferimento. Allinea poi i dati di sequenziamento al riferimento in modo da determinare il luogo del genoma che ha originato le read.

Questo tool è stato creato con l'obiettivo di allineare RNA-seq di media-grande lunghezza, a differenza dei suoi competitori che, essendo creati a partire da allineatori per dati di DNA, hanno un maggiore tasso di errore. Star infatti tenta di risolvere problemi di allineamento dovuti agli eventi di splicing che avvengono durante la creazione delle molecole di mRNA. Per farlo deve tentare di creare allineamenti accurati di read contenenti mal-accoppiamenti, inserzioni o delezioni causati da variazioni genomiche o errori di sequenziamento. Lo fa mappando contemporaneamente sequenze derivate da regioni genomiche non contigue unite da eventi di splicing. Il processo di allineamento di star avviene in due fasi. La prima fase o *seed search* consiste della ricerca sequenziale di *Maximal Mappable Prefix MMP*. Data una sequenza R , una regione i e un genoma di riferimento G , $MMP(R, i, G)$ è la più lunga sottostringa $(R_i, R_{i+1}, \dots, R_{i+MML-1})$ che corrisponde esattamente a una o più sottostringhe di G . MML è la massima lunghezza mappabile. La seed search permette quindi un'identificazione di giunzioni di splicing senza nessuna conoscenza a priori. L'implementazione attraverso *Uncompressed suffix arrays* causa alla complessità dell'algoritmo di scalare logaritmicamente con la lunghezza del genoma di riferimento. Gli array sono non compressi per permettere tempi di ricerca più veloci, ma causano un aumento del consumo di memoria (circa 27GB per il genoma umano).

La seconda fase o *clustering, stitching and scoring* consiste nel costruire allineamenti dell'intera sequenza di read unendo tutti i seed allineati al genoma nella prima fase. Viene scelto un seed ancora a cui tutti gli altri sono raggruppati insieme: tutti quelli che si trovano al di sotto di una certa distanza formano una *genomic window*. Il seed ancora viene scelto limitando il numero di loci genomici a cui si allinea. Tutti i seed che sono stati mappati nella *genomic window* sono successivamente uniti insieme assumendo un modello lineare locale di trascrizione. Il processo di unione viene guidato da un sistema di punteggi che penalizza mal-accoppiamenti, inserzioni, delezioni e *splice junction gap*. Star ha come output un file sam (sequence alignment map) contenente le sequenze di input allineate rispetto al genoma di riferimento. La sezione di allineamento dell'output è formata da record contenenti campi separati da tabulazioni con informazioni sulla sequenza, su dove è stata allineata e sulla qualità dell'allineamento. Crea inoltre una stringa *CIGAR* utile a valle della pipeline.

Parametro	Valore
Genoma di riferimento	GRCh38
outSamstrandField	intronMotif
outSAMunmapped	None
outReadsUnmapped	fastx
outFilterScoreMinOverLread	0.33
outFilterMatchNminOverLread	0.33

Tabella 3.2: Parametri utilizzati per star

3.1.3 Samtools

I samtools [11] sono un insieme di programmi necessari per interagire con i dati di allineamento. Nella pipeline sono utilizzati per compiere delle operazioni sui file sam generati da star (§3.1.2) in modo da prepararli prima che possano essere utilizzati successivamente per generare i dati di sbilanciamento allelico (§3.3). In particolare svolgono le operazioni di ordinamento, indicizzazione e compressione dei file di input. Questo avviene attraverso due programmi: samtools sort [12] e samtools index [13]. Il primo ordina gli allineamenti secondo le coordinate di inizio e comprime implicitamente l'input in formato bam (binary alignment map). Il secondo invece crea, a partire dall'output del programma sort un indice in formato bai di tale file, permettendo efficienti operazioni di accesso casuale al file bam. L'output finale è un file bam ordinato e indicizzato che può essere utilizzato come input di aseq.

3.1.4 Deduplicazione e ricalibrazione

La deduplicazione e la ricalibrazione sono due processi di elaborazione dei dati di RNA-seq che tentano di risolvere errori presenti nei file bam che sono stati allineati attraverso star (§3.1.2). Questi due passaggi vengono svolti attraverso una serie di tools eseguiti sequenzialmente come si nota nella figura 3.2.

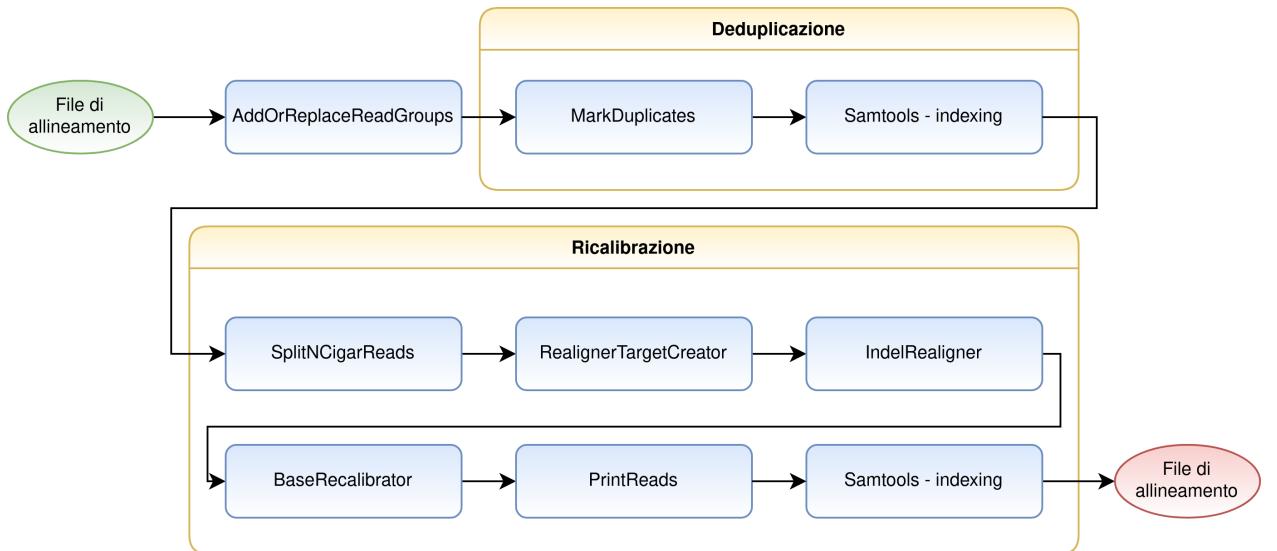


Figura 3.2: Pipeline di deduplicazione e ricalibrazione

Il primo passaggio viene svolto dal tool *AddOrReplaceReadGroups* [14] di Picard [15], un passaggio di preparazione delle read necessario a tutti i tool successivi. Questo assegna tutte le read in un file a un nuovo read-group settando un campo nel file bam, in modo da assegnare tutte le read a un genotipo specifico.

Parametro	Valore
RGID	1
RGLB	lib1
RGPL	ILLUMINA
RGPU	unit1
RGSM	20

Tabella 3.3: Parametri utilizzati per AddOrReplaceReadGroups

Deduplicazione

Il processo di deduplicazione e le motivazioni dietro alla sua utilità sono definite in [16]. Si definiscono come read duplicate in un file bam delle read che si generano da un singolo frammento di RNA. Possono originarsi durante la preparazione del campione, per esempio durante la costruzione della libreria attraverso PCR o risultare da un singolo cluster di amplificazione, identificato incorrettamente come cluster multipli dal sensore ottico dello strumento di sequenziamento.

Il processo di deduplicazione è stato svolto attraverso il tool *MarkDuplicates* di Picard. Il programma compara le sequenze nelle posizioni 5' sia delle read che dei read-pairs. Dopo che ha trovato tutte le read duplicate queste vengono ordinate secondo la somma dei punteggi di qualità delle basi. La read con il punteggio più alto viene considerata primaria, le altre duplicati. Grazie all'opzione *REMOVE_DUPLICATES=true* tutte le sequenze duplicate vengono rimosse. Viene infine ricreato l'indice del file bam attraverso *samtools index*.

Parametro	Valore
REMOVE_DUPLICATES	true
MAX_FILE_HANDLES_FOR_READ_ENDS_MAP	1000
VALIDATION_STRINGENCY	LENIENT
ASSUME_SORTED	true

Tabella 3.4: Parametri utilizzati per MarkDuplicates

Ricalibrazione

La ricalibrazione viene svolta da una serie di tool facenti parte della suite Gatk [17]. Il processo si compone di due fasi:

1. Riallineamento degli indels definito in [18].
2. Ricalibrazione delle basi basata sul punteggio di qualità definito in [19].

Il primo passaggio, svolto dal tool *SplitNCigarReads* [20], progettato specificatamente per dati di RNA-seq, è necessario per il corretto funzionamento degli step successivi. L'esecuzione di star ha generato nel file bam per ogni record una stringa *CIGAR* che descrive come una base in ogni read è mappata rispetto al genoma di riferimento. Il valore *N* corrisponde a una base saltata sul genoma di riferimento. Nel caso di RNA-seq tali basi possono corrispondere o a sequenze introniche, non presenti nel RNA a causa dello splicing o a sequenze di "overhang" che potrebbero portare a falsi positivi. *SplitNCigarReads* elimina le basi corrispondenti a una *N* da una read separandola in due: una che finisce a sinistra e l'altra che inizia a destra della base rimossa. Come risultato gli esoni del RNA vengono separati in segmenti diversi e gli overhang eliminati in modo da non causare falsi positivi. Dopo questo lavoro di processamento inizia la fase di riallineamento degli indels.

Parametro	Valore
R	Homo_Sapiens.GRCh38.dna.primary_assembly.fa
U	ALLOW_N_CIGAR_READS

Tabella 3.5: Parametri utilizzati per SplitNCigarReads

Riallineamento locale degli indel Il riallineamento locale degli indels è necessario in quanto permette di correggere errori sistematici causati dall'allineatore genomico. Una limitazione di questi allineatori infatti è che considerano ogni read in maniera indipendente: le strategie di assegnazione dei punteggi limitano la loro abilità di allineare accuratamente in presenza di indel. Il processo di allineamento locale considera invece tutte le read che attraversano una posizione in modo da ottenere un consenso di alto punteggio che supporta la presenza di un evento di indel. Il riallineamento viene svolto attraverso l'utilizzo di due tool: *RealignerTargetCreator* e *IndelRealigner*. Il primo prende come input un file bam ordinato e indicizzato e a partire da esso genera un file di output formato da una lista a una colonna contenente gli intervalli. Ogni record di questo file degli intervalli rappresenta un potenziale sito dove è avvenuto un indel. Infine se gli intervalli sono prossimali vengono uniti in un intervallo unico. Il secondo tool prende come input lo stesso bam di *RealignerTargetCreator* e il file di intervalli da esso generato e svolge un riallineamento locale sulle read coincidenti con l'intervallo target usando consensi dagli indel presenti nell'allineamento originale. L'output è un file bam ordinato e indicizzato.

Parametro	Valore
S	SILENT
R	Homo_Sapiens.GRCh38.dna.primary_assembly.fa
nt	10

Tabella 3.6: Parametri utilizzati per *RealignerTargetCreator*

Parametro	Valore
targetIntervals	file prodotto da <i>RealignerTargetCreator</i>
R	Homo_Sapiens.GRCh38.dna.primary_assembly.fa

Tabella 3.7: Parametri utilizzati per *IndelRealigner*

Ricalibrazione delle basi basata sul punteggio di qualità Il processo di ricalibrazione delle basi basato sul punteggio di qualità serve a risolvere errori generati durante il sequenziamento. Si definisce il punteggio di qualità di una base come una stima dell'errore emesso dalla macchina di sequenziamento: esprimono quanto questa è confidente che ha chiamato la base corretta ogni volta. Questi punteggi sono soggetti a varie sorgenti di errori tecnici dovuti alla fisica o alla chimica di come una reazione di sequenziamento funziona o a difetti nell'equipaggiamento. Gli errori portano pertanto a una sottostima o sovrastima (tipicamente la seconda) del punteggio di qualità fornito dal macchinario di sequenziamento. Per tentare di risolvere si utilizza un algoritmo di machine learning per modellare gli errori empiricamente in modo da modificare i valori di qualità per aumentarne la veridicità. La ricalibrazione delle basi avviene attraverso l'utilizzo di due tool.

Il primo è *BaseRecalibrator* [21] e costruisce un modello di covarianza da un file bam e da un insieme di varianti conosciute in un file vcf (variant calling format) e lo salva in un file. Le varianti conosciute sono usate per mascherare basi ai siti di variazioni aspettate in modo da non considerarle come errori. Al di fuori di queste eccezioni ogni mal-accoppiamento viene contato come un errore. Per costruire il modello di ricalibrazione il tool tabula i dati del file bam secondo il read group, il punteggio di qualità, il ciclo della macchina che ha prodotto la base, la base corrente e successiva. Successivamente si conta il numero di basi e quanto spesso queste hanno un mal-accoppiamento con la base di riferimento.

Il secondo tool *PrintReads* [22] applica il modello creato dal primo al file bam di input, aggiornandolo così ai punteggi di qualità migliorati. Infine viene ricreato l'indice del file bam attraverso *samtools index*.

In questo modo si è creato un file bam pronto per essere utilizzato da aseq.

Parametro	Valore
knownSites	lista di VCF contenente gli SNP conosciuti
R	Homo_Sapiens.GRCh38.dna.primary_assembly.fa
nct	10

Tabella 3.8: Parametri utilizzati per BaseRecalibrator

Parametro	Valore
nct	10
R	Homo_Sapiens.GRCh38.dna.primary_assembly.fa
BQSR	Modello output di BaseRecalibrator

Tabella 3.9: Parametri utilizzati per PrintReads

3.2 Ottenimento degli SNP di interesse

Una volta ottenuti i file di allineamento si rende necessario ottenere una lista di SNP dei quali si vuole ottenere il valore di sbilanciamento allelico. Per ottenere questi dati sono stati utilizzati il file gtf² contenente informazioni sulla porzione esonica del genoma umano e un insieme di vcf divisi per cromosoma contenenti le informazioni riguardo le varianti presenti negli esseri umani.

Dopo aver ristretto i vcf agli SNP sono stati intersecati con il file gtf. L'intersezione è stata voluta dal tool *bedtools intersect* [23], facente parte della suite bedtools [24]. Questo tool ha permesso di trovare ogni record del vcf presente anche nel file gtf. L'intersezione è stata svolta con diversi file gtf in modo da trovare quello che permettesse di conservare più informazioni. L'origine dei file gtf è descritta in §2.4. I risultati dell'intersezione sono descritti nella figura 3.3, è stato utilizzato pertanto utilizzato il file *Homo_sapiens.GRCh38.103.exonic.gtf*. In questo modo è stato creato un insieme di file vcf, uno per ogni cromosoma, contiene ogni SNP presente nella parte esonica del genoma.

La restrizione alla parte esonica del genoma non causa alcuna perdita di potere predittivo in quanto si stanno considerando dati di RNA-seq, ma permette una significativa diminuzione del carico computazionale svolto da aseq (§3.3). I VCF risultanti da questo processo sono pronti per essere utilizzati come input di aseq.

Parametro	Valore
wa	non richiesto
u	non richiesto
a	VCF file
b	GTF file

Tabella 3.10: Parametri utilizzati per bedtools intersect

²Sezione dati input

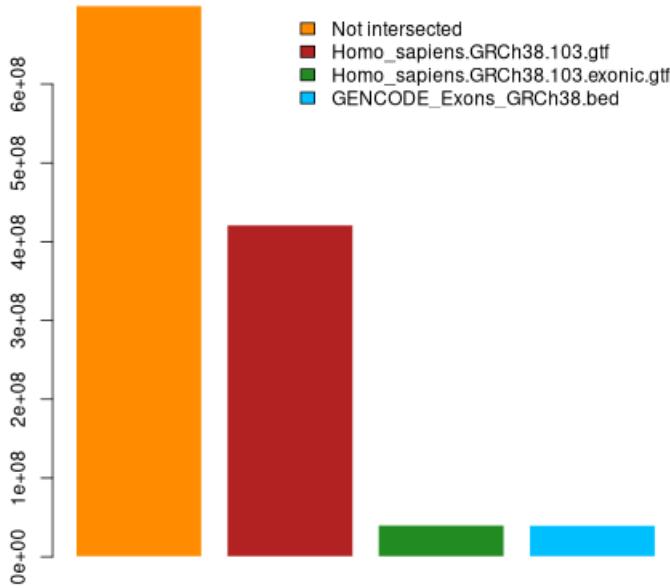


Figura 3.3: Numero di SNP mantenuti dopo l’intersezione

3.3 Calcolo dei dati di sbilanciamento allelico

Per computare i valori di sbilanciamento allelico è stato utilizzato aseq [25]. Questo tool tenta di risolvere alcune limitazioni di altri programmi di analisi di espressione allelo-specifica. Non richiede infatti informazioni genomiche dei genitori dell’individuo e non si basa unicamente sui dati di RNA-seq. Aseq in particolare accoppia dati di sequenziamento trascrittomici di nuova generazione e genomici in modo da superare queste limitazioni. La sua implementazione sfrutta le API di samtools, permettendo rapide funzionalità di accesso casuale su file di allineamento inidicizzati e ne aumenta il potere computazionale attraverso il multi-threading.

Delle modalità che aseq fornisce è stata utilizzata quella principale di analisi ASE (allelic-specific expression), ponendo limitazioni solo sulla qualità delle basi e la qualità delle letture dei file di allineamento. In questo modo aseq non restringe l’output agli eventi di ASE che individua ma ritorna il valore del pileup per ogni SNP datogli in input. Il pileup è un formato che riassume le chiamate delle basi rispetto a una sequenza di riferimento. Pertanto aseq computa per ogni SNP nella lista, a partire dal nome, dalla base canonica e dalla base alternativa, il coverage per tale SNP, la qualità della chiamata della base e la frazione allelica.

Questo permette di applicare all’output diversi filtri durante l’analisi cosicchè da poter trovare i valori soglia ottimali per ottenere risultati significativi in un secondo momento. In questo modo non si ripete continuamente la computazione dei valori di pileup snellendo il carico computazionale dell’analisi. Non si sfrutta pertanto il potere di individuazione di eventi di ASE di aseq, ma solo il suo veloce engine computazionale che permette di aumentare l’efficienza della creazione dei dati di pileup per una singola base. Ora, dopo aver prodotto i file di allineamento dai dati di RNA-seq dei campioni e la lista di SNP esonici di interesse si può procedere all’esecuzione di aseq.

In particolare i file di allineamento sono divisi in due insiemi: uno creato dopo l’allineamento con star (§3.1.2) e l’altro dopo il processo di deduplicazione e ricalibrazione (§3.1.4). Ognuno di questi insiemi possiede un file per campione³. La lista di SNP si trova invece in un insieme di file VCF, un file per ogni autosoma, uno per il cromosoma X e uno per il DNA mitocondriale. Perciò per ogni file di allineamento viene eseguito aseq per ogni file VCF, ottenendo in questo modo i dati di pileup di

³Riferimento a lista di campioni

ogni SNP per ogni campione divisi per cromosoma. L'output di aseq è un file di testo delimitato da tabulazioni, formato che rende facilmente ottimizzabili le successive analisi attraverso tool come awk, sed o framework come pandas per python o tidyverse per R.

Parametro	Valore
bam	file bam indicizzato prodotto precedentemente
vcf	file vcf ottenuto precedentemente
mbq	20
mrq	20

Tabella 3.11: Parametri utilizzati per aseq

4 Analisi dei dati

4.1 Conta degli SNP trovati con ASEQ

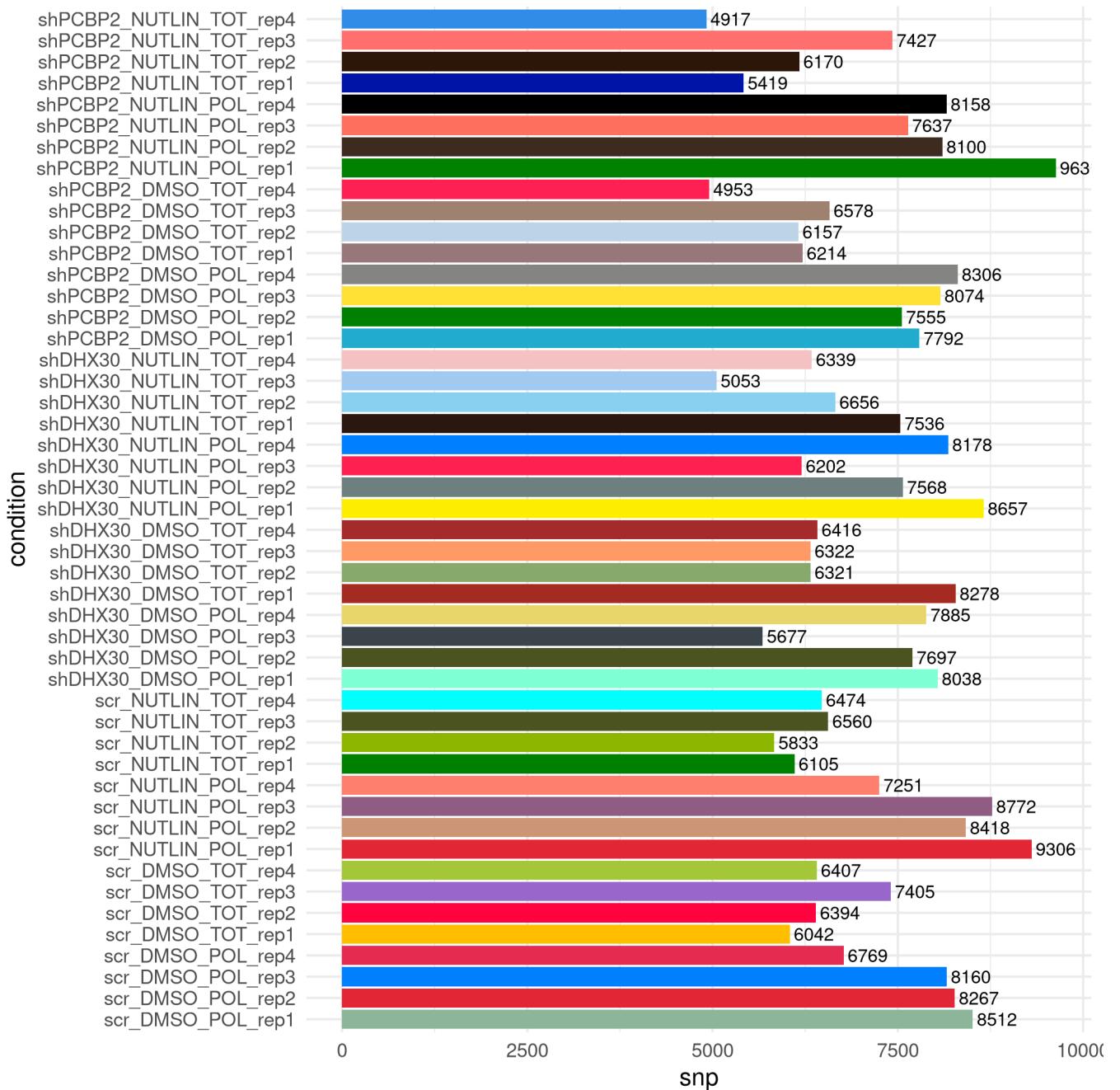
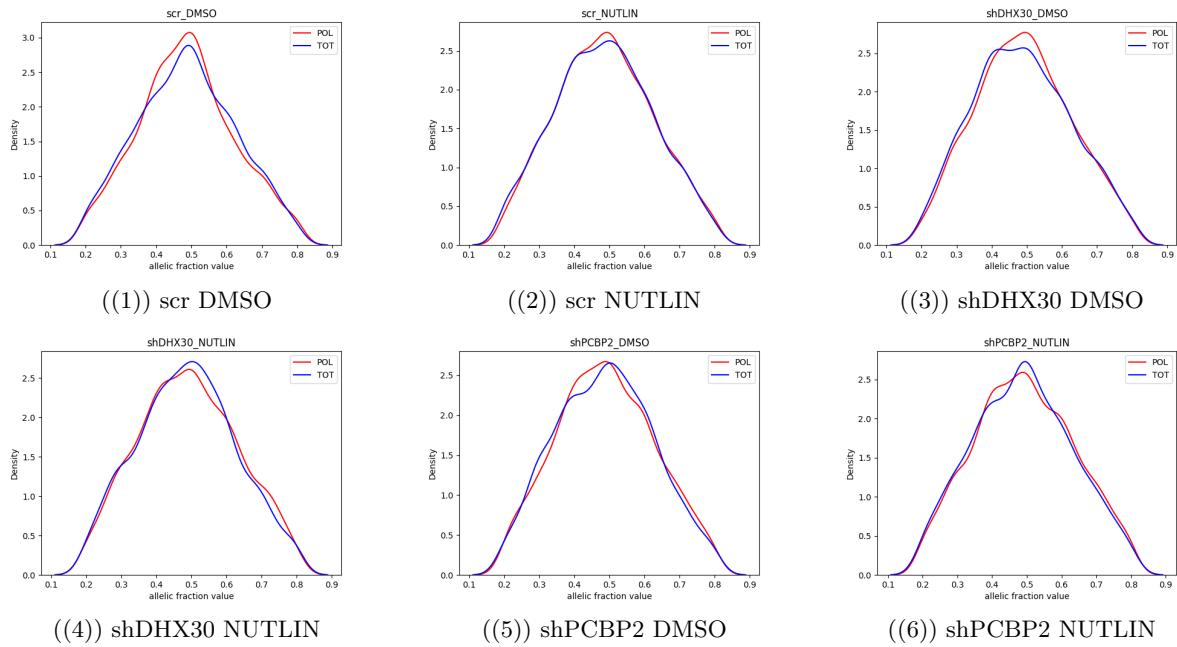


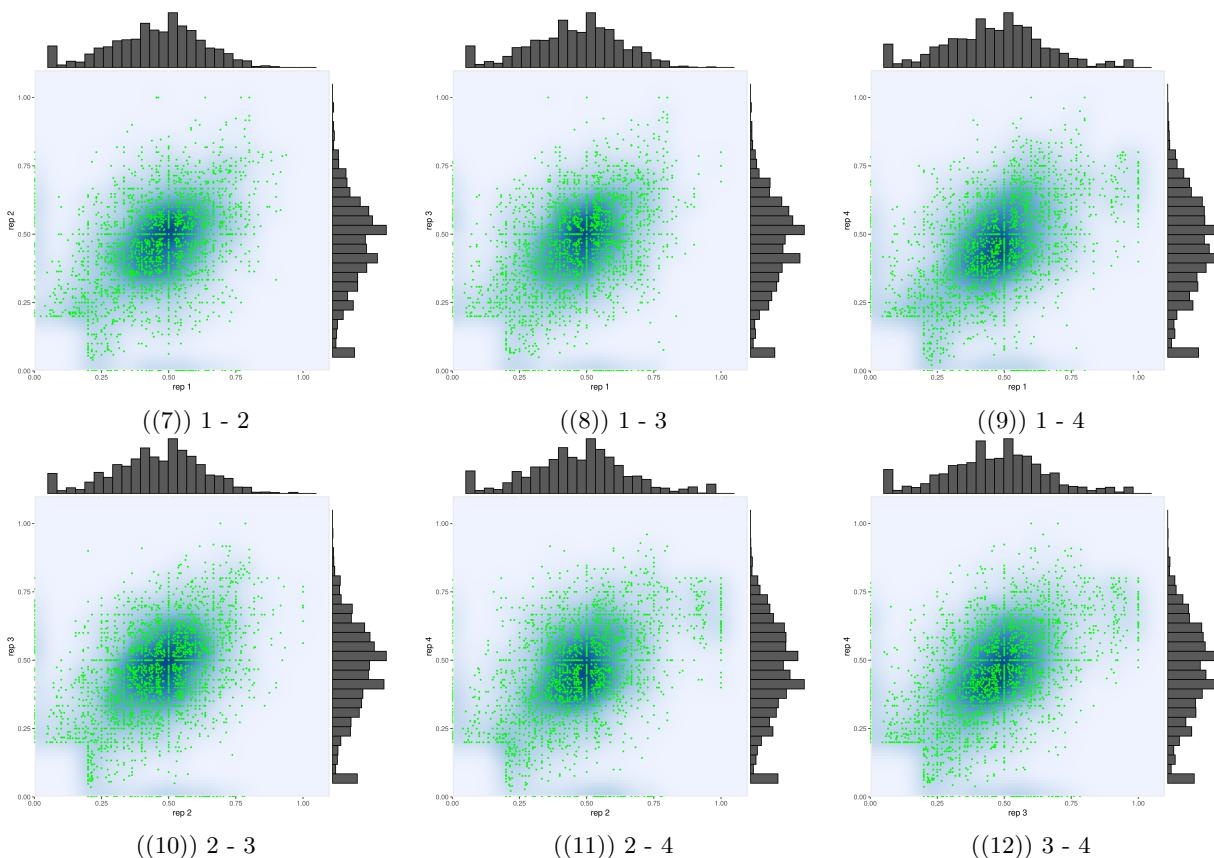
Figura 4.1: Conta per campione degli SNP eterozigoti trovati con aseq ($0.2 < af < 0.8$ e coverage ≥ 10)

4.1.1 Distrubuzione degli SNP



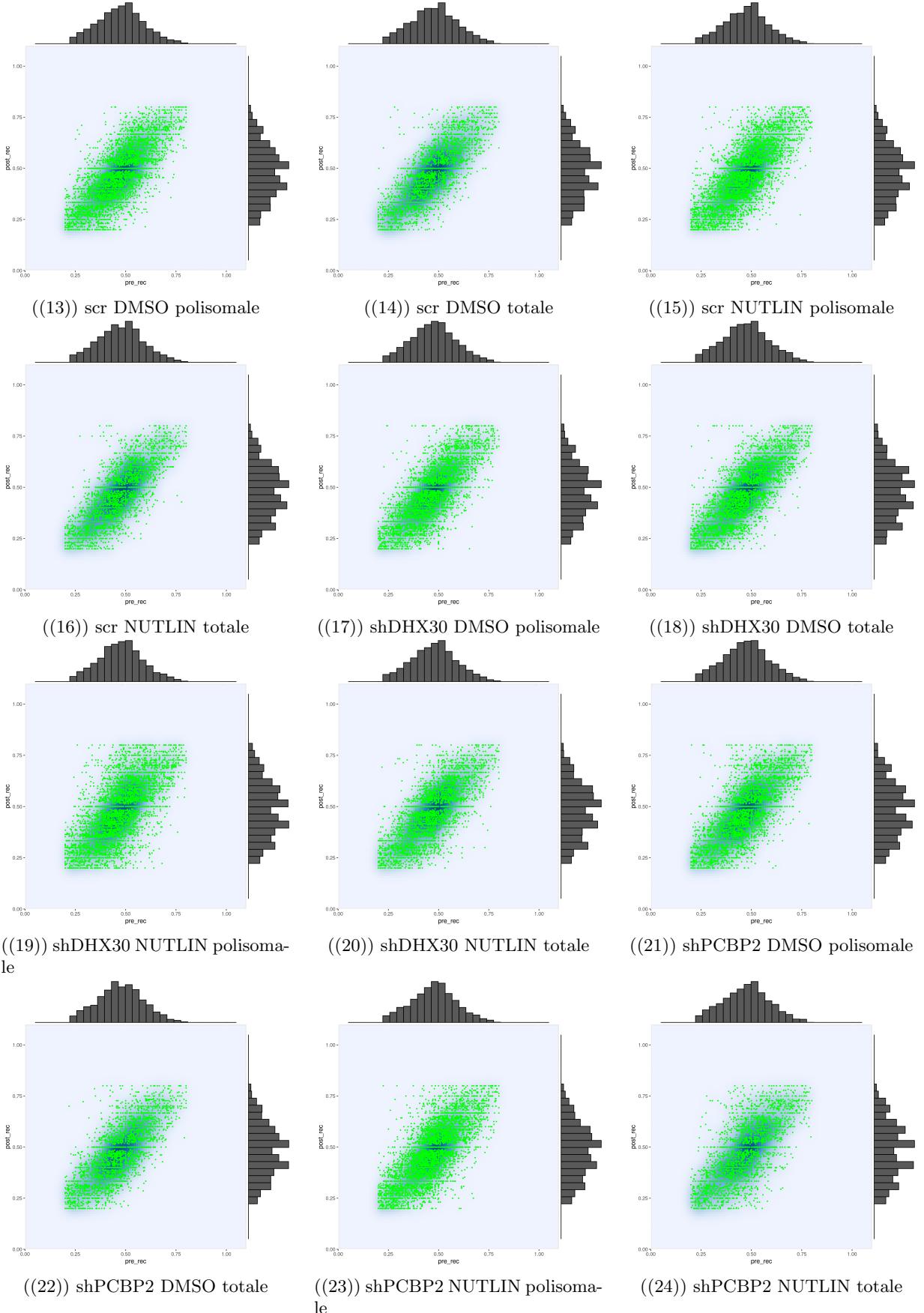
4.2 Qualità dei campioni

Confronto tra replicato di un campione in modo da verificare come cambiano i valori di AF tra un replicato e l'altro.



4.3 Considerazioni sulla recalibrazione

Discussione dei risultati di ASEQ prima e dopo la recalibrazione.



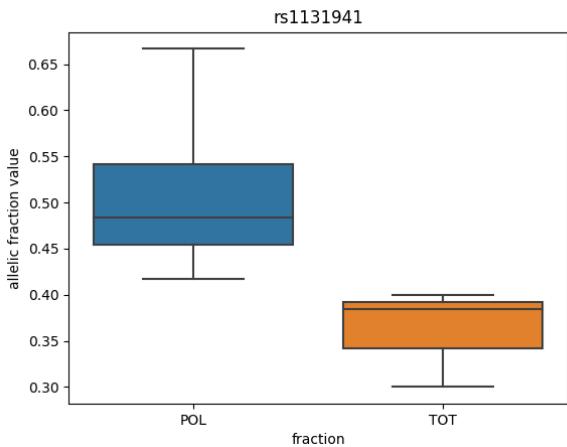
4.4 Ottenere i dati per gli SNP di interesse

Discussione degli SNP con i dati necessari per lo studio e scelta degli SNP di interesse. Ovvero come è stata ottenuta la lista da cellminer, i t-test. Magari qua posso mettere un barplot con la conta degli

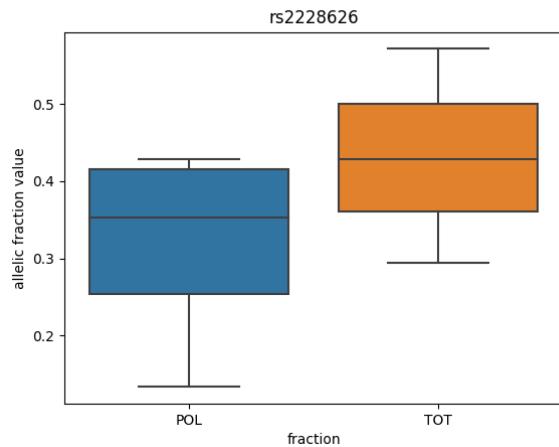
SNP trovati.

4.4.1 Boxplot

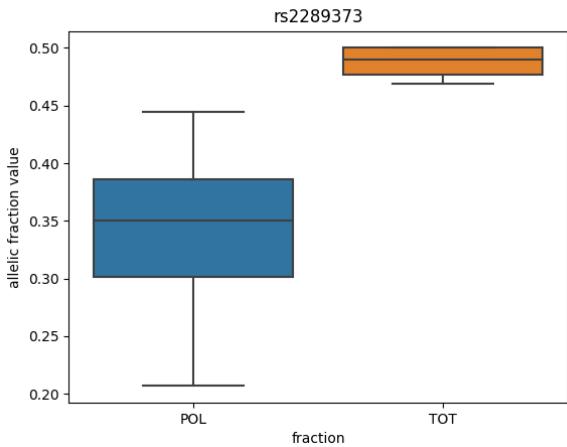
I boxplot degli SNP trovati.



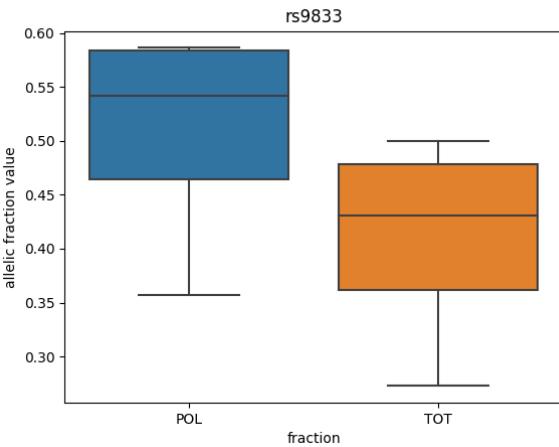
((25)) scr DMSO rs1131941



((26)) scr DMSO rs2228626



((27)) scr DMSO rs2289373



((28)) shPCBP2 NUTLIN rs9833

4.5 Caratterizzazione degli SNP di interesse

Magari una lista contenente gli SNP caratterizzati con snpEff

4.6 Conclusioni

Bibliografia

- [1] Caterina Marchioretti, Samuel Valentini, Alessandra Bisio, Annalisa Rossi, Alessandro Romanel, and Alberto Inga. TransSNP: a new class of functional SNPs that affect mRNA translation potential revealed by frac-seq-based allelic imbalance. *Submitted*, 2021.
- [2] Nathaniel Robichaud, Nahum Sonenberg, Davide Ruggero, and Robert J. Schneider. Translational control in cancer. *Cold Spring Harbor perspectives in biology*, 11(7):a032896, Jul 2019.
- [3] J. Kasteri, D. Das, X. Zhong, L. Persaud, A. Francis, H. Muhamam, and M. Sauane. Tranlation Control by p53. *Cancers (Basel)*, 10(5), May 2018.
- [4] Héloïse Chassé, Sandrine Boulben, Vlad Costache, Patrick Cormier, and Julia Morales. Analysis of translation using polysome profiling. *Nucleic acids research*, 45(3):e15–e15, Feb 2017.
- [5] M. G. Brattain, W. D. Fine, F. M. Khaled, J. Thompson, and D. E. Brattain. Heterogeneity of malignant cells from a human colonic carcinoma. *Cancer Res*, 41(5):1751–1756, May 1981.
- [6] Dario Rizzotto, Sara Zaccara, Annalisa Rossi, Matthew D. Galbraith, Zdenek Andrysik, Ahwan Pandey, Kelly D. Sullivan, Alessandro Quattrone, Joaquín M. Espinosa, Erik Dassi, and Alberto Inga. Nutlin-induced apoptosis is specified by a translation program regulated by pcbp2 and ddx30. *Cell Reports*, 30(13):4355–4369.e6, 2020.
- [7] Alessandro Romanel. Allele-specific expression analysis in cancer using next-generation sequencing data. *Krasnitz A. (eds) Cancer Bioinformatics. Methods in Molecular Biology*, 1878:125–137, 2019.
- [8] O. Tange. Gnu parallel - the command-line power tool. ;*login: The USENIX Magazine*, 36(1):42–47, Feb 2011.
- [9] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 04 2014.
- [10] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 10 2012.
- [11] Samtools. <http://www.htslib.org/>. ultimo accesso 01/08/2021.
- [12] Documentazione per samtools sort. <http://www.htslib.org/doc/samtools-sort.html>. ultimo accesso 01/08/2021.
- [13] Documentazione per samtools index. <http://www.htslib.org/doc/samtools-index.html>. ultimo accesso 01/08/2021.
- [14] Addorreplacegroups. <https://gatk.broadinstitute.org/hc/en-us/articles/360057440331-AddOrReplaceReadGroups-Picard->. ultimo accesso 06/08/2021.
- [15] Picard. <https://broadinstitute.github.io/picard/>. ultimo accesso 01/08/2021.

- [16] Markduplicates. <https://gatk.broadinstitute.org/hc/en-us/articles/360057439771-MarkDuplicates-Picard->. ultimo accesso 06/08/2021.
- [17] Gatk. <https://gatk.broadinstitute.org>. ultimo accesso 01/08/2021.
- [18] Indel realignment. [https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/\(howto\)_Perform_local_realignment_around_indels.md](https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/(howto)_Perform_local_realignment_around_indels.md). ultimo accesso 06/08/2021.
- [19] Base quality score recalibration. <https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-Recalibration-BQSR->. ultimo accesso 06/08/2021.
- [20] Splitncigarreads. <https://gatk.broadinstitute.org/hc/en-us/articles/360036734471-SplitNCigarReads>. ultimo accesso 06/08/2021.
- [21] Baserecalibrator. <https://gatk.broadinstitute.org/hc/en-us/articles/360036898312-BaseRecalibrator>. ultimo accesso 06/08/2021.
- [22] Printreads. <https://gatk.broadinstitute.org/hc/en-us/articles/360036883571-PrintReads>. ultimo accesso 06/08/2021.
- [23] Bedtools intersect. <https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html>. ultimo accesso 06/08/2021.
- [24] Bedtools. <https://bedtools.readthedocs.io/en/latest/>. ultimo accesso 06/08/2021.
- [25] Alessandro Romanel, Sara Lago, D. Prandi, A. Sboner, and F. Demichelis. Aseq: fast allele-specific studies from next-generation sequencing data. *BMC Medical Genomics*, 8, 2015.

Allegato A Titolo primo allegato

A.1 Titolo

A.1.1 Sottotitolo

Allegato B Titolo secondo allegato

B.1 Titolo

B.1.1 Sottotitolo