



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 Polimorfismi a singolo nucleotide	3
1.1.1 Espressione genica allelo-specifica	3
1.2 TransSNP	4
1.2.1 Cambi nelle regioni UTR nelle cellule di cancro	4
1.2.2 Profilamento polisomico	5
1.2.3 Sequenziamento	5
2 Linea cellulare e dati di partenza	6
2.1 HCT116	6
2.1.1 DHX30	6
2.1.2 PCBP2	6
2.2 Trattamenti	6
2.2.1 DMSO	6
2.2.2 Nutlin	6
2.3 Dati di partenza	6
2.3.1 Dati di RNA-seq	6
2.3.2 Dati WES	6
3 Processamento dei dati	7
3.1 Pre-processamento e allineamento dei dati RNA-seq	8
3.1.1 Trimmomatic	8
3.1.2 Star	9
3.1.3 Samtools	10
3.1.4 Deduplicazione e ricalibrazione	10
3.2 Ottenimento degli SNP di interesse	13
3.3 Calcolo dei dati di sbilanciamento allelico	13
4 Analisi dei dati	15
4.1 Identificazione delle varianti	15
4.2 Conta degli SNP trovati con ASEQ	16
4.2.1 Distribuzione degli SNP	17
4.3 Qualità dei campioni	17
4.4 Considerazioni sulla recalibrazione	17
4.5 Ottenere i dati per gli SNP di interesse	18
4.5.1 Boxplot	19
4.6 Analisi degli sbilanciamenti di frazione allelica	19
4.7 Conclusioni	19
Bibliografia	19

A Titolo primo allegato	22
A.1 Titolo	22
A.1.1 Sottotitolo	22
B Titolo secondo allegato	23
B.1 Titolo	23
B.1.1 Sottotitolo	23

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 e perciò essere una fonte potenziale di variazione inter-individuale nel rischio di cancro³.

1.1 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.

1.1.1 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.

¹Riferimento a linea cellulare

²Ce lo metto? Devo citare in qualche modo?

³Magari qua ci va un termine tecnico?

1.2 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 traduzionale. 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. Si nota infatti come la regolazione traduzionale 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'altro⁴. In questo modo si crea un catalogo di SNP codificantili e negli UTR associati e cause potenziali con alterazioni nel potenziale di traduzione degli mRNA.

1.2.1 Cambi nelle regioni UTR nelle cellule di cancro

Come evidenziato in [2] sequenze e motivi strutturali presenti negli mRNA determinano una 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 codificantili 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 questi motivi di sequenza o strutturali in alcune delle 5' UTR mediano 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.

Si nota pertanto come nelle regioni non codificantili degli mRNA si trovano motivi e sequenze di fondamentale importanza per il benessere della cellula e mutazioni come gli SNP possono andare a distruggerli o ad intacciarne l'efficacia, rendendola pertanto più prona a trasformarsi in una cellula di cancro.

⁴Non riesco a spiegare bene il concetto

1.2.2 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 [3]. 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 provetta⁵. 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 parallelo⁶.

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.

1.2.3 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.

⁵Provetta è il termine giusto? Nel paper si parla di “tube”

⁶Sto prendendo dal paper, magari per la mia linea cellulare queste cose sono state fatte in modo diverso

2 Linea cellulare e dati di partenza

Descrizione delle linee cellulari e dei materiali utilizzati.

2.1 HCT116

Descrizione della linea cellulare e motivazione del suo utilizzo.

2.1.1 DHX30

Funzione di DHX30, cosa ne è stato fatto nei campioni e motivazioni dietro alla scelta.

2.1.2 PCBP2

Funzione di PCBP2, cosa ne è stato fatto nei campioni e motivazioni dietro alla scelta.

2.2 Trattamenti

2.2.1 DMSO

Descrizione del trattamento e motivazioni.

2.2.2 Nutlin

Descrizione del trattamento e motivazioni.

2.3 Dati di partenza

2.3.1 Dati di RNA-seq

2.3.2 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 [4] 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 ?? 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 [5]. 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.

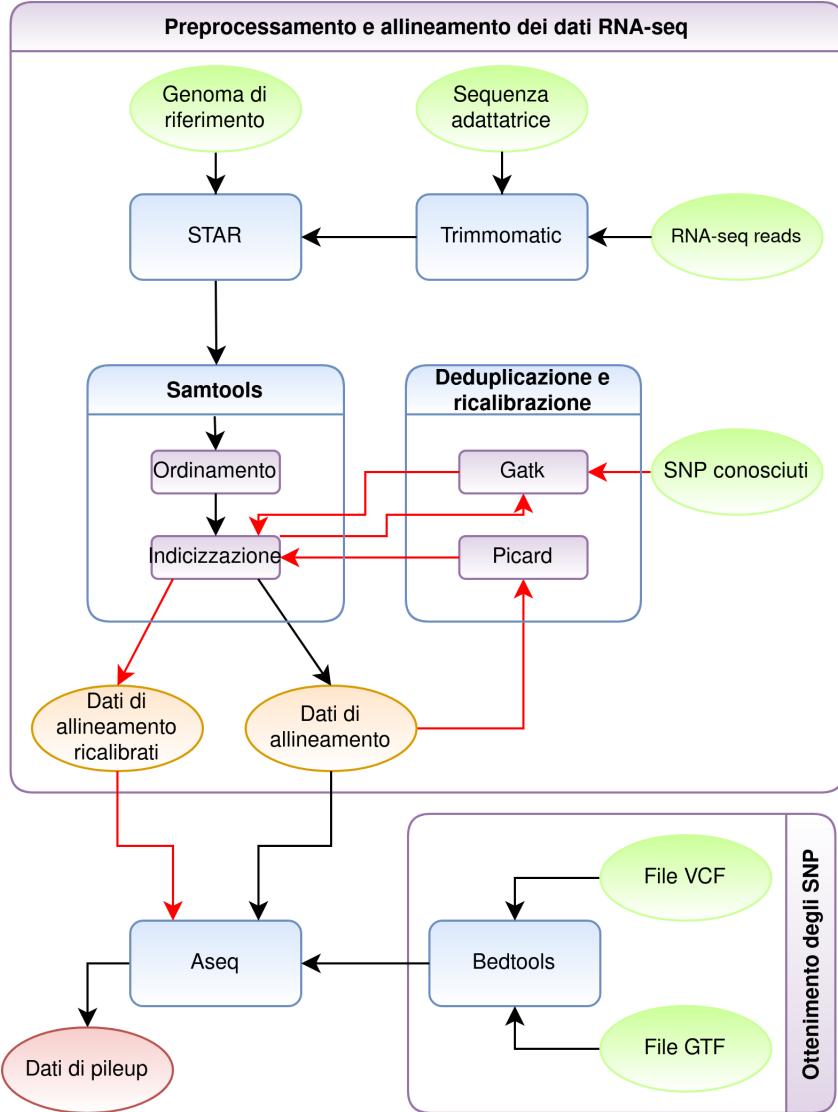


Figura 3.1: Pipeline per l’ottenimento dei dati di allelic imbalance

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 [6], 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 scansiona 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 [7], 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 [8] 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 [9] e samtools index [10]. 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.

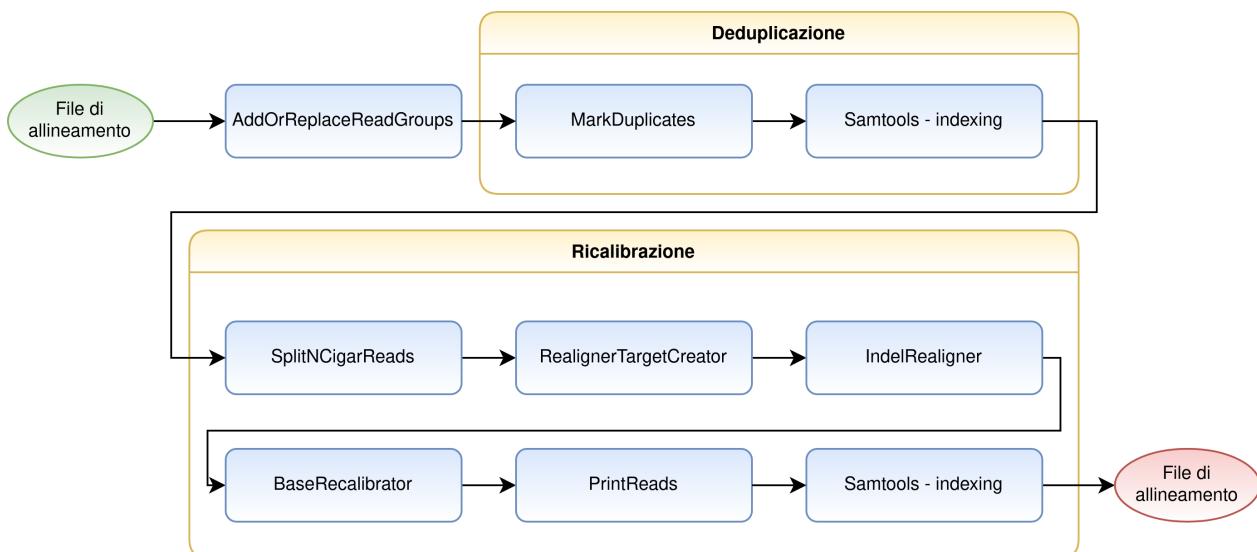


Figura 3.2: Pipeline di deduplicazione e ricalibrazione

Il primo passaggio viene svolto dal tool *AddOrReplaceReadGroups* [11] di Picard [12], 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 [13]. 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_DUPLICATE=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 [14]. Il processo si compone di due fasi:

1. Riallineamento degli indels definito in [15].
2. Ricalibrazione delle basi basata sul punteggio di qualità definito in [16].

Il primo passaggio, svolto dal tool *SplitNCigarReads* [17], 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 sequence 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 bamordinato 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* [18] 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* [19] 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* [20], facente parte della suite bedtools [21]. Questo tool ha permesso di trovare ogni record del vcf presente anche nel file gtf. In questo modo è stato creato un insieme di file vcf, uno per ogni cromosoma, che 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

3.3 Calcolo dei dati di sbilanciamento allelico

Per computare i valori di sbilanciamento allelico è stato utilizzato aseq [22]. 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 (allele-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

²Sezione dati input

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 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

³Riferimento a lista di campioni

4 Analisi dei dati

4.1 Identificazione delle varianti

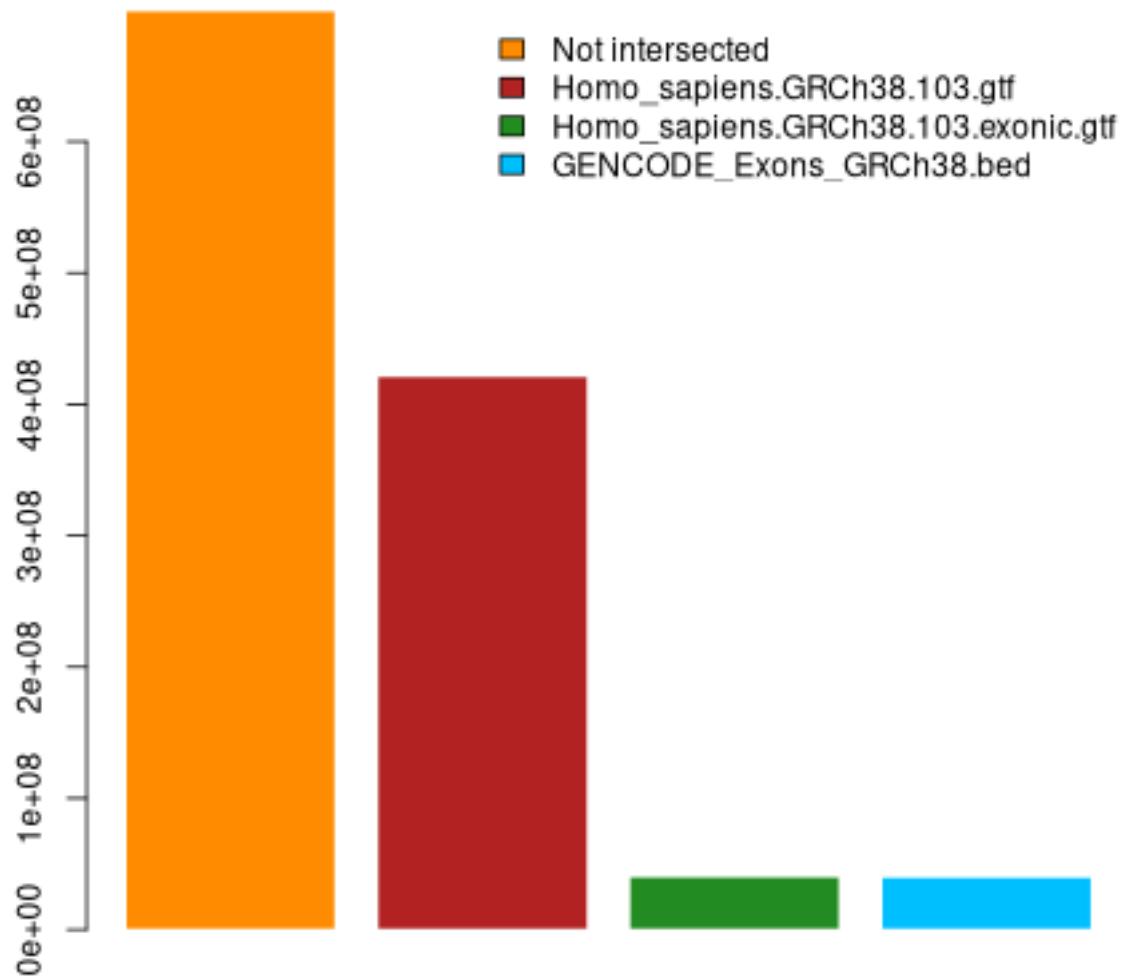


Figura 4.1: SNP mantenuti dopo l'intersezione con i VCF

4.2 Conta degli SNP trovati con ASEQ

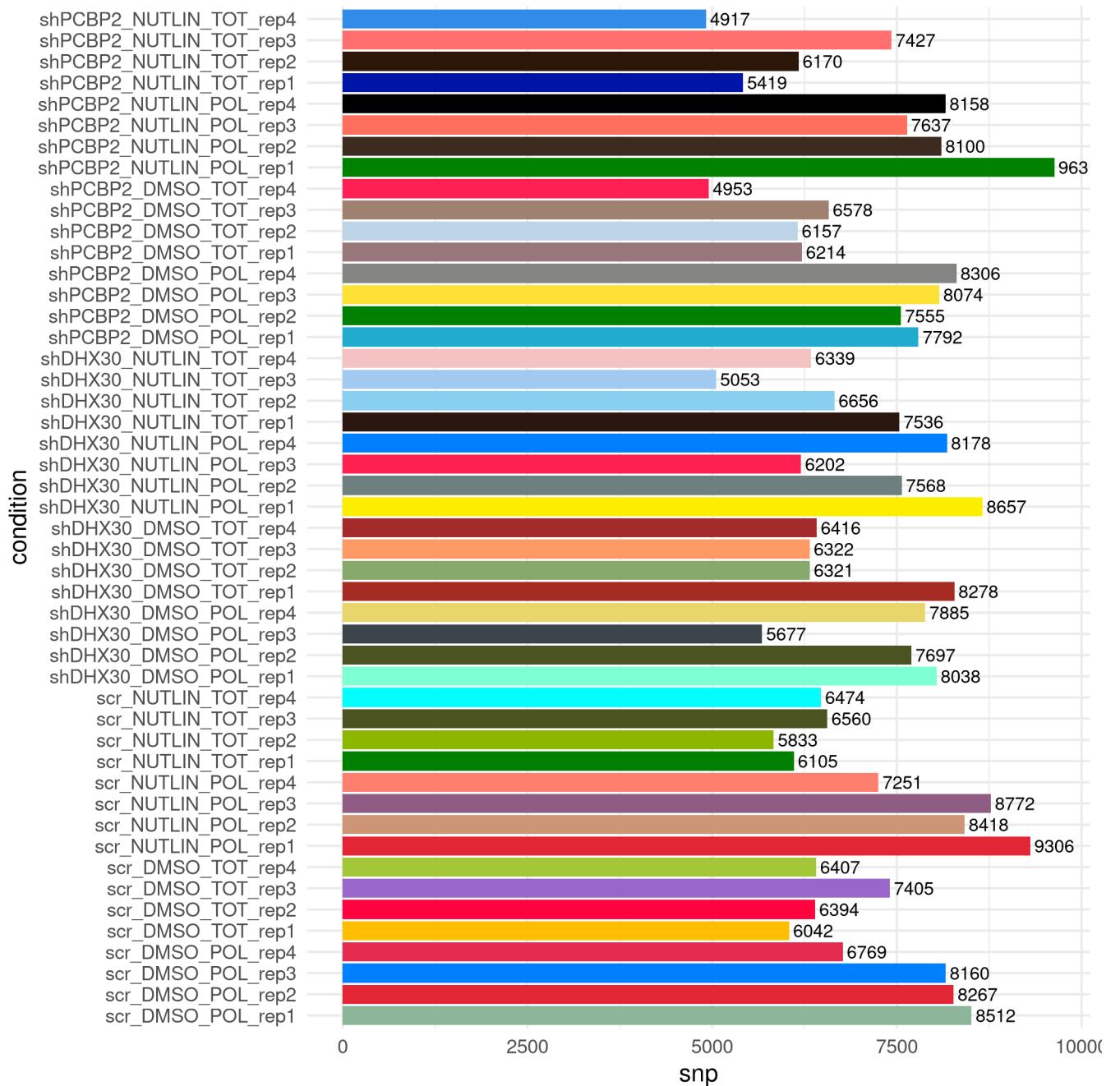
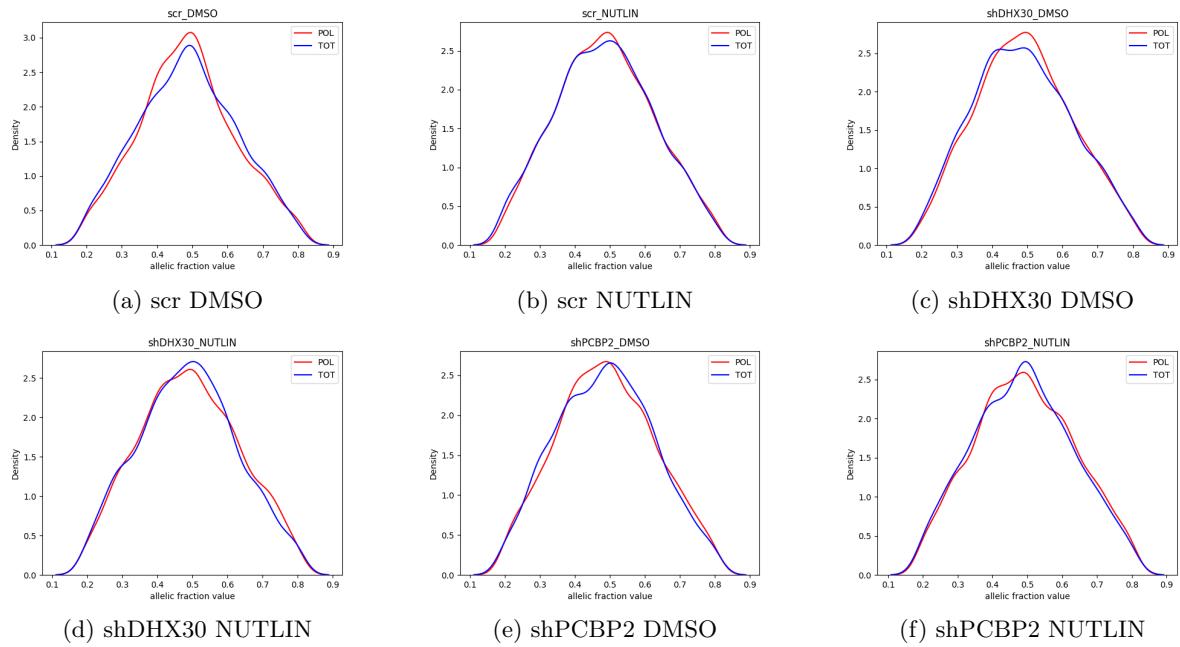


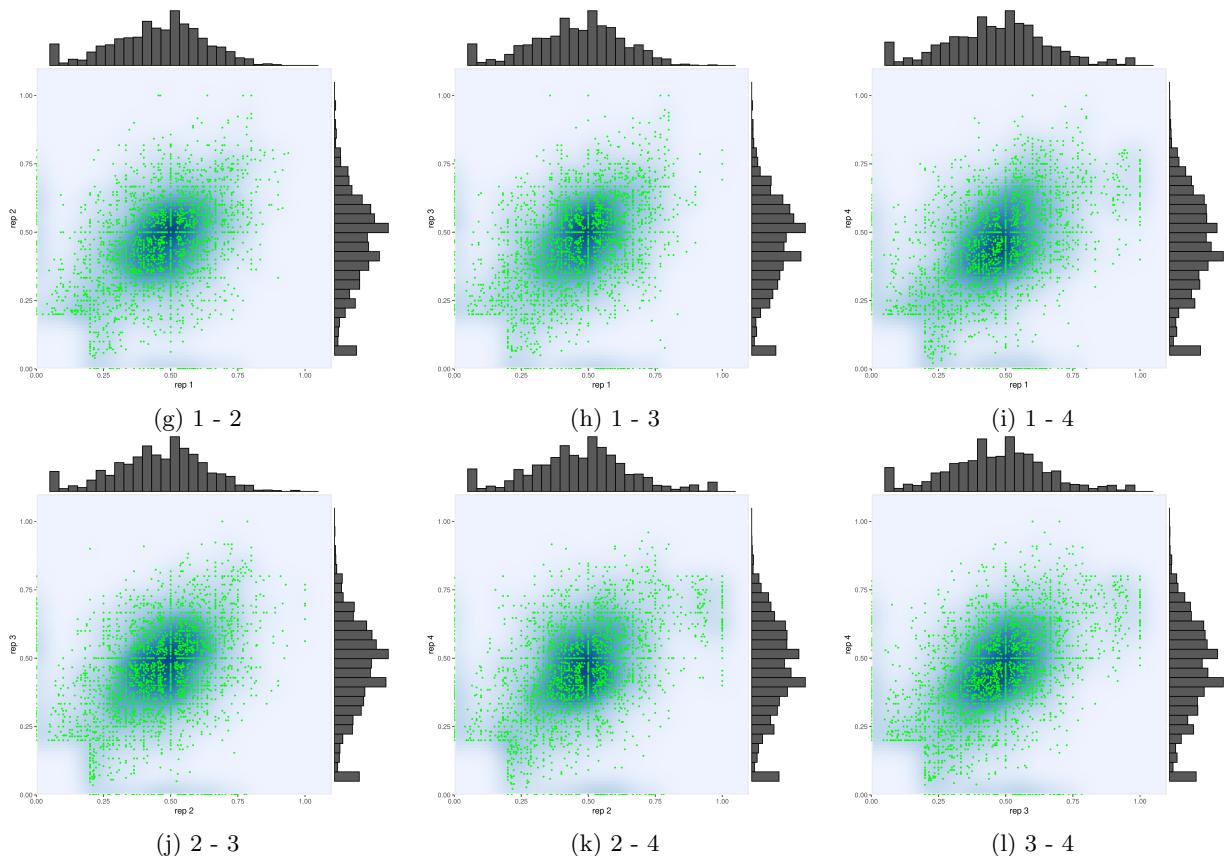
Figura 4.2: Conta per campione degli SNP eterozigoti trovati con aseq ($0.2 < af < 0.8$ e coverage ≥ 10)

4.2.1 Distribuzione degli SNP



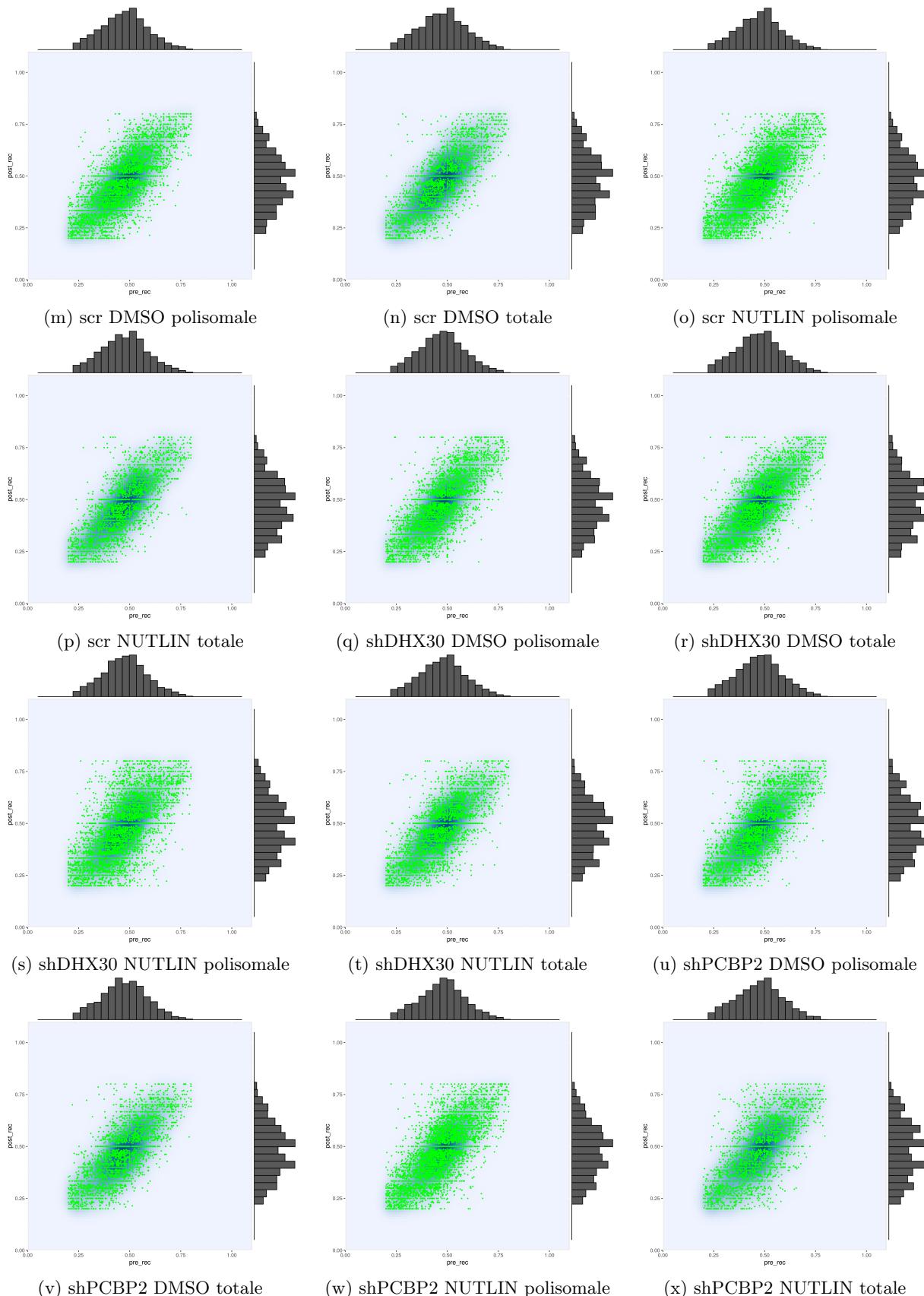
4.3 Qualità dei campioni

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



4.4 Considerazioni sulla recalibrazione

Discussione dei risultati di ASEQ prima e dopo la recalibrazione.



4.5 Ottenere i dati per gli SNP di interesse

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

4.5.1 Boxplot

I boxplot degli SNP trovati.

4.6 Analisi degli sbilanciamenti di frazione allelica

Analisi finali.

4.7 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] 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.
- [4] 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.
- [5] O. Tange. Gnu parallel - the command-line power tool. ;*login: The USENIX Magazine*, 36(1):42–47, Feb 2011.
- [6] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 04 2014.
- [7] 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.
- [8] Samtools. <http://www.htslib.org/>. ultimo accesso 01/08/2021.
- [9] Documentazione per samtools sort. <http://www.htslib.org/doc/samtools-sort.html>. ultimo accesso 01/08/2021.
- [10] Documentazione per samtools index. <http://www.htslib.org/doc/samtools-index.html>. ultimo accesso 01/08/2021.
- [11] Addorreplacegroups. <https://gatk.broadinstitute.org/hc/en-us/articles/360057440331-AddOrReplaceReadGroups-Picard->. ultimo accesso 06/08/2021.
- [12] Picard. <https://broadinstitute.github.io/picard/>. ultimo accesso 01/08/2021.
- [13] Markduplicates. <https://gatk.broadinstitute.org/hc/en-us/articles/360057439771-MarkDuplicates-Picard->. ultimo accesso 06/08/2021.
- [14] Gatk. <https://gatk.broadinstitute.org>. ultimo accesso 01/08/2021.
- [15] 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.
- [16] Base quality score recalibration. <https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-Recalibration-BQSR->. ultimo accesso 06/08/2021.

- [17] Splitncigarreads. <https://gatk.broadinstitute.org/hc/en-us/articles/360036734471-SplitNCigarReads>. ultimo accesso 06/08/2021.
- [18] Baserecalibrator. <https://gatk.broadinstitute.org/hc/en-us/articles/360036898312-BaseRecalibrator>. ultimo accesso 06/08/2021.
- [19] Printreads. <https://gatk.broadinstitute.org/hc/en-us/articles/360036883571-PrintReads>. ultimo accesso 06/08/2021.
- [20] Bedtools intersect. <https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html>. ultimo accesso 06/08/2021.
- [21] Bedtools. <https://bedtools.readthedocs.io/en/latest/>. ultimo accesso 06/08/2021.
- [22] 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