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

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

Questo capitolo è volto a descrivere i processi biologici considerati durante il progetto. Cito principalmente dal draft paper sui transSNPS

1.1 TransSNPs

Definizione di SNP e loro impatto. Descrizione degli SNP considerati in questo esperimento.

1.2 Sbilanciamento allelico

Definizione di sbilanciamento allelico e perchè viene considerato.

1.2.1 Profilamento polisomico

Come si identifica lo sbilanciamento allelico.

2 Linee cellulari e condizioni

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.

3 Processamento dei dati

Lo scopo dell'analisi è quello di individuare SNP che causano un cambio di potenziale di traduzione agli mRNA che li contengono. In quanto gli eventi regolatori del processo di traduzione avvengono nella porzione *3'-UTR* del mRNA ci si concentra sull'analisi di SNP presenti in tale luogo. Come dettagliato in [**?**] per per superare il rumore intrinseco nella chiamata di SNP e per la coverage dei dati di RNA-sequencing viene progettata una pipeline di analisi comparativa di sbilanciamento allelico tra mRNA totali e polisomiali estratti e sequenziati a partire dallo stesso campione cellulare. Questo inoltre viene indipendentemente genotipizzato per SNP eterozigoti Il processamento dei dati si configura come una pipeline di analisi di espressione allelo specifica nel cancro. Ci si basa pertanto sul lavoro sviluppato in [1], adattandolo in modo da riuscire ad evidenziare un potenziale traduzionale differenziale allelo-specifico. Si parte pertanto da letture di sequenziamento del trascrittoma della linea cellulare *TODO inserisci riferimento a sezione con spiegazione linea HCT116* con lo scopo ultimo di caratterizzare SNP che causano il cambio di potenziale di traduzione e che potrebbero essere pertanto coinvolti nel cancro.

3.1 Pipeline

La pipeline pertanto si definisce di n fasi:

1. Preprocessamento e allineamento dei dati di RNA-sequencing.
2. Identificazione di SNP informativi dai dati *WES*.
3. Analisi di espressione allelo-specifica.
4. Identificazione di livelli di traduzione differenziale di mRNA contenenti gli SNP identificati.

Il processo computazionale viene descritto visivamente in *Metti immagine*

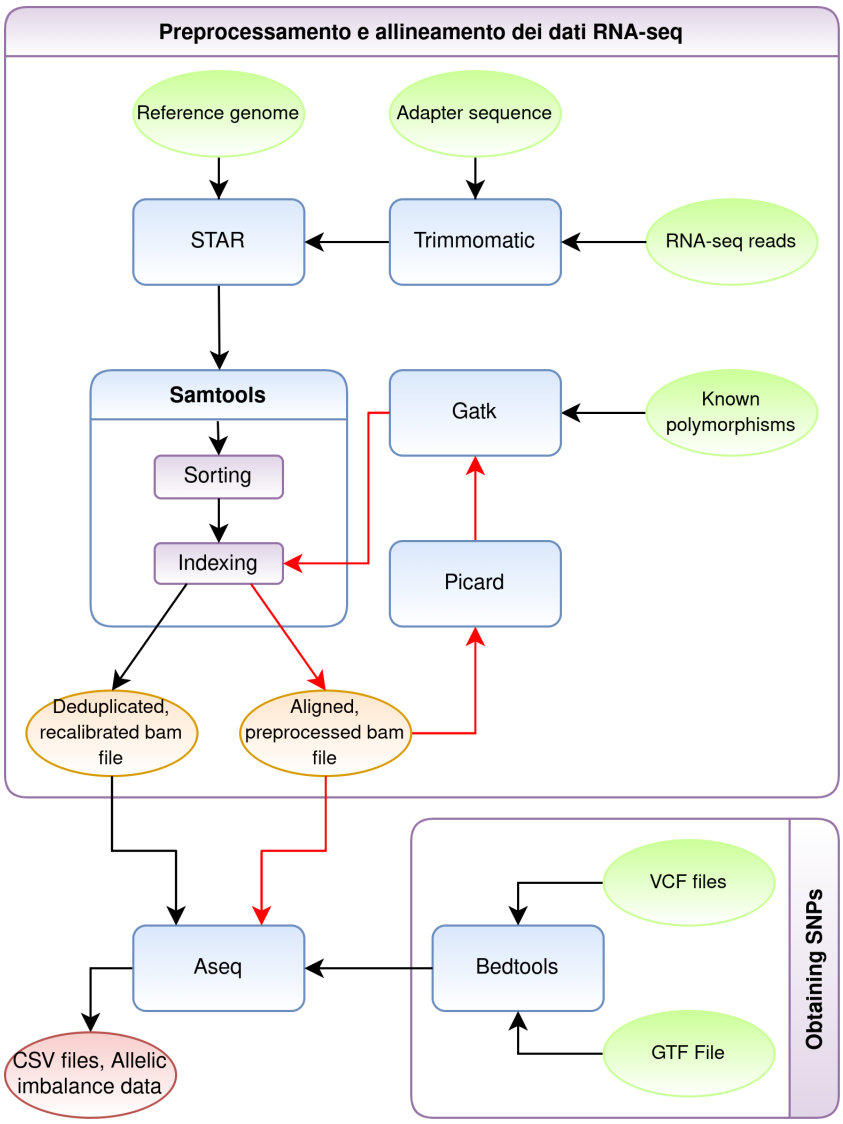

Figura 3.1: Pipeline

3.2 Dati disponibili

3.2.1 Sequenze biologiche

Descrizione dei fastq.

3.2.2 Genoma di riferimento

Descrizione del genoma di riferimento.

3.2.3 Variant call

Descrizione dei vcf.

3.2.4 Struttura dei geni

Descrizione del gtf.

3.3 Troncatura e allinamento

Descrizione del processo e perchè viene fatto.

3.3.1 Troncatura

Trimmomatic, cosa fa come è stato usato.

3.3.2 Allineamento
STAR, cosa fa come è stato usato.

3.3.3 Ordinamento
SAMTOOLS SORT cosa fa come è stato usato.

3.3.4 Indicizzazione
SAMTOOLS index cosa fa come è stato usato.

3.4 Deduplicazione, riallinamento e recalibrazione
Descrizione del processo e perchè viene fatto

3.4.1 Deduplicazione
Come sopra.

3.4.2 Riallineamento e recalibrazione
Come sopra.

3.5 Ottenere le varianti alleliche
Intersezione tra VCF e GTF.

3.6 Ottenere i dati delle frazioni alleliche
ASEQ cosa fa come viene usato.

3.6.1 Filtrare le frazioni alleliche
Condizioni di filtraggio per i risultati di ASEQ.

3.7 Ottenere gli SNP nel 3'-UTR
Filtraggio del gtf e intersezione con i VCF

4 Analisi dei dati

4.1 Conta degli SNP trovati con ASEQ
Discussione dei risultati di ASEQ.

4.2 Considerazioni sulla recalibrazione
Discussione dei risultati di ASEQ prima e dopo la recalibrazione

4.3 Ottenere i dati per gli SNP di interesse
Discussione degli SNP con i dati necessari per lo studio e scelta degli SNP di interesse.

4.4 Analisi degli sbilanciamenti di frazione allelica
Analisi finali.

4.5 Conclusioni

Bibliografia

[1] Alessandro Romanel. Allele-specific expressione analysis in cancer using next-generation sequencing data. *Krasnitz A. (eds) Cancer Bioinformatics. Methods in Molecular Biology*, 1878:125–137, 2019.

Allegato A Titolo primo allegato

A.1 Titolo

A.1.1 Sottotitolo

Allegato B Titolo secondo allegato

B.1 Titolo

B.1.1 Sottotitolo