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

Definito in **CAPITOLO 1** l'obiettivo di questo progetto si deve pertanto definire una pipeline che, da dati di sequenziamento di RNA e di *WES* (**SEMPRE DEFINITO NEL CAPITOLO 1**) sia in grado di fornire le istanze di espressione allelo-specifica. Questa pipeline viene definita basandosi su quella presentata in [2] 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. Ottenimento dei dati di sbilanciamento allelico.

La figura 3 è una visualizzazione della pipeline e dei tool utilizzati.
Essendo infine che questo lavoro prevede il processamento di un gran numero di file, unito al fatto che i tool utilizzzati sono stati implementati con una integrazione con le pipe di unix e con la possibilità di essere eseguiti in multi-threading è stato di fondamentale importanza l'utility parallel [3]. In particolare nel caso di STAR[1] è stato notato una diminuzione lineare delle performance per thread del tool. L'abbassamento delle performance è stato risolto grazie a parallel, lanciando più istanze parallele del tool, ognuna di esse con pochi thread. 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.

[1] §3.3.2

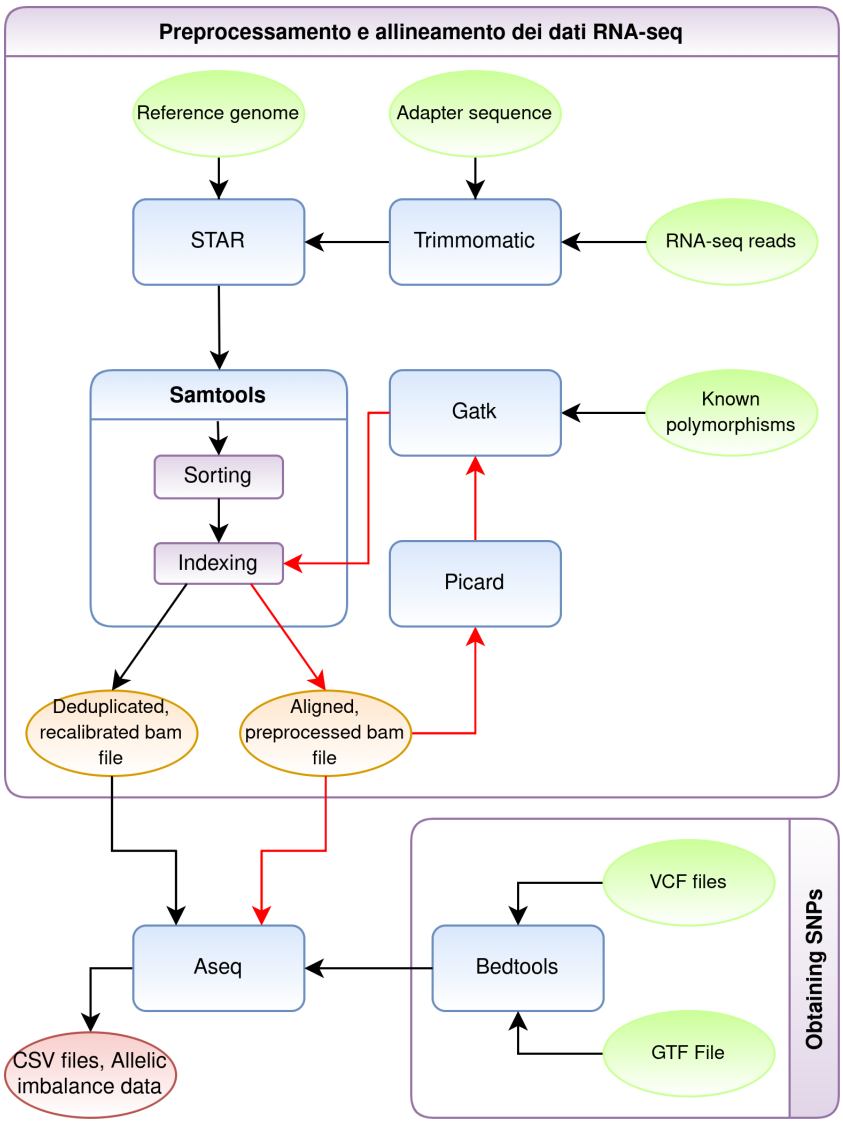

Figura 3.1: Pipeline

3.1 Pre-processamento e allineamento dei dati RNA-seq

Il primo passo della pipeline appenda definita è il pre-processamento e l'allineamento dei dati RNA-seq. L'input della pipeline sono dei file *FASTQ*, ovvero file di testo che contengono i dati di sequenziamento raccolti in laboratorio tipicamente compressi. Questi file vengono modificati dai tool Trimmomatic, Star e Samtools in modo da produrre un output già utilizzabile per ottenere i dati di sbilanciamento allelico. Opzionalmente l'output può subire ulteriori modifiche prima di passare alla prossima fase: in particolare può essere deduplicato e ricalibrato.

3.1.1 Trimmomatic

Trimmomatic [1] è il primo tool ad essere utilizzato. Il suo obiettivo è quello di eliminare dalle reads la sequenza adattatrice, o suoi frammenti, utilizzata per rendere possibile il sequenziamento. Essendo le RNA-seq fornite a single-ends[2] viene utilizzata la simple mode del tool. Questa scansiona ogni read[3] dalla terminazione $5'$ alla $3'$ per determinare la presenza della sequenza adattatrice. Utilizza il mdetodo "seed and extend" per trovare corrispondenze iniziali, anche non perfette, tra la read e la sequenza adattatrice. Successivamente svolge un allineamento locale e se questo ha uno score maggiore di una soglia viene rimosso insieme alla porzione successiva ad esso. Questa modalità permette di identificare ogni sequenza adattatrice in ogni luogo della read a patto che l'allineamento sia abbastanza lungo e la

[2]Spiega cosa sono
[3]Non so se metterla in italiano

read abbastanza accurata. Si nota però come nelle regioni dove solo una corta corrispondenza parziale è possibile come alle estremità della read e pertanto 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.

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] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 04 2014.

[2] 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.

[3] O. Tange. Gnu parallel - the command-line power tool. *;login: The USENIX Magazine*, 36(1):42–47, Feb 2011.

Allegato A Titolo primo allegato

A.1 Titolo

A.1.1 Sottotitolo

Allegato B Titolo secondo allegato

B.1 Titolo

B.1.1 Sottotitolo