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 §1 l'obiettivo di questo progetto si deve definire 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 [1] 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 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 pemette lo sfruttamento delle pipe di unix e la possibilità di un esecuzione in multi-threading è stato di fondamentale importanza l'utility parallel [2]. 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 moltissimi thread, di averne diverse in contemporanea con meno thread. In questo modo aumenta il tempo di esecuzione di un'istanza di STAR, 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.

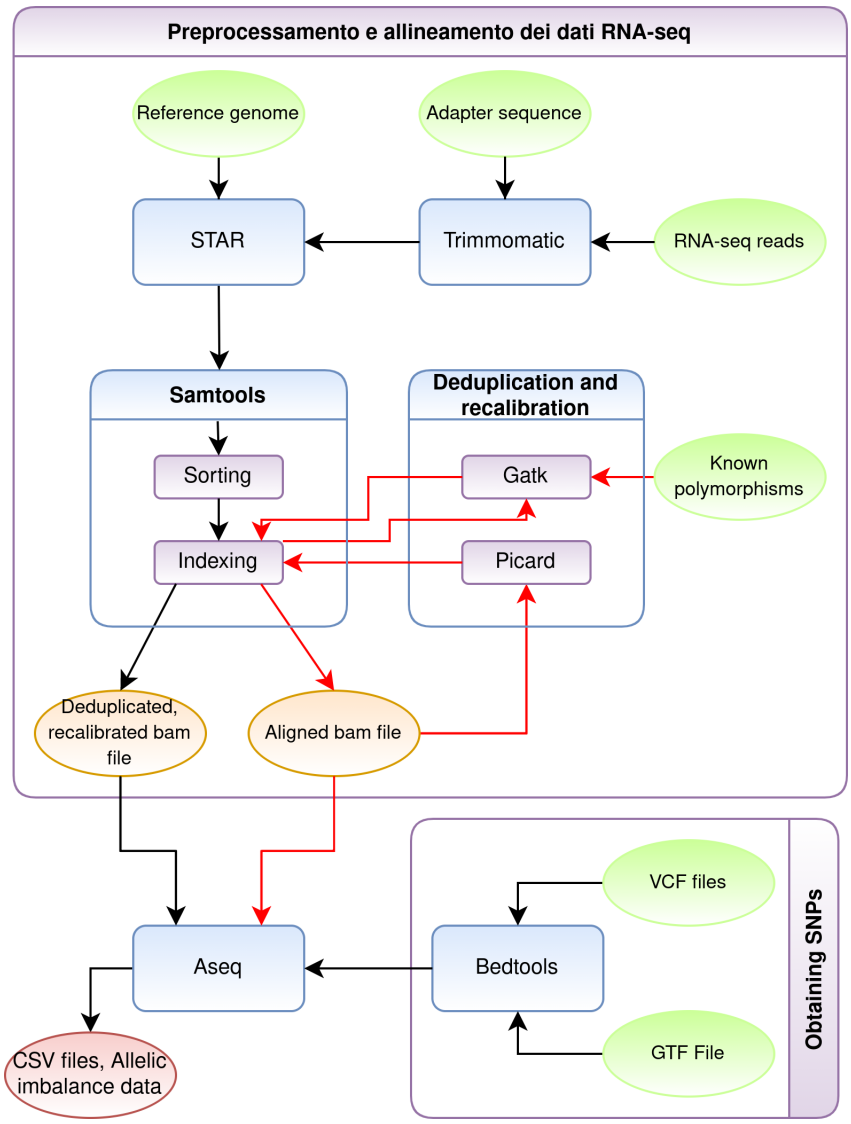

Figura 3.1: Pipeline per l'ottenimento dei dati di allelic imbalance

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 dei *FASTQ*, file di testo che contengono i dati di sequenziamento raccolti in laboratorio. Questi file subiscono un processo di preparazione in modo da ottenere dei file di allineamento utilizzabileiper 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 [3] 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-ends[1] 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 abbastanza accurata. Si nota

[1] Punta a spiegazione nel capitolo 1

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 fastq file con le sequenze adattatrici rimosse.

3.1.2 Star

Star (spliced transcript alignment to a riferimento) [4] 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 e producendo i file di allineamento. 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 $(R_i, R_{i+1}, \ldots, R_{i+MML-1})$ 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 mappato 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 sam file, un file di testo delimitato da tabulazioni contenente una riga per allineamento con tutte le informazioni necessarie per la sua identificazione.[2]

3.1.3 Samtools

I samtools [5] sono un insieme di programmi necessari per interagire con i dati di allineamento. Nella pipeline sono utilizzati per compiere delle operazioni sui sam file 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 [6] e samtools index [7]. Il primo ordina gli allineamenti secondo le coordinate di inizio e comprime implicitamente l'input in formato bam. 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 bam file. L'output finale è un bam file 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 bam file che sono stati allineati attraverso star (§3.1.2). Questi due

[2] Approfondire

passaggi vengono svolti attraverso una serie di tools eseguiti sequenzialmente come si nota nella figura 3.1.4.

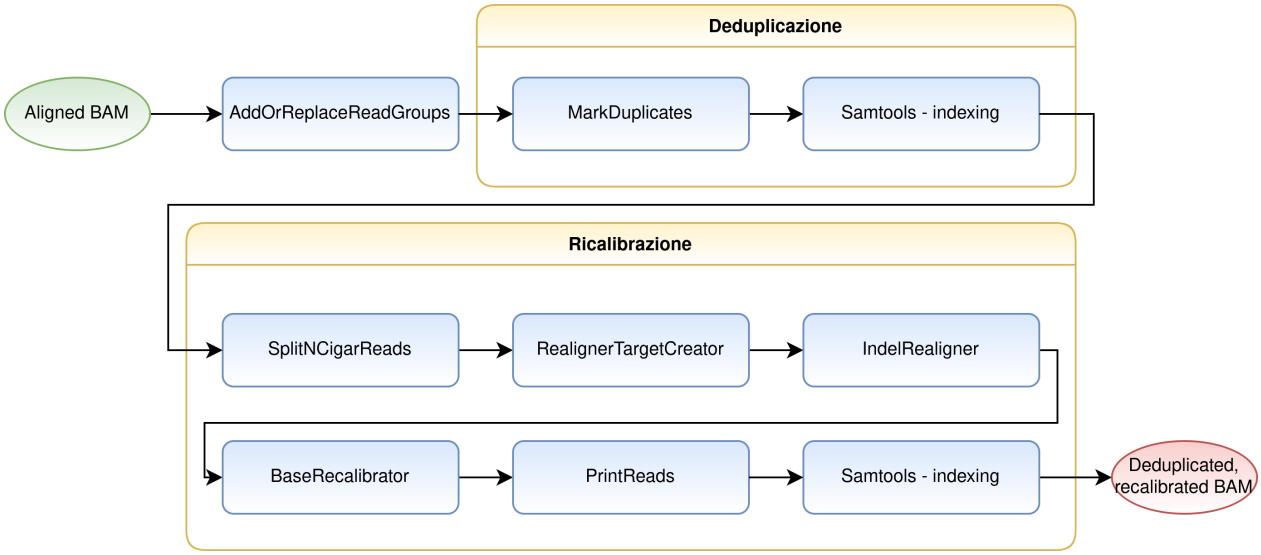

Figura 3.2: Pipeline di deduplicazione e ricalibrazione

Il primo passaggio viene svolto dal tool *AddOrReplaceReadGroups* [8] di Picard [9], 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 BAM file, in modo da assegnare tutte le read a un genotipo specifico.

Deduplicazione

Il processo di deduplicazione e le motivazioni dietro alla sua utilità sono definite in [10] Si definiscono come read duplicate in un BAM file 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 tutti le read duplicata 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 BAM file attraverso *samtools index*.

Ricalibrazione

La ricalibrazione viene svolta da una serie di tool facenti parte della suite Gatk [11]. Il processo si compone di due fasi:

1. Riallineamento degli indels definito in [12].
2. Ricalibrazione delle basi basata sul punteggio di qualtità definito in [13].

Il primo passaggio, svolto dal tool *SplitNCigarReads* [14], progettato specificatamente per dati di RNA-seq, è necessario per il corretto funzionamento degli step successivi. In un file BAM di RNA-seq infatti si trovano stringhe *CIGAR* che descrivono come una base in ogni read è mappata. Il valore *N* corrisponde a una base saltata sul genoma di riferimento. Nel caso di RNA-seq tali basi possono corrispondere o a sequence intronic, 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 destra e l'altra che inizia a sinistra della base

rimossa. Il risultato di questo processo è la separazione degli esoni in read diverse e un'eliminazione degli overhang. Come risultato gli esoni del RNA vengono separati in segmenti e gli overhang eliminati in modo da non causare falsi positivi.
Dopo questo lavoro di processamento inizia la fase di riallineamento degli indels.

Riallineamento locale degl 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 BAM file ordinato e indicizzato e a partire da esso genera un file di output formato da una lista a una colonna contenente gli intervalli. Ogni reecord 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 BAM ordinato e indicizzato.

Ricalibazione 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 machine learning per modellarli empiricamente a modificare i valori di qualità migliorarne la veridicità. La ricalibrazione delle basi avviene attraverso l'utilizzo di due tool. Il primo è *BaseRecalibrator* [15] e costruisce un modello di covarianza da un BAM e da un insieme di varianti conosciute in un file VCF 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 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* [16] applica il modello creato dal primo al BAM di input, aggiornandolo così ai punteggi di qualità migliorati. Infine viene ricreato l'indice del BAM file attraverso *samtools index*. In questo modo si è creato un file BAM pronto per essere utilizzato da ASEQ.

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 GTF[3] contenente informazioni sulla porzione esonica del genoma umano e da un insieme di VCF divisi per cromosoma contenenti le informazioni riguardo le varianti ottenute negli essere umani. Dopo aver ristretto i VCF agli SNP sono stati intersecati con il file GTF. L'intersezione è stata volta dal tool *bedtools intersect* [17], facente parte della suite bedtools [18]. Questo tool ha permesso di trovare ogni record del VCF presente anche nel 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

[3]Sezione dati input

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.

3.3 Calcolo dei dati di sbilanciamento allelico

Per computare i valori di sbilanciamento allelico è stato utilizzato aseq [19]. 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 di nuova generazione trascrittomici 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 inidicizzate e ne aumenta il potere computazionale attraverso il multi-threading. Delle modalità che aseq fornisce è stata utilizzata quella principale di analisi ASE, 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. Questo permette di applicare all'output diversi filtri successivamente durante l'analisi cosicchè da poter trovare i valori soglia ottimali per ottenere risultati significativi in un secondo momento senza ripetere continuamente la computazione del pileup e snellendo il carico computazionale dell'analisi. Non si sfrutta pertanto il potere di individuazione di eventi di ASE da parte 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 campione[4]. 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.

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

[4] Riferimento a lista di campioni

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.

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

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

[4] 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.

[5] Samtools. http://www.htslib.org/. ultimo accesso 01/08/2021.

[6] Documentazione per samtools sort. http://www.htslib.org/doc/samtools-sort.html. ultimo accesso 01/08/2021.

[7] Documentazione per samtools index. http://www.htslib.org/doc/samtools-index.html. ultimo accesso 01/08/2021.

[8] Addorreplacereadgroups. https://gatk.broadinstitute.org/hc/en-us/articles/360057440331-AddOrReplaceReadGroups-Picard-. ultimo accesso 06/08/2021.

[9] Picard. https://broadinstitute.github.io/picard/. ultimo accesso 01/08/2021.

[10] Markduplicates. https://gatk.broadinstitute.org/hc/en-us/articles/360057439771-MarkDuplicates-Picard-. ultimo accesso 06/08/2021.

[11] Gatk. https://gatk.broadinstitute.org. ultimo accesso 01/08/2021.

[12] Indel realignment. https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/(howto)_Perform_local_realignment_around_indels.md. ultimo accesso 06/08/2021.

[13] Base quality score recalibration. https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-Recalibration-BQSR-. ultimo accesso 06/08/2021.

[14] Splitncigarreads. https://gatk.broadinstitute.org/hc/en-us/articles/360036734471-SplitNCigarReads. ultimo accesso 06/08/2021.

[15] Baserecalibrator. https://gatk.broadinstitute.org/hc/en-us/articles/360036898312-BaseRecalibrator. ultimo accesso 06/08/2021.

[16] Printreads. https://gatk.broadinstitute.org/hc/en-us/articles/360036883571-PrintReads. ultimo accesso 06/08/2021.

[17] Bedtools intersect. https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html. ultimo accesso 06/08/2021.

[18] Bedtools. https://bedtools.readthedocs.io/en/latest/. ultimo accesso 06/08/2021.

[19] 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