Bioinformatica II

LM Biologia Evoluzionistica, Università di Padova

Docenti: Dr. Giorgio Valle, Dr. Stefania Bortoluzzi

Esercitazione 6

Padova, 21 dicembre 2017

GUIDA

Mappaggio e variant calling con reads Illumina

Obbiettivo dell'esercitazione

Capire come mappare le reads prodotte da un sequenziatore di nuovo generazione su una sequenza di riferimento e chiamare le varianti.

Il primo passaggio è la creazione dell'indice della sequenza di riferimento con il comando:

```
bwa index -p ref ref.fasta
```

Questo comando produrrà vari file di output con diverse estensioni che costituiscono l'indice della sequenza di riferimento (un trascritto o il genoma su cui vogliamo mappare le reads).
Possiamo visualizzare l'elenco dei file prodotti con il comando:

```
ls -l
```

bwa ha vari sottocomandi che possono essere elencati lanciando semplicemente il comando:

```
bwa
```

Una volta ottenuto l'indice della sequenza di riferimento possiamo effettuare il mappaggio vero e proprio. In questo caso useremo reads paired end prodotti con la tecnologia illumina, quindi le reads si troveranno in due file diversi.
Per mappare le reads nel primo file usare il seguente comando:

```
bwa aln -f 1.sai ref 1.fastq
```

e per mappare il secondo file di reads:

```
bwa aln -f 2.sai ref 2.fastq
```

I file prodotti dal comando aln di bwa sono in un formato "intermedio" caratterizzato dall'estensione .sai
Per utilizzare i risultati dell'allineamento con altri programmi i file .sai devono essere convertiti in un formato standard utilizzato da gran parte dei programmi bioinformatici: sam
Per convertire i .sai in un sam si usa il sottocomando sampe di bwa:

```
bwa sampe ref 1.sai 2.sai 1.fastq 2.fastq > mapping.sam
```

I file sam di solito vengono compressi in un formato binario (non di testo e quindi comprensibile solo al computer) che si chiama bam e che sta per binary sam. Per portare a termine questa conversione si usa il programma samtools che è un pacchetto software per la manipolazione e l'estrazione di informazione dai file sam/bam.

```
samtools view -S -b mapping.sam > mapping.bam
```

Per ordinare le reads del file BAM in base alla loro posizione nel genoma usiamo il comando sort di samtools:

```
samtools sort mapping.bam > sorted.bam
```

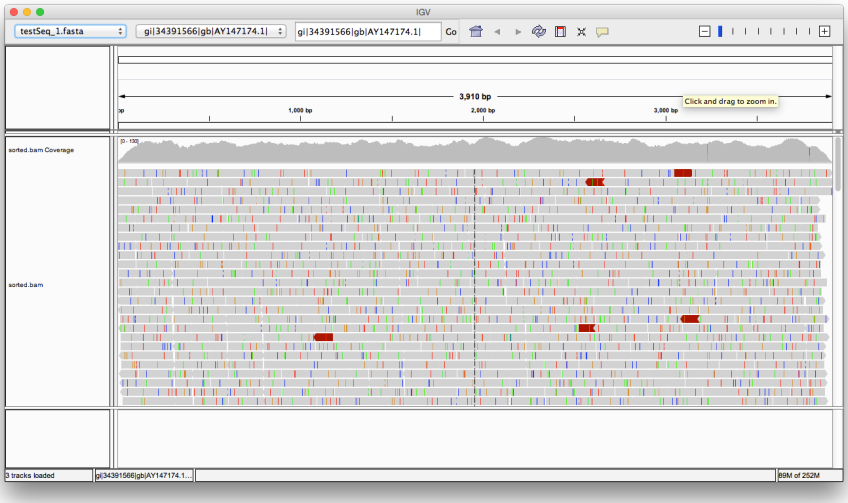
Ora creiamo l'indice per il nostro file BAM ordinato:

```
samtools index sorted.bam
```

Per visualizzare le reads allineate al genoma usiamo il comando tview di samtools, ma prima bisogna creare l'indice per il nostro genoma:

```
samtools faidx ref.fasta
samtools tview -p hg19_dna:15892-15987 sorted.bam ref.fasta
```

Un programma molto utilizzato per visualizzare gli allineamenti in vari formati, tra cui i sam/bam è IGV (Integrative genomics viewer) che può essere scaricato dal seguente indirizzo: http://www.broadinstitute.org/igv/



La chiamata delle varianti può essere fatta utilizzando il programma mpileup di samtools:

```
samtools mpileup -v -u -f ref.fasta sorted.bam > mapping.vcf
```

```
samtools tview -p hg19_dna:15894-15896 sorted.bam ref.fasta
```