Laurea in Biotecnologie Sanitarie, Universita' di Padova
Docente: Dr. Stefania Bortoluzzi

Corso di Bioinformatica

http://compgen.bio.unipd.it/~stefania/Didattica/AA2011-2012/Bioinformatica_BTS/Bioinfo_BTS.html

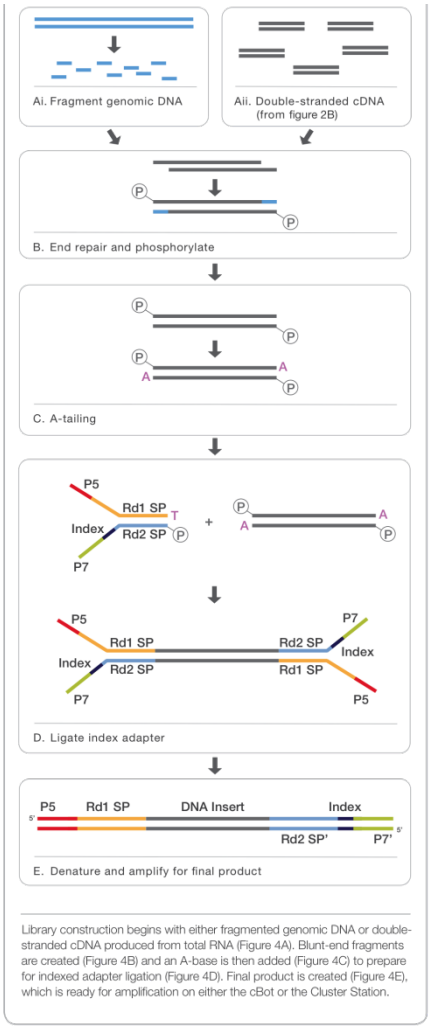
Per ogni punto dell'esercitazione copiare e incollare le informazioni richieste in un file sotto forma di breve testo con figure, in modo da assemblare un traccia di relazione con commenti sintetici.
Alla fine inviare via mail la relazione prodotta nel corso dell'esercitazione (stefania.bortoluzzi@unipd.it).
Salvare il file della relazione ogni 5'!

IV ESERCITAZIONE

Analisi del trascrittoma in organismi modello a partire da dati di deep sequencing: mappaggio delle reads al genoma ed analisi dei risultati per l'identificazione di geni espressi.

In questa esercitazione svolgeremo una parte delle analisi necessarie per identificare i geni espressi in un campione biologico di un organismo modello (genoma disponibile) a partire da dati di sequenziamento massivo del trascrittoma (RNA-seq) con tecnologia Illumina. In particolare, usando un web server molto potente (Galaxy), effettueremo il mappaggio delle reads di sequenza al DNA genomico con due programmi diversi ed integreremo i risultati con alcune track di annotazione del genoma mediante il sistema IGV (Integrative Genomics Viewer) per identificare esoni espressi in geni noti, e scoprire possibili nuovi geni.

Figura 1. Schema della contruzione di una libreria Paired End

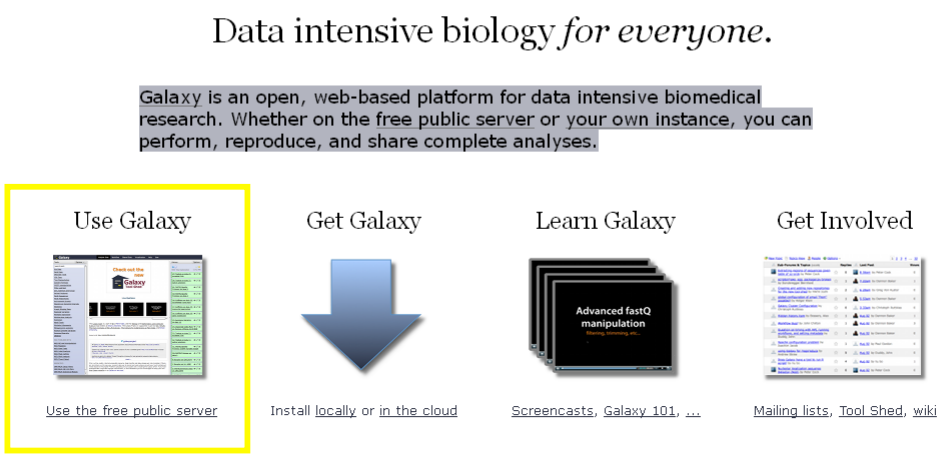


1. Scaricare i dati necessari per l'esercitazione

Aprire il browser Firefox, all'indirizzo **compgen.bio.unipd.it/~enrico** scaricare il file .zip disponibile online, salvandolo in una cartella col vostro nome in "Documenti".
Ci vorranno circa 5 minuti.

Decomprimendo il file troverete una cartella contenente:
Raw_data:
SRR167673_chr10_1.fastq.zip
SRR167673_chr10_2.fastq.zip
Si tratta dei dati grezzi di sequenziamento RNA-seq Illumina con metodica "paired end a partire da RNA di tessuto adiposo di maiale. Nei due files (in quanto il sequenziamento è stato fatto "paired end") sono incluse reads in formato FASTQ.
Si tratta di circa 500.000 reads (solo una parte del dataset originale) di 90 basi ciascuna per file. Le reads appaiate distano 200 sulla sequenza originale.
I file sono zippati.
Mapping_results:
bowtie_sscg9.65_chr10_mapped_reads.bam
bowtie_sscg9.65_chr10_mapped_reads.bam.bai
tophat_sscg9.65_chr10_mapped_reads.bam
tophat_sscg9.65_chr10_mapped_reads.bam.bai
Risultati del mappaggio, da usare più tardi, o "di scorta".

2. Usare Galaxy project http://g2.bx.psu.edu/

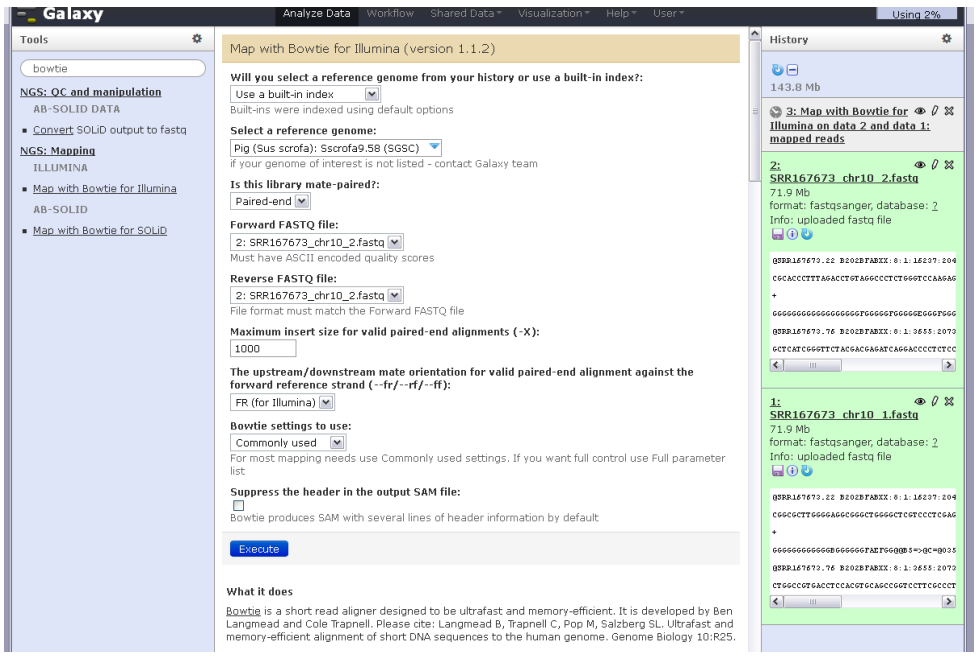


Data Upload in Galaxy: scegliere "Get data" → "Upload data from your computer"
Scegliere il file da caricare (entrambi i fastq zippati, uno alla volta) e cliccare su "Execute".
Attenzione!!! ***Please do not use your browser's "stop" or "reload" buttons until the upload is complete, or it may be interrupted.***
Il sistema prima esegue l'upload, poi dezippa automaticamente i files. Quando appare il colore verde l'operazione e' completata. Cliccando "sull'occhio" potete vedere la parte iniziale del file. **Copiare ed incollare le prime due reads e commentare il formato.**

Cliccando "sulla matitina" è possibile editare le proprietà del file.
Cambiare il formato scegliendo "fastqsanger" per entrambi i files.

Mappaggio delle reads al genoma di maiale usando Bowtie.
Sulla sinistra, scrivere "bowtie": il sistema fornisce diverse opzioni, scegliere "Map with Bowtie for Illumina".
Compare il form per lanciare il programma di mappaggio.
Scegliere come genoma di riferimento "Pig (Sus scrofa): Sscrofa9.58 (SGSC)" (dato che e' disponibile non e' necessario fornire un indice della sequenza genomica).
Specificare che la libreria e' "mate paired" e usare il file #1 come forward e il #2 come reverse.
Cliccare su "Execute".

Il mappaggio e' un processo computazionalmente intensivo, data la numerosità delle reads e la dimensione del genoma. Il programma Bowtie e' uno dei programmi di mappaggio più veloci, ma non permette la divisione delle reads (unspliced mapping)



Se eseguito localmente, questo mappaggio potrebbe necessitare di un tempo variabile tra mezz'ora e alcune ore, a seconda della potenza e del numero dei processori disponibili. Noi stiamo eseguiamo il mappaggio usando le potenti risorse computazionali del server remoto.

Durante il tempo di esecuzione del mappaggio: **breve presentazione sugli algoritmi di mappaggio di dati RNAseq.**

I risultati del mappaggio consistono in un file in formato SAM(Sequence Alignment/Map), che e' "human readable", ma molto grande. Usare il menu a destra per visualizzare la parte iniziale del file SAM.

Quanti Mb occupa il file SAM?

```
@HD VN:1.0 SO:unsorted@SQ SN:chrM LN:16613
@SQ SN:chr1 LN:295534705
@SQ SN:chr2 LN:140138492
@SQ SN:chr3 LN:123604780
@SQ SN:chr4 LN:136259946
@SQ SN:chr5 LN:100521970
@SQ SN:chr6 LN:123310171
@SQ SN:chr7 LN:136414062
@SQ SN:chr8 LN:119990671
@SQ SN:chr9 LN:132473591
@SQ SN:chr10 LN:66741929
@SQ SN:chr11 LN:79819395
@SQ SN:chr12 LN:57436344
@SQ SN:chr13 LN:145240301
@SQ SN:chr14 LN:148515138
@SQ SN:chr15 LN:134546103
@SQ SN:chr16 LN:77440658
@SQ SN:chr17 LN:64400339
@SQ SN:chr18 LN:54314914
@SQ SN:chrX LN:125876292
@PG   ID:Bowtie   VN:0.12.7   CL:"bowtie   -q   -X   1000   --fr   -p   8   -S   --phred33-quals
/galaxy/data/Sscrofa9.58/bowtie_index/Sscrofa9.58                                    -1
/galaxy/main_pool/pool4/files/004/269/dataset_4269327.dat                            -2
/galaxy/main_pool/pool5/files/004/269/dataset_4269329.dat"

SRR167673.30399     chr10  11375393     255    90M    =      11375475     172
       CGGGTTTACAGACTGGCAAACAGGCAGCCAGGTAGCCTGGCAACATGGGCTTTGGTCTCGCTGCAGAAGCGGACACGGGGCA
TCCGGTCA
       GGGGEGGGGGGGGGGGGGGGGFGGFFFDDEEFAEEFFDEFGGEEDADEEEDGEDGEGEEEDA=EDE:ECCDDCECDDEFEED
CDE?B=A=      XA:i:0 MD:Z:75T14 NM:i:1
```

Convertire il file in formato binario compresso

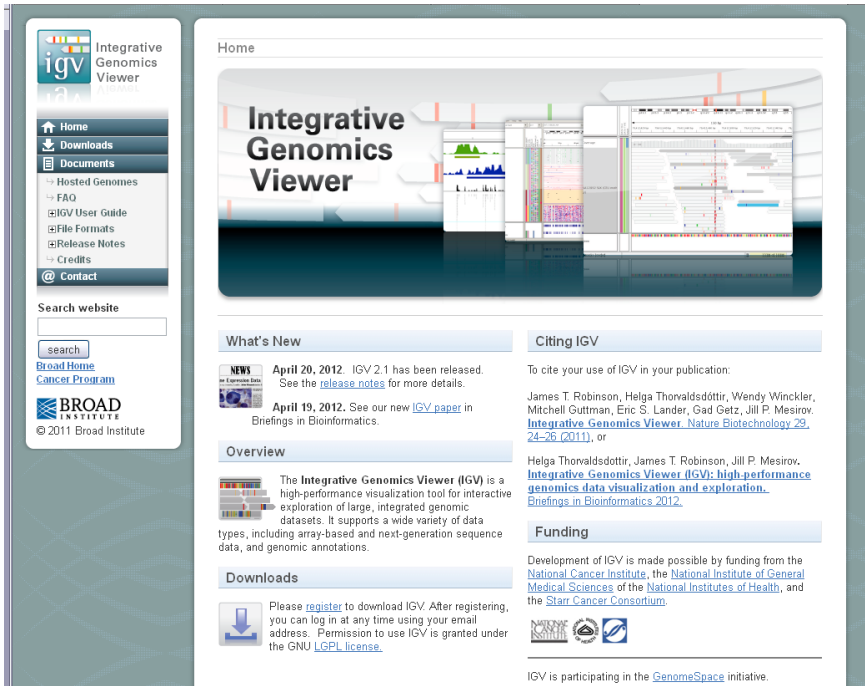
Nel menu a sinistra scrivere "SAMtools" e scegliere "SAM-TO-BAM": Human readable SAM (Sequence Alignment Mapping) CONVERSION TO Binary compressed file BAM (Binary Alignment Mapping).

Quanti Mb occupa il file BAM?

Scaricare il file BAM e l'indice nella propria cartella.

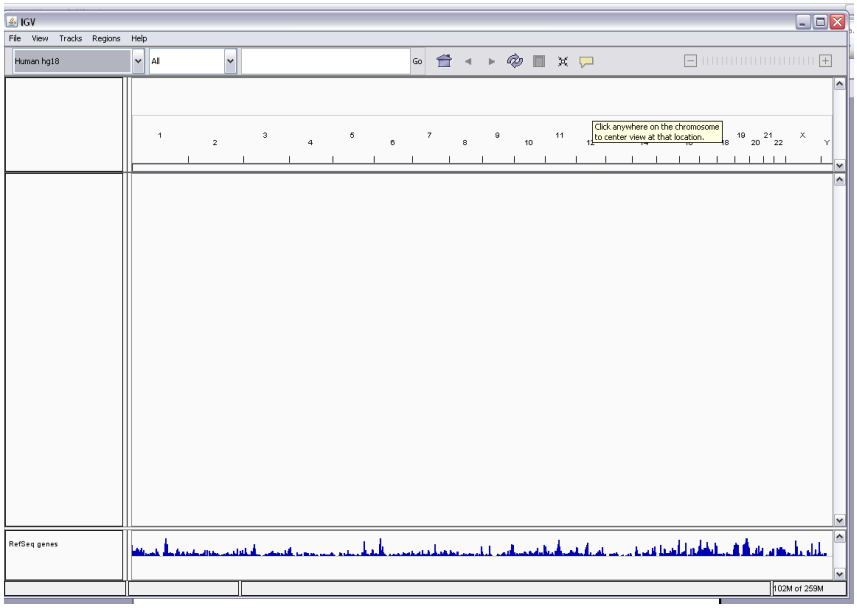
3. Visualizzazione e "analisi" del mappaggio con IGV (Integrative genomics Viewer).

The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.



Sulla sinistra scegliere "Downloads" e usare l'indirizzo mail compgen.unipd@gmail.com già preparato.
Prima di tutto lanciare IGV usando JAVA Web Start, scegliere "Launch with 750 MB" (minima "potenza" permessa, ma la massima utilizzabile con i PC disponibili), aprire il file e dare "esegui".

Quando si entra in IGV, si visualizza di default il genoma umano (24 cromosomi), con già caricata la track delle annotazioni corrispondente, quale la track "RefSeq" genes in basso.
Dal menu a tendina è possibile scegliere un cromosoma oppure il genoma mitocondriale.



Nel menu in lato a sinistra scegliere il genoma di maiale "Sus scofa 9.56".
Quanti cromosomi vedete?
IGV ci servirà per visualizzare gli allineamenti ottenuti tra le reads RNA-seq, che in ultima analisi rappresentano frammenti di RNA trascritti in tessuto adiposo di maiale, sulla sequenza genomica, e in integrazione con alcune annotazioni della stessa, quali le posizioni e la struttura dei geni noti.
Espandere la track "Gene in basso" usando il tasto destro.

Caricare il file degli allineamenti ottenuti con Bowtie, in formato BAM sorted: "Load from File".
Caricare anche il file `tophat_sscg9.65_chr10_mapped_reads.bam` presente nella cartella fornita.
Scegliere cromosoma 10, verrà visualizzata una schermata con l'intero cromosoma, con le reads non visibili. **Quante Mb è lungo il cromosoma 10?**

Aumentando lo zoom visualizziamo gli allineamenti e contemporaneamente vediamo i geni noti in maggiore dettaglio. Oltre un certo "ingrandimento" visualizziamo la sequenza genomica come una barra con i nucleotidi colorati.
Usando lo zoom o le coordinate visualizzare una finestra che comprenda circa 20 Kb.
Scorrere il cromosoma a partire dall'inizio (ptel; coordinate: 10:1-20,001).

Nella parte superiore vedrete due tracks per ciascun mappaggio:
- Coverage: diagramma a barre che indica il numero di sequenze allineate per nucleotide
- Allineamenti: dettaglio degli allineamenti reads-genoma con gli SNPs colorati.

Confrontare i risultati di mappaggio ottenuti con Bowtie e con Tophat.
Secondo il mappaggio con TopHat:
- trovare un gene espresso(coverage superiore a 10) e uno "non espresso".
- Esistono regioni "introniche" che risultano espresse? Cercarne una lunga almeno 50 nt.
- Cercare un esone con almeno uno SNP e visualizzarlo insieme con la traduzione del messaggero nella regione intorno.