

Esercitazione con Galaxy

Accesso a Galaxy:

<http://galaxyproject.org/> (e cliccare su "Use Galaxy") oppure <https://orione.crs4.it>

ESERCIZIO 1

Ricostruire i trascritti e le coding sequences (CDS) del gene umano ARHGAP4 annotato sulla regione ENCODE ENm006.

File di input:

- 1) Sequenza della regione ENCODE ENm006 in formato FASTA → ENm006.fa.
- 2) File in formato GTF dei trascritti del gene umano ARHGAP4 annotati sulla regione ENCODE ENm006 → ARHGAP4.gtf

Step 1: caricare il file GTF di ARHGAP4 e il file FASTA della regione ENm006.

Tool:

"GetData" → "Upload file from your computer"

Dopo l'esecuzione del job:

- cambiare in "ARHGAP4 annotation" il nome del dataset caricato per l'annotazione GTF di ARHGAP4.

Step 2: ricostruzione dei trascritti

a) Estrarre dal dataset "ARHGAP4 annotation" i soli *record* corrispondenti alla *feature* "exon"

Tool:

"Filter and sort" → "Extract features from GFF data" in (sottosezione GFF),
OPPURE

"Filter and sort" → "Filter data on any column using simple expressions"

Dopo l'esecuzione del job:

- cambiare il nome del dataset prodotto in "ARHGAP4 exons"

b) Ordinare i *record* del dataset "ARHGAP4 exons" in maniera tale che, per ogni trascritto i *record* siano ordinati per valore crescente di start della *feature*

Tool:

"Filter and sort" → "Sort data in ascending or descending order"

Dopo l'esecuzione del job:

- cambiare il nome del dataset prodotto in "ARHGAP4 sorted exons"

c) Estrarre in formato FASTA le sequenze di ognuno degli esoni del dataset "ARHGAP4 sorted exons"

Tool:

"Fetch Alignments/Sequences" → "Extract genomic DNA using coordinates from assembled/unassembled genomes"

Step 3: ricostruzione delle CDS

ESERCIZIO 2

Allineare reads NGS al genoma.

Step 1: recuperare da Sequence Read Archive (SRA) il dataset FASTQ con con identificatore SRR1517298.

Tool:

“Get Data” → “EBI SRA ENA SRA”

In alternativa:

scaricare il dataset direttamente all’indirizzo <http://www.ebi.ac.uk/ena> e caricarlo poi in Galaxy usando il tool “Get Data” → “Upload File from your computer”

Step 2: convertire il dataset in formato Sanger FASTQ

Tool:

“NGS: QC and manipulation” → “FASTQ Groomer convert between various FASTQ quality formats”

Tipo di input → “Illumina 1.3–1.7”

Step 3: allineare il dataset al genoma usando il software BWA (Burrows-Wheeler Aligner)

Tool:

“NGS: mapping” → “BWA - map short reads (< 100 bp) against reference genome”

Genoma → “Human (Homo Sapiens) (b38): hg38”