

CORSO DI METODI MOLECOLARI E BIOINFORMATICA

LM Biologia Evoluzionistica, Università di Padova
Anno accademico 2021/22

Computational Genomics Lab. (compgen.bio.unipd.it)



Prof. Stefania Bortoluzzi



Andrea Binatti, Post. Doc
(andrea.binatti@unipd.it)



Enrico Gaffo, Post. Doc
(enrico.gaffo@unipd.it)

Programma del corso

- 20/10 Genome browser (UCSC)
- 27/10 Struttura sistema Unix + macchina virtuale
- 03/11 Comandi Unix
- 10/11 Comandi Unix
- 18/11 Comparazione di biosequenze ed allineamenti multipli (Blast, Clustal Omega)
- 26/11 Analisi di dati DNA-seq
- 03/12 Analisi di dati RNA-seq

Programma del corso

- 20/10 Genome browser (UCSC)
- 27/10 Struttura sistema Unix + macchina virtuale
- 03/11 Comandi Unix
- 10/11 Comandi Unix
- 18/11 Comparazione di biosequenze ed allineamenti multipli (Blast, Clustal Omega)
- 26/11 Analisi di dati DNA-seq
- 03/12 Analisi di dati RNA-seq

Cos'è un “genome browser”?

È un programma informatico che permette di “navigare” lungo il genoma

Perchè è utile un genome browser?

- Rappresenta graficamente il genoma
- Annota il genoma con dati biologici (espressione genica, regolazione dell'espressione genica, struttura dei geni, varianti genomiche, analisi comparative etc.)
- Interpretazione dei risultati o dei dati in un contesto genomico
- Ci permette di ottenere facilmente dati biologici su una regione di interesse
- Forniscono strumenti informatici per analizzare il genoma

Genome browsers disponibili in rete



Genome Browser



U.S. National Library of Medicine



National Center for Biotechnology Information

Genome Data Viewer

UCSC Microbial Genome Browser



Artemis

PlantGDB

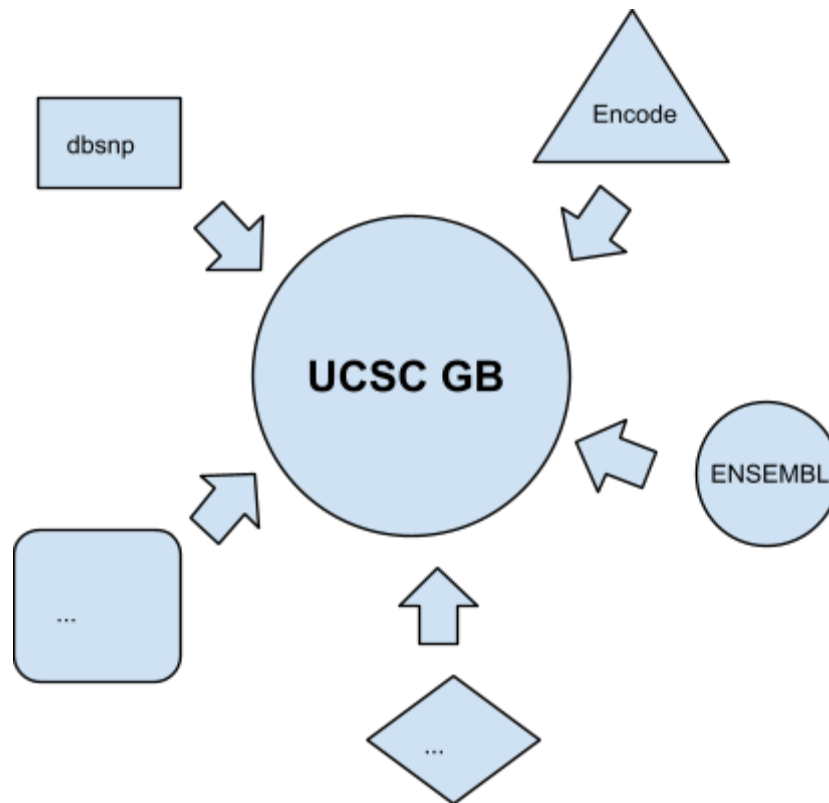


GUIDA DELL'ESERCITAZIONE 1

Parte 1: Introduzione all' UCSC Genome Browser

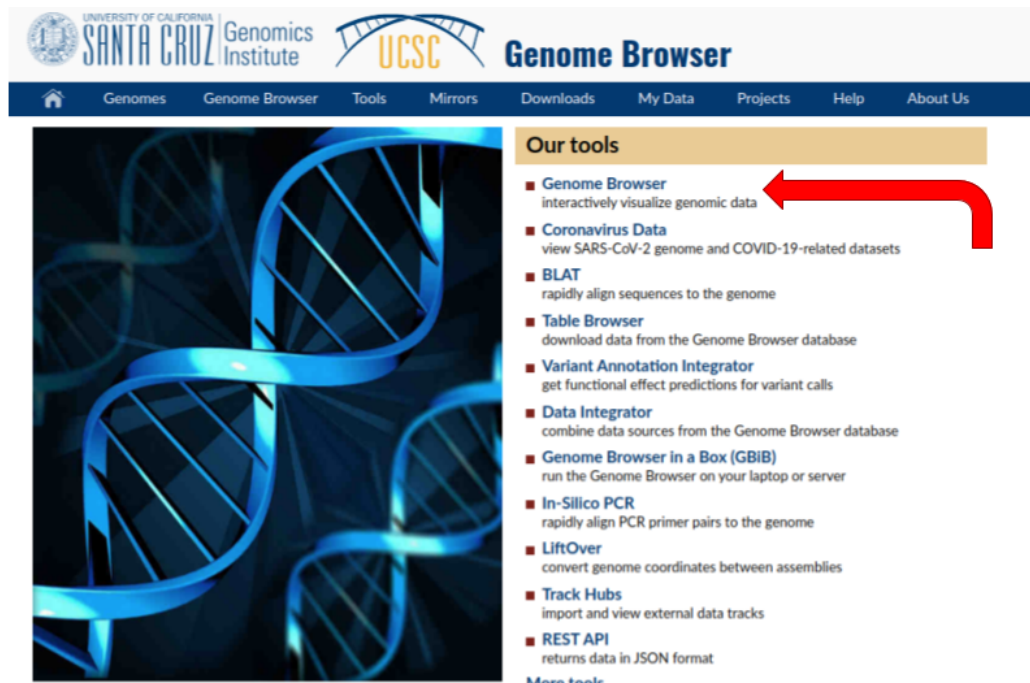
Obiettivo dell'esercitazione

Capire il funzionamento e l'utilità di un genome browser attraverso l'utilizzo dell' UCSC Genome Browser (<http://genome.ucsc.edu/>)

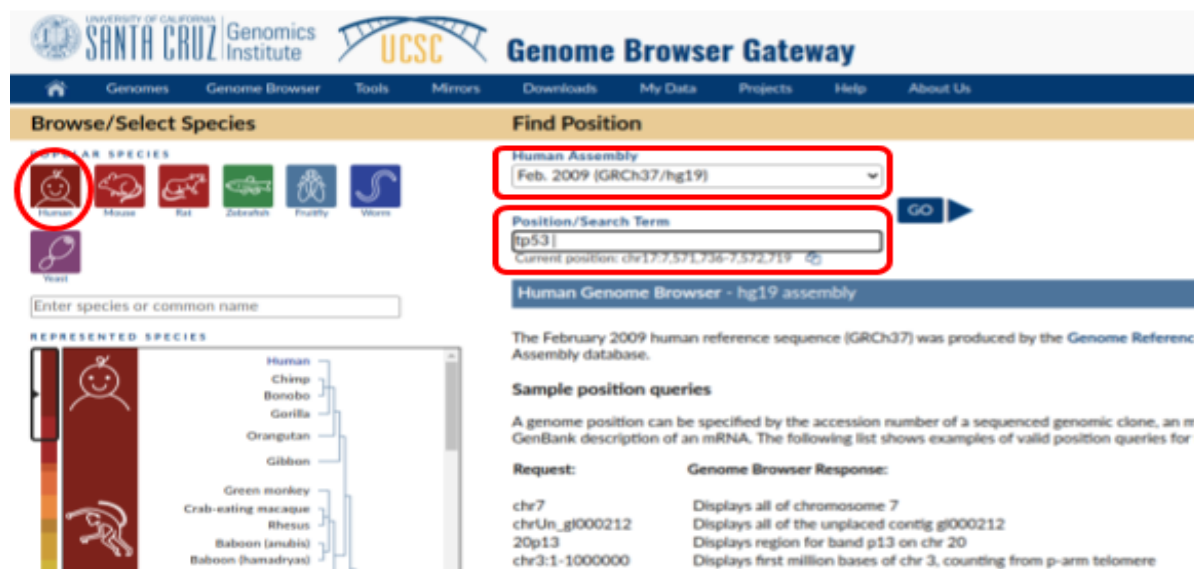


Ricerche di base

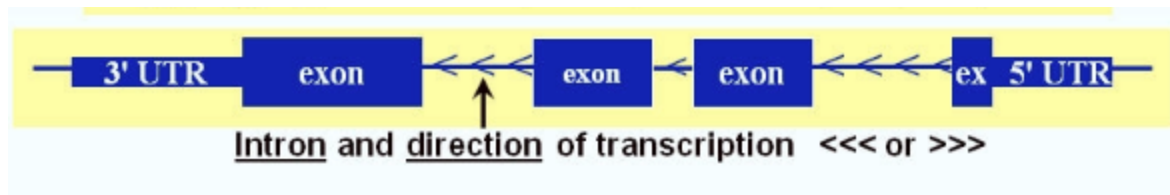
1 - Cliccare su **Genome Browser** nella barra laterale oppure **Genomes** nella barra superiore.



2 - Cercare il gene tp53 (umano) utilizzando la versione di febbraio 2009 dell'assemblaggio



3 - Nella pagina dei risultati della ricerca cliccare la entry “TP53 (uc002gij.3) at chr17:7571720-7590868”

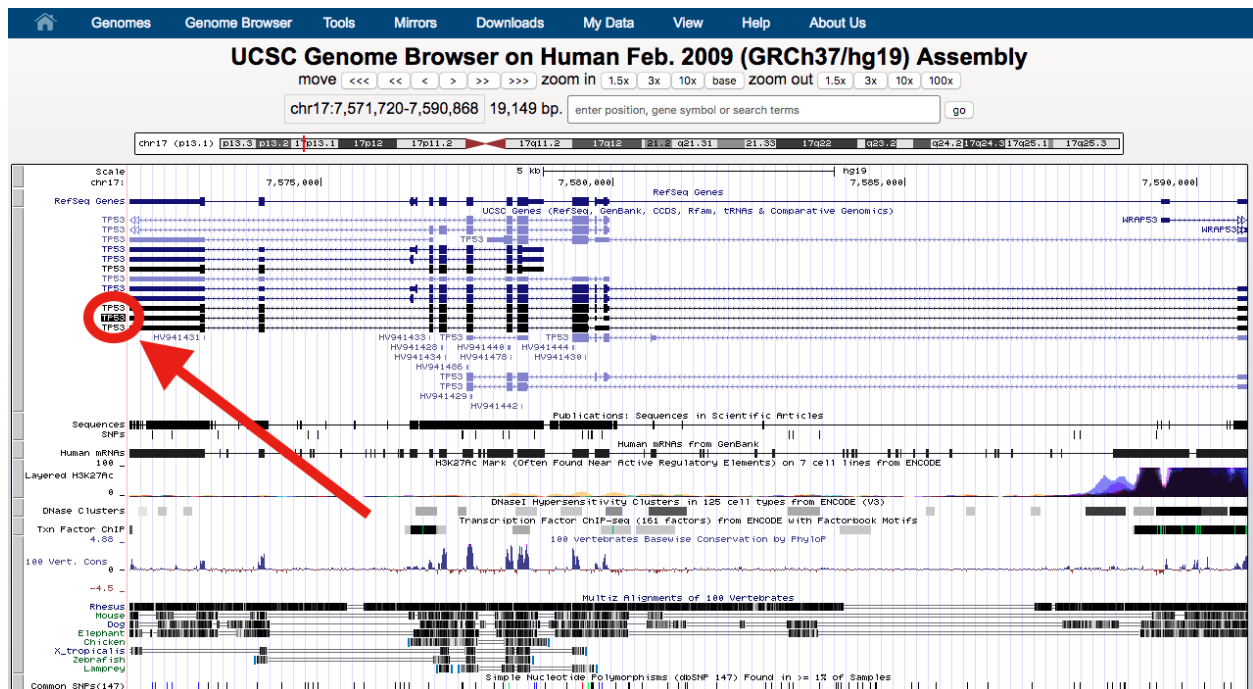


4 - Includere 1000 basi a monte (promotore putativo) del trascritto scelto: “chr17:7,571,720-7,591,868”

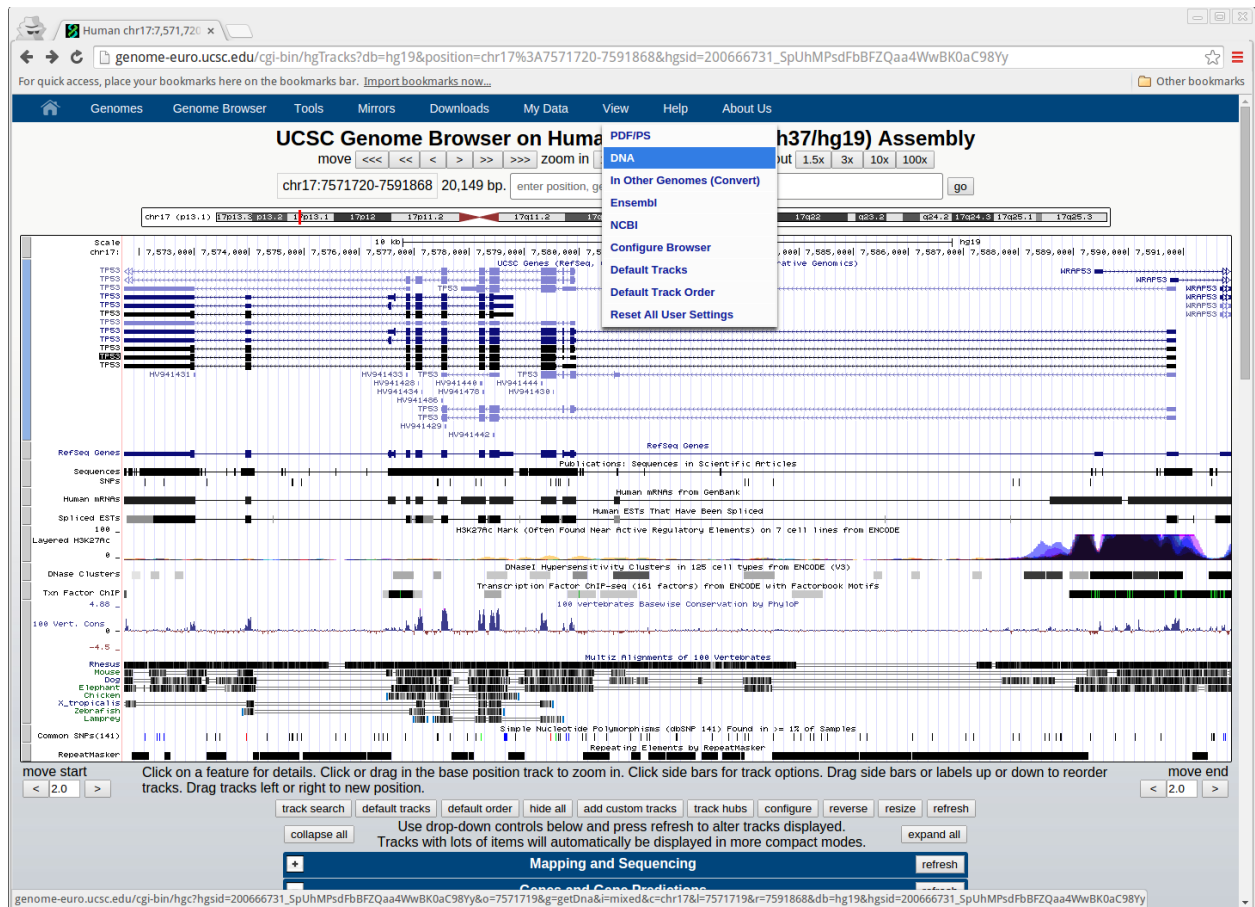
5 - Provare a cambiare l’ordine con cui le track vengono visualizzate nel “viewer”

6 - Cambiare i diversi livelli di visualizzazione della track Spliced ESTs

7 - Visualizzare la pagina sui dettagli dell’isoforma da noi scelta cliccando sulla riga con il corrispondente trascritto.



8 - Visualizzare la sequenza di DNA corrispondente alla regione nel "viewer".



Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

Get DNA in Window (hg19/Human)

Get DNA for

Position

Note: This page retrieves genomic DNA for a single region. If you would prefer to get DNA for many items in a particular track, or get DNA with formatting options based on gene structure (introns, exons, UTRs, etc.), try using the [Table Browser](#) with the "sequence" output format.

Sequence Retrieval Region Options:

Add extra bases upstream (5') and extra downstream (3')

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

Sequence Formatting Options:

- ☒ All upper case.
- ☐ All lower case.
- ☐ Mask repeats: ☒ to lower case ☐ to N
- ☐ Reverse complement (get '-' strand sequence)
- ☒ get DNA. [extended case/color options](#)

Note: The "Mask repeats" option applies only to "get DNA", not to "extended case/color options".

9 - A partire da Extended DNA Case/Color Options scegliere di visualizzare UCSC Genes in rosso (254) e Spliced ESTs in verde (254). Cosa rappresentano le regioni colorate in giallo?

10 - A partire dalla pagina sui dettagli dell'isoforma scelta (punto 7) ottenere la sequenza proteica in formato FASTA.

[Home](#)
[Genomes](#)
[Genome Browser](#)
[Tools](#)
[Mirrors](#)
[Downloads](#)
[My Data](#)
[Projects](#)
[Help](#)
[About Us](#)

Human Gene TP53 (uc002gij.3) Description and Page Index

Description: Homo sapiens tumor protein p53 (TP53), transcript variant 1, mRNA.
RefSeq Summary (NM_001276760): This gene encodes a tumor suppressor protein containing transcriptional activation, DNA target genes, thereby inducing cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. Mutations in the splicing of this gene and the use of alternate promoters result in multiple transcript variants and isoforms. Additional isoforms 12032546, 20937277). [provided by RefSeq, Dec 2016].

Transcript (Including UTRs)
Position: hg19 chr17:7,571,720-7,590,868 **Size:** 19,149 **Total Exon Count:** 11 **Strand:** -

Coding Region
Position: hg19 chr17:7,572,927-7,579,569 **Size:** 6,643 **Coding Exon Count:** 8

Page Index	Sequence and Links	UniProtKB Comments	Genetic Associations	MalaCards	CTD
Gene Alleles	RNA-Seq Expression	Microarray Expression	RNA Structure	Protein Structure	Other Species
GO Annotations	mRNA Descriptions	Pathways	Other Names	GeneReviews	Model Information
Methods					

Data last updated: 2013-06-14

Sequence and Links to Tools and Databases

Genomic Sequence (chr17:7,571,720-7,590,868)		mRNA (may differ from genome)		Protein (354 aa)	
Gene Sorter	Genome Browser	Other Species FASTA	Gene interactions	Table Schema	BioGPS
CGAP	Ensembl	Entrez Gene	ExonPrimer	GeneCards	H-INV
HGNC	Lynx	MGI	neXtProt	OMIM	PubMed
Reactome	Stanford SOURCE	UniProtKB			

Esercizio di riepilogo

OBIETTIVO: visualizzare gli SNPs coding synonymous e missense del gene NOTCH1 del topo

1- Visualizzare gli SNPs (dbSNP versione 142) coding synonymous (**Coding - Synonymous**), nel gene NOTCH1 di topo e colorarli di blu

2- Vogliamo ora identificare tutti gli SNP (dbSNP versione 142) **missense variants (Coding - NonSynonymous)** nel gene NOTCH1 di topo e visualizzarli in rosso. Definizione di **missense variant**:

“A genetic alteration in which a single base pair substitution alters the genetic code in a way that produces an amino acid that is different from the usual amino acid at that position. Some missense variants (or mutations) will alter the function of the protein. Also called missense mutation”. From NCI Dictionary of Genetics Terms.

3- “Zoommare” questa posizione: chr2:26,467,082-26,468,407 e dare un’occhiata alla sezione **Coding annotations by dbSNP** dei due SNP trovati.

Amino Acid	SLC	DNA codons
Isoleucine	I	ATT, ATC, ATA
Leucine	L	CTT, CTC, CTA, CTG, TTA, TTG
Valine	V	GTT, GTC, GTA, GTG
Phenylalanine	F	TTT, TTC
Methionine	M	ATG
Cysteine	C	TGT, TGC
Alanine	A	GCT, GCC, GCA, GCG
Glycine	G	GGT, GGC, GGA, GGG
Proline	P	CCT, CCG, CCA, CCG
Threonine	T	ACT, ACC, ACA, ACG
Serine	S	TCT, TCC, TCA, TCG, AGT, AGC
Tyrosine	Y	TAT, TAC
Tryptophan	W	TGG
Glutamine	Q	CAA, CAG
Asparagine	N	AAT, AAC
Histidine	H	CAT, CAC
Glutamic acid	E	GAA, GAG
Aspartic acid	D	GAT, GAC
Lysine	K	AAA, AAG
Arginine	R	CGT, CGC, CGA, CGG, AGA, AGG
Stop codons	Stop	TAA, TAG, TGA

Parte 2: Uso dell'UCSC Genome Browser tables

Obiettivo dell'esercitazione

L'obiettivo dell'esercitazione è capire come effettuare ricerche avanzate utilizzando il “*table browser*” dell'UCSC Genome Browser e visualizzare i risultati delle ricerche con le “*custom tracks*”. La funzione table browser permette di interagire in maniera quasi diretta con le tabelle del database MySQL che costituiscono lo scheletro dell'UCSC GB.

1 - Cliccare su Tables Browser nella barra di navigazione



Il Genome browser è composto da tabelle di dati, alcune di queste tabelle sono primarie e contengono informazione di posizione, per esempio la tabella degli UCSC Genes. Ci sono inoltre tabelle ausiliari che possono anche non avere informazioni di posizione.

Identificare nel genoma umano le simple repeats con sequenza esatta CAG

1 - Scegliere la tabella simpleRepeats utilizzando l'assemblaggio del 2009 del genoma umano.

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use with diverse computational tools. Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

clade: genome: assembly:

group: track:

table:

region: ☒ genome ☐ ENCODE Pilot regions ☐ position

identifiers (names/accessions):

filter:

intersection:

correlation:

output format: Send output to ☐ [Galaxy](#) ☐ [GREAT](#) ☐ [GenomeSpace](#)

output file: (leave blank to keep output in browser)

file type returned: ☒ plain text ☐ gzip compressed

To reset all user cart settings (including custom tracks), [click here](#).

La track simple repeats contiene solo una tabella (simpleRepeats). Quando ci sono più tabelle per una track, la tabella principale con informazioni di posizione genomica appare in prima posizione nella lista delle tabelle.

2 - Cliccare sul tasto “describe table schema” per vedere una descrizione della tabella

3 - Cliccare su “summary/statistics” per ottenere il numero di simple repeats presenti nel genoma umano.

4 - Creare un filtro per ottenere soltanto simple repeats la cui sequenza sia CAG

Filter on Fields from hg19.simpleRepeat

bin	is	ignored	0	
chrom	does	match	*	AND
chromStart	is	ignored	0	AND
chromEnd	is	ignored	0	AND
name	does	match	*	AND
period	is	ignored	0	AND
copyNum	is	ignored	0	AND
consensusSize	is	ignored	0	AND
perMatch	is	ignored	0	AND
perIndel	is	ignored	0	AND
score	is	ignored	0	AND
A	is	ignored	0	AND
C	is	ignored	0	AND
G	is	ignored	0	AND
T	is	ignored	0	AND
entropy	is	ignored	0	AND
sequence	does	match	CAG	

AND Free-form query:

5 - Selezionare l'opzione "all fields from selected table" nel campo "output format" e cliccare il tasto "get output".

Identificare simple repeats con sequenza esatta CAG che si trovano su geni UCSC

1 - Cliccare sul tasto "create" nella sezione "intersection" per raggiungere la pagina di creazione delle intersezioni.

2 - Scegliere l'opzione "All Simple Repeats records that have any overlap with UCSC Genes" e cliccare sul tasto "submit"

3 - Cliccare su "summary/statistics" per ottenere il numero di simple repeats identificate

4 - Scegliere l'opzione "hyperlinks to Genome Browser" nella sezione "output format" e cliccare il tasto "get output"

5 - Cliccare sul link "[trf at chr4:3076604-3076667](#)" (Gene HTT)

6 - Cliccare su [trf at chr12:7045880-7045938](#) (Gene ATN11)

7 - Cliccare su [trf at chr17:17697094-17697134](#) (Gene RAI1)

8 - Cliccare su [trf at chr19:46273463-46273524](#) (DMPK)

Creazione di custom tracks

1 - In “output format” nella pagina principale del table browser scegliere “custom track” e poi cliccare su “get output”

2 - Rinominare “SRepeatsGenes” la custom track e cambiare la descrizione a “Intersection of simple CAG repeats with Genes”. Infine, cliccare su “get custom track in genome browser”

Output simpleRepeat as Custom Track

Custom track header:

name= SRepeatsGenes

description= Intersection of simple CAG repeats with Genes

visibility= pack

url=

Create one BED record per:

☒ Whole Gene

☐ Upstream by 200 bases

☐ Downstream by 200 bases

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream regions are specified, only the region closest to the feature will be used.

get custom track in table browser get custom track in file

get custom track in genome browser cancel

3 - Spostarsi sul gene “HTT (Homo sapiens huntingtin (HTT), mRNA.)” e zoomare sul primo esone al 5’

4 - Tornare alla pagina principale del table browser e notare che le custom track sono disponibili per la creazione di filtri e intersezioni.

5 - Cliccare su “My Data” sulla barra di navigazione superiore e scegliere l’opzione “custom tracks” per visualizzare la pagina di gestione delle custom tracks.

6 - Collegarsi al sito di CompGen (http://compgen.bio.unipd.it/~stefania/Didattica/AA2021-2022/MMOL_BIOINFO_BE/MMOL_BIOINFO_BE.html) e scaricare il file esercitazione1.zip cliccando su “Guida” della riga “l’esercitazione BIOINFORMATICA”

7 - Decomprimere il file esercitazione1.zip e aprire il file BED con l’editor di testo:

- Informazioni riguardanti il display di default della nostra custom track
browser position chr4:56010000-56030000
browser pix 800
browser hide all
browser pack snp130
browser full knownGene

- Caratteristiche della track

track name="Items" description="Track per bioinfo2 bioevo" visibility=2
color=200,0,200 useScore=1 db=hg19

- Sequenze che verranno rappresentate dalla "custom track" in formato BED

chr4	56010000	56015000	Item1	900	+
chr4	56014000	56019000	Item2	200	+
chr4	56017000	56023000	Item3	800	-
chr4	56021000	56028000	Item4	300	-

9 - Cliccare sul tasto "add custom track" e incollare la custom track sull'apposito campo.

Add Custom Tracks

clade Mammal
genome Human
assembly Mar. 2006 (NCBI36/hg18)

Display your own data as custom annotation tracks in the browser. Data must be formatted in [BED](#), [bigBed](#), [bedGraph](#), [broadPeak](#), [narrowPeak](#), or [PSI](#) formats. To configure the display, set [track](#) and [browser](#) line attributes as described provided via only a URL or embedded in a track line in the box below. Publicly available custom tracks are listed [here](#).

Paste URLs or data:

Or upload: Choose File No file chosen

Submit

browser position chr4:56010000-56030000
browser pix 800
browser hide all
browser pack snp130
browser full knownGene
track name="Items" description="Track per bioinfo2 bioevo" visibility=2
color=200,0,200 useScore=1 db=hg18

Clear

Optional track documentation:

Or upload: Choose File No file chosen

Clear

Click [here](#) for an HTML document template that may be used for Genome Browser track descriptions.

10 - Cliccare chr4, la posizione di default della nostra custom track, per visualizzare gli elementi

Esercizio

Utilizzando le tabelle dell'UCSC Genome Browser:

1. Fare una custom track tramite il table browser per rappresentare le sottosequenze dell'isoforma "uc002gij.3" (tabella UCSC Gene) che si sovrappongono ad almeno un mRNA (tabella all_mrna) e vederle nel genome browser.
2. Ottenere la sequenza di DNA della custom track ed evidenziare in giallo le sequenze che si sovrappongono tra UCSC gene track e la track del trascritto "uc002gij.3". Cosa rappresentano le sequenze gialle?