Universitat Oberta de Catalunya

UOC Universitat Oberta de Catalunya

Universitat de Barcelona

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA Y BIOESTADÍSTICA

Genómica computacional

PEC1. Manejo de los navegadores genómicos

Noviembre de 2022

Alumno: Demetrio Muñoz Álvarez

Ejercicio 1. Anotar el gen de estudio [25%]

1. En primavera de 2017 conocimos la siguiente noticia a través de la mayoría de medios de comunicación. Leed el texto con atención, dado que utilizaremos esta información como base para nuestro ejercicio.

2. Conectaos a PUBMED y localizad el resumen (en inglés, Abstract) de esta publicación mencionada en el artículo anterior:
http://www.ncbi.nlm.nih.gov/pubmed

Con los nombres de los autores que menciona la noticia del apartado anterior podemos localizar el artículo al que hace referencia mediante el buscador PUBME. A continuación, se muestra una captura de pantalla con el título y Abstract de dicho artículo:

Differential Aging Analysis in Human Cerebral Cortex Identifies Variants in TMEM106B and GRN that Regulate Aging Phenotypes

Herve Rhinn [1], Asa Abeliovich [2]

Affiliations + expand

PMID: 28330615 DOI: 10.1016/j.cels.2017.02.009

Free article

Abstract

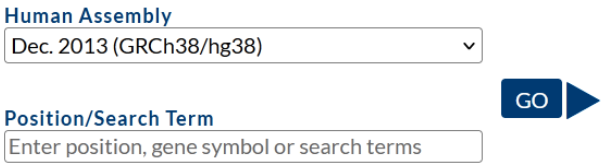
Human age-associated traits, such as cognitive decline, can be highly variable across the population, with some individuals exhibiting traits that are not expected at a given chronological age. Here we present differential aging (Δ-aging), an unbiased method that quantifies individual variability in age-associated phenotypes within a tissue of interest, and apply this approach to the analysis of existing transcriptome-wide cerebral cortex gene expression data from several cohorts totaling 1,904 autopsied human brain samples. We subsequently performed a genome-wide association study and identified the TMEM106B and GRN gene loci, previously associated with frontotemporal dementia, as determinants of Δ-aging in the cerebral cortex with genome-wide significance. TMEM106B risk variants are associated with inflammation, neuronal loss, and cognitive deficits, even in the absence of known brain disease, and their impact is highly selective for the frontal cerebral cortex of older individuals (>65 years). The methodological framework we describe can be broadly applied to the analysis of quantitative traits associated with aging or with other parameters.

Keywords: TMEM106B; aging; brain; genome-wide association study; genomics; inflammaging; inflammation; microglia; progranulin.

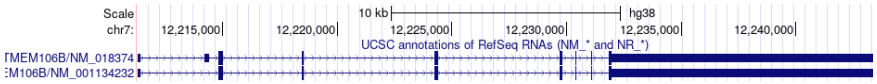
Copyright © 2017 The Authors. Published by Elsevier Inc. All rights reserved.

3. Con el servidor genómico UCSC, buscad las anotaciones del gen TMEM106B (Homo sapiens, hg38). Utilizaremos la anotación suministrada por RefSeq (track NCBI RefSeq, subtrack UCSC RefSeq). Anotad la localización genómica y los genes más cercanos: http://genome.ucsc.edu/

Dentro del servidor genómico UCSC podemos seleccionar con que genoma y versión de este queremos trabajar, a su vez podemos buscar una posición especifica dentro del genoma seleccionado.

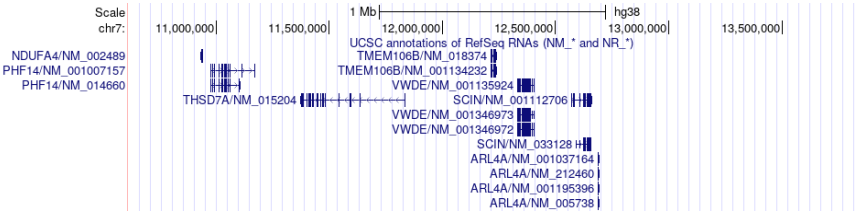


En “Position” vamos a introducir el gen TMEM106B, con lo que obtendremos una ventada de visualización con bastante información. En este ejercicio solo vamos a trabajar con la anotación “**RefSeq**”, se localiza justo debajo de la imagen de visualización en el apartado “**Genes and Gene predictions**”. Seleccionamos el track “**NCBI RefSeq**”, dentro de este track podemos añadir o eliminar subtracks. Para la solución de este apartado seleccionamos el subtrack “**USCD RefSeq**”. Con lo que obtendremos la siguiente imagen al visualizar el gen, vemos dos versiones del gen:



El gen se encuentra en la posición: **chr7:12,211,294-12,243,367**

En la ventana de visualización podemos hacer “**zoom out**” y ver los genes cercanos que se encuentran próximos a la región del gen que estamos visualizando, como se muestra en la siguiente imagen:



Los genes mas cercanos: **VWE, SCIN, ARL4A, THSD7A, PHF14, NDUF4**

4. Dentro de la ficha del gen según RefSeq, accede al registro de Entrez Gene y de OMIM para describir que funciones desempeña según Gene Ontology y en que enfermedades puede estar involucrado.

En la ventana de visualización del gen TMEM106B podemos seleccionar el track del gen, haciendo esto podemos acceder a "**RefSeq Gene**", donde se nos facilitan varios enlaces con información específica para este gen:

RefSeq Gene

RefSeq Gene TMEM106B

RefSeq: NM_018374.4 **Status:** Validated
Description: Homo sapiens transmembrane protein 106B (TMEM106B), transcript variant 1, mRNA.
CCDS: CCDS5358.1
CDS: 3' complete
OMIM: 613413
Entrez Gene: 54664
PubMed on Gene: TMEM106B
PubMed on Product: transmembrane protein 106B
GeneCards: TMEM106B
AceView: TMEM106B

En el enlace "**OMIM**" podemos localizar una explicación de las funciones del gen, como se muestra a continuación:

▼ **Description**

TMEM106B controls the size, number, motility, trafficking, and acidification of lysosomes (Klein et al., 2017).

► **Cloning and Expression**

► **Mapping**

▼ **Gene Function**

Chen-Plotkin et al. (2012) found that mature processed microRNAs (miRNAs) of the MIR132 (610016)/MIR212 (613487) cluster bound to 2 identical target sites in the 3-prime UTR of the TMEM106B transcript and downregulated TMEM106B mRNA and protein expression. TMEM106B expression was significantly increased, and MIR132/MIR212 expression significantly decreased, in FTLD-TDP brain compared with normal brain. Overexpression of TMEM106B in human cell lines and rodent neurons caused abnormalities in late endosome-lysosome morphology and impaired their acidification. Overexpression of TMEM106B also increased intracellular, but not extracellular, progranulin (PGRN, or GRN; 138945) level.

Brady et al. (2013) found that overexpression of human TMEM106B in N2A mouse neuroblastoma, T98G human glioblastoma, and NSC-34 mouse motor neuron cell lines induced lysosome enlargement and impaired lysosomal degradation of EGFR (131550) and PGRN. Enlarged lysosomes induced by TMEM106B overexpression retained the ability to fuse with incoming endosomes, and TMEM106B overexpression did not cause apoptosis. Knockdown of Tmem106b in N2A cells had no apparent effect on lysosome size or morphology.

Brady et al. (2014) found that TMEM106B underwent intramembrane proteolysis. The lysosomal luminal C-terminal domain of TMEM106B was removed by lysosomal cysteine proteinases to generate a stable N-terminal fragment of approximately 127 amino acids that included the transmembrane region and cytosolic N terminus. Subsequently, lysosomal membrane proteinase SPLL2A (608238) cleaved the N-terminal fragment at 2 places within the transmembrane segment around amino acid 106 to yield an unstable intracellular domain.

Using coimmunoprecipitation analysis, Klein et al. (2017) found that mouse Tmem106b interacted with the lysosomal V-ATPase via subunit AP1 (ATP6V0A1; 192130).

Seleccionando "**Entrez Gen**" accedemos a bastante información del gen, buscamos el apartado "**Gene Ontology**" y visualizamos las funciones y los procesos en lo que está involucrado el gen:

Gene Ontology Provided by GOA

Function	Evidence Code	Pubs
enables ATPase binding	IPI	PubMed
enables molecular_function	ND	
enables protein binding	IPI	PubMed

Process	Evidence Code	Pubs
involved_in dendrite morphogenesis	IBA	PubMed
involved_in dendrite morphogenesis	IMP	PubMed
involved_in lysosomal lumen acidification	IEA	
involved_in lysosomal protein catabolic process	IEA	
involved_in lysosomal transport	IBA	PubMed
involved_in lysosome localization	IBA	PubMed
involved_in lysosome localization	IMP	PubMed
involved_in lysosome organization	IBA	PubMed
involved_in neuron cellular homeostasis	IEA	
involved_in positive regulation of dendrite development	IEA	
involved_in positive regulation of hydrolase activity	IEA	
involved_in regulation of lysosome organization	IEA	

Buscando por la palabra clave "**disease**" podemos ver como el gen TMEM106B se relaciona con enfermedades de **degeneración neuronal** como la Demencia, Alzheimer o Esclerosis. Esta información ha sido consultada en los artículos de referencia que nos muestra la "**Entrez Gen**". Para buscar información de las enfermedades se ha usado el siguiente artículo:

- *Chang, Andrew et al. "Homotypic fibrillization of TMEM106B across diverse neurodegenerative diseases." Cell* vol. 185,8 (2022): 1346-1355.e15. doi:10.1016/j.cell.2022.02.026

5. Extraed de las anotaciones de RefSeq la región codificante (CDS, únicamente los exones) del gen TMEM106B humano. Repetid el mismo procedimiento para obtener la secuencia CDS ortóloga en el ratón (Mus musculus, mm10).

Devuelta a "**RefSeq Gene**" seleccionamos el enlace CCDS que nos llevara a la siguiente pantalla:

Consensus CDS Gene CCDS5358.1

Gene	TMEM106B
Description	transmembrane protein 106B
Sequences	CDS, protein, genomic
CCDS database	CCDS5358.1

Seleccionamos la opción CDS que nos devolverá la secuencia de la región codificante para el gen TMEM106B humano:

>CCDS5358.1 TMEM106B humano

ATGGGAAAGTCTCTTTCTCATTTGCCTTTGCATTCAAGCAAAGAAGATGCTTATGATGGAGTCACATCTGAAAACATG
AGGAATGGACTGGTTAATAGTGAAGTCCATAATGAAGATGGAAGAAATGGAGATGTCTCTCAGTTTCCATATGTGGA
ATTTACAGGAAGAGATAGTGTCACCTGCCCTACTTGTCAGGGAACAGGAAGAATTCCTAGGGGGCAAGAAAACCAA
CTGGTGGCATTGATTCCATATAGTGATCAGAGATTAAGGCCAAGAAGAACAAAGCTGTATGTGATGGCTTCTGTGTTT
GTCTGTCTACTCCTTTCTGGATTGGCTGTGTTTTTCCTTTTCCCTCGCTCTATCGACGTGAAATACATTGGTGTAAAATC
AGCCTATGTCAGTTATGATGTTCAGAAGCGTACAATTTATTTAAATATCACAAACACACTAAATATAACAAACAATA
ACTATTACTCTGTCGAAGTTGAAAACATCACTGCCCAAGTTCAATTTTCAAAAACAGTTATTGGAAAGGCACGCTTAA
ACAACATAACCATTATTGGTCCACTTGATATGAAACAAATTGATTACACAGTACCTACCGTTATAGCAGAGGAAATG
AGTTATATGTATGATTTCTGTACTCTGATATCCATCAAAGTGCATAACATAGTACTCATGATGCAAGTTACTGTGACA
ACAACATACTTTGGCCACTCTGAACAGATATCCCAGGAGAGGTATCAGTATGTCGACTGTGGAAGAAACACAACTTA
TCAGTTGGGGCAGTCTGAATATTTAAATGTACTTCAGCCACAACAGTAA

Repetiremos el mismo proceso para el gen TMEM106B, pero en este caso en el genoma del ratón y buscaremos la secuencia de la región codificante de este gen:

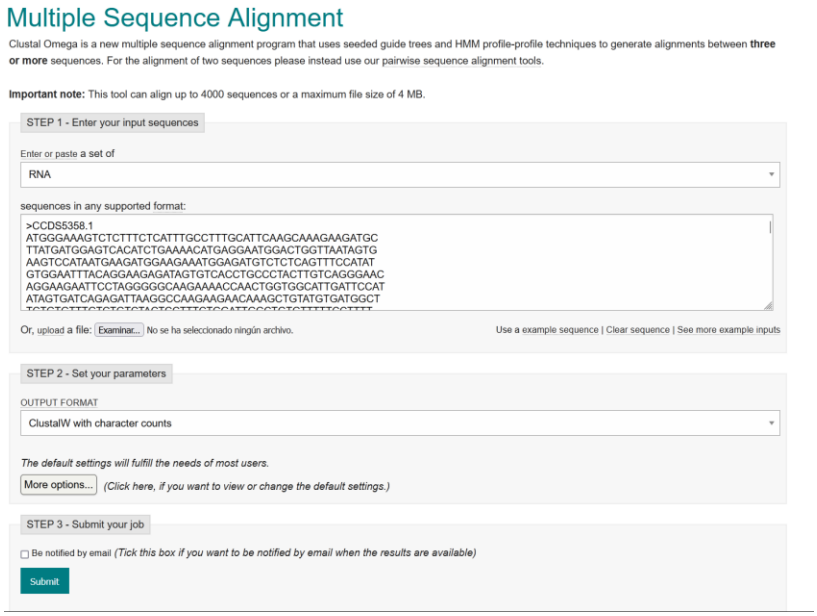
>CCDS19914.1 TMEM106B ratón

ATGGGAAAGTCTCTTTCTCACTTACCTTTGCATTCAAATAAAGAAGATGGCTATGATGGCGTTACATCGACAGACAAT
ATGAGAAATGGATTGGTTAGCAGTGAAGTGCACAACGAAGACGGAAGAAATGGAGATGTCTCTCAGTTCCCATATGT
GGAATTTACTGGAAGAGATAGTGTCACTTGTCCCACTTGCCAAGGAACAGGAAGAATTCCTAGGGGACAAGAAAACC
AACTGGTGGCATTGATTCCATATAGTGATCAGCGGTTACGGCCAAGAAGAACAAAGCTGTATGTGATGGCGTCTGTG
TTTGTCTGCCTGCTCCTGTCTGGATTGGCTGTGTTTTTTCTTTTCCCTCGATCTATTGAGGTGAAGTACATTGGAGTAA
AATCAGCCTATGTCAGCTACGACGCTGAAAAGCGAACCATATATTTAAATATCACGAACACACTAAATATAACAAAT
AATAACTATTATTCTGTTGAAGTTGAAAACATCACTGCTCAAGTCCAGTTTTCAAAAACCGTGATTGGAAAGGCTCGT
TTAAACAACATAACTAACATTGGCCCACTTGATATGAAGCAGATTGATTATACGGTACCCACAGTTATTGCAGAGGA
AATGAGTTACATGTATGATTTCTGTACACTGCTCTCCATCAAAGTGCACAACATAGTACTCATGATGCAAGTTACTGT
AACAACAGCATACTTTGGACACTCTGAGCAGATATCTCAGGAAAGGTACCAGTATGTCGACTGTGGAAGGAACACGA
CTTACCAGTTGGCCCAGTCTGAGTATCTAAATGTCCTTCAGCCACAACAATAA

6. El programa CLUSTAL Omega realiza alineamientos globales de dos o más secuencias. Conectaos al servidor de CLUSTAL Omega para alinear las dos secuencias CDS (humana y de ratón) obtenidas en el paso anterior.

http://www.ebi.ac.uk/Tools/msa/clustalo/

El enlace facilitado nos lleva a la herramienta "**CLUSTAL Omega**", con esta herramienta podremos alinear las secuencias CDS que hemos obtenido en el apartado anterior. En la herramienta vemos una serie de campos, en nuestro caso estamos trabajando con RNA lo cual debemos indicar en el campo "**STEP-1**", posteriormente pegaremos las secuencias CDS una de tras de otra. En "**STEP-2**" dejaremos la opción que aparece y por último en "**STEP-3**" lanzaremos el proceso para nuestras secuencias:



Cuando el proceso termina obtendremos la siguiente imagen, donde vemos el alineamiento de las dos secuencias que hemos introducido:

```
CLUSTAL O(1.2.4) multiple sequence alignment

CCDS5358.1       ATGGGAAAGTCTCTTTCTCATTTGCCTTTGCATTCAAGCAAAGAAGATGCTTATGATGGA   60
CCDS19914.1      ATGGGAAAGTCTCTTTCTCACTTACCTTTGCATTCAAATAAAGAAGATGGCTATGATGGC   60
                 ******************** ** *************  **********  ********

CCDS5358.1       GTCACATCT---GAAAACATGAGGAATGGACTGGTTAATAGTGAAGTCCATAATGAAGAT   117
CCDS19914.1      GTTACATCGACAGACAATATGAGAAATGGATTGGTTAGCAGTGAAGTGCACAACGAAGAC   120
                 ** *****    ** ** ***** ****** ******  ******** ** ** *****

CCDS5358.1       GGAAGAAATGGAGATGTCTCTCAGTTTCCATATGTGGAATTTACAGGAAGAGATAGTGTC   177
CCDS19914.1      GGAAGAAATGGAGATGTCTCTCAGTTCCCATATGTGGAATTTACTGGAAGAGATAGTGTC   180
                 ************************** ****************** ***************

CCDS5358.1       ACCTGCCCTACTTGTCAGGGAACAGGAAGAATTCCTAGGGGGCAAGAAAACCAACTGGTG   237
CCDS19914.1      ACTTGTCCCACTTGCCAAGGAACAGGAAGAATTCCTAGGGGACAAGAAAACCAACTGGTG   240
                 ** ** ** ***** ** ************************ ******************

CCDS5358.1       GCATTGATTCCATATAGTGATCAGAGATTAAGGCCAAGAAGAACAAAGCTGTATGTGATG   297
CCDS19914.1      GCATTGATTCCATATAGTGATCAGCGGTTACGGCCAAGAAGAACAAAGCTGTATGTGATG   300
                 ************************* * *** ******************************

CCDS5358.1       GCTTCTGTGTTTGTCTGTCTACTCCTTTCTGGATTGGCTGTGTTTTTCCTTTTCCCTCGC   357
CCDS19914.1      GCGTCTGTGTTTGTCTGCCTGCTCCTGTCTGGATTGGCTGTGTTTTTTCTTTTCCCTCGA   360
                 ** *************** ** ***** ********************* ***********

CCDS5358.1       TCTATCGACGTGAAATACATTGGTGTAAAATCAGCCTATGTCAGTTATGATGTTCAGAAG   417
CCDS19914.1      TCTATTGAGGTGAAGTACATTGGAGTAAAATCAGCCTATGTCAGCTACGACGCTGAAAAG   420
                 ***** ** ***** ******** ********************** ** ** * * * ***

CCDS5358.1       CGTACAATTTATTTAAATATCACAAACACACTAAATATAACAAACAATAACTATTACTCT   477
CCDS19914.1      CGAACCATATATTTAAATATCACGAACACACTAAATATAACAAATAATAACTATTATTCT   480
                 ** ** ** *************** ********************* *********** ***

CCDS5358.1       GTCGAAGTTGAAAACATCACTGCCCAAGTTCAATTTTCAAAAACAGTTATTGGAAAGGCA   537
CCDS19914.1      GTTGAAGTTGAAAACATCACTGCTCAAGTCCAGTTTTCAAAAACCGTGATTGGAAAGGCT   540
                 ** ********************* ***** ** *********** ** ***********

CCDS5358.1       CGCTTAAACAACATAACCATTATTGGTCCACTTGATATGAAACAAATTGATTACACAGTA   597
CCDS19914.1      CGTTTAAACAACATAACTAACATTGGCCCACTTGATATGAAGCAGATTGATTATACGGTA   600
                 ** *************** *  ***** *************** ** ******** ** ***

CCDS5358.1       CCTACCGTTATAGCAGAGGAAATGAGTTATATGTATGATTTCTGTACTCTGATATCCATC   657
CCDS19914.1      CCCACAGTTATTGCAGAGGAAATGAGTTACATGTATGATTTCTGTACACTGCTCTCCATC   660
                 ** ** ***** ****************** ****************** *** * *******

CCDS5358.1       AAAGTGCATAACATAGTACTCATGATGCAAGTTACTGTGACAACAACATACTTTGGCCAC   717
CCDS19914.1      AAAGTGCACAACATAGTACTCATGATGCAAGTTACTGTAACAACAGCATACTTTGGACAC   720
                 ******** ***************************** ****** ********** ***

CCDS5358.1       TCTGAACAGATATCCCAGGAGAGGTATCAGTATGTCGACTGTGGAAGAAACACAACTTAT   777
CCDS19914.1      TCTGAGCAGATATCTCAGGAAAGGTACCAGTATGTCGACTGTGGAAGGAACACGACTTAC   780
                 ***** ******** ***** ***** ********************* ***** *****

CCDS5358.1       CAGTTGGGGCAGTCTGAATATTTAAATGTACTTCAGCCACAACAGTAA       825
CCDS19914.1      CAGTTGGCCCAGTCTGAGTATCTAAATGTCCTTCAGCCACAACAATAA       828
                 *******  ******** *** ******* *************** ***
```

7. Repetid este mismo alineamiento global, utilizando ahora las correspondientes proteínas de cada gen (que previamente debéis recuperar de la entrada de RefSeq). Valorad el grado de homología entre estas dos secuencias tanto a nivel genómico como a nivel de proteína.

Devuelta a "**RefSeq Gene**" seleccionamos el enlace CCDS que nos llevara a la siguiente pantalla:

Consensus CDS Gene CCDS19914.1

Gene	Tmem106b
Description	transmembrane protein 106B
Sequences	CDS, protein, genomic
CCDS database	CCDS19914.1

En este caso vamos a seleccionar "**protein**" tanto en el genoma humano como en el de ratón, lo que nos dará las siguientes secuencias:

>CCDS19914.1_prot length=275

ratón

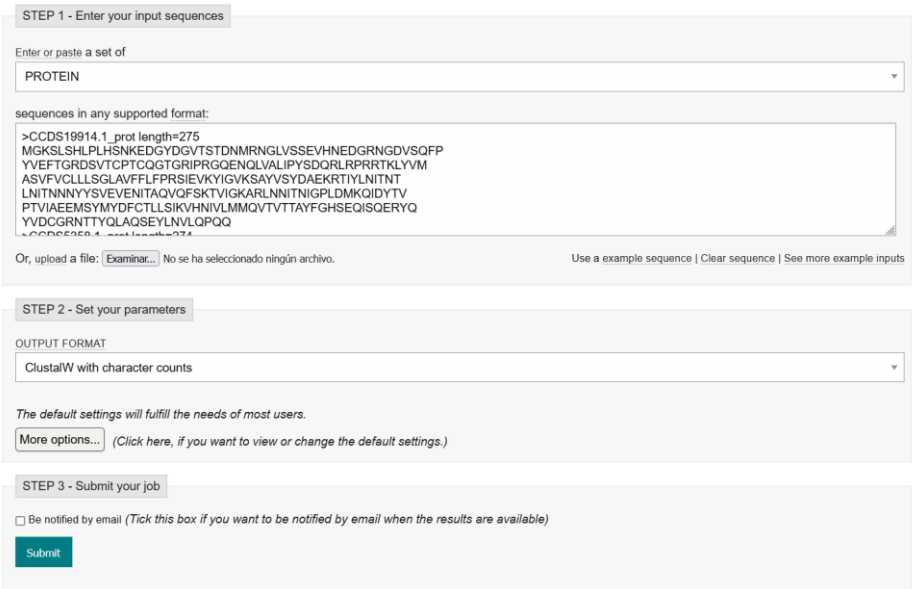
MGKSLSHLPLHSNKEDGYDGVTSTDNMRNGLVSSEV
HNEDGRNGDVSQFPYVEFTGRDSVTCPTCQGTGRIPR
GQENQLVALIPYSDQRLRPRRTKLYVMASVFVCLLLS
GLAVFFLFPRSIEVKYIGVKSAYVSYDAEKRTIYLNIT
NTLNITNNNYYSVEVENITAQVQFSKTVIGKARLNNIT
NIGPLDMKQIDYTVPTVIAEEMSYMYDFCTLLSIKVH
NIVLMMQVTVTTAYFGHSEQISQERYQYVDCGRNTT
YQLAQSEYLNVLQPQQ

>CCDS5358.1_prot length=274

humano

MGKSLSHLPLHSSKEDAYDGVTSENMRNGLVNSEVH
NEDGRNGDVSQFPYVEFTGRDSVTCPTCQGTGRIPRG
QENQLVALIPYSDQRLRPRRTKLYVMASVFVCLLLSG
LAVFFLFPRSIDVKYIGVKSAYVSYDVQKRTIYLNITN
TLNITNNNYYSVEVENITAQVQFSKTVIGKARLNNITII
GPLDMKQIDYTVPTVIAEEMSYMYDFCTLISIKVHNIV
LMMQVTVTTTYFGHSEQISQERYQYVDCGRNTTYQL
GQSEYLNVLQPQQ

Repetimos el proceso del apartado anterior en "**CLUSTAL Omega**", solo cambiando el campo en "**STEP-1**", en este caso trabajamos con proteínas, el resto del procedimiento es el mismo:



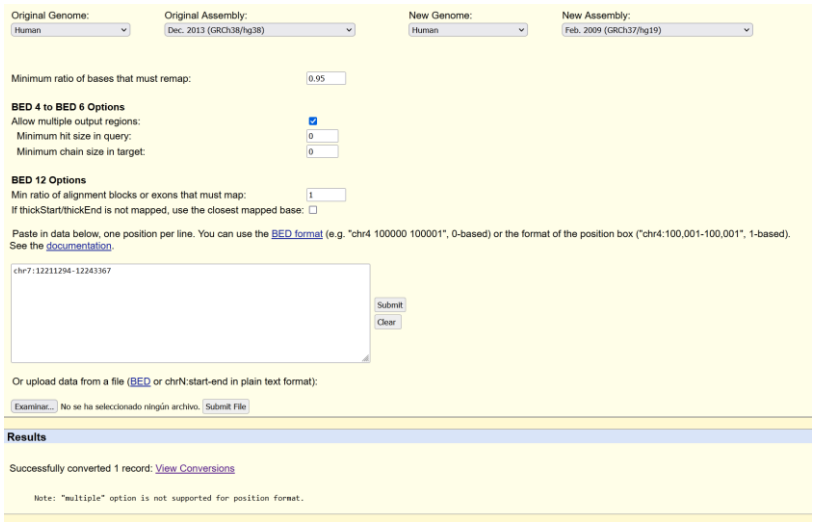
Terminado el proceso obtenemos la siguiente imagen para el alineamiento de proteínas:

```
CLUSTAL O(1.2.4) multiple sequence alignment


CCDS19914.1_prot      MGKSLSHLPLHSNKEDGYDGVTSTDNMRNGLVSSEVHNEDGRNGDVSQFPYVEFTGRDSV      60
CCDS5358.1_prot       MGKSLSHLPLHSSKEDAYDGVTS-ENMRNGLVNSEVHNEDGRNGDVSQFPYVEFTGRDSV      59
                      ************.***.****** :*******.***************************

CCDS19914.1_prot      TCPTCQGTGRIPRGQENQLVALIPYSDQRLRPRRTKLYVMASVFVCLLLSGLAVFFLFPR      120
CCDS5358.1_prot       TCPTCQGTGRIPRGQENQLVALIPYSDQRLRPRRTKLYVMASVFVCLLLSGLAVFFLFPR      119
                      ************************************************************

CCDS19914.1_prot      SIEVKYIGVKSAYVSYDAEKRTIYLNITNTLNITNNNYYSVEVENITAQVQFSKTVIGKA      180
CCDS5358.1_prot       SIDVKYIGVKSAYVSYDVQKRTIYLNITNTLNITNNNYYSVEVENITAQVQFSKTVIGKA      179
                      **.**************..*****************************************

CCDS19914.1_prot      RLNNITNIGPLDMKQIDYTVPTVIAEEMSYMYDFCTLLSIKVHNIVLMMQVTVTTAYFGH      240
CCDS5358.1_prot       RLNNITIIGPLDMKQIDYTVPTVIAEEMSYMYDFCTLISIKVHNIVLMMQVTVTTTYFGH      239
                      ****** ******************************.*****************.****

CCDS19914.1_prot      SEQISQERYQYVDCGRNTTYQLAQSEYLNVLQPQQ       275
CCDS5358.1_prot       SEQISQERYQYVDCGRNTTYQLGQSEYLNVLQPQQ       274
                      **********************.************
```

8. Emplead la herramienta LiftOver para averiguar las coordenadas de este gen en la versión hg19 del genoma humano y en la versión mm9 del genoma del ratón, respectivamente. https://genome.ucsc.edu/cgi-bin/hgLiftOver

Con la herramienta "**LiftOver**" podemos buscar las posiciones de los genes en otras versiones del mismo genoma o buscar las posiciones de los genes en otros genomas.

Partimos de nuestro genoma original, en este caso la versión h38 del genoma humano, lo especificamos en el campo "**Original Genome**" y "**Original Assembly**", luego introducimos en que genoma y versión queremos buscar la nueva posición de nuestro gen de interés en la versión h19 del genoma humano, en el campo "**New Genome**" y "**New Assembly**". Mas abajo introducimos las coordenadas originales del gen y lanzamos el proceso:



De la posición **chr7:12211294-12243367** de la versión h38 obtenemos que el gen se encuentra en la posición **chr7:12250920-12282993** en la versión h19 del genoma humano.

De la misma manera, repetimos el proceso cambiando los campos "**Original Genome**" y "**Original Assembly**" y "**New Genome**" y "**New Assembly**" al genoma del ratón para encontrar la posición del gen en la versión mm9 a partir de la versión mm10.

De la posición original de la versión mm10 **chr7:12250881-12276882** obtenemos la posición en la versión mm9 **chr7:12836230-12862231** del genoma de esta especie.

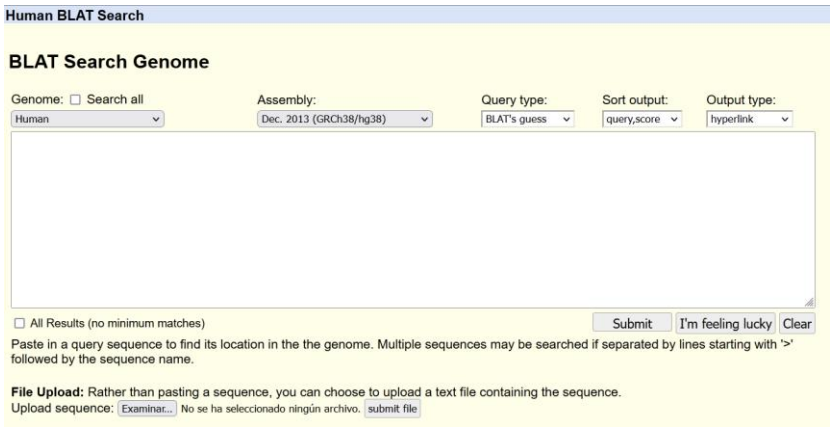
Ejercicio 2. La herramienta BLAT del navegador de UCSC [25%]

1. La aplicación BLAT es una herramienta muy popular disponible dentro del servidor genómico de UCSC. Localizad su página web y definid en pocas palabras cuál es la función principal de este programa:

http://genome.ucsc.edu

En la cabecera de opciones del servidor UCSC podemos localizar la herramienta "**BLAT**" en la pestaña Tools:

Una vez seleccionada la herramienta podemos visualizar su interfaz:



BLAT es un algoritmo que se encarga de encontrar regiones en el genoma similares a las secuencias con las que estamos trabajando.

2. Emplead BLAT para identificar la ubicación de la secuencia CDS humana del gen TMEM106B dentro del genoma humano (hg38, cromosoma, coordenadas).

En la herramienta "**BLAT**" introducimos la secuencia CDS obtenida anteriormente y lanzamos la búsqueda, lo que nos devuelve los siguientes resultados:

BLAT Search Results

Go back to chr7:12214811-12231975 on the Genome Browser.

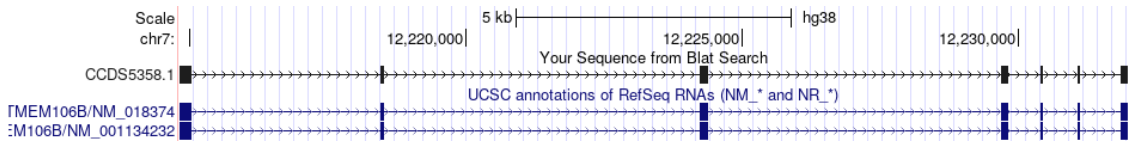
Custom track name: blat CCDS5358.1 (1)

Custom track description: blat on CCDS5358.1

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN	
browser details	CCDS5358.1	819	1	825	825	100.0%	chr7	+	12214811	12231975	17165	
browser details	CCDS5358.1	26	241	274	825	75.9%	chr10	-	105719014	105719042	29	
browser details	CCDS5358.1	22	12	33	825	100.0%	chr22_ML143380v1_fix	-	308706	308727	22	What is chrom fix?
browser details	CCDS5358.1	22	12	33	825	100.0%	chr21_ML143377v1_fix	-	328525	328546	22	What is chrom fix?
browser details	CCDS5358.1	22	12	33	825	100.0%	chr21	-	7132471	7132492	22	
browser details	CCDS5358.1	22	286	307	825	100.0%	chr2	+	30809266	30809287	22	
browser details	CCDS5358.1	22	112	133	825	100.0%	chr12	+	13388871	13388892	22	
browser details	CCDS5358.1	21	103	123	825	100.0%	chr12	+	88639591	88639611	21	
browser details	CCDS5358.1	20	137	156	825	100.0%	chr10	-	73466839	73466858	20	

Obtenemos una lista con todas las similitudes que ha encontrado BLAT, vamos a utilizar la que más "**Score**" tiene, además podemos ver en que cromosoma y posición se encuentra. Si de esta secuencia pulsamos en "browser" podemos visualizar la secuencia que ha generado BLAT en el buscador y compararla con el gen que estamos trabajando. Como se puede ver en la siguiente imagen la secuencia seleccionada coincide perfectamente con el gen TMEM106B que es de donde proviene la secuencia CDS que hemos utilizado:



3. Emplead BLAT para localizar con la secuencia CDS humana del gen TMEM106B dónde se encuentra la ubicación de este mismo gen en el genoma del ratón (mm10, cromosoma, coordenadas).

En este caso vamos a utilizar la misma secuencia CDS que el anterior apartado, sin embargo, vamos a marcar la casilla en "**Genome**" de "**Search all**" (podemos especificar directamente en que especie y versión queremos la búsqueda), lo que nos dará una lista de todas las especias que tienen alguna similitud con la secuencia CDS introducida. En

esta lista vamos a buscar la especie “**Mouse**” y “**mm10**”, de nuevo se nos muestra una lista con los “Scores” de la secuencia CDS para el genoma ratón, usamos la que más puntuación tiene:

BLAT Search Genome

Genome: ☑ Search all — Human | Assembly: Dec. 2013 (GRCh38/hg38) | Query type: BLAT's guess | Sort output: query,score | Output type: hyperlink

```
>CCDS5358.1
ATGGGAAAGTCTCTTTCTCATTTGCCTTTGCATTCAAGCAAAGAAGATGC
TTATGATGGAGTCACATCTGAAAACATGAGGAATGGACTGGTTAATAGTG
AAGTCCATAATGAAGATGGAAGAAATGGAGATGTCTCTCAGTTTCCATAT
GTGGAATTTACAGGAAGAGATAGTGTCACCTGCCCTACTTGTCAGGGAAC
AGGAAGAATTCCTAGGGGGCAAGAAAACCAACTGGTGGCATTGATTCCAT
ATAGTGATCAGAGATTAAGGCCAAGAAGAACAAAGCTGTATGTGATGGCT
TCTGTGTTTGTCTGTCTACTCCTTTCTGGATTGGCTGTGTTTTTCCTTTT
CCCTCGCTCTATCGACGTGAAATACATTGGTGTAAAATCAGCCTATGTCA
GTTATGATGTTCAGAAGCGTACAATTTATTTAAATATCACAAACACACTA
AATATAACAAACAATAACTATTACTCTGTCGAAGTTGAAAACATCACTGC
CCAAGTTCAATTTTCAAAAACAGTTATTGGAAAGGCACGCTTAAACAACA
TAACCATTATTGGTCCACTTGATATGAAACAAATTGATTACACAGTACCT
ACCGTTATAGCAGAGGAAATGAGTTATATGTATGATTTCTGTACTCTGAT
```

☐ All Results (no minimum matches) — Submit | I'm feeling lucky | Clear

CCDS5358.1	Proboscis monkey	nasLar1	37	chrUn_JMHX01214576v1
CCDS5358.1	Pika	ochPri3	35	JH802073
CCDS5358.1	Mouse	mm10	34	chr6
CCDS5358.1	Manatee	triMan1	30	JH594848
CCDS5358.1	Rat	rn6	27	chr4
CCDS5358.1	Rock hyrax	proCap1	22	scaffold_7457
CCDS5358.1	S. cerevisiae	sacCer3	19	chrXII

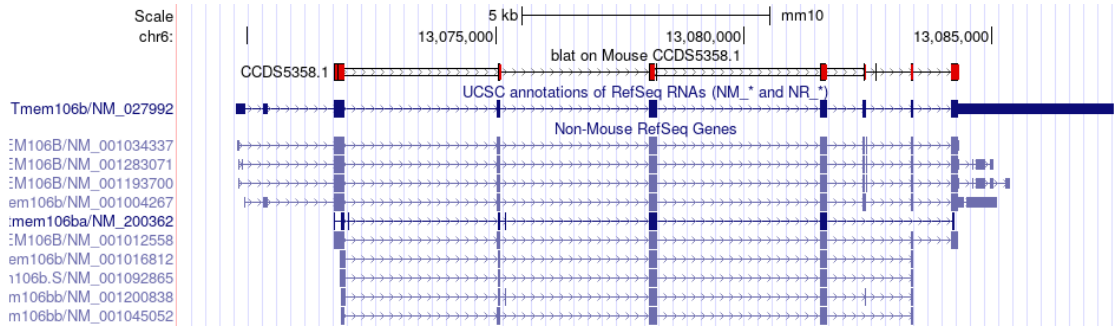
Custom track name: blat Mouse CCDS5358.1

Custom track description: blat on Mouse CCDS5358.1

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	CCDS5358.1	605	1	825	825	90.3%	chr6	+	13071744	13084326	12583
browser details	CCDS5358.1	25	116	146	825	96.3%	chr4	+	51834959	51834991	33
browser details	CCDS5358.1	23	432	454	825	100.0%	chr6	+	39862355	39862377	23
browser details	CCDS5358.1	22	424	445	825	100.0%	chr18	-	81910476	81910497	22

Podemos visualizar en el navegador la secuencia CDS humano que coincide parcialmente con el gen en el genoma del ratón (mm10), en la posición **chr6: 13071744-3084326** (el color rojo en la secuencia nos indica que el genoma y la secuencia buscada tiene diferentes bases en esas posiciones):



4. Emplead BLAT para localizar con la secuencia CDS humana del gen TMEM106B dónde se encuentra la ubicación de este mismo gen en el genoma del pollo (galGal4, cromosoma, coordenadas).

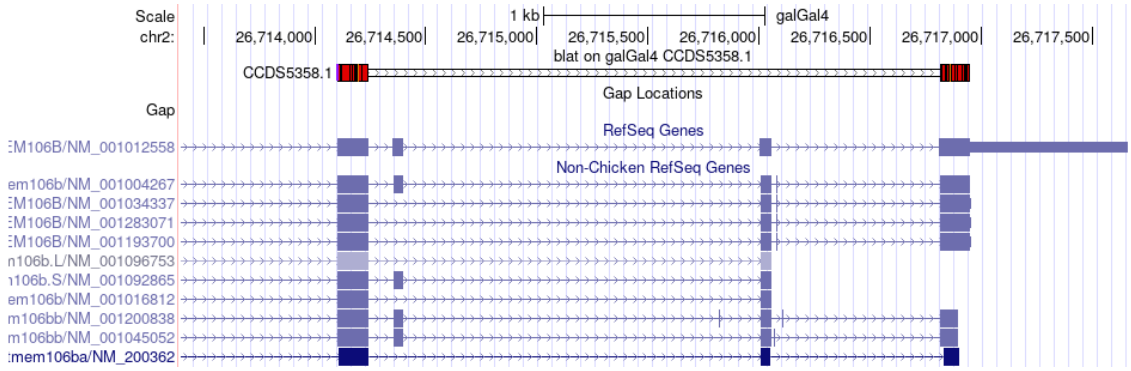
Repetimos la operación anterior esta vez con el genoma del pollo, en la versión gelGal4, la secuencia se encuentro en la posición **chr2: 26714104-26716947**. La visualizamos también en el navegador:

Custom track name: blat galGal4 CCDS5358.1

Custom track description: blat on galGal4 CCDS5358.1

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	CCDS5358.1	199	442	825	825	87.0%	chr2	+	26714104	26716947	2844
browser details	CCDS5358.1	21	295	315	825	100.0%	chr17	+	2890161	2890181	21
browser details	CCDS5358.1	20	481	500	825	100.0%	chr1	-	181592342	181592361	20
browser details	CCDS5358.1	20	683	702	825	100.0%	chr1	-	315556	315575	20
browser details	CCDS5358.1	20	399	418	825	100.0%	chr11	+	9398973	9398992	20
browser details	CCDS5358.1	19	10	28	825	100.0%	chr1	-	119351928	119351946	19
browser details	CCDS5358.1	19	150	168	825	100.0%	chr1	-	86318390	86318408	19
browser details	CCDS5358.1	19	21	39	825	100.0%	chr22	+	274816	274834	19
browser details	CCDS5358.1	19	654	672	825	100.0%	chr1	+	117514051	117514069	19
browser details	CCDS5358.1	19	118	136	825	100.0%	chr1	+	107887569	107887587	19



5. Ahora emplead BLAT con la proteína humana del gen TMEM106B sobre el mismo genoma del pollo (galGal4, cromosoma, coordenadas). Razonad sobre las diferencias entre este resultado y el obtenido en el punto anterior.

Para este apartado cambiamos la secuencia CDS y usamos la proteína del gen sobre el genoma del pollo, galGal4. Obteniendo lo siguiente:

Custom track name: blat galGal4 CCDS5358.1_prot

Custom track description: blat on galGal4 CCDS5358.1_prot

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	CCDS5358.1_prot	508	23	274	274	89.3%	chr2	++	26708501	26716944	8444
browser details	CCDS5358.1_prot	95	41	94	274	79.7%	chrUn_JH375554	+-	90879	91154	276

En la secuencia de proteínas obtenemos un mayor "**Score**" y una mayor similitud que si usamos la secuencia CDS, puede deberse a que la expresión de proteínas coincida aun no teniendo el mismo código de bases. Es decir, en la secuencia CDS tenemos diferentes bases al comparar con el genoma objetivo, aun así, la expresión de la proteína será la misma, aunque se difiera en alguna base. Lo que podría explicar la diferencia entre las dos búsquedas.

Ejercicio 3. La herramienta Table browser del navegador de UCSC [40%]

1. La aplicación Table browser es una herramienta muy útil para acceder a los datos de las pistas que constituyen el entorno gráfico disponible dentro del servidor genómico de UCSC. Localizad su página web y definid en pocas palabras cuál es la función principal de este programa: http://genome.ucsc.edu

En la cabecera de opciones del servidor UCSC podemos localizar la herramienta "Table Browser" en la pestaña Tools:

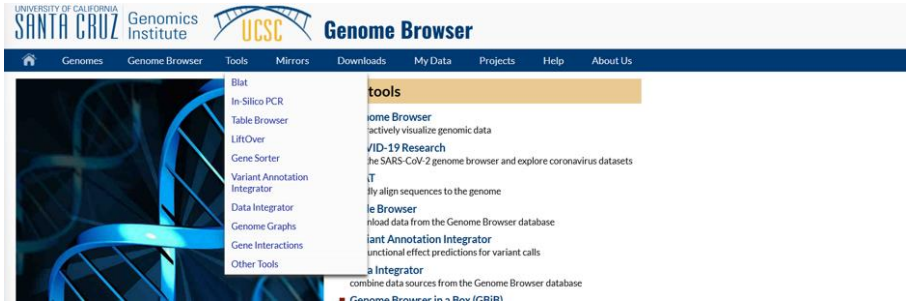


Table browser nos permite tener una interfaz grafica para realizar consultas y manipular los resultados que vayamos añadiendo en el buscador genómico. Podemos obtener información específica del genoma completo o de una parte especifica, con esto podemos crear nuestros propios ficheros que posteriormente podremos añadir al buscador genómico y visualizar nuestras propias tracks.

2. Imaginemos un escenario real: estamos trabajando con el genoma humano (ensamblado hg38) y se nos plantean una serie de cuestiones prácticas. Encontrad la manera de responder a estas preguntas empleando el navegador de tablas sobre la pista RefSeq genes (track NCBI RefSeq, table UCSC RefSeq). Debéis añadir una breve descripción en el informe de cómo habéis logrado llegar a la solución.

En la interfaz de "**Table browser**" vamos a rellenar los campos según nos especifica el ejercicio para ir resolviendo las cuestiones:

Número de pares de bases del genoma completo

Con los parámetros que hemos establecido en la cuestión anterior pulsamos el botón inferior "**summary/statistics**" nos devuelve dos tablas de información sobre el genoma o la región que especifiquemos. En este caso hemos consultado el genoma humano completo en la versión hg38. Obteniendo las siguientes tablas:

refGene (refGene) Summary Statistics

item count	88,819
item bases	1,449,357,225 (46.59%)
item total	5,762,756,485 (185.25%)
smallest item	21
average item	64,882
biggest item	2,298,757
block count	896,726
block bases	97,040,402 (3.12%)
block total	288,493,938 (9.27%)
smallest block	1
average block	322
biggest block	91,671

Region and Timing Statistics

region	genome
bases in region	3,272,116,950
bases in gaps	161,348,343
load time	0.17
calculation time	1.05
free memory time	0.00
filter	off
intersection	off

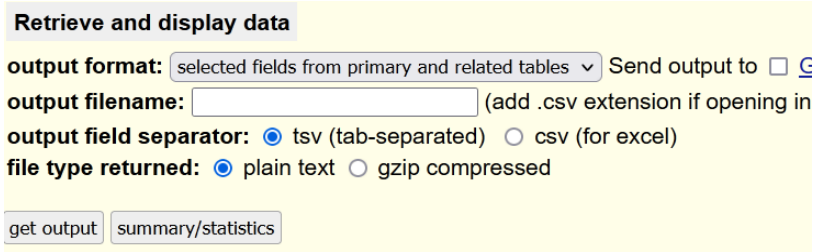
“**Bases in region**” nos facilita el numero de pares de bases en el genoma completo, 3.272.116.950 pares de bases.

Número de transcritos en total a lo largo del genoma.

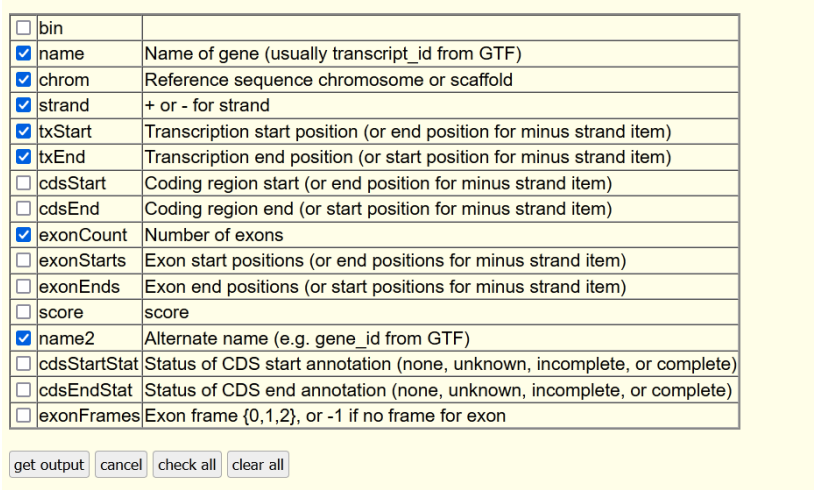
El numero total de transcritos nos lo da el dato “**item count**”, con 88.819 transcritos en el genoma.

Listado de todos los transcritos en el genoma (únicamente captura de los primeros cinco). Debéis mostrar solo los siguientes atributos: código de RefSeq del transcrito, cromosoma, hebra, inicio/final del transcrito, número de exones y nombre del gen.

De nuevo, en la interfaz inicial de esta herramienta en el apartado “Retrieve and display data” vamos a elegir en “output format” la opción “selected fields from primary...”:



Finalmente pulsamos el botón “**get output**”, nos dirigimos a la siguiente tabla donde podemos especificar las características que queremos obtener de los transcritos para el genoma humano completo:



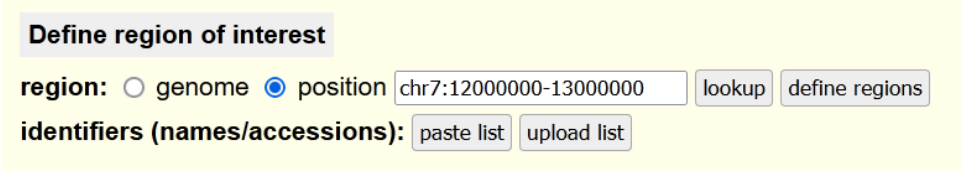
☐	bin	
☑	name	Name of gene (usually transcript_id from GTF)
☑	chrom	Reference sequence chromosome or scaffold
☑	strand	+ or - for strand
☑	txStart	Transcription start position (or end position for minus strand item)
☑	txEnd	Transcription end position (or start position for minus strand item)
☐	cdsStart	Coding region start (or end position for minus strand item)
☐	cdsEnd	Coding region end (or start position for minus strand item)
☑	exonCount	Number of exons
☐	exonStarts	Exon start positions (or end positions for minus strand item)
☐	exonEnds	Exon end positions (or start positions for minus strand item)
☐	score	score
☑	name2	Alternate name (e.g. gene_id from GTF)
☐	cdsStartStat	Status of CDS start annotation (none, unknown, incomplete, or complete)
☐	cdsEndStat	Status of CDS end annotation (none, unknown, incomplete, or complete)
☐	exonFrames	Exon frame {0,1,2}, or -1 if no frame for exon

Volvemos a pulsar el botón “**get output**” y nos mostrara el listado de transcritos con las características que hemos definido anteriormente:

#name	chrom	strand	txStart	txEnd	exonCount	name2
NM_000299	chr1	+	201283451	201332993	15	PKP1
NM_001276351	chr1	-	67092165	67134970	8	C1orf141
NM_001005337	chr1	+	201283505	201332989	14	PKP1
NM_001276352	chr1	-	67092165	67134970	9	C1orf141
NR_075077	chr1	-	67092165	67134970	10	C1orf141

Número total y listado de los transcritos en la región chr7:12000000-13000000.

En este apartado vamos a realizar un procedimiento similar, en el interfaz principal de la herramienta, en vez de elegir el genoma completo vamos a especificar la siguiente región:



Obtenemos 16 transcritos para esta región:

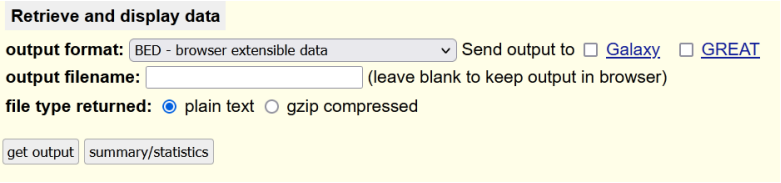
item count	16
item bases	243,432 (24.34%)
item total	739,829 (73.98%)
smallest item	3,326
average item	46,239
biggest item	89,462
block count	197
block bases	32,945 (3.29%)
block total	81,698 (8.17%)
smallest block	50
average block	415
biggest block	11,531

Realizamos el proceso del apartado anterior con esta región definida y obtenemos la lista con los 16 transcritos de estas coordenadas:

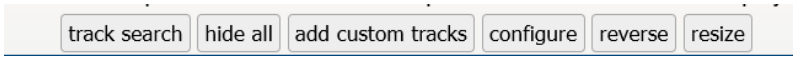
#name	chrom	strand	txStart	txEnd	exonCount	name2
NM_018374	chr7	+	12211293	12243367	9	TMEM106B
NM_001134232	chr7	+	12211293	12243367	8	TMEM106B
NM_001135924	chr7	-	12330884	12403865	29	VWDE
NM_001346973	chr7	-	12330884	12403941	27	VWDE
NM_001346972	chr7	-	12330884	12403941	27	VWDE
NR_144534	chr7	-	12330884	12403941	30	VWDE
NR_136261	chr7	+	12496428	12541135	6	LOC102725191
NR_136262	chr7	+	12497245	12541135	4	LOC102725191
NR_136263	chr7	+	12504396	12541135	4	LOC102725191
NM_001112706	chr7	+	12570719	12660181	16	SCIN
NR_156701	chr7	+	12570719	12660181	15	SCIN
NM_033128	chr7	+	12589522	12653603	14	SCIN
NM_001037164	chr7	+	12686826	12690933	2	ARL4A
NM_212460	chr7	+	12686826	12690933	2	ARL4A
NM_001195396	chr7	+	12687285	12690933	2	ARL4A
NM_005738	chr7	+	12687632	12690958	2	ARL4A

Extraed la custom track con los datos de la pregunta anterior (en formato BED), para añadirle una cabecera (track name=...) que os permita darle el nombre y el color que os resulte más atractivo. Cargar posteriormente la pista en el navegador (hg38) y comprobad que los exones encajan con los visualizados en la pista RefSeq original.

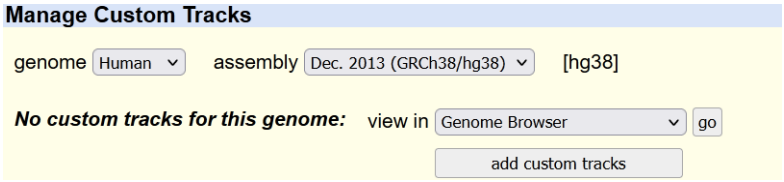
Con los datos anteriores vamos a cambiar el "output format" para obtener un BED. El URL de esta petición la guardamos para crear nuestra custom track.



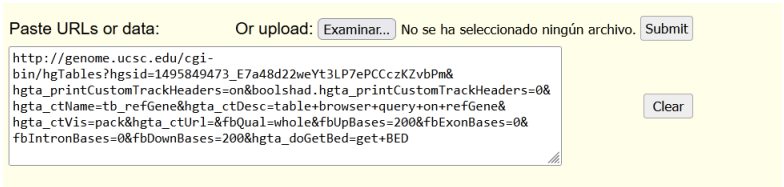
Para crear la custom track nos dirigimos al navegador genómico y buscamos la cabecera donde aparecen las opciones de las distintas Tracks, pulsaremos el botón "add custom tracks":



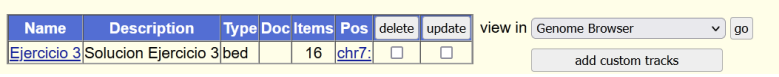
Visualizaremos las siguientes opciones y volveremos a pulsar el botón "add custom tracks":



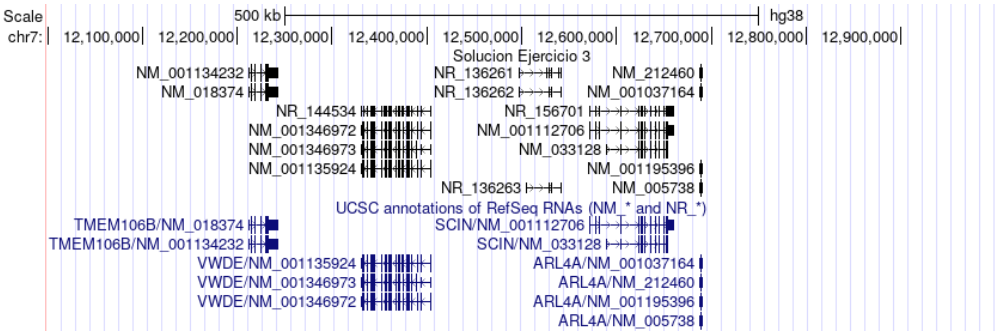
De nuevo, nos mostrara otra pantalla con varias opciones y varios campos, la URL con el BED que hemos generado la copiamos en el campo "**Paste URL sor data**" y pulsaremos el botón "**Submit**":



Ya tenemos creado la “Custom track” creada, pulsamos el botón “go” seleccionando la opción “Genome Browser”.

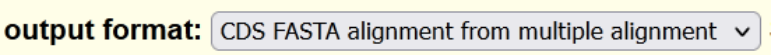


En este punto deberíamos observar nuestra “custom track” (color negro) con los 16 transcritos en el navegador genómico y podremos compararlos con la región que hemos usado para estos ejercicios (color azul):

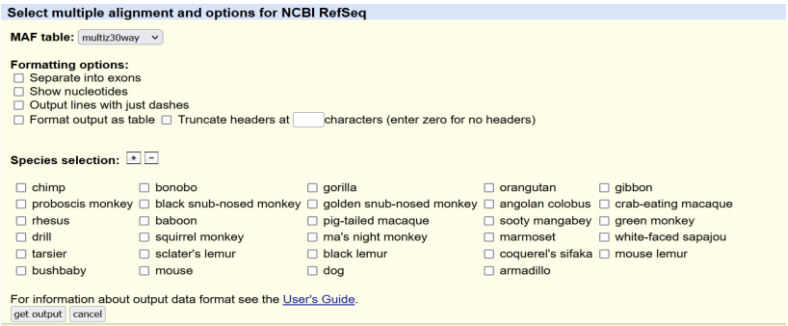


Extraed ahora la secuencia CDS de todos estos transcritos con el Table browser. Escoged una de las secuencias al azar y realizad un BLAT con ella para comprobar que encaja correctamente con el transcrito de RefSeq anotado en esa localización.

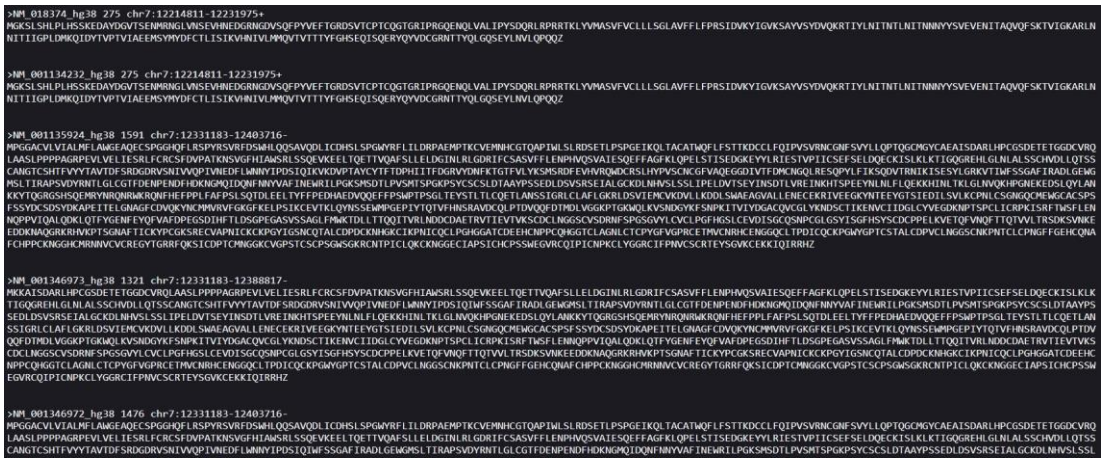
Con los datos anteriores vamos a cambiar el “output format” a “CDS FASTA alignment...” para obtener un listado de las secuencias CDS de los 16 transcritos:



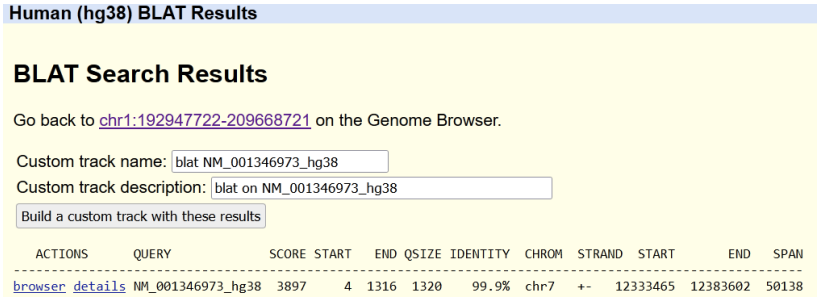
Pulsamos “**get ouput**” nos redirigimos a la siguiente pantalla, donde no seleccionaremos ninguna especia especifica. Pulsamos “**get ouput**” y nos mostrara la lista con las secuencias:



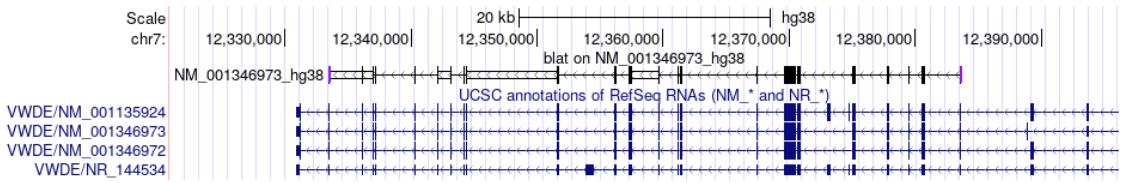
De la lista de los transcritos mostraremos una muestra de las primeras secuencias CDS:



Para realizar el último paso elegimos la secuencia del transcrito NM_001346973_hg38 con la que ejecutaremos el proceso BLAT:



Creamos la Custom track con este BLAT y mostramos el navegador genomico para comparar este transcrito con el transcrito anotado en esa posicion para RefSeq:



Ejercicio 4. La herramienta Biomart del navegador ENSEMBL [10%]

1. Estudiad el modo de funcionamiento de la herramienta Biomart. Mostrad un ejemplo de cómo interrogar el genoma humano (hg38) para la búsqueda de datos sobre la región chr12:7,680,240-7,905,217. Por ejemplo, cómo obtener el listado de términos de Gene ontology para los genes que se encuentran en ese lugar del genoma o el listado de SNPs anotados en los mismos genes. http://www.ensembl.org/biomart/martview