

PEC 1 - Genómica computacional

Alejandro Keymer

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>

Al revisar el artículo es fácil hacer una simple búsqueda con el título y los autores en PubMed y obtener la ficha del artículo en referencia.

The screenshot shows a web browser window with the following details:

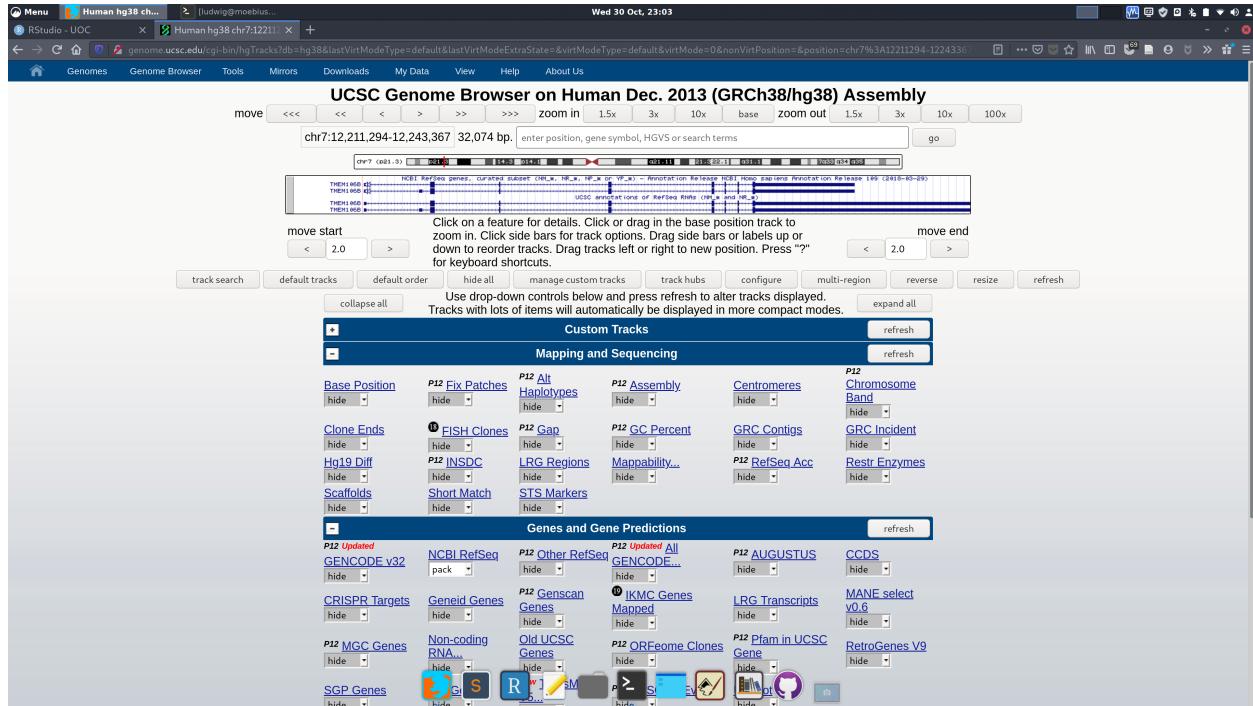
- Address Bar:** https://www.ncbi.nlm.nih.gov/pubmed/?term=Differential+Aging+Analysis+in+Human+Cerebral+Corex+Identifies+Variants+in+TMEM106B+and+GRN+that+Regulate+Aging+Phenotypes
- Search Bar:** Differential Aging Analysis in Human Cerebral Cortex Identifies Variants in TMEM106B and GRN that Regulate Aging Phenotypes
- Left Sidebar:** Shows 'Format: Abstract' and 'See 1 citation found by title matching your search: Cell Syst., 2017 Apr 26;4(4):404-415.e5. doi: 10.1016/j.cels.2017.02.009. Epub 2017 Mar 18.'
- Article Summary:** 'Differential Aging Analysis in Human Cerebral Cortex Identifies Variants in TMEM106B and GRN that Regulate Aging Phenotypes.' by Bhim H¹, Abeliovich A².
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 specific 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.
- Right Sidebar:** Includes sections for 'Full text links', 'Save items', 'Similar articles', 'Cited by 19 PubMed Central articles', and 'See reviews...'.

3. Con el servidor genómico UCSC, buscad las anotaciones del gen TMEM106 (*Homo sapiens*, hg38). Utilizaremos la anotación suministrada por RefSeq (track RefSeq Genes llamada también UCSC RefSeq). Anotad la localización genómica y los genes más cercanos:

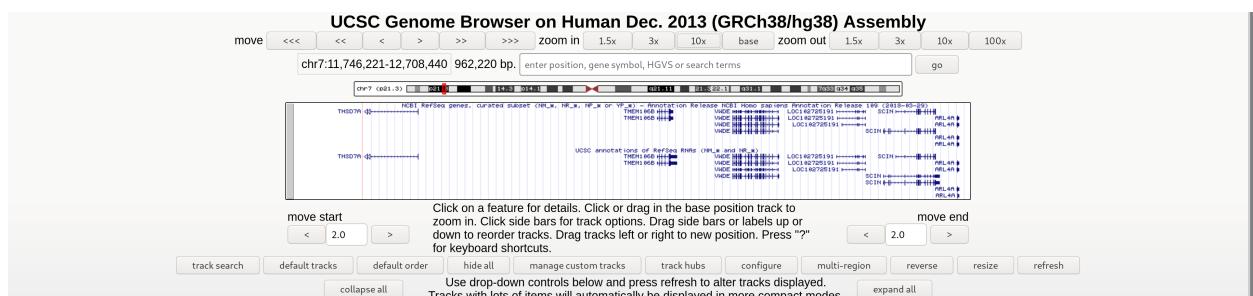
<http://genome.ucsc.edu>

En primer lugar utilice el navegador genómico para hacer seleccionar la base *Homo sapiens*, hg38 y hacer la búsqueda del gen TMEM106B. Luego filtro las vistas para que sólo aparezcan las pistas de RefSeq y accedo a la información detallada. De la pantalla principal del navegador o de la del detalle de RefSeq podemos conocer la ubicación del gen, que tiene dos variantes.

La ubicación genómica del gen TMEM106 es: chr7:12,211,294 - 12,243,367



Para poder apreciar mejor la región donde se encuentra este gen se puede hacer un *zoom out*, lo que permite apreciar algunos de los genes que se encuentran antes y después del gen TMEM106B. Algunos de los genes cercanos en la región son:



- **PHF14**, posición chr7:10973872-11107749. *PHD finger protein 14*
- **THSD7A**, posición chr7:11370365-11832198. *thrombospondin type 1 domain containing 7A*
- **VWDE**, posición chr7:12330885-12403941. *von Willebrand factor D and EGF domains*
- **LOC102725191**, no caracterizado (aún...)
- **SCIN**, posición chr7:12570720-12660181. *scinderin*
- **ARL4A**, posición chr7:12686827-12690933. *ADP ribosylation factor like GTPase 4A*

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 qué enfermedades puede estar involucrado.

Desde la ficha del gen podemos acceder al enlace directo a la página de *Entrez Gene*. En la sección de *Gene Ontology* aparece la función del gen en un formato reducido. Una de las funciones que se puede destacar es la de *ligando de proteínas* que tiene un enlace a un artículo de PubMed (<https://www.ncbi.nlm.nih.gov/pubmed/24357581>). El nombre del artículo (y el abstract) hablan de una “función en el control del tráfico dendrítico de lysosomas”.

RefSeq Gene THSD7A

RefSeq: NM_015204.3 **Status:** Reviewed
Description: Homo sapiens thrombospondin type 1 domain containing 7A (THSD7A), mRNA.
CCDS: CCDS47543.1
CDS: Full length
OMIM: 612249
Entrez Gene: 221981
PubMed on Gene: THSD7A
PubMed on Product: thrombospondin type-1 domain-containing protein 7A precursor
GeneCards: THSD7A
GeneView: THSD7A

Summary of THSD7A

The protein encoded by this gene is found almost exclusively in endothelial cells from placenta and umbilical cord. The encoded protein appears to interact with alpha(V)beta(3) integrin and paxillin to inhibit endothelial cell migration and tube formation. This protein may be involved in cytoskeletal organization. Variations in this gene may be associated with low bone mineral density in osteoporosis. [provided by RefSeq, Aug 2010]. Sequence Note: This RefSeq record was created from transcript and genomic sequence data to make the sequence consistent with the reference genome assembly. The genomic coordinates used for the transcript record were based on transcript alignments. Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene record to access additional publications. ##Evidence-Data-START## RNAseq introns :: mixed/partial sample support SAMEA1965299, SAMEA1966682 [ECO:0000350] ##Evidence-Data-END## ##RefSeq-Attributes-START## MANE Ensembl match :: ENST00000423059.9/ENSP00000406482.2 RefSeq Select criteria :: based on conservation, expression, longest protein ##RefSeq-Attributes-END##

mRNA/Genomic Alignments

BROWSER	SIZE	IDENTITY	CHROMOSOME	STRAND	START	END	QUERY	START	END	TOTAL
browser	10655	108.0%	7	-	11378365	11832198	NM_015204	1	10655	10655

[View details of parts of alignment within browser window.](#)

Position: chr7:11370365-11832198
Band: 7p21.3
Genomic Size: 461834
Strand: -
Gene Symbol: THSD7A
CDS Start: complete
CDS End: complete

Links to sequence:

La sección **GeneRif** permite obtener otros resultados que relacionan el gen con funciones específicas. En esta sección destaca la asociación que se hace de:

- Variaciones genéticas del gen se asocian a diferentes fenotipos de envejecimiento.
- Asociado a hipomielinización cerebral.
- Asociado a la demencia fronto-temporal.

La ficha de OMIM especifica más las funciones;

- El gen TMEM106B controla el tamaño, el número, la movilidad, el tráfico y la acidificación de los lisosomas (Klein et al., 2017)
- La sobreexpresión del gen TMEM106B se asocia a anomalías en las fases tardías de la morfología de los endosomas y lisosomas... (Chen-Plotkin et al., 2012)
- La sobre expresión del gen TMM106B se asoció a Neuroblastoma N2A en ratones, Glioblastoma T98G en humanos y en otras enfermedades de los lisosomas en ratones. (Brady et al., 2013)
- Se ha asociado una mutación del gen TMEM106B a la Leucodistrofia hipomielinizante 16 (Simons et al. 2017)

Como se puede concluir, la función del gen está relacionada con el funcionamiento y regulación de los lisosomas. Alteraciones de este gen se relacionan con un desequilibrio en la función de los lisosomas que se puede traducir a diferentes condiciones; desde un fenotipo de envejecimiento diferente, a enfermedades neuro-degenerativas graves

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).

Para extraer la región codificante, en primer lugar seccionamos el gen desde el navegador. En la página del detalle del gen, podemos seleccionar la sección **Genomic Sequence**. En esta sección podemos configurar una consulta para que nos devuelva solo la secuencia codificante CDS.

Get Genomic Sequence Near Gene

Note: if you would prefer to get DNA for more than one feature of this track at a time, try the [Table Browser](#) using the output format sequence.

Sequence Retrieval Region Options:

- Promoter/Upstream by bases
- 5' UTR Exons
- CDS Exons
- 3' UTR Exons
- Introns
- Downstream by bases
- One FASTA record per gene.
- One FASTA record per region (exon, intron, etc.) with extra bases upstream (5') and extra downstream (3')
- Split UTR and CDS parts of an exon into separate FASTA records

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:

- Exons in upper case, everything else in lower case.
- CDS in upper case, UTR in lower case.
- All upper case.
- All lower case.
- Mask repeats: to lower case to N

The yellow area contains the following DNA sequence:

```

>hg38_refGene_NM_001134232 range=chr7:12214811-12231975 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATGGGAAAGTCTCTTCTATTGCCTTGCAATTCAAAGCAAAGAACATGC
TTATGATGGAGTCACATCTGAAAACATGAGGAATGGACTGGTTAATAGTG
AAGTCCATAATGAAGATGGAAGAAATGGAGATGTCTCTCAGTTCCATAT
GTGGAATTACAGGAAGAGATAGTGTACCTGCCCTACTTGTCAAGGAAC
AGGAAGAATTCTAGGGGCAAGAAAACCACACTGGTGGCATTGATTCCAT
ATAGTGATCAGAGATTAAGGCCAAGAACAAAGCTGTATGTGATGGCT
TCTGTGTTGCTGTCTACTCCTTCTGGATTGGCTGTGTTTCCTTT
CCCTCGCTCTATCGACGTAAACATGGTGTAAAATCAGCCTATGTCA
GTTATGATGTTCAGAACGGTACAATTATTTAAATATCACAAACACACTA
AATATAACAAACAATAACTATTACTCTGCGAACATTGAAACATCACTGC
CCAAGTTCAATTTCAAAAACAGTTATTGGAAAGGCACGCTTAAACAACA
TAACCATTATTGGTCACTTGATATGAAACAAATTGATTACACAGTACCT
ACCGTTATAGCAGAGGAATGAGTTATATGTATGATTTCTGTACTCTGAT
ATCCCATCAAAGTGCATAACATAGTACTCATGATGCAAGTTACTGTGACAA
CAACATACTTGGCACTCTGAACAGATATCCCAGGAGAGGTATCAGTAT
GTCGACTGTGAAAGAACACAACATTATCAGTTGGGGCAGTCTGAATATT
AAATGTAACCTCAGCCACAACAGTAA

```

- La secuencia de exones de TMEM106B en humano es:

```
>hg38_refGene_NM_001134232 range=chr7:12214811-12231975 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATGGGAAAGTCTCTTCTATTGCCTTGCAATTCAAAGCAAAGAACATGC
TTATGATGGAGTCACATCTGAAAACATGAGGAATGGACTGGTTAATAGTG
AAGTCCATAATGAAGATGGAAGAAATGGAGATGTCTCTCAGTTCCATAT
GTGGAATTACAGGAAGAGATAGTGTACCTGCCCTACTTGTCAAGGAAC
AGGAAGAATTCTAGGGGCAAGAAAACCACACTGGTGGCATTGATTCCAT
ATAGTGATCAGAGATTAAGGCCAAGAACAAAGCTGTATGTGATGGCT
TCTGTGTTGCTGTCTACTCCTTCTGGATTGGCTGTGTTTCCTTT
CCCTCGCTCTATCGACGTAAACATGGTGTAAAATCAGCCTATGTCA
GTTATGATGTTCAGAACGGTACAATTATTTAAATATCACAAACACACTA
AATATAACAAACAATAACTATTACTCTGCGAACATTGAAACATCACTGC
CCAAGTTCAATTTCAAAAACAGTTATTGGAAAGGCACGCTTAAACAACA
TAACCATTATTGGTCACTTGATATGAAACAAATTGATTACACAGTACCT
ACCGTTATAGCAGAGGAATGAGTTATATGTATGATTTCTGTACTCTGAT
ATCCCATCAAAGTGCATAACATAGTACTCATGATGCAAGTTACTGTGACAA
CAACATACTTGGCACTCTGAACAGATATCCCAGGAGAGGTATCAGTAT
GTCGACTGTGAAAGAACACAACATTATCAGTTGGGGCAGTCTGAATATT
AAATGTAACCTCAGCCACAACAGTAA
```

- Para obtener la región ortóloga en el ratón, cambiamos la base de datos del navegador y busco la ubicación del gen TMEM106B. Una vez encontrado el gen, seleccionamos el gen y desde la página del detalle del gen vuelvo a hacer la consulta por la región codificante CDS.

```
>mm10_refGene_NM_027992 range=chr6:13071744-13084326 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATGGGAAAGTCTCTTCTACTTACCTTGCAATTCAAATAAAAGAACATGC
CTATGATGGCCTTACATCGACAGACAATATGAGAAATGGATTGGTTAGCA
GTGAAGTGCACAACGAAGACGGAAGAAATGGAGATGTCTCTCAGTTCCCA
TATGTGGAATTACTGGAAAGAGATAGTGTCACTTGTCCCACTTGCCAAGG
AACAGGAAGAACCTAGGGACAAGAAAACCAACTGGTGGCATTGATTC
CATATAGTGTACCGGGTTACGGCCAAGAAGAACAAAGCTGTATGTGATG
GGCTCTGTGTTGCTGCTGCTCTGCTGGATTGGCTGTGTTTTCT
```

```

TTTCCCTGATCTATTGAGGTAACTACATGGAGTAAAATCAGCCTATG
TCAGCTACGACGCTGAAAAGCGAACCATATATTTAAATATCACGAACACA
CTAAATATAACAAATAATAACTATTATTCTGTTGAAGTTGAAAACATCAC
TGCTCAAGTCCAGTTTCAAAAACCGTGATTGGAAAGGCTGTTAAACA
ACATAACTAACATTGCCACTTGATATGAAGCAGATTGATTACCGTA
CCACAGTTATTGCAGAGGAATGAGTTACATGTATGATTCTGTACACT
GCTCTCCATCAAAGTCACAAACATAGTACTCATGATGCAAGTTACTGTAA
CAACAGCATACTTGGACACTCTGAGCAGATATCTCAGGAAAGGTACAG
TATGTCGACTGTGGAAGGAACACGACTTACAGTTGGCCAGTCTGAGTA
TCTAAATGTCCTTCAGCCACAACAATAA

```

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/>

En primer lugar entro a la página. Utilizando la información de la pregunta anterior, realizo la consulta con las dos secuencias anteriores.

The screenshot shows the Clustal Omega results page. At the top, there's a header with the date 'Wed 30 Oct, 23:11' and a link to 'Results < Clustal...'. Below the header, there's a navigation bar with links to 'Input form', 'Web services', 'Help & Documentation', 'Bioinformatics Tools FAQ', 'Feedback', and 'Share'. The main content area is titled 'Clustal Omega' and shows a 'Results for job clustalo-l20191030-221114-0599-31932964-p2m'. Under this, there are tabs for 'Alignments' (which is selected), 'Result Summary', 'Phylogenetic Tree', 'Results Viewers', and 'Submission Details'. Below the tabs, there's a 'Download Alignment File' button. The main part of the page displays a multiple sequence alignment between two genes. The alignment shows the sequence for hg38.refGene_NM_001134232 and mm10.refGene_NM_027992. The sequences are aligned vertically, with identical bases shown in green and differences in red. The alignment is 600 characters long, with positions 60 through 660 labeled. The alignment is presented in a table format with columns for each nucleotide position and rows for each gene's sequence. At the bottom of the alignment table, there's a note about cookies and privacy, followed by a 'I agree, dismiss this banner' button.

Lo que podemos observar, es que hay un ajuste bastante bueno entre las dos secuencias de genes, siendo las diferencias generalmente puntuales (en general de una sola base)

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.

Utilizo la misma metodología que en la pregunta 5, pero al momento de generar la consulta, solicito la Predicted Protein.

```

>NP_001127704 length=274
MGKSLSHLPLHSSKEDAYDGVTSNMRNGLVNSEVHNEDGRNGDVSQFPY
VEFTGRDSVTCPCTCQGTGRIPRGQENQLVALIPYSDQRRLPRRTKLYVMA

```

```
SVFVCLLLSGLAVFFLFPRSIDVKYIGVKSAYVSYDVQKRTIYLNITNTL
NITNNNYYSEVENITAQVQFSKTVIGKARLNNITIIGPLDMKQIDYTV
TVIAEEMS MYDFCTLISIKVHNIVLMMQVTVTYYFGHSEQISQERYQY
VDCGRNTTYQLGQSEYLNVLQPQQ
```

```
>NP_082268 length=275
MGKSLSHLPLHSNKEDGYDGVSTDNRNGLVSSEVHNEDGRNGDVSQFP
YVEFTGRDSVTCPTCQGTGRIPRGQENQLVALIPYSQRLRPRRTKLYVM
ASVFVCLLLSGLAVFFLFPRSIEVKYIGVKSAYVSYDAEKRTIYLNITNT
LNITNNNYYSEVENITAQVQFSKTVIGKARLNNITNIGPLDMKQIDYTV
PTVIAEEMS MYDFCTLSSIKVHNIVLMMQVTVTAYFGHSEQISQERYQ
YVDCGRNTTYQLAQSEYLNVLQPQQ
```

Posteriormente vuelvo a la página de CLUSTAL y realizo la consulta de alineamiento usando esta vez las dos proteínas

The screenshot shows the Clustal Omega results page. At the top, there are tabs for Alignments, Result Summary, Phylogenetic Tree, Results Viewers, and Submission Details. The 'Alignments' tab is selected. Below the tabs, there are buttons for 'Download Alignment File' and 'Hide Colors'. The main content area displays a sequence alignment for 'CLUSTAL O(1,2,4) multiple sequence alignment'. The alignment includes four sequences: NP_081127794, NP_082268, NP_081127794, and NP_082268. The sequences are aligned across 128 positions. The alignment shows high conservation between the first two proteins and lower conservation between the last two. A note at the bottom says 'PLEASE NOTE: Showing colors on large alignments is slow.'

El ajuste de las proteínas es bastante bueno y de hecho parece mejor que el de ADN. Esto tiene sentido en la medida de que las hebras de ADN que las codifican tienen diferencias puntuales. Como la codifican de la transcripción permite que un aminoácido puede estar codificado por *más de una combinación* de tripletes de bases, es de esperar que la proteína se ajuste mejor que el ADN que la codifica, ya que es probable que una diferencia puntual de una base en un codón *no cambie* el aminoácido codificado.

8. Emplead la herramienta LiftOver para averiguar las coordenadas del mismo 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 se puede obtener fácilmente la ubicación del gen en las versiones hg19 de humano y con ésta, la versión mm9 del ratón.

Lift Genome Annotations

This tool converts genome coordinates and genome annotation files between assemblies. The input data can be pasted into the text box, or uploaded from a file. If a pair of assemblies cannot be selected from the pull-down menus, a direct lift between them is unavailable. However, a sequential lift may be possible. Example: lift from Mouse, May 2004, to Mouse, Feb. 2006, and then from Mouse, Feb. 2006 to Mouse, July 2007 to achieve a lift from mm5 to mm9.

Original Genome:	Original Assembly:	New Genome:	New Assembly:
Human	Dec. 2013 (GRCh38/hg38)	Human	Feb. 2009 (GRCh37/hg19)

Minimum ratio of bases that must remap:

BED 4 to BED 6 Options:

Allow multiple output regions:

Minimum hit size in query:

Minimum chain size in target:

BED 12 Options:

Min ratio of alignment blocks or exons that must map:

If thickStart/thickEnd is not mapped, use the closest mapped base:

Paste in data ([BED](#) or chrN:start-end formats):

```
chr7:12211294-12243367
```

Or upload data from a file ([BED](#) or chrN:start-end in plain text format):

No file selected.

Results

Successfully converted 1 record: [View Conversions](#)

Secuencia original human - hg38:

chr7:12211294-12243367

Secuencia para la versión human - hg19:

chr7:12250920-12282993

Secuencia para la versión mouse - mm9:

chr6:13019841-13042696

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>

BLAT es una herramienta que permite localizar secuencias de bases o proteínas en un set de genoma determinado. La herramienta permite obtener resultados a una consulta utilizando algoritmos de alineación que resultan rápidos y eficientes para las consultas de secuencias de bases y proteínas.

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

Utilizo la secuencia CDS del ejercicio 1 y hacemos la consulta para la secuencia. El resultado son varios candidatos, siendo el primero el que tiene un *score* mayor y el que efectivamente corresponde a la pista del gen TMEM106B que encontramos en el ejercicio 1.

genome.ucsc.edu/cgi-bin/hgBlat

Genomes Genome Browser Tools Mirrors Downloads My Data Projects Help About Us

Human (hg38) BLAT Results

BLAT Search Results

Go back to [chr7:11746221-12708440](#) on the Genome Browser.

Custom track name:

Custom track description:

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	DSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN	
browser	details	YourSeq	819	1	825	825	100.0%	chr7	+	12214811	12231975	17165
browser	details	YourSeq	22	29	237	825	100.0%	chr21	-	10132471	10132492	22
browser	details	YourSeq	22	12	33	825	100.0%	chr21	-	7132471	7132492	22
browser	details	YourSeq	22	286	307	825	100.0%	chr2	+	30889266	30890287	22
browser	details	YourSeq	22	286	307	825	100.0%	chr2	+	1332471	1332492	22
browser	details	YourSeq	21	183	123	825	100.0%	chr12	-	88639591	88639611	21
browser	details	YourSeq	20	137	156	825	100.0%	chr10	-	73466839	73466858	20

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, como en el anterior, utilizo la secuencia CDS del ejercicio 1. A diferencia del anterior, selecciono la base de datos de *mouse*, *mm19*. De esta manera realizo la consulta.

genome.ucsc.edu/cgi-bin/hgBlat

Genomes Genome Browser Tools Mirrors Downloads My Data Projects Help About Us

Mouse (mm10) BLAT Results

BLAT Search Results

Go back to [chr12:56694976-56714605](#) on the Genome Browser.

Custom track name:

Custom track description:

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	DSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN	
browser	details	YourSeq	605	1	825	825	99.3%	chr6	+	13071744	13084326	12583
browser	details	YourSeq	25	116	146	825	96.3%	chr4	+	51834959	51834991	33
browser	details	YourSeq	23	432	454	825	100.0%	chr6	+	39882355	39882377	23
browser	details	YourSeq	22	424	445	825	100.0%	chr18	-	81930478	81930497	22

El mejor candidato en mm10 se ubica en: chr6, 13071744-13084326

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).

genome.ucsc.edu/cgi-bin/hgBlat

Genomes Genome Browser Tools Mirrors Downloads My Data Projects Help About Us

Chicken (galGal4) BLAT Results

BLAT Search Results

Go back to [chr1:51849854-51853553](#) on the Genome Browser.

Custom track name:

Custom track description:

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	DSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN	
browser	details	YourSeq	199	442	825	825	87.0%	chr2	+	26714104	26716947	2844
browser	details	YourSeq	21	295	315	825	100.0%	chr17	+	2890161	2899181	21
browser	details	YourSeq	20	592	595	825	100.0%	chr17	-	16032471	18132492	20
browser	details	YourSeq	28	683	702	825	100.0%	chr1	+	315556	315575	20
browser	details	YourSeq	20	399	418	825	100.0%	chr11	+	9389873	9398992	20

Utilizo la misma metodología anterior, cambiando la base a la de *chicken*, *galGal4*.

El mejor candidato en pollo (galGal4) se ubica en: chr2, 26714104 - 26716947

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

BLAT Search Results

Go back to [chr1:51849854-51853553](#) on the Genome Browser.

Custom track name:

Custom track description:

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	SIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	YourSeq	415	23	274	274	92.3%	chr2	+	26708501	26716944	8444
browser details	YourSeq	95	41	94	274	79.7%	chrUn_JH37554	+	90879	91154	276

En este caso, comienzo con la secuencia de proteína del ejercicio 1 y buscamos la mejor alineación en la base *galGal4*. > el mejor candidato en pollo (galGal4) se ubica en: **chr2, 26708501 – 26716944**

En este caso la codificación desde la proteína no coincide del todo con la que utiliza la región CDS del gen. A continuación pongo las pistas sobre el navegador para poder visualizar mejor las diferencias.

UCSC Genome Browser on Chicken Nov. 2011 (ICGSC Gallus_gallus-4.0/galGal4) Assembly

move <<< << < > >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr2:26,705,060-26,717,857 12,798 bp enter position, gene symbol or search terms go

chr2

TMEM106B

from_CDS (filter activated)

From_protein (filter activated)

RefSeq Genes

move start < 2.0 > move end < 2.0 >

track search default tracks default order hide all manage custom tracks track hubs configure multi-region reverse resize refresh

Una primera vista confirma la información de la posición. La pista codificada desde la proteína comienza antes y se ajusta algo **mejor** que la pista codificada de la región CDS

Con un zoom mayor a una de las zonas codificantes permiten ver que la región codificada desde la proteína se ajusta mejor a la región del gen TMEM106B y tiene menos bases que no coinciden. La interpretación que hago de esto es que cada aminoácido de la proteína está representada por uno o **mas** codones de tripletes de bases. De esta manera un mismo aminoácido se puede codificar por dos o mas secuencias diferentes. En este caso es posible que secuencias diferentes del CDS terminen codificando la misma secuencia de aminoácido y de esta forma se explica que aunque las regiones CDS no calcen del todo, la proteína final si que termina calzando mejor (lo que hace que la transcripción inversa sea a su vez calce mejor).

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>

Para acceder a la página de Table browser se accede desde el menu tools de la página del navegador UCSC. La herramienta Table browser permite obtener la información del navegador de pistas de genes UCSC, en

un formato de texto que es un formato más compatible para otros softwares. El buscador permite realizar además búsquedas complejas a las tablas de datos de genes.

The screenshot shows the Table Browser interface with the following search parameters:

- clade: Mammal
- genome: Human
- assembly: Dec. 2013 (GRCh38/hg38)
- group: Genes and Gene Predictions
- track: NCBI RefSeq
- table: UCSC RefSeq (refGene)
- region: chr7:12,211,294-12,243,367
- identifiers (names/accessions):
- filter: create
- subtrack merge: create
- intersection: create
- correlation: create
- output format: all fields from selected table
- output file: (leave blank to keep output in browser)
- file type returned: plain text

Below the search bar, there is a note: "To reset all user cart settings (including custom tracks), click here."

Using the Table Browser

This section provides brief line-by-line descriptions of the Table Browser controls. For more information on using this program, see the [Table Browser User's Guide](#).

- **clade:** Specifies which clade the organism is in.
- **genome:** Specifies which organism data to use.
- **assembly:** Specifies which version of the organism's genome sequence to use.
- **group:** Selects the type of tracks to be displayed in the track list. The options correspond to the track groupings shown in the Genome Browser. Select 'All Tracks' for an alphabetical list of all available tracks in all groups. Select 'All Tables' to see all tables including those not associated with a track.
- **database:** (with "All Tables" group option) Determines which database should be used for options in table menu.
- **track:** Selects the annotation track data to work with. This list displays all tracks available in the genome. Icons for tracks are shown below the list, though some may not be available when the region is set to genome due to the data provider's restrictions on sharing.

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.

- Número de pares de bases del genoma completo. Para la primera consulta seleccionamos el genoma y tablas correspondientes, dejamos la región genome y hacemos la consulta de **summary/statistics**

refGene (refGene) Summary Statistics

item count	82,864
item bases	1,440,975,236 (46.54%)
item total	5,128,305,053 (165.64%)
smallest item	21
average item	61,888
biggest item	2,298,757
block count	806,362
block bases	95,921,067 (3.10%)
block total	261,133,851 (8.43%)
smallest block	1
average block	324
biggest block	91,671

Region and Timing Statistics

region	genome
bases in region	3,257,347,282
bases in gaps	161,348,343
load time	35.66
calculation time	2.19
free memory time	0.00
filter	on
intersection	off

Según esta consulta el número total de bases es de: **item bases: 1,440,975,236**

- Número de tránscritos en total a lo largo del genoma.

DE la misma consulta anterior se obtiene que el número de transcritos total es de: **item count: 82,864**

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

En este caso utilizo la misma consulta, pero en este caso cambiamos `output format` para solicitar sólo aquellos atributos solicitados.

```
#hg38.refGene.bin hg38.refGene.name hg38.refGene.chrom hg38.refGene.strand hg38.refGene.txStart
0 NM_000299 chr1 + 201283451 201332993 15 PKP1
0 NM_001276351 chr1 - 67092165 67134970 8 C1orf141
0 NM_001276352 chr1 - 67092165 67134970 9 C1orf141
0 NM_001005337 chr1 + 201283505 201332989 14 PKP1
0 NR_075077 chr1 - 67092165 67134970 10 C1orf141
```

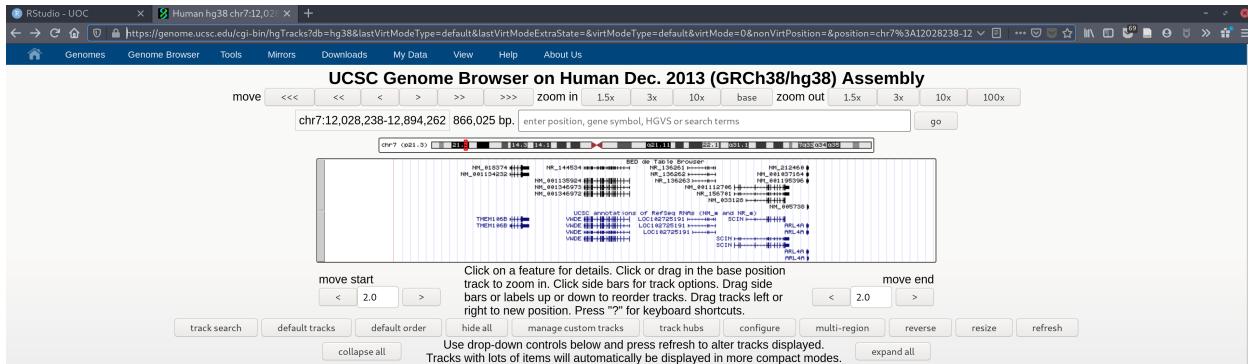
- Número total y listado de los transcritos en la región chr7:12000000-13000000.

En esta consulta restringimos la búsqueda a la región solicitada `chr7:12000000-13000000` y hacemos la consulta. La consulta entrega 16 transcritos, con un total de 243,432 bases.

```
#hg38.refGene.bin hg38.refGene.name hg38.refGene.chrom hg38.refGene.strand hg38.refGene.txStart
678 NM_001134232 chr7 + 12211293 12243367 8 TMEM106B
678 NM_018374 chr7 + 12211293 12243367 9 TMEM106B
679 NM_001346972 chr7 - 12330884 12403941 27 VWDE
679 NM_001346973 chr7 - 12330884 12403941 27 VWDE
679 NR_144534 chr7 - 12330884 12403941 30 VWDE
679 NM_001135924 chr7 - 12330884 12403865 29 VWDE
680 NR_136261 chr7 + 12496428 12541135 6 LOC102725191
680 NR_136262 chr7 + 12497245 12541135 4 LOC102725191
680 NR_136263 chr7 + 12504396 12541135 4 LOC102725191
10 NR_156701 chr7 + 12570719 12660181 15 SCIN
10 NM_001112706 chr7 + 12570719 12660181 16 SCIN
681 NM_033128 chr7 + 12589522 12653603 14 SCIN
681 NM_001037164 chr7 + 12686826 12690933 2 ARL4A
681 NM_212460 chr7 + 12686826 12690933 2 ARL4A
681 NM_001195396 chr7 + 12687285 12690933 2 ARL4A
681 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.

En primer lugar cambio la opción de `Output Format` para generar el formato BED. Agregamos un nombre (en este caso “Consulta_PEC”) y copio el texto. Posteriormente vuelvo al navegador UCSC y desde aquí importo un *track personalizado*. Al mirar el track podemos ver que efectivamente corresponde con la región del track original e incluye varios genes, incluido el TMEM106b.



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

En primer lugar cambio el *output format* a **sequence** y solicito solo las regiones CDS. Copio el texto de solo una de las secuencias.

```
>hg38_refGene_NM_001112706 range=chr7:12570787-12652715 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATGGCGCGGGAGCTATAACACAGAAGAGTCGCCCCGGCGGCAGCAGGC
GGGGCTGCAGGTCTGGAGGATTGAGAACGCTGGAGCTGGTCCCCGTCCCC
AGAGCGCTCACGGCACTTCTACGTCCCCGATGCCAACCTGGTCTGCAC
ACGGCCAAGACGAGCGAGGCTTACCTACCACCTGCACTTCTGGCTCGG
AAAGGAGTGTCCCAGGATGAAAGCACAGCTGCCATCTCACTGTTC
AGATGGATGACTATTTGGTGGCAAGCCAGTCAGAACATAGAGAACTTCAA
GGATATGAGTCTAATGACTTGTAGCTATTTCAAAGCGGTCTGAAATA
CAAGGCTGGAGGCGTGGCATCTGGATTAAATCATGTTCTTACGAACGACC
TGACAGCCAAGAGGCTCTACATGTGAAGGGCTGTAGAGTGGTAGAGGCC
ACAGAAAGTCCCTTAGCTGGACAGTTCAACAAGGGTACTGCTTCAT
CATTGACCTTGGCACCGAAATTATCAGTGGTGTGGTCTCTCGTCAACA
AATATGAACGCTGAAGGCAAACCAGGTAGCTACTGGCATTGGTACAAT
GAAAGGAAAGGAAGGTCTGAACTAATGGTGTGGAAAGAACGGAAAGTGAACC
CTCAGAACTTATAAAGGTCTAGGGAAAAGCCAGAGCTCCAGATGGAG
GTGATGATGATGACATTATAGCAGACATAAGTAACAGAAAATGGCTAAA
CTATACATGGTTCAAGATGCAAGTGGCTCATGAGAGTACTGTGGTGGC
AGAAGAAAACCCCTCTCAATGGAATGCTGTCTGAAGAACATGCTTTA
TTTGGACCAACGGGCTGCCAACAAATTTCGTATGGAAAGGTAAAGAT
GCTAATCCCCAAGAGAGGAAGGGCTGAATGAAGACAGCTGAAGAACATTCT
ACAGCAAATGAATTATTCCAAGAACATCCAAATTCAAGTTCTCCAGAAG
GAGGTGAAACACCAATCTCAACAGTTTTAAGGACTGGAGAGATAAA
GATCAGAGTGTGGCTCCGGAAAGTTATGTCACAGAGAAAAGTGGCTCA
AATAAAACAAATTCCCTTGATGCCCTAAATTACACAGTCTCCGAGA
TGGCAGCCCAGCACAATATGGTGGATGATGGTCTGGCAAAGTGGAGATT
TGGCGTGTAGAAAACAATGGTAGGATCCAAGTTGACCAAAACTCATATGG
TGAATTCTATGGTGGTACTGCTACATCATACTCTACACCTATCCCAGAG
GACAGATTATCTACCGTGGCAAGGAGCAAATGCCACACGAGATGAGCTG
ACAACATCTCGTCTCTGACTGTTAGTGGATCGGTCCCTGGAGGACA
GGCTGTGCAGATCGAGTCTCCAAGGCAAAGAGCCTGTTCACCTACTGA
GTTTGTCAAAGACAAACCGCTCATTATTTACAAGAACATCAAAG
AAAGGAGGTCAAGGCACCTGCTCCCCCTACACGCCCTTCAAGTCCGGAG
AAACCTGGCATCTACACAGAATTGTGGAGGTTGATGTTGATGCAAATT
CACTGAATTCTAACGATGTTTGTCTGAAACTGCCACAAAATAGTGGC
TACATCTGGTAGGAAAAGGTCTAGGCCAGGAGGAGGAGAAAGGAGCAGA
GTATGTAGCAAGTGTCTAAAGTGCACAAACCTTAAGGATCCAAGAAGGCG
```

```

AGGAGGCCAGAGGAGTTCTGGAATTCCCTTGGAGGGAAAAAAAGACTACCA
ACCTCACCACTACTGGAAACCCAGGCTGAAGACCACCTCGGCTTA
CGGCTGCTCTAACAAAATGGAAGATTGTTATTGAAGAGATTCCAGGAG
AGTTCACCCAGGGATATTAGCTGAAGATGATGTCTAGATGCT
TGGAACAGATAATTATTGGATTGGAAAGATGCTAATGAAGTTGAGAA
AAAAGAATCTCTGAAGTCTGCCAAAATGTACCTTGAGACAGACCCCTCTG
GAAGAGACAAGAGGACACCAATTGTCATCATAAAACAGGGCCATGAGCCA
CCACACATTACAGGCTGGTCCTGGGATTCAGCAAGTGGTAA

```

Accedo a la herramienta BLAT y desde aquí realizo la consulta para este transcríto. Con el resultado de esta consulta creo una nueva pista personalizada y vuelvo al navegador para mirar las pistas originales, la consulta de la pregunta anterior (BED de Table Browser) y la consulta del trascrito con la herramienta BLAT (Consulta_BLAT).

Se puede observar que efectivamente el transcríto corresponde a unos de los genes de la región. En este caso corresponde al gen SCIN

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>

La herramienta Biomart es otra herramienta que permite hacer consultas a bases de datos genéticas. En este caso tenemos una barra de herramientas lateral que permite modificar los parámetros de la consulta, y arriba en el menú podemos elegir la base y tablas a consultar.

Para realizar búsquedas específicas, en este caso utilizamos el filtro en el menú lateral **Filters** y filtramos la *Region* solicitada en la pregunta.

Para obtener distintos tipos de atributos se puede modificar el menu **Attributes**. Para obtener información de las SNP y anotaciones se puede mirar la sección **Variants** del menú. En este caso que miramos los elementos de *Gene Ontology*, vamos a la sección **Features > External** donde podemos seleccionar los campos a consultar.

The screenshot shows the Ensembl Biomart interface. The URL is <https://www.ensembl.org/biomart/martview/e1cd7a542125857adde97a2cb536784>. The main content area displays a table of GO terms for the gene ENSG00000184344. The table has columns: Gene stable ID, Gene stable ID version, Transcript stable ID, Transcript stable ID version, GO term accession, GO term name, GO term definition, GO term evidence code, and GO domain. The data includes rows for various GO terms like positive regulation of fat cell differentiation, cytoplasm, extracellular space, etc., with details such as GO:0045600, GO:0005737, GO:0005615, and IAA.

Gene stable ID	Gene stable ID version	Transcript stable ID	Transcript stable ID version	GO term accession	GO term name	GO term definition	GO term evidence code	GO domain
ENSG00000184344	ENSG00000184344.4	ENST00000329913	ENST00000329913.4	GO:0045600	positive regulation of fat cell differentiation	Any process that activates or increases the frequency, rate or extent of adipocyte differentiation.	ISS	biological_process
ENSG00000184344	ENSG00000184344.4	ENST00000329913	ENST00000329913.4	GO:0005737	cytoplasm	All of the contents of a cell excluding the plasma membrane and nucleus, but including other subcellular structures.	IDA	cellular_component
ENSG00000184344	ENSG00000184344.4	ENST00000329913	ENST00000329913.4	GO:0005737	cytoplasm	All of the contents of a cell excluding the plasma membrane and nucleus, but including other subcellular structures.	IEA	cellular_component
ENSG00000184344	ENSG00000184344.4	ENST00000329913	ENST00000329913.4	GO:0008083	growth factor activity	The function that stimulates a cell to grow or proliferate. Most growth factors have other actions besides the induction of cell growth, including differentiation.	IEA	molecular_function
ENSG00000184344	ENSG00000184344.4	ENST00000329913	ENST00000329913.4	GO:0005615	extracellular space	That part of a multicellular organism outside the cells proper, usually taken to be outside the plasma membranes, and occupied by fluid.	IEA	cellular_component
ENSG00000184344	ENSG00000184344.4	ENST00000329913	ENST00000329913.4	GO:0005576	extracellular region	The space external to the outermost structure of a cell. For cells without external protective or external encapsulating structures, this refers to space outside of the cell membrane, and includes the interstitial fluid, host cell environment outside an intracellular parasite.	IEA	cellular_component
ENSG00000184344	ENSG00000184344.4	ENST00000329913	ENST00000329913.4	GO:0007275	multicellular organism development	The biological process whose specific outcome is the progression of a multicellular organism over time from an initial condition (e.g. a zygote or a young adult) to a later condition (e.g. a multicellular animal or an aged adult).	IEA	biological_process
ENSG00000184344	ENSG00000184344.4	ENST00000329913	ENST00000329913.4	GO:0005125	cytokine activity	Functions to control the survival, growth, differentiation and effector function of tissues and cells.	IEA	molecular_function
ENSG00000184344	ENSG00000184344.4	ENST00000329913	ENST00000329913.4	GO:0005615	extracellular space	That part of a multicellular organism outside the cells proper, usually taken to be outside the plasma membranes, and occupied by fluid.	IBA	cellular_component
ENSG00000184344	ENSG00000184344.4	ENST00000329913	ENST00000329913.4	GO:0019901	protein kinase binding	Interacting selectively and non-covalently with a protein kinase, any enzyme that catalyzes the transfer of a phosphate group, usually from ATP, to a protein	IPI	molecular_function