

Navegadores genómicos

Enrique Blanco García

2 créditos ECTS
Enrique Blanco García
Módulo 1

Índice

Introducción	3
Objetivos	5
1. Navegadores genómicos	6
1.1. Conceptos básicos	6
1.2. Filosofía de la navegación genómica	10
1.3. El navegador genómico UCSC: estructura	14
1.4. El navegador genómico UCSC: pistas	20
1.5. El navegador genómico UCSC: anotaciones propias	31
1.6. El navegador genómico UCSC: grupos de pistas propios	39
1.7. El navegador genómico UCSC: el navegador de tablas	41
1.8. El navegador genómico UCSC: distribución local de ficheros	44
1.9. El navegador genómico ENSEMBL	47
1.10. Na	50
Resumen	51
Actividades	52
Ejercicios de autoevaluación	53
Solucionario	54
Bibliografía	56

Introducción

La consecución de las secuencias de los genomas de innumerables especies desde finales del pasado siglo XXI hasta la actualidad constituye un hito sin precedentes en la historia del progreso científico. Gracias al trabajo conjunto de miles de investigadores de todo el mundo, podemos explorar desde nuestro propio ordenador estas secuencias biológicas para descifrar las claves del código genético e incluso editar la propia secuencia del genoma mediante la tecnología CRISPR para realizar posteriores experimentos de perturbación génica global. En paralelo con el incesante progreso de las tecnologías informáticas para el tratamiento masivo de datos, la democratización en el acceso universal a esta clase de información ha transformado radicalmente la investigación en biología molecular y biomedicina durante las dos últimas décadas. En consecuencia, en el camino para explicar numerosos enigmas hasta ahora sin resolver, el enfoque científico clásico sustentado en la elaboración de hipótesis está derivando en aproximaciones más pragmáticas dirigidas por el análisis no supervisado de datos a gran escala. De este modo, dentro de un determinado contexto biológico, es factible monitorizar el funcionamiento del inventario completo de los genes del genoma simultáneamente para descubrir nuevos mecanismos de regulación entre ellos. Por otro lado, las mismas tecnologías de análisis pueden aplicarse para realizar el seguimiento del transcriptoma completo en varios contextos biológicos aparentemente antagónicos con el objetivo de identificar aquellos elementos diferenciales que conforman la firma molecular de cada distinto escenario.

El progreso exponencial experimentado por la ciencia desde finales del siglo XX se sustenta, en gran parte, en la implementación de herramientas eficientes de transmisión de información sobre la plataforma de la Red. Hoy día, para un investigador resulta extremadamente sencillo explorar desde su propio ordenador personal la totalidad de las anotaciones biológicas realizadas sobre una región concreta del genoma por otros miembros de la comunidad científica. Esto permite complementar de forma ágil y veraz la información que se publica en las revistas científicas, con la ventaja de estar incluso mejor actualizada. Gracias al progreso de las tecnologías de la información en la elaboración de interfaces *web* que simplifican y uniformizan la interrogación de las anotaciones existentes sobre cualquier genoma. Los navegadores genómicos son una poderosa herramienta para inferir computacionalmente nuevo conocimiento a partir de los datos aportados por otros medios más tradicionales, como los resultados obtenidos en un entorno experimental. Los enormes avances conseguidos en los estudios de regulación genómica más recientes, de hecho, no pueden comprenderse sin la contribución capital de esta clase de aplicaciones. Todavía más excitante resulta la facilidad con la que desde nuestro escritorio de trabajo podemos acceder y contrastar datos genómicos determinados en muestras de pacientes para un voluminoso universo de enfermedades y contextos genéticos. Asimismo, tenemos a nuestra disposición numerosos recursos genómicos para estudiar distintos actores asociados a la variabilidad genética (p.e. polimorfismos, variaciones en el número de copias, etc.) y combinar su identificación con todas estas clases de informaciones en nuestro trabajo.

Mediante una serie de protocolos razonablemente establecidos, cualquier investigador puede obtener en pocos minutos el cartografiado completo del genoma perfectamente actualizado. Con estas regulaciones generalmente aceptadas por la comunidad de investigadores, el conocimiento existente sobre el genoma de una especie (la secuencia de nucleótidos y el mapa de anotaciones sobre ésta) resulta convenientemente actualizado, gracias a la incesante actividad científica que intenta caracterizar con mayor precisión cada escenario biológico que ocurre dentro del entorno celular. Es de resaltar que la información contenida en versiones anteriores a la actual de cada genoma se congela y resta a disposición de cualquier investigación posterior para mantener la reproducibilidad de cualquier tipo de resultados a lo largo de los años. Existen herramientas de fácil acceso que permiten interoperar con anotaciones entre diferentes ensamblados del mismo genoma. Pero la navegación genómica no es unidireccional: los científicos pueden fácilmente integrar en estos entornos de trabajo las anotaciones obtenidas con sus propios experimentos de secuenciación masiva, realizar comparaciones con los datos existentes suministrados por los grandes consorcios y compartir estas sesiones de trabajo con sus colaboradores. En suma, gracias a estas herramientas podemos acceder inmediatamente a la zona de influencia genómica de un gen en particular para agilizar el análisis y la integración de numerosas fuentes de información disponibles sobre esa región, convirtiendo estas herramientas de navegación en indispensables para la investigación científica.

A lo largo de este módulo mostraremos al estudiante los mecanismos esenciales para explorar con garantías los genomas utilizando estos portales *web*, incidiendo especialmente en la correcta interpretación del increíble volumen de información que contienen. Diferentes navegadores genómicos surgieron desde la publicación de los primeros genomas. En las próximas líneas tomaremos como referencia el navegador de la Universidad de Santa Cruz de California (UCSC) en Estados Unidos por constituirse en el recurso universal en este campo desde su creación en paralelo con las primeras secuenciaciones genómicas. Además, múltiples proyectos internacionales surgidos con posterioridad a la consecución de la secuencia del genoma humano depositan sus datos en este repositorio, cuyo manejo resulta enormemente sencillo para cualquier investigador. No obstante, los principales fundamentos de manejo que enseñaremos para este servidor pueden aplicarse en cualquier otra aplicación similar. Por consiguiente, para complementar estas informaciones, mencionaremos otros navegadores genómicos de notable relevancia, como el portal ENSEMBL, diseñado en el Instituto Europeo de Bioinformática (EBI) en Hinxton (Reino Unido). La parte final de este módulo la dedicaremos a presentar varios de los principales recursos existentes (GTEx, TCGA y DEPMAP) para el acceso a información biomédica asociada a variabilidad y expresión genética tanto en individuos sanos como en pacientes de cáncer en distintos tejidos/órganos afectados y fases de la enfermedad.

Objetivos

Con el programa de contenidos establecido en este módulo, una vez finalizada la etapa de aprendizaje, el estudiante debe lograr la consecución de los siguientes objetivos relacionados con el manejo de los principales navegadores genómicos existentes:

1. Conocer el elenco de bancos de datos genómicos existentes.
2. Descubrir cómo se almacena computacionalmente un genoma.
3. Manejar las opciones básicas de los navegadores genómicos.
4. Aprender a interpretar los datos suministrados por éstos.
5. Enriquecer las anotaciones existentes con nuevas informaciones.
6. Comparar sistemáticamente anotaciones existentes en una misma región.
7. Acceder a la información genómica disponible para pacientes y variabilidad genética.

1. Navegadores genómicos

1.1. Conceptos básicos

El genoma de una célula es un repositorio de secuencias de ADN empaquetado en forma de cromosomas. Este material hereditario codifica los genes que una vez descifrados resultan útiles para la síntesis de proteínas y de transcritos no codificantes. Junto con los genes, cohabitan en el genoma otros elementos funcionales que regulan la activación de los mismos, y proporcionan, además, una cierta estructuración a la cromatina. Para modelar este complejo escenario biológico dentro de un entorno informático, la secuencia de nucleótidos de cada cromosoma de un genoma se almacena, en primer lugar, en un fichero de texto. Junto con las secuencias, es necesario cartografiar cada cromosoma aportando un segundo tipo de datos denominados anotaciones. Las anotaciones son necesarias para indicar la ubicación exacta de aquellos elementos cifrados en una secuencia concreta. El catálogo de elementos biológicos identificables dentro de un genoma está constituido primordialmente por:

- Genes con sus exones.
- Sitios de unión de factores de transcripción.
- Inicios de transcripción génica.
- Marcas de modificación de histonas.
- Regiones con niveles específicos de metilación de ADN.
- Transposones y otras regiones repetitivas.
- Polimorfismos y variantes de número de copias.

En función de los nuevos datos aportados por investigaciones más recientes, constantemente ampliamos y mejoramos nuestro conocimiento sobre cualquier función biológica o componente celular. En consecuencia, parece natural pensar que tanto la secuencia como las anotaciones de cualquier genoma necesitan renovarse. Para favorecer la reproducibilidad, por regla general suele distribuirse la última versión de cada genoma (secuencia y anotaciones) junto con el repositorio de anteriores distribuciones. Cada nueva versión de un genoma establece una codificación propia de referencia, mientras que las coordenadas de un elemento funcional pueden variar entre versiones cuando la secuencia de base resulta modificada, posiblemente debido a mejoras en la secuenciación de ese organismo. Como veremos más adelante, existen ficheros de conversión de coordenadas entre versiones (en inglés, *liftover*) para comparar correctamente distintas anotaciones. Es fundamental mencionar en todo momento cuál estamos empleando en nuestros análisis bioinformáticos (la Tabla 1 muestra las distribuciones genómicas más recientes de varios organismos modelo).

Tabla 1. Distribuciones actualizadas de múltiples genomas (a 1 de Septiembre de 2022).

<i>Homo sapiens</i>	hg38 (GRCh38)	Diciembre 2013
<i>Pan troglodytes</i>	panTro6 (Clint_PTRv2)	Enero 2018
<i>Bos taurus</i>	bosTau9 (ARS-UCD1.2)	Abril 2018
<i>Mus musculus</i>	mm10 (GRCm38)	Diciembre 2011
<i>Rattus norvegicus</i>	rn6 (RGSC 6.0)	Julio 2014
<i>Gallus gallus</i>	galGal6 (GRCg6a)	Marzo 2018
<i>Danio rerio</i>	danRer11 (GRCz11)	Mayo 2017
<i>Fugu rubripes</i>	fr3 (FUGU5)	Octubre 2011
<i>Drosophila melanogaster</i>	dm6 (BDGP6 + ISO1 MT)	Agosto 2014
<i>Anopheles gambiae</i>	anoGam3 (AgamP3)	Octubre 2006
<i>Caenorhabditis elegans</i>	ce11 (WBcel235)	Febrero 2013
<i>Saccharomyces cerevisiae</i>	sacCer3 (S288c)	Abril 2011

Cuando analiza la secuencia de una región cromosómica para explorar su contenido o anotar nuevos elementos, el bioinformático accede a estas informaciones mediante un programa especial denominado navegador genómico (en inglés, *genome browser*). Generalmente, tanto la secuencia como las anotaciones de las múltiples versiones de un genoma se distribuyen de forma pública a través de la Red. Con estos programas, todo el conocimiento aportado por distintos grupos de investigación sobre el genoma está integrado en una sola herramienta, facilitando enormemente el intercambio de información. Los navegadores genómicos suministran los datos que poseen sobre cada genoma a través de potentes interfaces gráficas que favorecen su legibilidad. Estas fotografías del genoma, no obstante, son generadas a partir de simples ficheros de texto que contienen las coordenadas de las anotaciones. Cada navegador habitualmente implementa funciones adicionales para permitir interrogar sistemáticamente estas bases de datos. Esto permite a los usuarios generar fácilmente subconjuntos de las anotaciones originales que cumplen determinados criterios e integrarlos automáticamente dentro del navegador en la vista actual. También es importante recordar que estos archivos suelen distribuirse separadamente en su formato original para realizar tratamientos bioinformáticos de forma local.

Dada su enorme versatilidad, existen varios puntos de entrada en un navegador genómico para acceder a la información de un elemento en particular. En la Tabla 2 mostramos algunos ejemplos a la hora de localizar la misma región genómica que contiene un gen de interés para nosotros. Según el método escogido y el nivel de detalle de nuestra búsqueda, es probable que el navegador genómico identifique más de un posible resultado. En ese caso, la exploración visual de las alternativas debería ser suficiente para elegir correctamente la región deseada. Una vez realizada satisfactoriamente la búsqueda, el navegador genómico nos proporcionará una fotografía de las anotaciones disponibles en forma de pistas o carriles (en inglés, *tracks*). Con este mecanismo de visualización, el bioinformático puede comparar fácilmente diferentes elementos anotados sobre una misma región genómica.

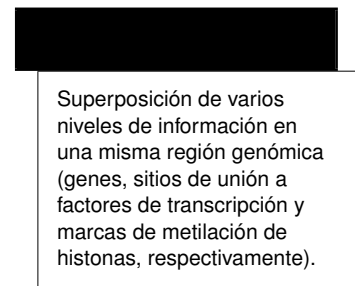
Dependiendo de los datos que conozcamos sobre el objeto de la búsqueda, debemos escoger el modo más adecuado. Generalmente, es posible acceder desde varios lugares distintos a la misma información.

Tabla 2. Puntos de acceso alternativos a un mismo gen (a 1 de Septiembre de 2022).

Región	Coordenadas	chr2:80301878-80304752 (hg38)
Gen	Abreviatura	<i>LRRTM1</i>
Gen	Nombre completo	<i>leucine-rich repeat transmembrane neuronal protein 1 precursor</i>
Gen	Código RefSeq	NM_178839
Proteína	Identificador	LRRTM1
Tránsito	Identificador	BC045113
Secuencia	FASTA	ATGGATTTCCTGCTGCTCGGTC...

El navegador genómico proyecta sobre cada pista de su visor genómico la ubicación de los elementos biológicos conocidos en ambas orientaciones de la molécula de ADN simultáneamente. Las anotaciones en hebras distintas pueden presentarse gráficamente integradas dentro de la misma pista o estar separadas en cuadrantes distintos de la imagen, según el navegador. La propia secuencia del genoma es una pista cuyo contenido corresponde a la base observada en cada nucleótido de la región visualizada (únicamente cuando el tamaño de ésta sea inferior a la resolución de la imagen, el navegador muestra explícitamente la secuencia de bases). Para gestionar de forma óptima la proyección simultánea de varias pistas, el navegador superpone la información en diferentes niveles (ver Figura 1). El usuario puede elegir el elenco de pistas que desea visualizar en cada momento, así como en qué orden desea disponerlas en la ventana gráfica.

Figura 1. Representación gráfica de anotaciones en forma de pistas.



En tiempo real, el navegador compone una fotografía del genoma en función de nuestras necesidades a partir de los ficheros de anotaciones almacenados en su propia base de datos. El usuario puede configurar esta superposición gráfica de pistas, escogiendo entre las diversas pistas disponibles precisamente aquellas que más le interesan. Por regla general, las anotaciones están agrupadas en bloques conceptuales biológicamente relacionados. Una vez establecido el contenido del nuevo mapa de anotaciones, el navegador realiza una actualización de la imagen. También es posible establecer el nivel de detalle gráfico de cada pista para conseguir representaciones gráficas que faciliten una comparación óptima. (desplegar o compactar pistas, ver Figura 2). Existen aplicaciones auxiliares en los propios navegadores genómicos para realizar estas comparaciones de forma cuantitativa, teniendo en cuenta la correlación entre la ubicación de los elementos de diferentes anotaciones.

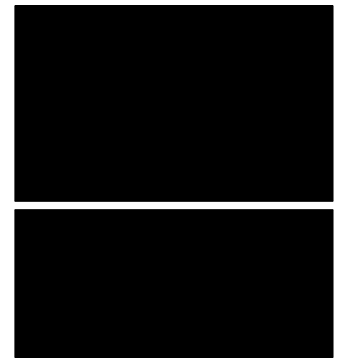
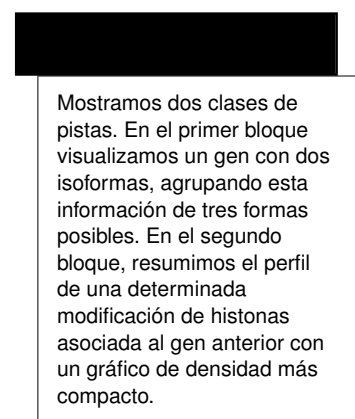


Figura 2. Mostrando diversos niveles de información con pistas.

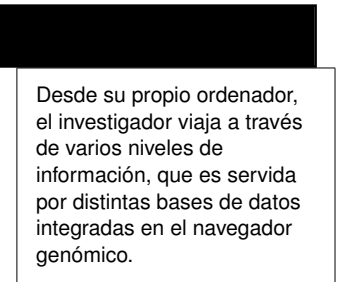


1.2. Filosofía de la navegación genómica

En la actual era post-genómica, los investigadores han desplazado el foco de atención desde las secuencias individuales hacia el análisis a gran escala del genoma. Con el objetivo de tener una visión más amplia de las anotaciones agrupadas entorno a una región del genoma, aparecen los navegadores genómicos, que integran distintas fuentes de información dispersas por la Red. Estos sistemas redefinen, por tanto, la comunicación con estos complejos repositorios de vastos volúmenes de información. Actualmente navegamos por heterogéneos paisajes genómicos codificados en el interior de los cromosomas, concentrando nuestra interés únicamente sobre determinadas áreas.

Desde el genoma completo hasta la secuencia individual de una proteína, el usuario de estos servicios puede recuperar todas las anotaciones con diversos niveles de detalle. Lógicamente, esta amalgama de recursos también es accesible de forma aislada, menguando sin embargo la efectividad del análisis bioinformático fuera de estos entornos integrados. Como muestra la Figura 3, desde la representación gráfica servida por el navegador genómico podemos visitar las anotaciones particulares suministradas por otro tipo de recursos más específicos.

Figura 3. Paradigma de la comunicación con navegadores genómicos.



Desde su propio ordenador, el investigador viaja a través de varios niveles de información, que es servida por distintas bases de datos integradas en el navegador genómico.

Por ejemplo, cuando analizamos los genes anotados en una región genómica en concreto, podemos acceder desde la pista individual a una nueva pantalla que contiene información recopilada por múltiples fuentes de información. Muchos de estos repositorios, que denominaremos **bases de datos primarias**, constituyeron en su momento el germen de los actuales navegadores genómicos. Cuando no era posible llevar a cabo análisis globales de un genoma completo, la unidad de búsqueda de información era precisamente la secuencia individual (cuya anotación era suministrada por diferentes miembros de la propia comunidad científica). Estas primeras colecciones de secuencias, a pesar de su enorme valor, contenían frecuentes errores que podían conducir a importantes excesos de redundancia. Estos catálogos reciben actualmente un tratamiento de validación mucho más cuidadoso, basado en la verificación manual llevada a cabo por expertos entrenados en este tipo de tareas. A continuación ofrecemos una selección de los recursos primarios más populares. A lo largo de estos materiales analizaremos detalladamente el contenido de estos repositorios de información genómica, cuyo acceso está de hecho integrado dentro de cualquiera de los navegadores genómicos existentes.

El analista bioinformático dispone de un amplio abanico de navegadores genómicos (ver Tabla 3). Los **navegadores genéricos** integran la secuencia y las anotaciones de múltiples genomas, presentando todos los datos, por tanto, dentro de un marco común uniforme que facilita su manipulación. El navegador genómico de UCSC o la plataforma ENSEMBL son las aplicaciones genéricas más populares. Esta familia de herramientas permite cruzar sus pistas de anotaciones, facilitando de este modo la integración y visibilidad de diferentes fuentes de información desde cualquier lugar. El código de la mayoría de estas plataformas se distribuye gratuitamente junto con todos sus bancos de datos. Profundizando en esta filosofía, la herramienta Gbrowse ofrece todas las funciones de un navegador genómico genérico adaptable a cualquier genoma, cediendo al usuario la responsabilidad de introducir las anotaciones que deben respetar ciertas pautas sobre el formato de los ficheros. El navegador genómico IGV constituye otro ejemplo de esta clase de herramientas, ejecutándose como novedad en forma de aplicación independiente de cualquier navegador *web* en nuestro ordenador.

En determinadas especies se han implementado **navegadores específicos** suministrados por dichos consorcios de secuenciación para acceder exclusivamente a sus anotaciones. Estos recursos están optimizados exclusivamente para trabajar con ese genoma, convirtiéndose en un repositorio de referencia para la comunidad de estudio de ese organismo. Los ejemplos más conocidos de estos programas están dedicados a distintos organismos modelos como FlyBase para la mosca de la fruta, los navegadores de ratón y rata o los portales para la secuenciación de numerosas especies vegetales. Los navegadores genéricos, no obstante, comparten con éstos las mismas distribuciones de los genomas (ver Figura 4).

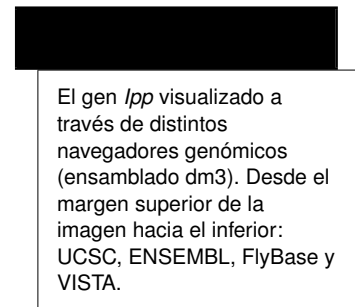
La comparación entre las anotaciones de varios genomas en regiones ortólogas puede proporcionar una información muy valiosa sobre la ubicación de ciertos elementos funcionales conservados evolutivamente. Por ejemplo, la identificación de algunas secuencias preservadas a lo largo de millones de años permite refinar la anotación de genes y elementos asociados a su regulación.

Esta prestigiosa revista científica publica semestralmente ediciones especiales sobre nuevos servidores y bases de datos en el ámbito genómico y proteómico.

Tabla 3. Listado de recursos genómicos esenciales.

GenBank	http://www.ncbi.nlm.nih.gov/genbank
Gene	https://www.ncbi.nlm.nih.gov/gene
RefSeq	http://www.ncbi.nlm.nih.gov/refseq
Homologene	http://www.ncbi.nlm.nih.gov/homologene
Gene Ontology	http://www.geneontology.org
Kyoto Encyclopedia of Genes and Genomes	http://www.genome.jp/kegg
WikiPathways	https://www.wikipathways.org
Eukaryotic Promoter Database	https://epd.epfl.ch
Jaspar	http://jaspar.genereg.net
dbVar	https://www.ncbi.nlm.nih.gov/dbvar
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP
Gene Expression Omnibus	http://www.ncbi.nlm.nih.gov/geo
OMIM	http://www.ncbi.nlm.nih.gov/omim
UCSC	http://genome.ucsc.edu
ENSEMBL	http://www.ensembl.org
NCBI	http://www.ncbi.nlm.nih.gov/sites/genome
WashU Epigenome Browser	https://epigenomegateway.wustl.edu
GBrowse	http://gmod.org/wiki/Gbrowse
JBrowse	https://jbrowse.org
IGV	https://software.broadinstitute.org/software/igv
Mouse Genome Informatics Database	http://www.informatics.jax.org
Rat Genome Database	http://rgd.mcw.edu
FlyBase	http://flybase.org
WormBase	http://www.wormbase.org
Saccharomyces Genome Database	http://www.yeastgenome.org
UCSC microbial Genome Browser	http://microbes.ucsc.edu/
GTEx	https://gtexportal.org
TCGA	https://portal.gdc.cancer.gov
DEPMap	https://depmap.org
ENCODE	https://www.encodeproject.org
modENCODE	http://www.modencode.org
Roadmap Epigenomics Project	http://www.roadmapepigenomics.org
VISTA	http://pipeline.lbl.gov
enhancerVISTA	http://enhancer.lbl.gov

Figura 4. Visualización de una región genómica.



VISTA es uno de los portales pioneros en realizar este tipo de comparaciones, facilitando enormemente la búsqueda a los investigadores (ver Figura 4). En otro orden de cosas, el análisis en profundidad del genoma humano resulta potencialmente interesante dentro del campo de la biomedicina y la salud, debido a que la mayoría de enfermedades documentadas poseen un importante componente genético. No es casual, por tanto, que sean necesarios también **navegadores genómicos avanzados** que proporcionen acceso a las anotaciones referentes a estos desórdenes hereditarios. Ejemplos de estos portales biomédicos son los proyectos de secuenciación del cáncer, expresión diferencial en tejidos o la detección del conjunto de polimorfismos existentes en el ser humano (Tabla 3).



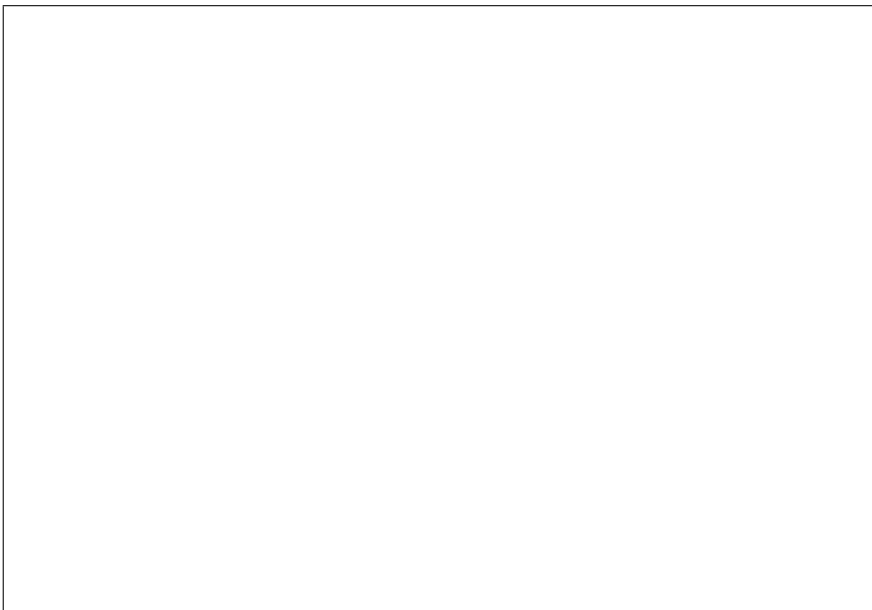
Dada su enorme popularidad, fraguada básicamente por su simplicidad de uso, en estos materiales analizaremos el navegador genómico de UCSC como ejemplo de navegador de propósito general. A medida que profundicemos en su estudio, ello nos servirá también para investigar el contenido de muchas de las bases de anotaciones primarias. El portal genómico de UCSC distribuye, además de un vasto conjunto de genomas, las anotaciones generadas en el marco de distintos proyectos internacionales de secuenciación, aumentando por tanto su relevancia como objeto de estudio.

1.3. El navegador genómico UCSC: estructura

El navegador de la Universidad de Santa Cruz de California (UCSC, de ahora en adelante) fue diseñado aproximadamente hace 20 años con el objeto de construir una herramienta que proporcionara al usuario un acceso sencillo a las anotaciones existentes sobre el genoma humano, que estaba siendo secuenciado en aquel momento. Desde su nacimiento, este portal ha incorporado numerosas mejoras a su implementación inicial, convirtiéndose en un recurso de referencia utilizado para distribuir los resultados de innumerables proyectos internacionales de secuenciación. Todas las anotaciones sobre un genoma base de referencia se suministran en forma de pistas de datos, gráficamente representadas sobre una línea horizontal en paralelo con la propia secuencia de cada cromosoma.



Figura 5. Pantalla inicial de UCSC.



Para iniciar la exploración debemos seleccionar primero el organismo y la versión del genoma que deseamos explorar. El navegador UCSC ofrece un amplio catálogo de especies para explorar, accesible desde la parte izquierda de su página principal. Ilustraremos el funcionamiento del navegador trabajando en la versión hg38 del ensamblado del genoma humano. A continuación, es necesario introducir la información suficiente para que el navegador UCSC pueda identificar el objeto de nuestra búsqueda (e.g. genes, regiones o cromosomas). Vamos a centrarnos en un gen humano denominado *LRRTM1* (abreviatura de *leucine-rich repeat transmembrane neuronal protein 1 precursor*).

Figura 6. Parámetros de búsqueda en UCSC.



En función del tipo de identificador especificado, los resultados de la búsqueda serán más o menos concretos, siendo preciso incluir en algunas ocasiones información adicional para caracterizar mejor aquello que deseamos localizar. En la Tabla 4 mostramos distintos ejemplos con diferentes tipos de valores aceptados por el motor de búsqueda del navegador UCSC para acabar encontrando el mismo gen en todos los casos.

Tabla 4. Objetos de búsqueda de UCSC (hg38, a 1 de Septiembre de 2022).

Cromosoma	chr2
Región	chr2:80301878-80304752
Gen	<i>LRRTM1</i>
Tránsito	NM_178839
Proteína	NP_849161

Con los parámetros de búsqueda acotados, el sistema proporciona el listado de todas las pistas que posean algún nexo en común con la información suministrada. En este caso, encontramos múltiples pistas relacionadas con el gen *LRRTM1* en la distribución hg38 del genoma humano. Dado que existen varios catálogos de genes generados con diferentes herramientas y distintos grados de fiabilidad, no debe resultar extraño obtener un número elevado de resultados positivos en nuestras búsquedas. RefSeq y GENCODE son actualmente los repositorios de referencia mas extendidos. En ambos casos el sistema nos informa de la anotación principal propuesta para este gen, suministrando tambien el listado completo de pistas con menor o mayor grado de fiabilidad utilizado para generar dicha anotación tipo (p.e. las pistas NCBI RefSeq curated y predicted). Procederemos a enumerar los resultados más relevantes obtenidos (ver Figura 7):

1. Pista `RefSeq Genes`: la anotación de referencia propuesta por el consorcio RefSeq.
2. Pista `NCBI RefSeq genes, curated`: anotación del mismo transcrito validada por expertos.
3. Pista `NCBI RefSeq genes, predicted`: anotación del mismo transcrito generada computacionalmente.
4. Pista `Non-human RefSeq Genes`: varias formas ortólogas del gen identificadas por comparación de secuencia en las anotaciones servidas por RefSeq para otras especies.
5. Pista `Gencode Genes`: las anotaciones de referencia propuestas por el consorcio ENCODE/GENCODE.
6. Pista `Basic Annotation from GENCODE`: anotación mínima del gen propuesta por el consorcio ENCODE/GENCODE.
7. Pista `Comprehensive Annotation from GENCODE`: anotación extendida del gen propuesta por el consorcio ENCODE/GENCODE.
8. Pista `International Knockout Mouse Consortium`: forma ortóloga del gen humano en ratón según expertos.
9. Pista `Non-Human mRNA Aligned Results`: ARN mensajeros homólogos identificados por comparación de secuencias.

El proyecto RefSeq produce un conjunto depurado de los transcritos pertenecientes a todos los genes conocidos, evitando la redundancia y las ambigüedades existentes dentro del resto de bases de datos.

El proyecto ENCODE es tratado en profundidad en el módulo *Anotación de genomas*.

~~LRRTM1 at chr2:80301878-80304752 - (NM_178839) leucine-rich repeat~~

~~transmembrane neuronal protein 1 precursor~~

Figura 7. Selección de pistas para el gen *LRRTM1* (a 1 de Septiembre de 2022).

NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*)

NM_178839.5 at chr2:80301878-80304752

NCBI RefSeq genes, predicted subset (XM_* or XR_*)

XM_017003986.3 at chr2:80288356-80304752

...

Non-Human RefSeq Genes

LRRTM1 at chr2:80301917-80304282 - (NM_001257467) leucine-rich repeat

transmembrane neuronal protein 1 precursor

...

Gencode Genes

LRRTM1 (ENST00000295057.4) at chr2:80301878-80304752

- Homo sapiens leucine rich repeat transmembrane neuronal 1 (LRRTM1), mRNA. (from RefSeq NM_178839)

...

Basic Gene Annotation Set from GENCODE Version 41
(Ensembl 107)

LRRTM1 at chr2:80301878-80304752

LRRTM1 at chr2:80301880-80304274

Comprehensive Gene Annotation Set from GENCODE Version 41
(Ensembl 107)

LRRTM1 at chr2:80288351-80304393

...

International Knockout Mouse Consortium Genes Mapped to
Human Genome

Lrrtml_68078 at chr2:80301878-80304362

Non-Human Aligned mRNA Search Results

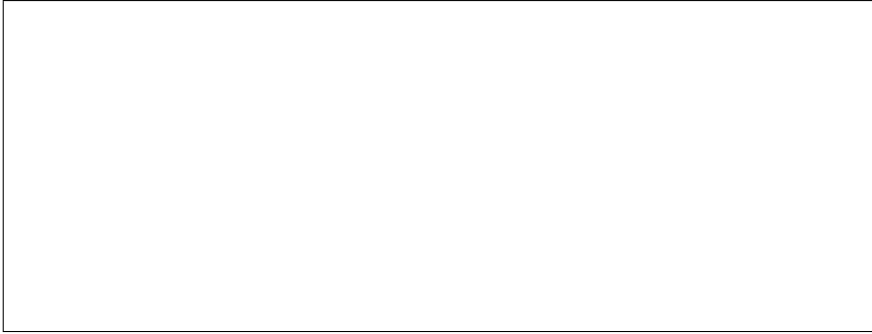
Para [BC027893](#) la representación gráfica del genoma humano que contiene nuestro gen,

vamos a seleccionar la pista de [Mus musculus leucine-rich repeat transmembrane neuronal 1, mRNA](#)

7), (cdna clone MGC:38174 IMAGE:5321979) completa cds [BC126503](#)

toda la comunidad científica debido a su precisión.

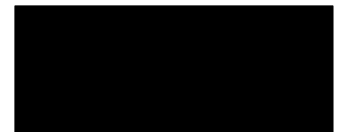
Figura 8. Pantalla inicial de navegación de UCSC (a 1 de Septiembre de 2022).



Una vez UCSC reconoce nuestra petición, abre la pantalla de navegación principal (Figura 8), para automáticamente enviarnos a la región (80,301,878-80,304,752) del cromosoma 2 (hg38) que contiene nuestro gen. El visor genómico está centrado por defecto en la zona que contiene exactamente todos los exones de nuestro gen. Junto con la pista RefSeq seleccionada, es habitual que el navegador UCSC active en nuestra primera conexión un conjunto predefinido de pistas considerado de interés general por su contenido. Para mayor claridad en las explicaciones a continuación, recomendamos esconder todas las pistas, a excepción de la pista RefSeq que hemos elegido anteriormente. Observamos los dos exones del gen *LRRTM1* en forma de cajas rectangulares. Para indicar el sentido de traducción de cada gen, el navegador introduce múltiples flechas a lo largo de sus intrones. En este caso, la traducción de este gen debe realizarse de derecha a izquierda (hebra negativa de la molécula de ADN). En ambos extremos del gen, los exones presentan un trazo más fino para distinguir la fracción codificante (CDS, *coding sequence* en inglés) de la fracción no traducible del transcrito (UTR, *UnTRanslated*). Curiosamente, nuestro gen *LRRTM1* está ubicado en el interior del intrón de otro gen de mayor tamaño denominado *CTNNA2*, codificado en la hebra contraria.

Cada pista puede ocultarse individualmente posicionándonos encima de ella, presionando el botón derecho del ratón y utilizando la opción *hide*. Alternativamente, el botón *hide all* situado debajo del visor principal oculta el conjunto completo de pistas activo.

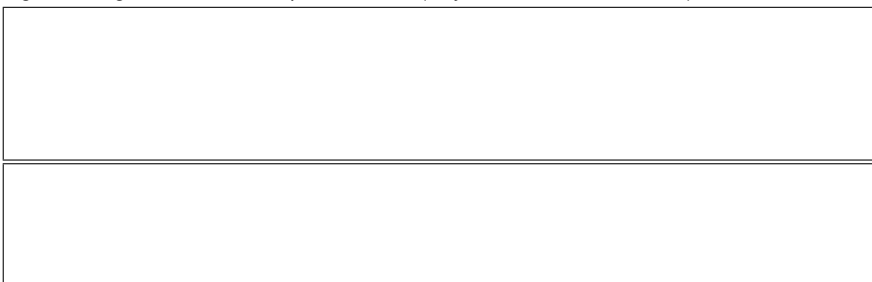
Si ocasionalmente deseamos reiniciar el navegador con las pistas por defecto utilizaremos la opción *Reset All User Settings* en la pestaña *Genome browser* del menú principal en la parte superior.



UCSC permite realizar (a) ampliaciones de posición con diferente resolución para acercar o alejar la escena, (b) desplazamientos horizontales de diferente tamaño a izquierda o derecha de la localización actual, (c) saltos hacia una región concreta mediante la caja de coordenadas y (d) búsquedas rápidas dirigidas por un identificador catalogado en este navegador. Para modificar la vista panorámica inicial podemos alterar la perspectiva usando tres modos de resolución (1.5x, 3x y 10x). Mediante un alejamiento o *zoom out* podemos obtener una visión global del paisaje genómico que rodea a nuestro gen de interés. En otras circunstancias puede ser relevante acceder a fragmentos locales de las anotaciones empleando un acercamiento o *zoom in*.

El botón *base* intensifica al máximo el acercamiento para visualizar directamente la secuencia de nucleótidos.

Figura 9. Jugando con el enfoque de UCSC (alejamiento o acercamiento).



Los botones de desplazamiento (</<<<<< y >/>>>>>) generan una nueva fotografía de las regiones del genoma inmediatamente anteriores o posteriores a esta ventana. Cada movimiento automáticamente refresca la imagen, actualizando el contenido de las pistas que estemos explorando. Si combinamos la herramienta de ampliación para fijar el ancho de la ventana gráfica (número de bases) con el desplazamiento de ésta (solapamiento entre la ventana actual y la siguiente) disponemos de un mecanismo muy efectivo para visitar cada sección del cromosoma.

Figura 10. Desplazamiento por el genoma con UCSC (izquierda o derecha).



El botón *reverse*, ubicado en la parte inferior (ver Figura 8), permite obtener la anotación gráfica de la región actual seleccionando como referencia de coordenadas la hebra alternativa del cromosoma. Como podemos ver en la Figura 11, con esta opción podemos visualizar nuestro gen de interés *LRRTMI* mostrando los exones de acuerdo con su propia orientación (obteniendo la ordenación natural de izquierda a derecha con el margen izquierdo de la imagen indicando el punto de inicio de la transcripción del gen).

El botón *configure* permite el acceso a una pantalla donde podemos modificar los parámetros gráficos del visor genómico.

Figura 11. Cambiando la hebra de ADN con UCSC.



El menú azul superior de la pantalla principal (Figura 8) contiene enlaces hacia varias aplicaciones complementarias. El botón denominado *PDF/PS* permite realizar una fotografía de la región actualmente visitada, incluyendo todas las pistas activadas en ese instante. Esta opción es verdaderamente útil para generar imágenes de alta calidad, que podemos modificar para incluir en nuestras publicaciones científicas (Figura 12).

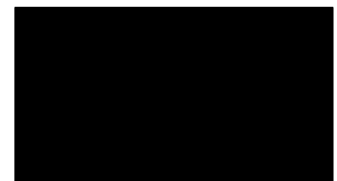


Figura 12. Fotografiando una region del genoma.



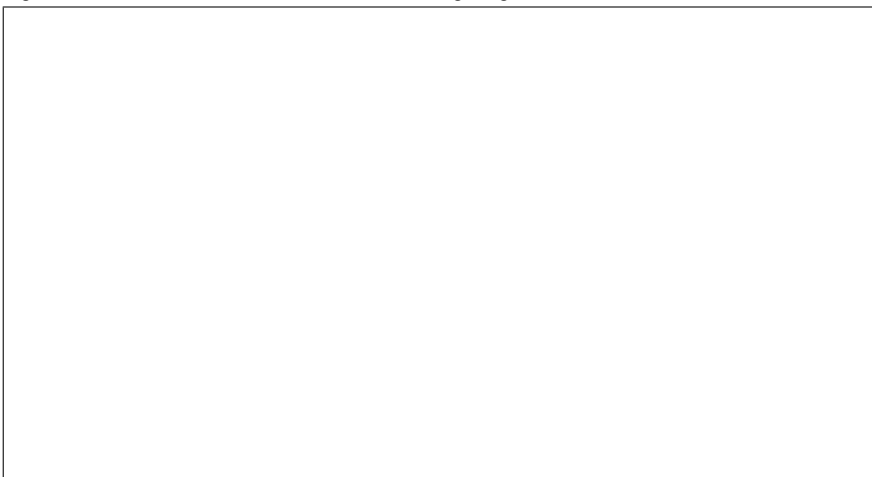
El botón *configure* situado debajo del visor principal del navegador permite al usuario modificar múltiples parámetros gráficos. Por ejemplo, podemos jugar con la resolución de la imagen en pixels, alterar el espacio para la zona del visor dedicada a mostrar los nombres de las pistas, escoger la fuente para los textos, borrar la cuadrícula azul que sirve de referencia para las pistas, ocultar la descripción de las pistas, o redefinir el comportamiento de la selección de regiones con el ratón. Esta función de configuración posibilita también el acceso a una pantalla donde el usuario puede activar u ocultar el muestrario completo de pistas del navegador.

Figura 13. Opciones de configuración general del navegador.



El marco de trabajo establecido por el navegador genómico delimita una serie de acciones que nos permiten modificar dinámicamente la representación gráfica de los elementos anotados en una región genómica. Sin embargo, no debemos olvidar que todas estas anotaciones en forma de pistas contienen implícitamente información sobre la localización exacta de dichos elementos. Para obtener directamente la secuencia de nucleótidos fotografiados en la actual ventana, podemos utilizar el botón denominado *DNA*. De este modo accederemos a la secuencia real sobre la cual estamos superponiendo las anotaciones, junto con las regiones flanqueantes a ambos lados de ésta (ver Figura 14).

Figura 14. Extracción de la secuencia de una región genómica.



1.4. El navegador genómico UCSC: pistas

Para mostrar únicamente la información relevante en cada caso, el usuario puede configurar dinámicamente el inventario de pistas que desea estudiar, seleccionando el modo más conveniente de visualización para trabajar con cada una durante el análisis. La mayoría de los navegadores genómicos permiten modificar la configuración actual de las pistas mediante un conjunto de opciones de visibilidad agrupadas en distintos bloques conceptuales (e.g. genes, regulación, conservación, etc.). En UCSC podemos modificar simultáneamente el comportamiento de varias pistas, siendo necesario presionar posteriormente el botón *refresh* (en inglés, refrescar) para reflejar el efecto final de la nueva configuración sobre la región mostrada en el visor. Para cada distribución de un genoma existe un conjunto determinado de pistas con las anotaciones registradas sobre su secuencia. No obstante, los bloques principales de opciones suelen conservarse dentro de los servicios ofrecidos por el mismo navegador. El listado de bloques proporcionados por UCSC para la versión hg38 del genoma humano puede encontrarse en la Tabla 5.

El usuario puede modificar el orden relativo entre las pistas mostradas en el visor genómico de UCSC arrastrándolas verticalmente hacia su nueva posición. También es posible realizar selecciones horizontales para acercar el foco de interés.

Tabla 5. Bloques de pistas de UCSC (hg38, a 1 de Septiembre de 2022).

<i>Mapping and Sequencing</i>	Ensamblado y secuenciación
<i>Genes and Gene Predictions</i>	Genes y predicción génica
<i>Phenotype and Literature</i>	Fenotipos y enfermedades
<i>COVID-19</i>	Información sobre COVID-19
<i>Single Cell RNA-seq</i>	Transcriptómica celular
<i>mRNA and EST</i>	Tránscrios experimentales
<i>Expression</i>	Expresión génica
<i>Regulation</i>	Regulación génica
<i>Comparative Genomics</i>	Genómica comparada
<i>Variation</i>	Polimorfismos y variación genómica
<i>Repeats</i>	Regiones repetitivas

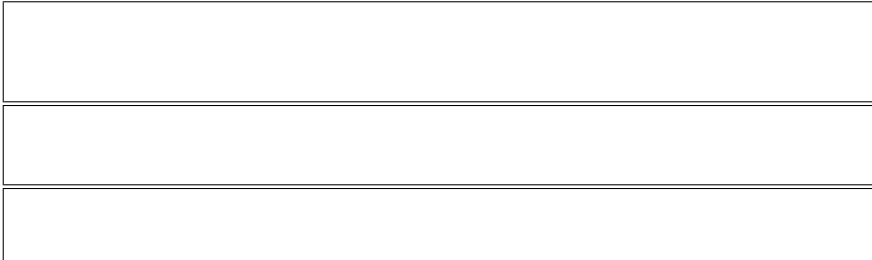
En función del tipo de elemento biológico, el espacio físico en la pantalla o el tipo de análisis comparativo, en ocasiones debemos modificar la manera en que debe plasmarse gráficamente en nuestro navegador genómico la información sobre las anotaciones. Disponemos de cinco modos de visualización para mostrar una determinada pista sobre la región genómica actual (enumerados de menor a mayor cantidad de información):

- *hide* (oculto): no muestra las anotaciones (suele ser la opción por defecto).
- *dense* (compacto): muestra las anotaciones comprimidas en una sola pista.
- *squish* (reducido): muestra las anotaciones en varias pistas reducidas.
- *pack* (agrupado): muestra las anotaciones en varias pistas agrupadas.
- *full* (completo): muestra toda la información disponible para esa pista.

Cada una de estas opciones permite activar gráficamente un conjunto específico de informaciones sobre la pista de trabajo. La visualización de las distintas formas alternativas de un gen es un ejemplo típico que denota la importancia de realizar una selección apropiada de estos datos. En la Figura 15, el estudiante puede efectivamente verificar cómo podemos plegar y desplegar a nuestro gusto la representación gráfica de las tres isoformas anotadas para el gen humano *AKR1D1*, según el modo de visualización seleccionado para la pista RefSeq Genes en el ensamblado hg38.



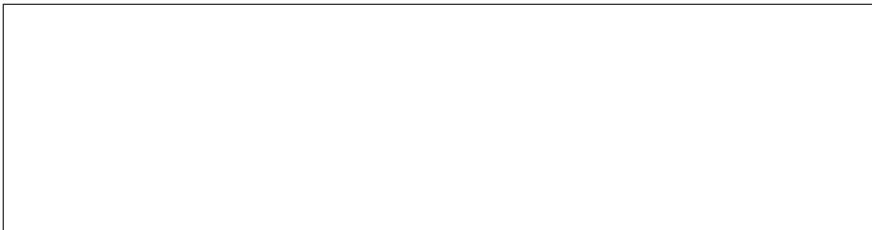
Figura 15. Configurando la visualización de una pista (hg38, a 1 de Septiembre de 2022).



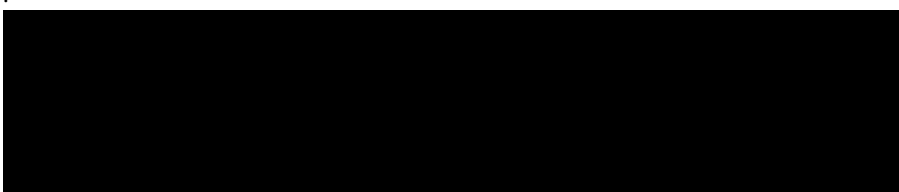
Primero desplegamos la información básica utilizando *pack*, después compactamos las tres isoformas cambiando a *squish* y finalmente plegamos las tres anotaciones en una sola pista con *dense*.

Para empezar, vamos a proceder a estudiar el bloque de opciones *Genes and Gene Predictions* (en inglés, genes y predicciones génicas). Este bloque, mostrado en la Figura 18 para el ensamblado humano hg38, nos permite acceder a todos los genes que diferentes repositorios han localizado sobre la región que estamos visualizando en un momento dado. En particular, podemos encontrar la pista de anotaciones NCBI RefSeq que estamos empleando a modo de ejemplo en este módulo.

Figura 18. Bloque de pistas de anotación de genes de UCSC (a 1 de Septiembre de 2022).



En la búsqueda anterior de la Figura 7 hemos obtenido varios resultados asociados al consorcio RefSeq de anotación génica. Dado que este proyecto internacional genera sus datos mediante diferentes herramientas, sus anotaciones se distribuyen en varias pistas agrupadas. Para gestionar estas situaciones los diseñadores de UCSC han creado un formato de pista denominado *supertrack* o *composite* (en inglés, superpista o pista compuesta). Cada subpista puede configurarse de forma independiente aunque también existen opciones de visualización aplicables al conjunto completo.



The NCBI RefSeq Genes composite track shows human protein-coding and non-protein-coding genes taken from the NCBI RNA reference sequences collection (RefSeq). All subtracks use coordinates provided by RefSeq, except for the UCSC RefSeq track, which UCSC produces by realigning the RefSeq RNAs to the genome.

Figura 17. Información asociada a la pista NCBI RefSeq (a 1 de Septiembre de 2022).

The color shading indicates the level of review the RefSeq record has undergone: predicted (light), provisional (medium), or reviewed (dark), as defined by RefSeq.

The UCSC RefSeq Genes track is constructed using the same methods as previous RefSeq Genes tracks. RefSeq RNAs were aligned against the human genome using BLAT. Those with an alignment of less than 15% were discarded. When a single RNA aligned in multiple places, the alignment having the highest base identity was identified. Only alignments having a base identity level within 0.1% of the best and at least 96% base identity with the genomic sequence were kept.

This track was produced at UCSC from data generated by scientists worldwide and curated by the NCBI RefSeq project.

References

Kent WJ. BLAT - the BLAST-like alignment tool. *Genome Res.* 2002 Apr;12(4):656-64.

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* 2014 Jan;42(Database issue):D756-63.

En el caso de las superpistas, además de dicho texto descriptivo tendremos acceso a su propia pantalla de configuración desde el mismo enlace (alternativamente, mediante el ratón

actuando sobre la imagen, podemos acceder a esta misma pantalla). Por ejemplo, cuando seleccionamos NCBI RefSeq en el bloque de opciones de información genética, esta nueva pantalla, como lista de superpistas individuales, nos muestra:

La subpista UCSC RefSeq contiene el conjunto de anotaciones que hemos optado por emplear como referencia a lo largo de este texto (se corresponde con la pista RefSeq Genes en anteriores figuras).

Figura 18. Configuración de la superpista NCBI RefSeq (a 1 de Septiembre de 2022).



Volvamos a la pantalla principal del navegador genómico con el visor gráfico activo. Para analizar los datos concretos asociados a las anotaciones gráficas mostradas en el visor, el usuario debe seleccionar con el ratón sobre el elemento de la pista que desea explorar. Por ejemplo, podemos acceder a una fuente adicional de información asociada a cada anotación RefSeq cuando seleccionamos encima de alguno de los exones del gen *LRRTM1* en la pista RefSeq Genes (Figura 8). Podemos ver un resumen del contenido de esta entrada en la Figura 19. Descendiendo desde la parte superior de la ficha, en primer lugar nos encontramos con el identificador asignado por RefSeq a este gen (NM_178839). A continuación, el navegador nos proporciona los enlaces que ha recopilado para este gen en distintas bases de datos primarias (e.g. Entrez, Pubmed, OMIM, GeneCards, etc.).

RefSeq: NM_178839.5 Status: Validated

Description of RefSeq sapiens LRRTM1 (leucine-rich repeat transmembrane neuronal 1 (LRRTM1), mRNA).

transmembrane neuronal 1 (LRRTM1), mRNA.

CCDS: [CCDS1966.1](#)

CDS: full length

OMIM: [610867](#)

Entrez Gene: [347730](#)

PubMed on Gene: [LRRTM1](#)

PubMed on Product: [leucine-rich repeat transmembrane neuronal protein 1 precursor](#)

GeneCards: [LRRTM1](#)

AceView: [LRRTM1](#)

Summary of LRRTM1

Position: [chr2:80301878-80304752](#)

Band: 2p12

Genomic Size: 2875

Strand: -

Gene Symbol: LRRTM1

CDS Start: complete

CDS End: complete

Links to sequence:

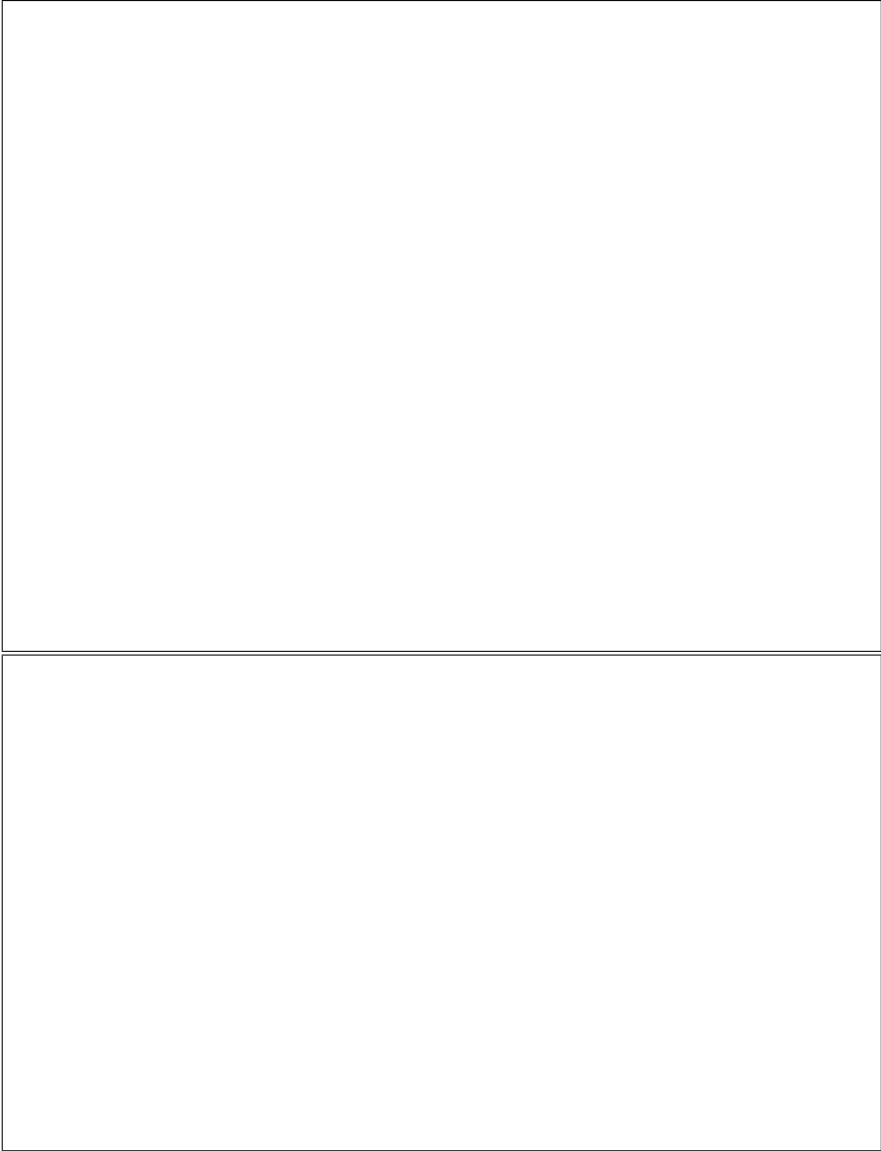
- * [Predicted Protein](#)
- * [mRNA Sequence \(may be different from the genomic sequence\)](#)
- * [Genomic Sequence from assembly](#)
- * [CDS FASTA alignment from multiple alignment](#)

Podemos regresar a la panorámica actual mediante el enlace *Genome Browser* desde el menú superior de color azul.

Podemos ejecutar acciones diferentes cuando pinchamos en distintas partes de la imagen dinámica, en función de su ubicación.

Antes de continuar el recorrido por la ficha de anotaciones de RefSeq para nuestro gen, proponemos al estudiante profundizar en alguno de estos repositorios. El enlace hacia Entrez Gene (347730), por ejemplo, nos permite acceder a otra ficha de información recopilada por este recurso también gestionado por el instituto NCBI (*National Center for Biotechnology Information*, centro nacional para la información biotecnológica de Estados Unidos). Apreciamos en esta nueva ficha que cada anotación recopilada sobre el gen *LRRTM1* posee un enlace que nos permite llegar a la fuente original de esos datos. En la Figura 20 mostramos únicamente el resumen y la bibliografía referente a este gen, omitiendo el resto de informaciones. El usuario puede estudiar la estructura exónica del gen o acceder al navegador genómico suministrado por el propio NCBI para navegar por la región colindante.

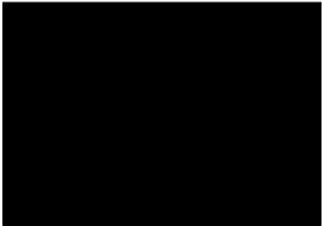
Figura 20. Información en Entrez sobre el gen *LRRTM1*.



Especialmente interesante por su importancia en estudios de análisis masivo de datos genómicos resulta la anotación recogida principalmente en la literatura científica de las funciones biológicas que desempeña este gen en la célula (ver Tabla 6).

Tabla 6. Funciones biológicas del gen *LRRTM1* (extracto a 1 de Septiembre de 2022).

<i>comportamiento locomotriz</i>	Proceso biológico	computacional
<i>internalización de receptores</i>	Proceso biológico	computacional
<i>organización de sinapsis</i>	Proceso biológico	computacional
<i>NO superficie celular</i>	Componente celular	bibliográfica
<i>matriz extracelular</i>	Componente celular	bibliográfica
<i>retículo endoplasmático</i>	Componente celular	bibliográfica
<i>cono de crecimiento</i>	Componente celular	bibliográfica
<i>axón</i>	Componente celular	bibliográfica



Desde aquí podemos acceder a la secuencia de este transcrito depositada en GenBank. Este repositorio, pionero durante la década de los ochenta, ofrece información sobre millones de secuencias genómicas y proteómicas. Cualquier ficha de GenBank está estructurada siguiendo un formato notablemente rígido. En cada línea, el tipo de característica anotada se indica primero, mientras que su valor concreto se escribe a continuación. Estas fichas generalmente incluyen información descriptiva y anotaciones sobre una determinada región para acabar con la secuencia completa de ésta. Podemos apreciar en la Figura 21 las coordenadas de los dos exones del ARN mensajero del gen *LRRTM1* (1..601 y 602..2602) así como su fracción codificante (661..2229).

LOCUS NM_178839 2602 bp mRNA linear PRI 13-MAR-2022
DEFINITION Homo sapiens leucine rich repeat transmembrane neuronal 1 (LRRTM1), mRNA.
ACCESSION NM_178839
VERSION NM_178839.5
SOURCE Homo sapiens (human)

GenBank proporciona también herramientas para manipular las secuencias (e.g. extraer subsecuencias dentro de un cierto rango de coordenadas).

Figura 21. Extracto de la secuencia NM_178839 en GenBank (a 1 de Septiembre de 2022).

ORGANISM Homo sapiens	
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.	
REFERENCE	1 (bases 1 to 2602)
AUTHORS	Beste C, Arning L, Gerding WM, Epplen JT, Mertins A, Roder MC, Bless JJ, Hugdahl K, Westerhausen R, Gunturkun O and Ocklenburg S.
TITLE	Cognitive Control Processes and Functional Cerebral Asymmetries: Association with Variation in the Handedness-Associated Gene LRRTM1
JOURNAL	Mol Neurobiol 55 (3), 2268-2274 (2018)
PUBMED	28321770
REMARK	GeneRIF: Functional cerebral asymmetries in the language domain are associated with the rs6733871 LRRTM1 polymorphism during cognitive control and top-down attention mechanisms.
(...)	
COMMENT	VALIDATED REFSEQ: This record has undergone validation or preliminary review. The reference sequence was derived from AC016670.7, BI756350.1, AY182024.1 and AA977181.1. On Nov 22, 2018 this sequence version replaced NM_178839.4.
(...)	
FEATURES	Location/Qualifiers
source	1..2602 /organism=Homo sapiens /mol_type=mRNA /db_xref=taxon:9606 /chromosome=2 /map=2p12
gene	1..2602 /gene=LRRTM1 /note=leucine rich repeat transmembrane neuronal 1
exon	1..601
(...)	
exon	602..2602
(...)	
CDS	661..2229 /gene=LRRTM1
/translation="MDFLLGLCLYWLLRRPSGVVLCLLGACFQMLPAAP	
(...)	
SCTCHQQPARECEV	
ORIGIN	

Dentro de la ficha de anotaciones de RefSeq que proporciona el navegador UCSC (ver Figura 19), observamos también un interfaz denominado *Links to sequence* (en inglés, enlaces a la secuencia) que permite extraer directamente la secuencia de nucleótidos y aminoácidos de cada gen. Mediante el enlace *predicted protein* (en inglés, proteína predicha) podemos ver el producto de la traducción de este gen, en formato FASTA (Figura 22).

Figura 22. Secuencia de aminoácidos de RRTM1.

```

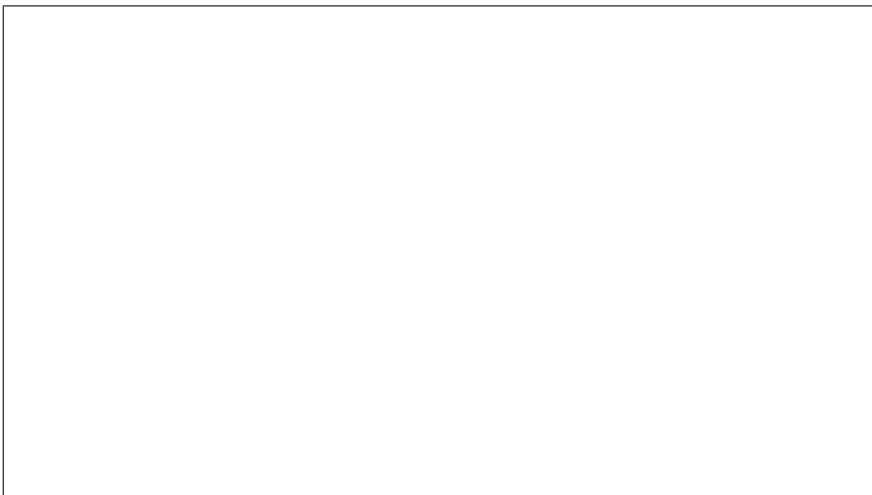
MDFLLLGLCLYWLLRRPSGVVLCLLGACFQMLPAAPSGCPQLCRCEGRLLYCEALNLTEAPHNISGLIG
LSLRYNLSSELRAQQTGLMQLTWLYLDHNHICSVQGDADFQKLRRVKELTLSSNQITQLPNTTFRPMPN
LRSVDLSYNKLQALAPDLFHGLRKLTTLHMRANAIQFVPVRIFQDCRSCLKFLDIGYNQLKSLARNSFAG
LFLKLTLEHLEHNDLVKVNFAHFPRLLISLHSLCLRRNKVAIVVSSLDWVWNLEKMDLSGNEIEYMEPHVF
ETVPHLQSLQSDNSRLTYIEPRIILNSWKSLSITLAGNLWDCGRNVCALASWLNNFQGRYDGNLQCASP
EYAQGEDVLDAVYAFHLCEDGAEPSTGHLLSAVTNRSDLGPPASSATTADGGEGQHDGTFEPATVALP
GGEHAENAVQIHKVVTGTMALIFSFLIVVLVLYVSWKCFPASLRQLRQCFVTQRRKQKQKQTMHQMAAM
SAQEYYVDYKPNHIEGALVIINEYGSCTCHQQPARECEV

```

Una secuencia en formato FASTA está compuesta por un encabezamiento, cuyo primer carácter es el símbolo ">", que informa de su origen biológico, junto con la propia secuencia a continuación, agrupada en líneas con el mismo número de caracteres.

Como comentábamos anteriormente, podemos extraer diferentes partes de la secuencia de este gen, haciendo uso en este caso de la opción *Genomic sequence from assembly* (en inglés, secuencia genómica del ensamblado). El navegador genómico nos proporciona ahora el acceso a una nueva pantalla donde podemos elegir los componentes de la estructura génica que más nos interesen (ver Figura 23). De este modo, es posible obtener un fragmento de la secuencia promotora del gen (Promoter/Upstream), la región no traducible inicial o final (5'UTR y 3'UTR), la región codificante (CDS) o bien, la secuencia del único intrón de este gen (Introns). Cuando activamos varias opciones simultáneamente podemos combinar las secuencias resultantes que constituyen el propio gen. Cada elemento génico puede separarse en diferentes secuencias en formato FASTA o integrarse en una única secuencia de salida para su posterior análisis. Finalmente, el usuario del navegador genómico puede optar por rellenar (*padding* en inglés) la secuencia original de cada parte del gen con un número determinado de nucleótidos antes (*upstream*) y después (*downstream*) de cada elemento o de la región completa. Podemos configurar el uso de mayúsculas y minúsculas en todo momento para denotar cambios de tipo de región del gen en la secuencia FASTA resultante.

Figura 23. Extraer diferentes regiones de un gen con UCSC (pista RefSeq Genes).



A continuación mostramos en la Figura 24 la secuencia completa del gen *LRRTM1*, presentada en el sentido de lectura original para facilitar la comparación con otros genes anotados en la hebra positiva. Cada exón de este gen está incluido en un fichero FASTA distinto. Utilizamos letras minúsculas para denotar la región no traducible y letras mayúsculas para indicar el fragmento codificante, que empieza con el codón ATG y finaliza con el codón TGA. El navegador UCSC genera suficiente información en el encabezamiento de cada secuencia para una fácil identificación. Podemos reconocer aquí la distribución

UCSC realiza automáticamente la operación de complementar las bases e invertir su secuencia para mostrar los datos en cualquier orientación.

>hg38_refGene NM_178839_0 range=chr2:80304152-80304752
del genoma (hg19), la pista (refGene equivale a RefSeq Genes), el código RefSeq (NM_178839), el exón (0 de 1) y su ubicación genómica.

gacactcagtgcagcagtgctgcctgtgcaagctcgccgcgcacactgcctggtggaggggaag
gagcccgccgcgcgcctcgccgctccccgcgcgcgcctccgcacctccccaccgcccgcgcgcc
gccgcgccgcgcgcgcgcaagcatgagtgcgcgcgcctctctgcagctgcccgccgcgcgaatgg
Figura 24. Secuencia del transcrito del gen *LRRTM1* (hg38)

caggctgtttccgcggagtaaaagggtggcgccggtcagtggtcggtttccaatgacggacatta
accagactgtcagatcctggggagtgcgcagccccgagtttggagttttttccccccacaacg
tcacagtcggaactgcagagggaaaggaaggcggcaggaaggcgaagctcgggctccggcacg
tagttgggaaacttgccgggtcctagaagtgcctccccgccttgccggccgccttgagccc
cgagccgagcagcaaaagtgcacattgtgcgcctgccagatccgcggccgcggaccggggct
gcctcggaacacagaggggtcttctctcgccctgcataaattagcctgcacacaaaggag
cagctgaatggaggttgtcactctctggaaaagg

>hg38_refGene NM_178839_1 range=chr2:80301878-80303878

atttctgaccgagcgtttccaatggacattctccagtctctctggaaagattctcgctaATGG
ATTTCTGCTGCTCGGTCTCTGTCTATACTGGCTGCTGAGGAGGCCCTCGGGGGTGGTCTTGT
GTCTGCTGGGGGCTGCTTTTCAGATGCTGCCCCGCCGCCAGCGGGTGCCCGCAGCTGTGCC
GGTGCAGAGGGGCGGTGCTGTACTGCGAGGCGCTCAACCTCACCGAGGCGCCCCACAACCTGT
CCGGCCTGCTGGGCTTGTCCCTGCGCTACAACAGCCTCTCGGAGCTGCGCGCCGGCCAGTTCA
CGGGGTTAATGCAGCTCACGTGGCTCTATCTGGATCACAAATCACATCTGCTCCGTGCAGGGGG
ACGCCTTTCAGAACTGCGCCGAGTTAAGGAACCTCACGCTGAGTTCCAACCAGATCACCCAAC
TGCCCAACACCACCTTCCGGCCCATGCCAACCTGCGCAGCGTGGACCTCTCGTACAACAAGC
TGCAGGCGCTCGCGCCCGACCTCTTCCACGGGCTGCGGAAGCTCACACGCTGCATATGCGGG
CCAACGCCATCCAGTTTGTGCCCCGTGCGCATCTTCCAGGACTGCCGAGCCTCAAGTTTCTCG
ACATCGGATACAATCAGCTCAAGAGTCTGGCGCGCAACTCTTTCGCCGGCTTGTTTAAGCTCA
CCGAGCTGCACCTCGAGCACAACGACTTGGTCAAGGTGAACTTCGCCCACTTCCCGCCCTCA
TCTCCCTGCACTCGCTCTGCCTGCGGAGGAACAAGGTGGCCATTGTGGTCAGCTCGCTGGACT
GGGTTTGAACCTGGAGAAAATGGACTTGTGCGGCAACGAGATCGAGTACATGGAGCCCCATG
TGTTTCGAGACCGTGCCGCACCTGCAGTCCCTGCAGCTGGACTCCAACCGCCTCACCTACATCG
AGCCCCGGATCCTCAACTCTTGAAGTCCCTGACAAGCATCACCTGGCCGGGAACCTGTGGG
ATTGCGGGCGCAACGTGTGTGCCCTAGCCTCGTGGCTCAACAACCTCCAGGGGCGCTACGATG
GCAACTTGCAGTGCGCCAGCCCGAGTACGCACAGGGCGAGGACGTCCTGGACGCCGTGTACG
CCTTCCACCTGTGCGAGGATGGGGCCGAGCCACCAGCGGCCACCTGCTCTCGGCCGTACCA
ACCGCAGTGATCTGGGGCCCCCTGCCAGCTCGGCCACCACGCTCGCGGACGGCGGGGAGGGGC
AGCACGACGGCACATTCGAGCCTGCCACCGTGGCTCTTCCAGGCGGCGAGCACGCCGAGAAGC
CCGTGCAGATCCACAAGGTGGTCACGGGCACCATGGCCCTCATCTTCTCCTTCATCGTGG
TCCTGGTGCTCTACGTGTCTGGAAGTGTTCAGCCAGCCTCAGGCAGCTCAGACAGTGCT
TTGTACGCGAGCGCAGGAAGCAAAAGCAGAAACAGACCATGCATCAGATGGCTGCCATGTCTG
CCCAGGAATACTACGTTGATTACAAACCGAACCACATTGAGGGAGCCCTGGTGATCATCAACG
AGTATGGCTCGTGTACCTGCCACCAGCAGCCCGCGAGGGAATGCGAGGTGTGAttgtccagtg
ggctctcaacccatgcgctaccaaatacgccctgggcagccgggacggggccggcgggcaccagg
ctgggggtctccttgtctgtgctctgatatgctccttgactgaaactttaaggggatctctccc
agagacttgacatttttagctttattgtgtcttaaaaaacaaaagcgaattaaaacacacaaaa
aaccaccacccacacaccttcaggacagctctatcttaatttcatatgagaactccttcctccc
tttgaagatctgtccatattcaggaatctgagagtgtaaaaaaggtaccaatcattgattttt
tttttttttgtaaaactaaaatgttttaaaataaaatagcatttacagtt

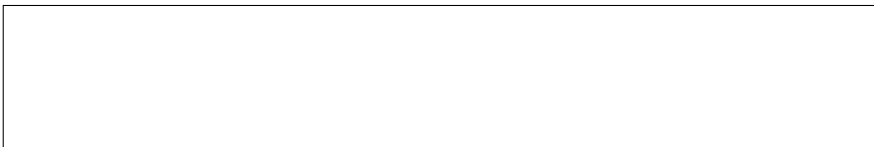
Junto con los transcritos humanos revisados manualmente, el proyecto RefSeq produce anotaciones de referencia para varias especies de vertebrados. Esta información resulta extremadamente útil para identificar regiones funcionales conservadas a lo largo de la evolución en otros organismos. De este modo, mediante las oportunas comparaciones entre las distintas anotaciones de RefSeq, el propio navegador UCSC habilita una pista adicional dentro del bloque de anotaciones génicas denominada *Other RefSeq* para mostrar las formas ortólogas de cada gen en otros genomas (Figura 25). Con esta información, el usuario puede optar por seleccionar el mismo transcrito en otra especie y el navegador genómico realiza el cambio de genoma de referencia en el visor principal para acceder al gen ortólogo anotado en esta segunda especie. Ya sea bien a causa del algoritmo de alineamiento o porque algunas regiones del transcrito humano no se conservan igual de bien que otras, es frecuente que en algunas especies sólo visualicemos una fracción de la secuencia génica alineada con el genoma de referencia.

Figura 25. Formas ortólogas del gen humano *KCMF1* en RefSeq (hg38)



A nivel general, es posible evaluar la conservación evolutiva a nivel de secuencia explorando todas las posiciones de cada cromosoma, no únicamente las pertenecientes a los genes. Con este objetivo, el bloque de pistas de genómica comparada (en inglés, *comparative genomics*) permite activar la visualización de esta información en nuestro navegador genómico. Como podemos ver en la Figura 26, disponemos de varias pistas que albergan los resultados de varias comparaciones del genoma que estamos utilizando de referencia (humano, en este caso la distribución NCBI38/hg38) contra los genomas ensamblados de otros vertebrados.

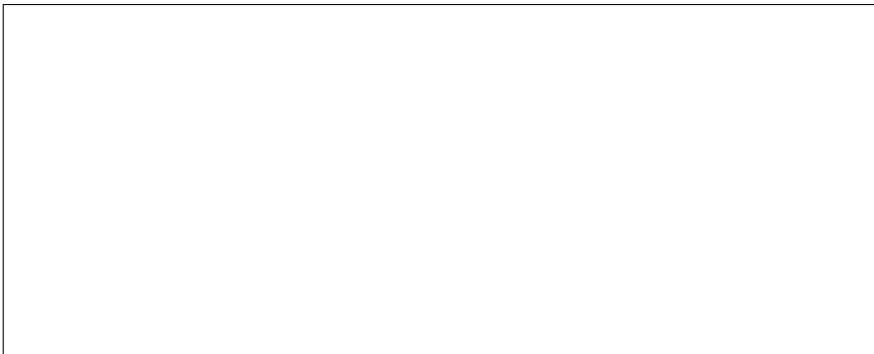
Figura 26. Bloque de genómica comparada



La comparación entre organismos cuya evolución difiere a partir de determinados puntos del árbol filogenético permite descubrir aquellos elementos funcionales conservados entre distintas secuencias. Para proporcionar estos datos con celeridad, todos los análisis están precalculados de antemano en el interior del navegador. Con estas pistas, por ejemplo, podemos establecer la existencia de regiones codificantes a través de los alineamientos de las secuencias homólogas (ver Figura 27). En estas representaciones gráficas resulta especialmente interesante destacar cómo la conservación del genoma de una especie respecto al genoma humano es mayor cuanto más cerca está evolutivamente de nosotros. Por ejemplo, la práctica totalidad de la secuencia del genoma de macaco (subpista *Rhesus* en verde) es idéntica a la del genoma humano, mientras que a medida que nos alejamos en el árbol filogenético de especies, observamos que sólo la región donde se ubica el gen *LRRTM1* continúa precisamente conservada en todos los casos (Figura 27).

Figura 27. Conservación de secuencias funcionales con la superpista *Conservation*

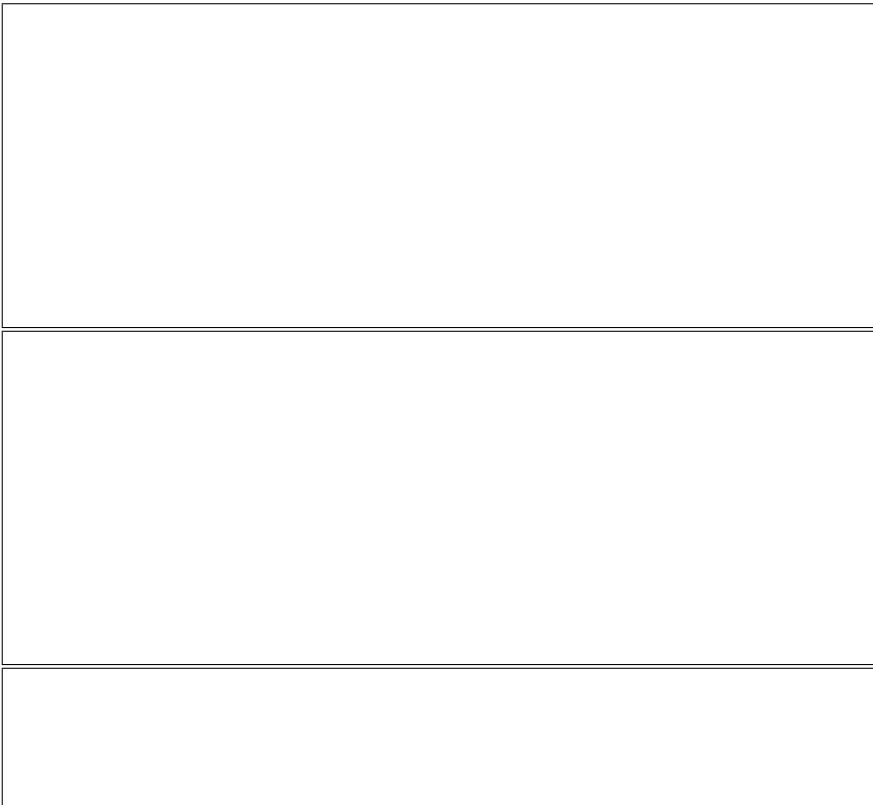
Con el navegador UCSC podemos configurar numerosos parámetros para incluir diferentes comparaciones. Para ello, el usuario debe pinchar sobre el enlace coloreado en azul que indica el nombre de cada pista de este bloque. La superpista *Conservation* contiene alineamientos entre la mayoría de los grupos de especies más representativos. Dichas comparaciones se han realizado empleando varias aplicaciones bioinformáticas distintas (p.e. *phastCons*). De este modo, en la pantalla de configuración, por ejemplo, disponemos de la opción de activar/desactivar la estrategia de comparación, incorporar o retirar genomas en el visor genómico, modificar el color del gráfico de picos en función de las pautas de lectura o activar diferentes modos de alineamiento para la identificación de regiones conservadas:

Figura 28. Configuración de la superpista *Conservation*

Aprovecharemos la subpista *phastCons* perteneciente a la superpista *Conservation* para introducir varios conceptos sobre la configuración gráfica de las pistas que contienen distribuciones continuas de valores en UCSC. En la Figura 29 observamos la pantalla de configuración de esta subpista. Para acceder a esta pantalla podemos emplear el enlace coloreado en azul *Element Conservation (phastCons)* de la pantalla principal (panel superior) o directamente con el botón derecho del ratón sobre la propia pista en el navegador genómico. Entre los parámetros más relevantes encontramos el modo de escalado vertical (*Data view scaling*): autoescalado (*auto-scale* en inglés) para mostrar los datos ocupando todo el espacio en el eje Y disponible o determinación de rangos máximo y mínimo (*range setting*) para indicarle al navegador los puntos para recortar y visualizar esa parte de la distribución de valores. En términos de espacio físico en pantalla que ocupará la parte de la pista anteriormente seleccionada, podemos jugar también con el valor de la altura de la pista (*track height*, en píxeles). La casilla *Negate values* genera una versión especular de la misma distribución de valores que en determinadas situaciones puede resultar ventajosa para mostrar estas informaciones de forma más sugerente.

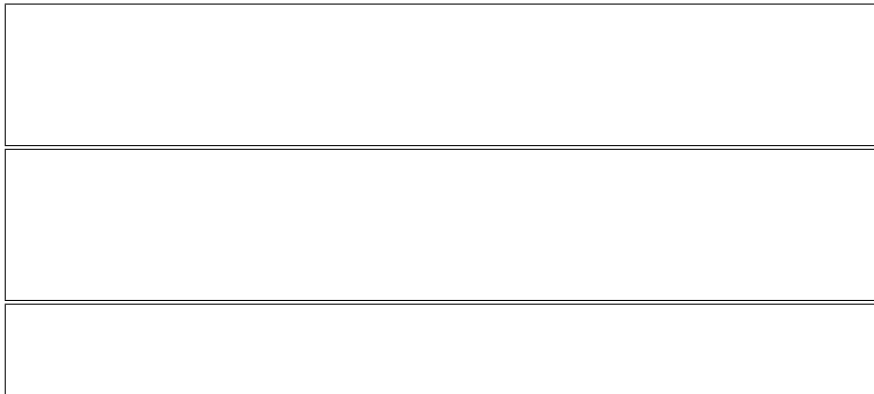
Figura 29. Configuración de la subpista `phastCons`.

La herramienta BLAT (accesible desde la barra de menú superior de color azul) complementa perfectamente el bloque de pistas de conservación. BLAT (*Blast-like alignment tool*) es un programa de alineamiento de secuencias que identifica las regiones más similares a aquella propuesta por el usuario en un determinado genoma. A diferencia de otros programas de alineamiento que veremos más adelante, BLAT está especializado en la detección de secuencias prácticamente idénticas (hecho que aumenta su velocidad de respuesta de forma notable). Podemos usar esta aplicación para identificar, por ejemplo, cuál es la ubicación de cualquier secuencia codificada dentro del mismo genoma.

Figura 30. Empleando BLAT para identificar la ubicación del gen *LRRTM1*.

BLAT proporciona una clasificación de posibles ubicaciones de nuestra secuencia ordenadas por similitud. Aquella secuencia que seleccionamos como resultado, es posteriormente integrada como una nueva pista dentro del navegador genómico. Resulta especialmente interesante el empleo de BLAT para localizar la posición de la región más similar a nuestra secuencia en otro genoma. Con este procedimiento, podemos detectar para el gen humano *LRRTM1*, su correspondiente ortólogo anotado en otras especies:

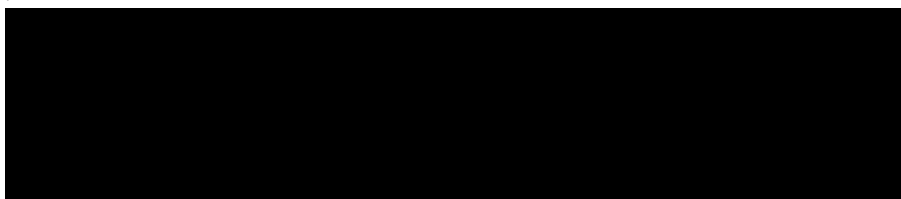
Figura 31. Utilizando BLAT para identificar formas ortólogas del gen *LRRTM1*.



Presentamos el gen humano (hg38) en la parte superior de la imagen, su ortólogo en ratón (mm10) en el centro y en el genoma de la rata (rn6) en la parte inferior. En este caso hemos empleado la proteína humana para llevar a cabo las búsquedas.

1.5. El navegador genómico UCSC: anotaciones propias

Los navegadores genómicos proporcionan un enorme volumen de información. El estudio exhaustivo de la cartografía existente para los distintos elementos funcionales codificados en el genoma resulta fundamental para abordar cualquier problema biológico. Sin embargo, cualquier grupo de investigación durante el proceso de elaboración de hipótesis, habitualmente genera sus propios resultados en función de los experimentos llevados a cabo. En consecuencia, es fundamental que el bioinformático conozca los entresijos del manejo de estos navegadores genómicos para enriquecer las informaciones originales con los resultados experimentales obtenidos en el laboratorio. En última instancia, estas mismas herramientas permiten gestionar eficientemente el almacén permanente de estas nuevas configuraciones de pistas e implementan mecanismos para la compartición de los datos entre grupos colaboradores, potenciando exponencialmente la utilidad de esta clase de visualizaciones.



La introducción de anotaciones propias dentro de un navegador genómico resulta esencial para realizar comparaciones entre el conocimiento existente y los nuevos resultados experimentales, favoreciendo también la generación de representaciones gráficas de alta calidad de nuestros propios resultados. De hecho, una vez el usuario introduce sus propias pistas, éstas son automáticamente integradas dentro de la imagen proporcionada por el navegador. Este hecho abre todo un nuevo abanico de opciones de trabajo para realizar el análisis comparativo de las pistas. Podemos importar múltiples pistas propias dentro del navegador, utilizando el navegador para estudiar el comportamiento de nuestros resultados. De este modo, todas las herramientas disponibles para llevar a cabo posteriores análisis sobre pistas, tanto a nivel cualitativo como cuantitativo, pueden trabajar indistintamente sobre las pistas distribuidas por defecto con el navegador y las pistas propias del usuario.

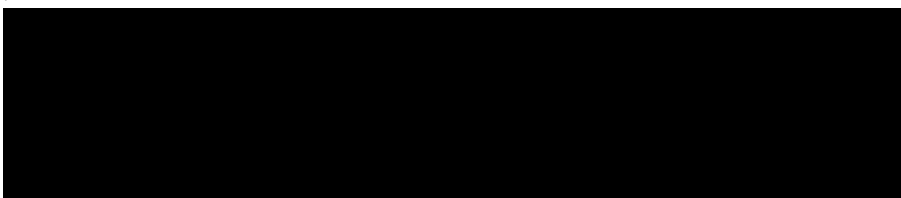
Dentro del entorno de trabajo del navegador UCSC, las pistas subidas al sistema por el usuario reciben el nombre de *custom tracks* (en inglés, pistas adaptadas). Desde la pantalla principal del visor genómico, el usuario debe presionar el botón *Add custom tracks* (añadir pistas) para incorporar una nueva pista de datos (regresar a la Figura 8). Si la operación de carga de nuestra pista finaliza satisfactoriamente, ésta es incorporada al inventario de pistas propias y podemos visualizar todos sus elementos a lo largo del genoma junto con el resto de anotaciones convencionales. Además, el botón de carga pasa a denominarse *Manage custom tracks* (gestionar pistas en inglés, ver Figura 33). Dado que estos datos son procesados como cualquier pista, podemos hacer uso de las opciones habituales. Resultan particularmente útiles para la mayoría de usuarios la generación de figuras en alta resolución mostrando datos de nuestras pistas y la comparación con otras bases de datos primarias. Aquellas aplicaciones del propio navegador que generan nuevos resultados ofrecen habitualmente la opción de estructurarlos para ser mostrados como pistas propias dentro de nuestro entorno de trabajo. Por ejemplo, como puede apreciarse en las anteriores Figuras 30 y 31, el programa BLAT añade una nueva pista al visor gráfico con los resultados de cada búsqueda.

UCSC permite la carga directa de ficheros comprimidos (p.e. archivos gzip), para facilitar una rápida transferencia de los datos a través de la Red.

Figura 33. Carga y visualización de pistas propias.

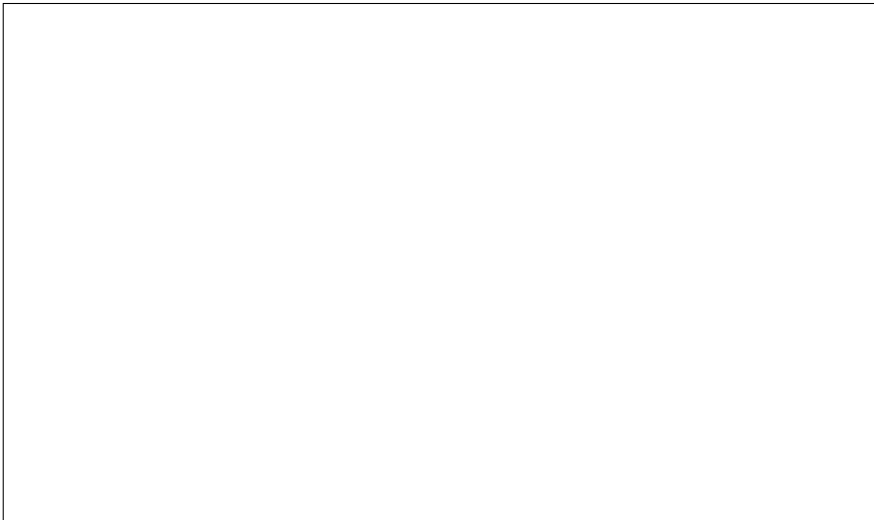


El navegador UCSC asocia un código interno a la conexión realizada desde nuestro ordenador, gestionando internamente las pistas estándar que estamos visualizando y el conjunto de pistas propias que hemos cargado en el sistema. De este modo, estos resultados únicamente pueden ser visualizados desde nuestro ordenador, no siendo accesibles para el resto de la comunidad de usuarios. Por regla general, el navegador mantiene en memoria nuestras pistas durante uno o dos días únicamente, para no almacenar un volumen excesivo de datos. Lógicamente, en la mayoría de ocasiones es necesario acceder a estas nuevas pistas durante más tiempo. Por esta razón, el navegador UCSC ofrece un sistema gratuito de sesiones de usuario protegidas bajo contraseña.

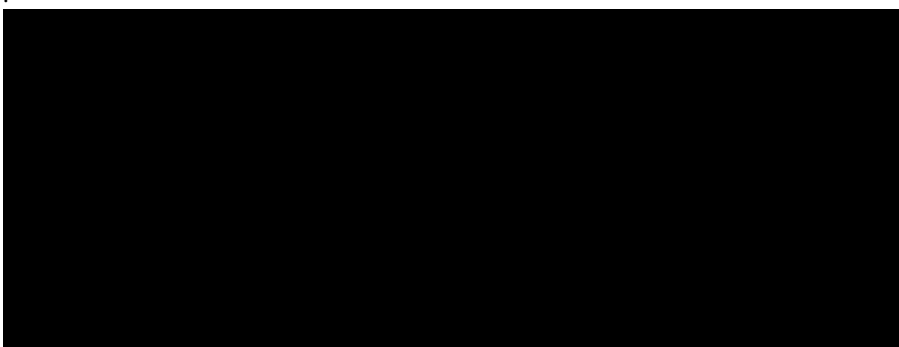
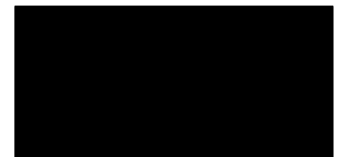


Mediante el enlace *MyData->MySessions* del menú principal, el usuario registrado puede guardar el estado actual del navegador (*Save Settings* en inglés) en una nueva sesión o sobrescribir una ya existente. A dicha sesión se asignará a una dirección de internet concreta, que podremos compartir para facilitar la comunicación de datos entre nuestros diversos colaboradores según nuestra conveniencia.

Figura ?? . Gestor de sesiones del navegador UCSC.



El bioinformático puede crear sus propias pistas en numerosos formatos dentro de UCSC. En estos materiales vamos a focalizar nuestro interés en los más extendidos. Previamente a la carga en el sistema, el navegador UCSC procesa sintácticamente nuestras pistas para verificar su corrección. En caso de detectarse un error de formato, el navegador nos informará apropiadamente para subsanarlo con facilidad.



En primer lugar, el usuario debe configurar el comportamiento inicial del navegador empleando la palabra clave `browser`. Para ello, es necesario introducir las coordenadas de la región genómica e indicar el conjunto de pistas convencionales que deseamos activar para comparar con nuestras anotaciones (utilizando distintos modos de visualización).

Figura 34. Encabezamiento de pistas propias (1).

```
browser position cromosoma:inicio-final
browser [hide/dense/pack/full] pista 1
...
browser [hide/dense/pack/full] pista n
```

A continuación, dentro del mismo fichero que subiremos al navegador, introduciendo la palabra clave `track` debemos definir los parámetros de visualización básicos de cada nueva pista: asignar un nombre y una descripción, definir un modo de visibilidad o seleccionar un color. Tras cada línea de encabezamiento debemos aportar las coordenadas de nuestras anotaciones genómicas para integrarlas en la visualización final (denotado por puntos suspensivos en la Figura 35). Cada línea corresponde exclusivamente a una anotación efectuada por el usuario, introduciéndose el carácter tabulador “\t” para separar los distintos atributos. Para cada tipo de característica genómica debemos elegir el formato de anotación más adecuado. Para elementos genómicos delimitados por pares de coordenadas (inicio,fin) es apropiado utilizar el formato BED o el estándar GFF. En cambio, para visualizar una determinada característica biológica a lo largo de todas las posiciones del genoma debemos emplear los formatos BEDGRAPH o WIG.

El usuario puede guardar más de una pista dentro de un mismo fichero. Es posible mezclar pistas codificadas en el mismo o en diferentes formatos en su interior.

Recomendamos registrar en todo momento la versión del genoma en la que estamos trabajando. Nuestras anotaciones propias deben interpretarse dentro de esa misma secuencia de referencia.

Figura 35. Encabezamiento de pistas propias (2).

```
track name=nombre1 description=descripcion visibility=[0-4] color=R,G,B
...
track name=nombreN description=descripcion visibility=[0-4] color=R,G,B
...
```

El formato BED (*Browser Extensible Format*, del inglés formato extensible del navegador) es útil para plasmar en pantalla aquellas anotaciones constituidas por elementos genómicos definidos por un rango de coordenadas. Dentro de este grupo de características biológicas están clasificados los transcritos, los exones o los sitios de unión a factores de transcripción. En su formato más simple, podemos codificar anotaciones proporcionando su localización en el genoma. Para añadir atributos adicionales (e.g. el nombre, el color o un valor asociado a su fiabilidad) debemos extender la línea actual con esos valores. Es posible introducir un rango de coordenadas adicional para definir distintos grosores en la anotación gráfica (e.g. regiones codificantes y no traducibles de los genes en la Figura 36).

Figura 36. Definición básica y extensión del formato BED

```
cromosoma inicio final
...
cromosoma inicio final nombre score hebra inicio2 final2
```

Hemos ilustrado con un ejemplo real el comportamiento de las anotaciones codificadas en formato BED (ver Figura 37). Tomando como referencia la información que conocemos sobre la ubicación del gen *LRRTM1* en el ensamblado hg38, analizado a lo largo de este capítulo, hemos generado nuestras propias pistas para resaltar distintas regiones en su interior. A partir de las coordenadas para el propio transcrito, sus exones y la región codificante, podemos generar tres pistas diferentes (pistas 1, 2 y 3 en la Figura 37), resaltando un área diferente del gen en cada caso. Para denotar con un trazo de distinto grosor la frontera entre la región codificante y la región no traducible, debemos extender el formato BED para añadir una segunda pareja de coordenadas (inicio del CDS, pista 4). Finalmente, para emular la propia representación de la pista de RefSeq (con los intrones marcados con una línea recta entre exones y la orientación del gen indicada mediante flechas), enriquecemos el formato añadiendo más información a continuación sobre el número de exones, la longitud de éstos y sus puntos de inicio (bloques en la pista 5, Figura 37).

Figura 37. Anotaciones del gen *LRRTM1* en el formato BED (hg38).

```

browser position chr2:80301877-80304752
browser hide all
browser pack refGene
track name=gene description="1.GENE" visibility=1 color=10,10,10
chr2 80301877 80304752
track name=cds description="2.CDS" visibility=1 color=50,50,50
chr2 80302250 80303819
track name=exons description="3.EXONS" visibility=1 color=100,100,100
chr2 80301877 80303878
chr2 80304151 80304752
track name=utr description="4.UTR" visibility=1 color=150,150,150
chr2 80301877 80303878 EXON 1 - 80302250 80303819
chr2 80304151 80304752 EXON 2 - 80304151 80304151
track name=full description="5.FULL" visibility=2 color=200,200,200
chr2 80301877 80304752 GENE 1 - 80302250 80303819 0 2 2001,601 0,2274

```

Un aspecto importante a destacar es la coexistencia de dos sistemas de coordenadas de referencia distintos dentro del navegador UCSC. En primer lugar, una vez estamos interactuando con el visor gráfico en la *web*, las anotaciones vienen posicionadas en los cromosomas asumiendo que la primera posición de cada cromosoma es la posición número uno. Sin embargo, en paralelo, internamente el navegador almacena los datos de cada pista considerando que la primera posición de cada cromosoma es la posición cero. De esta manera, la diferencia entre la primera alternativa denominada **totalmente cerrado** (*fully-closed* en inglés) y la segunda, que recibe el nombre de **medio abierto** (*half-open* en inglés), estriba en la forma de indicar la primera coordenada de cada anotación. Por ejemplo, el visor gráfico de UCSC nos informa con el sistema totalmente cerrado que nuestro gen de interés *LRRTM1* está ubicado en la región chr2:80301878-80304752 (pista RefSeq Genes (hg38)). Pero en la Figura 37 hemos creado precisamente la primera pista empleando la fórmula (chr2,80301877,80304752), basada en la notación medio abierta. Si tenemos en cuenta este detalle de precisión cuando generemos nuestras propias pistas, el navegador será el encargado posteriormente de realizar el cambio entre sistemas en los momentos apropiados. La ventaja de emplear una anotación *half-open* básicamente consiste en que para el cálculo de la longitud de una región en formato BED solo es preciso hacer una resta, mientras que para la nomenclatura *fully-closed* es preciso sumar una unidad además.

De hecho, ambas aproximaciones también pueden citarse como **1-based** y **0-based** respectivamente, por su primera posición de referencia en cada cromosoma.

El estándar GFF (*General Feature Format*, en inglés formato de características generales) fue concebido para permitir la transmisión eficiente de anotaciones génicas entre diferentes aplicaciones bioinformáticas. Permite generar representaciones gráficas similares al formato BED, dedicando cada línea a un elemento genómico distinto. Posee la particularidad de que podemos asociar un identificador propio individualmente a cada anotación. Si utilizamos un mismo nombre para un conjunto de anotaciones relacionadas, el navegador interpretará que éstas pertenecen al mismo grupo (e.g. los exones de un gen). Es posible asignar una puntuación a cada registro, reflejándose gráficamente en cajas de diferentes grosores o con una gama de colores de distintas intensidades.

El formato GFF presenta la siguiente sintaxis básica:

Figura 38. Definición del formato GFF

```
# comentario
secuencia origen característica inicio final valor hebra pauta grupo
```

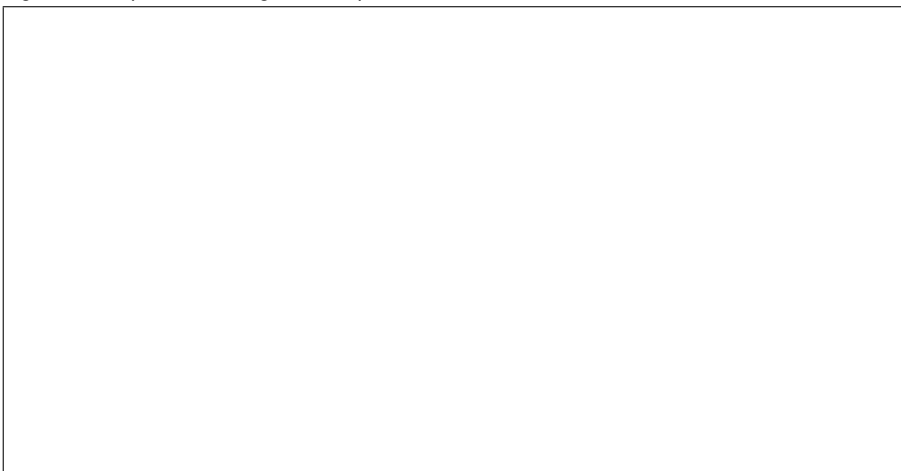
En el siguiente ejemplo hemos codificado el mismo gen *LRRTM1* en el formato GTF (Gene Transfer Format, una adaptación del formato GFF versión 2 aceptada por el navegador UCSC). Nótese como especificamos cada componente génico en líneas separadas y empleamos la notación medio abierta (*0-based*) para especificar la posición de cada elemento. En el formato GTF el último campo permite albergar dos identificadores para cada registro.

Figura 39. Anotación del gen *LRRTM1* en el formato GFF (hg38).

```
browser position chr2:80301877-80304752
browser hide all
browser pack refGene
track name=GENE_GFF-GTF
chr2 hg38_refGene stop_codon 80302251 80302253 0.000000 - . gene_id "NM_178839"; transcript_id "NM_178839"
chr2 hg38_refGene CDS 80302254 80303819 0.000000 - 0 gene_id "NM_178839"; transcript_id "NM_178839"
chr2 hg38_refGene start_codon 80303817 80303819 0.000000 - . gene_id "NM_178839"; transcript_id "NM_178839"
chr2 hg38_refGene exon 80301878 80303878 0.000000 - . gene_id "NM_178839"; transcript_id "NM_178839"
chr2 hg38_refGene exon 80304152 80304752 0.000000 - . gene_id "NM_178839"; transcript_id "NM_178839"
```

Este formato está ampliamente extendido entre los desarrolladores de aplicaciones bioinformáticas. Por ejemplo, existen numerosos programas que procesan anotaciones en formato GFF para construir representaciones gráficas de calidad (e.g. los mapas de anotaciones producidas en publicaciones que anuncian la secuenciación de los genomas más populares). Por ejemplo, gracias al programa GFF2PS, las predicciones pueden volcarse en PostScript, fácilmente convertible a otros formatos de impresión más populares. El eje central delimita si las predicciones se encuentran en la hebra positiva o negativa de la molécula de ADN. Para personalizar la representación gráfica, el programa GFF2PS permite introducir un fichero de configuración para asignar colores y otros atributos gráficos a cada característica genómica anotada.

Figura 40. Representación gráfica de predicciones con GFF2PS



El formato BEDGRAPH es útil para subir a nuestra sesión del navegador genómico información sobre elementos cuya ubicación más probable viene definida por una distribución continua de valores que varía en intensidad a lo largo de distintas regiones del genoma. Los resultados de experimentos de secuenciación masiva para cartografiar la posición de las modificaciones de histonas y factores de transcripción en los genomas o para reconstruir los mapas transcriptómicos de expresión génica son claros ejemplos de esta familia de anotaciones. Para codificar estos elementos correctamente en el navegador, el bioinformático debe asignar un valor numérico a cada posición en los cromosomas. Esta cantidad representa la intensidad de la señal biológica anotada por una máquina de secuenciación durante el análisis de datos genómicos a gran escala.

Para introducir este tipo de pistas en el navegador debemos indicar explícitamente en su encabezamiento que el tipo de pista es `bedGraph`. Dado que es una pista continua, es posible modificar los mismos parámetros de visualización asociados que ya estudiamos en el caso de la subpista `phastCons` (p.e. *auto-scale* en la Figura 29).

Figura 41. Definición básica del formato BEDGRAPH

```
track type=bedGraph name=nombre description=descripcion
    autoScale=on|off maxHeightPixels=max:default:min
    color=R,B,G altColor=R,G,B visibility=[hide/dense/full]
    cromosoma inicio final valor
```

...

A continuación, mostramos en formato *full* y *dense* una pista de ejemplo que produce un pequeño pico en la región promotora del gen *LRRTM1*. El estudiante puede comprobar alejándose y acercándose con el encuadre como el pico es generado realmente mediante una sucesión de valores con distintas intensidades asociados a ventanas consecutivas (*bins* en inglés) del genoma.

Figura 42. Carga de una pista en formato BEDGRAPH (hg38).

```
track type=bedGraph name=profile1 description="PROFILE" color=250,150,10 altColor=10,10,200
chr2 80304800 80304900 10
chr2 80304900 80305000 20
chr2 80305000 80305100 50
chr2 80305100 80305200 -20
chr2 80305200 80305300 -40
chr2 80305300 80305400 20
chr2 80305400 80305500 40
chr2 80305500 80305600 -10
chr2 80305600 80305700 20
chr2 80305700 80305800 10
```

El formato WIG (*wiggle* en inglés, agitado) es una alternativa más flexible que el formato BEDGRAPH para representar el mismo tipo de información. Ahora podemos asignar intensidades a intervalos de posiciones de diferente tamaño sin necesidad de especificarlas explícitamente todas ellas. La palabra clave `variableStep` debe utilizarse cuando la distancia física entre dos anotaciones consecutivas de nuestra pista es variable. En cambio, cuando los propios datos están ubicados a intervalos regulares de distancia podemos introducir la instrucción `fixedStep`. Cuando un cierto rango de posiciones contiguas presenta el mismo valor, es posible resumir varias anotaciones en una única línea mediante la palabra clave `span`.

Los archivos WIG pueden convertirse en ficheros BIGWIG, que ocupan notablemente menos espacio, con las utilidades de UCSC.

Figura 43. `variableStep` y `fixedStep` en formato WIG

```

coordenada   valor
...
variableStep chrom=cromosoma span=largo
coordenada   valor
...
fixedStep chrom=cromosoma start=inicio span=largo step=paso
coordenada   valor
...

```

Ahora procedemos a cargar una pista en formato WIG para anotar una hipotética función de probabilidad asociada a la localización del final de transcripción de nuestro gen *LRRTM1* (utilizamos dos modos distintos de visualización).

Figura 44. Carga de una pista en formato WIG (hg38).

```

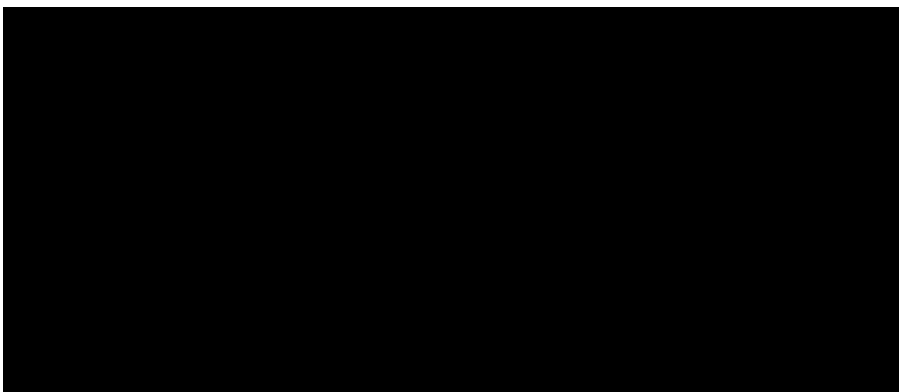
browser position chr2:80301877-80304752
browser hide all
browser pack refGene
track type=wiggle_0 name=TES1 description=TES1 color=50,150,50 visibility=2
variableStep chrom=chr2
80301877      1
80301878      4
80301879      9
80301880     50
80301881    100
80301882     50
80301883      9
80301884      4
80301885      1

track type=wiggle_0 name=TES2 description=TES2 color=50,150,50 visibility=1
variableStep chrom=chr2
80301877      1
80301878      4
80301879      9
80301880     50
80301881    100
80301882     50
80301883      9
80301884      4
80301885      1

```

1.6. El navegador genómico UCSC: grupos de pistas propios

La integración de nuestras propias pistas de datos en el navegador otorga gran flexibilidad a nuestros análisis. No obstante, una vez el usuario comienza a dar de alta nuevas pistas, el bloque de *custom tracks* puede crecer y ser difícil localizar y gestionar nuestros propios datos. Para evitar estas situaciones, el propio navegador UCSC implementa dos mecanismos para que sea factible agrupar nuestros catálogos de pistas para compartir con nuestro entorno de trabajo. En un extremo, vamos a ver que resulta extremadamente sencillo generar nuestras propias superpistas para ser gestionadas de forma inmediata en el visor gráfico como una sola unidad. Por otro lado, aunque requiera una mayor complejidad técnica, también es posible diseñar nuestros propios bloques de pistas.



A continuación hemos creado una nueva superpista constituida por la pista de conservación *phastCons* y la pista de medición del contenido en GC del genoma. Para ello, primero hemos accedido al constructor de colecciones de pistas en el enlace *My Data->Track Collection Builder*. Después, hemos buscado y seleccionado en la parte izquierda de la pantalla ambas pistas, para añadirlas a la nueva superpista que se va configurando en paralelo en la parte derecha según nuestras elecciones. Cuando regresamos al visor principal ambas pistas aparecen agrupadas juntas unidas por el marco común de color gris a su izquierda.

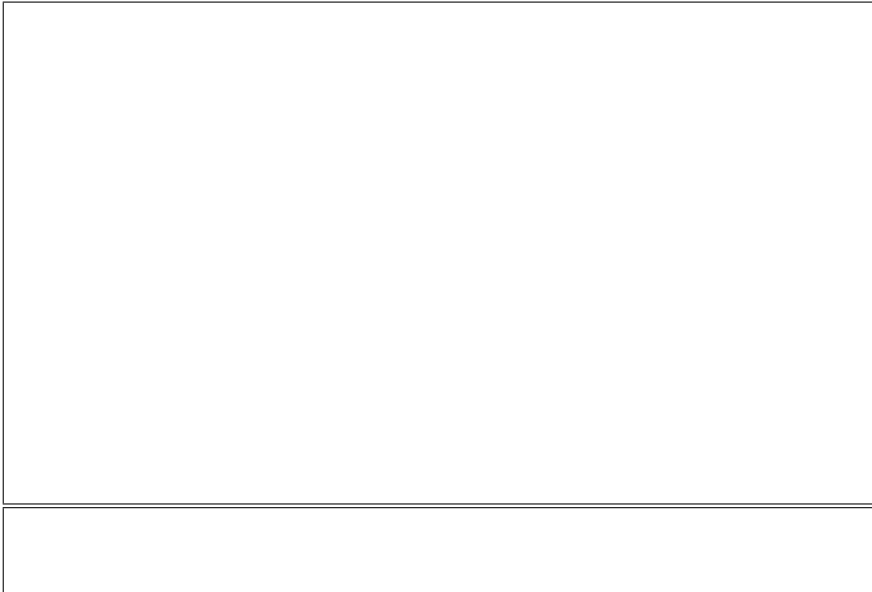
El marco gris a la izquierda de las pistas constituye un acceso alternativo a la pantalla de configuración de cualquier pista.

Figura 45. Creación y manejo de una superpista con el navegador UCSC.



Una vez generada la superpista podemos acceder a su pantalla de configuración conjunta donde es posible editar los atributos gráficos de forma conjunta o separada en función de nuestras necesidades. Simultáneamente un nuevo bloque de pistas aparece en el menú principal para integrar esta superpista al catálogo actual de pistas de nuestra sesión de trabajo. Además de mostrarse de forma separada, en función del tipo de pistas propias que añadamos, es posible configurar también el modo en que ambas anotaciones se superpongan visualmente con la característica *Merge method* (método de fusión en inglés).

Figura 46. Creación y manejo de una superpista con el navegador UCSC.



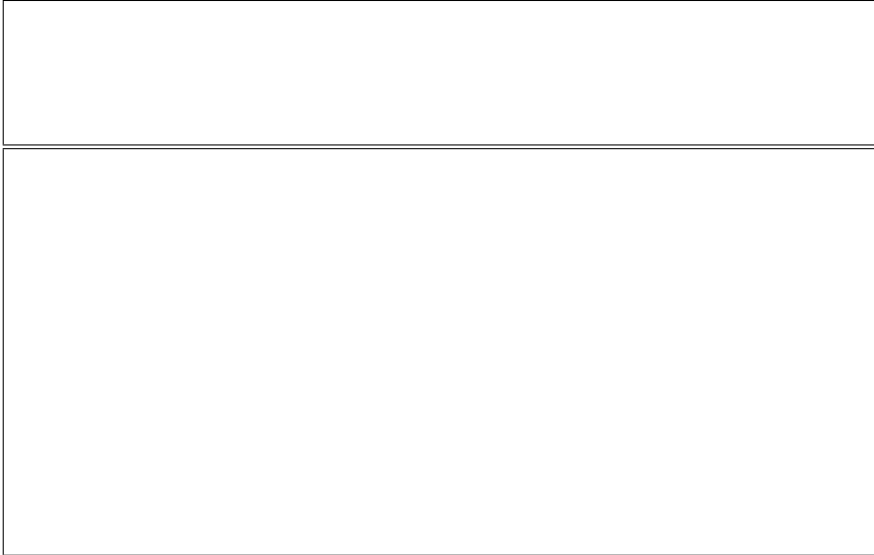
Como hemos visto con anterioridad en la Tabla 5, el navegador UCSC ofrece una serie de bloques de pistas predefinidas con anotaciones mapeadas sobre la versión del genoma en la que estamos trabajando. A pesar de que el repertorio inicial de pistas es voluminoso, es cierto que a menudo es preciso trabajar con datos generados específicamente por consorcios especializados en determinadas características genómicas. Para acceder a estos nodos de pistas externos podemos utilizar el botón *track hubs* debajo del visor gráfico principal. En la pantalla de selección de nuevos bloques de pistas se mostrará el inventario de completo bloques para su interrogación. En el siguiente ejemplo vamos a activar el nodo de pistas asociadas al recurso primario EPD (Eukaryotic Promoter Database). Para incluir este nuevo bloque en el navegador debemos proceder a conectarlo. Es importante activar el *track hub* en la misma versión del ensamblado que estamos visualizando en nuestra sesión de trabajo actual. Aquí seleccionaremos el ensamblado hg38 para el recurso EPD.

Figura 47. Activación de un nuevo nodo de pistas en el navegador UCSC.



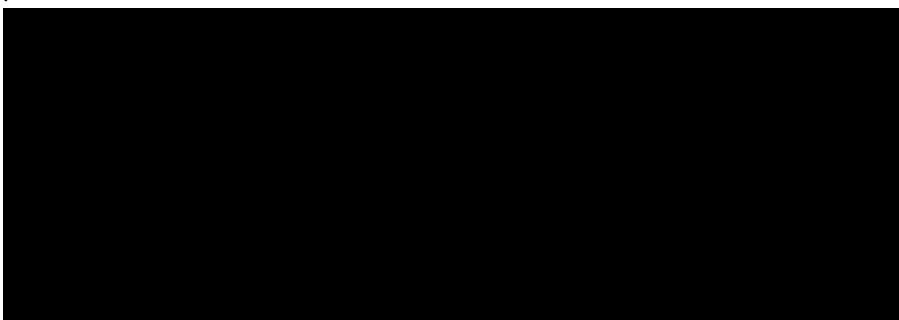
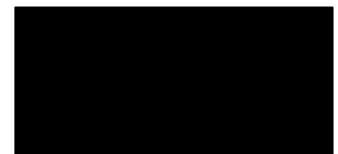
Una vez efectuada esta conexión el bloque de pistas de EPD pasa a formar parte de nuestra sesión de trabajo. Esto significa que podemos visualizar cualquiera de sus pistas en la pantalla principal del navegador. En este caso (ver Figura 48), las pistas generadas por la base de datos EPD incluyen información de secuenciación masiva útil para anotar con alta precisión el inicio de transcripción de los genes humanos.

Figura 48. Visualización de un nuevo nodo de pistas en nuestro navegador UCSC.



1.7. El navegador genómico UCSC: el navegador de tablas

En las secciones anteriores hemos visto que el navegador genómico UCSC ofrece un potente interfaz gráfico para interactuar con los elementos anotados a lo largo de los cromosomas. Si bien esta metodología de trabajo es enormemente útil en multitud de escenarios, no es menos cierto que en determinadas situaciones el usuario de estas aplicaciones necesita trabajar con las anotaciones de las pistas de forma global para extraer nuevas conclusiones en forma de estadísticas o nuevas pistas. Internamente, los datos empleados para construir las pistas que visualizamos en el navegador están almacenadas dentro del servidor de UCSC en forma de tablas. En este contexto, una tabla es un fichero de texto asociado a una pista cuya información está estructurada en filas (anotaciones) y columnas (atributos de dichas anotaciones) que es gestionado por un *software* denominado administrador de base de datos. El navegador de tablas enmascara la complejidad técnica de esta implementación informática, facilitando el acceso a la información genómica mediante un interfaz gráfico amigable.

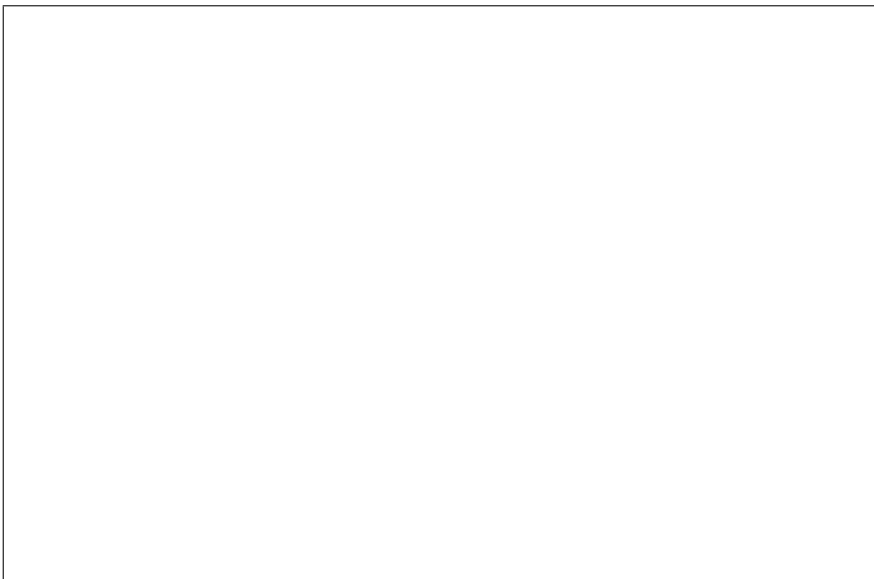


El estudiante puede acceder a la aplicación *Table browser* (en inglés, navegador de tablas), desde el enlace *Tools* (herramientas en inglés) situado dentro de la barra de menú superior de color azul. En la pantalla principal debemos elegir en primer lugar el genoma y la versión del ensamblado. A continuación seleccionamos la pista de trabajo que vamos a interrogar y el tipo de análisis que precisamos realizar (selección, comparación, etc.). Finalmente, una vez hemos configurado el formato de salida (*output format* en inglés), procederemos a ejecutar la acción presionando el botón *get output*. Los resultados pueden obtenerse en forma de fichero de texto o dar lugar a nuevas pistas que serán reintegradas al inventario actual de nuestra sesión de trabajo. A continuación trabajaremos con la tabla UCSC RefSeq genes perteneciente a la superpista NCBI RefSeq sobre la región del genoma donde se ubica nuestro gen de interés *LRRTM1*.

El usuario puede reubicar las pregunta utilizando nombres de elementos anotados en la pista también.

Durante cualquier análisis podemos presionar el botón *summary/statistics* para obtener en pantalla el número total de elementos que coinciden con el filtro actual.

Figura 49. Pantalla principal del navegador de tablas.



Si ejecutamos la pregunta (*query* en inglés) por defecto obtendremos el listado completo de genes anotados por RefSeq y sus isoformas con todas sus características. Aunque dicho listado pueda resultar obviamente interesante para ser analizado por otros programas informáticos, resulta conveniente seleccionar solamente alguno de los atributos de la pista para facilitar su lectura en pantalla cuando estamos visualizando directamente los datos. Para ello debemos cambiar el formato de salida de la opción *all fields from primary and related tables* a la opción *selected fields from primary and related table* (en inglés, mostrar todos los campos de las tablas o seleccionar únicamente algunos, respectivamente).

Los resultados son visualizados por pantalla por defecto, pero el usuario puede modificar este hecho para recuperar un archivo de texto tabulado con toda la información.

Figura 50. Opciones de formato de salida del navegador de tablas.

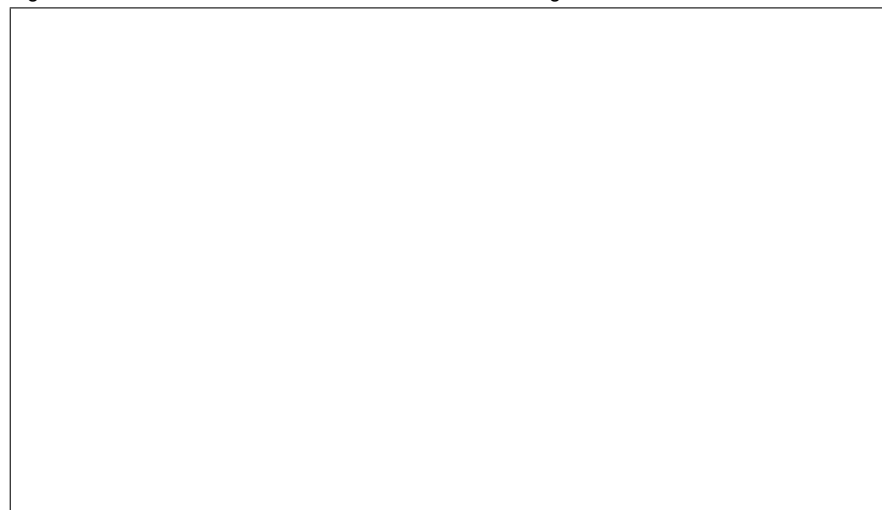


Figura 51. Listado filtrado de los transcritos anotados por RefSeq en la región del gen *LRRTM1* (a 1 de Septiembre de 2022).

#chrom	strand	txStart	txEnd	exonCount	name2
chr2	+	79512933	80648863	18	CTNNA2
chr2	+	79512933	80648863	17	CTNNA2
chr2	+	79513050	80648780	19	CTNNA2
chr2	+	79523182	80648863	18	CTNNA2
chr2	-	80301877	80304752	2	LRRTM1

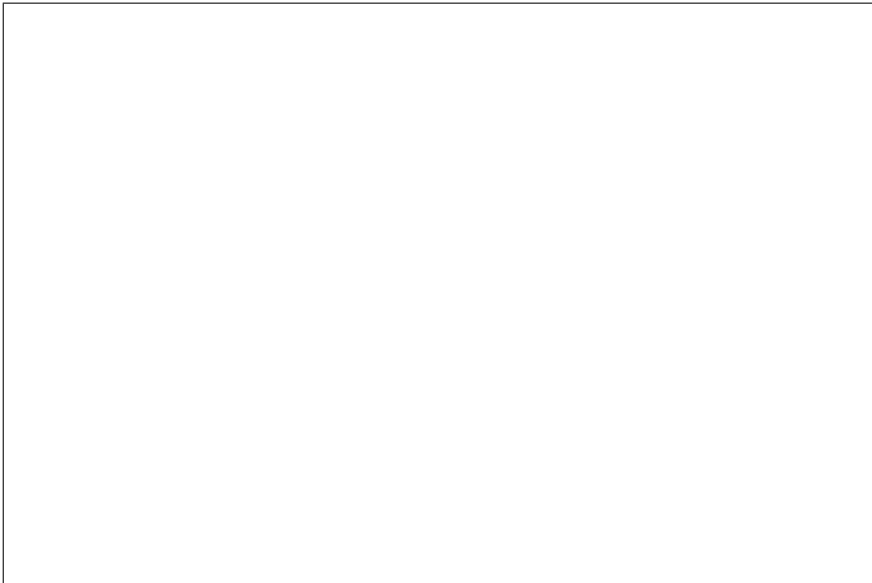
Además de la generación de listados de genes, el bioinformático puede emplear esta aplicación para descargar las secuencias de distintos elementos génicos para su posterior análisis computacional con otros programas informáticos. Por ejemplo, en el marco de la investigación sobre diferentes catálogos de genes, es posible extraer sistemáticamente la secuencia de uno o varios tipos de diferentes clases de regiones de estos (p.e. la región codificante para proteína de todos los genes o su posible región promotora de la transcripción). También podremos obtener las secuencias en formato FASTA agrupadas en un único fichero o separadas por gen o por tipo de característica para llevar a cabo un análisis bioinformático posterior.

Figura 52. Extracción sistemática de secuencias del navegador de tablas.



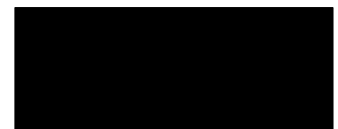
El mismo navegador de tablas permite ejecutar fácilmente múltiples acciones que involucren varias pistas (e.g. calcular intersecciones o correlaciones). Por ejemplo, podemos utilizar esta herramienta para calcular la correlación entre las anotaciones servidas por el proyecto RefSeq y por el consorcio GENCODE. El cálculo de solapamiento entre elementos se realizará posición a posición tanto para características discretas (p.e. exones) como valores continuos (p.e. dos réplicas del mismo experimento de secuenciación a gran escala).

Figura 53. Cálculo de correlaciones con el navegador de tablas.



1.8. El navegador genómico UCSC: distribución local de ficheros

UCSC sirve anotaciones para un vasto conjunto de genomas de distintas especies. Como hemos visto a lo largo de estos materiales el acceso interactivo al navegador gráfico ofrece multitud de posibilidades al investigador. No obstante, para ejecutar protocolos de análisis bioinformático a gran escala involucrando miles de genes resulta mas conveniente obtener una copia local de la distribución de los genomas estudiados. El enlace *Genome Data* en el menú superior del navegador permite el acceso a la página que contiene la colección completa de especies organizadas por grupo taxonómico (Figura 54). Una vez elegido un organismo, podremos descargar libremente en nuestro ordenador los mismos ficheros de texto que son precisamente interpretados por el interfaz gráfico de UCSC cuando trabajamos con el visor principal (secuencias y anotaciones). En cualquier distribución de un genoma en particular encontraremos básicamente la siguiente información (Figura 55): (i) el conjunto completo de cromosomas de esta distribución (*Genome sequence files*), (ii) el conjunto de datos de un cromosoma en particular (*Sequence data by chromosome*), (iii) las anotaciones de cada pista mostrada por el navegador gráfico en ficheros individuales (*Annotations*), (iv) los ficheros de conversión entre distintas versiones (*LiftOver files*) y (v) las comparaciones con otros genomas (*Pairwise alignments/Multiple alignments*). Cada enlace nos abre el acceso a un nuevo directorio para escoger los ficheros deseados.



This page contains links to sequence and annotation downloads for the genome assemblies featured in the UCSC Genome Browser. Downloads

are also available via the Genome Browser FTP server. For access to the most recent assembly of each genome, see the current genomes directory. To query and download data in JSON format, use our JSON API. To view descriptions of annotations, use the "describe table schema" button in the Table Browser. Previous versions of certain data are available from our track archive. For data hosted in Public Hubs the files exist on external sites, with GenArk (Genome Archive) Public Hub species found here.

All tables in the Genome Browser are freely usable for any purpose except as indicated in the README.txt files in the download directories. To view restrictions specific to a particular data set, click on the corresponding download link and review the README text. These data were contributed by many researchers, as listed on the Genome Browser credits page. Please acknowledge the contributor(s) of the data you use.

Human

Mouse

Mammals

Other vertebrates

Deuterostomes

Insects

Nematodes

Human genome Dec. 2013 (GRCh38/hg38)

Other genomes

Other downloads

- * Genome sequence files and select annotations (2bit, GTF, etc)
- * Sequence data by chromosome
- * Annotations
- * SNP-masked fasta files
- * LiftOver files
- * Pairwise alignments
- * Multiple alignments
- * Patches
- * Data archive

El conjunto de secuencias de nucleótidos de los cromosomas en cada genoma está empaquetado en un único archivo comprimido. En la Figura 56 podemos ver el contenido una vez descomprimido de la distribución hg38 del genoma humano, centrando nuestro interés en acceder a la secuencia del cromosoma 17. Para ello, hemos procedido a descargar y desempaquetar el fichero hg38.chromFa.tar.gz, visualizando después el fichero chr17.fa presentado en formato FASTA. La carpeta Annotations contiene todas las pistas utilizadas por el interfaz gráfico del navegador UCSC (mostramos una pequeña parte de éstas en la Figura 57). En particular, el fichero refGene.txt.gz que contiene las anotaciones de cada gen humano según el consorcio RefSeq, es procesado por UCSC para generar la pista gráfica RefSeq Genes para cada versión del ensamblado:

También están disponibles las secuencias enmascaradas de cada cromosoma, filtrando las regiones de baja complejidad.

Junto con la versión estándar de cada cromosoma es habitual encontrar también fragmentos de secuencias alternativas de cromosomas o sin ubicar exactamente todavía.

Figura 56. Descarga de los cromosomas humanos (hg38, a 1 de Septiembre de 2022).

```

hg38.chromFa.tar.gz          2014-01-23 17:18 938M

hg38.chromFa/
chr10.fa  chr14.fa  chr18.fa  chr21.fa  chr4.fa  chr8.fa  chrY.fa
chr11.fa  chr15.fa  chr19.fa  chr22.fa  chr5.fa  chr9.fa
chr12.fa  chr16.fa  chr1.fa  chr2.fa  chr6.fa  chrM.fa
chr13.fa  chr17.fa  chr20.fa  chr3.fa  chr7.fa  chrX.fa

AGTGAGGAGTAAACTGAAAAACAGGCCTGATTGTCCTTTTGTAGTGGA
AAATTGAGCTTTTATTAGAAGGCAGCCTGGACTGGTTGTGTACCTCCTA
ACCTCAGATGCCAGCTCACCAGATGAGTTTCACCGTGACCTAAGCCACAT
CCTTGGGATACCTCAGAGGTATTATTATTTTTTGTCAAAATGTTTTTAT
AGCTATTATAATGATTTTCTCTGATTGCAAAAATAAGTGCTTGACTAGAG
TGACTGAAAGCACACAGGCAAGAAATAAAACCACCCGTGATTTTACCACT
CACTGTGGTCATGTCAGTAGTTTTTCTAGCATCTGTCTTATTTTTATTA
TATTCTGTCCTCTGTTGAATTCCCTGCCTTTTGTTCCTCCAGACGGC
GTCTCACTCTGTGGCCAGGCTGGAGTACAGTGGCGCAATCTTGGCTCAC
TGCAGCCTCCACCTCCAGGTTTAGGCAATTCTCTGCTTCAGCCTCCTT
AGTAGCAGGGACTACAGGTGCACACTACCACACATGGCTAATTTTTGTAT
TTTTAGTAGAGAAGGGTTTTTACCATTGTTGGCCAGGCTGGTTTCGAACTC
CTGACCTCAAGTGATCCACCCACCTCGGCCTCCCAAAGTGCTAGGATTGT
AGGCGTGAGCCACCGTGCTGGGTCTGAATACCTGCTTTTGAAGCTTCG
TGTGGTATCATGAGCAGCCTCTGTGTGCGCGTGTGTTAGGAGTTGTGGCG
TGACGCCAGCCTGAATCATCCCTCAAGGACATCTGCAGAAGCAGCGTG
AATGTTCTCCCCATGCCCTGCCTGTGCCGTGTCTAGTTCAGGAAACACA

```

Figura 57. Descarga de anotaciones de genes de RefSeq (a 1 de Septiembre de 2022).

```

affyGnflh.txt.gz          2015-05-11 01:50 596K
...
refGene.sql               2020-08-17 18:56 1.9K
refGene.txt.gz            2020-08-17 18:56 8.3M
...
xenoRefSeqAli.sql         2020-08-17 19:17 2.1K
xenoRefSeqAli.txt.gz      2020-08-17 19:17 17M
1197 NM_178839 chr2 - 80301877 80304752 80302250 80303819 2
80301877,80304151, 80303878,80304752, 0 LRRTM1 cpl cpl 0,-1,

```

Cada línea del fichero `refGene.txt` contiene la anotación suministrada por RefSeq para un transcrito concreto. Las formas alternativas del mismo gen están codificadas en líneas distintas. Recomendamos precaución para interpretar las anotaciones de genes codificados en la hebra negativa de la molécula de ADN. En estos casos, las coordenadas iniciales y finales de las anotaciones deben intercambiarse para tener en cuenta este hecho (el campo número cinco contendría el final del gen y el campo número seis el inicio). Una conversión similar debería realizarse con las coordenadas de cada exón individualmente. El terminal de la plataforma LINUX es el método más eficiente para analizar distintas pistas de anotaciones de forma sistemática. A continuación mostramos la descripción de los atributos pertenecientes a cada registro de este fichero (tomando como ejemplo la línea de la anotación correspondiente al gen *LRRTM1* mostrada en la anterior Figura 57).

Tabla 7. Atributos del gen *LRRTM1* en `refGene.txt` (a 1 de Septiembre de 2022).

1	1197	Identificador interno
2	NM_178839	Código asignado por RefSeq
3	chr2	Cromosoma
4	–	Hebra
5	80301877	Inicio del transcrito
6	80304752	Fin del transcrito
7	80302250	Inicio de la región codificante
8	80303819	Fin de de la región codificante
9	2	Número total de exones
10	80301877, 80304151,	Coordenadas iniciales de los exones
11	80303878, 80304752,	Coordenadas finales de los exones
12	0	Puntuación
13	LRRTM1	Nombre común abreviado del gen
14	cmp1	Anotación del inicio del CDS (completa)
15	cmp1	Anotación del fin del CDS (completa)
16	0, -1,	Pauta de lectura de los exones

1.9. El navegador genómico ENSEMBL

ENSEMBL es otro navegador ampliamente extendido que está gestionado por el Instituto Europeo de Bioinformática (EBI) en Hinxton (Reino Unido). ENSEMBL también sirve las anotaciones en forma de pistas que pueden mostrarse u ocultarse dentro del visor gráfico. El usuario puede cargar sus propias pistas en ENSEMBL para integrarlas junto con la información convencional. La plataforma ENSEMBL produce también desde sus orígenes su propia anotación génica de referencia, colaborando intensamente con varios consorcios para la curación manual de estos datos. Como se aprecia en la Figura 58, para acceder a la anotación de un gen el usuario debe seleccionar primero la distribución apropiada del genoma para introducir posteriormente su nombre en la caja de búsqueda.

Figura 58. Accediendo a las anotaciones en ENSEMBL: pantalla principal.

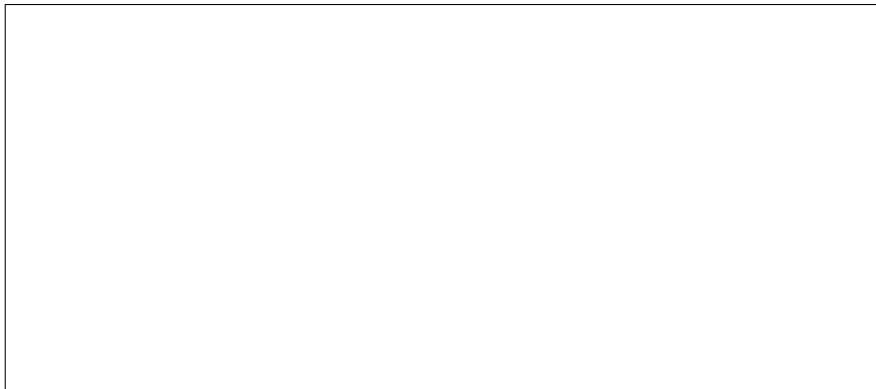
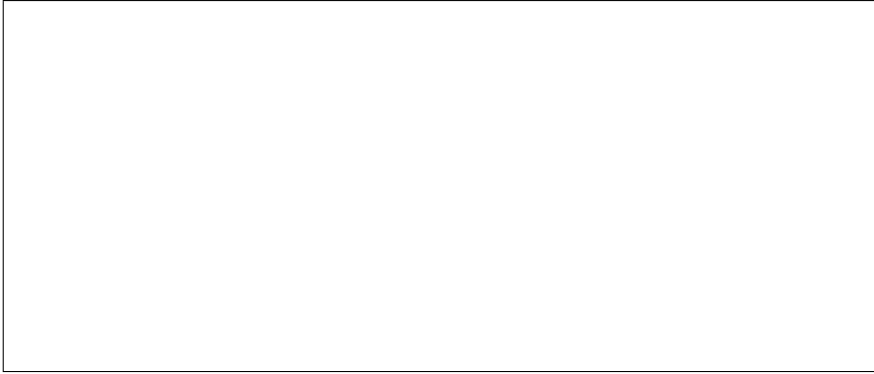


Figura 59. Accediendo a las anotaciones en ENSEMBL: versión hg38.



ENSEMBL fue pionero en presentar la información sobre los genes y los transcritos de forma estructurada. Los genes, como unidades que agrupan múltiples ARN mensajeros, reciben un identificador con las iniciales *ENSG* (*ENSEMBL gene*), mientras que los diferentes productos de éste, transcritos y proteínas, poseen sus propios códigos *ENST* y *ENSP*, respectivamente. Es posible acceder a toda la información recopilada sobre cada gen o visualizar la secuencia de éste en pantalla (Figura 60). ENSEMBL reconstruye dinámicamente el visor genómico en el momento de la actualización. A medida que desplazamos el puntero de nuestro ratón sobre ciertas secciones de la imagen, aparece un resumen de la información asociada a cada pista (incluyendo enlaces hacia otras bases de datos). ENSEMBL genera igualmente imágenes para impresión en formato de alta calidad.

Tanto UCSC como ENSEMBL sirven sus propias anotaciones junto con un conjunto de pistas comunes proporcionadas por otros recursos. La anotación génica de cada navegador puede encontrarse como una pista más de datos en el otro programa.

Para modificar el inventario de pistas el usuario debe presionar el botón *Configure this page* (en inglés, configurar esta página).

Figura 60. Anotación del gen *LRRTM1* en ENSEMBL (a 1 de Septiembre de 2022).

BIOMART es una de las utilidades más potentes de ENSEMBL. Esta aplicación combina registros de numerosas fuentes con el objetivo de extraer nuevo conocimiento sobre los genes y otras pistas de anotaciones almacenadas en su base de datos. Como se aprecia en la Figura 61, la pantalla de BIOMART está dividida principalmente en dos áreas claramente diferenciadas: el menú (a la izquierda) y la zona de trabajo (ocupando la mayoría de la pantalla). El usuario debe en primer lugar establecer un genoma de referencia para trabajar sobre éste. La unidad de trabajo en BIOMART es el conjunto de genes de dicho genoma (denominado *Dataset* en el menú, Figura 61).

Dado que las acciones que podemos hacer empleando BIOMART o el navegador de tablas de UCSC son ligeramente diferentes, recomendamos que el estudiante realice prácticas intensivas con ambos.

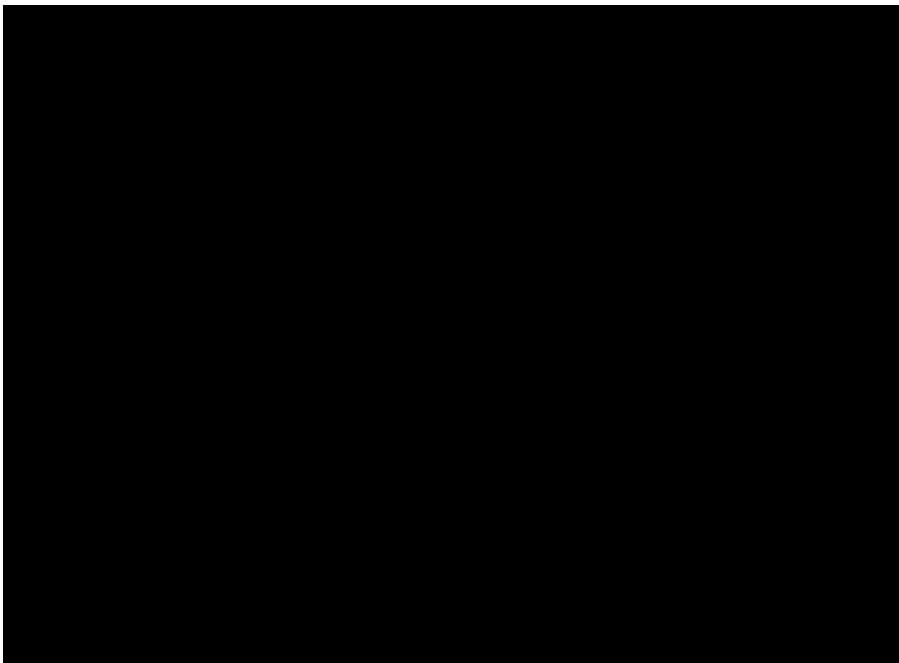
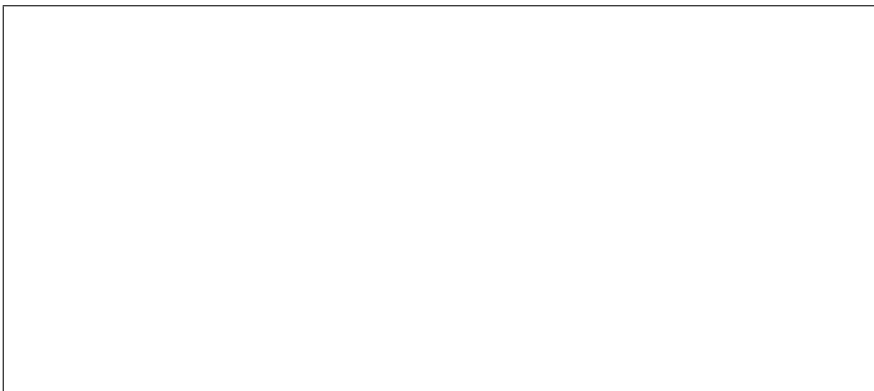
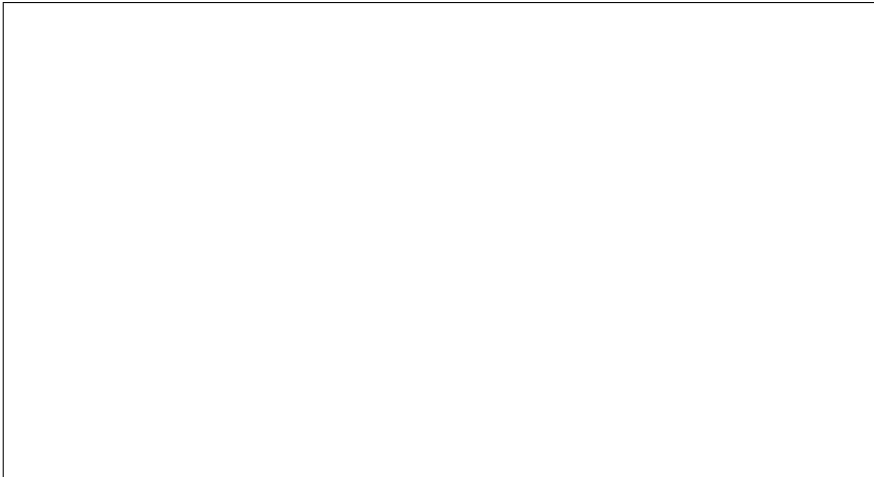


Figura 61. El interfaz de trabajo de BIOMART: opciones de filtros.



Inicialmente, podremos realizar operaciones sobre todos los transcritos anotados por ENSEMBL en el genoma seleccionado (68324 ARN mensajeros en la distribución GRCh38 del genoma humano, a 1 de Septiembre de 2022). Para actualizar la cifra total de genes de trabajo, debemos presionar el botón *Count* tras definir un nuevo filtro sobre los datos. Cuando hemos finalizado con el filtrado, mantenemos en el área de trabajo precisamente el grupo de genes que cumplen con todos los criterios especificados. En ese instante, justo antes de generar los resultados, es necesario realizar la selección de los atributos. Una vez terminados los procesos de filtrado y formateado de los resultados, debemos utilizar el botón *Results* para recibir el archivo final.

Figura 62. El interfaz de trabajo de BIOMART: opciones de atributos.



BIOMART exhibe siempre una vista preliminar más reducida de la información solicitada. De este modo, el usuario puede introducir cambios en las preguntas realizadas sobre el catálogo de genes sin necesidad de descargar completamente en cada ocasión el fichero final de resultados (Figura 61).

Figura 63. El interfaz de trabajo de BIOMART: previsualización de resultados.



Los usuarios de ENSEMBL tienen a su disposición numerosos tutoriales, páginas de ayuda e incluso películas con ejemplos ilustrativos (con su propio portal de acceso en YouTube). Para cada función también existe específicamente un manual de asistencia propio.

1.10. Portales genómicos del cáncer

- La secuencia del genoma de un individuo posee diferencias con el genoma de otro organismo de la misma especie. - Las variantes genómicas pueden ser puntuales SNP o en bloque CNV - Mientras los SNPs pueden coincidir en regiones funcionales con poca probabilidad y su potencial consecuencia se estudia por estudios de asociación (GWAS) - Cada vez más se asume que los CNVs conteniendo genes completos implicados en determinadas funciones pueden ser responsables del cáncer (Referencias *Frontiers Genetics* 13 May 2021 31 cancer subtypes). Por ejemplo, aumento de copias de CNVs con oncogenes o disminución de CNVs de genes supresores de tumores

Captura de pantalla de UCSC con SNPs y CNVs Captura de pantalla en Ensembl (tabla)?

Caja apuntando a la asignatura Fundamentos de Biología Molecular

TEXTB: En paralelo con la identificación de diferencias entre individuos, se está prestando más relevancia al sesgo en el tipo de población humana para la que existen estudios genómicos. Referencia Nature 2022 Feb

Todo esto hace que además de los navegadores genómicos que suministran anotaciones sobre la secuencia modelo de cada organismo, existan navegadores específicos para recuperar información en términos de variabilidad genómica y en el contexto del cáncer.

TCGA DepMap

Resumen

Durante este capítulo hemos introducido los conceptos fundamentales sobre la navegación genómica. Trabajando con el servidor genómico UCSC hemos aprendido a configurar el programa, manipular las pistas de anotaciones e integrar nuestros propios datos para ser visualizados dentro del mismo entorno gráfico. El estudiante también se ha familiarizado con los diferentes ficheros que constituyen la base de cualquier distribución genómica. Finalmente, como alternativa a los sistemas de análisis basados en la plataforma LINUX, hemos mostrado como manipular estas informaciones fácilmente para extraer nuevo conocimiento, utilizando herramientas *web* como el navegador de tablas de UCSC o BIOMART.

Actividades

1. Busca en la Red la revista científica *Nucleic Acids Research*. En su interior, explora el contenido las últimas ediciones especiales sobre bases de datos y servidores *web*.
2. Selecciona un gen humano cualquiera con el navegador UCSC. Después, abre el navegador ENSEMBL para visualizar el mismo gen. Ahora, explora en paralelo las anotaciones disponibles para este gen en ambos servidores.
3. La herramienta *Liftover* de UCSC es extremadamente útil para los bioinformáticos. Explora qué operaciones realiza, cómo funciona y prueba su eficacia sobre la anotación concreta de un gen humano cuyas coordenadas hayan variado entre dos versiones de este genoma.
4. Aprende a crear una pista propia en formato BED para mostrar una caja artificial de 1000 nucleótidos de longitud en color rojo, concretamente en la posición 1000000 del cromosoma 21 del genoma humano (versión hg19). Ahora crea una pista WIG en las mismas coordenadas para representar un elemento artificial que posee un valor de diez unidades durante los primeros 300 nucleótidos, un valor de veinte unidades en los siguientes 400 nucleótidos y un valor de diez unidades nuevamente en los últimos 300 nucleótidos.
5. Activa las pistas de conservación entre vertebrados, elige un gen humano constituido por cinco o más exones y estudia el patrón de conservación filogenética en las diferentes regiones de este gen (exones, intrones, regiones codificantes, regiones no traducibles).
6. Dentro de la versión de UCSC para servir datos del proyecto ENCODE, activa una superpista que contenga múltiples modificaciones de histonas y visita este paisaje genómico a lo largo de un cromosoma humano concreto.
7. Inspecciona el navegador de modENCODE e intenta reproducir la Figura ??.
8. Con UCSC, realiza la descarga de la última distribución del genoma humano y crea un DVD que contenga la secuencia de sus cromosomas y las anotaciones más relevantes.
9. En el marco de la última distribución del genoma de la mosca de la fruta, abre el navegador de tablas de UCSC y calcula la correlación existente entre las anotaciones servidas por los consorcios FlyBase y RefSeq. Visualiza casos concretos de anotaciones divergentes y valora los resultados obtenidos.
10. Utiliza el navegador de tablas de UCSC para extraer las regiones promotoras de todos los genes humanos (longitud: 500 bps). Emplea el programa BLAT sobre un gen en cada hebra de ADN para validar que la descarga haya funcionado correctamente.
11. Utiliza la herramienta BIOMART de ENSEMBL para recuperar los identificadores de aquellos genes cuya forma ortóloga entre humano y ratón está documentada.

Ejercicios de autoevaluación

1. Enumera qué dos tipos de información definen cada distribución de un genoma.
2. Define cuál es la principal misión de un servidor genómico.
3. Enumera qué informaciones pueden emplearse para localizar una región del genoma.
4. ¿ Qué es una pista de datos en el contexto de un navegador genómico ?
5. Describe el procedimiento habitual para acceder a una región genómica con un servidor de genomas como UCSC o ENSEMBL.
6. Enumera las opciones básicas para configurar la visualización de un navegador.
7. ¿ Qué opciones posee habitualmente el usuario para mostrar una pista ?
8. Enumera cinco bloques de opciones en los que se agrupan las pistas de datos en el navegador genómico UCSC.
9. Define qué es el recurso RefSeq.
10. ¿ Gráficamente, cómo se representan los exones u otros elementos genómicos anotados sobre una pista en cualquier navegador genómico ?
11. Describe el contenido de un fichero FASTA.
12. ¿Cuál es la función del programa BLAT ?
13. Enumera qué informaciones necesitamos obtener para crear nuestra propia pista.
14. Explica la principal diferencia entre los formatos BED y WIG.
15. Define el concepto de superpista (existente en el navegador UCSC para ENCODE).
16. Define las funciones principales del navegador de tablas de UCSC.
17. BIOMART: define el protocolo de uso de esta herramienta para extraer anotaciones.

Solucionario

1. La secuencia del genoma y las anotaciones a lo largo de los cromosomas de éste.
2. Un servidor genómico proporciona herramientas para navegar a través de las anotaciones de los genomas.
3. Las coordenadas dentro de un determinado cromosoma, el nombre de un gen, el nombre de una proteína, una secuencia similar.
4. Una pista contiene una serie de anotaciones sobre cierto elemento biológico ubicado en determinadas regiones del genoma.
5. Primero, introducimos una determinada palabra clave que identifique nuestro objeto de búsqueda. Segundo, el servidor nos muestra una lista de posibles coincidencias en varias pistas de datos. Tercero, escogemos una pista para que el servidor construya una representación gráfica de la región seleccionada. Cuarto, utilizamos los botones de desplazamiento y diferentes enfoques para configurar esta vista. Quinto y último, visitamos los detalles de las anotaciones pinchando con el ratón sobre las pistas.
6. Podemos desplazarnos por una región, acercarnos o alejarnos a ésta o invertir la secuencia junto con sus anotaciones en la otra hebra de la molécula de ADN.
7. El conjunto de anotaciones proporcionadas por una pista de datos puede mostrarse compactado en una única línea o desplegarse en varias líneas del visor.
8. Los bloques de opciones más comunes son *Mapping and Sequencing Tracks*, *Phenotype and Disease Associations*, *Genes and Gene Prediction Tracks*, *mRNA and EST Tracks*, *Expression, Regulation, Comparative Genomics, Variation and Repeats*.
9. RefSeq es un consorcio público de anotación de productos génicos que fundamentalmente produce anotaciones de calidad evitando información redundante o errónea.
10. Los exones son representados en forma de cajas, mientras que los intrones se indican con una línea que une dos exones contiguos. Otras anotaciones pueden mostrarse también de este modo o exhibiendo una cierta distribución continua de valores a lo largo del genoma.
11. Una secuencia FASTA posee dos tipos de información: (a) la cabecera, que posee el símbolo ">", contiene información para clasificar el fichero; (b) la secuencia, a continuación, en forma de líneas de longitud fija (generalmente, 60 caracteres).
12. El programa BLAT busca secuencias prácticamente idénticas a una suministrada por el usuario, generalmente con el objetivo de identificar su ubicación exacta en el genoma.
13. Para crear una *custom track* recomendamos dotar a la pista de (a) instrucciones para el navegador sobre la ubicación de la región anotada, (b) instrucciones sobre la configuración de la visualización de la pista y (c) las coordenadas de nuestras anotaciones en el genoma.

14. El formato BED (*Browser Extensible Data*) es útil para representar anotaciones en el genoma que poseen un inicio y final concretos (e.g. exones). El formato WIG (*WIGgle*), por contra, es más adecuado para mostrar anotaciones continuas de una cierta característica del genoma.

15. Una superpista realiza una superposición gráfica de varias pistas de datos (generalmente en formato WIG), integrándolas en una única línea del visor genómico.

16. El navegador de tablas permite comparar los datos de diferentes pistas para calcular intersecciones, correlaciones o extraer información adicional a partir de éstas.

17. Con BIOMART podemos aplicar sucesivamente varios tipos de operaciones (filtrar, seleccionar atributos y combinar con otros conjuntos de datos) para delimitar claramente la fracción de las anotaciones que deseamos recuperar.

Bibliografía

- J.F. Abril y R. Guigó** (2000). *gff2ps: visualizing genomic annotations*. Bioinformatics 8:743-744.
- A. Barski et al.** (2007). *High-resolution profiling of histone methylations in the human genome*. Cell 129:823-37.
- D.A. Benson et al.** (2008). *GenBank* Nucleic Acids Research 36:D25-30.
- S.L. Berger** (2007). *The complex language of chromatin regulation during transcription*. Nature 447:407-12.
- S.E. Celniker et al.** (2009). *Unlocking the secrets of the genome*. Nature 459:927-930.
- M.S. Cline y W.J. Kent** (2009). *Understanding genome browsing*. Nature Biotechnology 27:153-155.
- P. Flicek et al.** (2010). *Ensembl's 10th year* Nucleic Acids Research 2010 38:D557-D562.
- J. Harrow et al.** (2006). *GENCODE: producing a reference annotation for ENCODE*. Genome Biology Suppl 1:S4.1-9.
- T. Hubbard et al.** (2002). *The Ensembl genome database project*. Nucleic Acids Research 2002 30:38-41.
- W.J. Kent et al.** (2002). *The human genome browser at UCSC*. Genome Research 12:996-1006.
- W.J. Kent** (2002). *BLAT: the BLAST-like alignment tool*. Genome Research 12:656-664.
- C. Mayor et al.** (2000). *VISTA : visualizing global DNA sequence alignments of arbitrary length*. Bioinformatics 16:1046-1047.
- Mouse Genome Database Group** (2008). *The Mouse Genome Database (MGD): mouse biology and model systems*. Nucleic Acids Res 36:D724-728.
- C.A. Ouzounis y A. Valencia** (2003). *Early bioinformatics: the birth of a discipline - a personal view*. Bioinformatics 19:2176-2190.
- K.R. Rosenbloom et al.** (2010). *ENCODE whole-genome data in the UCSC Genome Browser*. Nucleic Acids Research 38:D620-5.
- L.D. Stein et al.** (2002). *The generic genome browser: a building block for a model organism system database*. Genome Research 12:1599-610.
- The ENCODE Project Consortium** (2007). *Identification and analysis of functional elements in 1 % of the human genome by the ENCODE pilot project*. Nature 447:799-816.
- The FlyBase Consortium** (2009). *FlyBase: enhancing Drosophila Gene Ontology annotations*. Nucleic Acids Research 37:D555-D559.
- The modENCODE Consortium** (2010). *Identification of functional elements and regulatory circuits by Drosophila modENCODE*. Science 330:1787-1797.
- D.L. Wheeler et al.** (2004). *Database resources of the NCBI: update*. Nucleic Acids Research 32:D35-D40.