

Servidores genómicos

Enrique Blanco García

2 créditos ECTS
Enrique Blanco García
Módulo 1

Índice

Introducción	3
Objetivos	4
1. Servidores genómicos	5
1.1. Introducción	5
1.2. Filosofía de la navegación genómica	8
1.3. El navegador genómico UCSC: estructura	12
1.4. El navegador genómico UCSC: pistas	17
1.5. El navegador genómico UCSC: anotaciones propias	29
1.6. El navegador genómico UCSC: ENCODE	34
1.7. El navegador genómico UCSC: instalación local	37
1.8. El navegador genómico ENSEMBL	42
Resumen	46
Actividades	47
Ejercicios de autoevaluación	48
Solucionario	49
Bibliografía	51

Introducción

La secuenciación de los genomas de múltiples especies 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. El acceso universal a toda esta información está transformando radicalmente la investigación en biología molecular y biomedicina. En consecuencia, en el camino para explicar numerosos enigmas hasta ahora sin resolver, el enfoque científico clásico está derivando en nuevas aproximaciones más pragmáticas. Dicho de otro modo, en lugar de orientar el foco de investigación hacia un gen en particular, ahora podemos trabajar a gran escala, analizando simultáneamente todos los genes involucrados en un proceso biológico concreto. El progreso exponencial experimentado por la ciencia en la última década 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. 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.

Mediante una serie de convenios razonablemente establecidos, cualquier bioinformático puede obtener en pocos minutos el cartografiado completo del genoma para realizar nuevas contribuciones. 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 continuamente actualizado, gracias a la incesante actividad científica que intenta caracterizar con mayor precisión cada escenario biológico que ocurre dentro del entorno celular. 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. Numerosos servidores 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. 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 recursos genómicos de notable relevancia, como el portal ENSEMBL, diseñado en el Instituto Europeo de Bioinformática (EBI) en Hinxton (Reino Unido).

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 servidores genómicos existentes:

1. Conocer el elenco de bancos de datos genómicos existentes.
2. Analizar 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. Acceder a las anotaciones del proyecto internacional ENCODE.

1. Servidores genómicos

1.1. Introducción

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. Junto con éstos, cohabitan en el genoma otros elementos que regulan la activación de los genes, proporcionando 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 debe transformarse, 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, sitios de unión de factores de transcripción, inicios de transcripción, marcas de modificación de histonas, regiones repetitivas y polimorfismos.

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 cuál estamos empleando en nuestros análisis bioinformáticos (la Tabla 1 muestra las distribuciones genómicas más recientes de varios organismos modelo).

Una descripción detallada de los genomas y los elementos funcionales más comunes codificados en su interior puede encontrarse en el módulo Fundamentos de biología molecular.

En los materiales del módulo Fundamentos de informática en entornos bioinformáticos el estudiante puede encontrar varios ejemplos de anotación computacional de regiones genómicas.

Un listado exhaustivo de los genomas secuenciados en la actualidad puede encontrarse en el módulo Fundamentos de biología molecular.

Una distribución concreta de un genoma viene definida por:

- La secuencia de nucleótidos de cada cromosoma determinada por una versión concreta del ensamblado producido por el consorcio correspondiente.
- La serie de anotaciones pertenecientes a distintas características biológicas cartografiadas dentro la secuencia de los cromosomas por diferentes entidades de investigación.

Tabla 1. Distribuciones actualizadas de los genomas más utilizados

Genoma	Ensamblado	Anotaciones	Fecha
<i>Homo sapiens</i>	GRCh37	hg19	Febrero 2009
<i>Pan troglodytes</i>	CGSC 2.1	panTro2	Marzo 2006
<i>Bos taurus</i>	Baylor 4.0	bosTau4	Octubre 2007
<i>Mus musculus</i>	NCBI37	mm9	Julio 2007
<i>Rattus norvegicus</i>	Baylor 3.4	rn4	Noviembre 2004
<i>Gallus gallus</i>	WUGSC 2.1	galGal3	Mayo 2006
<i>Danio rerio</i>	Zv8	danRer6	Diciembre 2008
<i>Fugu rubripes</i>	JGI 4.0	fr2	Octubre 2004
<i>Drosophila melanogaster</i>	BDGP R5	dm3	Abril 2006
<i>Anopheles gambiae</i>	IAGC MOZ2	anoGam1	Febrero 2003
<i>Caenorhabditis elegans</i>	WS190	ce6	Mayo 2008
<i>Saccharomyces cerevisiae</i>	SGD	sacCer2	Junio 2008

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 servidor 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 servidores genómicos proporcionan 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. Estos archivos suelen distribuirse también de forma separada para realizar tratamientos bioinformáticos de forma local.

Un servidor genómico es una aplicación informática que proporciona herramientas para navegar eficientemente a lo largo de la secuencia de los genomas, permitiendo el acceso a las anotaciones cartografiadas en su interior.

Dada su enorme versatilidad, existen varios puntos de entrada en un servidor 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 servidor 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 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.

Lectura complementaria

M.S. Cline y W.J. Kent (2009). *Understanding genome browsing*. Nature Biotechnology 27:153-155.

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 entrada en un servidor genómico

Clase	Formato	Ejemplo
Región	Coordenadas	chr2:80529003-80531487
Gen	Abreviatura	<i>LRRTM1</i>
Gen	Nombre completo	<i>leucine-rich repeat transmembrane neuronal</i>
Gen	Código RefSeq	NM_178839
Proteína	Identificador	LRRTM1
Tránsito	Identificador	BC045113
Secuencia	FASTA	ATGGATTTCCTGCTGCTCGGTC . . .

Una pista de información contiene la colección de elementos biológicos de una determinada clase anotados en un servidor genómico para una región específica del genoma. Cada anotación está definida generalmente por las coordenadas exactas necesarias para posicionar dicho elemento en el interior de la secuencia y un valor numérico asociado al grado de representatividad de éste.

El servidor 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 eficiente 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

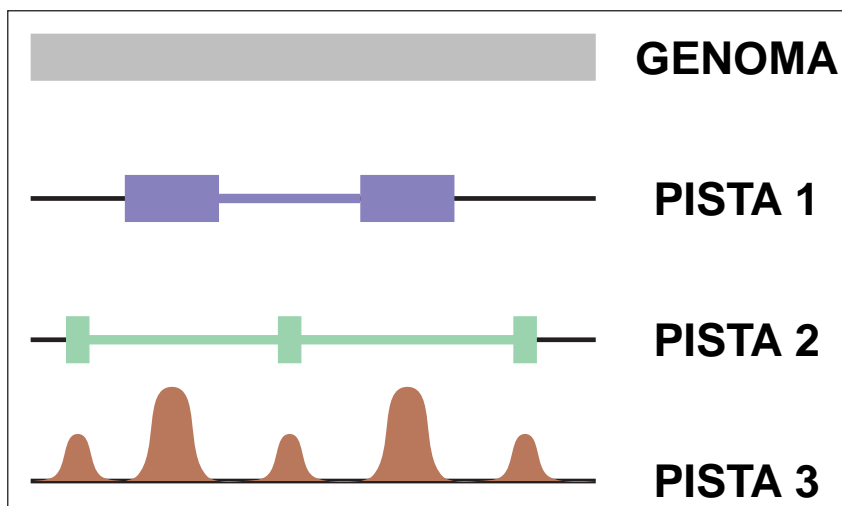


Figura 1

Superposición de varios niveles de información en una misma región genómica (genes, sitios de unión a factores de transcripción y marcas de metilación de histonas, respectivamente).

En tiempo real, el servidor 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. Es posible también 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.

Los ficheros de anotaciones pueden administrarse con gestores de bases de datos, ver módulo Fundamentos de informática en entornos bioinformáticos.

Los bloques de opciones disponibles pueden cambiar en función de la distribución del genoma habilitada.

Figura 2. Mostrando diversos niveles de información con pistas

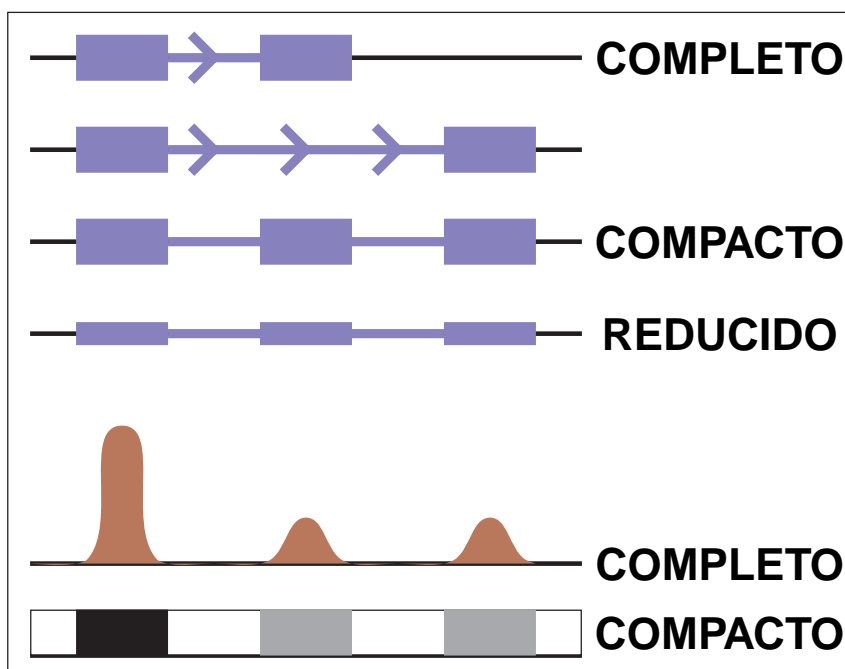


Figura 2

Mostramos dos clases de pistas. En el primer bloque visualizamos un gen con dos isoformas, agrupando esta información de tres formas posibles. En el segundo bloque, resumimos el perfil de una determinada modificación de histonas asociada al gen anterior con un gráfico de densidad más compacto.

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 servidores 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. 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 servidores 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 actualmente reciben un tratamiento de validación mucho más cuidadoso, basado en la verificación manual llevada a cabo por expertos entrenados en este tipo de tareas. Hemos realizado para el alumno una selección de los recursos primarios más populares en la Tabla 3. A lo largo de estos materiales analizaremos detalladamente el contenido de estos repositorios de información genómica, cuyo acceso está integrado dentro de cualquiera de los servidores genómicos existentes.

Lectura complementaria

C.A. Ouzounis y A. Valencia
(2003). *Early bioinformatics: the birth of a discipline - a personal view*. *Bioinformatics* 19:2176-2190.

Nucleic Acids Research

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

Figura 3. Paradigma de la comunicación con navegadores genómicos

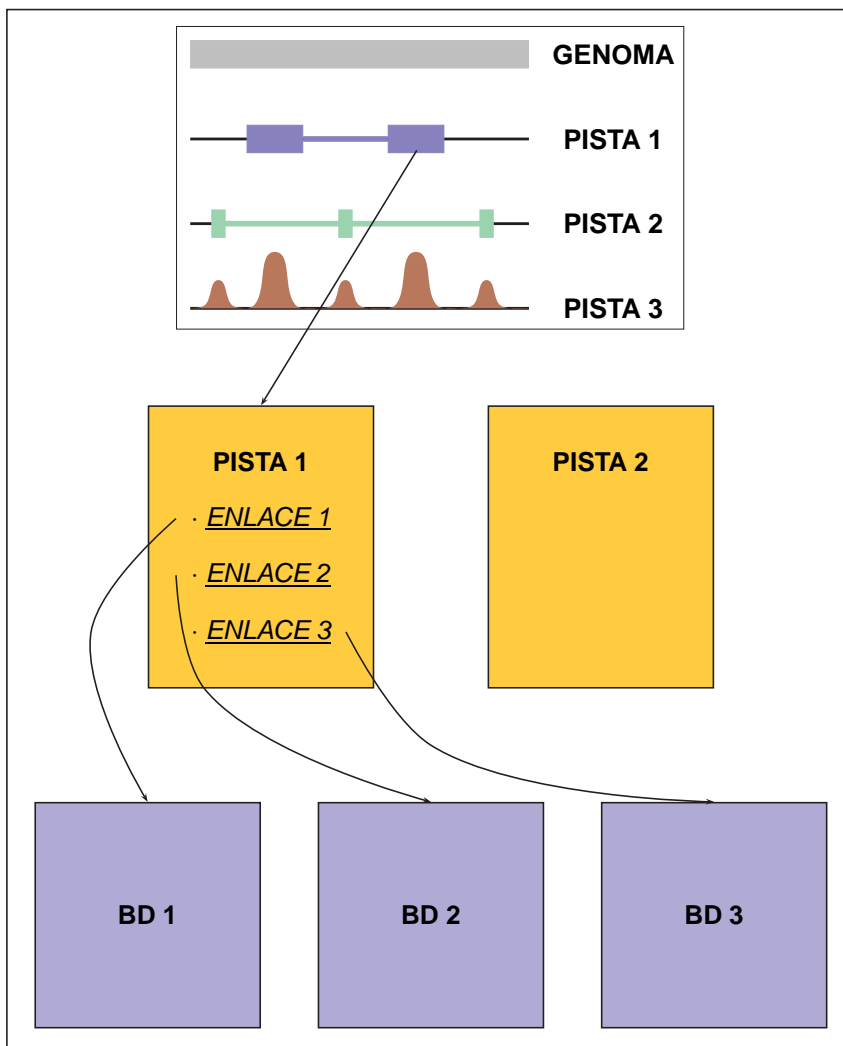


Figura 3

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.

Tabla 3. Listado de recursos genómicos esenciales

Recurso primario	Dirección
GenBank	http://www.ncbi.nlm.nih.gov/genbank
RefSeq	http://www.ncbi.nlm.nih.gov/refseq
Unigene	http://www.ncbi.nlm.nih.gov/unigene
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
Transfac	http://www.gene-regulation.com
Jaspar	http://jaspar.genereg.net
ORegAnno	http://www.oreganno.org
Gene Expression Omnibus	http://www.ncbi.nlm.nih.gov/geo
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP
OMIM	http://www.ncbi.nlm.nih.gov/omim
Pubmed	http://www.ncbi.nlm.nih.gov/pubmed
Navegador genómico	Dirección
UCSC	http://genome.ucsc.edu
ENSEMBL	http://www.ensembl.org
NCBI	http://www.ncbi.nlm.nih.gov/sites/genome
Vertebrate Genome Annotation	http://vega.sanger.ac.uk
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
Arabidopsis Information Resource	http://www.arabidopsis.org
Rice Genome Annotation Project	http://rice.plantbiology.msu.edu
International Wheat Genome Sequencing Consortium	http://www.wheatgenome.org
E.coli database	http://www.genome.wisc.edu
Gbrowse	http://gmod.org/wiki/Gbrowse
Navegador genómico avanzado	Dirección
VISTA	http://pipeline.lbl.gov
rVISTA	http://rvista.dcode.org
enhancerVISTA	http://enhancer.lbl.gov
modENCODE	http://www.modencode.org
International Cancer Genome Consortium	http://www.icgc.org
International HapMap Project	http://www.hapmap.org
Human Epigenome Consortium	http://www.epigenome.org

El analista bioinformático dispone de un amplio abanico de servidores genómicos (ver Tabla 3). Los servidores 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. Todos estas herramientas suelen 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. 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 servidores genéricos, no obstante, comparten con éstos las mismas distribuciones de los genomas (ver Figura 4).

Figura 4. Visualización de una región genómica

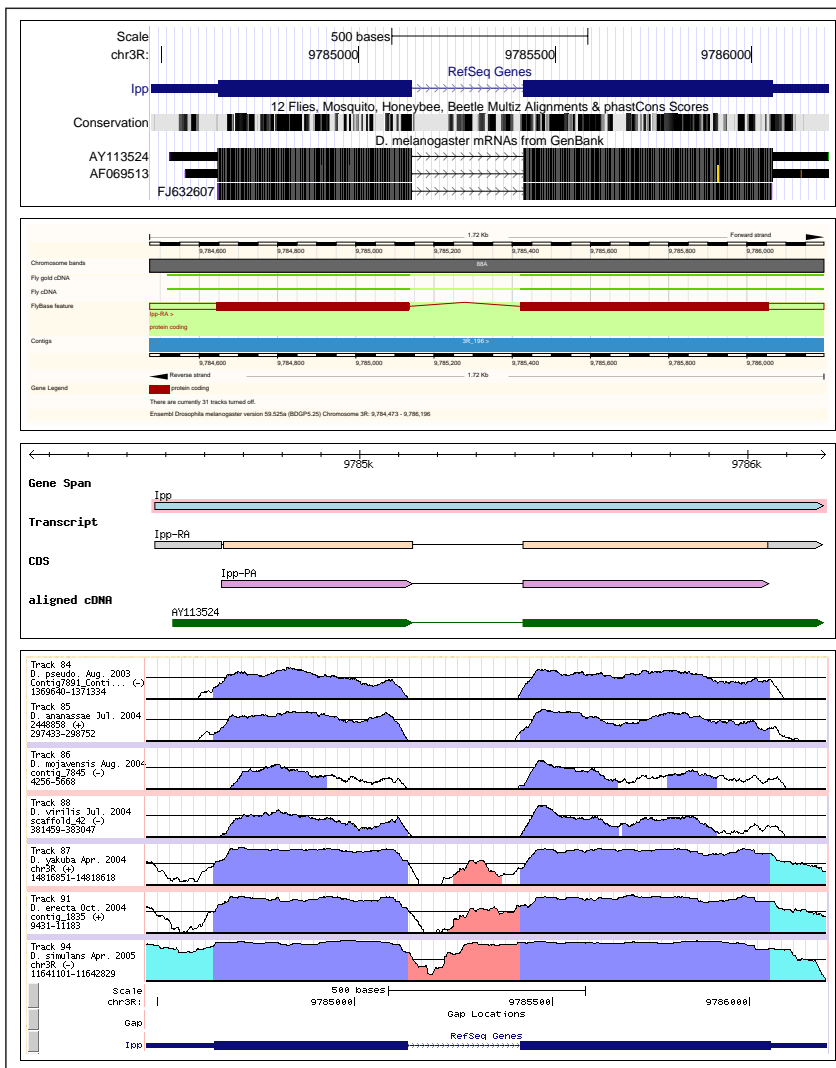


Figura 4

El gen *lpp* visualizado a través de distintos navegadores genómicos. Desde el margen superior de la imagen hacia el inferior: UCSC, ENSEMBL, FlyBase y VISTA.

Lecturas complementarias

W.J. Kent et al. (2002). *The human genome browser at UCSC*. Genome Research 12:996-1006.

P. Flicek et al. (2010). *Ensembl's 10th year*. Nucleic Acids Research 2010 38:D557-D562.

D.A. Benson et al. (2008). *GenBank*. Nucleic Acids Research 36:D25-30.

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 FlyBase Consortium (2009). *FlyBase: enhancing Drosophila Gene Ontology annotations*. Nucleic Acids Research 37:D555-D559.

Mouse Genome Database Group (2008). *The Mouse Genome Database (MGD): mouse biology and model systems*. Nucleic Acids Res 36:D724-728.

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. 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 que proporcionen un mapeo completo de las anotaciones referentes a estos desórdenes hereditarios. Ejemplos de estos portales biomédicos son el proyecto de secuenciación del cáncer 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 servidor 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 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 en su implementación inicial, convirtiéndose en un recurso de referencia utilizado, por ejemplo, para distribuir los resultados de la primera fase del proyecto ENCODE. 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.

El concepto de genómica comparada se trata con profundidad en el módulo Fundamentos de biología molecular.

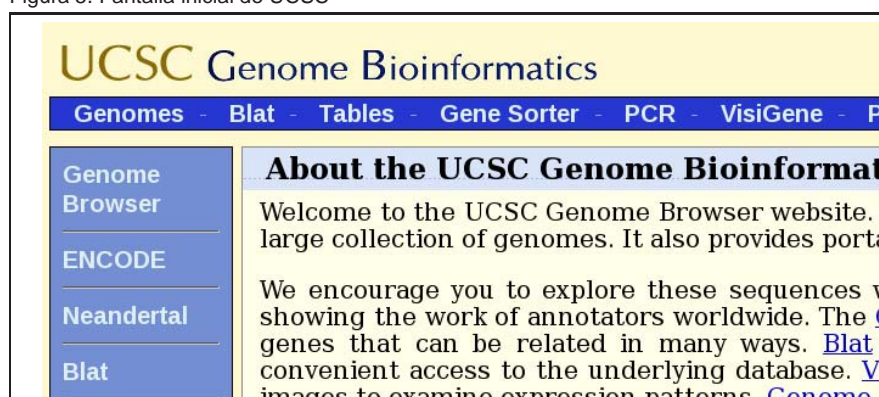
Lectura complementaria

C. Mayor et al. (2000). *VISTA: visualizing global DNA sequence alignments of arbitrary length*. *Bioinformatics* 16:1046-1047.

Lectura complementaria

W.J. Kent et al. (2002). *The human genome browser at UCSC*. *Genome Research* 12:996-1006.

Figura 5. Pantalla inicial de UCSC



Para iniciar la exploración debemos seleccionar primero el organismo y la versión del genoma que deseamos explorar. A continuación, es necesario introducir la información suficiente para que el navegador genómico 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*).

Figura 6. Parámetros de búsqueda en UCSC

clade	genome	assembly	
Mammal	Human	Feb. 2009 (GRCh37/hg19)	LRRTM1

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

Nombre	Ejemplo
Cromosoma	chr2
Región	chr2:80529002-80531487
Región	chr2:80529002+2485
Gen	<i>LRRTM1</i>
Tránsito	NM_178839
Proteína	NP_849161

Una vez acotados los parámetros de búsqueda, el sistema procesa estos datos y proporciona un listado de todas las pistas que posean algún nexo en común con la información suministrada. En este caso, encontramos siete pistas relacionadas con el gen *LRRTM1* en la distribución hg19 del genoma humano (a Noviembre de 2010). Para mostrar cómo interpretar esta información, procederemos a enumerar los distintos resultados obtenidos, analizando la ficha desde la parte superior de ésta (ver Figura 7):

1. Pista UCSC Genes: dos isoformas de este gen incluidas dentro de la anotación del genoma humano servida por el propio navegador UCSC.
2. Pista RefSeq Genes: una única anotación producida por el consorcio RefSeq.
3. Pista Non-human RefSeq Genes: cuatro formas ortólogas de este gen identificadas en las anotaciones servidas por RefSeq para otras especies.
4. Pista ENCODE Gencode: siete isoformas analizadas por el consorcio ENCODE.
5. Pista Human mRNA: Dos transcritos humanos obtenidos experimentalmente.
6. Pista Non-Human mRNA: Dos transcritos experimentales de otras especies.
7. Pista Vega: Siete formas alternativas reconocidas por el consorcio VEGA.

RefSeq

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.

Figura 7. Listado de pistas para el gen *LRRTM1* (resumen)

UCSC Genes	
LRRTM1 (uc002sok.1) at chr2:80529003-80531487	
- leucine rich repeat transmembrane neuronal 1	
LRRTM1 (uc002soj.3) at chr2:80515481-80531487	
- leucine rich repeat transmembrane neuronal 1	
RefSeq Genes	
LRRTM1 at chr2:80529003-80531487	
- (NM_178839) leucine-rich repeat transmembrane neuronal	
Non-Human RefSeq Genes	
LRRTM1 at chr2:80515605-80531552	
- (NM_001133111) leucine-rich repeat transmembrane neuronal	
...	
ENCODE Gencode Manual Gene Annotations (May 2010)	
LRRTM1 at chr2:80515483-80531518	
...	
Human Aligned mRNA Search Results	
BC045113 - Homo sapiens leucine rich repeat transmembrane neuronal 1, mRNA (cDNA clone MGC:47886 IMAGE:5200499), complete cds.	
...	
Non-Human Aligned mRNA Search Results	
BC027803 - Mus musculus, Lrrtml protein	
...	
Vega Protein Coding Annotations	
OTTHUMT00000252294 at chr2:80515483-80531518	
...	

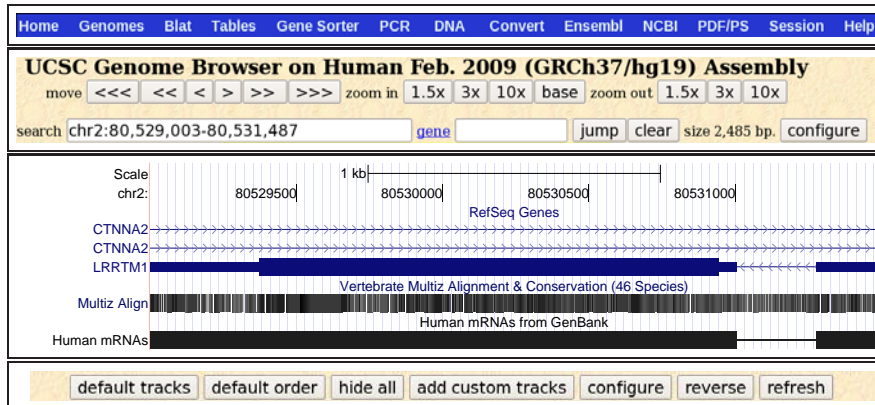
Como ocurre en este ejemplo, suele suceder con cierta frecuencia que para el mismo gen obtenemos anotaciones divergentes producidas por diferentes sistemas de anotación. Para acceder a la representación gráfica del genoma humano que contiene nuestro gen, vamos a seleccionar el enlace asociado a nuestro gen en la pista de RefSeq (RefSeq Genes en la Figura 7), cuyas anotaciones constituyen un estándar de referencia reconocido actualmente por toda la comunidad científica debido a su precisión. Una vez el servidor reconoce nuestra petición, abre la pantalla de navegación principal (ver Figura 8), para automáticamente enviarnos a la región (80,529,003-80,531,487) del cromosoma 2 que contiene nuestro gen (según el repositorio de RefSeq). El visor genómico está centrado por defecto en la región que contiene exactamente todos los exones de nuestro gen. Observamos los dos exones del gen *LRRTM1* en forma de cajas rectangulares junto con otras pistas básicas de información (la pista de conservación entre vertebrados y los ARN mensajeros humanos disponibles). Para indicar el sentido de traducción de cada gen, el servidor introduce múltiples flechas a lo largo de los 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 denominado *CTNNA2*, codificado en la hebra contraria.

Como veremos en esta misma sección, esta imagen posee contenido activo que nos permitirá navegar a través de estas pistas.

La primera conexión a UCSC suele mostrar un conjunto predefinido de pistas.

La estructura de los genes es analizada con más detalle en el módulo Fundamentos de biología molecular.

Figura 8. Pantalla inicial de navegación de UCSC

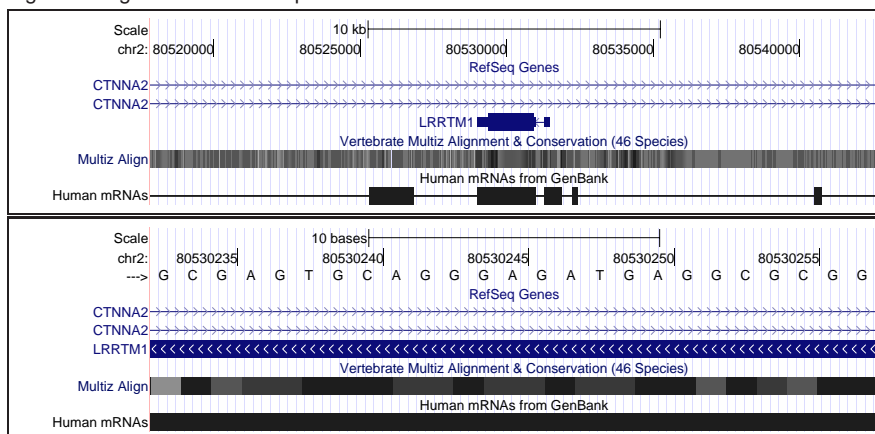


Los navegadores suelen proporcionar un conjunto sencillo de botones de movimiento (Figura 8). UCSC permite realizar movimientos utilizando (a) ampliaciones de posición con diferente resolución para acercar o alejar la escena, (b) desplazamientos horizontales de diferente tamaño a izquierda o derecha de la localización actual, (c) saltos hacia una región concreta mediante la caja de coordenadas y (d) búsquedas rápidas dirigidas por el identificador de cualquier elemento catalogado en este servidor.

Una vez accedemos a una pista concreta desde la pantalla inicial, el navegador nos muestra la anotación centrada ocupando toda la ventana gráfica. Para modificar la panorámica podemos alterar la perspectiva usando tres modos de resolución (1.5x, 3x y 10x). Por ejemplo, ocurre en ocasiones que es necesario utilizar *zoom out* para 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 *zoom in*.

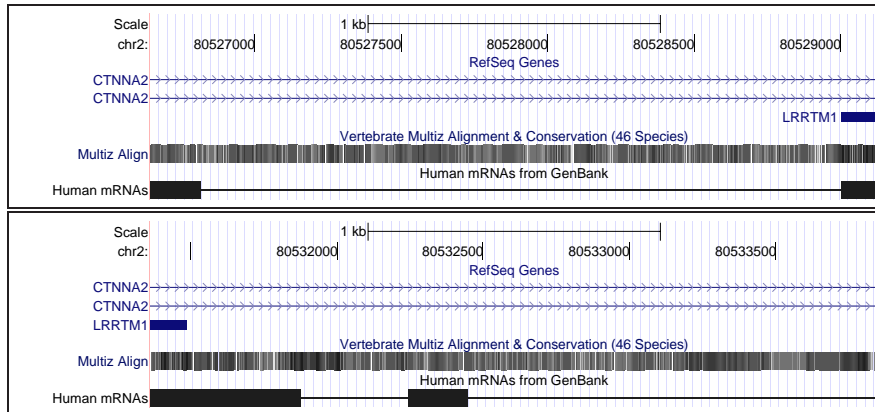
El botón *base* intensifica al máximo el acercamiento para visualizar directamente la secuencia de nucleótidos.

Figura 9. Jugando con el enfoque de UCSC



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.

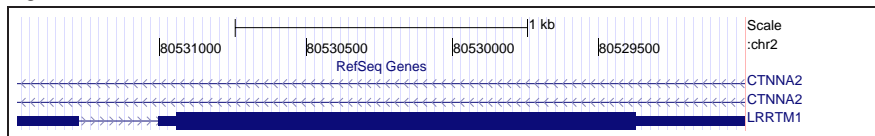
Figura 10. Desplazamiento por el genoma con UCSC



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 *LRRTM1* 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.

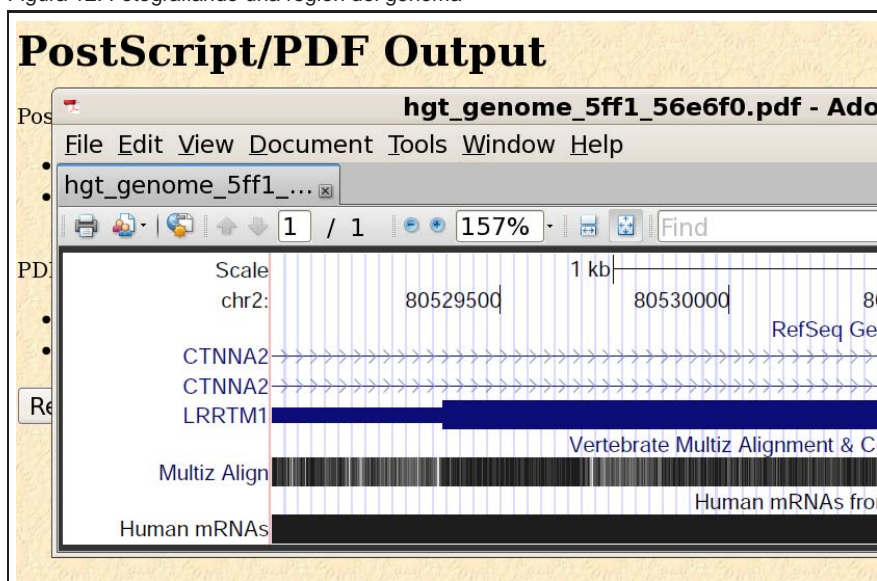
Figura 11. Cambiando la hebra de ADN con UCSC



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).

Esta opción, en combinación con la inclusión de nuestras propias pistas de datos, resulta especialmente útil para la composición de imágenes sobre anotaciones genómicas.

Figura 12. Fotografiando una region del genoma



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 13).

Figura 13. Extracción de la secuencia de una región genómica

Get DNA in Window

Get DNA for

Position

Note: if you would prefer to get DNA for features of a particular track or table, try the [Table Browser](#) using the output format sequence.

Sequence Retrieval Region Options:

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

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

Sequence Formatting Options:

☒ All upper case.
☐ All lower case.
☐ Mask repeats: ☒ to lower case ☐ to N
☐ Reverse complement (get '-' strand sequence)

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

1.4. El navegador genómico UCSC: pistas

Las pistas contienen anotaciones suministradas por distintos consorcios y grupos de investigación. Para mostrar únicamente aquella 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 servidores 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. En consecuencia, el navegador ofrece un conjunto de opciones ligeramente distinto entre diferentes genomas. No obstante, los bloques principales de opciones suelen conservarse dentro de los servicios ofrecidos por el mismo servidor. El listado de bloques proporcionados por UCSC para la versión hg19 del genoma humano puede encontrarse en la Tabla 5. Utilizaremos como referencia esta distribución a lo largo de las explicaciones incluidas en esta sección.

El usuario puede modificar el orden relativo entre las pistas mostradas en el visor genómico de UCSC arrastrándolas verticalmente hacia su nueva posición.

Tabla 5. Bloques de pistas de UCSC

Nombre	Descripción
Mapping and Sequencing Tracks	Ensamblado y secuenciación
Phenotype and Disease Associations	Fenotipos y enfermedades
Genes and Gene Prediction Tracks	Genes y predicción génica
mRNA and EST Tracks	Tránscritos experimentales
Expression	Expresión génica
Regulation	Regulación génica
Comparative Genomics	Genómica comparada
Variation and Repeats	Polimorfismos y 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 14, 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.

Más información sobre las formas alternativas de los genes en el módulo Fundamentos de biología molecular.

Figura 14. Configurando la visualización de una pista

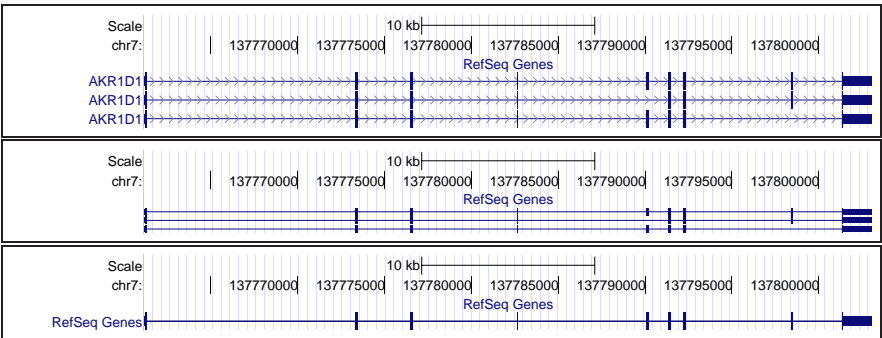


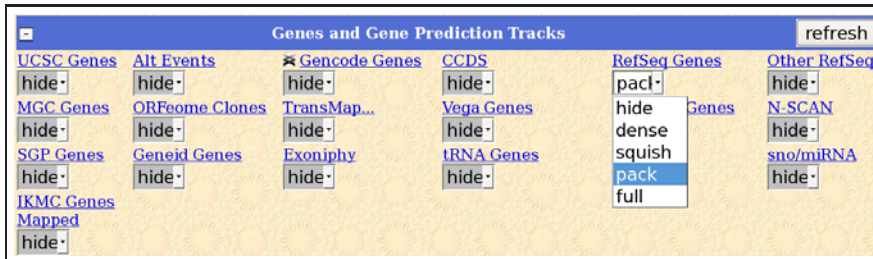
Figura 14

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 Prediction Tracks* (en inglés, pistas de genes y predicción génica). Este bloque, mostrado en la Figura 15, nos permite acceder a todos los datos que diferentes repositorios poseen sobre la región que estamos visualizando en un momento dado. En particular, podemos encontrar la pista de anotaciones RefSeq Genes producida por el consorcio RefSeq.

La identificación computacional de genes será introducida más adelante en estos mismos materiales.

Figura 15. Bloque de pistas de anotación de genes de UCSC



UCSC ofrece información específica sobre cada pista, accesible mediante el enlace coloreado en azul que muestra el nombre de cada pista. Por ejemplo, cuando seleccionamos *RefSeq Genes* en el bloque de opciones de información génica aparece una nueva pantalla que informa sobre el origen de esas anotaciones:

Figura 16. Información asociada a la pista RefSeq Genes

RefSeq Genes Track Settings

[View table schema](#)

Data last updated: 2010-10-13

Description

The RefSeq Genes track shows known human protein-coding and non protein-coding genes taken from the NCBI RNA reference sequences collection (RefSeq). The data underlying this track are updated daily.

Methods

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.

Credits

This track was produced at UCSC from RNA sequence 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, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 2005 Jan 1;33(Database issue):D501-4.

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* (Figura 8). Podemos ver un resumen del contenido de esta entrada en la Figura 17. 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 servidor nos proporciona los enlaces que ha recopilado para este gen en distintas bases de datos primarias (e.g. Entrez, Pubmed, OMIM, GeneCards, etc.).

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

Recomendamos al estudiante el análisis práctico detallado de todos los enlaces de esta entrada.

Figura 17. Anotaciones para la pista RefSeq Genes

```
RefSeq Gene LRRTM1
RefSeq: NM\_178839.4 Status: Validated
Description: Homo sapiens leucine rich repeat
              transmembrane neuronal 1 (LRRTM1), mRNA.
CCDS: CCDS1966.1
CDS: 3' complete
OMIM: 610867
Entrez Gene: 347730
PubMed on Gene: LRRTM1
PubMed on Product: leucine-rich repeat transmembrane neuronal
GeneCards: LRRTM1
AceView: LRRTM1
Stanford SOURCE: NM\_178839

Summary of LRRTM1
Position: chr2:80529003-80531487
Band: 2p12
Genomic Size: 2485
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

Data last updated: 2010-10-11
```

Antes de continuar analizando el contenido de la ficha de anotaciones que ofrece RefSeq para nuestro gen, proponemos al estudiante profundizar el recorrido 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, como RefSeq, 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 18 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.

Lectura complementaria

D.L. Wheeler *et al.* (2004). Database resources of the NCBI: update. Nucleic Acids Research 32:D35-D40.

Figura 18. Información en Entrez sobre el gen *LRRTM1*

LRRTM1 leucine rich repeat transmembrane neuronal 1 [*Homo sapiens*]
Gene ID: 347730, updated on 19-Sep-2010

Summary

Official Symbol LRRTM1 provided by [HGNC](#)
Official Full Name leucine rich repeat transmembrane neuronal 1 provided by [HGNC](#)
Primary source [HGNC:19408](#)
Locus tag UNQ675/PRO1309
See related [Ensembl:ENSG00000162951](#); [HPRD:14323](#); [MIM:610867](#)
Gene type protein coding
RefSeq status VALIDATED
Organism [Homo sapiens](#)
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia
Haplorrhini; Catarrhini; Hominidae; Homo
Also known as FLJ32082; LRRTM1

Bibliography

Related articles in PubMed

1. [Polymorphisms in leucine-rich repeat genes are associated with autism spectrum disorder susceptibility in populations of European ancestry.](#) Sousa I, et al. Mol Autism, 2010 Mar 25. PMID 20678249.
2. [Supporting evidence for LRRTM1 imprinting effects in schizophrenia.](#) Ludwig KU, et al. Mol Psychiatry, 2009 Aug. PMID 19626025.
3. [Understanding the genetics of behavioural and psychiatric traits will only be achieved through a realistic assessment of their complexity.](#) Francks C. Laterality, 2009 Jan. PMID 19125367.
4. [Where and what is the right shift factor or cerebral dominance gene? A critique of Francks et al. \(2007\).](#) Crow TJ, et al. Laterality, 2009 Jan. PMID 19125366.
5. [Editorial commentary: is LRRTM1 the gene for handedness?](#) McManus C, et al. Laterality, 2009 Jan. PMID 19125365.

[See all \(13\) citations in PubMed](#)

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

Tabla 6. Funciones biológicas del gen *LRRTM1*

Función	Evidencia
<i>protein binding</i>	computacional
<i>axon</i>	bibliográfica
<i>NOT cell surface</i>	bibliográfica
<i>endoplasmic reticulum</i>	bibliográfica
<i>endoplasmic reticulum membrane</i>	computacional
<i>growth cone</i>	bibliográfica
<i>integral to membrane</i>	computacional
<i>membrane</i>	computacional

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 19 las coordenadas de los dos exones del ARN mensajero del gen *LRRTM1* (1..211 y 212..2212) así como su fracción codificante (271..1839).

Gene Ontology es un diccionario de términos que permite describir cientos de eventos biológicos distintos. Estas definiciones se emplean para caracterizar los genes en base a evidencias experimentales. Más información en la siguiente dirección: www.geneontology.org

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

Figura 19. Entrada de la secuencia NM_178839 en GenBank

LOCUS	NM_178839 2217 bp mRNA linear PRI 19-SEP-2010
DEFINITION	Homo sapiens leucine rich repeat transmembrane neuronal 1 (LRRTM1), mRNA.
ACCESSION	NM_178839
VERSION	NM_178839.4 GI:86990455
SOURCE	Homo sapiens (human)
ORGANISM	Homo sapiens Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.
REFERENCE	2 (bases 1 to 2217)
AUTHORS	Ludwig,K.U., Mattheisen,M., Muhleisen,T.W., Roeske,D., Schmal,C., Breuer,R., Schulte-Korne,G., Muller-Myhsok,B., Nothen,M.M., Hoffmann,P., Rietschel,M. and Cichon,S.
TITLE	Supporting evidence for LRRTM1 imprinting effects in schizophrenia
JOURNAL	Mol. Psychiatry 14 (8), 743-745 (2009)
PUBMED	19626025
REMARK	GeneRIF: Results strengthen the evidence for an association of imprinted alleles of LRRTM1 with schizophrenia. Weaker supportive evidence was also obtained for a possible association of LRRTM1 with human brain asymmetry.
FEATURES	Location/Qualifiers
source	1..2217 /organism=Homo sapiens /mol_type=mRNA /db_xref=taxon:9606 /chromosome=2 /map=2p12
gene	1..2217 /gene=LRRTM1 /gene_synonym=FLJ32082 /note=leucine rich repeat transmembrane neuronal 1
exon	1..211 /gene=LRRTM1 /gene_synonym=FLJ32082 /inference=alignment:Splice /number=1
exon	212..2212 /gene=LRRTM1 /gene_synonym=FLJ32082 /inference=alignment:Splice /number=2
CDS	271..1839 /gene=LRRTM1 /translation=MDFULLGLCLYWLRRPSGVVLCCLGACFQMLPAAPS GCPQLCRCEGRLLYCEALNLTEAPHNLSGLLGLSLRYNSLSELRAGQFTG ... QHDGTFEPATVALPGGEHAENAVQIHKVVTTGTMALIFSFLIVVLVLYVSW KCFPASLRQLRQCFVTQRRKQKQKQTMHQMAAMSAQEYYVDYKPNHIEGA LVIINEYGSCTCHQQPARECEV
ORIGIN	1 ttgcgggtcc tagaagtcgc ctccccgcct tgcgggccgc...

Dentro de la ficha de anotaciones de RefSeq que proporciona el navegador UCSC (ver Figura 17), 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 20).

FASTA

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.

Figura 20. Secuencia de la proteína LRRTM1

```
>NP_849161 length=522
MDFLLGLCLYLWLLRRPSGVVLCLLGACFQMLPAAPSGCPQLCRCEGRLLYCEALNLTEAPHNLSGLLG
LSLRYNLSSELRAQQTGLMQLTWLYLDHNHICSVQGDADFQKLRRVKELTSSNQITQLPNTTFRPMPN
LRSVDLSYNKLQALAPDLFHLRLKLTTLHMRANAIQFVVRIFQDCRSCLKFLDIGYNQLKSLARNSFAG
LFLKLTLEHLEHNDLVKNFNAHFPRILISLHSLCLRRNKVAIVVSSLDWVWNLEKMDLSGNEIYMEPHVF
ETVPHLQSLQLDSNRLTYIEPRIINSWKSLSITLAGNLWDCGRNVCALASWLNQFQGRYDGNLQCASP
EYAQGEDVLDVAVYAFHLCEDGAEPSTGHLLSAVTNRSDLGPPASSATTADGGEGQHDGTFEPATVALP
GGEHAENAVQIHKVVTGTMALIFSFLIVVLVLYVSWKCFPASLRQLRQCFVTQRRKQKQKQTMHQMAM
SAQEYYVDYKPNHIEGALVIINEYGSCTCHQQPARECEV
```

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 21). De este modo, es posible obtener un fragmento de la secuencia promotora del gen (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.

Figura 21. Extraer diferentes regiones de un gen con UCSC

Genomic Sequence Near Gene

Get Genomic Sequence Near Gene

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

Sequence Retrieval Region Options:

☐ Promoter/Upstream by bases

☒ 5' UTR Exons

☒ CDS Exons

☒ 3' UTR Exons

☐ Introns

☐ Downstream by bases

☒ One FASTA record per gene.

☐ One FASTA record per region (exon, intron, etc.) with extra bases upstream (5') and extra downstream (3')

☐ Split UTR and CDS parts of an exon into separate FASTA records

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

Sequence Formatting Options:

☒ Exons in upper case, everything else in lower case.

☐ CDS in upper case, UTR in lower case.

☐ All upper case.

☐ All lower case.

☐ Mask repeats: ☒ to lower case ☐ to N

A continuación mostramos en la Figura 22 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 del genoma (hg19), la pista (refGene equivale a RefSeq Genes), el código RefSeq (NM_178839), el exón (0 o 1) y su ubicación genómica.

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

Figura 22. Secuencia del transcrito del gen *LRRTM1*

```
>hg19_refGene_NM_178839_0 range=chr2:80531277-80531487
ttgcgggtcctagaagtcgcctccccgccttgccggccgccttgccagccccgagccgagcag
caaagtgagacattgtgagcctgccagatccgcggccgcggaccggggtgcctcggaaca
cagagggtctctctcgcctgcataataattagcctgcacacaaaggagcagctgaatgga
ggttgctcactctctggaaaagg
>hg19_refGene_NM_178839_1 range=chr2:80529003-80531003
atttctgaccgagcgcttccaatggacattctccagctctctctggaaagattctcgctaATGG
ATTTCTGCTGCTCGGTCTCTGTCTATACTGGCTGCTGAGGAGGCCCTCGGGGGTGGTCTTGT
GTCTGCTGGGGGCTGCTTTCAGATGCTGCCCCGCGCCCCAGCGGTGCCCCGAGCTGTGCC
GGTGCAGAGGGCGGCTGCTGTACTGCGAGGCGCTCAACCTCACGAGGCGCCCCAACCTGT
CCGGCTGCTGGGCTTGTCCCTGCGCTACAACAGCCTCTCGGAGCTGCGCGCCGCCAGTTCA
CGGGGTTAATGCAGCTCACGTGGCTCTATCTGGATCACAATCACATCTGCTCCGTGCAGGGG
ACGCCTTTCAGAACTGCGCCGAGTTAAGGAACTCACGCTGAGTTCCAACAGATCACCAAC
TGCCCAACACCACCTTCCGGCCCATGCCAACCTGCGCAGCGTGGACCTCTCGTACAACAAGC
TGCAGGCGCTCGCGCCGACCTCTTCCACGGGCTGCGGAAGCTCACACGCTGCATATGCGGG
CCAACGCCATCCAGTTTGTGCCGTGCGCATCTTCCAGGACTGCCGAGCCTCAAGTTTCTCG
ACATCGGATACAATCAGCTCAAGAGTCTGGCGCGCAACTCTTTCGCGGCTTGTTTAAGCTCA
CCGAGCTGCACCTCGAGCACAACGACTTGGTCAAGGTGAAGTTGCGCCACTTCCGCGCCTCA
TCTCCCTGCACCTCGCTCTGCCTGCGGAGGAACAAGGTGGCCATTGTGGTCAGCTCGCTGGACT
GGGTTTGAACCTGGAGAAAATGGACTTGTGCGGCAACGAGATCGAGTACATGGAGCCCCATG
TGTTGAGACCGTGCCGACCTGCAGTCCCTGCGAGCTGGACTCCAACCGCCTCACCTACATCG
AGCCCCGGATCTCAACTCTTGAAAGTCCCTGACAAGCATCACCTGGCCGGGAACCTGTGGG
ATTGCGGGCGCAACGTGTGTGCCCTAGCCTCGTGGCTCAACAACTCCAGGGGCGCTACGATG
GCAACTTGCAGTGCGCCAGCCGGAGTACGCACAGGGCGAGGACGTCTGGACGCGGTGTACG
CCTTCCACCTGTGCGAGGATGGGGCCGAGCCACCAGCGGCCACTGCTCTCGGCCGTACCA
ACCGCAGTGATCTGGGGCCCCCTGCCAGCTCGGCCACCACGCTCGCGGACGGCGGGGAGGGG
AGCACGACGGCACATTGAGCCTGCCACCGTGGCTCTTCCAGGCGGCGAGCACGCCGAGAACG
CCGTGCAGATCCACAAGTGGTACGGGCACCATGGCCCTCATCTTCTCCTTCTCATCGTGG
TCCTGGTGCTCTACGTGTCTGGAAGTGTTTCCAGCCAGCCTCAGGCAGCTCAGACAGTGCT
TTGTACGCGAGCGCAGGAAGCAAAAGCAGAAACAGACCATGCATCAGATGGCTGCCATGTCTG
CCCAGGAATACTACGTTGATTACAAACCGAACCATTTAGGGAGCCCTGGTGATCATCAACG
AGTATGGCTCGTGTAACCTGCCACCAGCAGCCCGGAGGGAATGCGAGGTGTGAttgtcccagt
ggctctcaacccatgagctaccaaatacgccctgggcagccgggacgggcccggcgccaccagg
ctggggtctccttgctgtgctctgatgtctccttgactgaaactttaaggggatctctccc
agagacttgacatttttagctttattgtgtcttaaaaacaaaagcgaattaaaacacaacaaaa
aaccaccacccacaaccttcaggacagtctatcttaatttcatatgagaactccttcctccc
tttgaagatctgtccatattcaggaatctgagagtgtaaaaaggtaccaatcattgattttt
ttttttttgtaaactaaaatgttttaataaaaatagcattttacagt
```

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. 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 reconocer las formas ortólogas de cada gen en otros genomas (Figura 23). Con esta información, el usuario debe cambiar el genoma de referencia dentro del navegador y posteriormente utilizar el nuevo código suministrado por RefSeq para recuperar la secuencia homóloga de dicho ARN mensajero.

Los alineamientos de secuencias son explicados detalladamente más adelante en estos materiales.

Figura 23. Formas ortólogas del gen humano *KCMF1* en RefSeq

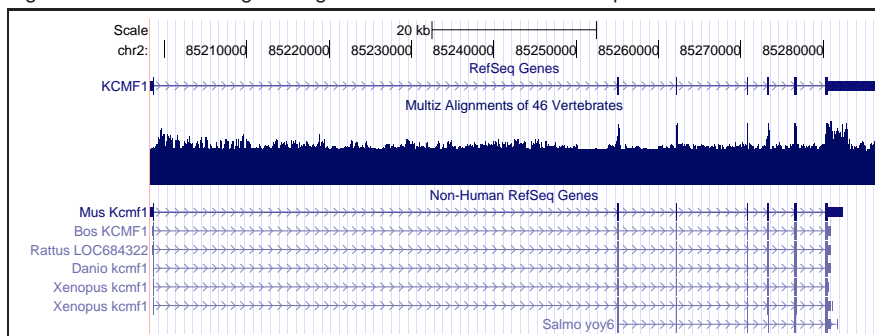


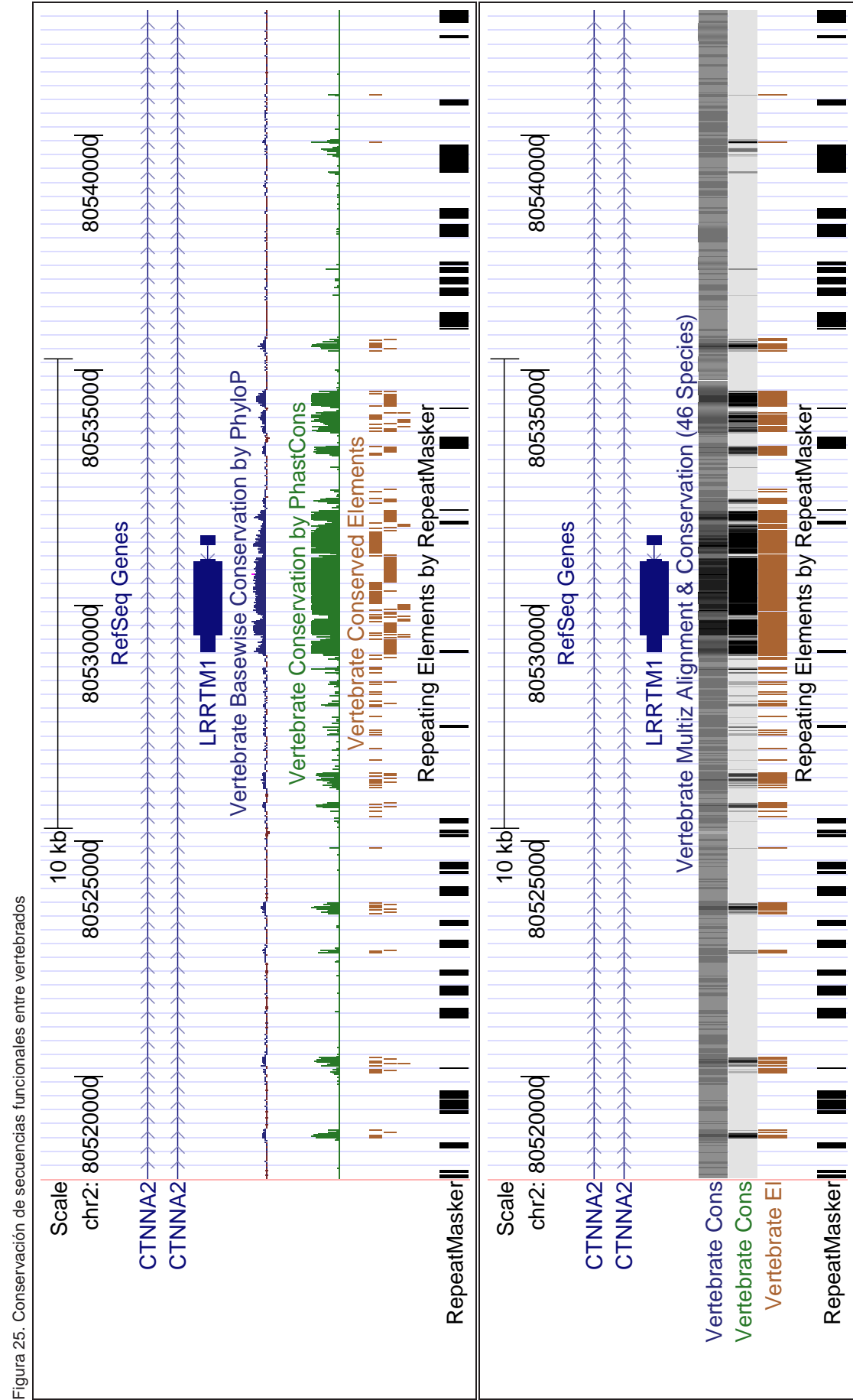
Figura 23

Pueden apreciarse nítidamente los picos de conservación precisamente sobre las regiones exónicas del gen en todas las especies.

Los bloques de pistas de genómica comparada (en inglés, *comparative genomics*) y de predicción génica son las secciones de anotaciones suministradas por el servidor genómico UCSC más frecuentemente utilizadas por los analistas bioinformáticos. Como podemos ver en la Figura 24, disponemos de múltiples comparaciones entre la secuencia que estamos utilizando de referencia (humano, en este caso la distribución NCBI36/hg18) y los genomas ensamblados de otros vertebrados. 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 servidor. Con estas pistas, por ejemplo, podemos establecer la existencia de regiones codificantes a través de los alineamientos de las secuencias homólogas (ver Figura 25).

Figura 24. Bloque de genómica comparada





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 pista *Conservation* contiene alineamientos entre la mayoría de los grupos de especies más representativos. En su interior, por ejemplo, tenemos la opción de 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 26. Configuración de la pista *Conservation*

Multiz Alignments Configuration

Species selection:

Primate

☒ chimp ☒ gorilla ☒ orangutan ☒ rhesus ☒ baboon
☒ marmoset ☒ tarsier ☒ mouse lemur ☒ bushbaby

Placental Mammal

☒ tree shrew ☒ mouse ☒ rat ☒ kangaroo rat ☒ guinea pig
☒ squirrel ☒ rabbit ☒ pika ☒ alpaca ☒ dolphin
☒ cow ☒ horse ☒ cat ☒ dog ☒ microbat
☒ megabat ☒ hedgehog ☒ shrew ☒ elephant ☒ rock hyrax
☒ tenrec ☒ armadillo ☒ sloth

Vertebrate

☒ wallaby ☒ opossum ☒ platypus ☒ chicken ☒ zebra finch
☒ lizard ☒ x. tropicalis ☒ tetraodon ☒ fugu ☒ stickleback
☒ medaka ☒ zebrafish ☒ lamprey

Multiple alignment base-level:
☐ Display bases identical to reference as dots
☒ Display chains between alignments

Codon Translation:
 Default species to establish reading frame:
☐ No codon translation
☒ Use default species reading frames for translation
☐ Use reading frames for species if available, otherwise no translation
☐ Use reading frames for species if available, otherwise use default species

Species List: Human, Chimpanzee, Gorilla, Orangutan, Rhesus, Baboon, Marmoset, Tarsier, Mouse lemur, Bushbaby, Tree shrew, Mouse, Rat, Kangaroo rat, Guinea pig, Squirrel, Rabbit, Pika, Alpaca, Dolphin, Cow, Horse, Cat, Dog, Microbat, Megabat, Hedgehog, Shrew, Elephant, Rock hyrax, Tenrec, Armadillo, Sloth, Wallaby, Opossum, Platypus, Chicken, Zebra finch, Lizard, X. tropicalis, Tetraodon, Fugu, Stickleback, Medaka, Zebrafish, Lamprey.

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 27. Funcionamiento de BLAT

BLAT Search Genome

Genome: Assembly: Query type:

>CDS
 ATGGATTTTCCTGCTGCTCGGTCTCTGTCTATACTGGCTGCTGAGGAGGCC

Scale: 20 bases
 chr2: 80530900 80530905 80530910 80530915 80530920 80530925 80530930 80530935 80530940
 --> GGCCTCCTCAGCAGCCAGTATAGACAGAGACCGAGCAGCAGGAATCCAT

Your Sequence from Blat Search

Blat Sequence

RefSeq Genes

CTNNA2
 CTNNA2
 LRRTM1

Lectura complementaria

W.J. Kent (2002). *BLAT: the BLAST-like alignment tool*.
 Genome Research 12:656-664.

Figura 27

Estamos identificando el inicio de la región codificante 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 mas similar a nuestra secuencia en otro genoma. Con este procedimiento, podemos detectar para el gen humano *LRRTM1*, su correspondiente ortólogo anotado dentro del genoma de ratón:

Figura 28. Utilizando BLAT para identificar genes ortólogos

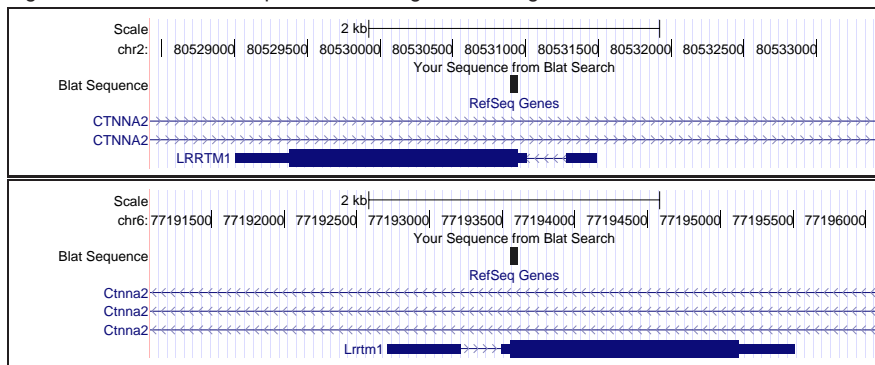


Figura 28

Presentamos el genoma humano en la parte superior de la imagen y el genoma de ratón en la parte inferior.

1.5. El navegador genómico UCSC: anotaciones propias

Los servidores 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.

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.

Una vez el usuario introduce sus propias pistas, éstas son automáticamente integradas dentro de la imagen proporcionada por el servidor. 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 servidor para estudiar el comportamiento de nuestros resultados. De este modo, todas las herramientas disponibles para llevar a cabo comparaciones entre distintas anotaciones, tanto a nivel cualitativo como cuantitativo, pueden aprovecharse directamente para contrastar el conocimiento de referencia con nuevos datos experimentales observados para cualquier característica genómica.

Dentro del entorno de trabajo del navegador UCSC, las pistas introducidas en el 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. 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.

El estudiante puede encontrar abundante información sobre la carga de nuevas pistas dentro de las páginas de ayuda del navegador UCSC.

El servidor UCSC ofrece también un sistema de sesiones de usuario protegidas bajo contraseña, facilitando la comunicación de datos entre diversos colaboradores.

UCSC permite la carga directa de ficheros comprimidos, para facilitar una rápida transferencia de los datos a través de la Red.

El módulo Fundamentos de informática en entornos bioinformáticos instruye al estudiante sobre la manipulación de ficheros con LINUX.

Una vez la operación de carga de nuestra pista finaliza satisfactoriamente, podemos visualizar todos sus elementos a lo largo del genoma junto con el resto de anotaciones convencionales. Esta completa integración de los nuevos datos resulta muy interesante para futuras operaciones. Por regla general, el navegador mantiene en memoria nuestras pistas durante uno o dos días únicamente, para no almacenar un volumen excesivo de datos. 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 y la comparación con otras bases de datos primarias.

El bioinformático puede crear sus propias pistas en numerosos formatos dentro de UCSC. En estos materiales vamos a focalizar nuestro interés en los formatos más populares para la integración de anotaciones genómicas propias. 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.

Cada pista debe poseer la siguiente estructura dividida en tres componentes.

- Configuración del navegador: listado de pistas visibles.
- Configuración de la pista adaptada: parámetros de visualización.
- Anotaciones de la pista adaptada: valores identificados.

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 29. Encabezamiento de pistas propias (1)

```
browser position cromosoma:inicio-final
browser [hide/dense/pack/full] pista 1
browser [hide/dense/pack/full] pista 2
...
browser [hide/dense/pack/full] pista n
```

A continuación, dentro del mismo fichero, introduciendo la palabra clave `track` debemos definir los parámetros de visualización de la nueva pista: asignar un nombre y una descripción, definir un modo de visibilidad o seleccionar un color.

Figura 30. Encabezamiento de pistas propias (2)

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

Finalmente, debemos aportar las coordenadas de nuestras anotaciones genómicas para integrarlas en la visualización final. Cada línea corresponde exclusivamente a una anotación efectuada por el usuario, introduciéndose el carácter tabulador “\t” para separar los distintos atributos. Podemos emplear varios formatos para codificar esta información. Para elementos genómicos delimitados por pares de coordenadas es necesario utilizar el formato BED o el estándar GFF. Para mostrar el comportamiento de una determinada característica biológica a lo largo del genoma debemos especificarlo en el formato WIG.

Recomendamos precaución con la versión del genoma. Nuestras propias anotaciones deben obtenerse con esa misma secuencia de referencia.

El formato BED (*Browser Extensible Format*, del inglés formato extensible del navegador) es útil para plasmar en pantalla aquellas anotaciones constituídas 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 31, el estudiante puede observar la sintaxis completa de este formato:

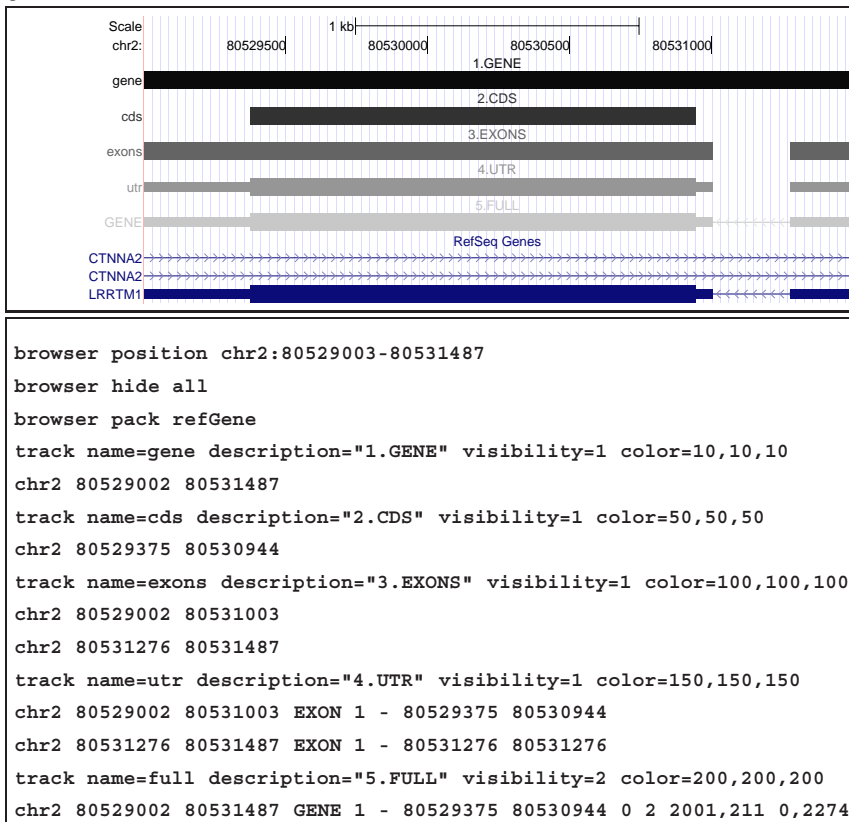
El usuario puede guardar más de una pista dentro de un mismo fichero. Es posible también codificar los datos en diferentes formatos.

Figura 31. 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 32). Tomando como referencia la información que conocemos sobre la ubicación del gen *LRRTM1*, 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 32), 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 y la longitud de éstos (pista 5, Figura 32).

Figura 32. Anotaciones en distintas versiones del formato BED



El formato WIG (*wiggle* en inglés, agitado) es útil para integrar en el 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 referentes a modificaciones de histonas y reconocimiento de factores de transcripción son claros ejemplos de esta familia de anotaciones. Para codificar estos elementos, el bioinformático debe asignar un valor numérico a cada posición de una determinada región genómica. Esta cantidad representa la intensidad de la señal biológica anotada en esta nueva pista. 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`.

Figura 33. Deficiones del formato WIG

```

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

```

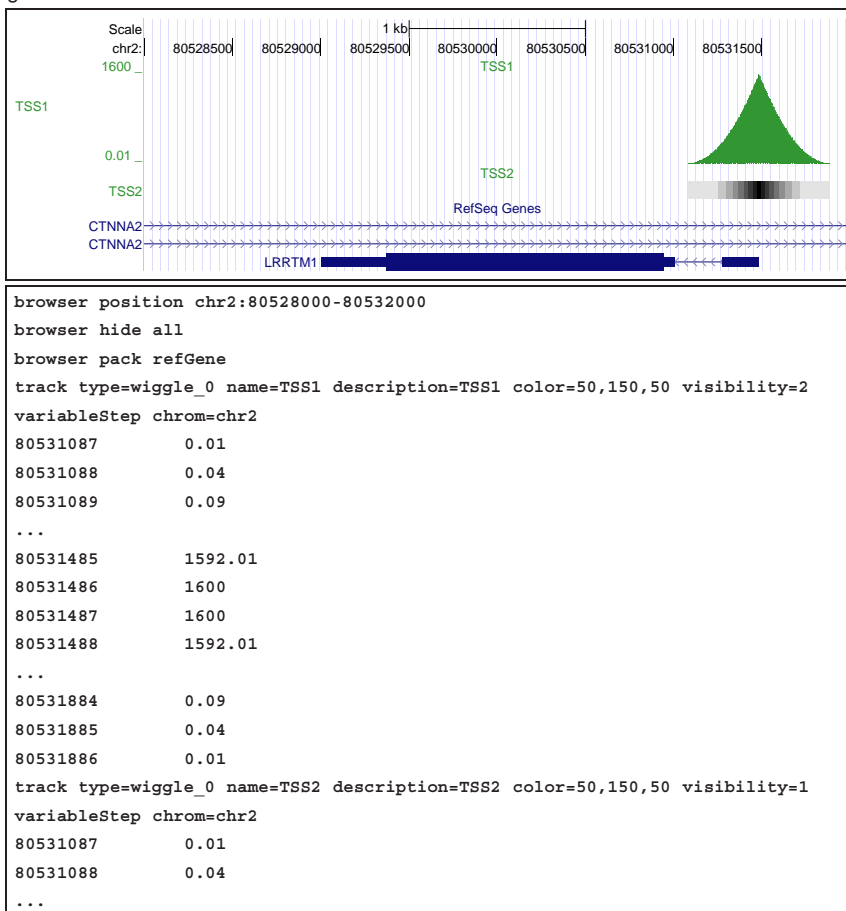
Lecturas complementarias

S.L. Berger (2007). *The complex language of chromatin regulation during transcription*. Nature 447:407-12.

A. Barski et al. (2007). *High-resolution profiling of histone methylations in the human genome*. Cell 129:823-37.

En la Figura 34 procedemos a cargar una pista en formato WIG para anotar una hipotética función de probabilidad asociada a la localización del posible inicio de transcripción de nuestro gen *LRRTM1* (mostramos las mismas anotaciones utilizando dos modos distintos de visualización).

Figura 34. Anotaciones en formato WIG



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 servidor interpretará que éstas pertenecen al mismo un 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:

El estudiante encontrará más información sobre el estándar GFF en la dirección <http://www.sanger.ac.uk/Software/formats/GFF>

Figura 35. Definición del formato GFF

```

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

```

Este formato está francamente 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 más populares. El eje central delimita si las predicciones se encuentran en la hebra positiva o negativa de la molécula de ADN:

Figura 36. Representación gráfica de predicciones con GFF2PS

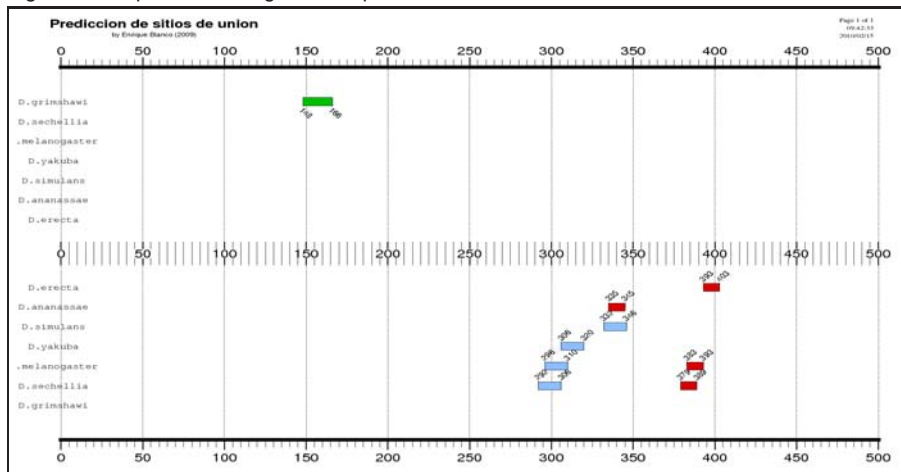


Figura 36

Para personalizar la representación gráfica, el programa GFF2PS permite introducir un fichero de configuración de colores. Ver el manual de usuario del programa para más detalles.

1.6. El navegador genómico UCSC: ENCODE

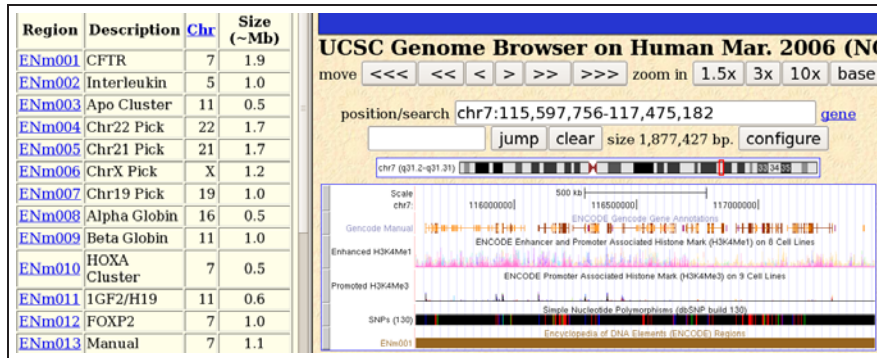
La obtención de la secuencia del genoma humano proporcionó a la comunidad científica mundial el acceso inmediato a una enorme cantidad de información inédita hasta ese momento. Este hito histórico reveló también la deficiente calidad de las anotaciones existentes. Un consorcio internacional de investigadores fundaron entonces el proyecto ENCODE (*Encyclopedia of DNA Elements*, en inglés enciclopedia de elementos del ADN), con la misión de producir para varias clases de tejidos una anotación mucho más precisa de todos los componentes codificados en el interior del genoma humano (e.g. genes, transcritos, microARNs, elementos regulatorios, etc.). Para lograr este objetivo, el mismo consorcio centró una parte sustancial de sus esfuerzos en el desarrollo de nuevas técnicas experimentales y computacionales que permitieron progresar espectacularmente en múltiples áreas de conocimiento. El navegador genómico UCSC fue escogido como herramienta para servir estos datos de forma pública. Las pistas de datos ENCODE se depositaban en este servidor inmediatamente después de llevar a cabo su anotación. Estos datos, no obstante, poseían un periodo de cuarentena (nueve meses desde su liberación) que impedía utilizarlas antes en otra publicación, para proteger a sus autores originales. En una primera fase del proyecto, entre los años 2003 y 2007, el consorcio ENCODE dirigió sus esfuerzos sobre el uno por ciento del genoma humano (la fase piloto del proyecto). Para acceder a las anotaciones de esta primera fase, el bioinformático debe seleccionar el enlace *ENCODE pilot project* en la página principal de UCSC. Una vez allí, puede escoger cualquiera de las 44 regiones habilitadas en esta fase del proyecto y abrir el visor genómico convencional (ver Figura 37).

Lecturas complementarias

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.

K.R. Rosenbloom et al. (2010).
ENCODE whole-genome data in the UCSC Genome Browser. Nucleic Acids Research 38:D620-5.

Figura 37. Explorando las regiones humanas anotadas en la fase piloto de ENCODE



Para acceder a través de UCSC a la anotación completa del genoma humano que está llevando a cabo este consorcio público debemos seleccionar el enlace *ENCODE* desde la página *web* inicial de este portal. Este proyecto está contribuyendo sustancialmente a incrementar nuestro conocimiento respecto a los patrones de expresión génica y la regulación de la transcripción en diferentes tejidos. En particular, el subproyecto GENCODE produjo un catálogo de genes de referencia mediante la revisión manual de anotaciones automáticas generadas por varios grupos de investigación.

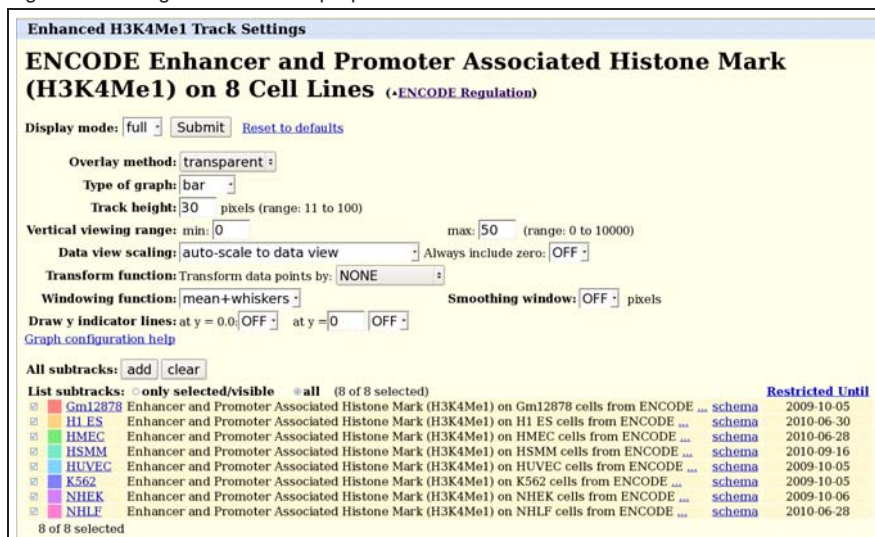
Para gestionar todo este volumen de información los diseñadores de UCSC han creado recientemente un nuevo formato de pista denominado *supertrack* (en inglés, superpista) que aglutina varias pistas de datos. La secuenciación masiva en varias líneas celulares de aproximadamente una docena de modificaciones post-transcripcionales de histonas puede visualizarse de forma óptima empleando este tipo de pistas. Cada marca posee su propia distribución a lo largo del genoma, presentando un patrón de picos asociable a distintos elementos genómicos. Si, como es esperable, crece el número de modificaciones estudiadas con este procedimiento parece evidente que la disposición horizontal clásica de pistas presentadas en el visor genómico de UCSC podría resultar impracticable. Para optimizar el espacio útil en pantalla y facilitar la interpretación de la información, UCSC proporciona estos datos integrados en una superpista, fusionando los patrones observados en todos los tejidos para una única marca (ver Figura 38):

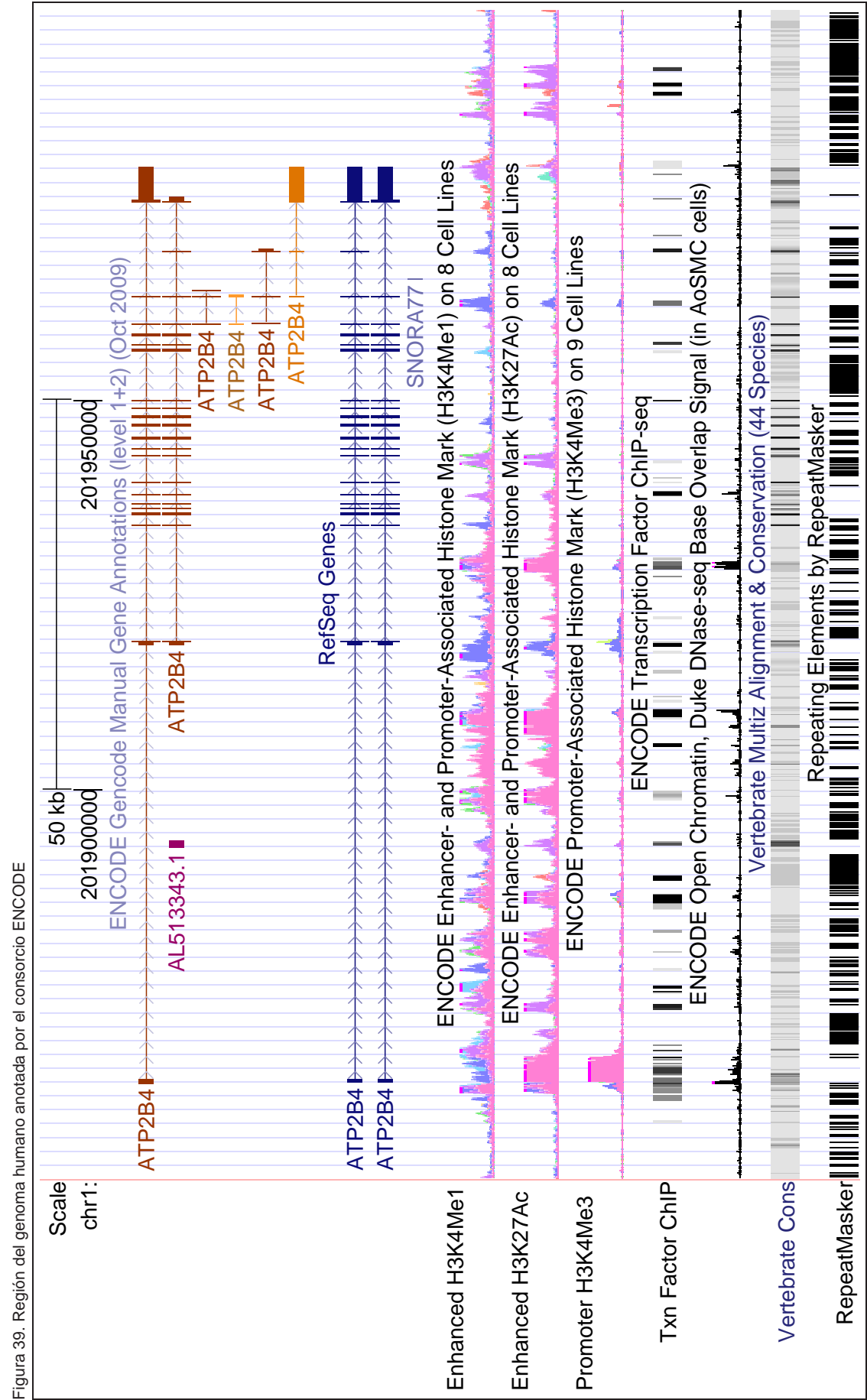
Lectura complementaria

J. Harrow *et al.* (2006).
GENCODE: producing a reference annotation for ENCODE. Genome Biology Suppl 1:S4.1-9.

La modificación H3K4Me3 marca los inicios de transcripción, mientras H3K4Me1 marca regiones reguladoras y H3K27Me3, las regiones de represión génica.

Figura 38. Configuración de la superpista H3K4Me1 en ENCODE





Estas anotaciones más sofisticadas se integran en una superpista denominada ENCODE Integrated Regulation Track (en inglés, pista de regulación de ENCODE). En esta superpista podemos manipular simultáneamente datos de secuenciación sobre histonas, factores de transcripción y transcritos, constituyendo una fuente de información inagotable para las investigaciones genómicas durante las próximas décadas. En la Figura 39 puede observarse la precisión de estas anotaciones para detectar un amplio conjunto de formas alternativas de un mismo gen previamente desconocidas. La marca ENCODE, como signo de excelencia científica, está en periodo de expansión hacia otros genomas, con la misión de proporcionar conjuntos similares de anotaciones para otros organismos. El proyecto modENCODE, en particular, representa sin ninguna duda una revolución similar dentro de la investigación de dos organismos modelo: la mosca de la fruta *Drosophila melanogaster* y el gusano *Caenorhabditis elegans*. Las anotaciones de este nuevo consorcio internacional están ya a disposición de ambas comunidades científicas gracias al portal dedicado a este proyecto (ver Figura 41).

Lecturas complementarias

S.E. Celniker *et al.* (2009).
Unlocking the secrets of the genome. Nature 459:927-930.

The modENCODE Consortium (2010).
Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330:1787-1797.

1.7. El navegador genómico UCSC: instalación local

UCSC sirve anotaciones para más de 50 genomas de distintas especies. 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 *Download genomes* (en inglés, descargar genomas) en la página principal permite recuperar los ficheros de texto interpretados por el interfaz gráfico de UCSC (secuencias y anotaciones):

Repasar estos conceptos en los materiales del módulo Fundamentos de informática en entornos bioinformáticos.

Figura 40. Página inicial de descargas de genomas en UCSC

Sequence and Annotation Downloads

This page contains links to sequence and annotation data downloads for the genome assemblies featured in the UCSC Genome Browser. Table downloads are also available via the Genome Browser FTP server. For quick access to the most recent assembly of each genome, see the current genomes directory. This directory may be useful to individuals with automated scripts that must always reference the most recent assembly.

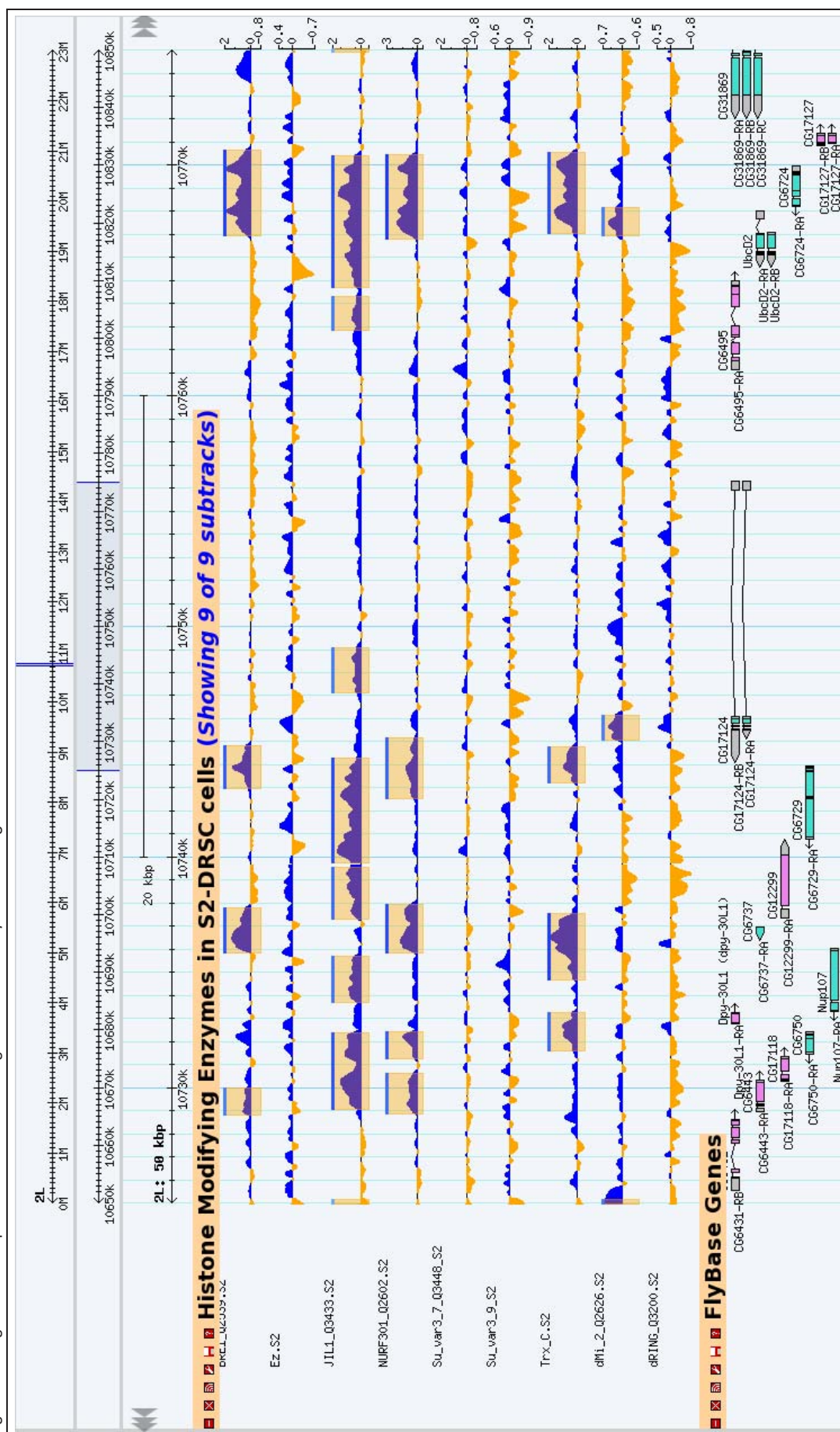
To view the current descriptions and formats of the tables in the annotation database, use the "describe table schema" button in the Table Browser. The Description of the annotation database page (no longer maintained) also provides descriptions of selected tables in the database.

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.

VERTEBRATES - Complete annotation sets

- Human
- Cat
- Chicken
- Chimpanzee
- (...)

Figura 41. Región anotada por modENCODE en el genoma de *Drosophila melanogaster*



En cualquier distribución de un genoma en particular básicamente podemos encontrar: (i) el conjunto completo de datos de la distribución que incluye los cromosomas y las anotaciones (Full data set), (ii) el conjunto completo de datos de un cromosoma en particular (Data set by chromosome), (iii) las anotaciones de cada pista en ficheros individuales (Annotation database), (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.

Figura 42. Descarga de anotaciones disponibles para el genoma humano

Human Genome Feb. 2009 (hg19, GRCh37)
* Full data set
* Data set by chromosome
* Annotation database
* GC percent data
* Protein database for hg19
* SNP131-masked FASTA files
* LiftOver files
* Pairwise Alignments
* Multiple Alignments

El conjunto de secuencias de nucleótidos de los cromosomas en cada genoma está empaquetado en un único archivo comprimido. En la Figura 43 podemos ver el contenido una vez descomprimido de la distribución hg19 del genoma humano, centrando nuestro interés en acceder a la secuencia del cromosoma 17. Para ello, hemos procedido a descargar y desempaquetar el fichero `chromFa.tar.gz`, visualizando después el fichero `chr17.fa` presentado en formato FASTA.

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

Figura 43. Descarga de la secuencia de los cromosomas humanos

<code>chromFa.tar.gz</code>	20-Mar-2009 09:21	905M
<code>chromFaMasked.tar.gz</code>	20-Mar-2009 09:30	477M
<code>est.fa.gz</code>	14-Oct-2010 01:07	1.4G
<code>refMrna.fa.gz</code>	14-Oct-2010 01:07	36M
<code>upstream1000.fa.gz</code>	14-Oct-2010 01:07	7.2M
<code>xenoMrna.fa.gz</code>	14-Oct-2010 01:08	910M
 <code>chromFa/</code>		
<code>chr10.fa</code>	<code>chr14.fa</code>	<code>chr18.fa</code>
<code>chr11.fa</code>	<code>chr15.fa</code>	<code>chr19.fa</code>
<code>chr12.fa</code>	<code>chr16.fa</code>	<code>chr1.fa</code>
<code>chr13.fa</code>	<code>chr17.fa</code>	<code>chr20.fa</code>
	<code>chr21.fa</code>	<code>chr22.fa</code>
	<code>chr23.fa</code>	<code>chr3.fa</code>
	<code>chr4.fa</code>	<code>chr5.fa</code>
	<code>chr6.fa</code>	<code>chr7.fa</code>
	<code>chr8.fa</code>	<code>chr9.fa</code>
	<code>chrX.fa</code>	<code>chrY.fa</code>
 >chr17		
AAGCTTCTCACCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGT		
CCCTGCTGAATGTGCTCTGGGGTCTCTGGGGTCTCACCCACGACCAACTC		
CCTGGGCTGGCACCAGGAGCTTAACAAACATCTGTCCAGCGAATACCT		
GCATCCCTAGAAAGTGAAGCCACCGCCCAAAGACACGCCCATGTCCAGCTT		
AACCTGCATCCCTAGAAAGTGAAGCCACCGCCCAAAGACACGCCCATGTCC		
AGCTTATTCTGCCAGTTCCTCTCCAGAAAGGCTGCATGGTTGACACACA		
GTGcctgcgacaaagctgaatgctatcatttaaaaactccttgctggttt		
gagaggcagaaaatgatattcatagttgctttactttgcatattttAAA		
ATTGTGACTTTCATGGCATAAATAATACTGGTTTATTACAGAAGCACTAG		
...		

La carpeta *Annotation database* contiene todas las pistas utilizadas por el interfaz gráfico del navegador UCSC (mostramos una pequeña parte de éstas en la Figura 44). 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 (ejemplo incluido en la misma imagen):

El fichero *refGene.txt*

Cada línea de éste archivo contiene la anotación suministrada por RefSeq para un transcrito concreto. Las formas alternativas del mismo gen están codificadas en líneas distintas.

Figura 44. La base de datos de anotaciones de UCSC

<u>all_bacends.sql</u>	24-May-2009 11:40	2.1K
<u>all_bacends.txt.gz</u>	24-May-2009 11:40	93M
<u>all_est.sql</u>	12-Sep-2010 16:33	2.3K
<u>all_est.txt.gz</u>	12-Sep-2010 16:34	374M
...		
<u>refGene.sql</u>	10-Oct-2010 20:40	2.0K
<u>refGene.txt.gz</u>	10-Oct-2010 20:40	3.7M
<u>refLink.sql</u>	10-Oct-2010 20:43	1.8K
<u>refLink.txt.gz</u>	10-Oct-2010 20:43	8.3M
...		
<u>xenoRefGene.sql</u>	10-Oct-2010 21:05	2.0K
<u>xenoRefGene.txt.gz</u>	10-Oct-2010 21:05	13M
<u>xenoRefSeqAli.sql</u>	10-Oct-2010 21:14	2.3K
<u>xenoRefSeqAli.txt.gz</u>	10-Oct-2010 21:14	13M


```

1199 NM_178839 chr2 - 80529002 80531487 80529375 80530944 2
80529002,80531276, 80531003,80531487, 0 LRRTM1 cmpl cmpl 0,-1,

```

Esta es la descripción de los atributos pertenecientes a cada registro de este fichero (tomando como ejemplo la línea de la anotación correspondiente al gen *LRRTM1*):

Tabla 7. Atributos del fichero *refGene.txt*

Campo	Ejemplo	Descripción
1	1199	Identificador interno
2	NM_178839	Código asignado por RefSeq
3	chr2	Cromosoma
4	-	Hebra
5	80529002	Inicio del transcrito
6	80531487	Fin del transcrito
7	80529375	Inicio de la región codificante
8	80530944	Fin de de la región codificante
9	2	Número total de exones
10	80529002,80531276,	Coordenadas iniciales de los exones
11	80531003,80531487,	Coordenadas finales de los exones
12	0	Puntuación
13	LRRTM1	Nombre común abreviado del gen
14	cmpl	Anotación del inicio del CDS (completa)
15	cmpl	Anotación del fin del CDS (completa)
16	0,-1,	Pauta de lectura de los exones

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.

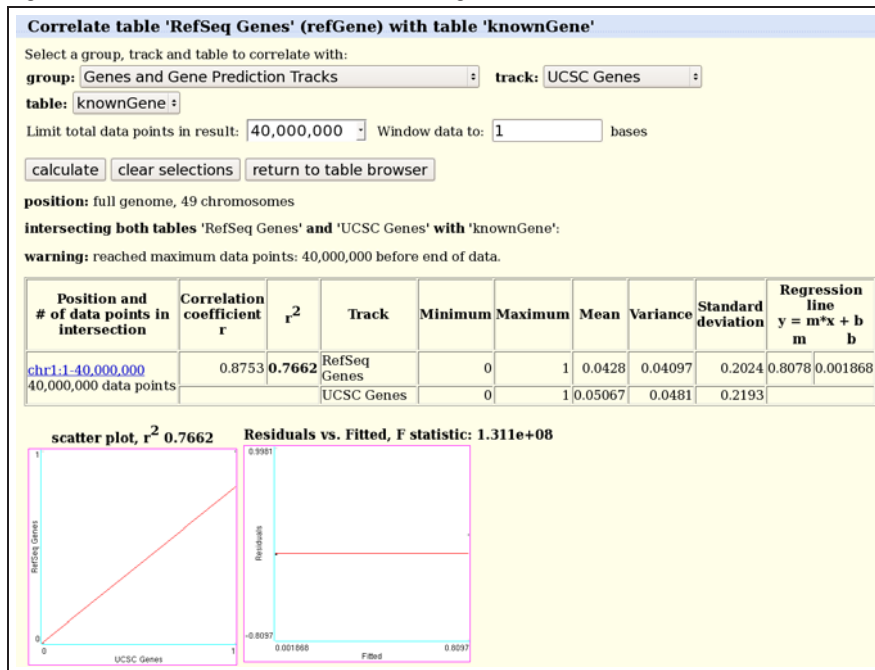
El terminal de la plataforma LINUX es el método más eficiente para analizar distintas pistas de anotaciones. Sin embargo, para aquellos usuarios que no poseen un dominio suficiente para trabajar con soltura en este tipo de entornos, el propio portal de UCSC permite gestionar las consultas enmascarando la complejidad de un gestor de base de datos. La aplicación *Table browser* (en inglés, navegador de tablas), accesible desde el enlace situado dentro de la barra de menú superior de color azul, ofrece un interfaz *web* para interrogar varias pistas simultáneamente. Podemos definir la región genómica donde dirigir nuestras demandas o trabajar con la totalidad de las anotaciones de las pistas. Con este interfaz de comandos podemos instruir al navegador para recuperar datos a gran escala. Por ejemplo, en el marco de la investigación sobre diferentes catálogos de genes, es posible extraer sistemáticamente la región codificante de todos los genes anotados en cualquier especie para llevar a cabo un análisis bioinformático posterior.

Estos conceptos están introducidos en el módulo Fundamentos de informática en entornos bioinformáticos.

Figura 45. El navegador de tablas de UCSC

Esta aplicación 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 la pista UCSC Genes proporcionada independientemente por el navegador UCSC (ver Figura 46). Este tipo de análisis permite establecer el nivel de solapamiento entre pistas de datos similares, evaluando globalmente la precisión de varios sistemas de anotaciones. Con esta aproximación podemos también calcular el grado de similitud entre dos réplicas del mismo experimento de secuenciación.

Figura 46. Cálculo de correlaciones con el navegador de tablas



1.8. 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 47, para acceder a la anotación de un gen el usuario debe seleccionar primero la distribución apropiada del genoma e introducir posteriormente su nombre en la caja de búsqueda.

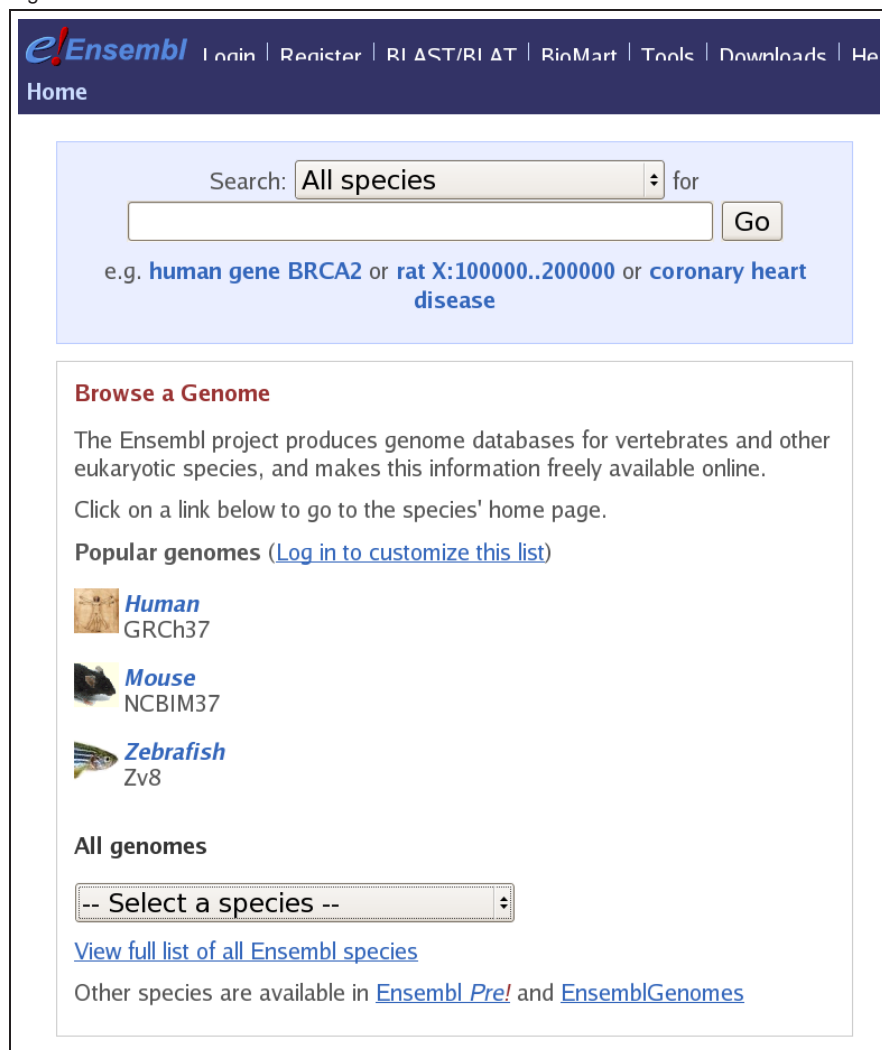
ENSEMBL fue pionero en presentar la información sobre los genes y los transcritos de forma estructurada. Los genes, como unidades que agrupan múltiples ARN mensajeros, reciben un identificador con las iniciales ENSG (*ENSEMBL gene*), mientras que los diferentes productos de éste, transcritos y proteínas, 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 (ver Figura 48). Para modificar el inventario de pistas o realizar cambios en su volcado por pantalla, el usuario debe presionar el botón *Configure this page* (en inglés, configurar esta página). A diferencia del servidor genómico UCSC, que genera imágenes estáticas de las pistas que permiten posteriormente acceder a más información, 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.

Lectura complementaria

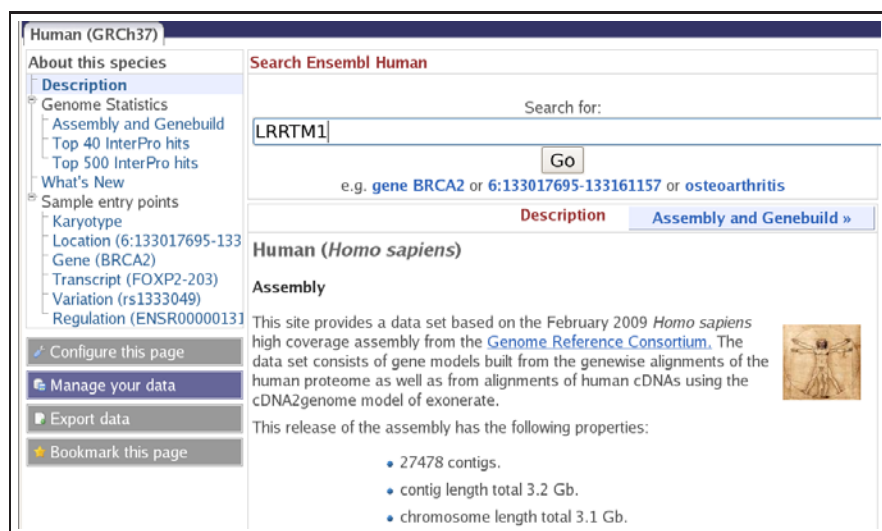
T. Hubbard et al. (2002). *The Ensembl genome database project.* Nucleic Acids Research 2002 30:38-41.

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

Figura 47. Accediendo a las anotaciones en ENSEMBL

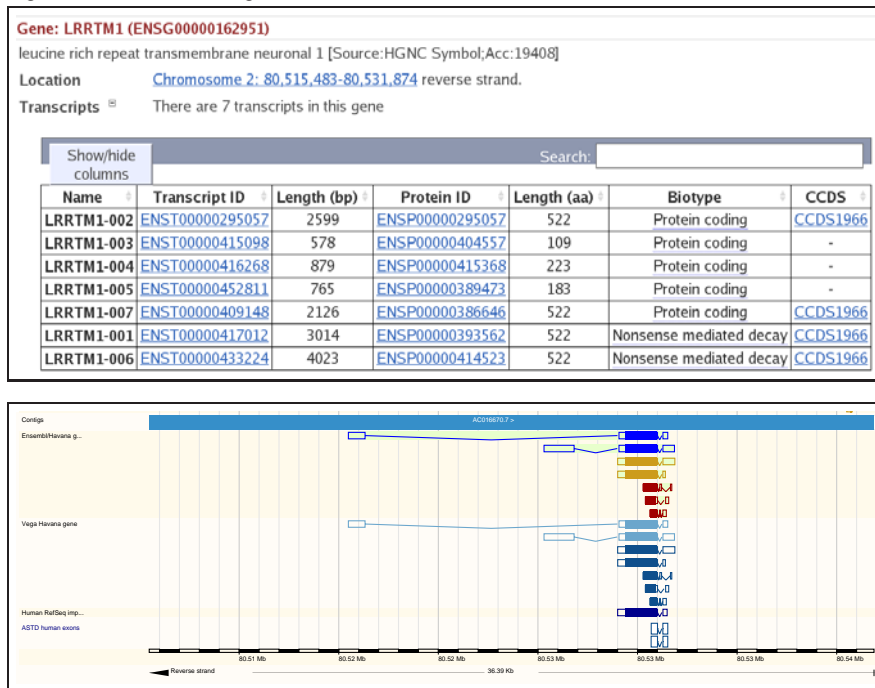


The screenshot shows the Ensembl homepage. At the top is a navigation bar with links: Home, Login, Register, BLAST/BLAT, BioMart, Tools, Downloads, and Help. Below this is a search box with a dropdown menu set to 'All species' and a 'Go' button. Examples of search queries are provided: 'human gene BRCA2', 'rat X:100000..200000', and 'coronary heart disease'. The main content area is titled 'Browse a Genome' and explains that the Ensembl project produces genome databases for vertebrates and other eukaryotic species. It lists 'Popular genomes' with links to Human (GRCh37), Mouse (NCBIM37), and Zebrafish (Zv8). There is also a section for 'All genomes' with a dropdown menu to 'Select a species' and a link to 'View full list of all Ensembl species'. At the bottom, it mentions other species available in 'Ensembl Pre!' and 'EnsemblGenomes'.



The screenshot shows the Ensembl Human (GRCh37) page. On the left is a sidebar with a table of contents: About this species, Description, Genome Statistics, Assembly and Genebuild, Top 40 InterPro hits, Top 500 InterPro hits, What's New, Sample entry points, Karyotype, Location (6:133017695-133017695), Gene (BRCA2), Transcript (FOXP2-203), Variation (rs1333049), and Regulation (ENSR00000133049). The main content area is titled 'Human (Homo sapiens)' and includes a search box for 'Search Ensembl Human' with a 'Go' button. Below the search box are tabs for 'Description' and 'Assembly and Genebuild'. The 'Description' tab is active, showing information about the assembly: 'This site provides a data set based on the February 2009 Homo sapiens high coverage assembly from the Genome Reference Consortium. The data set consists of gene models built from the pairwise alignments of the human proteome as well as from alignments of human cDNAs using the cDNA2genome model of exoneration.' It also lists the properties of the assembly: 27478 contigs, contig length total 3.2 Gb, and chromosome length total 3.1 Gb.

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 opción también existe específicamente un manual de asistencia propio.

Figura 48. Anotación del gen *LRRTM1* en ENSEMBL

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 49, 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 49). Inicialmente, podremos realizar operaciones sobre todos los transcritos anotados por ENSEMBL en el genoma seleccionado (51737 ARN mensajeros en la distribución GRCh3.7 del genoma humano).

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.

Sobre el catálogo de genes de trabajo podemos realizar tres tipos de acciones:

1. (Filtros) Reducir el conjunto de genes de trabajo, focalizando nuestro interés sobre aquellos que posean determinadas características. Podemos definir filtros sobre cromosomas, regiones genómicas, homología entre especies o funciones biológicas. También es posible introducir nuestra propia lista de genes (e.g. los resultados de un experimento de expresión).
2. (Atributos) Seleccionar el tipo de información que nos gustaría obtener del conjunto actual de genes. Podemos establecer los atributos generales de cada gen o definir la información sobre homología con otras especies que deseamos recuperar. Existe la opción de extraer sistemáticamente distintas partes de la estructura exónica de los genes.
3. (Dataset) Combinar las anotaciones actuales con otro genoma distribuido por ENSEMBL para reducir el conjunto inicial de resultados.

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. 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 49).

Figura 49. El interfaz de trabajo de BIOMART

Export all results to: File TSV ☐ Unique results only

Email notification to:

View: rows as ☐ Unique results only

Ensembl Gene ID	Chromosome Name	Gene Start (bp)	Gene End (bp)	Strand	GO Term Name (mf)
ENSG00000229617	GL000228.1	26525	27649	-1	transcription factor activity
ENSG00000217642	GL000228.1	72820	77376	1	transcription factor activity
ENSG00000224558	GL000228.1	79545	80054	1	transcription factor activity
ENSG00000233698	GL000228.1	82851	83360	1	transcription factor activity
ENSG00000228710	GL000228.1	86157	86666	1	transcription factor activity
ENSG00000233294	GL000228.1	89463	89972	1	transcription factor activity
ENSG00000234550	GL000228.1	92769	93278	1	transcription factor activity
ENSG00000232923	GL000228.1	96075	96584	1	transcription factor activity
ENSG00000226866	GL000228.1	99381	99890	1	transcription factor activity
ENSG00000230776	GL000228.1	102687	103196	1	transcription factor activity
ENSG00000233361	GL000228.1	105993	106502	1	transcription factor activity
ENSG00000231089	GL000222.1	84200	84790	-1	transcription factor activity
ENSG00000225958	GL000228.1	109299	109808	1	transcription factor activity
ENSG00000236207	GL000228.1	112605	113114	1	transcription factor activity
ENSG00000215754	GL000225.1	30131	36834	1	transcription factor activity
ENSG00000215754	GL000225.1	30131	36834	1	transcription factor activity
ENSG00000230380	GL000225.1	42803	53334	1	transcription factor activity
ENSG00000233794	GL000225.1	78792	114975	-1	transcription factor activity
ENSG00000184895	Y	2654896	2655740	-1	transcription factor activity
ENSG00000176679	Y	3447096	3448082	1	transcription factor activity

Please select columns to be included in the output and hit 'Results'

☐ Features ☐ Homologs
☐ Structures ☐ Variations
☐ Transcript Event ☒ Sequences

SEQUENCES:

Sequences (max 1)

☐ Unspliced (Transcript) ☐ 5' UTR
☐ Unspliced (Gene) ☐ 3' UTR
☒ Flank (Transcript) ☐ Exon sequences
☐ Flank (Gene) ☐ cDNA sequences
☐ Flank-coding region (Transcript) ☐ Coding sequence
☐ Flank-coding region (Gene) ☐ Protein

Upstream flank

☐ Upstream flank

Downstream flank

☐ Downstream flank

Header Information

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 41.
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.