Bases de Datos

Rodrigo Santamaría

Bases de Datos

Tipos e instituciones

Genes

Proteínas

Genomas

Publicaciones

Formatos

Instituciones

- National Center for Biotechnology Information (**NCBI**)
  - GenBank
- European Bioinformatics Institute (**EBI**)
  - EMBL Nucleotide Sequence Database
- National Institute of Genetics
  - DNA Database of Japan (DDBJ)
- Las tres comparten sus datos diariamente
  - Coordinadas por la International Nucleotide Sequence Database Collaboration (INSDC)

Otras bases de datos

- Hay otras muchas bases de datos (BBDD) que contienen datos sobre secuencias de ADN/proteínas:
  - Específicas de otras instituciones (p. ej. la UCSC o Swiss-Prot)
  - Específicas de un cromosoma u orgánulo
  - Específicas de familias de proteínas (p. ej. Pfam: Protein family database con miles de familias de proteínas homólogas)
  - Específicas de organismo. Por ejemplo:
    - SGD para S. cerevisiae
    - OMIM para H. sapiens y enfermedades
  - …

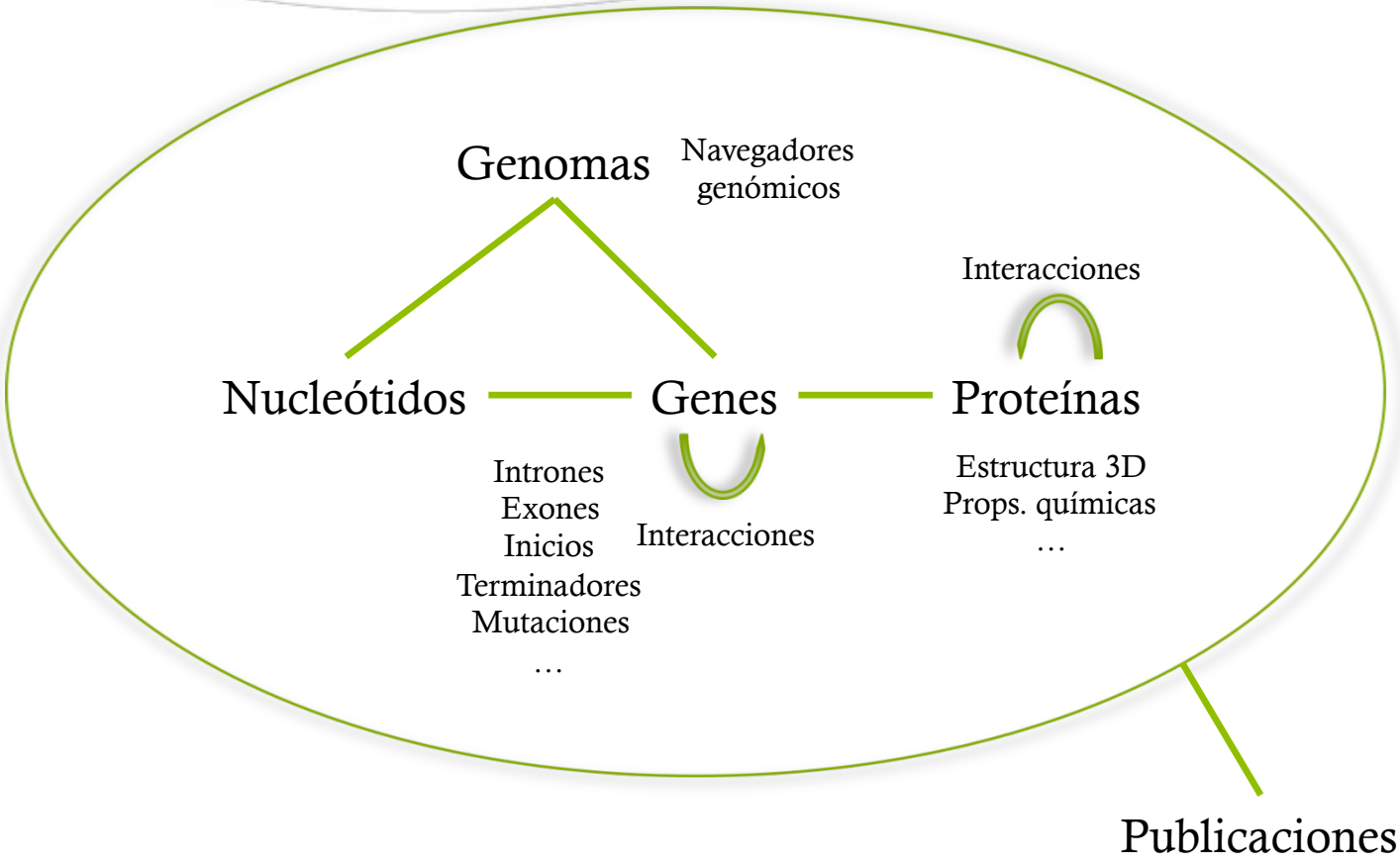
BBDD automáticas y curadas

- BBDD de construcción automática
  - Las entradas en la base de datos son realizadas de manera automática o manual por usuarios no especializados en la BD
    - Crecen rápidamente, pero su contenido no es siempre perfecto
- BBDD curadas
  - Las entradas se revisan a mano por expertos en la BD
    - Crecen más lentamente, pero ofrecen información fiable
- Casi todas las BBDD importantes tienen ambas versiones, o especifican en cada entrada el "grado de fiabilidad"

Información almacenada

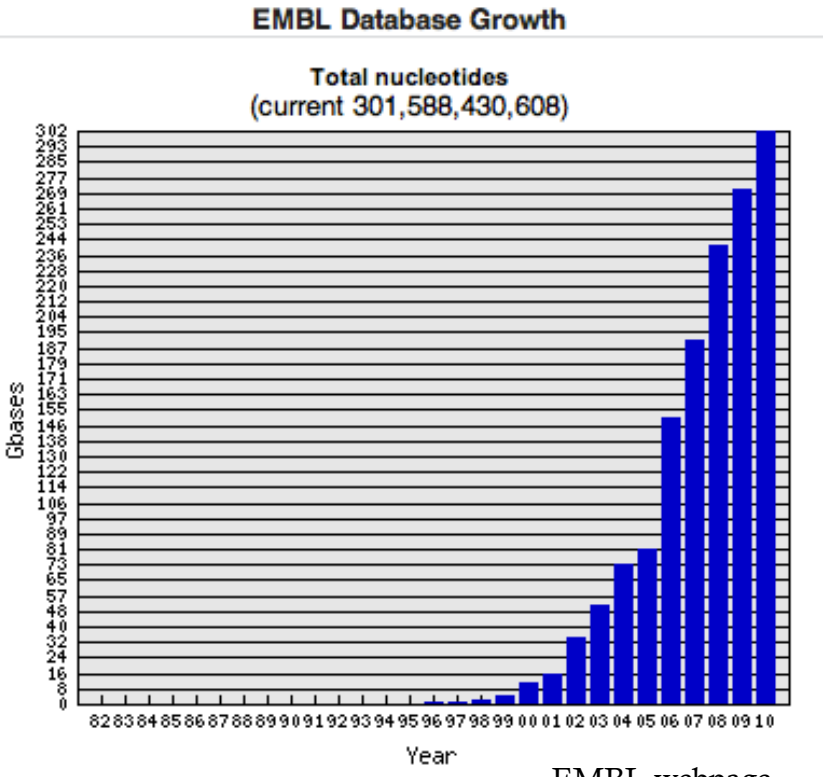
- La información almacenada siempre va a girar entorno a las secuencias, fundamentalmente:
  - **Nucleótidos**: orígenes, secuencias codificantes, genes, etc.
  - **Aminoácidos**: proteínas
  - **Genomas**: secuencias completas para organismos
  - **Publicaciones**: artículos científicos
- Información adicional relacionada con las secuencias
  - Expresión asociada
  - Anotaciones funcionales
  - Relaciones entre secuencias
  - …

Información almacenada: relaciones



Cantidad de información

- GenBank release 183 (abril 2011)
  - 191401393188 pares de bases
    - 191.4 Gbases
  - 200 entradas nuevas cada día
- GenBank y EMBL tienen tamaños y cuotas de crecimiento similares
  - Principalmente porque comparten mucha información



EMBL webpage

Bases de Datos

Tipos e instituciones

Genes

GenBank

Entrez

EMBL

Proteínas

Genomas

Publicaciones

Formatos

GenBank

- Colección anotada de secuencias del NCBI
- Las secuencias pueden ser de diversos tipos y alcances:
  - Secuencia de ADN, ARN, aminoácidos
  - Secuencia de transcrito, gen, cromosoma, genoma
  - Secuencia de mutación (SNP)
  - … hasta 40 BBDD distintas
- **PubMed**: complementa a GenBank con una colección anotada de artículos científicos
- **Entrez** es la herramienta del NCBI para facilitar las búsquedas

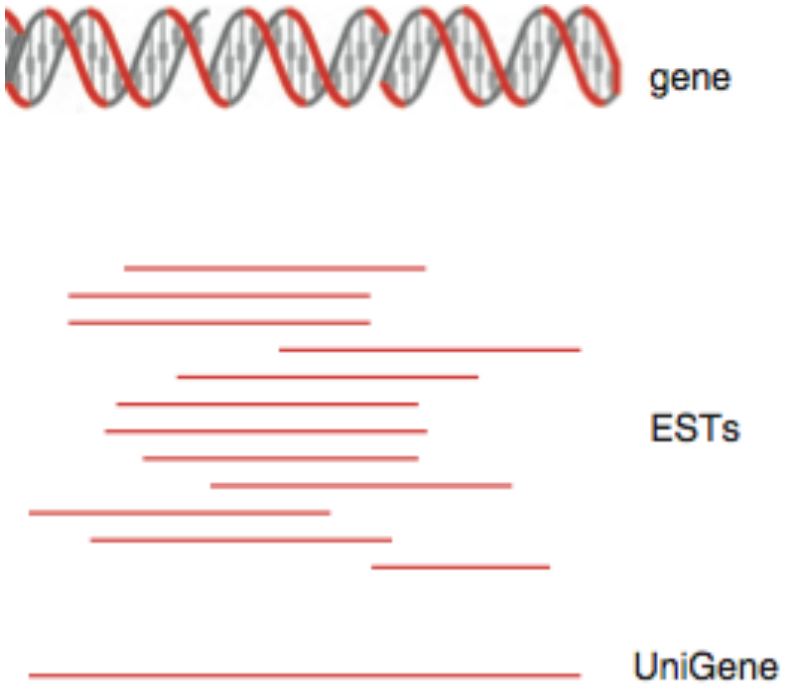

www.ncbi.nlm.nih.gov
Dirección web
NCBI
Resources
How To
My NCBI Sign In
NCBI
National Center for Biotechnology Information
Search
All Databases
Búsqueda en todas las BBDD
beta globin
Search
Clear
NCBI
Entrez, The Life Sciences Search Engine
HOME
SEARCH
SITE MAP
PubMed
All Databases
Human Genome
GenBank
Map Viewer
BLAST
Search across databases
beta globin
GO
Clear
Help
- Result counts displayed in gray indicate one or more terms not found
7993 PubMed: biomedical literature citations and abstracts
7006 PubMed Central: free, full text journal articles
2 Site Search: NCBI web and FTP sites
216 Books: online books
115 OMIM: online Mendelian Inheritance in Man
2179 Nucleotide: Core subset of nucleotide sequence records
2025 EST: Expressed Sequence Tag records
3 GSS: Genome Survey Sequence records
1760 Protein: sequence database
14 Genome: whole genome sequences
304 Structure: three-dimensional macromolecular structures
none Taxonomy: organisms in GenBank
654 SNP: single nucleotide polymorphism
none dbVar: Genomic structural variation
100 Gene: gene-centered information
none SRA: Sequence Read Archive
19 BioSystems: Pathways and systems of interacting molecules
3 HomoloGene: eukaryotic homology groups
3093 GENSAT: gene expression atlas of mouse central nervous system
993 dbGaP: genotype and phenotype
53 UniGene: gene-oriented clusters of transcript sequences
none CDD: conserved protein domain database
11 UniSTS: markers and mapping data
34 PopSet: population study data sets
3160 GEO Profiles: expression and molecular abundance profiles
48 GEO DataSets: experimental sets of GEO data
23 Epigenomics: Epigenetic maps and data sets
none Cancer Chromosomes: cytogenetic databases
61 PubChem BioAssay: bioactivity screens of chemical substances
none PubChem Compound: unique small molecule chemical structures
136 PubChem Substance: deposited chemical substance records
none Protein Clusters: a collection of related protein sequences
1 OMIA: online Mendelian Inheritance in Animals

Bases de Datos de GenBank

- **PubMed**: publicaciones científicas
  - PubMed Central
- **Nucleotide**: secuencias de nucleótidos
  - Gene, EST, UniGene, SNP
- **Protein**: secuencias de aminoácidos
  - Structure
- **Genome**: secuencias genómicas

Expressed Sequence Tags (ESTs)

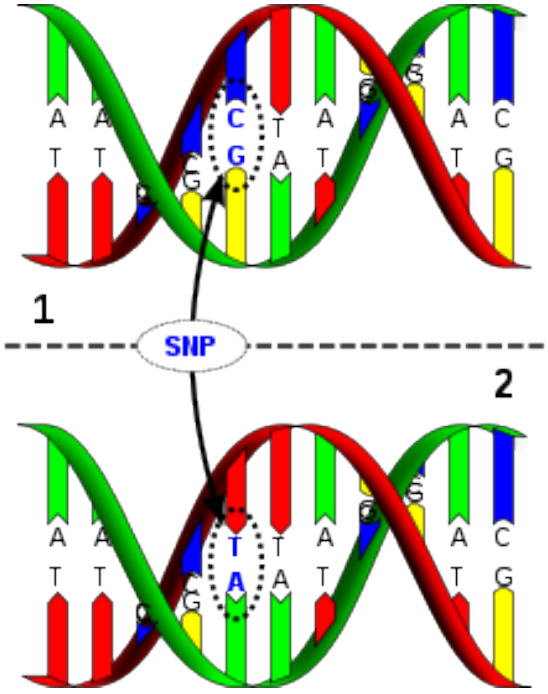
- BD con secuencias de ADN derivadas de secuencias expresadas de ARN
  - cDNA → RNA →DNA (EST)
  - Cada EST normalmente tiene un tamaño entre 300 y 800 bps
  - La secuenciación de ESTs llevó al descubrimiento de muchos genes
- UniGene: BD con clusters de ESTs redundantes



J. Pevsner, *Bioinformatics and Functional Genomics*. 2009

Single Nucleotide Polymorphism (SNP)

- Variación de un solo nucleótido en la cadena de ADN
- Implican el 90% de las mutaciones en humano
- Ocurre un SNP cada 1300 bases (en humano)
- Si ocurre en una región codificante, puede llegar el modificar el aminoácido (SNP nosinónimo) o no (SN sinónimo)



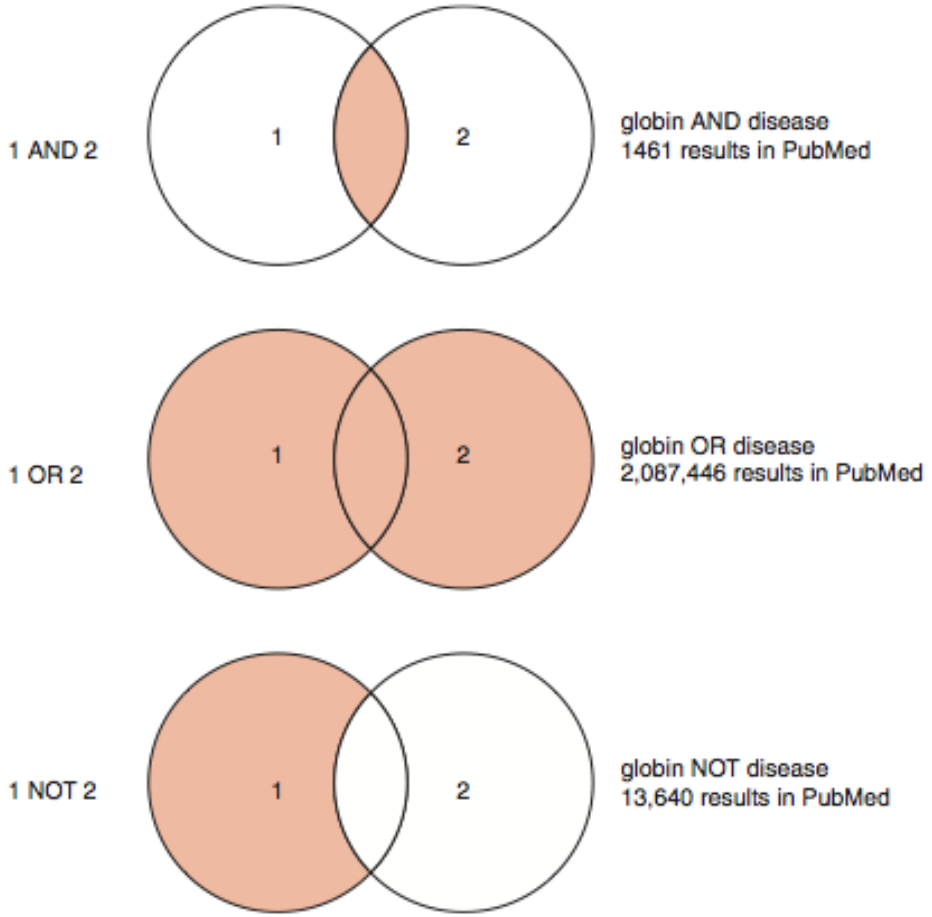
wikipedia

Entrez Gene
búsqueda centrada en genes

- La manera más sencilla de iniciar una búsqueda
  - A partir del nombre del gen (y opcionalmente su organismo)
    - O de una descripción más libre (p.ej. "cáncer de mama")
  - **Entrez Gene** nos dará información sobre
    - Localización cromosómica
    - Transcritos asociados (Nucleotide)
    - Productos génicos (Protein)
    - Artículos relacionados (PubMed) …
  - Ejercicio: Extraer información sobre el gen BRCA1 a partir de Entrez Gene

Búsqueda avanzada en Entrez

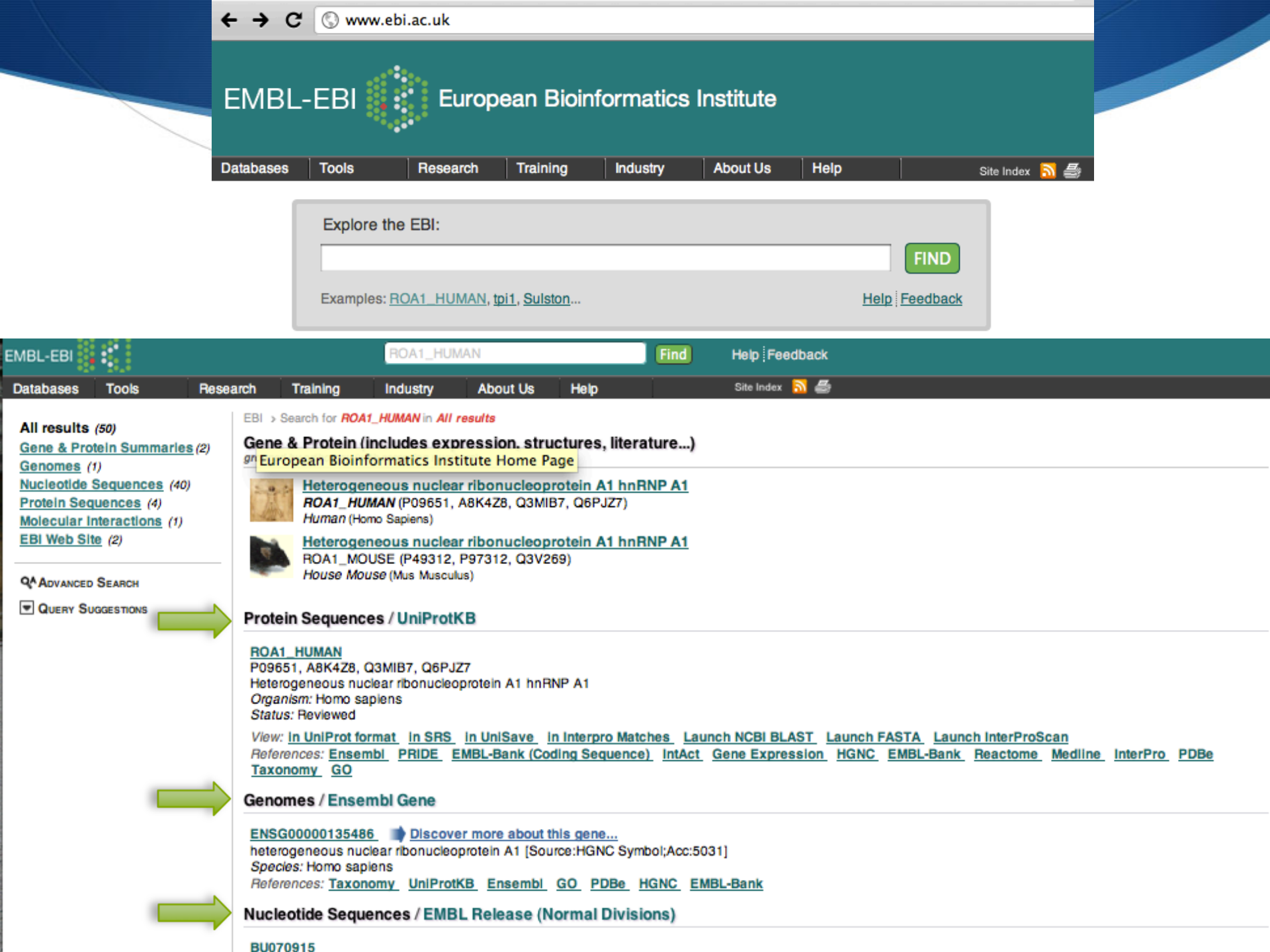
- Uso de **manuales**
  - Fundamental para cualquier herramienta bioinformática
- **Operadores booleanos**: AND, OR, NOT
  - horse OR horses
- **Filtros**: corchetes tras el nombre
  - horse[Organism]
  - BRCA1[Gene Name]
- **Comillas**: para coincidencia exacta si hay más de una palabra
  - "Equus caballus"



J. Pevsner, *Bioinformatics and Functional Genomics*. 2009

Bases de datos del EBI

- Estructura similar a NCBI:
  - Partimos de una búsqueda → filtramos según la base de datos
- Bases de datos principales:
  - Genoma: Ensembl
  - Proteína: UniProtKB
  - Nucleótido: EMBL
- No tiene BD sobre publicaciones


www.ebi.ac.uk
EMBL-EBI
European Bioinformatics Institute
Databases
Tools
Research
Training
Industry
About Us
Help
Site Index
Explore the EBI:
FIND
Examples: ROA1_HUMAN, tpi1, Sulston...
Help
Feedback
EMBL-EBI
ROA1_HUMAN
Find
Help
Feedback
Databases
Tools
Research
Training
Industry
About Us
Help
Site Index
All results (50)
Gene & Protein Summaries (2)
Genomes (1)
Nucleotide Sequences (40)
Protein Sequences (4)
Molecular Interactions (1)
EBI Web Site (2)
Advanced Search
Query Suggestions
EBI › Search for ROA1_HUMAN in All results
Gene & Protein (includes expression, structures, literature...)
European Bioinformatics Institute Home Page
Heterogeneous nuclear ribonucleoprotein A1 hnRNP A1
ROA1_HUMAN (P09651, A8K4Z8, Q3MIB7, Q6PJZ7)
Human (Homo Sapiens)
Heterogeneous nuclear ribonucleoprotein A1 hnRNP A1
ROA1_MOUSE (P49312, P97312, Q3V269)
House Mouse (Mus Musculus)
Protein Sequences / UniProtKB
ROA1_HUMAN
P09651, A8K4Z8, Q3MIB7, Q6PJZ7
Heterogeneous nuclear ribonucleoprotein A1 hnRNP A1
Organism: Homo sapiens
Status: Reviewed
View: In UniProt format In SRS In UniSave In Interpro Matches Launch NCBI BLAST Launch FASTA Launch InterProScan
References: Ensembl PRIDE EMBL-Bank (Coding Sequence) IntAct Gene Expression HGNC EMBL-Bank Reactome Medline InterPro PDBe Taxonomy GO
Genomes / Ensembl Gene
ENSG00000135486
Discover more about this gene...
heterogeneous nuclear ribonucleoprotein A1 [Source:HGNC Symbol;Acc:5031]
Species: Homo sapiens
References: Taxonomy UniProtKB Ensembl GO PDBe HGNC EMBL-Bank
Nucleotide Sequences / EMBL Release (Normal Divisions)
BU070915

EBI vs NCBI

- Ambas son complementarias
  - Comparten información
  - Cada vez las referencias de una a la otra son más frecuentes
- En ambos casos comenzamos con datos crudos
  - Que se organizan, analizan y muestran de forma diferente
- Datos crudos (BBDD primarias) y datos curados y anotados por expertos (BBDD secundarias)
- Es recomendable en estudios serios explorar la riqueza de recursos disponible en ambas
  - Explotando los métodos de búsqueda avanzada

Bases de Datos

Tipos e instituciones
Genes
Proteínas
UniProt
ExPASy
Nº de acceso
Genomas
Publicaciones
Formatos

UniProt

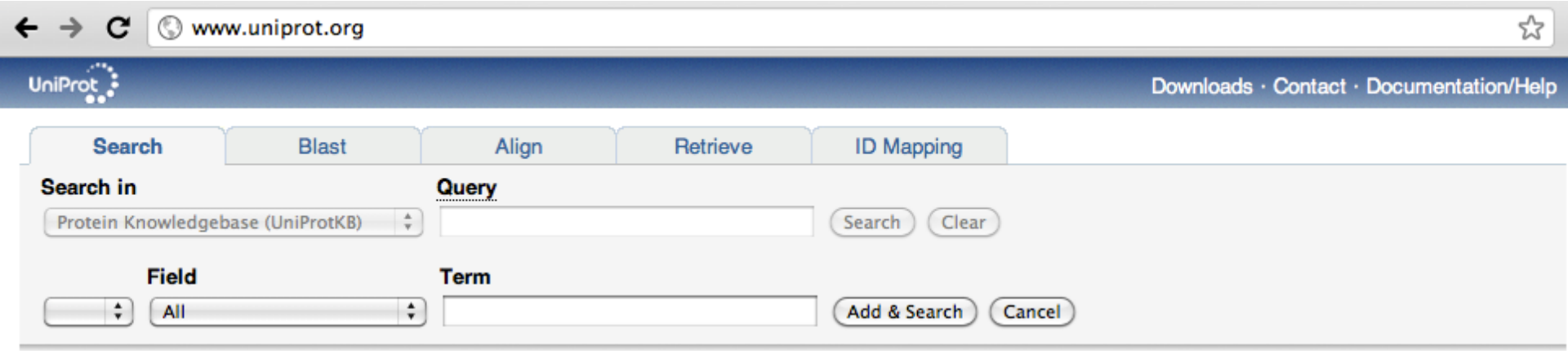
- UniProtKB es la BD más utilizada para búsquedas centradas en proteínas
  - El equivalente a GenBank para búsquedas centradas en genes.
- UniProt es un esfuerzo de unificación de tres bases de datos:
  - **Swiss-Prot**: la BD de proteínas mejor anotada por expertos
  - Translated EMBL (**TrEMBL**): proteínas que no están en Swiss-Prot, encontradas automáticamente
  - **Protein Sequence Database** (PSD): BD complementaria de proteínas anotadas por expertos del Protein Information Resource (PIR)

UniProt

- UniProt consta de tres componentes:
  - UniProt Knoweledgebase (**UniProtKB**) comprende
    - Swiss-Prot/PSD: bases de datos anotadas y revisadas manualmente
    - TrEMBL: base de datos anotada automáticamente y NO revisada
  - **UniRef**: clusters de proteínas similares para acelerar búsquedas
    - 50%, 90% ó 100% de similitud
  - **UniParc**: archivo estable y no redundante de secuencias de proteínas obtenidas de una gran variedad de fuentes.

ExPASy

- **Ex**pert **P**rotein **A**nalysis **Sy**stem
  - Compendio de herramientas de análisis en proteómica
  - Y para la búsqueda/recuperación de datos
  - Búsqueda avanzada: mediante una caja desplegable permite especificar filtros y operaciones booleanas



Números de acceso

- Propiedad esencial de las BBDD de secuencias
  - En general, de cualquier BD
- Nº de acceso: cadena de 4 a 12 números y/o caracteres alfabéticos asociados con una entrada de secuencia en la BD
  - También pueden identificar un experimento de expresión génica, una estructura de proteína, etc.
- Para una molécula determinada (p. ej. beta globina) puede haber cientos de números de acceso
  - Distintas proteínas homólogas
  - Distintos nombres para la misma proteína (sinónimos) → redundancia
  - Misma proteína para distintos organismos …

RefSeq y Ensembl

- Puede haber cientos de números de acceso distintos para un mismo gen
  - Las bases de datos son altamente redundantes
- Los proyectos RefSeq (NCBI) y Ensembl (EBI) tratan de mantener identificadores únicos para cada gen o producto génico, independientemente del nº de secuencias asociadas
  - Ejemplo: mioglobina humana en RefSeq
    - Tiene tres variantes: NM_005368, NM_203377, NM_2003378
    - Que dan lugar a tres proteínas: NP_005359, NP_976311, NP_976312

Números de acceso: ejemplos

Tipo de registros	Formato del número de acceso
Secuencia de nucleótidos de GenBank/EMBL/DDBJ	Una letra y 5 dígitos: X02775 Dos letras y 6 dígitos: AF025334
Secuencia de proteínas de SwissProt	Normalmente una letra y 5 dígitos: P12345
Secuencia de nucleótidos de RefSeq	Dos letras y seis dígitos separados por una línea: NM_006744, NT_008769
Secuencia de proteínas de RefSeq	Dos letras (NP) y seis dígitos separados por una línea: NP_006735
Secuencia de Ensembl	ENS+letra+11 dígitos → ENSG00000333504 La letra es P para proteína, T para transcrito, G para gen, etc.

Bases de Datos

Tipos e instituciones

Genes

Proteínas

Genomas

Navegadores de Genoma

UCSC Browser

Publicaciones

Formatos

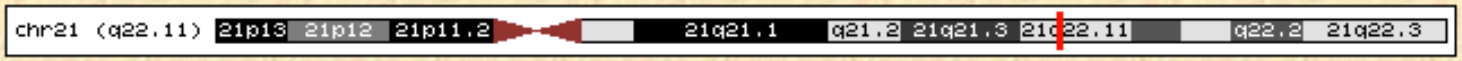
Navegadores de genoma

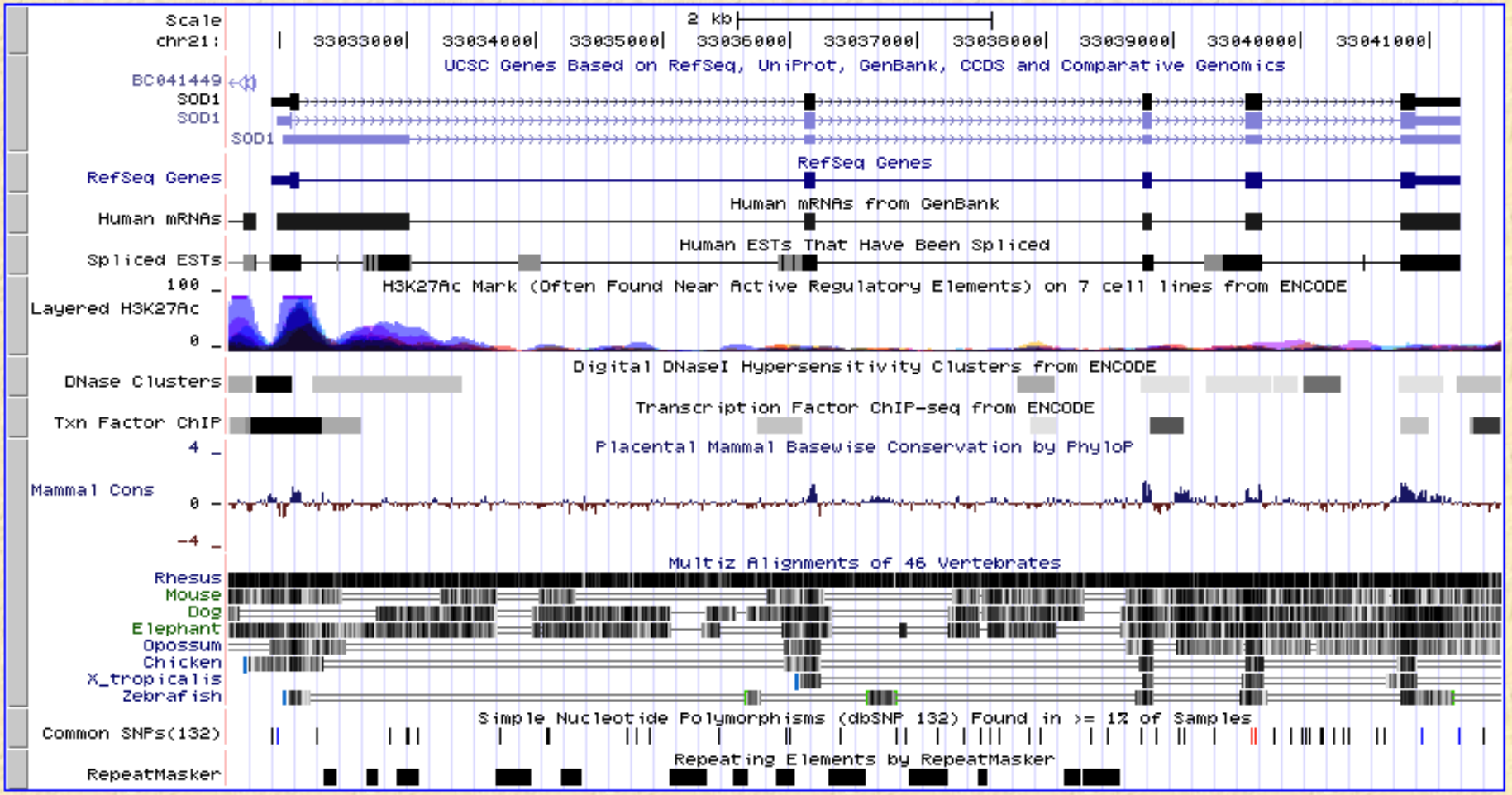
- **Navegador de genoma**: BD con una interfaz gráfica para representar secuencias y otros datos en función de su posición en los cromosomas
- Tres navegadores principales
  - NCBI Genome Browser
  - → Univ. de California, Santa Cruz (UCSC) Genome Browser
  - Ensembl Genome Browser
- Ejercicio: buscar e inspeccionar BRCA1 en UCSC

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

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

position/search chr21:33,031,597-33,041,570 gene jump clear size 9,974 bp. configure *2011 ENCODE Usability Survey*





move start < 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks.

move end < 2.0 >

Bases de Datos

Tipos e instituciones

Genes

Proteínas

Genomas

Publicaciones

PubMed

Medline

Formatos

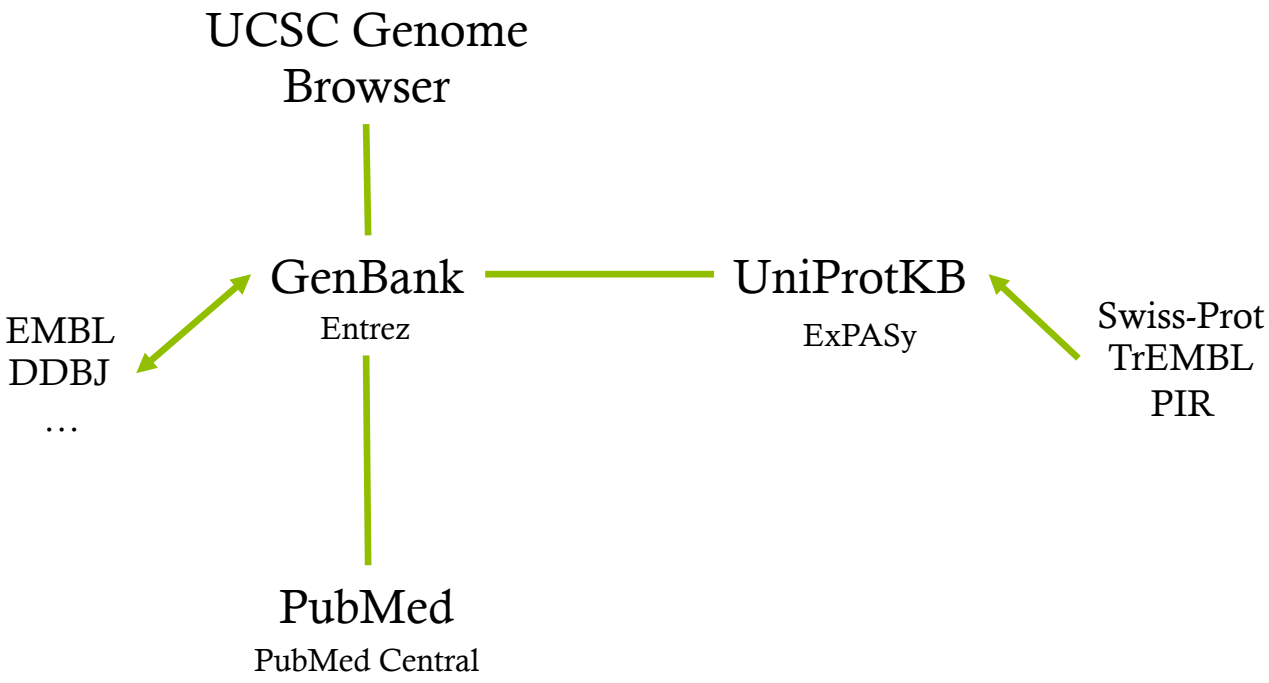
PubMed

- Acceso gratuito a citas de literatura biomédica desde 1965
  - Los editores de las revistas que participan en PubMed envían electrónicamente sus citas al NCBI antes o en el momento de su publicación
- MEDLINE: base de datos de citas y abstracts (resúmenes) de acceso online y gratuito, de carácter médico
  - Creado y administrado por la US National Library of Medicine (NLM)
  - Registros indexados por un vocabulario controlado (Medical Subject Headings – MeSH)

PubMed

- PubMed contiene las referencias de MEDLINE
  - Y referencias de revistas que no están en MEDLINE pero que son también revisadas por la NLM
- PMID: identificador de la referencia en PubMed
- PubMed Central (PMC): base de datos con artículos científicos completos y gratuitos
  - Gestionada por el NLM también
  - Actualmente contiene unos 2.2 millones de artículos

Resumen de las BBDD más usadas



NCBI

NIG

DDBJ

EMBL-EBI

GenBank Entrez

EBI

relacionadas

Gene

Protein

EMBL

SwissProt, TrEBML

unificación de nº de acceso

RefSeq

Genome Browser

Ensembl

Genome Browser

PubMed

PIR

UniProt

ExPASy

UCSC

Genome Browser

Institución

Compendios

Base de Datos

Motores de búsqueda

Recurso muy usado

Bases de Datos

Tipos e instituciones

Genes

Proteínas

Genomas

Publicaciones

Formatos

- FASTA
- GFF

Formatos

- Los formatos para compartir secuencias más exitosos son en **texto plano**
  - Parte del éxito de EMBL y GenBank
  - Su potencia es su facilidad de uso
    - Legibilidad
    - Facilidad de parseo y manipulación
- Aunque otros formatos (por ejemplo, XML) pueden ser más eficientes computacionalmente y correctos estructuralmente

Formatos

- Enorme variedad de formatos para representar datos de secuencia
  - Cada institución/grupo de investigación generaba sus propios formatos antes de pensar en la estandarización
- A día de hoy los estándares son los de las grandes instituciones
  - Formato GenBank (adoptado por EMBL y DDBJ)
- Y los más genéricos y sencillos (FASTA)

Códigos IUPAC

- Estándares sobre la representación de nucleótidos y aminoácidos

```
A = adenine
C = cytosine
G = guanine
T = thymine
U = uracil
R = G A (purine)
Y = T C (pyrimidine)
K = G T (keto)
M = A C (amino)
S = G C
W = A T
B = G T C
D = G A T
H = A C T
V = G C A
N = A G C T (any)
```

Amino Acid Code:	Three letter Code:	Amino Acid:
A	Ala	Alanine
B	Asx	Aspartic acid or Asparagine
C	Cys	Cysteine
D	Asp	Aspartic Acid
E	Glu	Glutamic Acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
X	Xaa	Any amino acid
Y	Tyr	Tyrosine
Z	Glx	Glutamine or Glutamic acid

FASTA

- Formato en texto plano para representación de secuencias
- Un fichero FASTA tiene una o más secuencias
- Cada secuencia está formada por dos líneas:

La primera línea es un comentario sobre la secuencia, comienza por ">"

```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDF
PEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADI
DGDGQVNYEEFVQMMTAK
```

La segunda línea es la secuencia en sí, usando el código estándar para aminoácidos y nucleótidos

FASTA

- Algunas consideraciones
  - No hay espacio entre “>” y la primera letra del comentario
  - Cada línea termina con un salto de línea
    - La línea de comentario sólo ocupa una línea
    - La línea de secuencia ocupa hasta la próxima línea que sea comentario
  - Se recomienda que las líneas tengan como máximo 80 caracteres
  - La extensión de un fichero FASTA genérico es .fasta
    - A veces se usa .fa o .fsa

Formato GenBank (GBFF)

- Formato compartido también por EMBL y DDBJ
  - Con pequeñas diferencias, sobre todo en la cabecera
- Información más detallada de cada secuencia
  - **Cabecera**: información sobre la secuencia
    - Identificadores, versión, fuente biológica, referencia, etc.
  - **Características**
    - Para cada sección de la secuencia:
      - comienzo, fin, longitud, dirección, tipo, cadena…

Cabecera

```
LOCUS       LISOD                   756 bp    DNA     linear   BCT 30-JUN-1993
DEFINITION  Listeria ivanovii sod gene for superoxide dismutase.
ACCESSION   X64011 S78972
VERSION     X64011.1  GI:44010
KEYWORDS    sod gene; superoxide dismutase.
SOURCE      Listeria ivanovii
  ORGANISM  Listeria ivanovii
            Bacteria; Firmicutes; Bacillales; Listeriaceae; Listeria.
REFERENCE   1  (bases 1 to 756)
  AUTHORS   Haas,A. and Goebel,W.
  TITLE     Cloning of a superoxide dismutase gene from Listeria ivanovii by
            functional complementation in Escherichia coli and characterization
            of the gene product
  JOURNAL   Mol. Gen. Genet. 231 (2), 313-322 (1992)
  MEDLINE   92140371
REFERENCE   2  (bases 1 to 756)
  AUTHORS   Kreft,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-APR-1992) J. Kreft, Institut f. Mikrobiologie,
            Universitaet Wuerzburg, Biozentrum Am Hubland, 8700 Wuerzburg, FRG
[...]
```

Fuente biológica

Referencia

Cabecera

- Es la sección que más varía entre las distintas BBDD
- LOCUS

```
LOCUS       LISOD      756 bp    DNA     linear   BCT 30-JUN-1993
```

nombre | longitud | tipo | división GenBank | fecha modificación

- DEFINITION
  - Resumen de la "biología" del registro en texto libre
  - Es la línea equivalente a la descripción en formato FASTA

```
DEFINITION  Listeria ivanovii sod gene for superoxide dismutase.
```

Cabecera

- ACCESSION
  - Clave primaria para referenciar un registro
  - Número que es citado en las publicaciones
- VERSION
- KEYWORDS
  - Vestigio "histórico"
  - Vocabulario no controlado
  - La política es no incluirlos

```
ACCESSION   X64011 S78972

VERSION     X64011.1  GI:44010

KEYWORDS    sod gene; superoxide
dismutase.
```

Cabecera

- SOURCE/ORGANISM
  - Nombre científico. Nombre común opcional en SOURCE
  - Taxonomía opcional en ORGANISM

```
  SOURCE      Listeria ivanovii
    ORGANISM  Listeria ivanovii
              Bacteria; Firmicutes; Bacillales; Listeriaceae; Listeria.
```

- REFERENCE

```
REFERENCE   1  (bases 1 to 756)
  AUTHORS   Haas,A. and Goebel,W.
  TITLE     Cloning of a superoxide dismutase gene from Listeria ivanovii by
            functional complementation in Escherichia coli and characterization
            of the gene product
  JOURNAL   Mol. Gen. Genet. 231 (2), 313-322 (1992)
  MEDLINE   92140371
REFERENCE   2  (bases 1 to 756)
  AUTHORS   Kreft,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-APR-1992) J. Kreft, Institut f. Mikrobiologie,
            Universitaet Wuerzburg, Biozentrum Am Hubland, 8700 Wuerzburg, FRG
```

→ Una o más referencias

→ La última referencia es la responsable del envío

Características (features)

- Es la sección más importante
- Cada característica representa una secuencia de algún tipo
  - La característica más importante es source
    - Debe aparecer siempre
    - Debe contener obligatoriamente la localización y los atributos organism y db_xref (referencia a su id taxonómico)
  - Algunas otras características son:

Key	Description
CDS	Protein-coding sequence
RBS	ribosome binding site
rep_origin	Origin of replication
protein_bind	Protein binding site on DNA
tRNA	mature transfer RNA

Sección de características

```
[...]
FEATURES             Location/Qualifiers
     source          1..756
                     /organism="Listeria ivanovii"
                     /strain="ATCC 19119"
                     /db_xref="taxon:1638"
                     /mol_type="genomic DNA"
     RBS             95..100
                     /gene="sod"
     gene            95..746
                     /gene="sod"
     CDS             109..717
                     /gene="sod"
                     /EC_number="1.15.1.1"
                     /codon_start=1
                     /transl_table=11
                     /product="superoxide dismutase"
                     /db_xref="GI:44011"
                     /db_xref="GOA:P28763"
                     /db_xref="InterPro:IPR001189"
                     /db_xref="UniProtKB/Swiss-Prot:P28763"
                     /protein_id="CAA45406.1"
                     /translation="MTYELPKLPYTYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVS
                     GHAELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGGAPTGNLK
                     AAIESEFGTFDEFKEKFNAAAAARFGSGWAWLVVNNGKLEIVSTANQDSPLSEGKTPV
                     LGLDVWEHAYYLKFQNRRPEYIDTFWNVINWDERNKRFDAAK"
     terminator      723..746
                     /gene="sod"
ORIGIN
        1 cgttatttaa ggtgttacat agttctatgg aaatagggtc tatacctttc gccttacaat
       61 gtaatttctt ..........
//
```

Característica

Localización

Atributos

Características

- Para cada característica se da su localización y una serie de atributos, con el formato:

```
característica    comienzo..fin
                  /atributo1=“valor1”
                  ...
                  /atributoN=“valorN”
```

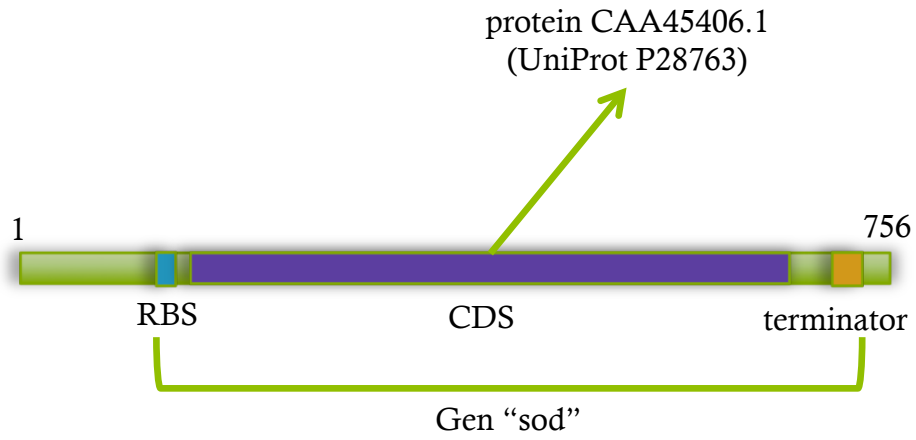
```
source     1..756
             /organism="Listeria ivanovii"
             /strain="ATCC 19119"
             /db_xref="taxon:1638"
             /mol_type="genomic DNA"
```

- La localización puede ser:
  - Completa: 687..3158
  - Parcial sobre el extremo 5’: <1..206
  - Parcial sobre el extremo 3’: 4821..5028>
  - La cadena complementaria: complement(3300..4037)

Características

- Algunos de los atributos (o calificadores) más importantes:
  - /organism – nombre del organismo de la secuencia
  - /gene – nombre del gen relacionado con la secuencia
  - /product – producto génico de la secuencia
  - /db_xref – referencia cruzada a otra base de datos
  - /direction – dirección de la replicación del ADN
  - /codon_start – primera base del primer codón completo (1,2 ó 3)
- Hay muchos más, consultad el manual para más detalle
  - http://www.ncbi.nlm.nih.gov/collab/FT/

```
[...]
FEATURES             Location/Qualifiers
     source          1..756
                     /organism="Listeria ivanovii"
                     /strain="ATCC 19119"
                     /db_xref="taxon:1638"
                     /mol_type="genomic DNA"
     RBS             95..100
                     /gene="sod"
     gene            95..746
                     /gene="sod"
     CDS             109..717
                     /gene="sod"
                     /EC_number="1.15.1.1"
                     /codon_start=1
                     /transl_table=11
                     /product="superoxide dismutase"
                     /db_xref="GI:44011"
                     /db_xref="GOA:P28763"
                     /db_xref="InterPro:IPR001189"
                     /db_xref="UniProtKB/Swiss-Prot:P28763"
                     /protein_id="CAA45406.1"
                     /translation="MTYELPKLPYTYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVS
                     GHAELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGGAPTGNLK
                     AAIESEFGTFDEFKEKFNAAAAARFGSGWAWLVVNNGKLEIVSTANQDSPLSEGKTPV
                     LGLDVWEHAYYLKFQNRRPEYIDTFWNVINWDERNKRFDAAK"
     terminator      723..746
                     /gene="sod"
ORIGIN
        1 cgttatttaa ggtgttacat agttctatgg aaatagggtc tatacctttc gccttacaat
       61 gtaatttctt ..........
//
```



FASTA y NCBI

- Las secuencias de GenBank pueden exportarse en FASTA
- La línea de cabecera de un fichero FASTA de GenBank es:

```
>gi|44010|emb|X64011.1| Listeria ivanovii sod gene for superoxide dismutase
>gi|ID|DATABASE|VERSION|DEFINITION
```

- La línea varía ligeramente según la base de datos de la que viene
  - Se mantiene estable en GenBank (gi), EMBL (emb) y DDBJ (dbj)
- Más info en
  - http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/formatdb_fastacmd.html

Preguntas para debate

- ¿Te parece sencillo el uso de las bases de datos? ¿Eres capaz de encontrar la información que buscas?
- ¿Cómo consideras la información encontrada? ¿Insuficiente? ¿Demasiado extensa? ¿Precisa?
- ¿Eres capaz de discernir lo que buscas entre los resultados de tu búsqueda? ¿Cómo crees que se refinan las bases de datos?
- ¿Has detectado errores en los resultados de la búsqueda? ¿Cómo crees que se solucionan estos errores?

Lecturas de apoyo

- Pevsner, 2009. Ch 2 *Access to Sequence Data and Literature Information.*
- http://www.ncbi.nlm.nih.gov/
  - Especialmente la sección "Get Started"
    - Tutoriales, ayuda sobre búsqueda avanzada, etc.
- Entrez (Maglott et al., 2004):
  - http://nar.oxfordjournals.org/content/33/suppl_1/D54.short
- RefSeq (Pruitt y Maglott, 2001)
  - http://nar.oxfordjournals.org/content/29/1/137.short

DNA Art es una empresa que, a partir de una muestra de saliva, analiza tu ADN y lo convierte en un objeto de arte

www.dna11.com