Bioinformática y análisis de datos ómicos

Tema 2: Gestión de Datos de Origen Biológico

Ignacio Varela Egocheaga
Departamento de Biología Molecular

Tema 2: Gestión de datos de origen biológico

2.1 Conceptos básicos de biología molecular.

2.2 Formatos comunes de archivos de información biológica

2.3 Bases de datos de información biomédica.

2.4 Acceso a las bases de datos.

Tema 2: Gestión de datos de origen biológico

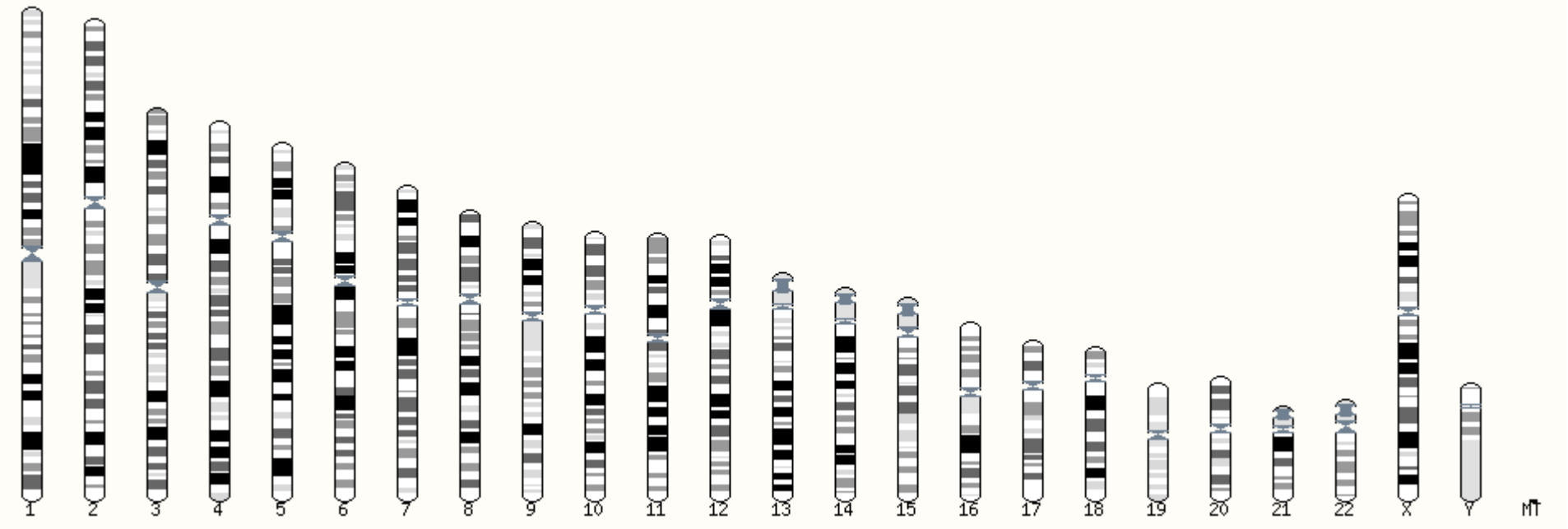
2.1 Conceptos básicos de biología molecular.

2.2 Formatos comunes de archivos de información biológica

2.3 Bases de datos de información biomédica.

2.4 Acceso a las bases de datos.

DNA/RNA en las células humanas



http://www.ensembl.org/Homo_sapiens/Location/Genome

El genoma humano está compuesto por 3.2 Gb divididas en 23-24 moléculas distintas (contenidas en **CROMOSOMAS**) y de las cuales tenemos 2 copias (**DIPLOIDE**).

La secuencia de cada individuo está compuesto por una mezcla de la secuencia materna y paterna y es única. De media cada individuo contiene alrededor de 3 millones de cambios frente a otro individuo.

Los cambios/mutaciones pueden ser heredados (**GERMINALES**) o producidos durante el desarrollo del individuo (**SOMÁTICOS**). De cada región del genoma tenemos 2 copias o **ALELOS**. Así un individuo para un determinado **LOCUS** puede ser **HOMOCIGOTO** (2 alelos iguales) o **HETEROCIGOTO** (2 alelos distintos).

Dogma central de la biología

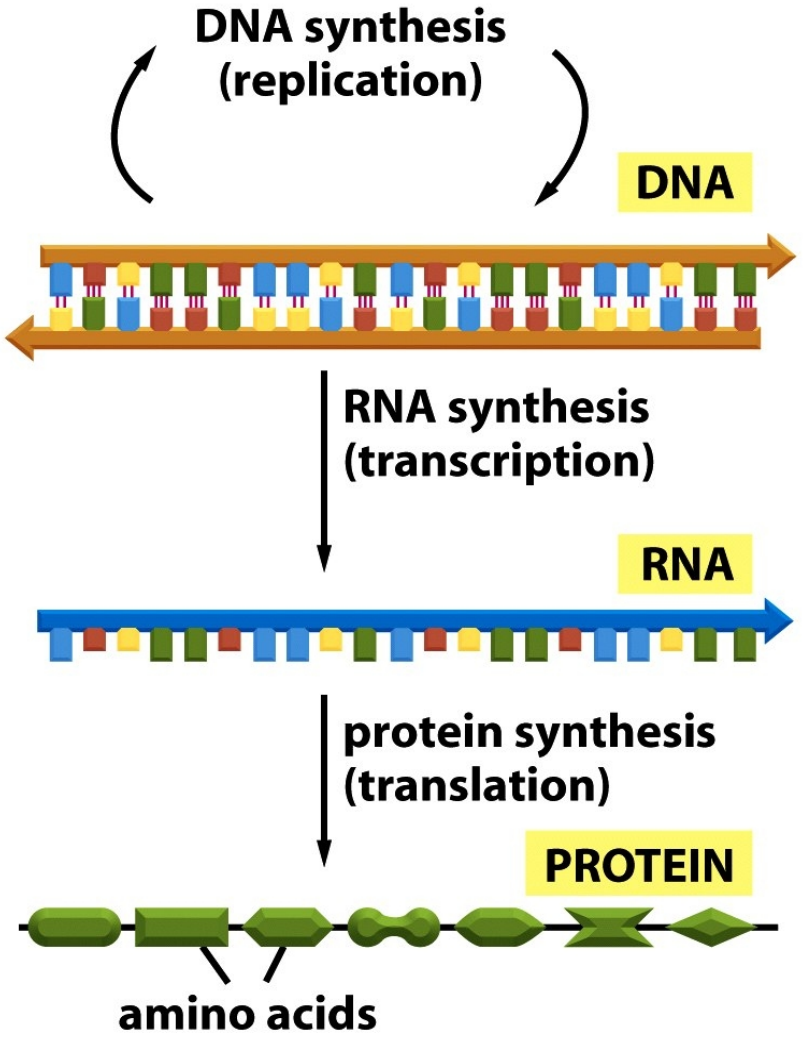
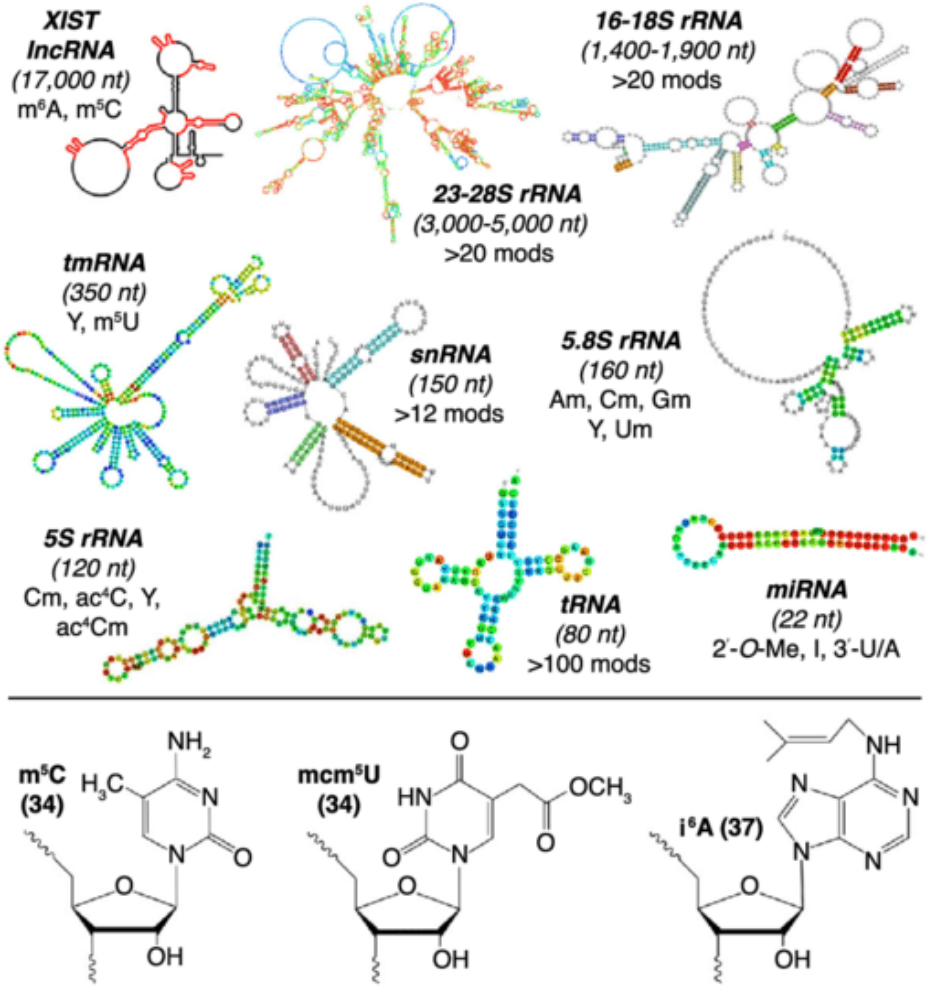


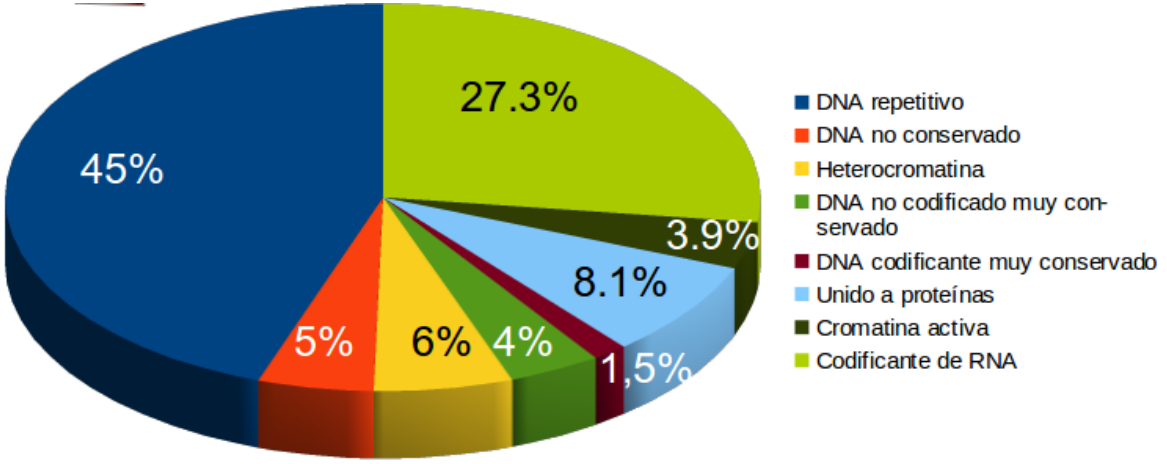
Figure 1-4 Molecular Biology of the Cell 5/e (© Garland Science 2008)



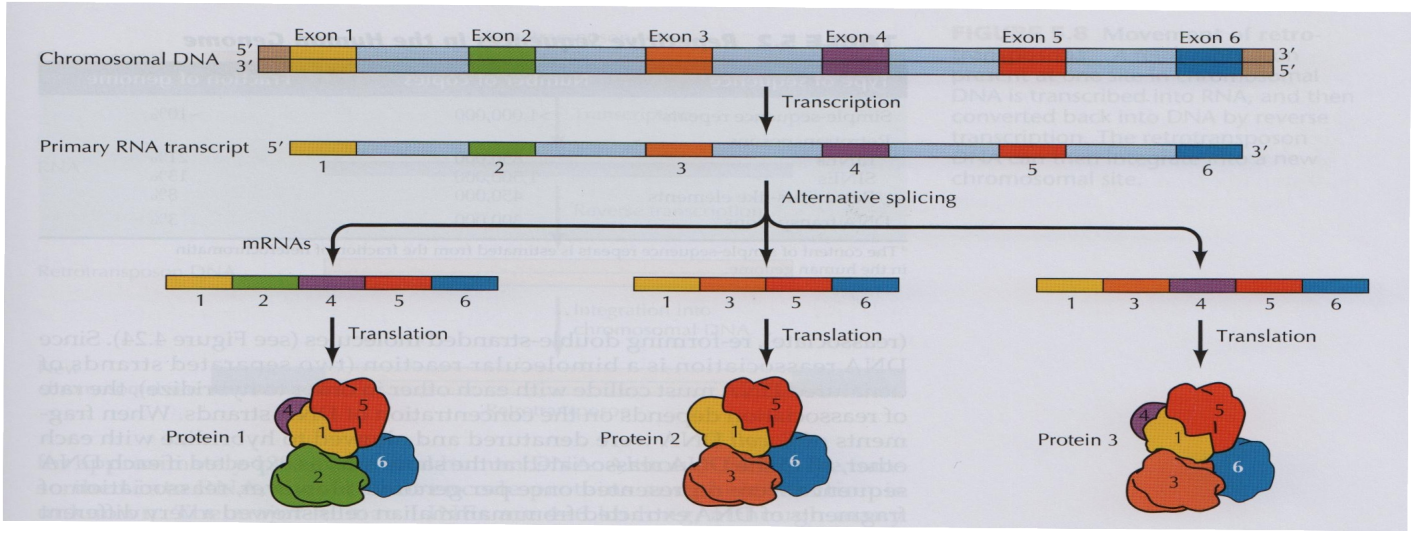
Chan C *et al.* Genome Biol. 19:228

DNA/RNA en las células humanas

Mayoría de DNA no codificante

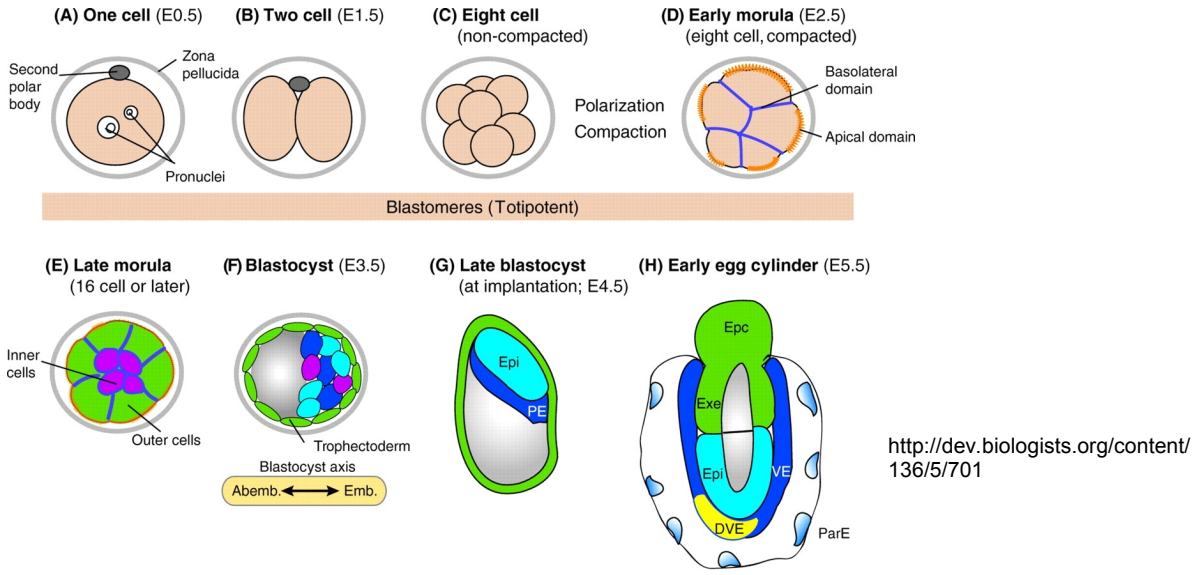


Los GENES son las regiones del DNA que producen RNA

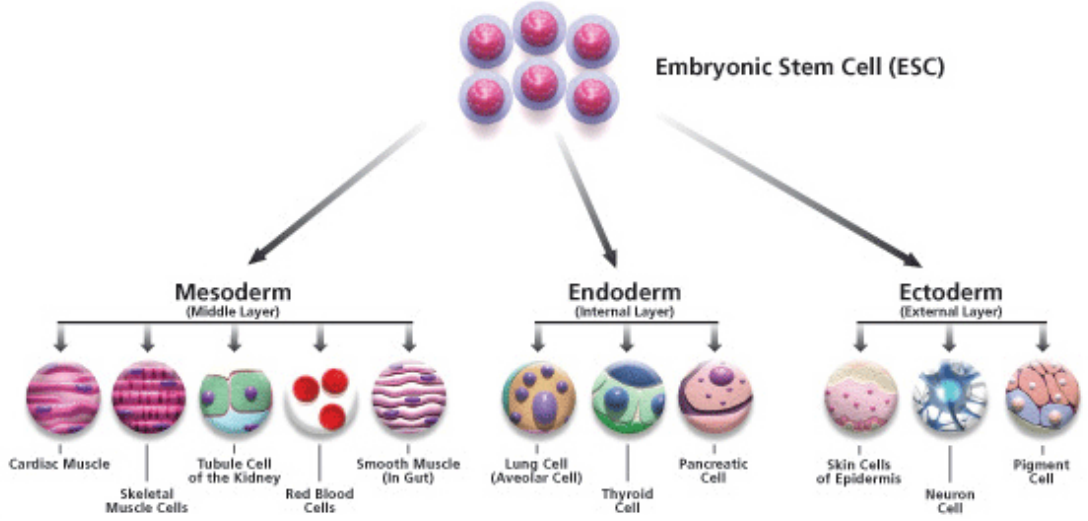


The Cell, A molecular approach. e5. Sinauer 2009

Desarrollo de los distintos linajes celulares



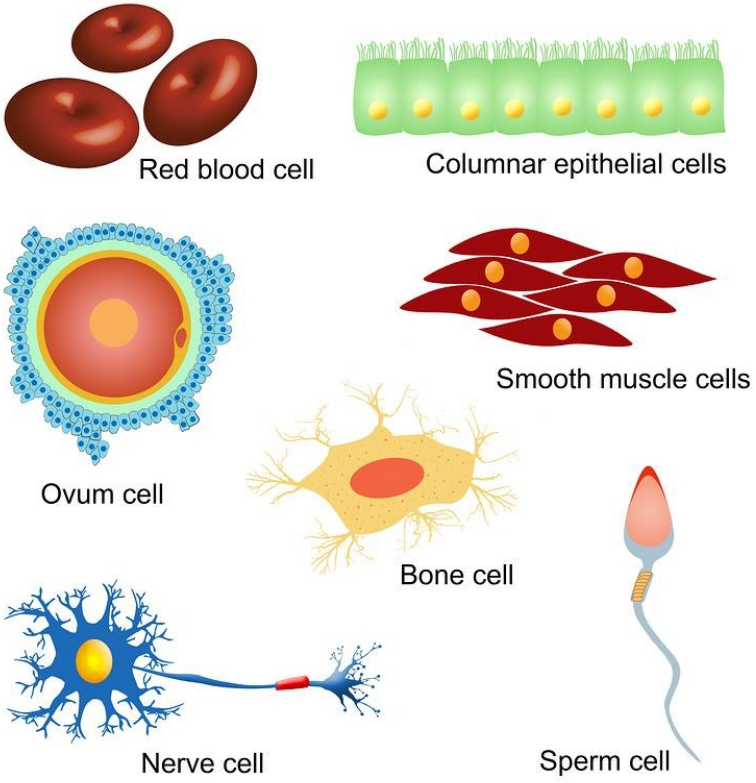
http://dev.biologists.org/content/136/5/701

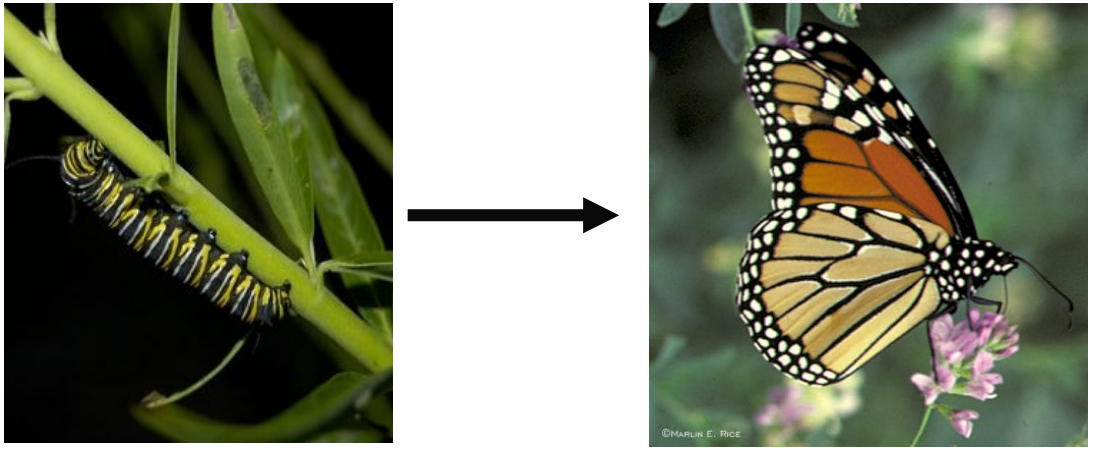


http://www.ebi.ac.uk/biomodels/ModelMonth/2010-06/fig1.png

Importancia de la regulación

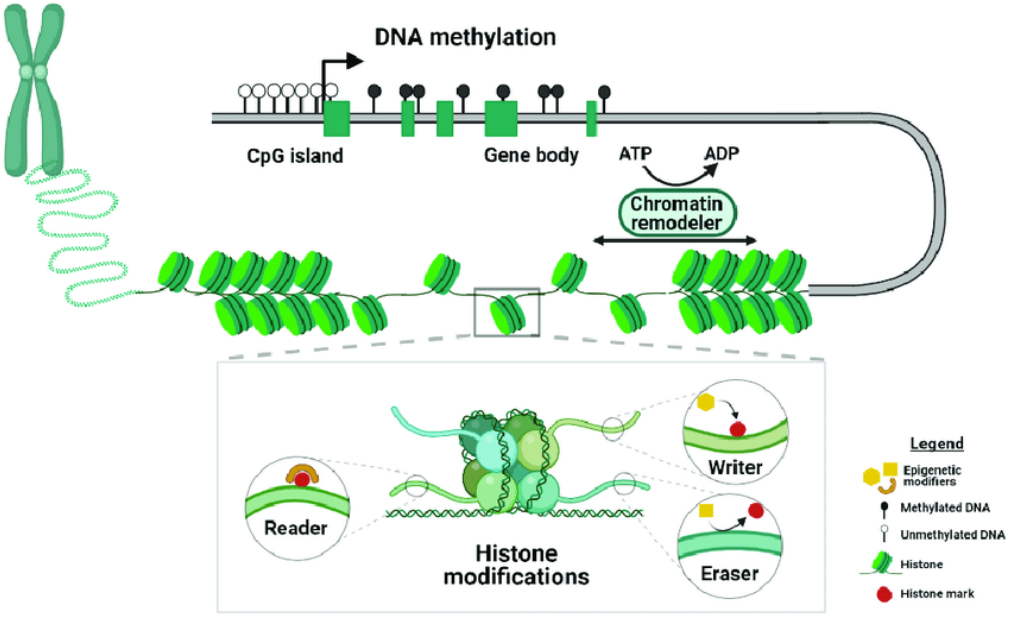
La combinación e intensidad de expresión de los genes de una célula determina fundamentalmente sus funciones. Estas células tienen el mismo **GENOMA** pero distinto **TRANSCRIPTOMA**



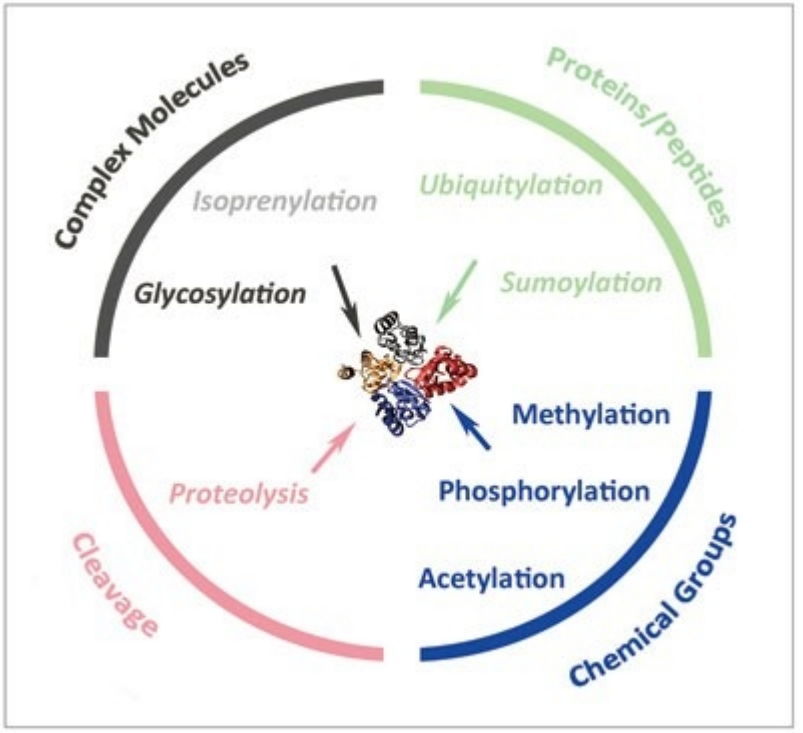


https://www.pinterest.es/pin/344103227749655379/

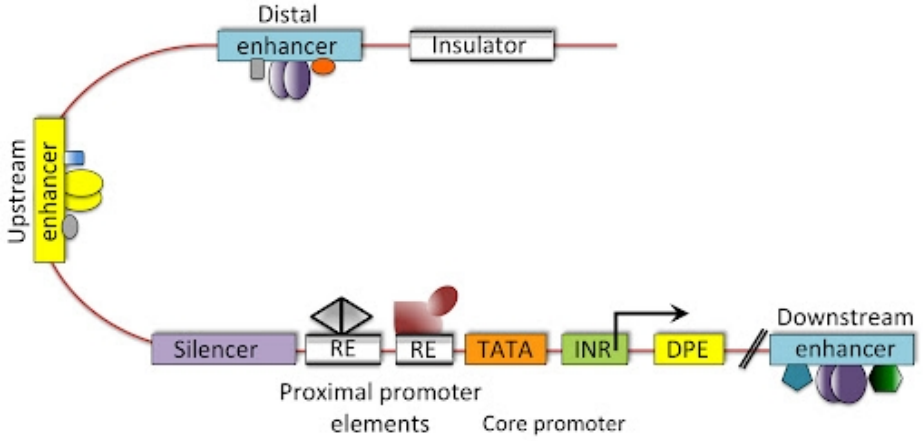
Niveles de regulación



Chiara DL *et al.* Cells 10:1074



Wang Y-C *et al.* Cell Research 24:143



Reinke V *et al.* Worm Book (2013)

Tema 2: Gestión de datos de origen biológico

2.1 Conceptos básicos de biología molecular.

2.2 Formatos comunes de archivos de información biológica

2.3 Bases de datos de información biomédica.

2.4 Acceso a las bases de datos.

Archivos de secuencia

FASTA (archivo codificación universal de texto. EXT: .fasta, .fa, .txt)

>Sequence 1 ◄——— **Nombre de la secuencia**
ACCCAGGCACAGGCATTAGTGCCCGTTGGAGAAAACAGGGGAATCCCGAAGAAATGGTGG
……
….
> Sequence 2
……
……..

Secuencia <120 caracteres. Muy común 60.

```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAKKADRLAAEG
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKDPNKPEHK
IPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQLDSKLTL
MGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKTEDFAAEVAAQL
>SEQUENCE_2
SATVSEINSETDFVAKNDQFIALTKDTTAHIQSNSLQSVEELHSSTINGVKFEEYLKSQI
ATIGENLVVRRFATLKAGANGVVNGYIHTNGRVGVVIAAACDSAEVASKSRDLLRQICMH
```

Nucleic Acid Code	Meaning
A	A
C	C
G	G
T	T
U	U
(i)	i
R	A or G (I)
Y	C, T or U
K	G, T or U
M	A or C
S	C or G
W	A, T or U
B	not A (i.e. C, G, T or U)
D	not C (i.e. A, G, T or U)
H	not G (i.e., A, C, T or U)
V	neither T nor U (i.e. A, C or G)
N	A C G T U
-	gap of indeterminate length

Amino Acid Code	Meaning
A	Alanine
B	Aspartic acid (D) or Asparagine (N)
C	Cysteine
D	Aspartic acid
E	Glutamic acid
F	Phenylalanine
G	Glycine
H	Histidine
I	Isoleucine
J	Leucine (L) or Isoleucine (I)
K	Lysine
L	Leucine
M	Methionine/Start codon
N	Asparagine
O	Pyrrolysine (rare)
P	Proline
Q	Glutamine
R	Arginine
S	Serine
T	Threonine
U	Selenocysteine (rare)
V	Valine
W	Tryptophan
Y	Tyrosine
Z	Glutamic acid (E) or Glutamine (Q)
X	any
*	translation stop
-	gap of indeterminate length

Archivos de secuencia

MAF (Multiple alignment file) (archivo codificación universal de texto. EXT: .maf, .txt)

Encabezamiento

```
##maf version=1 scoring=tba.v8
```

Version. Obligatorio
Scoring. Opcional
Program. Opcional

```
# tba.v8 (((human chimp) baboon) (mouse rat))
```

Parámetros. Opcional

Bloque de alineamiento (a)

```
a score=23262.0
```

score. Opcional
pass. Opcional

Secuencias alineadas (s)

```
s hg16.chr7     27707221 13 + 158545518 gcagctgaaaaca
s panTro1.chr6 28869787 13 + 161576975 gcagctgaaaaca
s baboon          249182 13 +   4622798 gcagctgaaaaca
s mm4.chr6      53310102 13 + 151104725 ACAGCTGAAAATA
```

Src.
Start.
Size.
Strand
Src size
Text

```
##maf version=1 scoring=tba.v8
# tba.v8 (((human chimp) baboon) (mouse rat))

a score=23262.0
s hg18.chr7     27578828 38 + 158545518 AAA-GGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s panTro1.chr6 28741140 38 + 161576975 AAA-GGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s baboon          116834 38 +   4622798 AAA-GGGAATGTTAACCAAATGA---GTTGTCTCTTATGGTG
s mm4.chr6      53215344 38 + 151104725 -AATGGGAATGTTAAGCAAACGA---ATTGTCTCTCAGTGTG
s rn3.chr4      81344243 40 + 187371129 -AA-GGGGATGCTAAGCCAATGAGTTGTTGTCTCTCAATGTG

a score=5062.0
s hg18.chr7     27699739 6 + 158545518 TAAAGA
s panTro1.chr6 28862317 6 + 161576975 TAAAGA
s baboon          241163 6 +   4622798 TAAAGA
s mm4.chr6      53303881 6 + 151104725 TAAAGA
s rn3.chr4      81444246 6 + 187371129 taagga

a score=6636.0
s hg18.chr7     27707221 13 + 158545518 gcagctgaaaaca
s panTro1.chr6 28869787 13 + 161576975 gcagctgaaaaca
s baboon          249182 13 +   4622798 gcagctgaaaaca
s mm4.chr6      53310102 13 + 151104725 ACAGCTGAAAATA
```

Archivos de modelos de genes

GTF/GFF (archivo codificación universal de texto. EXT: .gtf, .gff, .txt)

GFF/GTF File Format - Definition and supported options

The GFF (General Feature Format) format consists of one line per feature, each containing 9 columns of data, plus optional track definition lines. The following documen

The GTF (General Transfer Format) is identical to GFF version 2.

- Fields
- Track lines
- More information

Fields

Fields **must** be tab-separated. Also, all but the final field in each feature line must contain a value; "empty" columns should be denoted with a '.'

1. **seqname** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. **Important note**: the seqname must be one used assembly. See the example GFF output below.
2. **source** - name of the program that generated this feature, or the data source (database or project name)
3. **feature** - feature type name, e.g. Gene, Variation, Similarity
4. **start** - Start position* of the feature, with sequence numbering starting at 1.
5. **end** - End position* of the feature, with sequence numbering starting at 1.
6. **score** - A floating point value.
7. **strand** - defined as + (forward) or - (reverse).
8. **frame** - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
9. **attribute** - A semicolon-separated list of tag-value pairs, providing additional information about each feature.

```
1 transcribed_unprocessed_pseudogene  gene        11869 14409 . + . gene_id "ENSG00000223972"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "
1 processed_transcript                transcript  11869 14409 . + . gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; gene_name "DDX11L1"; gen
```

Archivos de regiones

BED (archivo codificación universal de texto. EXT: .bed, .txt)

Separado por tabulaciones

BED File Format - Definition and supported options

The BED format consists of one line per feature, each containing 3-12 columns of data, plus optional track definition lines.

- Required fields
- Optional fields
- Track lines
- BedGraph format

4. **name** - Label to be displayed under the feature, if turned on in "Configure this page".
5. **score** - A score between 0 and 1000. See track lines, below, for ways to configure the display s
6. **strand** - defined as + (forward) or - (reverse).
7. **thickStart** - coordinate at which to start drawing the feature as a solid rectangle
8. **thickEnd** - coordinate at which to stop drawing the feature as a solid rectangle
9. **itemRgb** - an RGB colour value (e.g. 0,0,255). Only used if there is a track line with the value o
10. **blockCount** - the number of sub-elements (e.g. exons) within the feature
11. **blockSizes** - the size of these sub-elements
12. **blockStarts** - the start coordinate of each sub-element

Required fields

The first three fields in each feature line are required:

1. **chrom** - name of the chromosome or scaffold. Any valid seq_region_name can be used, and chromosome names car
2. **chromStart** - Start position of the feature in standard chromosomal coordinates (i.e. first base is 0).
3. **chromEnd** - End position of the feature in standard chromosomal coordinates

```
chr1  213941196  213942363
chr1  213942363  213943530
chr1  213943530  213944697
chr2  158364697  158365864
chr2  158365864  158367031
chr3  127477031  127478198
chr3  127478198  127479365
chr3  127479365  127480532
chr3  127480532  127481699
```

Variantes del BED: , **bigBed** (binario, .bb), **BedGraph** (Texto, .bedgraph), **WIG** (Texto, .wig) o **BIGWIG** (binario,.bw)

Tema 2: Gestión de datos de origen biológico

2.1 Conceptos básicos de biología molecular.

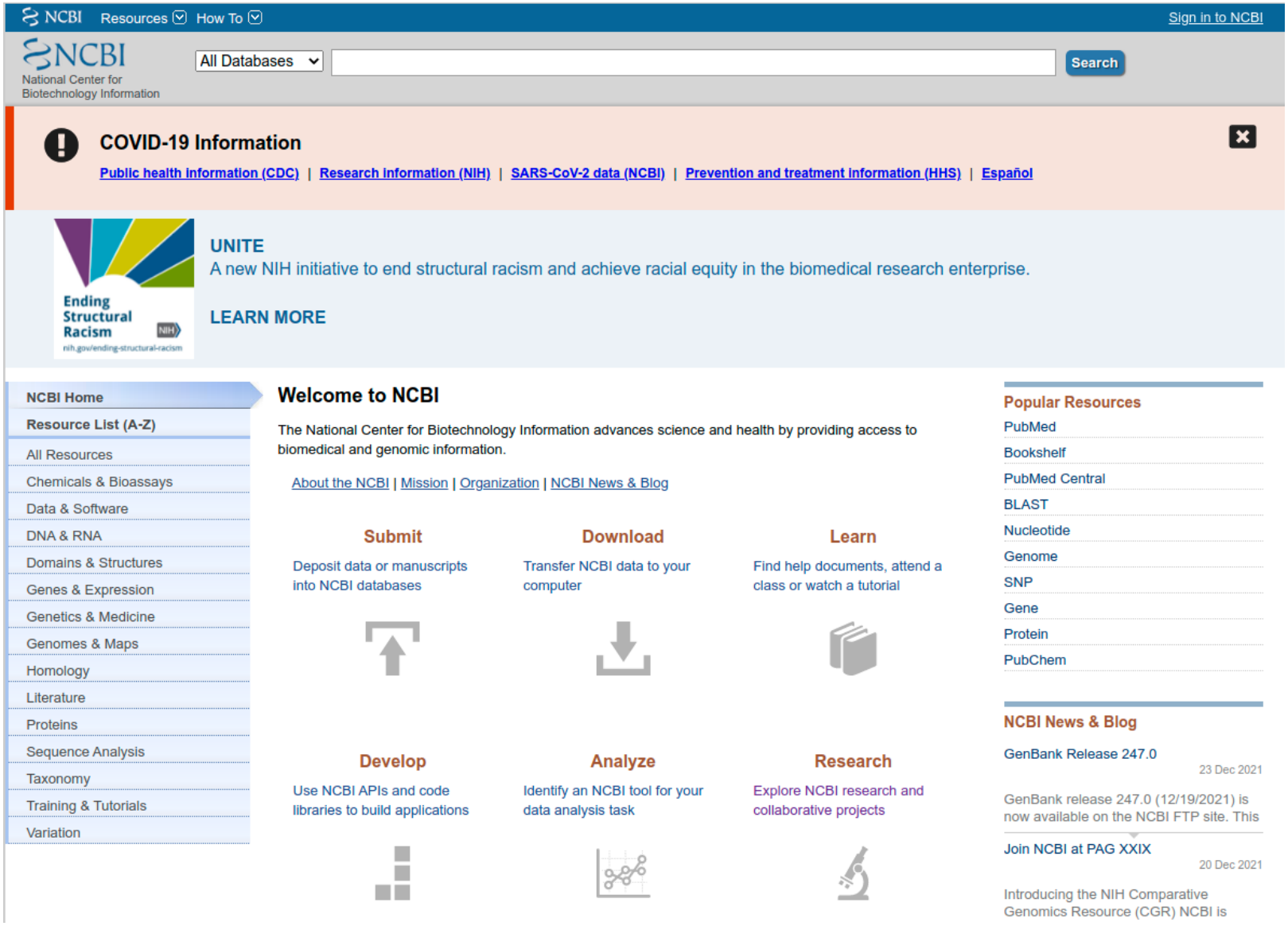
2.2 Formatos comunes de archivos de información biológica

2.3 Bases de datos de información biomédica.

2.4 Acceso a las bases de datos.

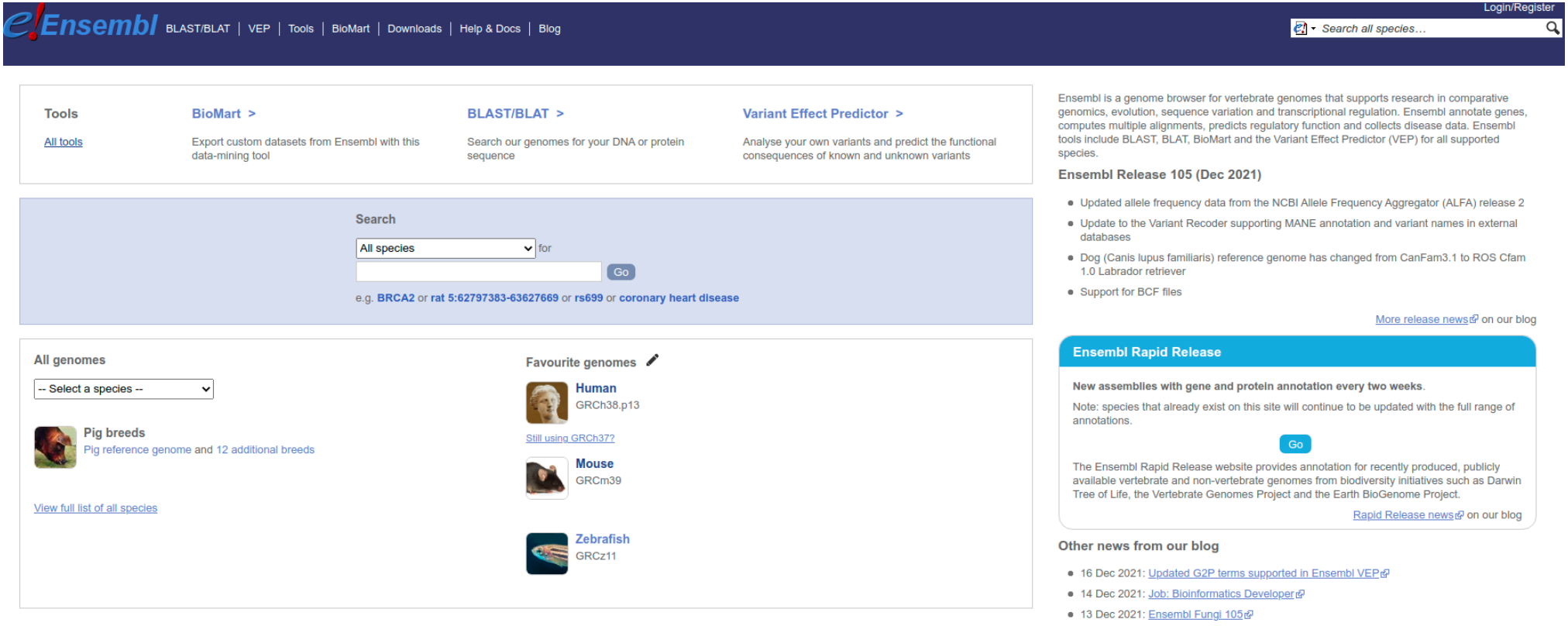
National Center for Biotechnology Information (NCBI)

https://www.ncbi.nlm.nih.gov/



Ensembl

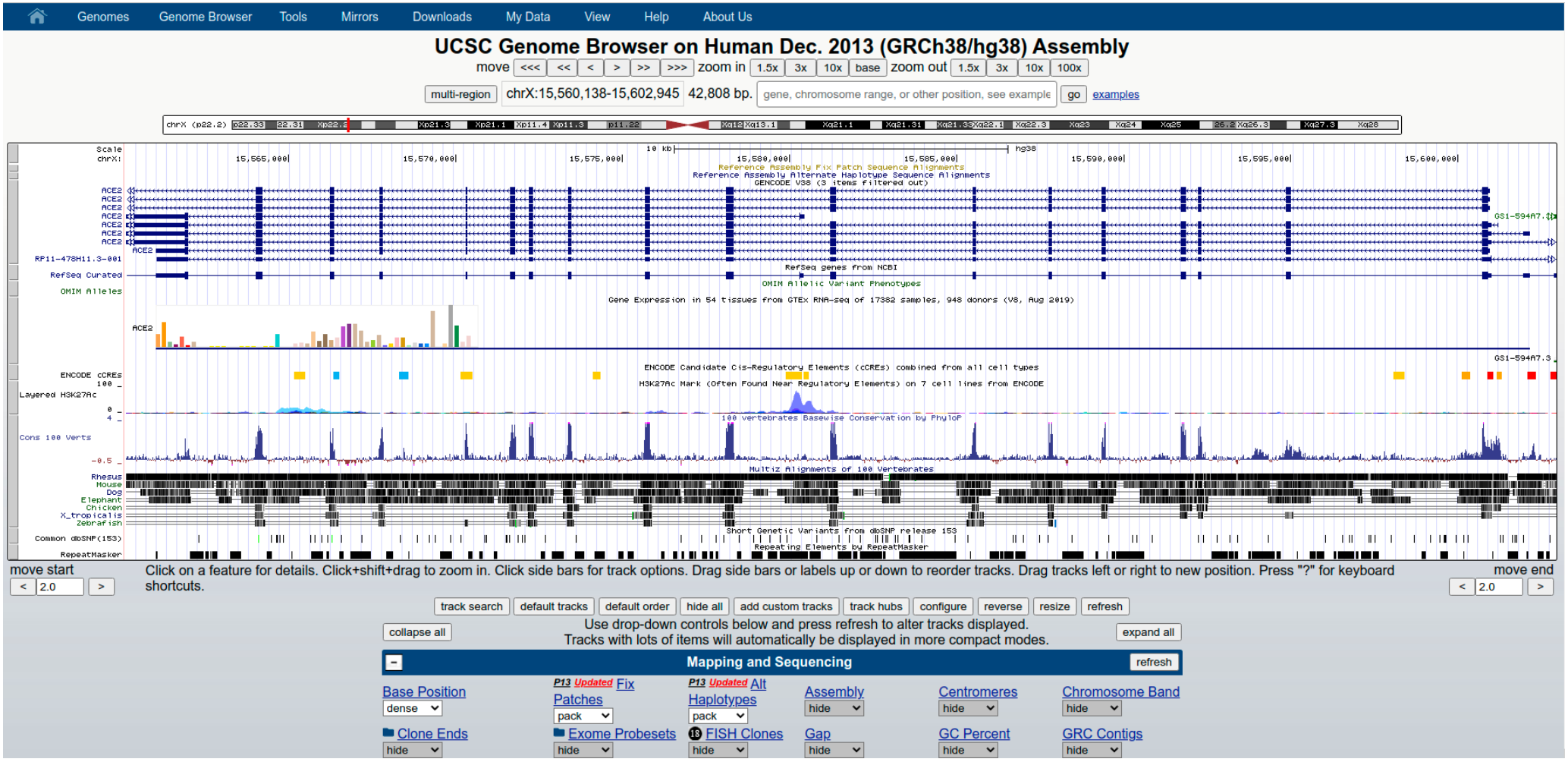
https://www.ensembl.org/index.html



Secuencias de genomas y genes.

UCSC

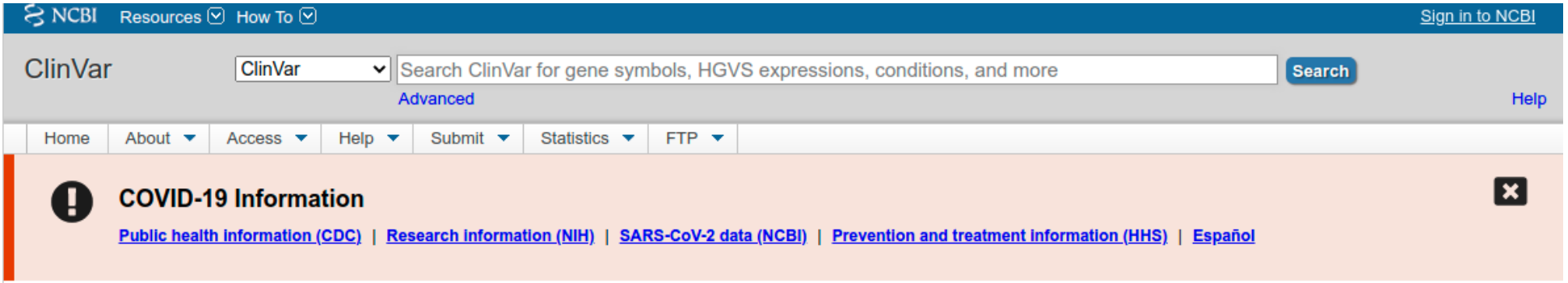
https://genome.ucsc.edu/index.html



Muchísima información centralizada sobretodo de genomas humano y ratón.

ClinVar

https://www.ncbi.nlm.nih.gov/clinvar/intro/



What is ClinVar?

ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence. ClinVar thus facilitates access to and communication about the relationships asserted between human variation and observed health status, and the history of that interpretation. ClinVar processes submissions reporting variants found in patient samples, assertions made regarding their clinical significance, information about the submitter, and other supporting data. The alleles described in submissions are mapped to reference sequences, and reported according to the HGVS standard. ClinVar then presents the data for interactive users as well as those wishing to use ClinVar in daily workflows and other local applications. ClinVar works in collaboration with interested organizations to meet the needs of the medical genetics community as efficiently and effectively as possible. Read more about using ClinVar.

ClinVar supports submissions of differing levels of complexity. The submission may be as simple as a representation of an allele and its interpretation (sometimes termed a variant-level submission), or as detailed as providing multiple types of structured observational (case-level) or experimental evidence about the effect of the variation on phenotype. A major goal is to support computational (re)evaluation, both of genotypes and assertions, and to enable the ongoing evolution and development of knowledge regarding variations and associated phenotypes. ClinVar is an active partner of the ClinGen project, providing data for evaluation and archiving the results of interpretation by recognized expert panels and providers of practice guidelines. ClinVar archives and versions submissions which means that when submitters update their records, the previous version is retained for review. Read more about submitting data to ClinVar.

The level of confidence in the accuracy of variation calls and assertions of clinical significance depends in large part on the supporting evidence, so this information, when available, is collected and visible to users. Because the availability of supporting evidence may vary, particularly in regard to retrospective data aggregated from published literature, the archive accepts submissions from multiple groups, and aggregates related information, to reflect transparently both consensus and conflicting assertions of clinical significance. A review status is also assigned to any assertion, to support communication about the trustworthiness of any assertion. Domain experts are encouraged to apply for recognition as an expert panel.

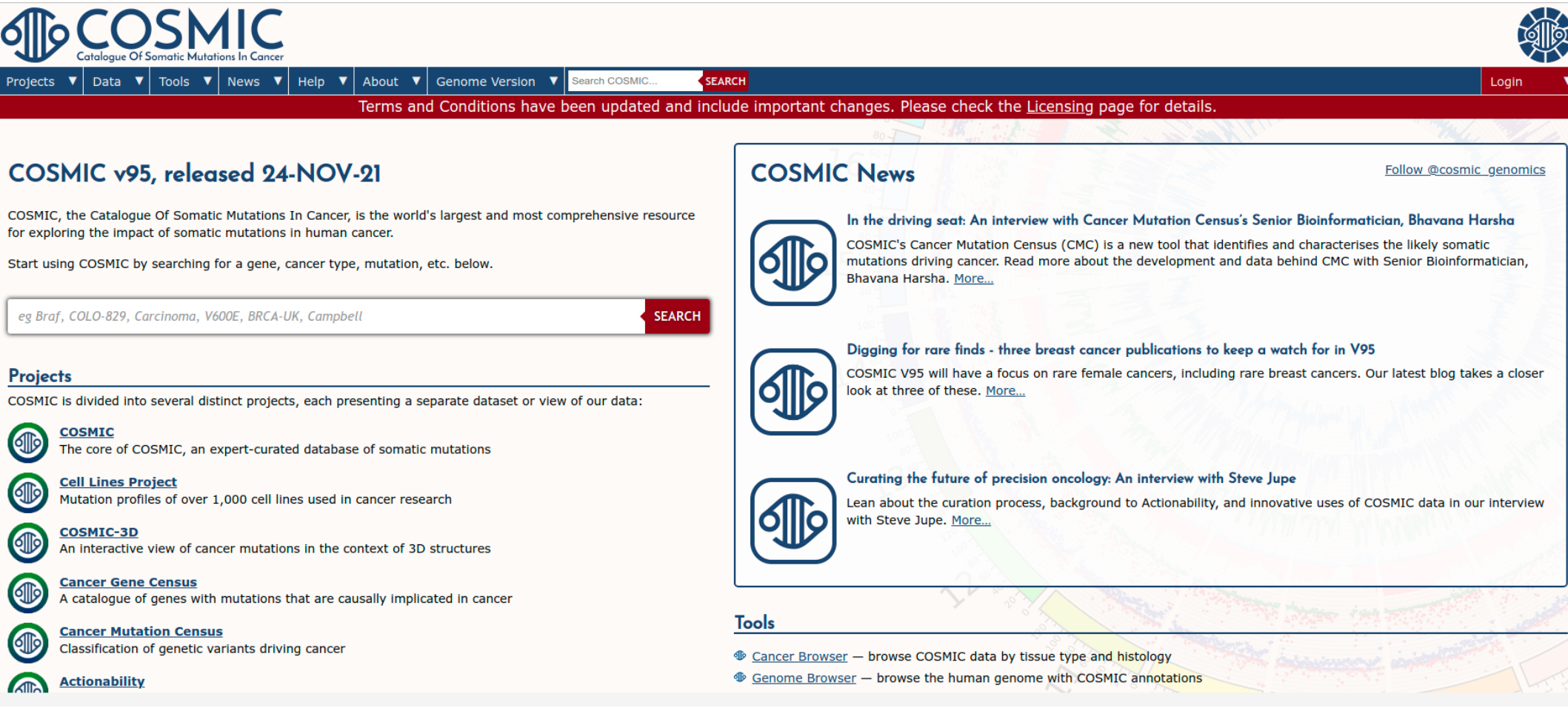
Accessions, with the format SCV000000000.0, are assigned to each submitted record. If there are multiple submitted records about the same variation/condition pair, they are aggregated within ClinVar's data flow and reported as a reference accession with the format RCV000000000.0. ***Because of this model, one variant will be included in multiple RCV accessions whenever different conditions are reported for that variant***. Submitted records for the same variation are also aggregated and reported as an accession with the format VCV000000000.0. This aggregation lets a user review all submitted data for a variant, regardless of the condition for which it was interpreted.

ClinVar archives submitted information, and adds identifiers and other data that may be available about a variant or condition from other public resources. However ClinVar neither curates content nor modifies interpretations independent of an explicit submission. If you have data that differs from what is currently represented in ClinVar, we encourage you to submit your data and the evidence supporting your interpretation. There is a submission wizard to guide you through that process.

If you are submitting variants that were interpreted as part of work funded by the NIH, please consult your program officer about expectations for submissions to ClinVar.

COSMIC

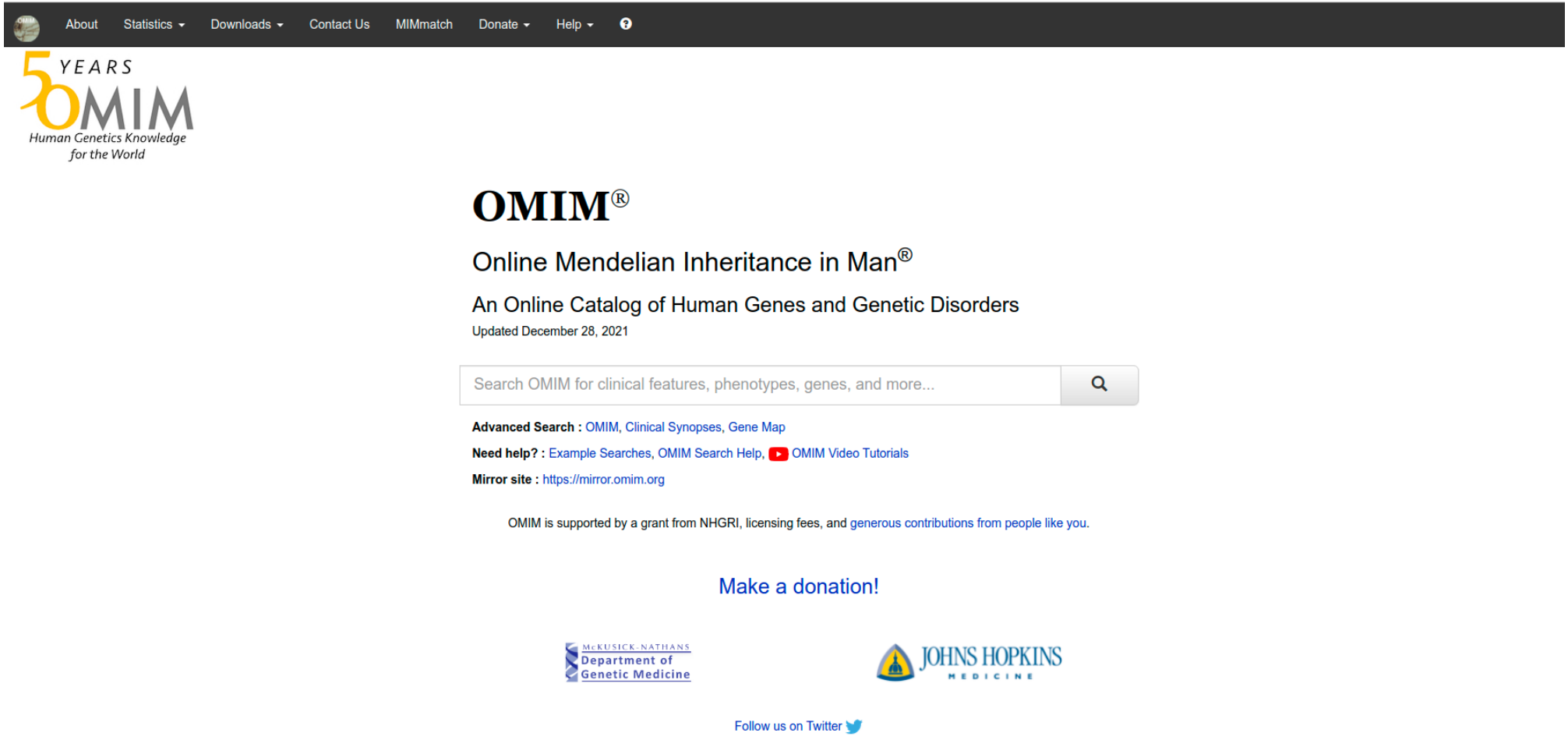
https://cancer.sanger.ac.uk/cosmic



Mutaciones en cáncer.

OMIM

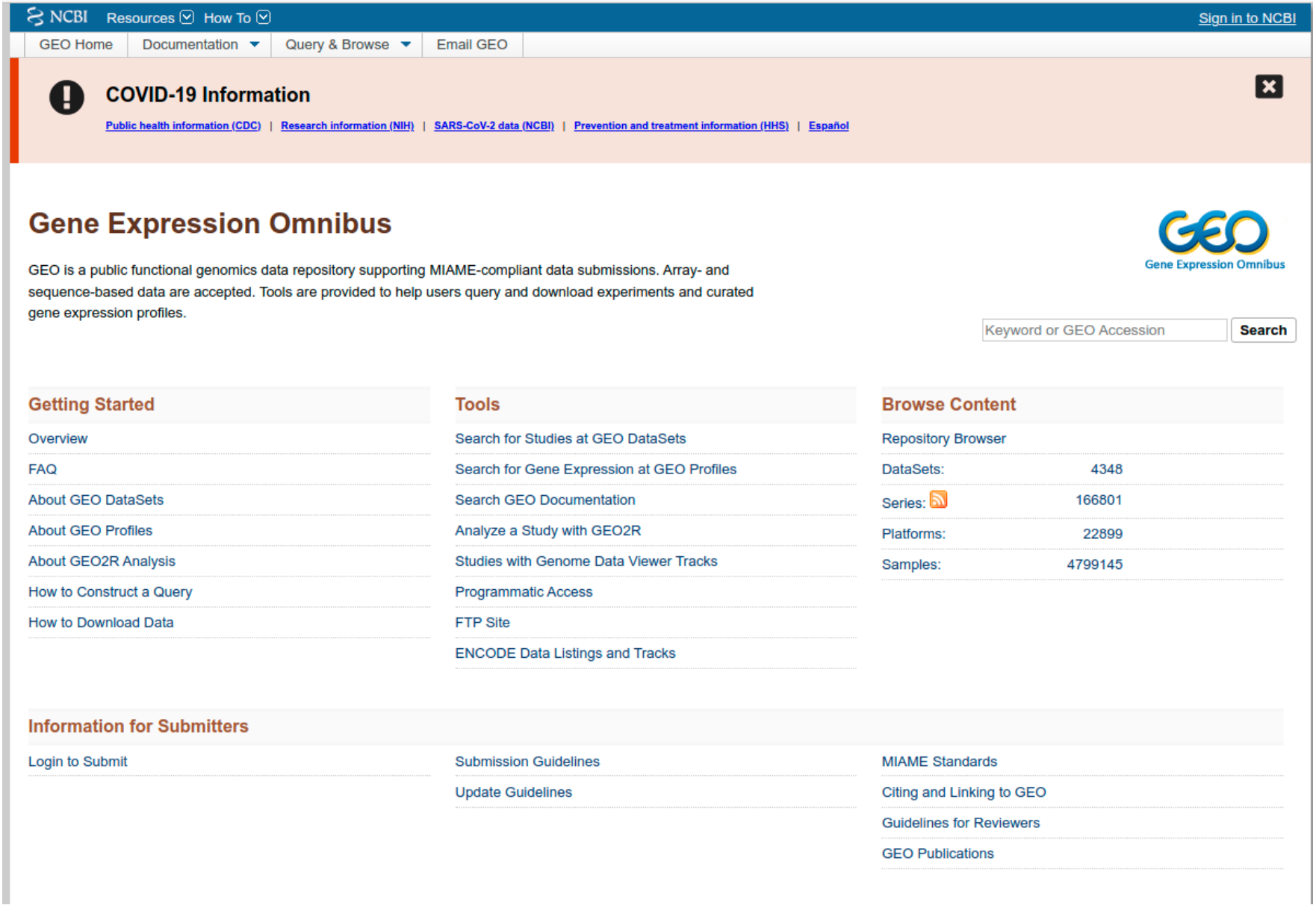
https://www.omim.org/



Mutaciones asociadas a enfermedades mendelianas.

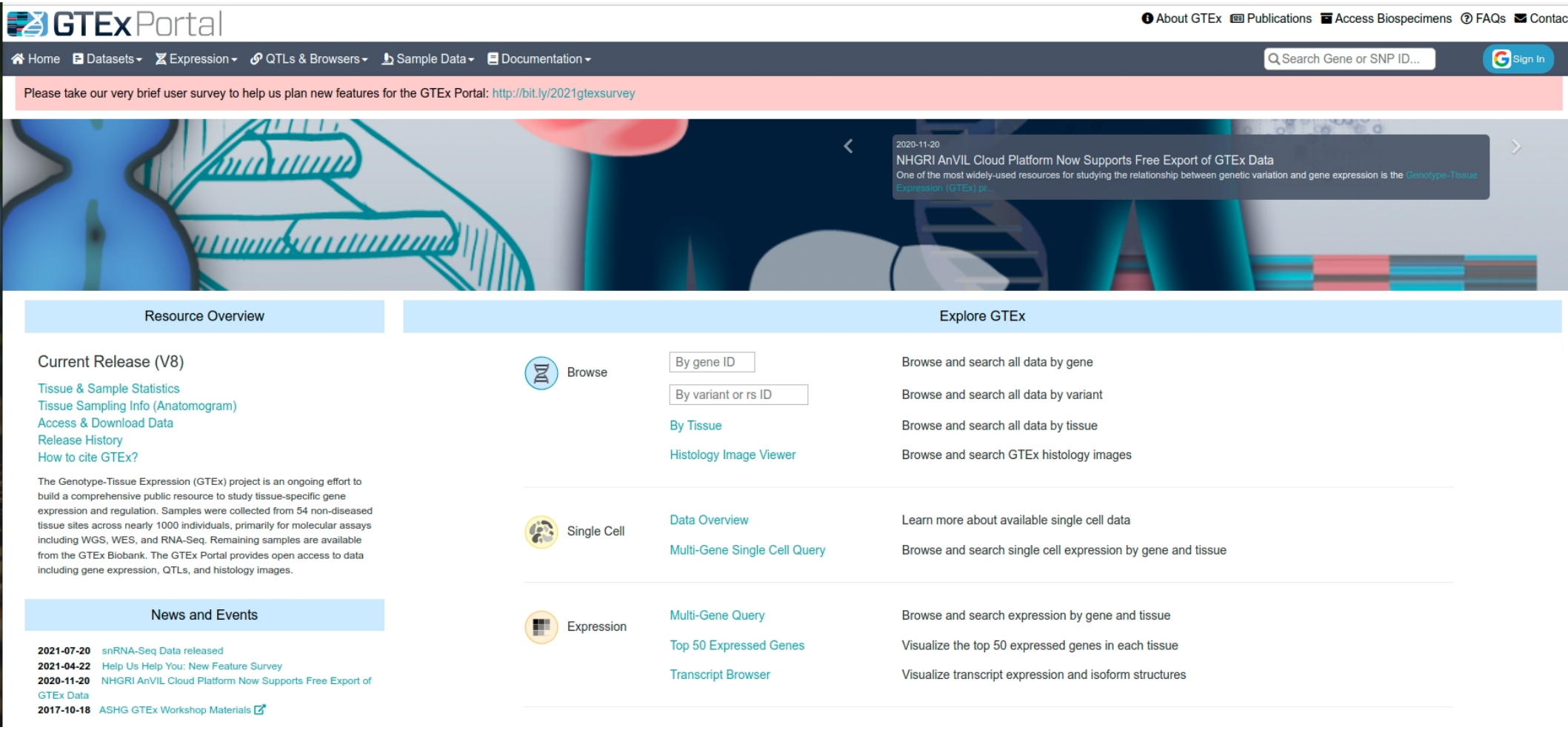
GEO

https://www.ncbi.nlm.nih.gov/geo/



GTEX

https://gtexportal.org/home/



Tema 2: Gestión de datos de origen biológico

2.1 Conceptos básicos de biología molecular.

2.2 Formatos comunes de archivos de información biológica

2.3 Bases de datos de información biomédica.

2.4 Acceso a las bases de datos.

Utilidades de descarga

https://genome.ucsc.edu/cgi-bin/hgTables

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

Table Browser

Use this tool to retrieve and export data from the Genome Browser annotation track database. You can limit retrieval based on data attributes and intersect or merge with data from another track, or retrieve DNA sequence covered by a track. More...

Select dataset

clade: Mammal **genome:** Human **assembly:** Dec. 2013 (GRCh38/hg38)
group: Genes and Gene Predictions **track:** GENCODE V38
table: knownGene describe table schema

Define region of interest

region: genome position chrX:15,560,138-15,602,945 lookup define regions
identifiers (names/accessions): paste list upload list

Optional: Subset, combine, compare with another track

filter: create
intersection: create

Retrieve and display data

output format: all fields from selected table Send output to Galaxy GREAT
output filename: (leave blank to keep output in browser)
file type returned: plain text gzip compressed

get output summary/statistics

- Ensembl protein families
- Ontologies
  - GO: Cellular component
  - GO: Biological process
  - GO: Molecular function
- Phenotypes
- Genetic Variation
  - Variant table
  - Variant image
  - Structural variants
- Gene expression
- Pathway
- Regulation
- External references
- Supporting evidence
- ID History
  - Gene history

Configure this page
Custom tracks
Export data
Share this page

Show/hide columns (1 hidden) Filter

Transcript ID	Name	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
ENST00000269305.9	TP53-201	2512	393aa	Protein coding	CCDS11118	P04637-1	NM_000546.6	MANE Select v0.95 Ensembl Canonical GENCODE basic APPRIS P1 TSL:1
ENST00000420246.6	TP53-204	2653	341aa	Protein coding	CCDS45606	P04637-2	-	GENCODE basic TSL:1
ENST00000622645.4	TP53-226	2653	302aa	Protein coding	CCDS73971	P04637-5	-	GENCODE basic TSL:1
ENST00000610292.4	TP53-219	2639	354aa	Protein coding	CCDS73969	P04637-4	-	GENCODE basic TSL:1
ENST00000455263.6	TP53-206	2580	346aa	Protein coding	CCDS45605	P04637-3	-	GENCODE basic TSL:1
ENST00000610538.4	TP53-220	2580	307aa	Protein coding	CCDS73970	P04637-6	-	GENCODE basic TSL:1
ENST00000620739.4	TP53-225	2579	354aa	Protein coding	CCDS73969	P04637-4	-	GENCODE basic TSL:1
ENST00000445888.6	TP53-205	2506	393aa	Protein coding	CCDS11118	P04637-1	-	GENCODE basic APPRIS P1 TSL:1
ENST00000619485.4	TP53-224	2506	354aa	Protein coding	CCDS73969	P04637-4	-	GENCODE basic TSL:1
ENST00000510385.5	TP53-213	2404	209aa	Protein coding	CCDS73968	P04637-8	-	GENCODE basic TSL:1
ENST00000618944.4	TP53-222	2404	182aa	Protein coding	CCDS73965	A0A087WXZ1	-	GENCODE basic TSL:1
ENST00000504290.5	TP53-208	2331	214aa	Protein coding	CCDS73967	P04637-9	-	GENCODE basic TSL:1
ENST00000610623.4	TP53-221	2331	187aa	Protein coding	CCDS73964	A0A087WT22	-	GENCODE basic TSL:1
ENST00000504937.5	TP53-209	2271	261aa	Protein coding	CCDS73966	P04637-7	-	GENCODE basic TSL:1
ENST00000619186.4	TP53-223	2271	234aa	Protein coding	CCDS73963	A0A087X1Q1	-	GENCODE basic TSL:1

Acceso directo a la base de datos

La mayoría de portales tienen las bases de datos accesibles vía FTP o MySQL

Complete datasets and databases

Many datasets, e.g. all genes for a species, are available to download in a variety of formats from our FTP site.

Entire databases are also available via FTP as MySQL dumps.

NCBI Resources How To Sign in to NCBI

GenBank Nucleotide Search

GenBank Submit Genomes WGS Metagenomes TPA TSA INSDC Other

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

FTP access to GenBank data

The ASN.1 and Flatfile forms of the data are available at NCBI's anonymous FTP server:

- ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1
- ftp://ftp.ncbi.nlm.nih.gov/genbank

A mirror of the GenBank FTP site at the NCBI is available at the University of Indiana, courtesy of the Bio-Mirror project:

- ftp://bio-mirror.net/biomirror/genbank/

Some users who experience slow FTP transfers of large files might realize an improvement in transfer rates from this alternate site when the volume of traffic at the NCBI is high. For a list of other Bio-Mirror nodes, visit:

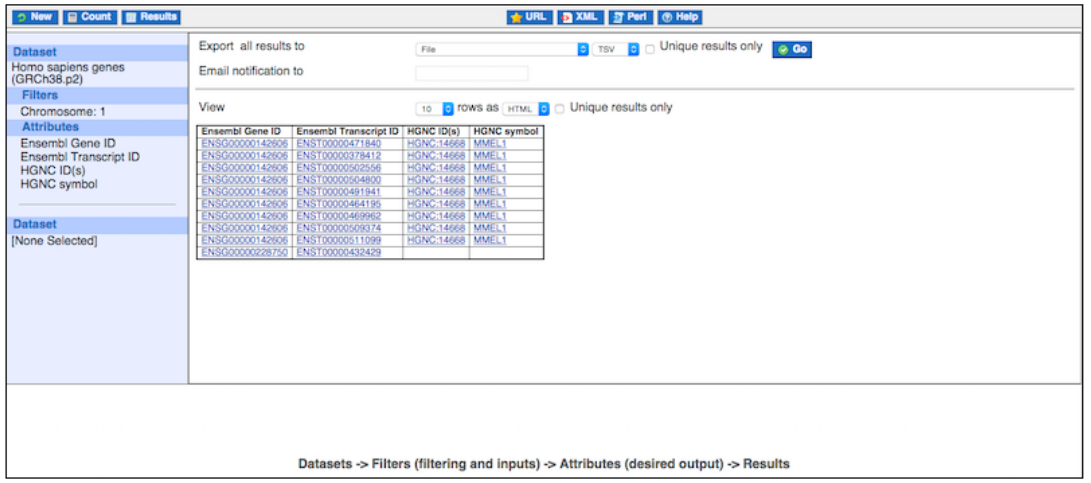
- http://www.bio-mirror.net/

For more complete details on what data are available, see the current GenBank release notes .

Aplicaciones intermediarias

Extracting data with BioMart

Tables of Ensembl data can be downloaded via the highly customisable BioMart data mining tool. The easy-to-use web-based tool allows extracti



BioMart tutorials and FAQs

- How to use BioMart
- BioMart tutorials: BioMart short videos and written materials.
- BioMart FAQs: BioMart Frequently Asked Questions.
  - How can I download all the proteins in the human proteome? Do I use BioMart?
  - Ensembl BioMart shows results for protein-coding genes when protein-associated attributes are chosen. Non-coding genes t
- Combining multiple species datasets: Example on how to combine multiple mart datasets inside one mart or accross different marts.

BioMart R package, RESTful access and API

- BiomaRt, Bioconductor R package
- BioMart RESTful access (Perl and wget)
- BioMart perl API

APIs

NCBI provides several public APIs that allow programmatic access to many databases and tools.

Entrez Programming Utilities (E-utilities)

The E-utilities are the public API to the NCBI Entrez system and allow access to all Entrez databases including PubMed, PMC, Gene, Nuccore and Protein. The E-utilities are a suite of eight server-side programs that accept a fixed URL syntax for search, link and retrieval operations. A companion package named Entrez Direct consists of several executables that allow the E-utilities to be called directly from a UNIX command line.

Documentation Quick Start Examples Entrez Direct

BLAST URL API

The BLAST API allows developers to submit BLAST searches for processing at NCBI or cloud service provider(s) using HTTPS. The API can then check the status of submitted searches and retrieve results when ready in several formats.

Overview Documentation

PubChem Power User Gateway (PUG)

PUG is a suite of APIs for the NCBI PubChem resource, and provides programmatic access to many PubChem functions including downloads of chemical and assay data, chemical structure searches and chemical standardization. PUG offers HTTP, REST and SOAP interfaces, and also integrates with the E-utilities to provide convenient access to other NCBI databases.

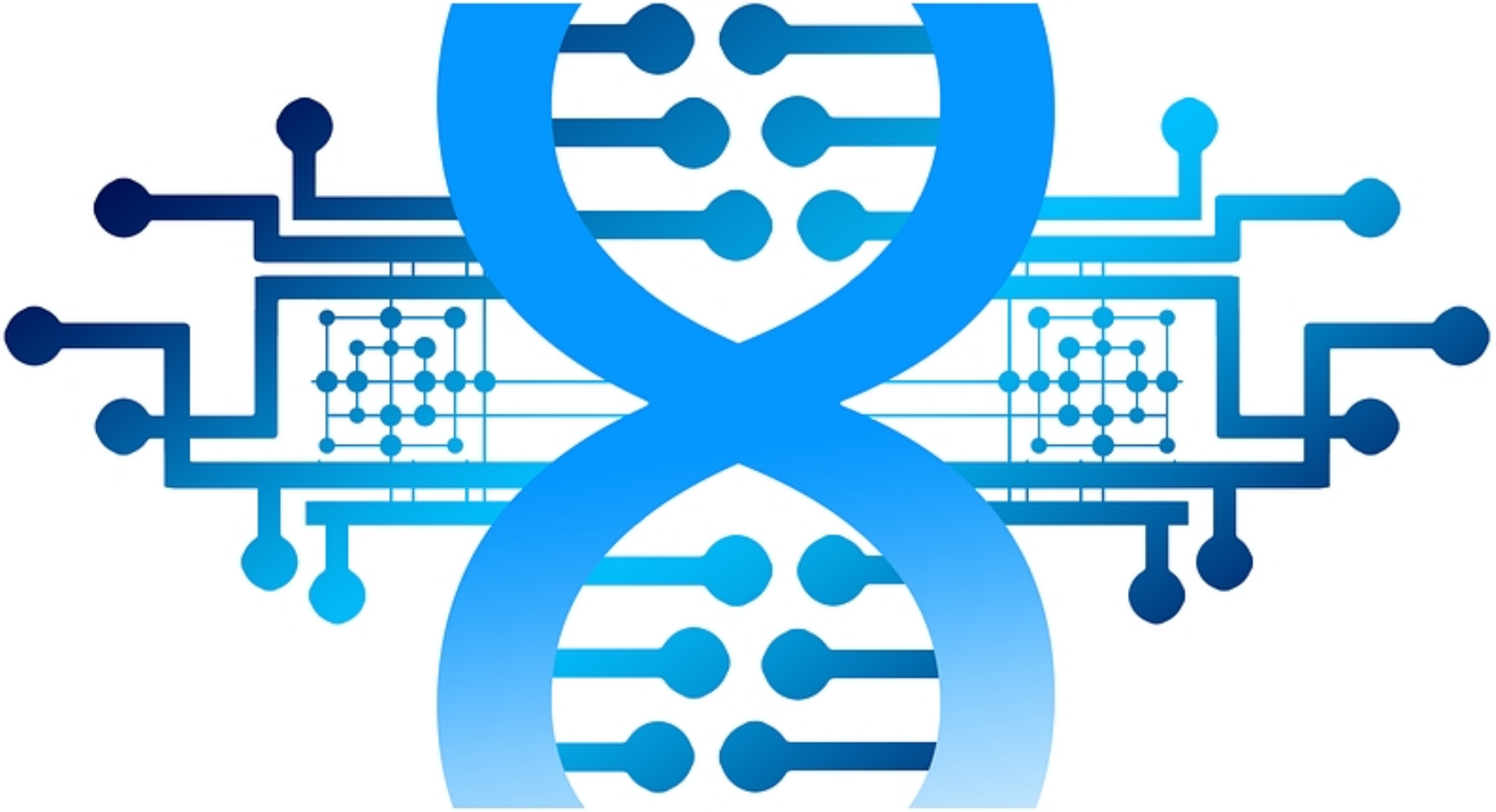
Documentation SOAP REST REST Tutorial

PubMed Central (PMC) APIs

PMC provides several APIs that provide programmatic access to various services that deal with PMC literature content, including file validation tools, Open Access web services, and an ID convertor that interconverts PMCID's, PMID's, Manuscript ID's, and DOI's.

Documentation

Bioinformática y análisis de datos ómicos



Tema 2: Gestión de datos de origen biológico

ignacio.varela@unican.es