LE BANCHE DATI

- I database relazionali

DataBase Management Systems

I “DataBase Management Systems” (DBMS) sono una collezione di programmi che permettono di immagazzinare, modificare ed estrarre informazioni da un database.

Esistono molti tipi differenti di DBMS progettati per lavorare su piccoli sistemi (personal computers) oppure su macchine molto potenti (piattaforme multiprocessore) o clusters.

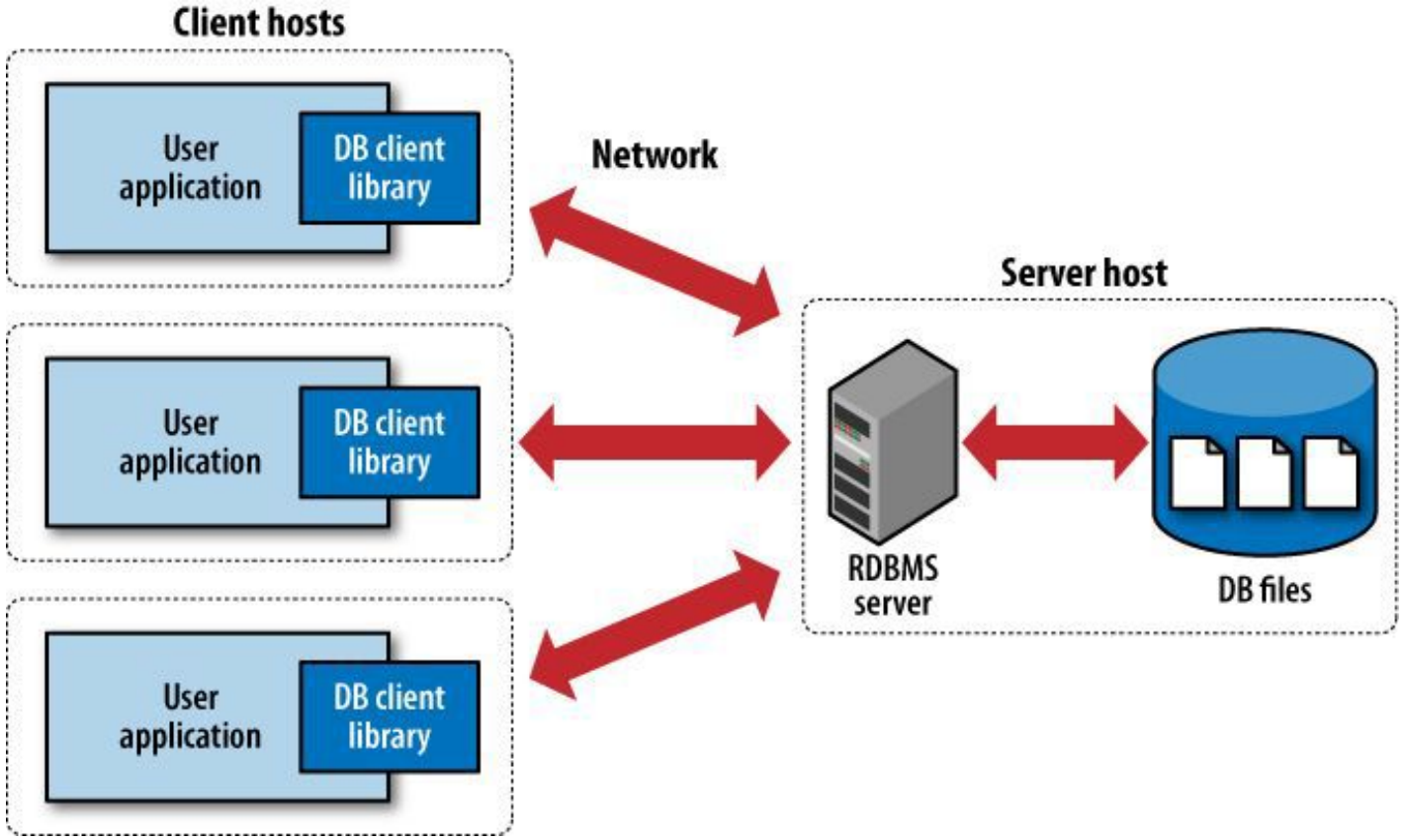
L’organizzazione interna di un DBMS influisce moltissimo sulla velocità’ e flessibilità’ di estrazione dei dati.

LE BANCHE DATI

- I database relazionali

La necessita' di condividere i dati mettendo un database centralizzato a disposizione di piu' utenti ha spinto la ricerca tecnologica a sviluppare programmi Relational-DBMS (RDBMS) di tipo client/server.

DataBase Management Systems
In questo tipo di sistema i dati si trovano su un computer che funge da database server e gli utenti interagiscono con i dati tramite applicazioni installate sui loro computer dette database client.

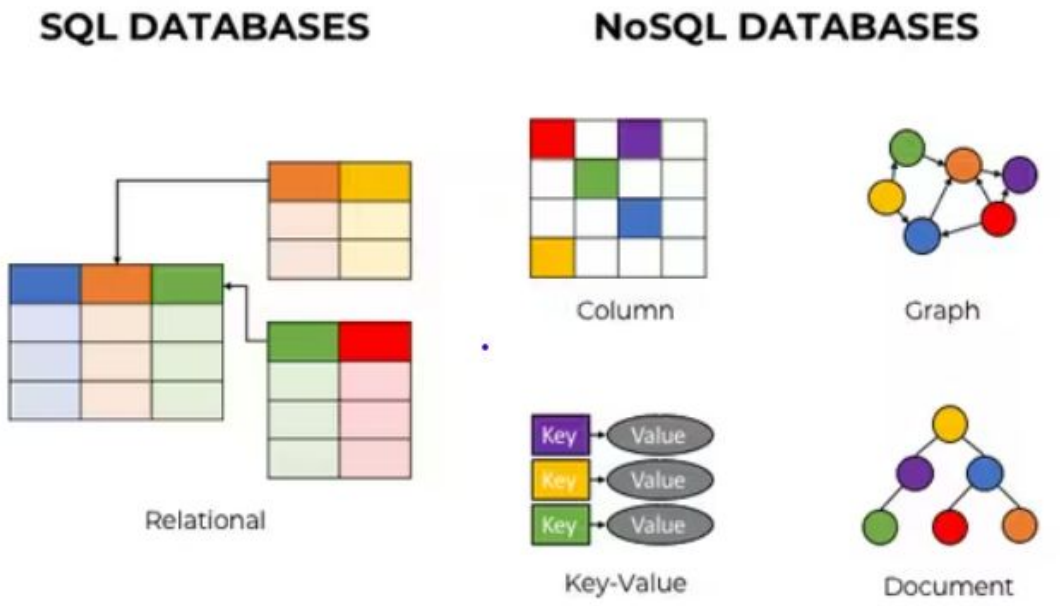


LE BANCHE DATI

- I database non relazionali

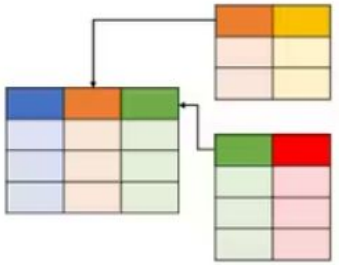
Un database **non relazionale**, o NoSQL (acronimo di Not Only SQL) non utilizza SQL (Structured Query Language) come linguaggio di interrogazione.

- I database non relazionali utilizzano altri tipi di modelli di dati, come il modello organizzato in documenti (in formato JSON), il modello a grafi o tabelle di dimensioni dinamiche, il modello a coppie chiave-valore.



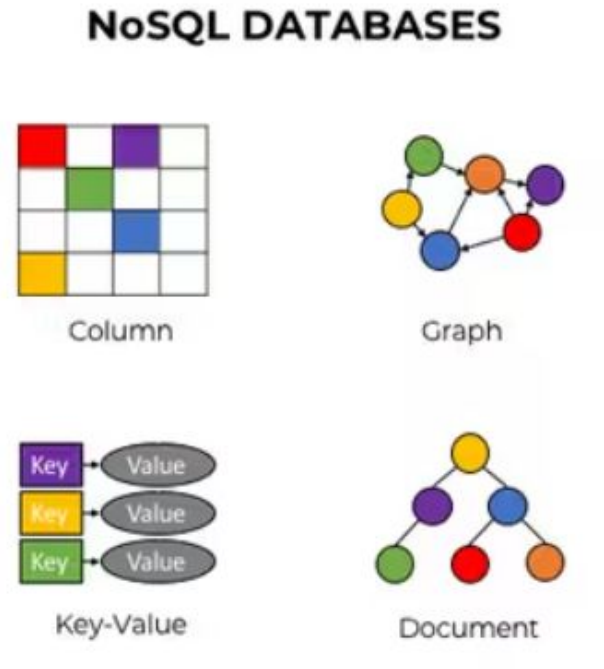
LE BANCHE DATI

I database relazionali, come MySQL, PostgreSQL e SQL Server, tendono a scalare verticalmente. Ciò significa che per gestire carichi di lavoro più grandi o migliorare le prestazioni, è necessario potenziare l'hardware del server esistente su cui il database è in esecuzione. Questo potenziamento può includere l'aggiunta di più CPU, RAM o spazio di archiviazione.



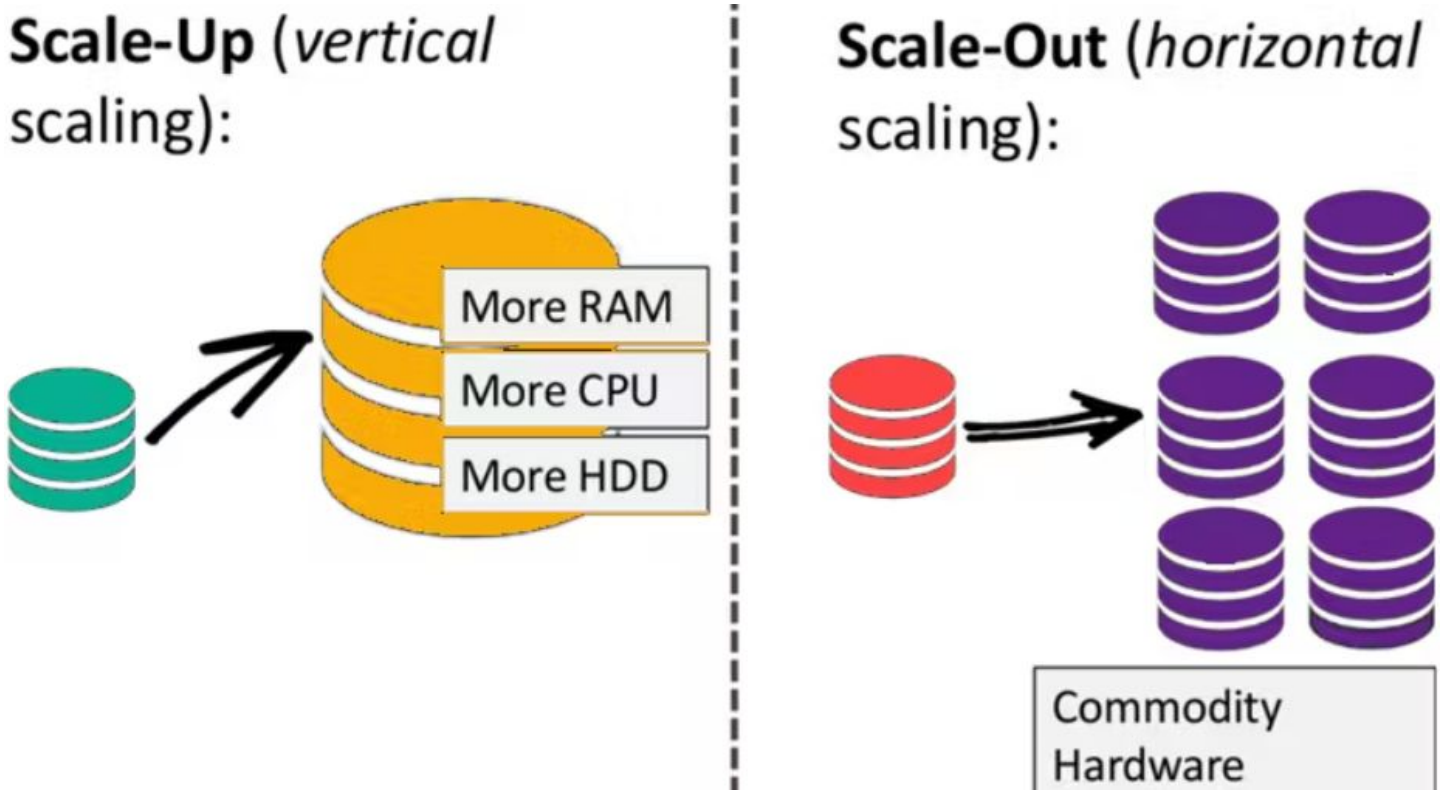
D'altra parte, i database NoSQL, come Cassandra, MongoDB e Neo4j, sono progettati per scalare orizzontalmente. Ciò significa che possono gestire più dati e traffico semplicemente aggiungendo più server o nodi alla rete del database. Questo approccio, noto anche come "sharding", consente di distribuire i dati e le richieste di elaborazione su molti server, invece di fare affidamento su un unico sistema potente.

LE BANCHE DATI



La scalabilità orizzontale è generalmente più flessibile e meno costosa della scalabilità verticale, poiché si possono aggiungere server standard o cloud senza dover investire in hardware costoso e specializzato. Inoltre, i database NoSQL sono spesso progettati con la tolleranza ai guasti e l'alta disponibilità in mente, il che significa che possono continuare a funzionare efficacemente anche quando alcuni nodi falliscono.

LE BANCHE DATI



Pertanto, i database NoSQL possono scalare all'infinito, rendendoli utili per l'archiviazione di Big Data. Aziende come Google, Amazon e Facebook hanno sviluppato i propri database NoSQL poiché non potevano più scalare verticalmente i propri database relazionali nel tempo.

LE BANCHE DATI

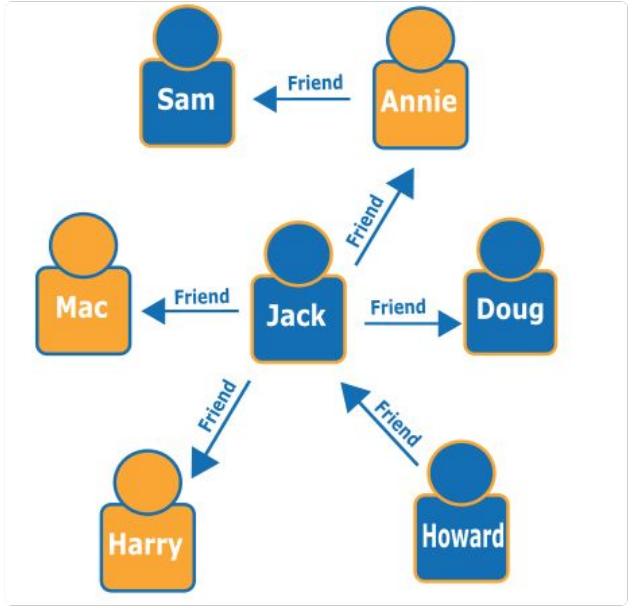
- I database non relazionali

NoSQL modello a grafo:

In un database a grafo, i dati sono rappresentati come **nodi** e le relazioni tra i dati sono rappresentate come **archi**.

Ogni nodo rappresenta un'entità o un concetto, mentre gli archi rappresentano le relazioni tra queste entità.

Ad esempio, in un database di social network, i nodi potrebbero rappresentare gli utenti, mentre gli archi potrebbero rappresentare le amicizie tra gli utenti.



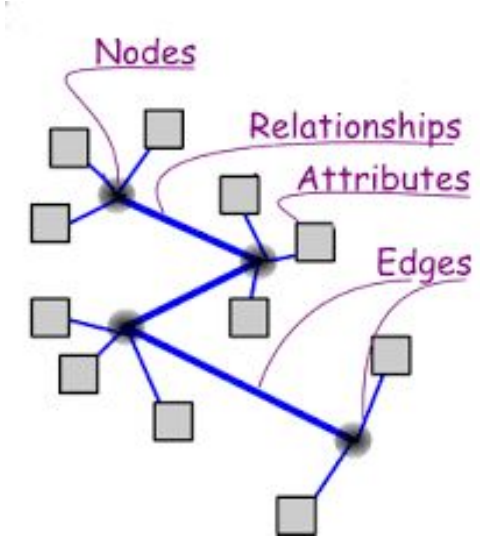
LE BANCHE DATI

- I database non relazionali

Un vantaggio dei database a grafo è la loro capacità di gestire dati altamente connessi e complessi, in cui le relazioni tra i dati sono altrettanto importanti quanto i dati stessi.

Inoltre, i database a grafo sono molto scalabili e possono gestire grandi quantità di dati e relazioni tra i dati in modo efficiente.

I database a grafo sono utilizzati in diverse applicazioni, tra cui social network, raccomandazioni di prodotti, motori di ricerca e sistemi di raccomandazione di contenuti.

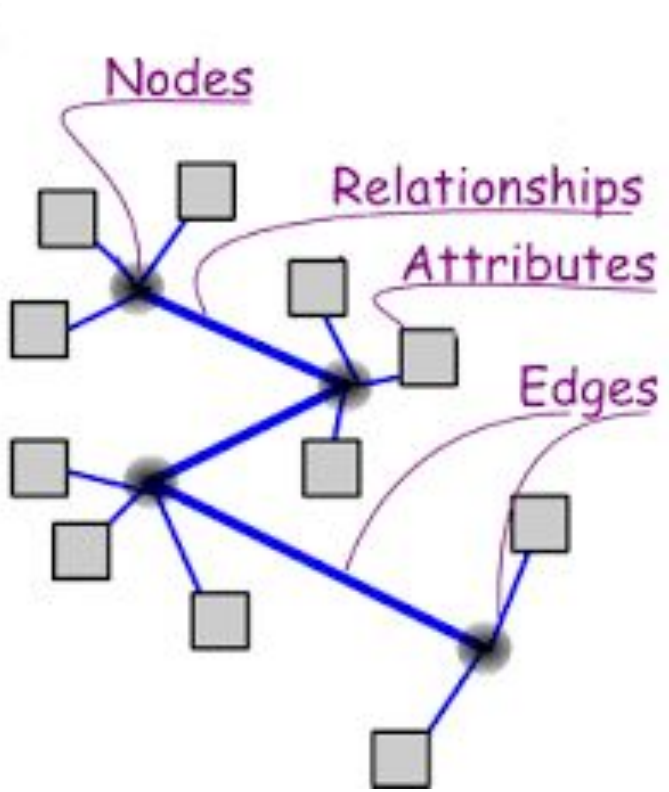


LE BANCHE DATI

- I database non relazionali

I componenti principali di un database NoSQL a grafo includono:

1. **Nodi**: sono gli oggetti fondamentali del database a grafo, rappresentano gli elementi dell'insieme che si vuole rappresentare, come ad esempio persone, luoghi, eventi, oggetti, etc.
2. **Archi**: sono le relazioni tra i nodi, che definiscono le connessioni e le interazioni tra gli oggetti rappresentati.
3. **Proprietà**: sono le informazioni aggiuntive associate a nodi e archi, che permettono di specificare ulteriori dettagli e metadati.
4. **Indici**: sono gli strumenti utilizzati per accelerare le ricerche all'interno de database, consentendo di trovare facilmente nodi e archi di interesse.
5. **Query language**: è il linguaggio utilizzato per interagire con il database, definire le ricerche e le interrogazioni per ottenere i dati desiderati.
6. **API**: sono le interfacce di programmazione (Application Programming Interface) utilizzate per accedere al database e interagire con esso da parte di applicazioni esterne o di altri sistemi.



LE BANCHE DATI

- API

Ecco alcuni esempi di API utilizzate per accedere e interagire con banche dati biologiche:

- La NCBI E-utilities API permette di effettuare ricerche su PubMed e altre banche dati della NCBI. Ad esempio, è possibile utilizzarla per scaricare sequenze nucleotidiche o proteiche, effettuare ricerche di parole chiave o creare filtri personalizzati per le ricerche.

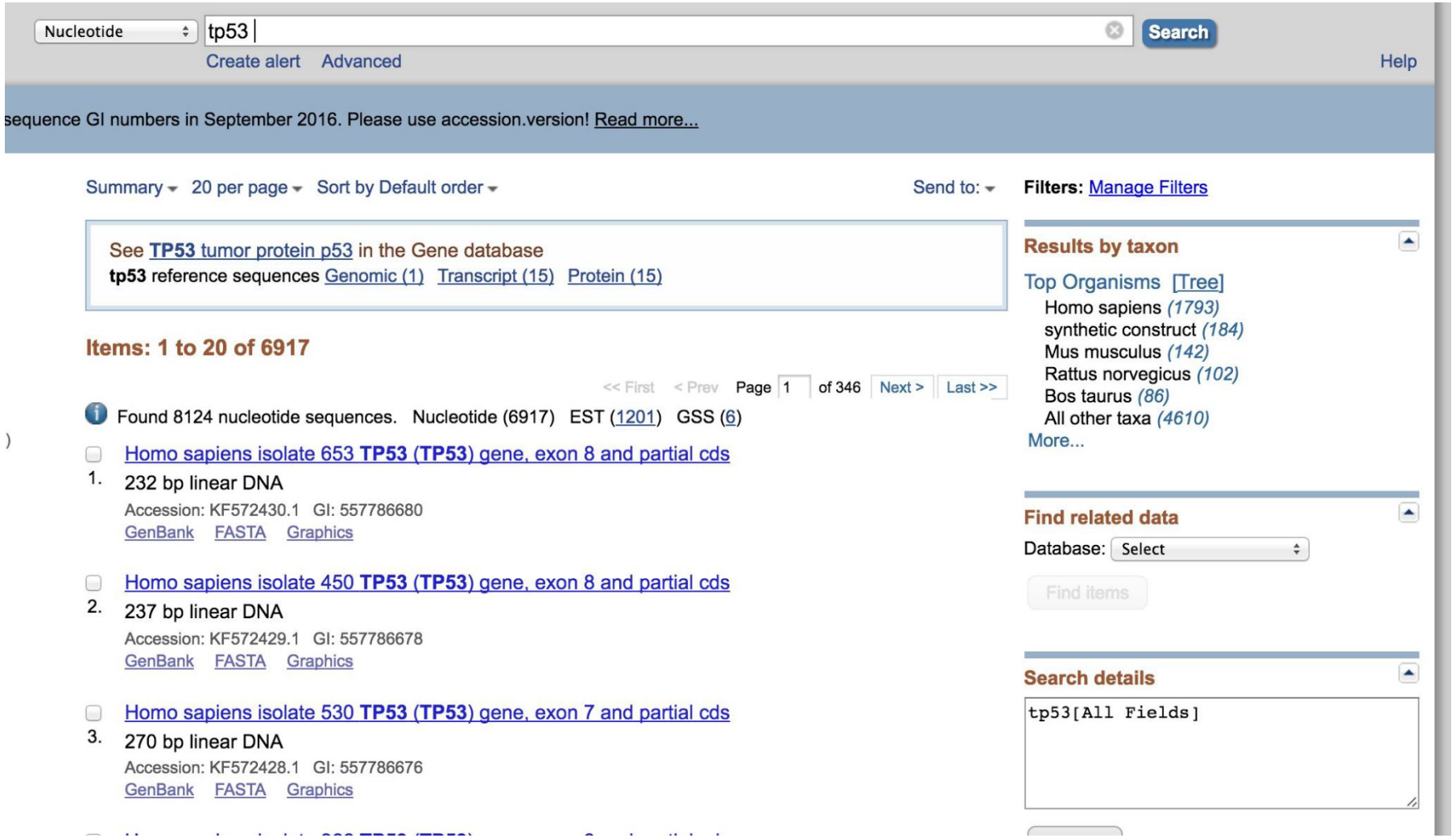
interfaccia di ricerca di Genbank

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

LE BANCHE DATI

- API

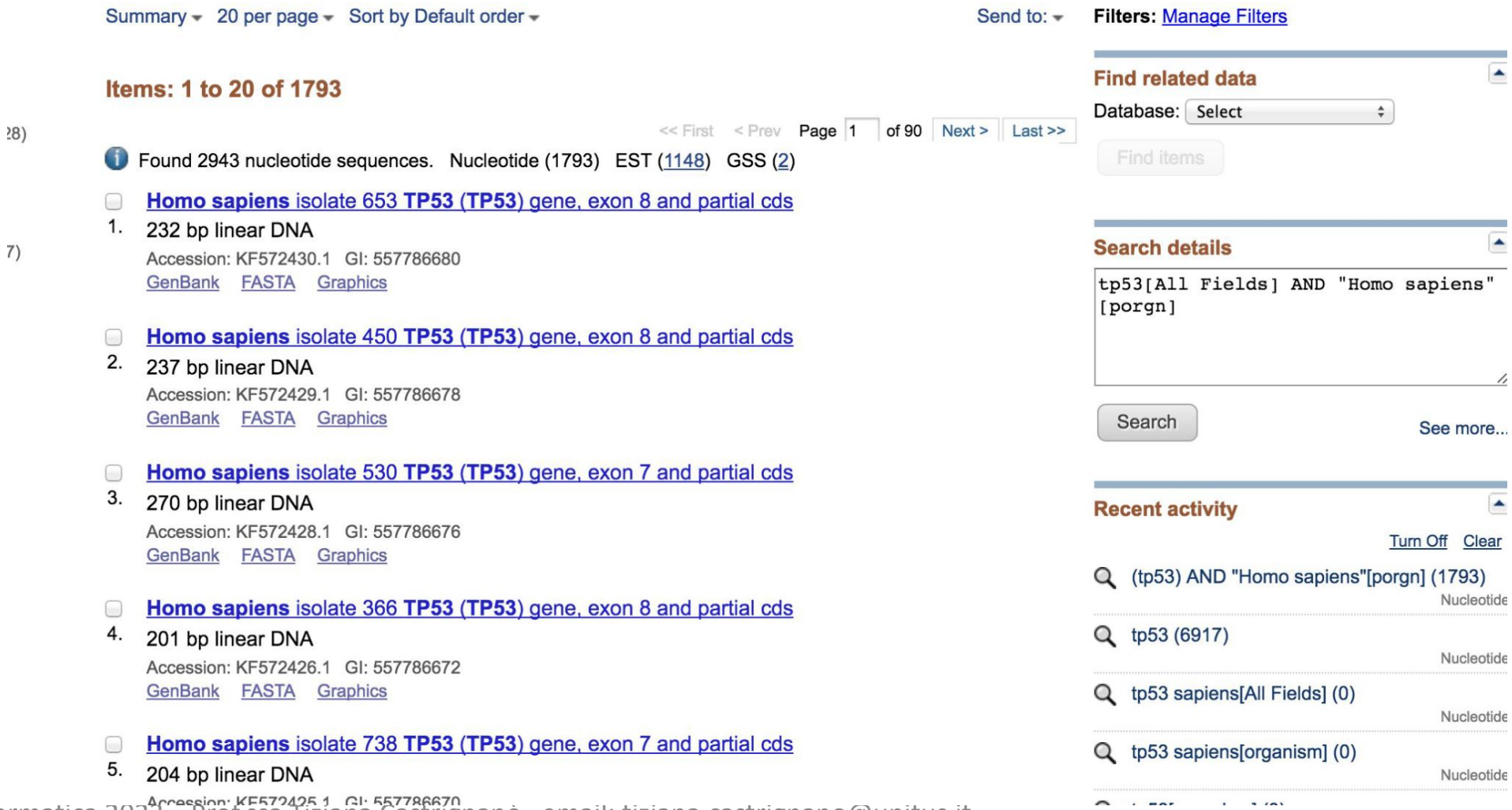
Se cerco una parola chiave generica,ad es. “Tp53” allora il risultato saranno tutte le entry che contengono quella parola in almeno uno dei loro campi.



LE BANCHE DATI

- API

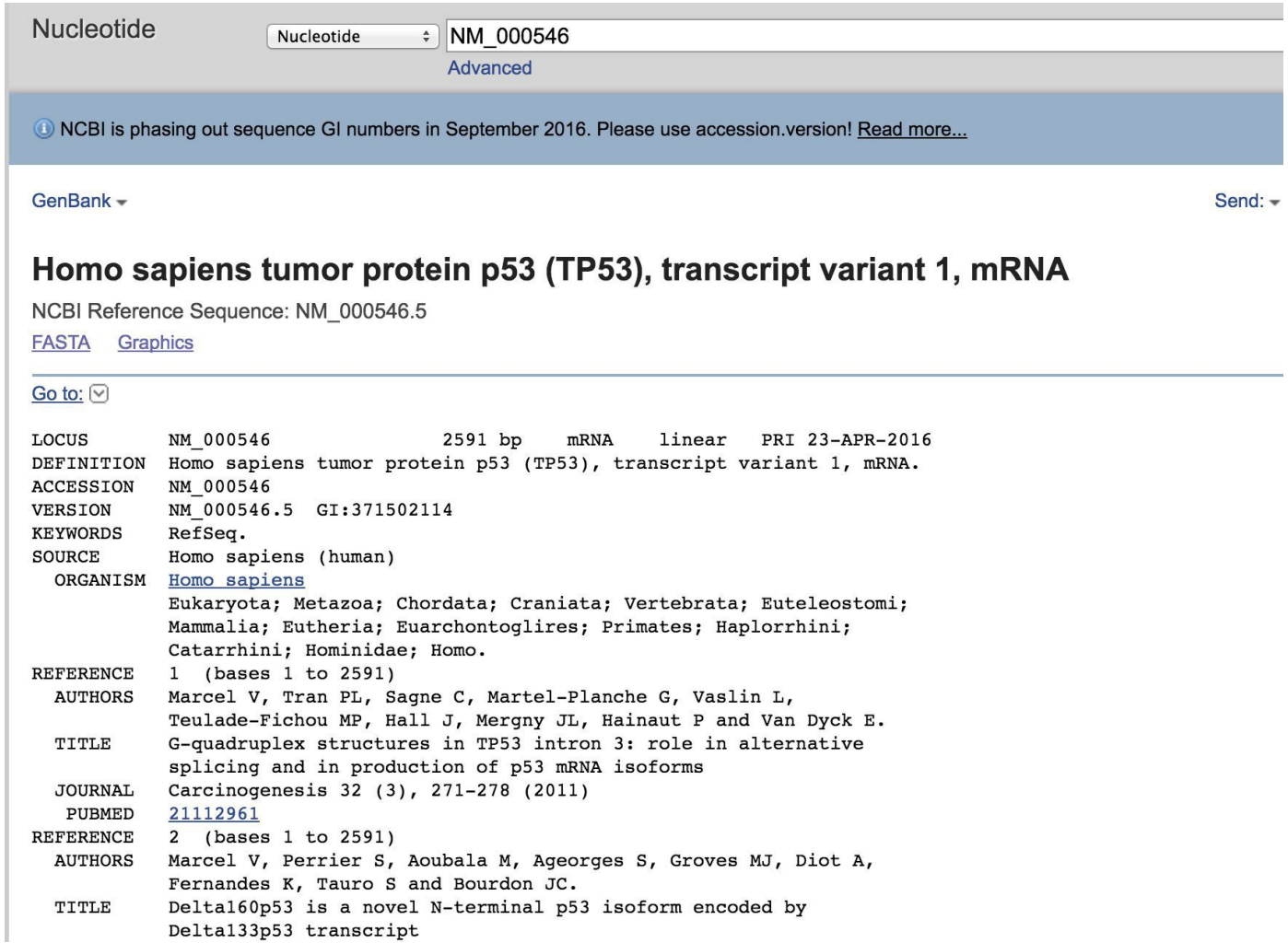
Posso usare il pannello a destra per filtrare i risultati per specie. Ad es. Queste sono le sequenze di uomo le cui entry contengono la parola tp53.



LE BANCHE DATI

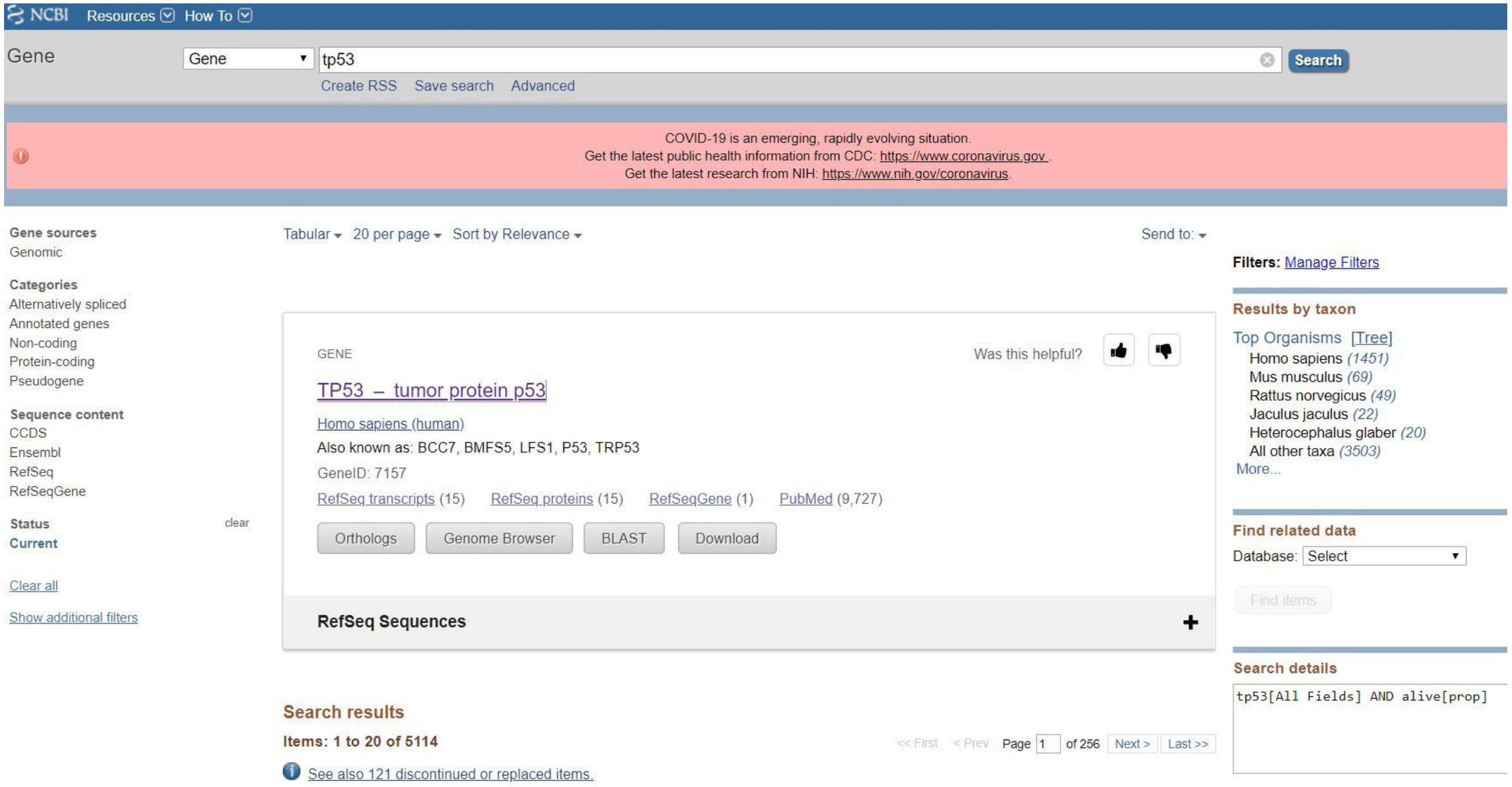
- API

Se invece conosco l'identificativo univoco della mia sequenza, ad es. NM_000546, posso scrivere quello e verrò portato direttamente alla entry corrispondente, visto che in questa caso c'è una ed una sola entry associata a quell'identificativo.



LE BANCHE DATI

- API



LE BANCHE DATI

- API

TP53 tumor protein p53 [*Homo sapiens* (human)]

Gene ID: 7157, updated on 15-Mar-2020

Summary

Official Symbol	TP53 provided by HGNC
Official Full Name	tumor protein p53 provided by HGNC
Primary source	HGNC:HGNC:11998
See related	Ensembl:ENSG00000141510 MIM:191170
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	P53; BCC7; LFS1; BMFS5; TRP53
Summary	This gene encodes a tumor suppressor protein containing transcriptional activation, DNA binding, and oligomerization domains. The encoded protein responds to diverse cellular stresses to regulate expression of target genes, thereby inducing cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. Mutations in this gene are associated with a variety of human cancers, including hereditary cancers such as Li-Fraumeni syndrome. Alternative splicing of this gene and the use of alternate promoters result in multiple transcript variants and isoforms. Additional isoforms have also been shown to result from the use of alternate translation initiation codons from identical transcript variants (PMIDs: 12032546, 20937277). [provided by RefSeq, Dec 2016]
Expression	Ubiquitous expression in spleen (RPKM 13.2), lymph node (RPKM 13.1) and 25 other tissues See more
Orthologs	mouse all

HGNC

HUGO Gene Nomenclature Committee

The resource for approved human gene nomenclature

Search all | Search symbols, keywords or IDs

Last updated: 2020-03-20

LE BANCHE DATI

- API

TP53 tumor protein p53 [*Homo sapiens* (human)]

Gene ID: 7157, updated on 15-Mar-2020

Summary

Official Symbol	TP53 provided by HGNC
Official Full Name	tumor protein p53 provided by HGNC
Primary source	HGNC:HGNC:11998
See related	Ensembl:ENSG00000141510 MIM:191170
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	P53; BCC7; LFS1; BMFS5; TRP53
Summary	This gene encodes a tumor suppressor protein containing transcriptional activation, DNA binding, and oligomerization domains. The encoded protein responds to diverse cellular stresses to regulate expression of target genes, thereby inducing cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. Mutations in this gene are associated with a variety of human cancers, including hereditary cancers such as Li-Fraumeni syndrome. Alternative splicing of this gene and the use of alternate promoters result in multiple transcript variants and isoforms. Additional isoforms have also been shown to result from the use of alternate translation initiation codons from identical transcript variants (PMIDs: 12032546, 20937277). [provided by RefSeq, Dec 2016]
Expression	Ubiquitous expression in spleen (RPKM 13.2), lymph node (RPKM 13.1) and 25 other tissues See more
Orthologs	mouse all

HGNC

Search all | Search symbols, keywords or IDs

Gene data | Tools | Downloads | VGNC | Contact us | More | Request symbol

The HGNC resources will be at risk daily between 3am and 9am GMT for approximately 1 hour.

Symbol report for TP53 Stable symbol

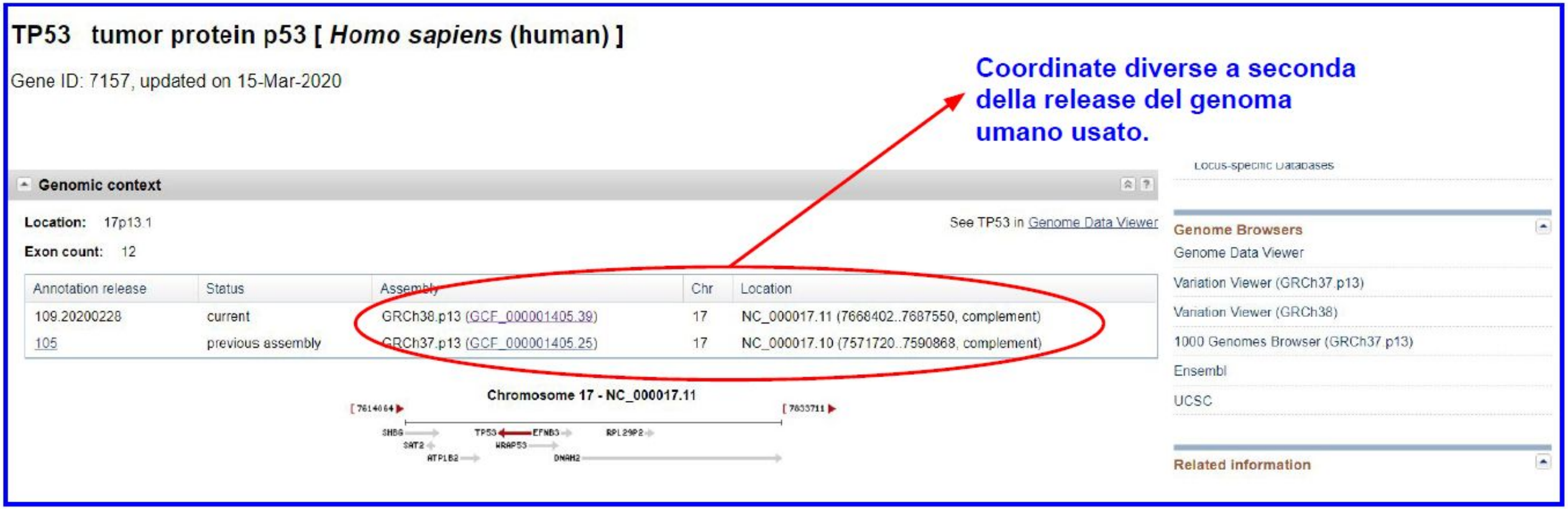
Report | HCOP homology predictions

HGNC data for TP53

Approved symbol	TP53
Approved name	tumor protein p53
Locus type	gene with protein product
HGNC ID	HGNC:11998
Symbol status	Approved
Alias symbols	p53; LFS1
Alias names	" Li-Fraumeni syndrome "
Chromosomal location	17p13.1

LE BANCHE DATI

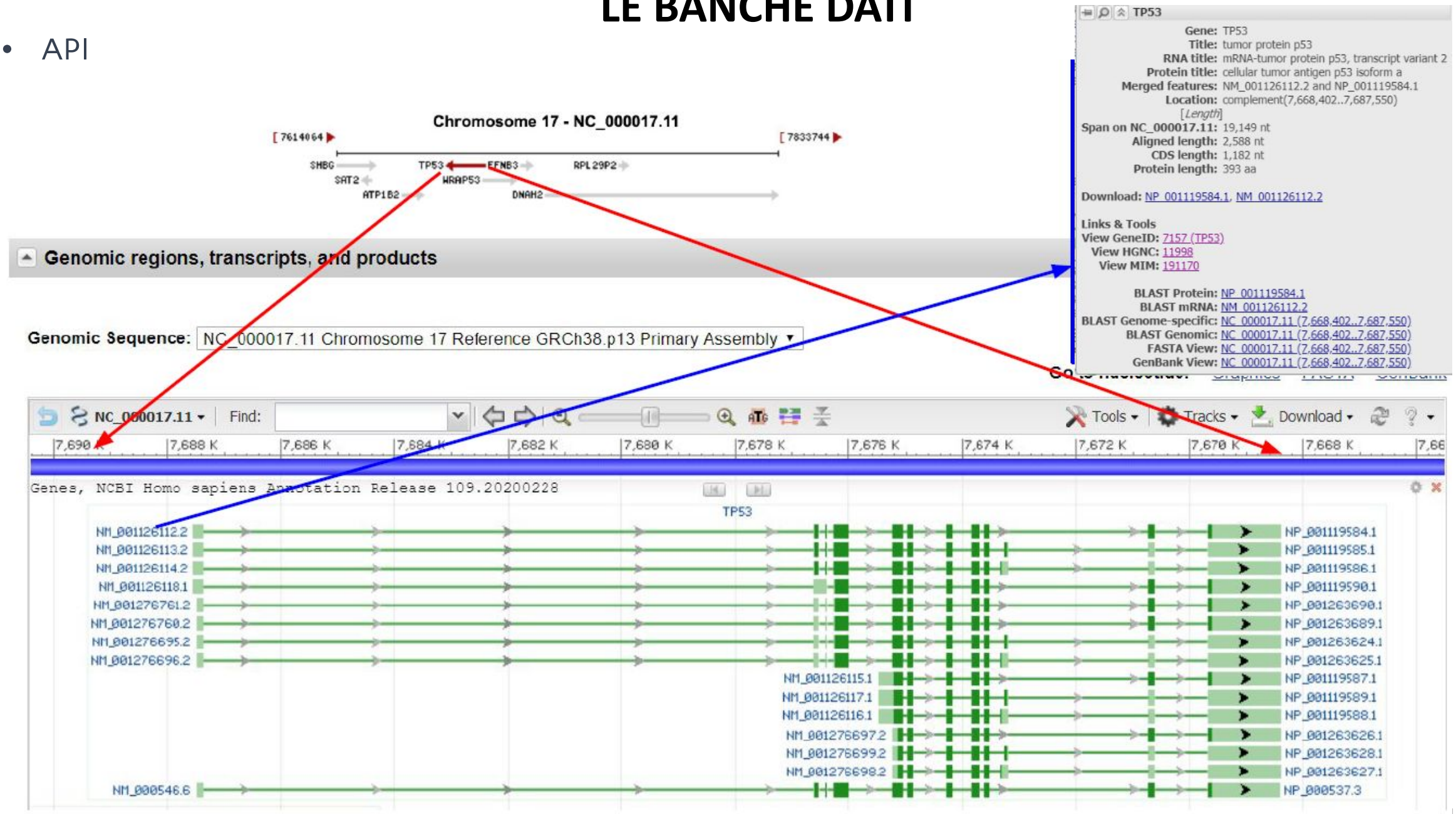
- API



- **GRCh38.p13:** Genome Reference Consortium Human Build 38, patch release 13 (GRCh38.p13)
- **GRCh37.p13:** Genome Reference Consortium Human Build 37, patch release 13 (GRCh37.p13)

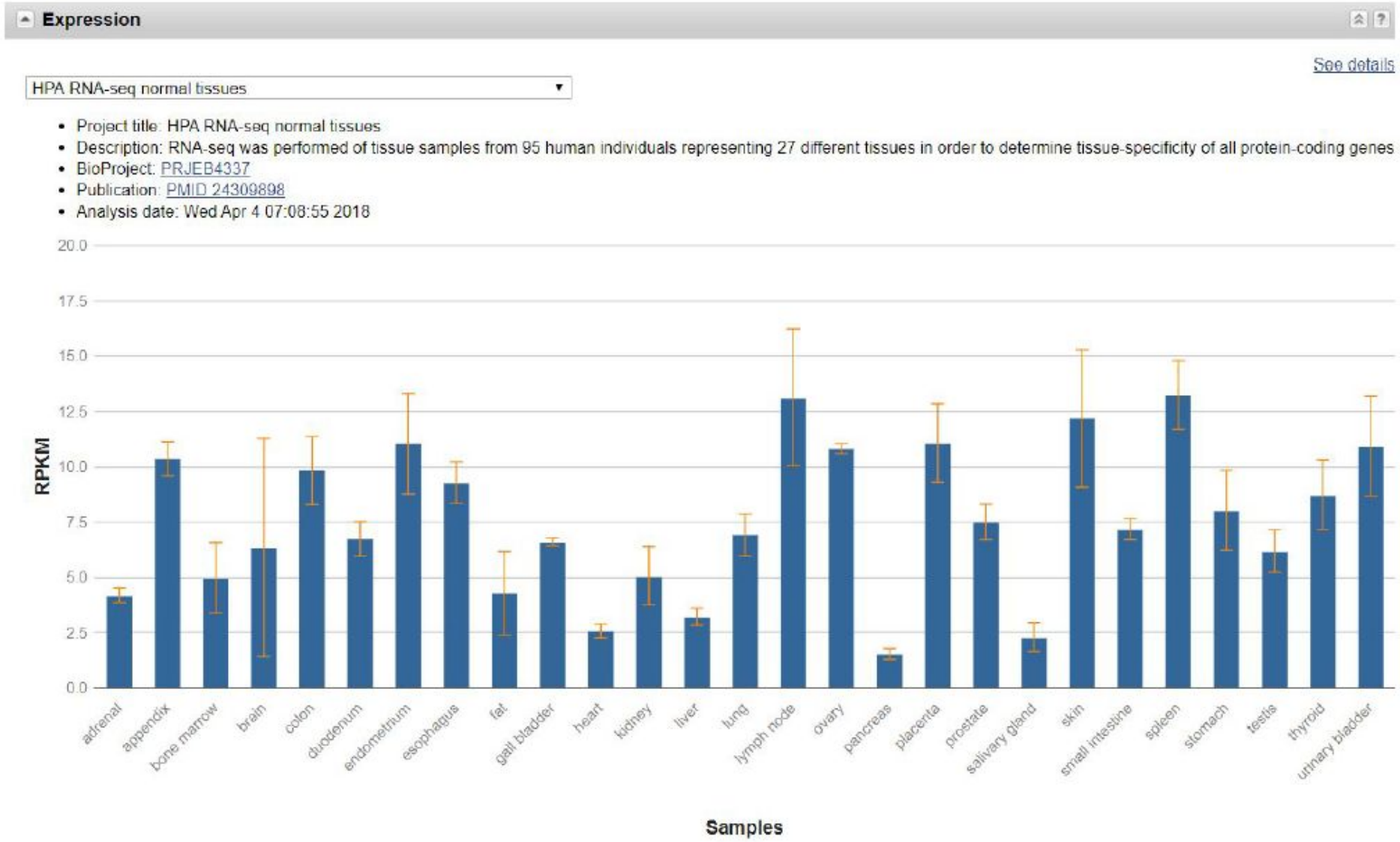
LE BANCHE DATI

- API



LE BANCHE DATI

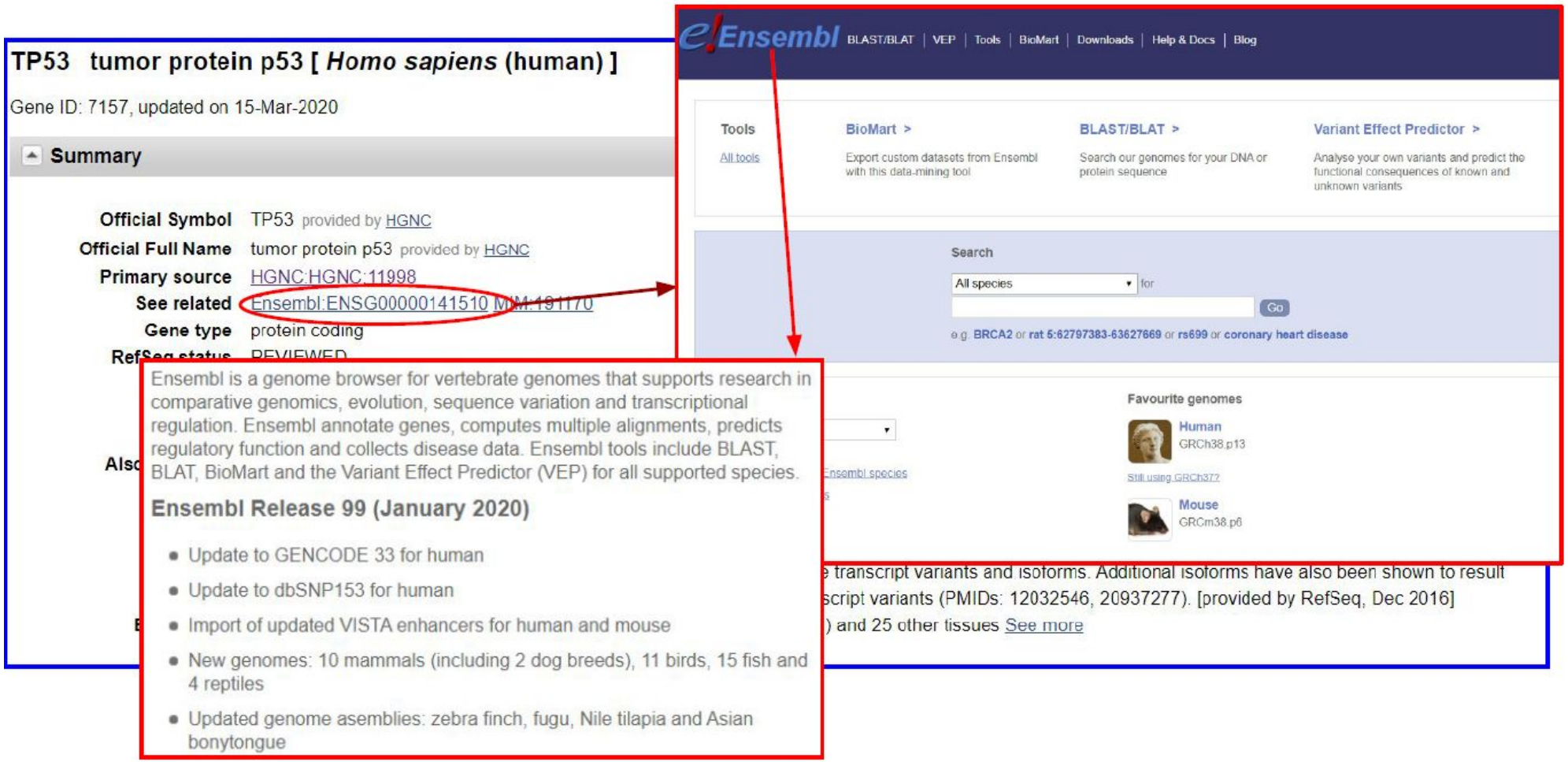
- API



LE BANCHE DATI

- API

La Ensembl API è utilizzata per accedere ai dati contenuti nella banca dati **Ensembl**, che contiene informazioni genomiche di varie specie. Con questa API è possibile effettuare ricerche di geni, trascritti, varianti genetiche e molto altro ancora.



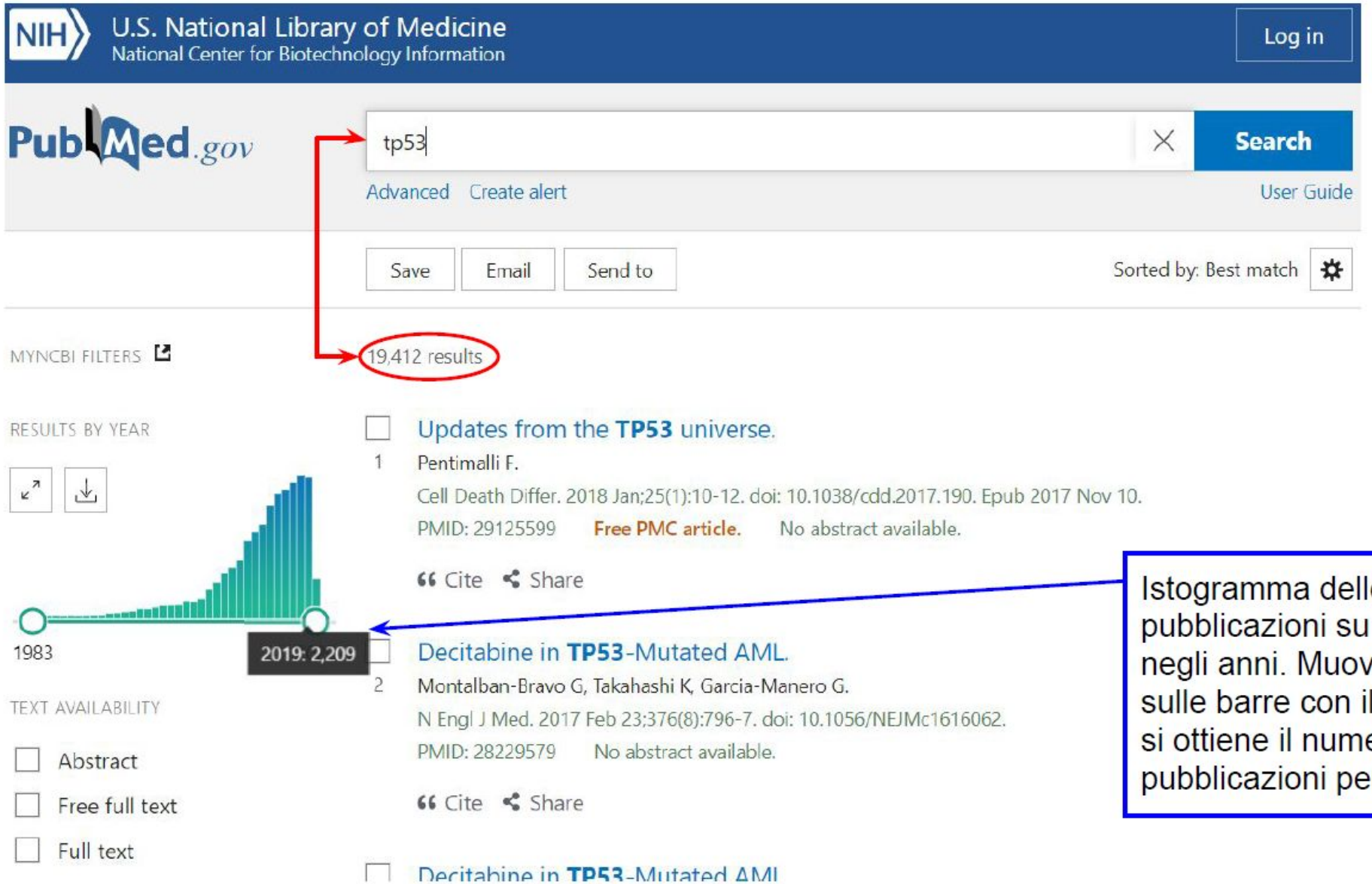
LE BANCHE DATI

- API

L'API di Pubmed è basata sul protocollo HTTP e consente di eseguire ricerche, recuperare i risultati delle ricerche, accedere ai dettagli dei singoli record.

LE BANCHE DATI

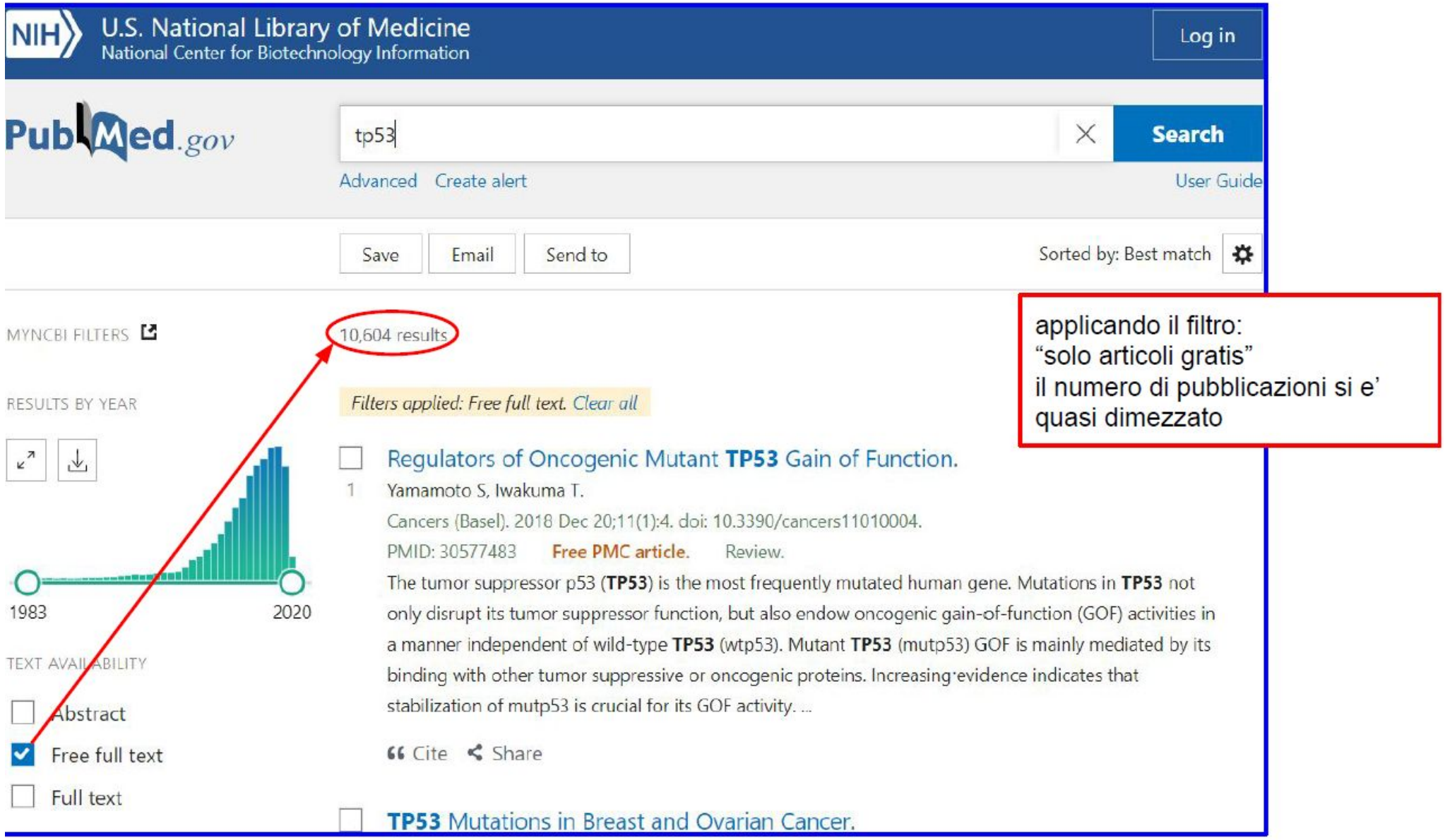
- API



Istogramma delle pubblicazioni su **tp53** negli anni. Muovendosi sulle barre con il cursore si ottiene il numero di pubblicazioni per anno

LE BANCHE DATI

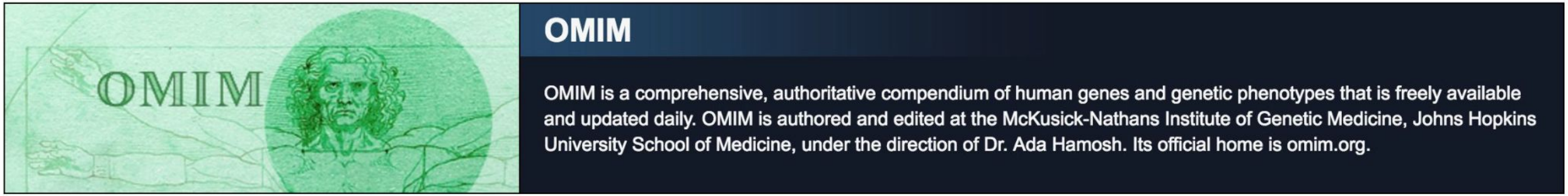
- API

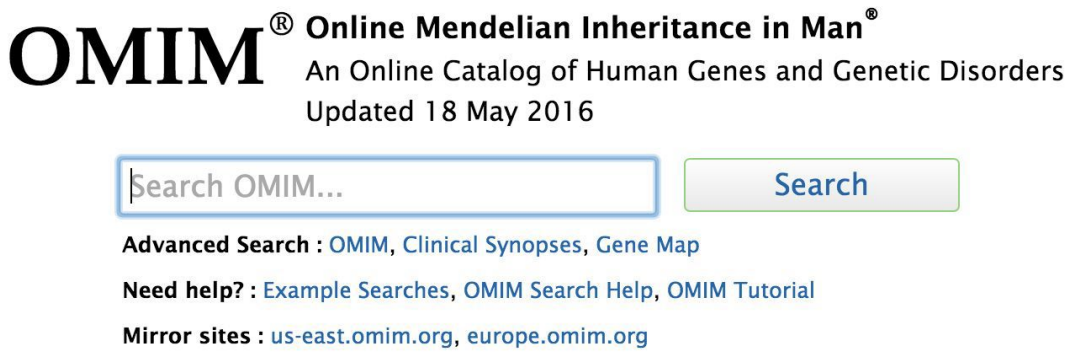


LE BANCHE DATI

- API

Omim (Online Mendelian Inheritance in Man) è una banca dati curata manualmente che cataloga le associazioni tra geni e fenotipi (leggi: malattie) umani. Ci sono quindi delle persone si occupano di leggersi la letteratura scientifica e di aggiornare manualmente il catalogo!



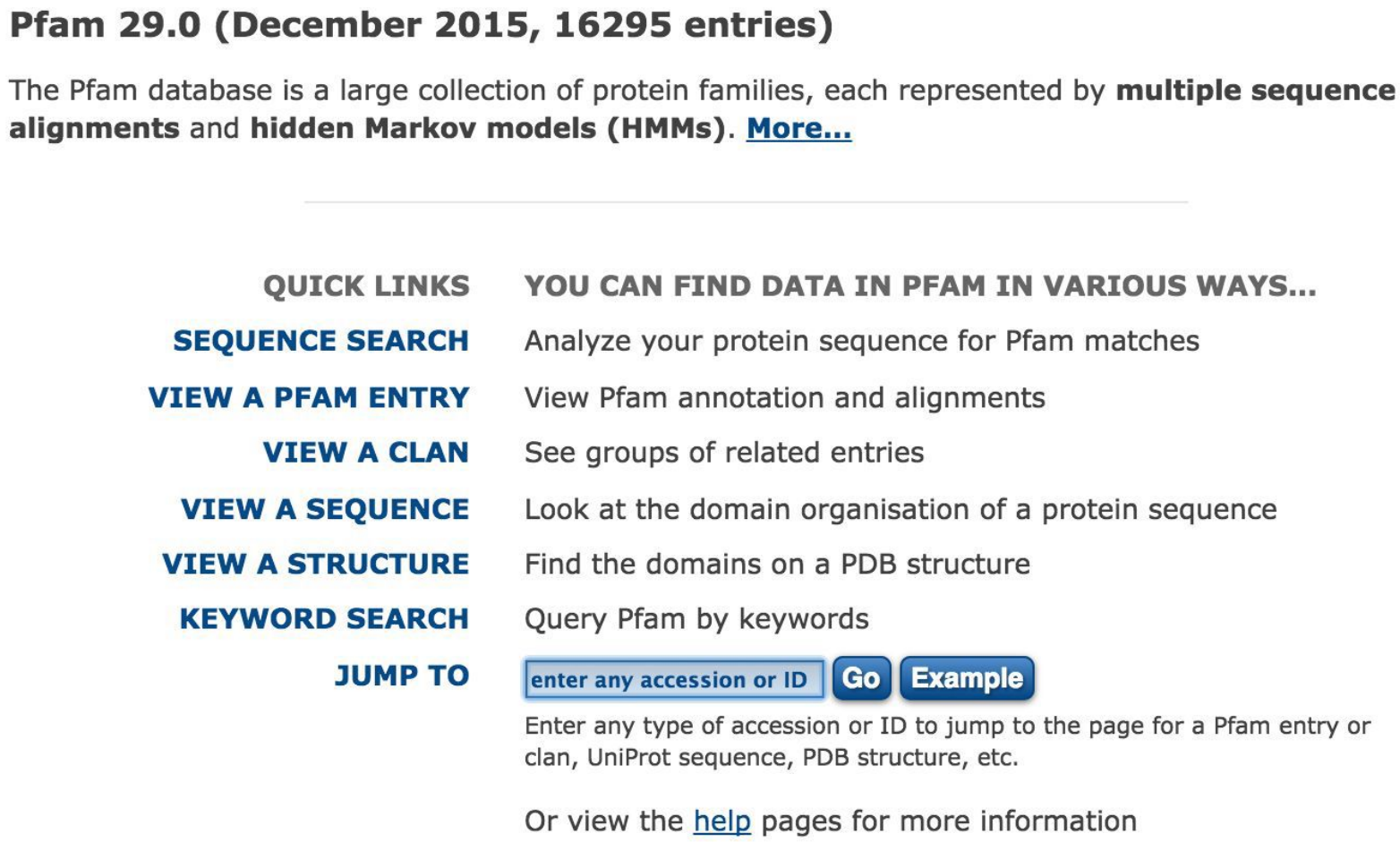


https://www.ncbi.nlm.nih.gov/omim

LE BANCHE DATI

- API

Pfam è una collezione di famiglie proteiche, si possono esaminare i multiallineamenti tra le stesse ed informazioni sui domini che le compongono.



LE BANCHE DATI

- API

Per ogni dominio potete visualizzare altre informazioni, ad esempio la sua struttura tridimensionale predetta o reale quando è nota.

No Wikipedia article | Pfam | InterPro

This tab holds the annotation information that is stored in the Pfam database. As we move to using Wikipedia as our main source of annotation, the contents of this tab will be gradually replaced by the Wikipedia tab.

Hedgehog amino-terminal signalling domain

Provide feedback

For the carboxyl Hint module, see PF01079. Hedgehog is a family of secreted signal molecules required for embryonic cell differentiation.

Literature references

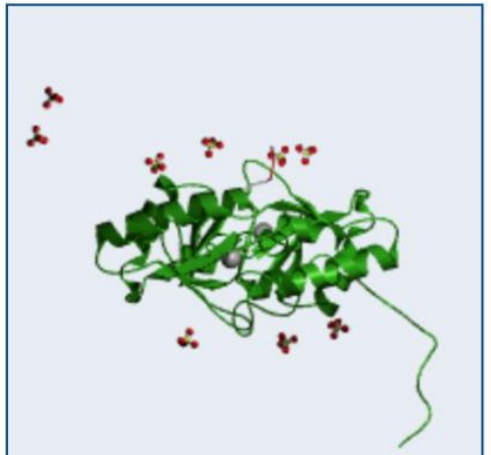
1. Hall TM, Porter JA, Beachy PA, Leahy DJ; , Nature 1995;378:212-216.: A potential catalytic site revealed by the 1.7-A crystal structure of the amino-terminal signalling domain of Sonic hedgehog. PUBMED:7477329 EPMC:7477329

Internal database links

SCOOP:	Peptidase_M15_3

External database links

MEROPS:	C46
SCOP:	1vhh



Example structure

PDB entry 3M1N: Crystal structure of Human Sonic Hedgehog N-terminal domain

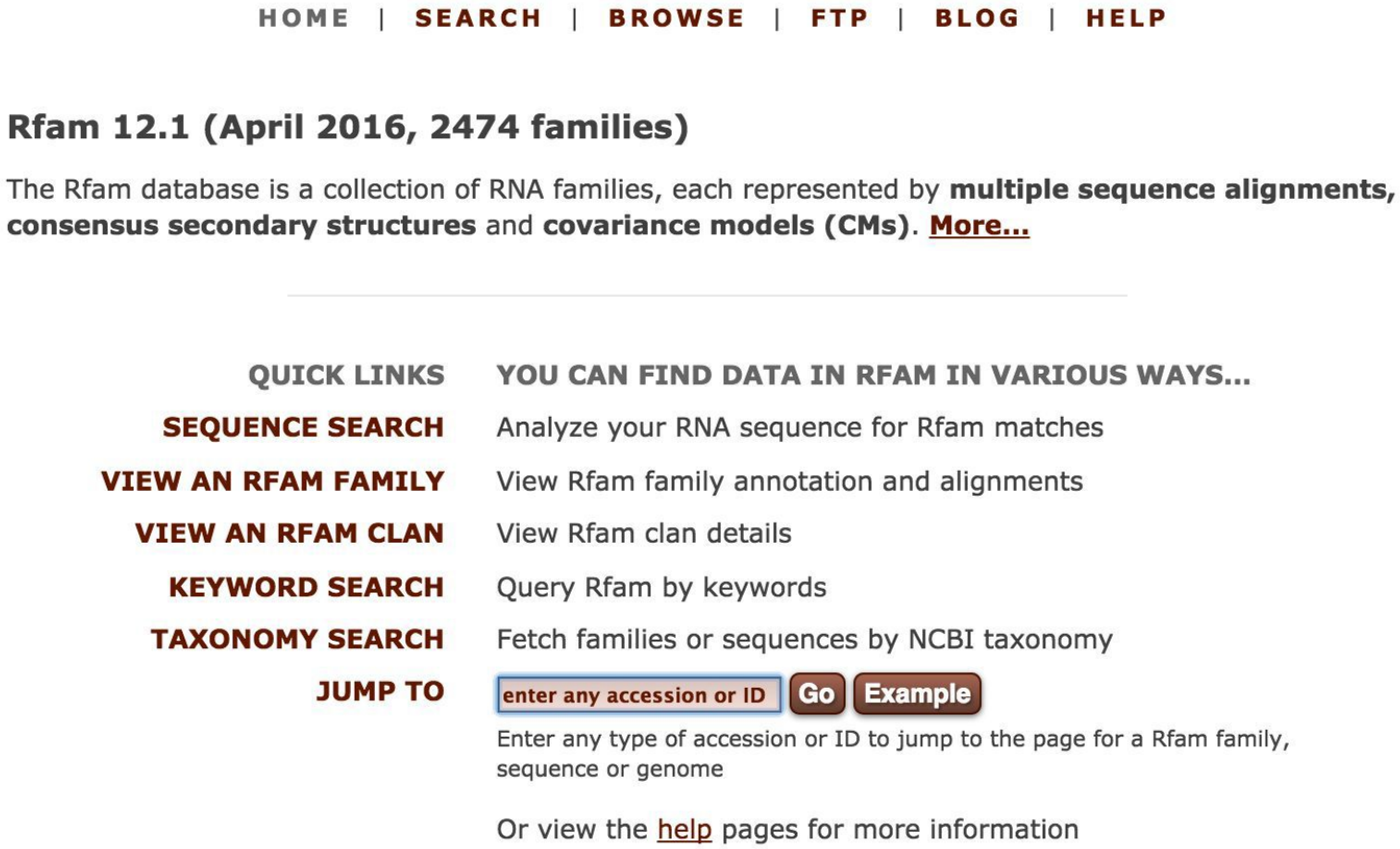
View a different structure:

3M1N

LE BANCHE DATI

- API

Rfam invece raccoglie in famiglie le strutture secondarie degli RNA. All'interno di una famiglia di RNA la conservazione della struttura secondaria rispetto alla primaria (sequenza) è ancora più marcata che non per le proteine.



LE BANCHE DATI

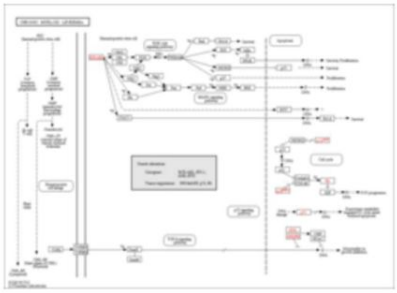
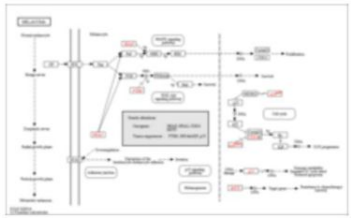
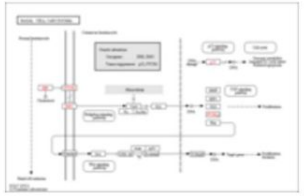
- API

● Kegg è una banca dati curata a mano di pathway , ossia di processi ed interazioni metaboliche. Possiamo fare ricerche per gene, malattia, ligando, farmaco, etc...

Pathway Text Search

Number of entries in a page 20 Hide thumbnail

Page : 1 Go of 2 Items : 1 - 20 of 37 Top Previous Next Bottom

Entry	Thumbnail Image	Name	Description	Object	Legend
map05220		Chronic myeloid leukemia	...on secondary abnormalities include mutations in TP53, RB, and p16/INK4A, or overexpression of genes ...	...3 (PTPN11) K06619 (ABL1), K08878 (BCR1) K04451 (TP53) K04451 (TP53) K13375 (TGFB1), K13376 (TGFB2), ...	p53 signaling pathway AML-EVI1 Bcl-xl Shp2 BCR-ABL p53 TGF-β signaling pathway Hematopoietic cell l...
map05218		Melanoma	...F/HMD2/p53) tumor suppressor pathways. MITF and TP53 are implicated in further melanoma progression.	...sphosphate) K04365 (BRAF) K05689 (CDHE) K04451 (TP53) K07828 (NRAS) K02089 (CDK4) K04503 (CCND1) K09...	p53 signaling pathway Melanogenesis BRAF ECAD p53 Adherens junction NRAS CDK4 CyclinD1 MITF PTEN RTK...
map05217		Basal cell carcinoma	...aling promotes cell proliferation. Mutations in TP53 are also found with high frequency (>50%) in sp...	...(PTCH1), K11101 (PTCH2) K04662 (BMP2_4) K04451 (TP53) K00182 (WNT2), K00312 (WNT3), K00408 (WNT4), K...	Cos2 Cholesterol GLI1 PTCH1 BMP p53 Wnt SHH APC PTCH1 Su (fu) Axin Dvl Frizzled BASAL CELL CARCINOM...

LE BANCHE DATI

- Riassumendo…

Le principali differenze tra i database **relazionali** e quelli **non relazionali** sono:

1. Struttura dei dati: i database relazionali organizzano i dati in tabelle con relazioni predefinite tra di esse, mentre i database non relazionali possono utilizzare diversi formati di archiviazione, come i documenti, le coppie chiave-valore o i grafi.

1. Flessibilità: i database non relazionali offrono maggiore flessibilità nella gestione dei dati. I database relazionali sono invece più rigidi nella struttura dei dati e richiedono uno schema predefinito.

1. Interrogazione dei dati: nei database relazionali l'interrogazione dei dati è eseguita attraverso il linguaggio SQL, mentre nei database non relazionali l'interrogazione dei dati può essere fatta attraverso una varietà di metodi e linguaggi specifici per ogni database.

LE BANCHE DATI

- Tipi di dati

Le banche dati biologiche possono contenere una vasta gamma di informazioni, tra cui:

1. Informazioni sulle **sequenze**: queste informazioni includono le sequenze di DNA, RNA e proteine, così come le loro annotazioni, come le regioni codificanti, le regioni non codificanti, i siti di legame dei fattori di trascrizione, le mutazioni e le varianti.

1. Informazioni sui **geni** e sui **trascritti**: queste informazioni riguardano la posizione dei geni nel genoma, la struttura dei trascritti, l'espressione genica, la regolazione genica e le variazioni genetiche.

1. Informazioni sulla **funzione** biologica delle proteine: queste informazioni includono le funzioni biologiche delle proteine, le interazioni proteina-proteina e proteina-ligando, le vie metaboliche, le pathway di segnalazione cellulare e altre informazioni relative alla biologia molecolare.

LE BANCHE DATI

- Tipi di dati

4. Informazioni sull'**organizzazione genomica**: queste informazioni riguardano l'organizzazione del genoma, come la mappatura fisica e genetica, la struttura cromosomica, le ripetizioni genomiche e le regioni di instabilità genomica.

4. Informazioni sulla **variazione genetica**: queste informazioni riguardano le variazioni genetiche tra gli individui, come i polimorfismi, le mutazioni e le varianti.

4. Informazioni sulla **letteratura scientifica**: molte banche dati biologiche includono anche informazioni sulla letteratura scientifica, come abstracts di articoli e link agli articoli completi.

LE BANCHE DATI

- Formato dei dati

Ci sono diversi formati di dati utilizzati nelle banche dati biologiche, ciascuno con le proprie caratteristiche e utilizzo. Alcuni esempi di formati di dati utilizzati nelle banche dati biologiche includono:

- FASTA: è un formato per rappresentare sequenze di DNA o proteine. Consiste in una descrizione di testo della sequenza, preceduta da un'intestazione che ne fornisce informazioni come l'identificatore e la fonte.

LE BANCHE DATI

- Formato dei dati

- GenBank: è un formato di file sviluppato dal National Center for Biotechnology Information (NCBI) per rappresentare sequenze di DNA. È uno standard internazionale utilizzato da molte banche dati biologiche e contiene informazioni sulla sequenza, la sua annotazione e i riferimenti bibliografici.

GENBANK FILE FORMAT

```
DEFINITION  Sodalis glossinidius str. 'morsitans', complete genome.
ACCESSION   NC_007712
VERSION     NC_007712.1  GI:85057978
KEYWORDS    .
```

LE BANCHE DATI

- Formato dei dati

- GFF (General Feature Format): è un formato di file per la descrizione di annotazioni genomiche. È spesso utilizzato per descrivere le posizioni di geni, trascritti, esoni e altre caratteristiche genomiche.

```
chr1 . mRNA          1300  9000  .  +  .  ID=mrna0001;Name=sonichedgehog
chr1 . exon          1300  1500  .  +  .  ID=exon00001;Parent=mrna0001
chr1 . exon          1050  1500  .  +  .  ID=exon00002;Parent=mrna0001
```

LE BANCHE DATI

- Formato dei dati

- PDB (Protein Data Bank): è un formato di file utilizzato per la descrizione della struttura tridimensionale delle proteine. Contiene informazioni sui singoli atomi che compongono la proteina, la loro posizione nello spazio e la loro interazione con altri atomi nella proteina.

```
ATOM      1  N   GLY A  32      -43.033  14.797   5.823  1.00179.38           N
ATOM      2  CA  GLY A  32      -41.820  14.137   5.372  1.00182.90           C
ATOM      3  C   GLY A  32      -40.823  13.915   6.493  1.00183.40           C
ATOM      4  O   GLY A  32      -41.157  14.065   7.669  1.00182.50           O
ATOM      5  N   THR A  33      -39.594  13.556   6.128  1.00184.33           N
ATOM      6  CA  THR A  33      -38.533  13.328   7.107  1.00183.62           C
ATOM      7  C   THR A  33      -38.428  14.501   8.079  1.00185.09           C
ATOM      8  O   THR A  33      -38.248  14.310   9.283  1.00182.96           O
ATOM      9  CB  THR A  33      -37.167  13.103   6.425  1.00177.96           C
```

LE BANCHE DATI

- Formato dei dati

- VCF (Variant Call Format): è un formato di file utilizzato per descrivere varianti genomiche, come singole sostituzioni di nucleotidi o inserzioni e delezioni di segmenti di DNA.

  -I file VCF contengono informazioni sulla **posizione** genomica delle varianti, la loro **frequenza** nelle popolazioni, l'**effetto** sulla funzione delle proteine e altre annotazioni pertinenti.

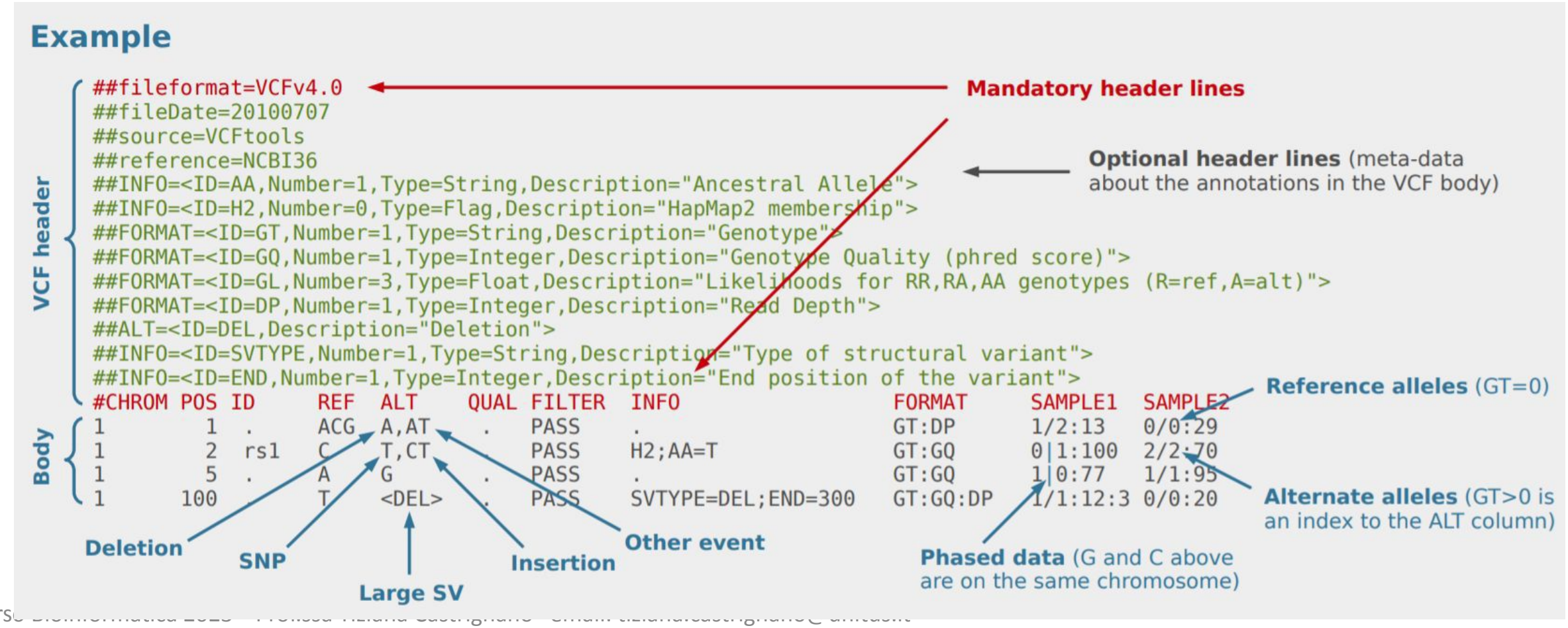
  -è stato sviluppato per essere estensibile e adattabile alle diverse esigenze degli utenti. Inoltre, i ricercatori possono anche definire campi personalizzati per memorizzare informazioni aggiuntive sulle varianti.

  -Il formato VCF è utilizzato in diversi ambiti della genomica, come la genomica delle popolazioni, la genomica clinica e la genomica comparativa.

LE BANCHE DATI

- Formato dei dati

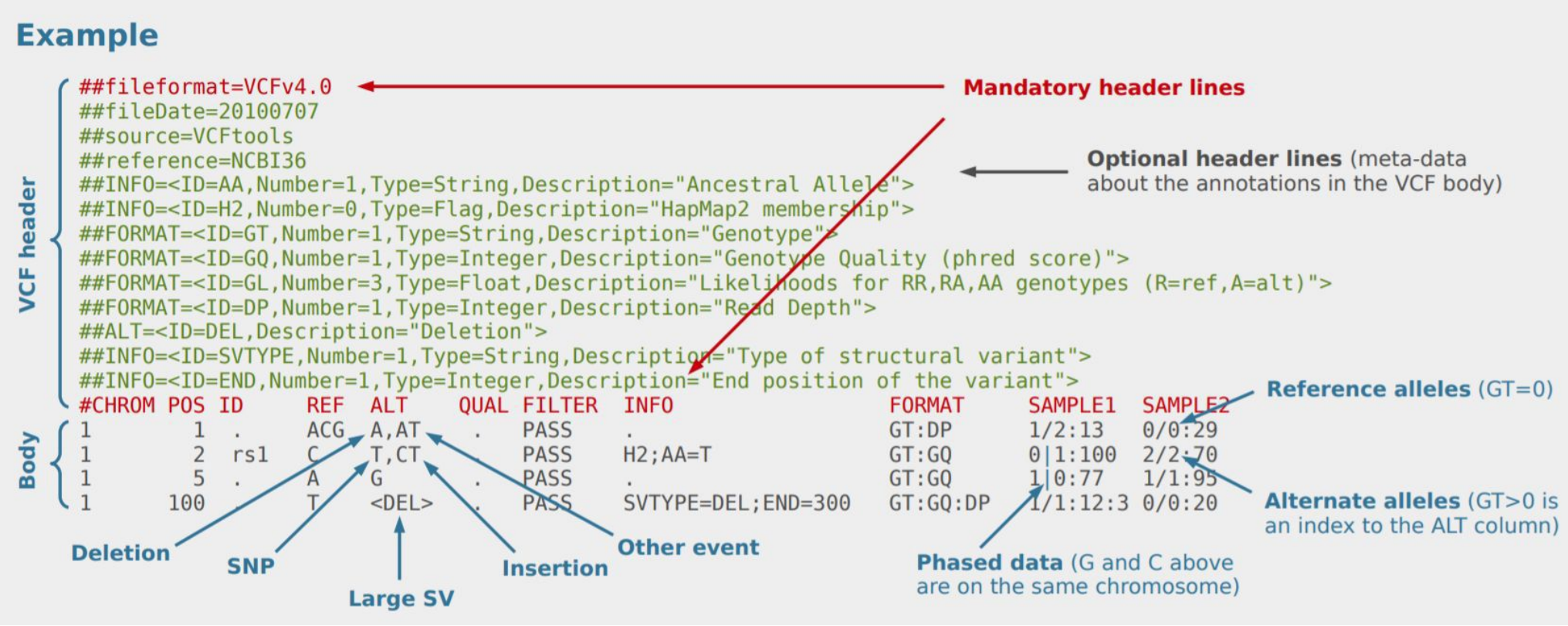
Il file VCF è costituito da un'intestazione (**header**) e da una serie di righe contenenti le informazioni sulle varianti (**body**).



LE BANCHE DATI

- Formato dei dati

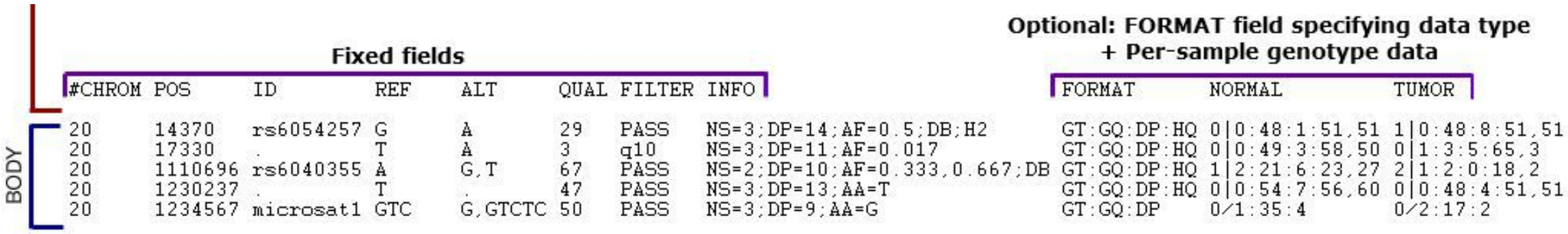
L'intestazione contiene le informazioni sulle annotazioni utilizzate nel file, come ad esempio i nomi dei campi e il formato dei dati. Le righe che seguono l'intestazione, rappresentano le varianti e contengono informazioni come il nome del cromosoma, la posizione genomica, il tipo di variante, la qualità della chiamata, la frequenza nella popolazione e le annotazioni sulle funzioni biologiche.



LE BANCHE DATI

- Formato dei dati

Nel dettaglio..
Le nove colonne del “body” di un file VCF sono:
1. CHROM: il cromosoma (contig) su cui si verifica la variante.
2. POS: Le coordinate genomiche in cui si verifica la variante. Per le delezioni, la posizione indicata sono le basi che precedono l'evento.
3. ID: Un identificatore della variante (se esiste). In genere un database dbSNP, se noto.
4. REF: L'allele di riferimento sul filamento anteriore.
5. ALT: l'allele o gli alleli alternativi sul filamento anteriore. Possono esserne presenti più di uno.

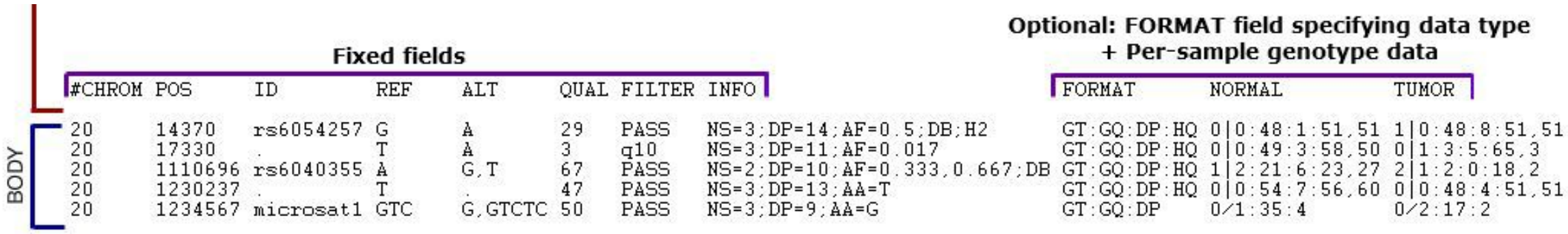


#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NORMAL	TUMOR
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51
20	1234567	microsat1	GTC	G,GTCTC	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2

LE BANCHE DATI

- Formato dei dati

6. QUAL: La probabilità che la variante REF/ALT esista in questo sito. È in scala Phred, come la qualità FASTQ e la qualità MAPQ.
qualità FASTQ e il campo MAPQ nel file SAM.
7. FILTRO: il nome dei filtri che la variante non ha superato, o il valore PASS se la variante ha superato tutti i filtri.
ha superato tutti i filtri. Se il valore FILTER è ., non è stato applicato alcun filtro al record.
8. INFO: Contiene le annotazioni specifiche del sito rappresentate in formato ID=VALORE.
9. FORMAT: annotazioni a livello di campione come TAGS separati da due punti.



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NORMAL	TUMOR
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51
20	1234567	microsat1	GTC	G,GTCTC	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2

LE BANCHE DATI

- Formato dei dati

HEADER

```
##fileformat=VCFv4.1
##fileDate=20090805
##tcgaversion=1.1
##vcfProcessLog=<InputVCF=<file1.vcf>,InputVCFSource=<caller1>,InputVCFVer=<1.0>,InputVCFParam=<a1,b>,InputVCFgeneAnno=<anno1.gaf>>
##reference=ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.fa
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
```

INFO meta-information

```
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
```

FILTER meta-information

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

FORMAT meta-information

```
##SAMPLE=<ID=NORMAL,Individual=TCGA-01-1000,File=TCGA-01-1000-1.bam,Platform=Illumina,Source=dbGAP,Accession=1234>
##SAMPLE=<ID=TUMOR,Individual=TCGA-01-1000,File=TCGA-01-1000-2.bam,Platform=Illumina,Source=dbGAP,Accession=4567>
##PEDIGREE=<Name_0=TUMOR,Name_1=NORMAL>
```

Fixed fields

Optional: FORMAT field specifying data type + Per-sample genotype data

BODY

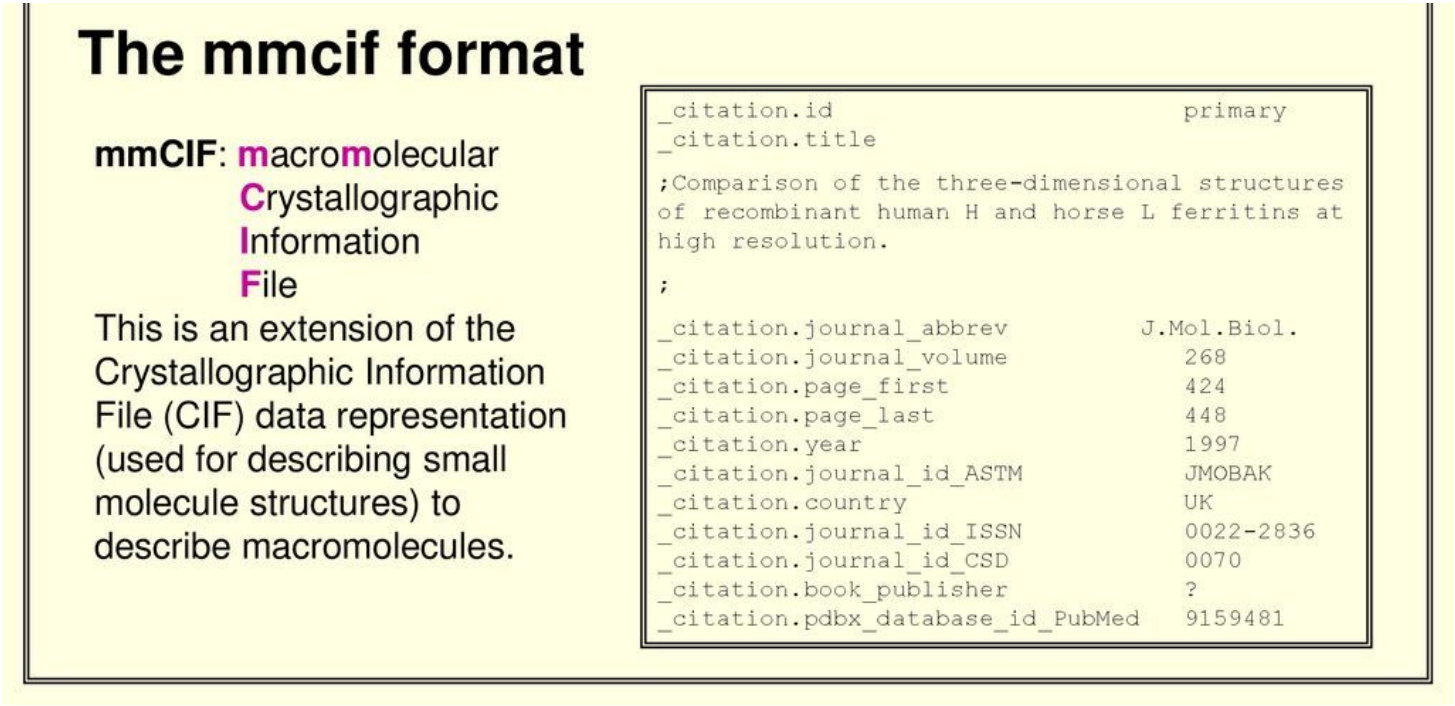
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NORMAL	TUMOR
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51
20	1234567	microsat1	GTC	G,GTCTC	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2

LE BANCHE DATI

- Formato dei dati

Esistono anche formati di file utilizzati per la descrizione della struttura e dell'allineamento di sequenze di RNA:

- formato mmCIF (macromolecular Crystallographic Information File) è un formato di file utilizzato per descrivere la struttura tridimensionale di macromolecole biologiche determinate attraverso la cristallografia a raggi X.



LE BANCHE DATI

- Formato dei dati

- formato Stockholm, invece, è utilizzato principalmente per rappresentare allineamenti di sequenze di RNA, ma può anche essere utilizzato per allineamenti di proteine o di DNA. Il formato è basato su un file di testo **tab-delimited**, dove ogni riga rappresenta una singola sequenza e le colonne rappresentano le posizioni allineate delle sequenze. Ogni sequenza ha anche una riga di metadati che fornisce informazioni sulle caratteristiche della sequenza.

```
# STOCKHOLM 1.0
#=GF ID pp
#=GF CLASS small
#=GF FAMILY pancreatic hormone
1bba APLEPEYPGDNATPEQMAQYAAELRRYINMLTRPRY
1ppt GPSQPTYPGDDAPVEDLIRFYDNLQQYLNVVTRHRY
1ron YPSKPDNPGEDAPAEDMARYYSALRHYINLITRQRY
//
```

LE BANCHE DATI

- Sincronizzazione dei dati

La sincronizzazione tra diverse banche dati biologiche può essere ottenuta tramite l'utilizzo di **mirror**.

- In informatica, un "mirror" è una replica di un sito web o di una banca dati che viene mantenuta in sincronia con l'originale, solitamente per scopi di backup o per garantire l'accesso rapido alle informazioni.

- Nel contesto delle banche dati biologiche, un mirror è una copia di una banca dati che viene mantenuta da un'organizzazione diversa da quella che gestisce l'originale, ma che viene sincronizzata regolarmente con l'originale per garantire che i dati siano sempre aggiornati.

LE BANCHE DATI

- Sincronizzazione dei dati

Un esempio di mirror è quello utilizzato da **NCBI**, che fornisce copie delle sue banche dati, come GenBank, Pubmed e altri, a siti mirror in tutto il mondo.

Questi siti mirror contengono una copia identica dei dati presenti nella banca dati originale e vengono sincronizzati regolarmente, generalmente ogni 24 ore, per garantire che i dati siano sempre aggiornati.

Anche **ENA** (European Nucleotide Archive), **DDBJ** (DNA Data Bank of Japan) e altre banche dati biologiche forniscono dei mirror dei loro database.

Questi mirror vengono aggiornati con una frequenza regolare e garantiscono che gli utenti possano accedere ai dati in modo rapido e affidabile, senza dover accedere direttamente alla banca dati originale.