

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.

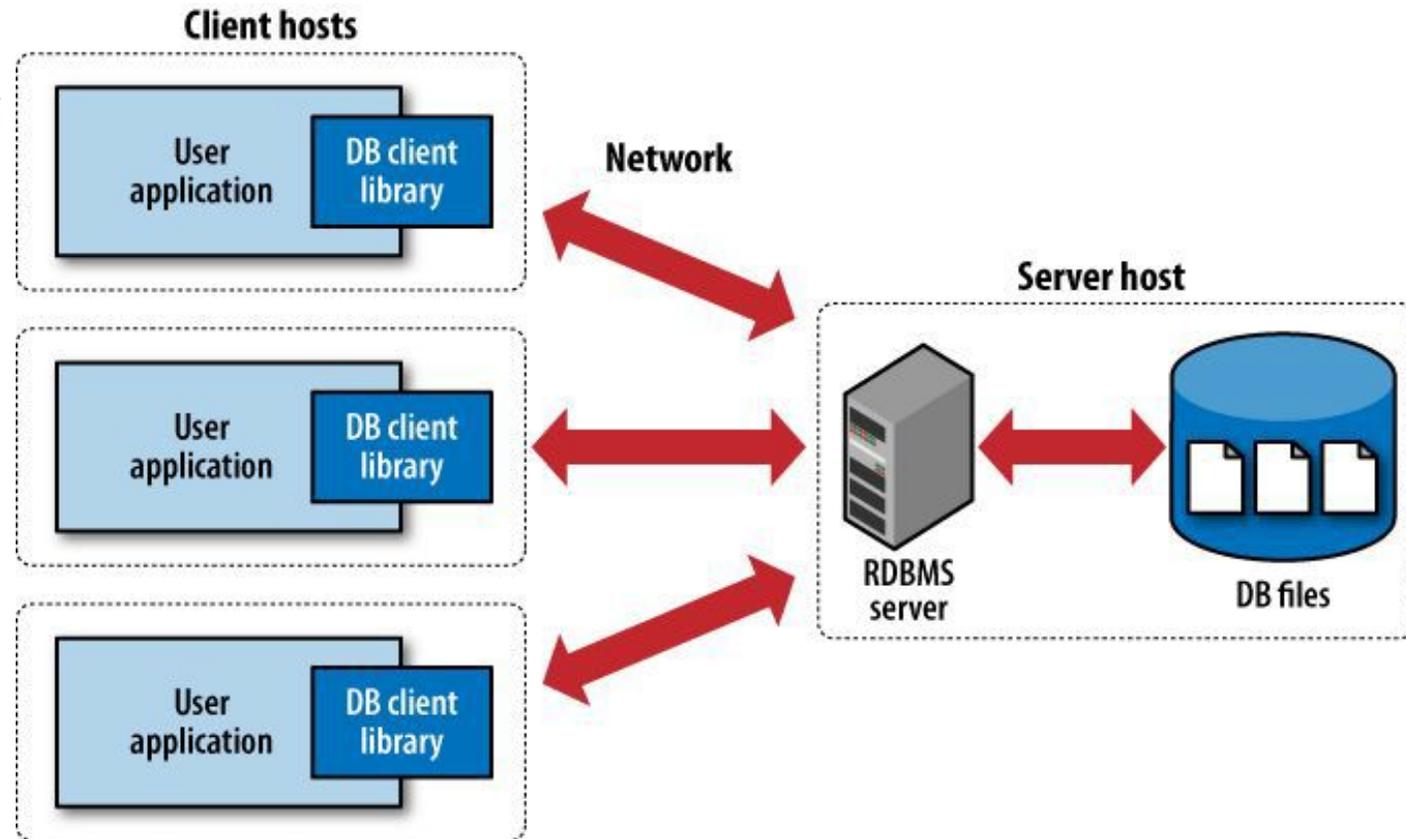


LE BANCHE DATI

- 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.



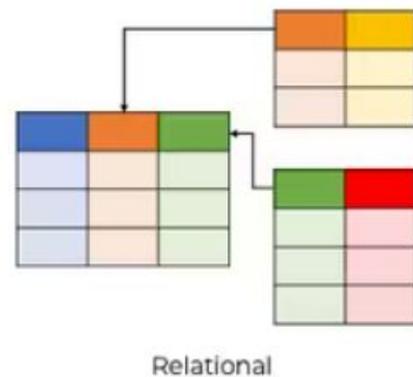
LE BANCHE DATI

- I database non relazionali

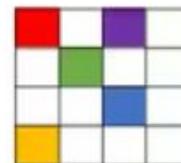
Un database **non relazionale**, o NoSQL (acronimo di Not Only SQL) non utilizza SQL (Structured Query Language) come linguaggio di interrogazione.

- 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.

SQL DATABASES



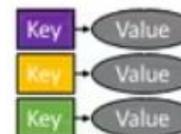
NoSQL DATABASES



Column



Graph



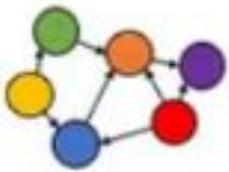
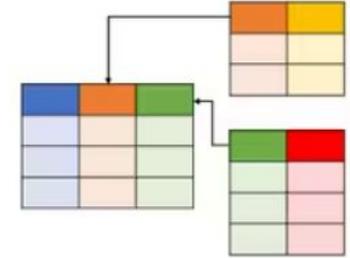
Key-Value



Document

LE BANCHE DATI

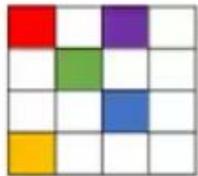
I database relazionali, come MySQL, PostgreSQL e SQL Server, tendono a scalare verticalmente. Ciò significa che per gestire carichi di lavoro più grandi o migliorare le prestazioni, è necessario potenziare l'hardware del server esistente su cui il database è in esecuzione. Questo potenziamento può includere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

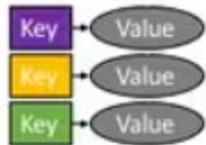
NoSQL DATABASES



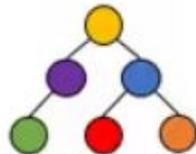
Column



Graph



Key-Value

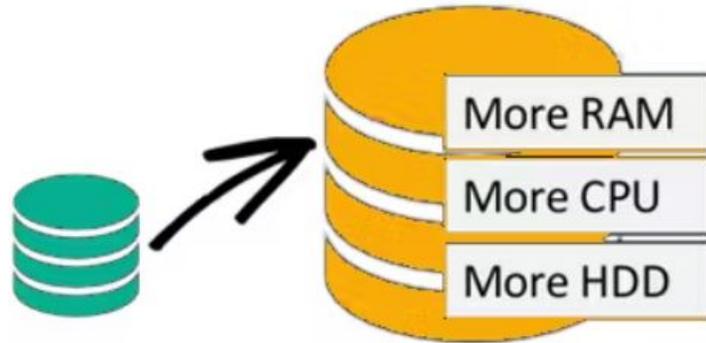


Document

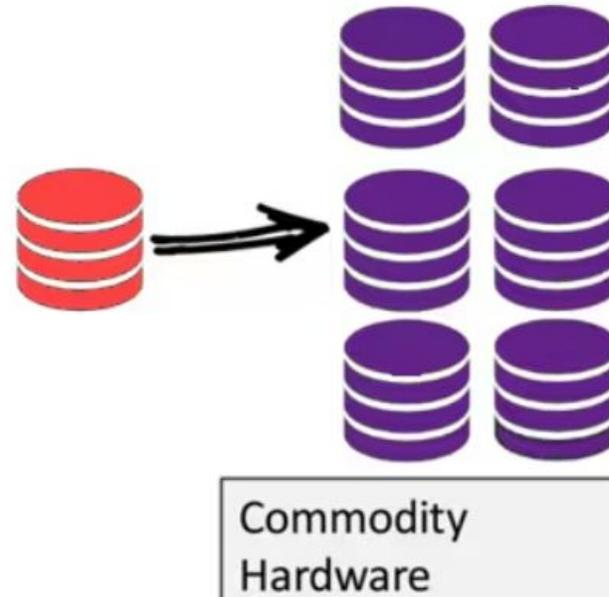
La scalabilità orizzontale è generalmente più flessibile e meno costosa della scalabilità verticale, poiché si possono aggiungere server standard o cloud senza dover investire in hardware costoso e specializzato. Inoltre, i database NoSQL sono spesso progettati con la tolleranza ai guasti e l'alta disponibilità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

LE BANCHE DATI

Scale-Up (*vertical*
scaling):



Scale-Out (*horizontal*
scaling):



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

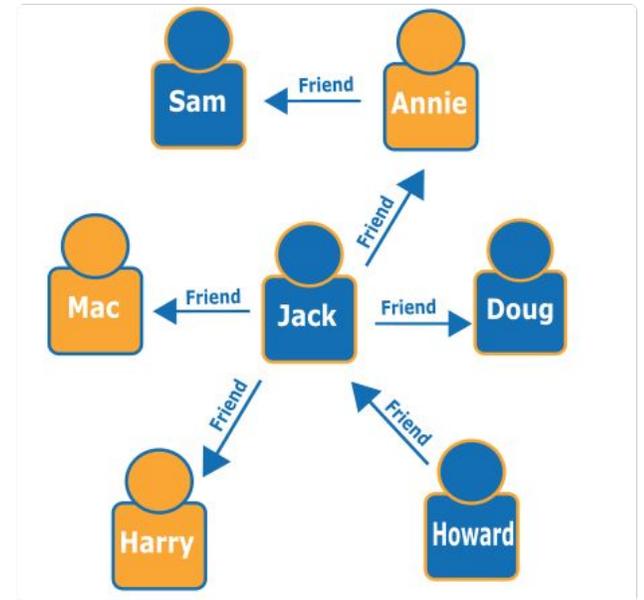
- I database non relazionali

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.



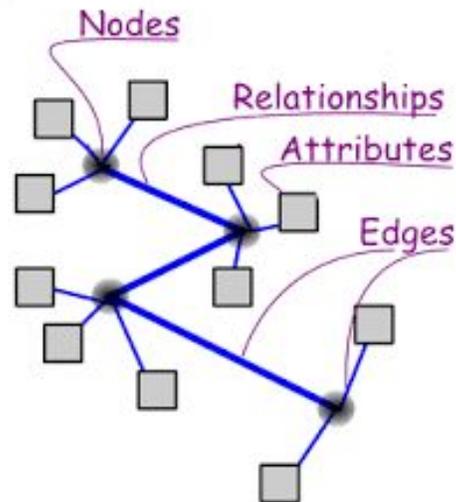
LE BANCHE DATI

- 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.

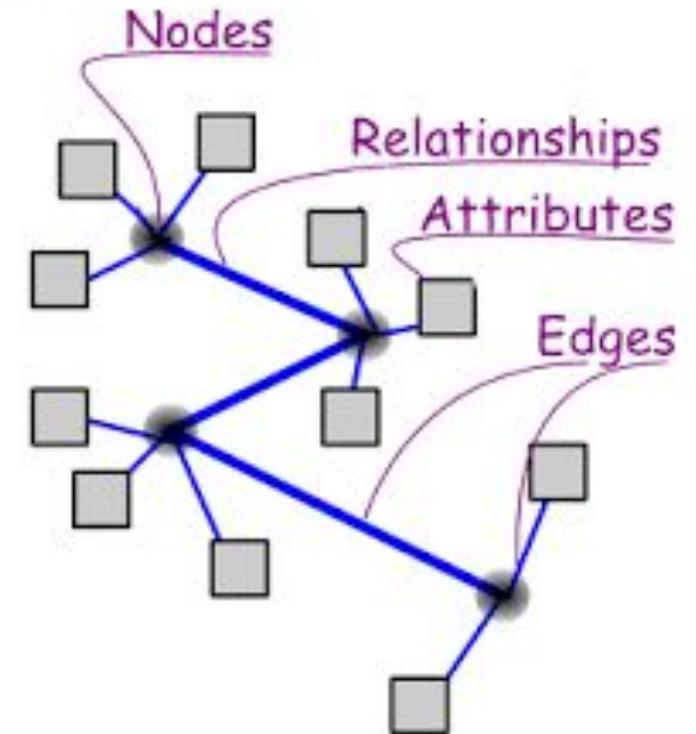


LE BANCHE DATI

- 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 del 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.

The screenshot shows a search interface for GenBank. At the top, there is a search bar with the text 'tp53' and a 'Search' button. Below the search bar, there are links for 'Create alert' and 'Advanced', and a 'Help' link on the right. A blue banner below the search bar contains the text: 'sequence GI numbers in September 2016. Please use accession.version! [Read more...](#)'. Below the banner, there are navigation options: 'Summary', '20 per page', 'Sort by Default order', and 'Send to'. There is also a 'Filters: [Manage Filters](#)' link. A box on the left contains the text: 'See [TP53 tumor protein p53](#) in the Gene database' and 'tp53 reference sequences [Genomic \(1\)](#) [Transcript \(15\)](#) [Protein \(15\)](#)'. Below this, it says 'Items: 1 to 20 of 6917'. There are navigation buttons: '<< First', '< Prev', 'Page 1 of 346', 'Next >', and 'Last >>'. A summary box shows 'Found 8124 nucleotide sequences. Nucleotide (6917) EST (1201) GSS (6)'. Below this, there are three search results, each with a checkbox and a link to the full record: 1. 'Homo sapiens isolate 653 TP53 (TP53) gene, exon 8 and partial cds' (232 bp linear DNA, Accession: KF572430.1, GI: 557786680); 2. 'Homo sapiens isolate 450 TP53 (TP53) gene, exon 8 and partial cds' (237 bp linear DNA, Accession: KF572429.1, GI: 557786678); 3. 'Homo sapiens isolate 530 TP53 (TP53) gene, exon 7 and partial cds' (270 bp linear DNA, Accession: KF572428.1, GI: 557786676). On the right side, there are sections for 'Results by taxon' (listing Homo sapiens, synthetic construct, Mus musculus, Rattus norvegicus, Bos taurus, and All other taxa) and 'Find related data' (with a 'Database: Select' dropdown and a 'Find items' button). At the bottom right, there is a 'Search details' section showing 'tp53[All Fields]'.

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.

Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾ Filters: [Manage Filters](#)

Items: 1 to 20 of 1793

<< First < Prev Page 1 of 90 Next > Last >>

Found 2943 nucleotide sequences. Nucleotide (1793) EST ([1148](#)) GSS ([2](#))

[Homo sapiens isolate 653 TP53 \(TP53\) gene, exon 8 and partial cds](#)

1. 232 bp linear DNA
Accession: KF572430.1 GI: 557786680
[GenBank](#) [FASTA](#) [Graphics](#)

[Homo sapiens isolate 450 TP53 \(TP53\) gene, exon 8 and partial cds](#)

2. 237 bp linear DNA
Accession: KF572429.1 GI: 557786678
[GenBank](#) [FASTA](#) [Graphics](#)

[Homo sapiens isolate 530 TP53 \(TP53\) gene, exon 7 and partial cds](#)

3. 270 bp linear DNA
Accession: KF572428.1 GI: 557786676
[GenBank](#) [FASTA](#) [Graphics](#)

[Homo sapiens isolate 366 TP53 \(TP53\) gene, exon 8 and partial cds](#)

4. 201 bp linear DNA
Accession: KF572426.1 GI: 557786672
[GenBank](#) [FASTA](#) [Graphics](#)

[Homo sapiens isolate 738 TP53 \(TP53\) gene, exon 7 and partial cds](#)

5. 204 bp linear DNA
Accession: KF572425.1 GI: 557786670

Find related data

Database:

Search details

```
tp53[All Fields] AND "Homo sapiens" [porgn]
```

Recent activity

[Turn Off](#) [Clear](#)

- (tp53) AND "Homo sapiens"[porgn] (1793)
Nucleotide

- tp53 (6917)
Nucleotide

- tp53 sapiens[All Fields] (0)
Nucleotide

- tp53 sapiens[organism] (0)
Nucleotide

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 questo caso c'è una ed una sola entry associata a quell'identificativo.

Nucleotide
[Advanced](#)

NCBI is phasing out sequence GI numbers in September 2016. Please use accession.version! [Read more...](#)

GenBank Send:

Homo sapiens tumor protein p53 (TP53), transcript variant 1, mRNA

NCBI Reference Sequence: NM_000546.5
[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS NM_000546 2591 bp mRNA linear PRI 23-APR-2016
DEFINITION Homo sapiens tumor protein p53 (TP53), transcript variant 1, mRNA.
ACCESSION NM_000546
VERSION NM_000546.5 GI:371502114
KEYWORDS RefSeq.
SOURCE Homo sapiens (human)
ORGANISM [Homo sapiens](#)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 2591)
AUTHORS Marcel V, Tran PL, Sagne C, Martel-Planche G, Vaslin L,
Teulade-Fichou MP, Hall J, Mergny JL, Hainaut P and Van Dyck E.
TITLE G-quadruplex structures in TP53 intron 3: role in alternative
splicing and in production of p53 mRNA isoforms
JOURNAL Carcinogenesis 32 (3), 271-278 (2011)
PUBMED [21112961](#)
REFERENCE 2 (bases 1 to 2591)
AUTHORS Marcel V, Perrier S, Aoubala M, Ageorges S, Groves MJ, Diot A,
Fernandes K, Tauro S and Bourdon JC.
TITLE Delta160p53 is a novel N-terminal p53 isoform encoded by
Delta133p53 transcript

LE BANCHE DATI

- API

The screenshot shows the NCBI Gene database search results for the gene **tp53**. The search bar at the top contains the text "Gene" and "tp53", with a "Search" button. Below the search bar, there is a red banner with a warning icon and text: "COVID-19 is an emerging, rapidly evolving situation. Get the latest public health information from CDC: <https://www.coronavirus.gov>. Get the latest research from NIH: <https://www.nih.gov/coronavirus>."

The main content area displays the gene details for **TP53 – tumor protein p53**. It includes the following information:

- GENE** (with a "Was this helpful?" feedback prompt)
- TP53 – tumor protein p53** (link)
- Homo sapiens (human)** (link)
- Also known as:** BCC7, BMFS5, LFS1, P53, TRP53
- GeneID:** 7157
- RefSeq transcripts (15)**, **RefSeq proteins (15)**, **RefSeqGene (1)**, **PubMed (9,727)**
- Buttons for **Orthologs**, **Genome Browser**, **BLAST**, and **Download**
- RefSeq Sequences** (with a "+" icon)

On the left side, there are navigation options: **Gene sources** (Genomic), **Categories** (Alternatively spliced, Annotated genes, Non-coding, Protein-coding, Pseudogene), **Sequence content** (CCDS, Ensembl, RefSeq, RefSeqGene), **Status** (Current), **Clear all**, and **Show additional filters**.

On the right side, there are additional filters and options: **Filters: Manage Filters**, **Results by taxon** (Top Organisms: [\[Tree\]](#), Homo sapiens (1451), Mus musculus (69), Rattus norvegicus (49), Jaculus jaculus (22), Heterocephalus glaber (20), All other taxa (3503), More...), **Find related data** (Database: Select, Find items), and **Search details** (tp53[All Fields] AND alive[prop]).

At the bottom, the **Search results** section shows **Items: 1 to 20 of 5114** and a pagination control: **<< First < Prev Page 1 of 256 Next > Last >>**. Below the pagination, there is a link: **See also 121 discontinued or replaced items.**

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

The screenshot shows the top portion of the HGNC website. At the top center is the HGNC logo in large yellow letters, with the text 'HUGO Gene Nomenclature Committee' underneath it. Below the logo is the tagline 'The resource for approved human gene nomenclature' in yellow. At the bottom of the header is a search bar with a yellow background. On the left of the search bar is a dropdown menu labeled 'Search all'. To the right of the dropdown is a search input field with the placeholder text 'Search symbols, keywords or IDs'. On the far right of the search bar are two icons: a question mark and a magnifying glass. Below the search bar, the text 'Last updated: 2020-03-20' is visible.

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

The screenshot shows the HGNC website interface. At the top, there is a search bar and a navigation menu. Below the navigation menu, a red banner indicates that HGNC resources will be at risk daily between 3am and 9am GMT for approximately 1 hour. The main heading is 'Symbol report for TP53' with a 'Stable symbol' tag. Below this, there are tabs for 'Report' and 'HCOP homology predictions'. The 'Report' tab is active, showing a table of HGNC data for TP53. The table includes fields such as 'Approved symbol', 'Approved name', 'Locus type', 'HGNC ID', 'Symbol status', 'Alias symbols', 'Alias names', and 'Chromosomal location'. The 'HGNC ID' field is highlighted with a red box, showing 'HGNC:11998'. An arrow points from this box to the 'Primary source' field in the summary table on the left.

LE BANCHE DATI

- API

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

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

Coordinate diverse a seconda della release del genoma umano usato.

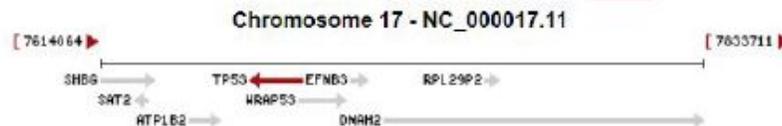
Genomic context

Location: 17p13.1

See TP53 in [Genome Data Viewer](#)

Exon count: 12

Annotation release	Status	Assembly	Chr	Location
109.20200228	current	GRCh38.p13 (GCF_000001405.39)	17	NC_000017.11 (7668402..7687550, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	17	NC_000017.10 (7571720..7590868, complement)



Locus-specific Databases

Genome Browsers

Genome Data Viewer

Variation Viewer (GRCh37.p13)

Variation Viewer (GRCh38)

1000 Genomes Browser (GRCh37.p13)

Ensembl

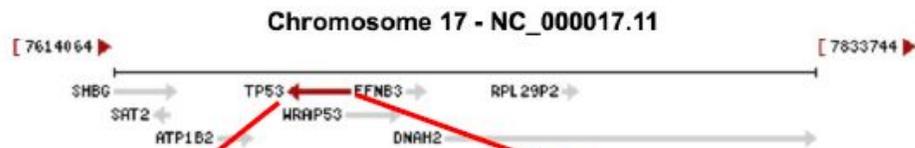
UCSC

Related information

- ❑ **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



TP53

Gene: TP53
Title: tumor protein p53
RNA title: mRNA-tumor protein p53, transcript variant 2
Protein title: cellular tumor antigen p53 isoform a
Merged features: NM_001126112.2 and NP_001119584.1
Location: complement(7,668,402..7,687,550)
[Length]
Span on NC_000017.11: 19,149 nt
Aligned length: 2,588 nt
CDS length: 1,182 nt
Protein length: 393 aa

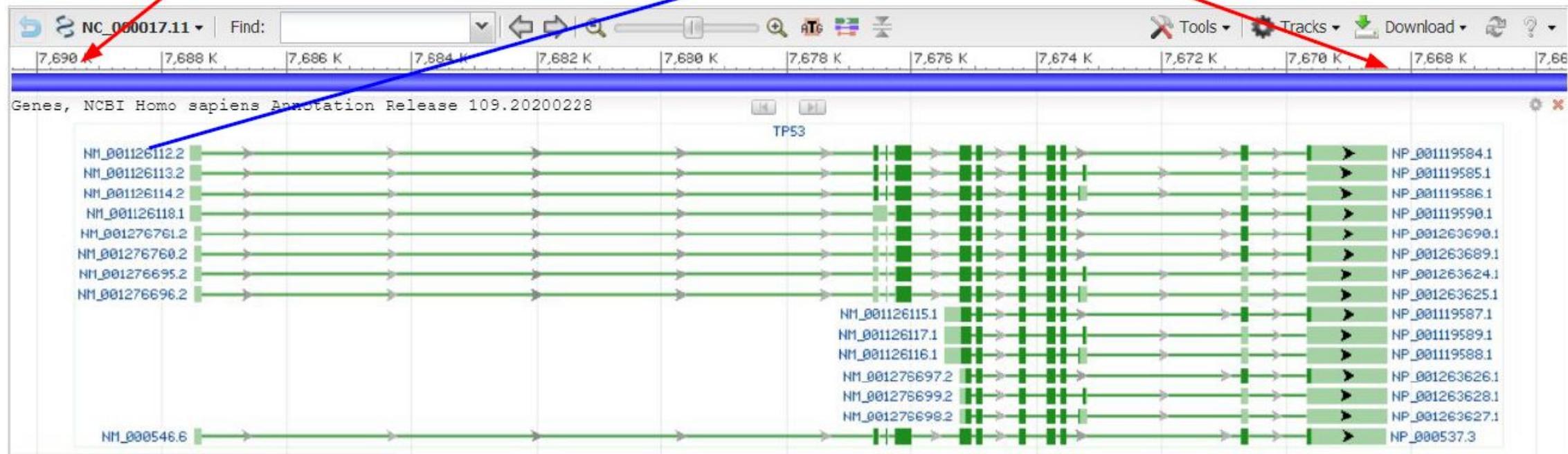
Download: [NP_001119584.1](#), [NM_001126112.2](#)

Links & Tools
View GeneID: [7157 \(TP53\)](#)
View HGNC: [11998](#)
View MIM: [191170](#)

BLAST Protein: [NP_001119584.1](#)
BLAST mRNA: [NM_001126112.2](#)
BLAST Genome-specific: [NC_000017.11 \(7,668,402..7,687,550\)](#)
BLAST Genomic: [NC_000017.11 \(7,668,402..7,687,550\)](#)
FASTA View: [NC_000017.11 \(7,668,402..7,687,550\)](#)
GenBank View: [NC_000017.11 \(7,668,402..7,687,550\)](#)

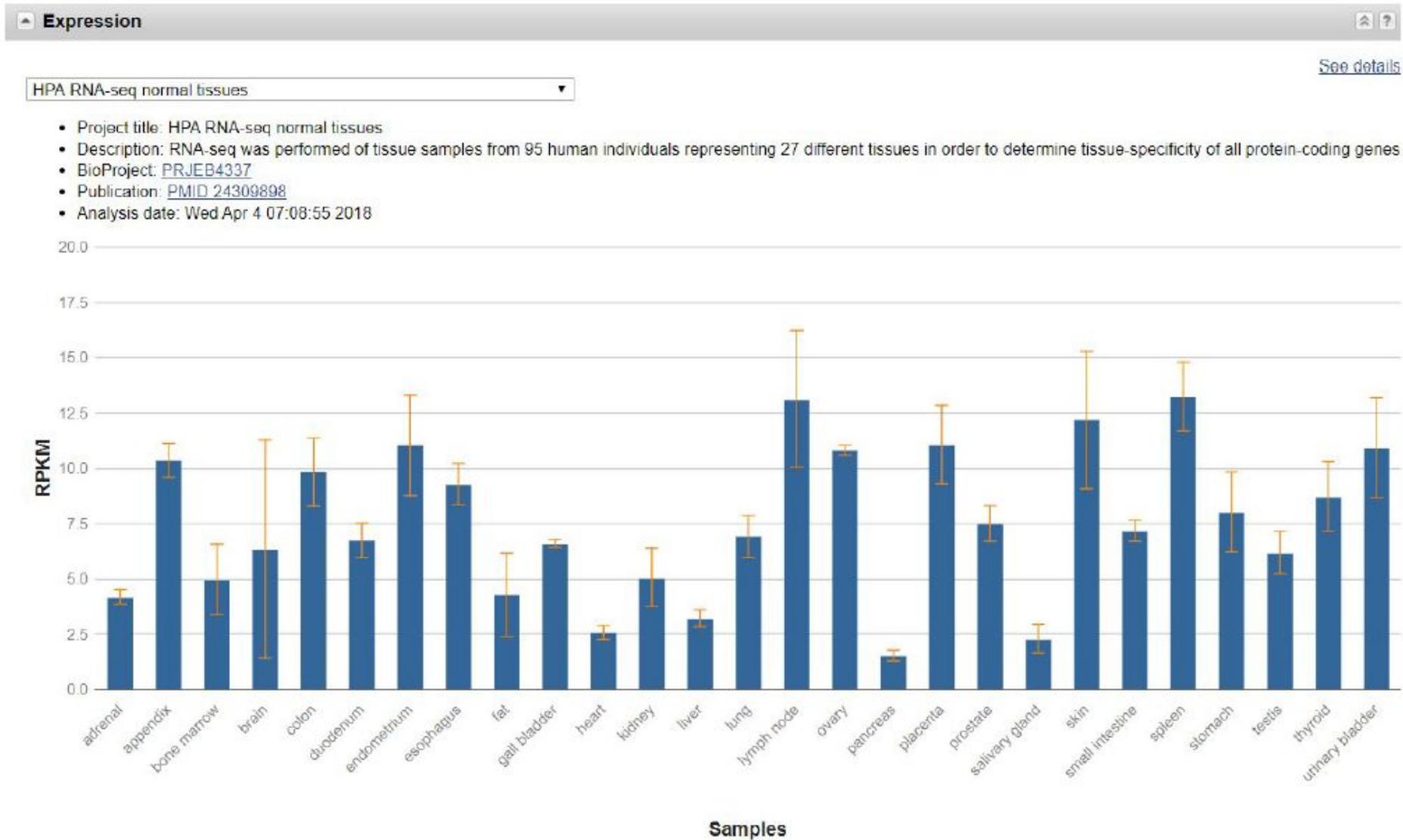
Genomic regions, transcripts, and products

Genomic Sequence:



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.

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

Also see

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 99 (January 2020)

- Update to GENCODE 33 for human
- Update to dbSNP153 for human
- Import of updated VISTA enhancers for human and mouse
- New genomes: 10 mammals (including 2 dog breeds), 11 birds, 15 fish and 4 reptiles
- Updated genome assemblies: zebra finch, fugu, Nile tilapia and Asian bonytongue

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Tools [All tools](#)

BioMart > Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT > Search our genomes for your DNA or protein sequence

Variant Effect Predictor > Analyse your own variants and predict the functional consequences of known and unknown variants

Search

All species for

e.g. [BRCA2](#) or [rat 6:62797383-63627669](#) or [rs699](#) or [coronary heart disease](#)

Favourite genomes

Human
GRCh38.p13
[Still using GRCh37?](#)

Mouse
GRCm38.p6

Ensembl species

... transcript variants and isoforms. Additional isoforms have also been shown to result from alternative splicing of transcript variants (PMIDs: 12032546, 20937277). [provided by RefSeq, Dec 2016] ... and 25 other tissues [See more](#)

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

The screenshot shows the PubMed.gov search results for the query 'tp53'. The search bar contains 'tp53' and the search button is labeled 'Search'. Below the search bar, there are options for 'Advanced' and 'Create alert', and a 'User Guide' link. The search results are sorted by 'Best match'. A red arrow points from the search bar to the search results, and another red arrow points from the search bar to the '19,412 results' count, which is circled in red. The results list includes:

- Updates from the **TP53** universe.
1 Pentimalli F.
Cell Death Differ. 2018 Jan;25(1):10-12. doi: 10.1038/cdd.2017.190. Epub 2017 Nov 10.
PMID: 29125599 **Free PMC article.** No abstract available.
Cite Share
- Decitabine in **TP53**-Mutated AML.
2 Montalban-Bravo G, Takahashi K, Garcia-Manero G.
N Engl J Med. 2017 Feb 23;376(8):796-7. doi: 10.1056/NEJMc1616062.
PMID: 28229579 No abstract available.
Cite Share
- Decitabine in **TP53**-Mutated AMI

On the left side, there is a 'RESULTS BY YEAR' section with a bar chart showing the number of publications per year from 1983 to 2019. The year 2019 is highlighted with a black box and labeled '2019: 2,209'. A blue arrow points from a text box to the 2019 bar in the chart.

Below the chart, there is a 'TEXT AVAILABILITY' section with checkboxes for 'Abstract', 'Free full text', and 'Full text'.

At the top left, the NIH logo and 'U.S. National Library of Medicine National Center for Biotechnology Information' are visible. A 'Log in' button is in the top right corner.

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

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

PubMed.gov

tp53 Search

Advanced Create alert User Guide

Save Email Send to Sorted by: Best match

MYNCBI FILTERS

RESULTS BY YEAR

10,604 results

Filters applied: Free full text. Clear all

Regulators of Oncogenic Mutant **TP53** Gain of Function.

1 Yamamoto S, Iwakuma T.
Cancers (Basel). 2018 Dec 20;11(1):4. doi: 10.3390/cancers11010004.
PMID: 30577483 **Free PMC article.** Review.

The tumor suppressor p53 (**TP53**) is the most frequently mutated human gene. Mutations in **TP53** not only disrupt its tumor suppressor function, but also endow oncogenic gain-of-function (GOF) activities in a manner independent of wild-type **TP53** (wtp53). Mutant **TP53** (mutp53) GOF is mainly mediated by its binding with other tumor suppressive or oncogenic proteins. Increasing evidence indicates that stabilization of mutp53 is crucial for its GOF activity. ...

Abstract
 Free full text
 Full text

TP53 Mutations in Breast and Ovarian Cancer.

“ Cite Share

applicando il filtro:
“solo articoli gratis”
il numero di pubblicazioni si e'
quasi dimezzato

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 leggerci la letteratura scientifica e di aggiornare manualmente il catalogo!



OMIM

OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. OMIM is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, under the direction of Dr. Ada Hamosh. Its official home is omim.org.

OMIM[®] Online Mendelian Inheritance in Man[®]
An Online Catalog of Human Genes and Genetic Disorders
Updated 18 May 2016

Advanced Search : [OMIM](#), [Clinical Synopses](#), [Gene Map](#)

Need help? : [Example Searches](#), [OMIM Search Help](#), [OMIM Tutorial](#)

Mirror sites : us-east.omim.org, europe.omim.org

<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.

Pfam 29.0 (December 2015, 16295 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM ENTRY	View Pfam annotation and alignments
VIEW A CLAN	See groups of related entries
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	<input type="text" value="enter any accession or ID"/> <input type="button" value="Go"/> <input type="button" value="Example"/>
	Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.
	Or view the help pages for more information

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](#) [↗](#)

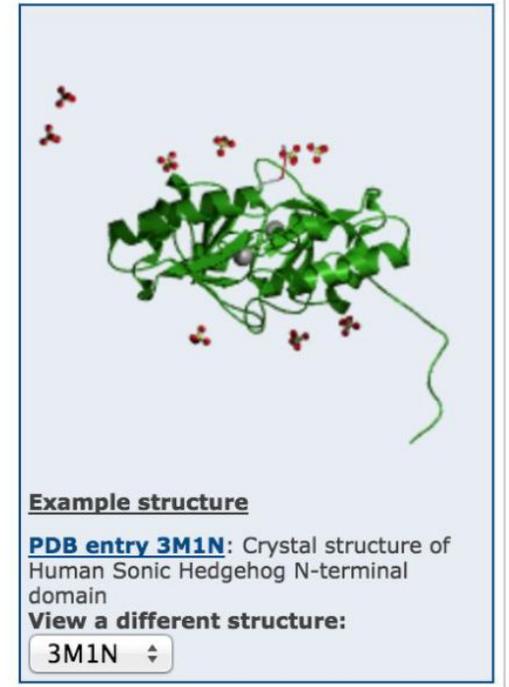
SCOP: [1vhh](#) [↗](#)

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.

[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [BLOG](#) | [HELP](#)

Rfam 12.1 (April 2016, 2474 families)

The Rfam database is a collection of RNA families, each represented by **multiple sequence alignments**, **consensus secondary structures** and **covariance models (CMs)**. [More...](#)

QUICK LINKS

[SEQUENCE SEARCH](#)

[VIEW AN RFAM FAMILY](#)

[VIEW AN RFAM CLAN](#)

[KEYWORD SEARCH](#)

[TAXONOMY SEARCH](#)

[JUMP TO](#)

YOU CAN FIND DATA IN RFAM IN VARIOUS WAYS...

Analyze your RNA sequence for Rfam matches

View Rfam family annotation and alignments

View Rfam clan details

Query Rfam by keywords

Fetch families or sequences by NCBI taxonomy

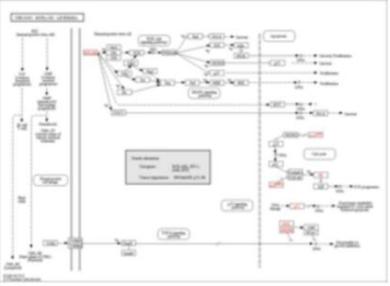
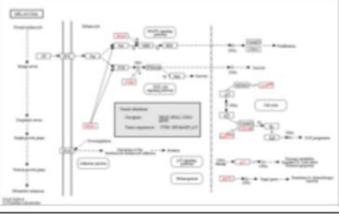
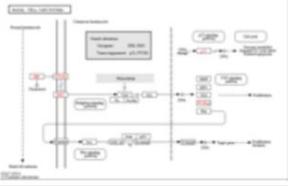
Enter any type of accession or ID to jump to the page for a Rfam family, sequence or genome

Or view the [help](#) pages for more information

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...

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 genes3 (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.

```
Header ● >VIT_201s0011g03530.1
Sequence ● AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
          ● GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
Header ● >VIT_201s0011g03540.1
Sequence ● CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
          ● AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
Header ● >VIT_201s0011g03550.1
Sequence ● CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
          ● GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```


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**).

Example

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Mandatory header lines (indicated by a red arrow pointing to the first line)

Optional header lines (meta-data about the annotations in the VCF body) (indicated by a grey arrow pointing to the remaining header lines)

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Reference alleles (GT=0) (indicated by a blue arrow pointing to the first '0' in the first row)

Alternate alleles (GT>0 is an index to the ALT column) (indicated by a blue arrow pointing to the '1' in the first row)

Deletion (indicated by a blue arrow pointing to the in the last row)

SNP (indicated by a blue arrow pointing to the A,AT in the first row)

Large SV (indicated by a blue arrow pointing to the T,CT in the second row)

Insertion (indicated by a blue arrow pointing to the G in the third row)

Other event (indicated by a blue arrow pointing to the in the last row)

Phased data (G and C above are on the same chromosome) (indicated by a blue arrow pointing to the | in the third row)

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.

Example

```

##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 . T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
  
```

VCF header

- Mandatory header lines** (indicated by a red arrow pointing to `##fileformat=VCFv4.0`)
- Optional header lines** (meta-data about the annotations in the VCF body) (indicated by a black arrow pointing to `##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">`)

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Reference alleles (GT=0) (indicated by a blue arrow pointing to the first column of the body)

Alternate alleles (GT>0 is an index to the ALT column) (indicated by a blue arrow pointing to the second column of the body)

Deletion (indicated by a blue arrow pointing to the ALT column of the last row)

SNP (indicated by a blue arrow pointing to the ALT column of the second row)

Large SV (indicated by a blue arrow pointing to the ALT column of the last row)

Insertion (indicated by a blue arrow pointing to the ALT column of the third row)

Other event (indicated by a blue arrow pointing to the ALT column of the first row)

Phased data (G and C above are on the same chromosome) (indicated by a blue arrow pointing to the vertical bar in the GQ field of the second row)

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.

Fixed fields									Optional: FORMAT field specifying data type + Per-sample genotype data		
#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.

		Fixed fields							Optional: FORMAT field specifying data type + Per-sample genotype data			
		#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NORMAL	TUMOR
BODY	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">
```

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

```
##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">
```

INFO meta-information

FILTER meta-information

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

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;DB
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T
20	1234567	microsat1	GTC	G,GTCTC	50	PASS	NS=3;DP=9;AA=G

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

FORMAT	NORMAL	TUMOR
GT:GQ:DP:HQ	0 0:48:1:51,51	1 0:48:8:51,51
GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3
GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2
GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:51,51
GT:GQ:DP	0/1:35:4	0/2:17:2

BODY

```
##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>
```

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.

The mmCIF format

mmCIF: macromolecular
Crystallographic
Information
File

This is an extension of the Crystallographic Information File (CIF) data representation (used for describing small molecule structures) to describe macromolecules.

```
_citation.id primary
_citation.title
;Comparison of the three-dimensional structures
of recombinant human H and horse L ferritins at
high resolution.
;
_citation.journal_abbrev J.Mol.Biol.
_citation.journal_volume 268
_citation.page_first 424
_citation.page_last 448
_citation.year 1997
_citation.journal_id_ASTM JMOBAK
_citation.country UK
_citation.journal_id_ISSN 0022-2836
_citation.journal_id_CSD 0070
_citation.book_publisher ?
_citation.pdbx_database_id_PubMed 9159481
```

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.

