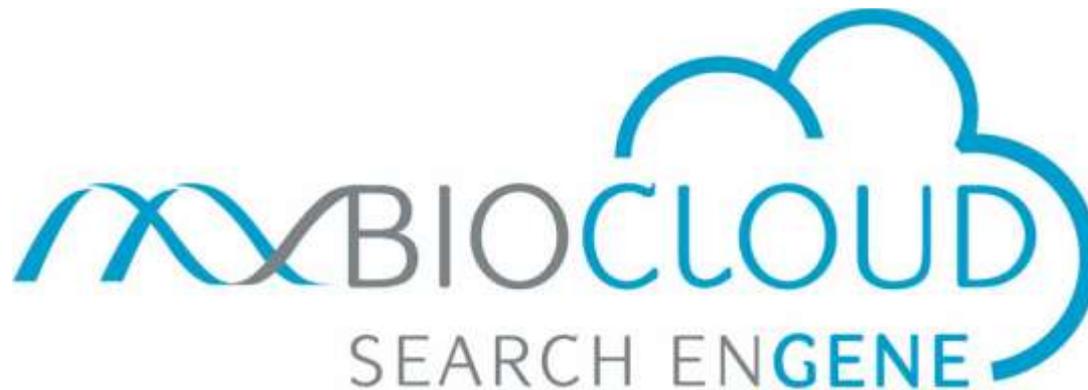




Giornate sugli sbocchi professionali
Del corso di Laurea in
Biotecnologie Industriali
(BIOTIN)

Oristano 23/24 Aprile 2013



URL <http://biocloud.unica.it>
email biocloud@unica.it

Emanuele Pascariello
emanuele.pascariello@gmail.com

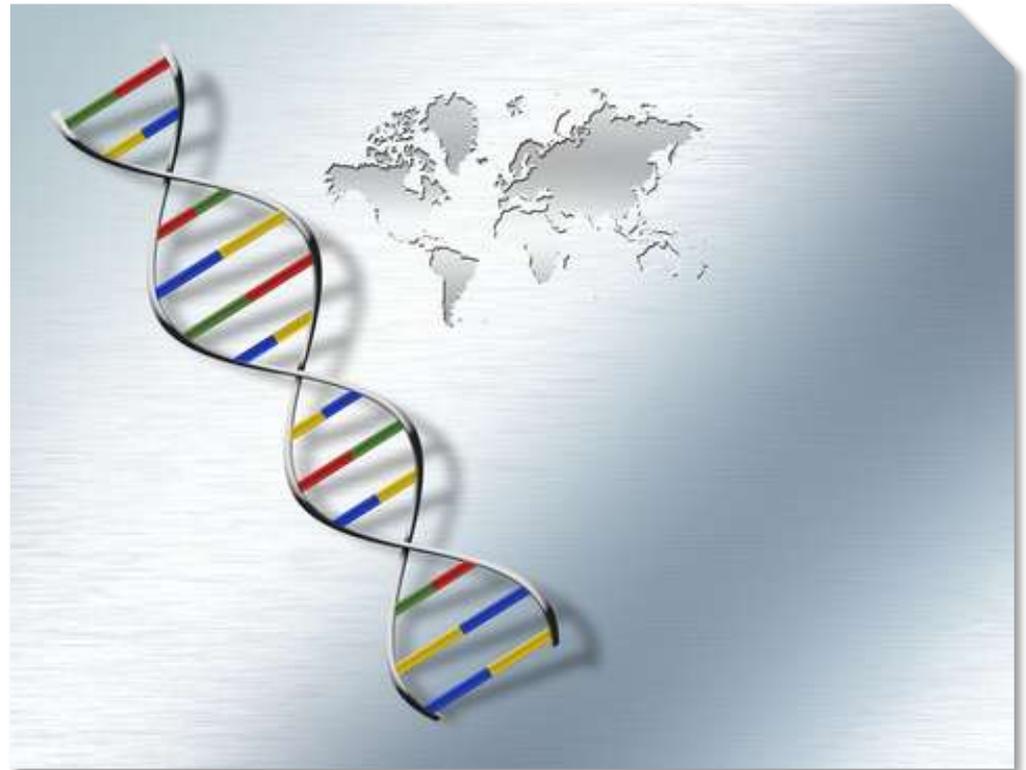
Scenario attuale

Negli ultimi 15 anni si è assistito ad una rivoluzione in campo Biomedico, che ha determinato la generazione di una enorme mole di dati. **Introduzione di tecnologie ad alto flusso**

Nuovi metodi di indagine

- ✓ Gene Expression profiling by array
- ✓ Expression profiling by high throughput sequencing
- ✓ SNP genotyping by SNP array
- ✓ Next Generation Sequencing

Generazione di una enorme mole di dati in tempi sempre più brevi ed a costi sempre inferiori.



Cosa è cambiato?

Scenario precedente:

La quantità di dati generati dalle tecnologie disponibili era inferiore alle nostre capacità di gestirli



High – Throughput Technologies



ATTCGCGATT TACGTAATCGAA
TAAGCGCTAA ATGCATTAGCTT

MERMLPLLALGLLAAGFCPAVLCHPN SPLDE
ENLTQENQDRGTHVDLGLASANVDFAFS

Annotazioni da processi manuali
Ed automatizzati



Scenario attuale:

La quantità di dati generati dalle tecnologie ad elevato flusso superano o mettono costantemente alla prova la nostra capacità di gestirli



Crescita del numero di Banche dati di tipo:

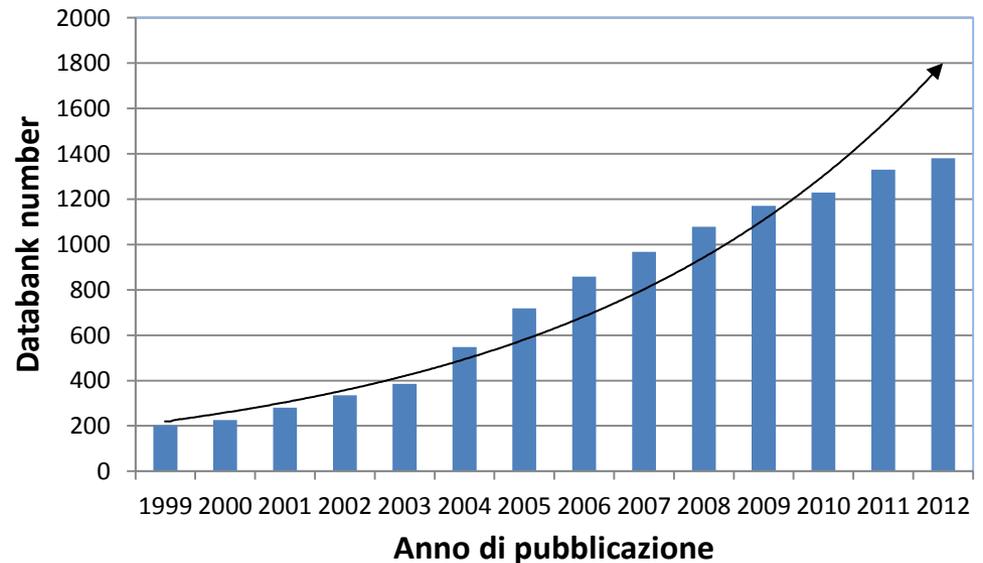
- ✓ **Primario** (DNA, RNA, Proteine)
- ✓ **Derivato** o specializzato (EST, SNP, Genomi, Microarray data, pathways, genetic disorders)

- ✓ I dati generati presentano elevata dimensionalità – struttura
- ✓ Sono quantitativamente molto consistenti
- ✓ Obsolescenza del dato
- ✓ Presentano fenomeni di ridondanza esterna e interna.

Problema: Molte banche dati diverse, molti Accession e riferimenti diversi.
Necessità di consultare differenti servizi/ banche dati per ottenere informazioni specifiche.



Databank number growth by year



Scopo del progetto

- ✓ Creazione di un servizio di integrazione di una parte consistente di queste informazioni provenienti da banche dati primarie e derivate
- ✓ Riorganizzazione con un approccio “modulare” del dato di origine
- ✓ Accesso mirato alle banche dati esterne
- ✓ Web usability
- ✓ Approccio *Pay as you go*



Intento

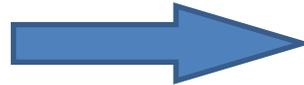
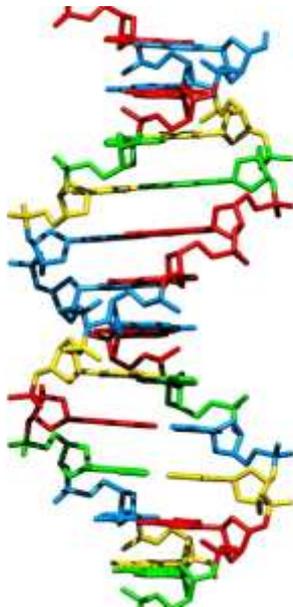
- ✓ Fornire un servizio per la ricerca di informazioni geniche, farmacologiche e fenotipiche nell'uomo, più facile e intuitivo
- ✓ Raccolta di grossa parte delle informazioni esterne in unico punto.
- ✓ Un **Hub** verso le maggiori fonti e Web-services a livello mondiale per ulteriori approfondimenti



Su cosa si basa Biocloud search enGene

Biocloud search enGene si basa sui dati contenuti nei principali database pubblici di Biomedicina.

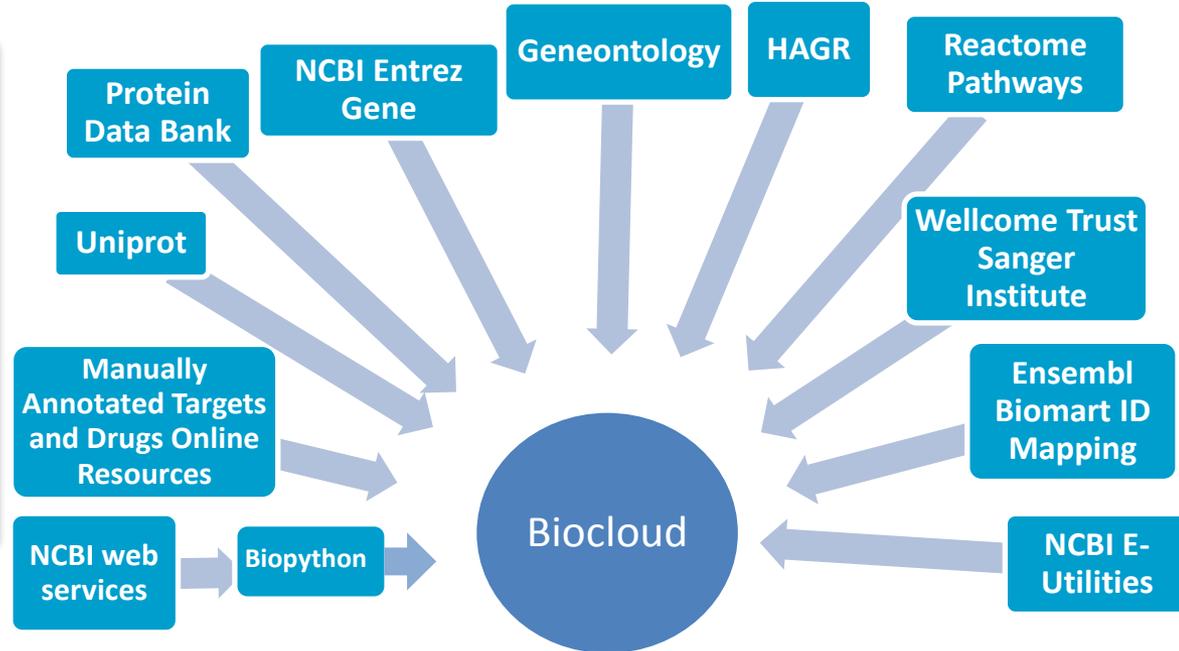
Nella versione attuale comprende annotazioni relative **all'*Homo Sapiens***



Come funziona

✓ **Redirect coerente con la query** verso le principali banche dati, primarie e derivate per ulteriori informazioni specifiche

✓ Eliminazione della necessità di orientarsi all'interno di tali servizi.

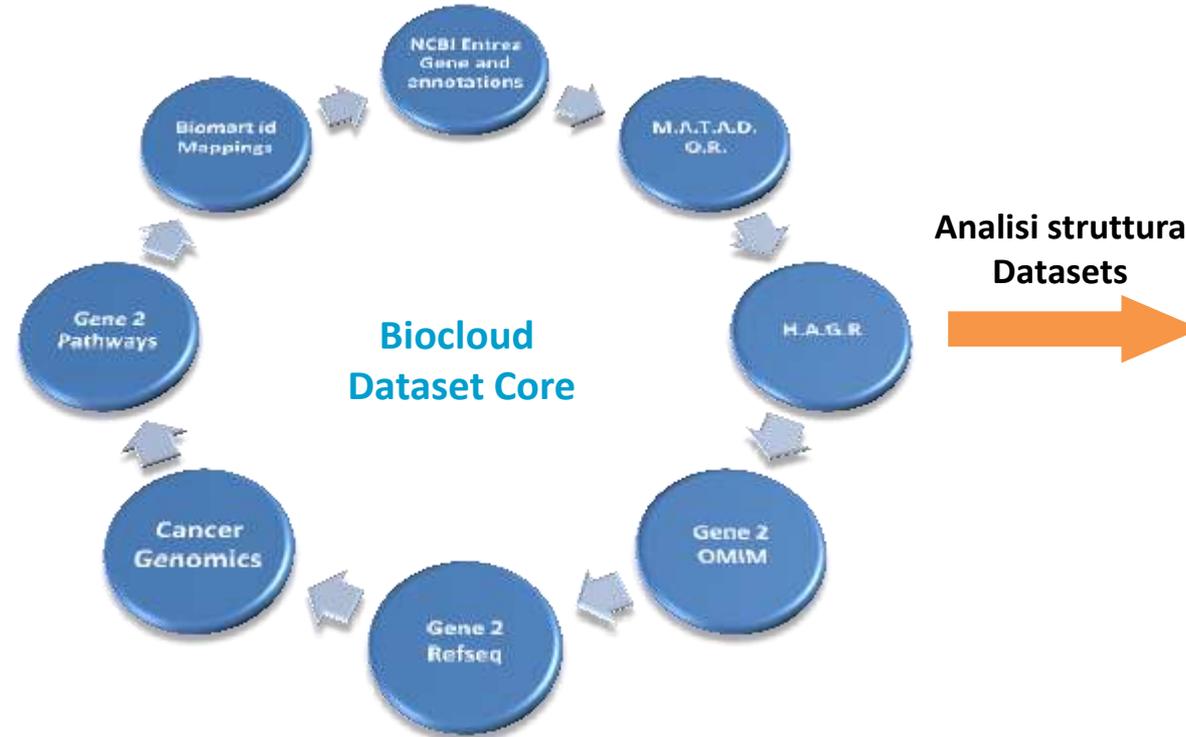


Local data



Come è fatto – Local data

L'applicazione è fortemente **Gene – Centric** e colleziona le annotazioni sui geni umani provenienti dalle banche dati tra le principali e maggiormente accreditate

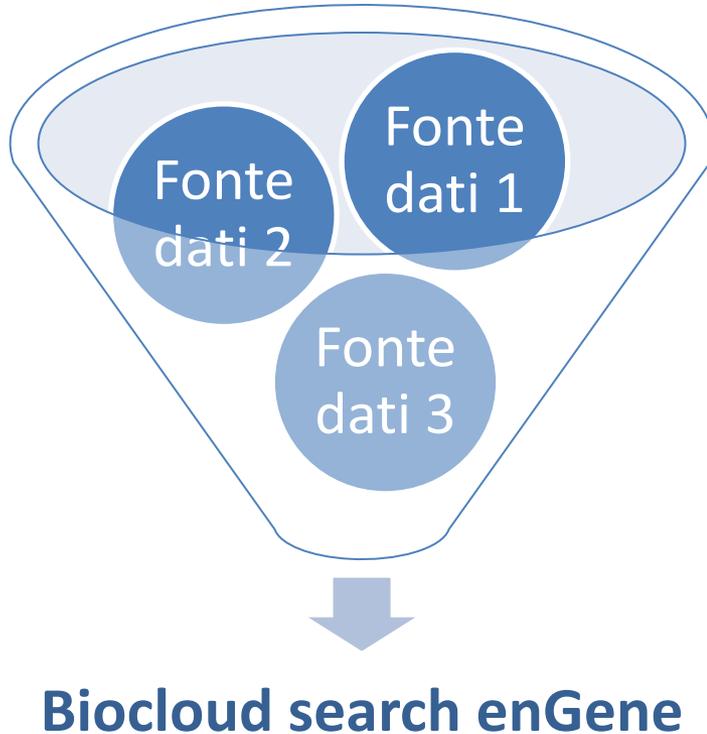


Relazioni tra i dataset

- ✓ Aging
- ✓ Fenotipo patologico
- ✓ Sequenze di riferimento nucleotidiche e aminoacidiche
- ✓ Strutture proteiche 3D
- ✓ Networks di interazioni
- ✓ Interazione farmaco – prodotto di espressione
- ✓ Definizione termini ontologici
- ✓ Pathways
- ✓ Aspetti genomici della sensibilità a chemioterapici oncologici
- ✓ Tipologia di relazione tra geni
- ✓ Relazioni tra geni omologhi

L'utilizzo di tecnologie cloud e di database non relazionali, permette di scalare con minori problemi di natura sistemistica l'enorme mole di dati in costante crescita.





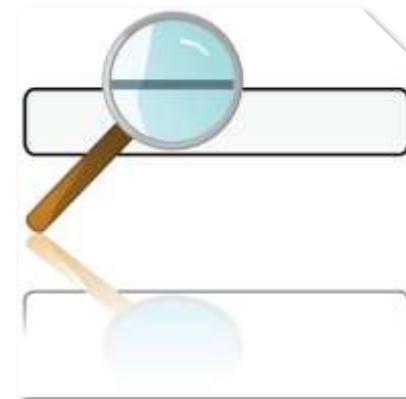
Viene fatta la cattura “al volo” di una notevole quantità di dati mentre altri sono ritenuti all’interno del datastore dell’applicazione ed usati come indice.

In questo modo è possibile by-passare il problema della enorme mole di dati con una efficienza maggiore.

Queries sui geni

Nella sua versione attuale Biocloud permette di eseguire delle queries sulla base dei seguenti criteri:

- ✓ Simbolo ufficiale del gene **HGNC** o identificativo univoco **Entrez gene ID**
- ✓ Identificativo dei prodotti di espressione **Uniprot ID**
- ✓ *bulk queries* Ricerca in base a criteri quali: **natura del gene - Cromosoma di appartenenza - Annotazioni relative ai processi Aging related - Annotazioni farmacogenomiche relative a sensibilità a chemioterapici oncologici**



Queries sui farmaci

- ✓ Queries su molecole farmacologicamente attive tramite **nome ufficiale** della molecola o tramite **Pubchem ID**

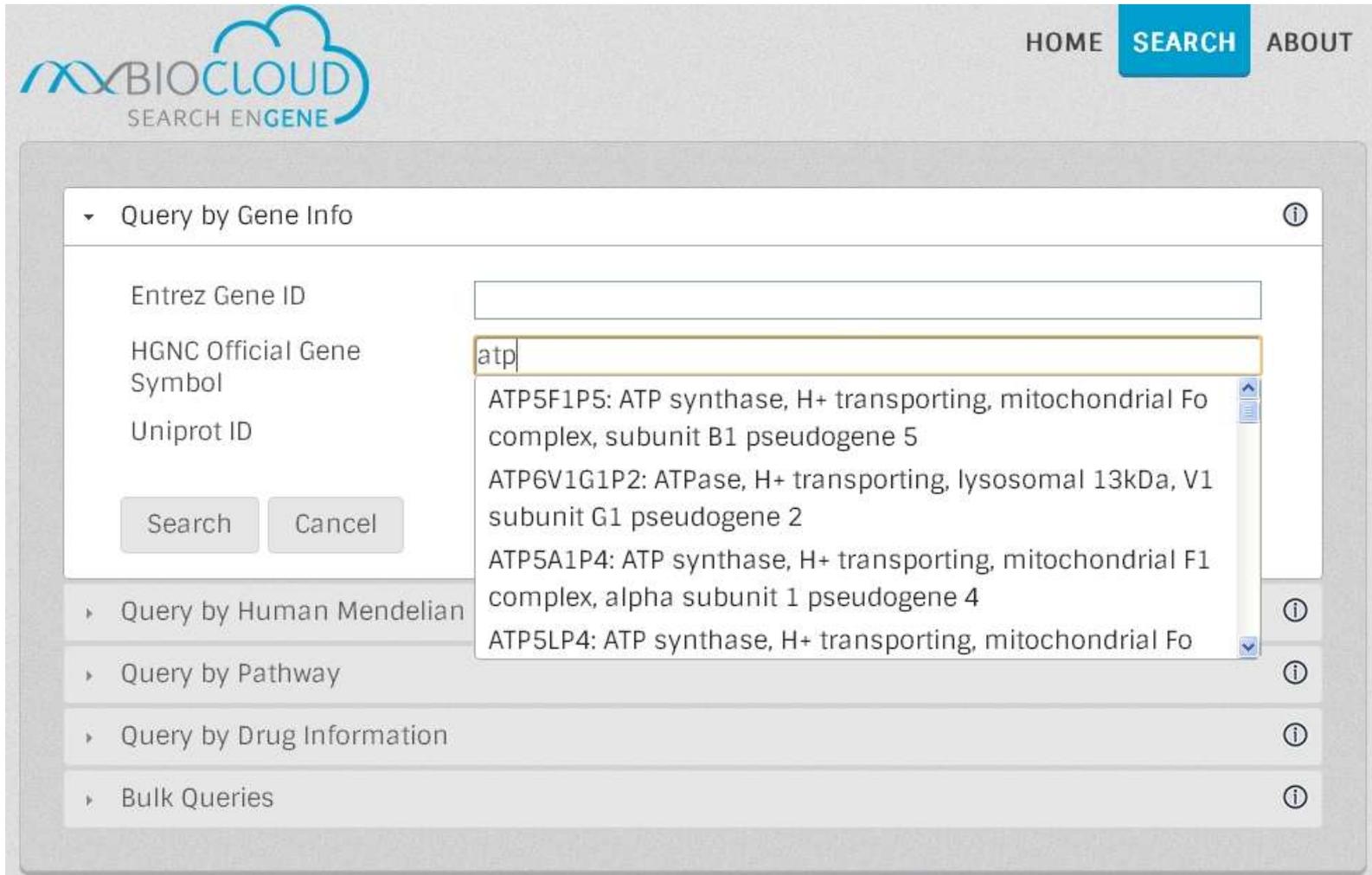
Queries sui fenotipi

- ✓ Queries basate su **fenotipi** associati a disordini genetici con trasmissione di tipo Mendeliano

Queries sui Pathways

- ✓ Queries basate su nomi descrittivi di pathways metabolici in cui sono coinvolti i geni

Query tramite l'utilizzo del
simbolo
ufficiale del gene

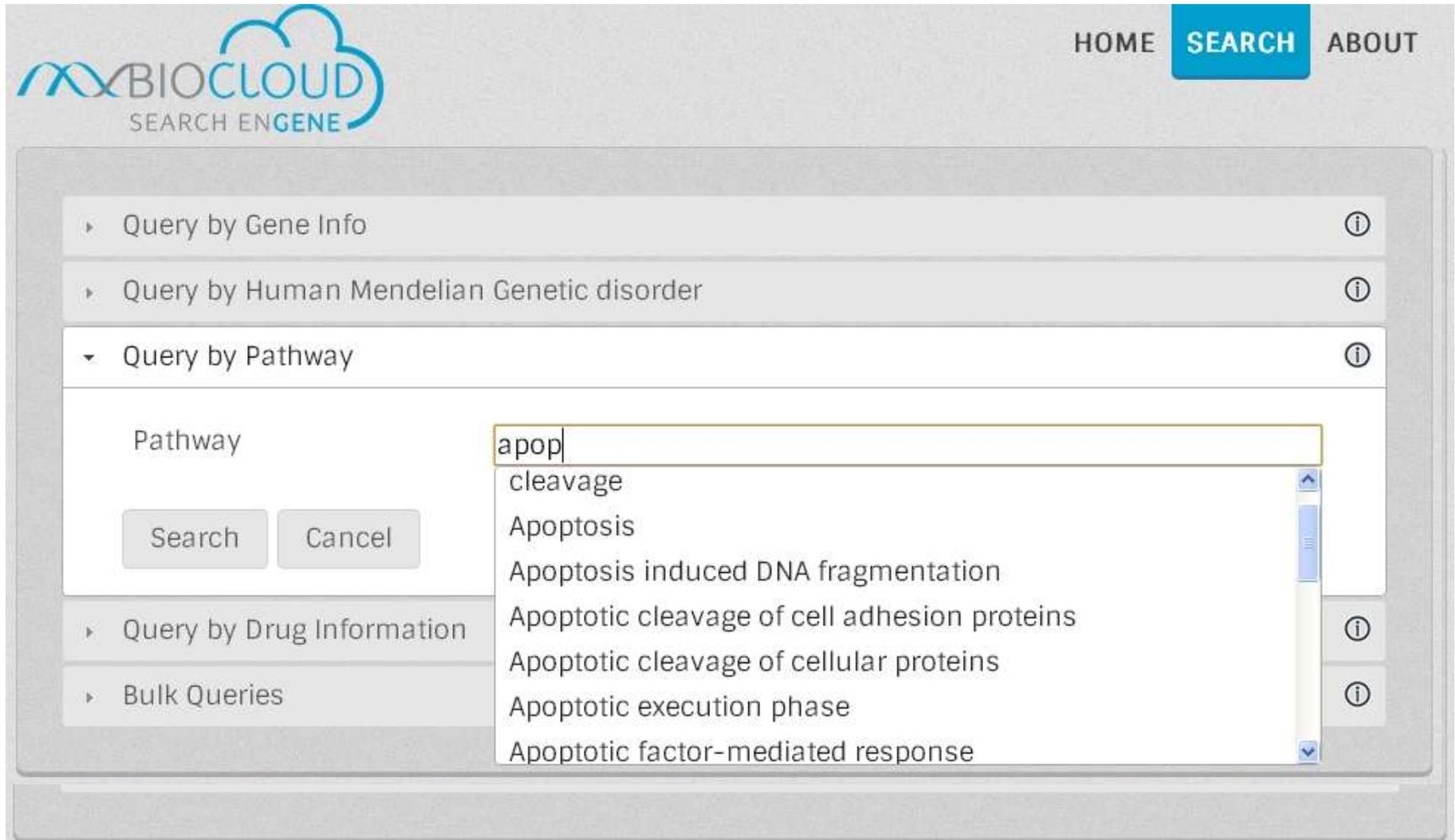


The screenshot shows the BIOCLOUD SEARCH ENGINE interface. At the top, there is a navigation bar with 'HOME', 'SEARCH', and 'ABOUT' links. The 'SEARCH' link is highlighted in blue. Below the navigation bar, the BIOCLOUD SEARCH ENGINE logo is displayed on the left. The main content area is titled 'Query by Gene Info' and contains several search options: 'Entrez Gene ID', 'HGNC Official Gene Symbol', and 'Uniprot ID'. The 'HGNC Official Gene Symbol' field is active, with the text 'atp' entered. A dropdown menu is open below this field, displaying a list of search results for 'atp':

- ATP5F1P5: ATP synthase, H+ transporting, mitochondrial Fo complex, subunit B1 pseudogene 5
- ATP6V1G1P2: ATPase, H+ transporting, lysosomal 13kDa, V1 subunit G1 pseudogene 2
- ATP5A1P4: ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit 1 pseudogene 4
- ATP5LP4: ATP synthase, H+ transporting, mitochondrial Fo

Below the search options, there are 'Search' and 'Cancel' buttons. At the bottom of the interface, there are additional search options: 'Query by Human Mendelian', 'Query by Pathway', 'Query by Drug Information', and 'Bulk Queries', each with an information icon (i) to its right.

Query tramite l'utilizzo del
Pathway metabolico
Con attività farmacologica



The screenshot displays the BIOCLOUD SEARCH ENGINE interface. At the top left is the logo, and at the top right are navigation links for HOME, SEARCH, and ABOUT. The main content area features a list of search options: 'Query by Gene Info', 'Query by Human Mendelian Genetic disorder', 'Query by Pathway' (which is expanded), 'Query by Drug Information', and 'Bulk Queries'. Each option has an information icon. The 'Query by Pathway' section contains a 'Pathway' input field with the text 'apop' and a dropdown menu listing the following pathways: 'cleavage', 'Apoptosis', 'Apoptosis induced DNA fragmentation', 'Apoptotic cleavage of cell adhesion proteins', 'Apoptotic cleavage of cellular proteins', 'Apoptotic execution phase', and 'Apoptotic factor-mediated response'. Below the input field are 'Search' and 'Cancel' buttons.

Query tramite la selezione di criteri quali:

- ✓ Gene biotype
- ✓ Chromosome belonging
- ✓ Aging Annotation
- ✓ Cancer Drug sensitivity

- ▶ Query by Gene Info ⓘ
- ▶ Query by Human Mendelian Genetic disorder ⓘ
- ▶ Query by Pathway ⓘ
- ▶ Query by Drug Information ⓘ
- ▼ Bulk Queries ⓘ

Gene type
Chromosome
Human ageing annotated genes
Cancer Drug sensitivity Annotaterd genes

- protein-coding
- miscRNA
- ncRNA
- rRNA
- snoRNA
- snRNA
- tRNA

Search

Cancel

Risultato query su Pathway metabolico



HOME SEARCH ABOUT

Search returned 11 items

Query Results

List of genes annotated participating in pathway [Apoptotic cleavage of cell adhesion proteins](#)

Show 10 entries Search:

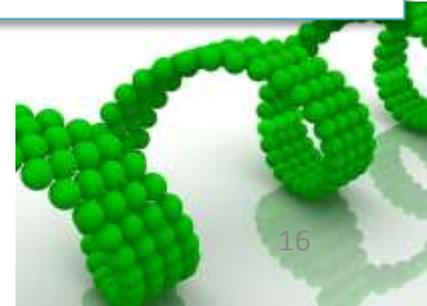
Entrez ID	HGNC Symbol	Gene Name	Type
100506658	OCLN	occludin	protein-coding
1499	CTNNB1	catenin (cadherin-associated protein), beta 1, 88kDa	protein-coding
1828	DSG1	desmoglein 1	protein-coding
1829	DSG2	desmoglein 2	protein-coding
1830	DSG3	desmoglein 3	protein-coding
1832	DSP	desmoplakin	protein-coding
5317	PKP1	plakophilin 1 (ectodermal dysplasia/skin fragility syndrome)	protein-coding
7082	TJP1	tight junction protein 1	protein-coding
836	CASP3	caspase 3, apoptosis-related cysteine peptidase	protein-coding
9414	TJP2	tight junction protein 2	protein-coding

Showing 1 to 10 of 11 entries

Queries eseguite sui geni

- ✓ Summary
- ✓ Full gene name
- ✓ Aliases gene name
- ✓ Taxonomy ID
- ✓ Posizione di **Start** e **Stop** del gene
- ✓ Numero degli **Esoni** del gene
- ✓ **Gene type**
- ✓ **HGNC** symbol
- ✓ Informazioni legate ad "**Aging**"
- ✓ **Mappa** completa del **Cromosoma** di appartenenza e riferimento cytoband del gene
- ✓ **Mappa** della posizione del **gene** sul cromosoma
- ✓ Informazioni **epigenomiche**
- ✓ Strutture proteiche **3D** relative al gene oggetto della query
- ✓ **Network di interazioni** note e presunte del prodotto di espressione del gene con altri prodotti di espressione
- ✓ Drug sensitivity – mutations in **Cancer**
- ✓ **Sequenza Aminoacidica** del prodotto di espressione in formato FASTA

- ✓ Tipologia delle relazioni tra il gene della query e altri geni presenti nel dataset delle annotazioni
- ✓ Elenco dei **termini ontologici** che annotano il gene oggetto della query
- ✓ Grafo di ciascun termine ontologico
- ✓ **Accession** Nucleotidiche e Aminoacidiche di riferimento
- ✓ Relazione con molecole farmacologicamente attive
- ✓ **Pathways** metabolici.
- ✓ Elenco dei **fenotipi** patologici annotati in O.M.I.M
- ✓ **Datasets e profili di espressione** correlati, da Gene Expression Omnibus relativi ad esperimenti di Microarray gene expression
- ✓ **SNP**
- ✓ **Homologene**
- ✓ Bibliografia Pubmed relativa al gene della query
- ✓ Link **diretti** a web services esterni



La struttura dei menu



HOME SEARCH ABOUT

TP53 - Entrez Gene ID 7157

- › General Information about TP53
- › Cross-database ID
- › Interaction network and Structures
- › Microarrays Gene Expression for TP53
- › Pathways for TP53
- › Ontology Terms for TP53
- › Uniprot entries for TP53 - FASTA Format
- › Gene to Gene relations
- › HomoloGene
- › Reference Sequences for TP53
- › SNP for TP53
- › Gene phenotype for TP53
- › Drug interactions for TP53
- › External References for TP53

Menu: general info



General Information about TP53

Summary | Genomic information

HGNC Official Gene Symbol	TP53
Gene Name	tumor protein p53
Also Known As	BCC7, LFS1, P53, TRP53
Other designations	antigen NY-CO-13, cellular tumor antigen p53, p53 tumor suppressor, phosphoprotein p53, transformation-related protein 53
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 (PMIDs: 12032546, 20937277). [provided by RefSeq, Feb 2013]
Gene Type	protein-coding ?
Organism	Homo sapiens
Taxonomy ID	9606
Aging possibly related	Yes
Drug Sensitivity - mutations in cancer	Yes

ABOUT

7157

Menu: Structures and interactions



HOME SEARCH ABOUT

TP53 - Entrez Gene ID 7157

› General Information about TP53

› Cross-database ID

▼ Interaction network and Structures

STRING interaction network

PDB Structures

3D Bio Assembly from Protein Data Bank

A repository for 3-D biological macromolecular structure data. PDB archives protein structures deduced from crystallography and nuclear magnetic resonance (NMR) experiments on protein structures.

Click the button to load all PDB IDs related to [TP53](#)

Load PDB IDs



"Pay as you go"

› Microarrays Gene Expression for TP53

› Pathways for TP53

TP53 - Entrez Gene ID 7157

- ▶ General Information about TP53
- ▶ Cross-database ID
- ▶ Interaction network and Structures
- ▶ Microarrays Gene Expression for TP53
- ▶ Pathways for TP53
- ▼ Ontology Terms for TP53

An organized hierarchy of terms produced by the Gene Ontology Consortium, used to describe biological processes, cellular component, and molecular function.

Click the button to show all the terms annotated to **TP53**

Load Terms



“Pay as you go”

Guide to GO Evidence Codes 

- ▶ Uniprot entries for TP53 - FASTA Format

Menu: Microarray datasets



HOME SEARCH ABOUT

TP53 - Entrez Gene ID 7157

- General Information about TP53
- Cross-database ID
- Interaction network and Structures
- ▾ Microarrays Gene Expression for TP53

Gene Expression Datasets and series from Omnibus (GEO)

Click the button to load all GEO Datasets and GEO Series related to TP53

Load Datasets



"Pay as you go"

Gene Expression Profiles across all datasets

Find specific expression profile

Add filter

Clear

🔗 See all GEO Expression profiles for TP53

- Pathways for TP53
- Ontology Terms for TP53

NCBI Resources How To Sign in to NCBI

GEO Profiles (TP53[Gene Symbol]) AND "homo sapiens"[Organism] AND colon tumor [Help](#)

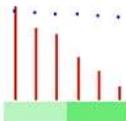
[Save search](#) [Limits](#) [Advanced](#)

Display Settings: Summary, 20 per page, Sorted by Subgroup effect [Send to:](#) [Filters: Manage Filters](#)

Results: 1 to 20 of 77 [<< First](#) [< Prev](#) Page of 4 [Next >](#) [Last >>](#)

TP53 - Methotrexate-resistant HT29 colon adenocarcinoma cell line

1. Annotation: TP53, **tumor** protein p53
Organism: **Homo sapiens**
Reporter: GPL570, 201746_at (ID_REF), GDS3330, 7157 (Gene ID), NM_000546
DataSet type: Expression profiling by array, count, 6 samples
ID: 54490195
[GEO DataSets](#) [Gene](#) [UniGene](#) [Profile neighbors](#) [Chromosome neighbors](#) [Sequence neighbors](#) [Homologene neighbors](#)



Profile data

Profile pathways

Find related data
Database:

Search details
TP53[Gene Symbol] AND "homo sapiens"[Organism] AND (colon[All Fields] AND tumor[All Fields])
 [See more...](#)

Recent activity

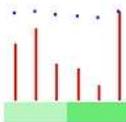
- (TP53[Gene Symbol]) AND "homo sapiens"[Organism] AND colon tumor (77) [GEO Profiles](#)
- Large airway epithelium response to cigarette smoking (HG-133A) [GDSBrowser](#)
- (GDS2490[ACCN]) AND GDS[filter] (1) [GDSBrowser](#)
- Large airway epithelium response to cigarette smoking (HG-133A) [GEO DataSets](#)
- Related DataSets for GEO Profiles (Select 33731615) (1) [GEO DataSets](#)

[See more...](#)

Important Links

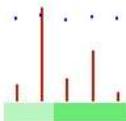
TP53 - Colon cancer progression

2. Annotation: TP53, **tumor** protein p53
Organism: **Homo sapiens**
Reporter: GPL96, 201746_at (ID_REF), GDS756, 7157 (Gene ID), NM_000546
DataSet type: Expression profiling by array, count, 6 samples
ID: 6167674
[GEO DataSets](#) [Gene](#) [UniGene](#) [Profile neighbors](#) [Chromosome neighbors](#) [Sequence neighbors](#) [Homologene neighbors](#)



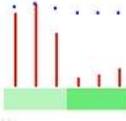
TP53 - PTEN deletion mutation effect on colon cancer cells

3. Annotation: TP53, **tumor** protein p53
Organism: **Homo sapiens**
Reporter: GPL96, 201746_at (ID_REF), GDS2446, 7157 (Gene ID), NM_000546
DataSet type: Expression profiling by array, count, 5 samples
ID: 32788874
[GEO DataSets](#) [Gene](#) [UniGene](#) [Profile neighbors](#) [Chromosome neighbors](#) [Sequence neighbors](#) [Homologene neighbors](#)



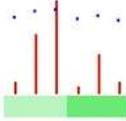
TP53 - Methotrexate-resistant HT29 colon adenocarcinoma cell line

4. Annotation: TP53, **tumor** protein p53
Organism: **Homo sapiens**
Reporter: GPL570, 211300_s_at (ID_REF), GDS3330, 7157 (Gene ID), K03199
DataSet type: Expression profiling by array, count, 6 samples
ID: 54499644
[GEO DataSets](#) [Gene](#) [UniGene](#) [Profile neighbors](#) [Chromosome neighbors](#) [Sequence neighbors](#) [Homologene neighbors](#)



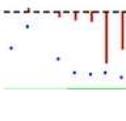
TP53 - Methotrexate resistant colon cancer cell line

5. Annotation: TP53, **tumor** protein p53
Organism: **Homo sapiens**
Reporter: GPL571, 211300_s_at (ID_REF), GDS3160, 7157 (Gene ID), K03199
DataSet type: Expression profiling by array, count, 6 samples
ID: 50274523
[GEO DataSets](#) [Gene](#) [UniGene](#) [Profile neighbors](#) [Chromosome neighbors](#) [Sequence neighbors](#) [Homologene neighbors](#)



TP53 - Colon tubular adenoma and carcinoma cells

6. Annotation: TP53, **tumor** protein p53
Organism: **Homo sapiens**
Reporter: GPL1818, 4176105 (ID_REF), GDS1777, NM_000546
DataSet type: Expression profiling by array, log ratio, 8 samples



- ▶ HomoloGene
- ▶ Reference Sequences for TP53
- ▶ SNP for TP53
- ▼ Gene phenotype for TP53

Gene - Morbid associations Data

Show entries

Search:

Phenotype accession	MIM Morbid Description
133239	ESOPHAGEAL CANCER
151623	LI-FRAUMENI SYNDROME 1; LFS1
202300	ADRENOCORTICAL CARCINOMA, HEREDITARY; ADCC
211980	LUNG CANCER
260500	PAPILLOMA OF CHOROID PLEXUS
275355	SQUAMOUS CELL CARCINOMA, HEAD AND NECK; HNSCC

Showing 1 to 6 of 6 entries

- ▶ Drug interactions for TP53
- ▶ External References for TP53

External References for TP53

See additional information about this gene in external databases

Show entries

Search:

Database	Description	Go to
Antibodypedia	an open-access database of publicly available antibodies against human protein targets	
CGAP	Cancer Genome Anatomy Project	
COSMIC	Catalogue of Somatic mutations in Cancer	
ENSEMBL	ENSEMBL gene	
Gene Cards	The human gene compendium	
HGNC	Hugo Gene Nomenclature Committee	
HPA	Human Protein Atlas - Antibody/Antigen	
HPA	Human Protein Atlas	
HPRD	Human Protein Reference Database	
iHOP	Information Hyperlinked over Proteins	
KEGG	Kyoto Encyclopedia of Genes and Genomes	
NCBI	Entrez Gene	
NCBI	Structures	
NCBI	Probe	
NCBI	Popset	
OMIM	Online Mendelian Inheritance in Man	
PubMed	PubMed entries	
STRING	Known and Predicted Protein-Protein Interactions	
Vega	Vertebrate Genome Annotation	
WikiGenes	Collaborative Publishing	

Showing 1 to 20 of 20 entries

Queries su molecole Farmacologicamente attive

- ✓ Drug name
- ✓ Pubchem Id e link diretti alla banca dati NCBI Pubchem
- ✓ Molecular weight
- ✓ IUPAC name
- ✓ Molecular formula
- ✓ Anatomical Therapeutic Chemical (ATC) Code
- ✓ Lista attività farmacologiche
- ✓ Struttura 2D della molecola
- ✓ Un elenco di prodotti di espressione con i quali interagisce la molecola e la tipologia di tali interazioni
- ✓ Pubmed Health entries

- ✓ STITCH – Network di interazioni note e predette tra la molecola e altri prodotti di espressione
- ✓ Pubchem Bioassay entries
- ✓ Sinonimi
- ✓ Side effects
- ✓ MeSH terms
- ✓ Bibliografia Pubmed correlata alla molecola oggetto della query

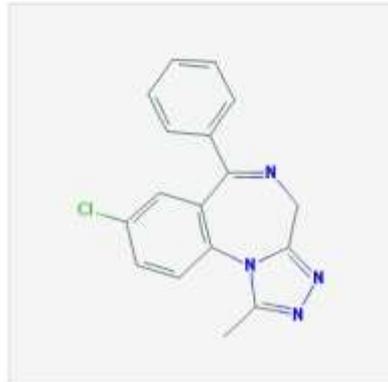


Drug detail

alprazolam - Pubchem ID 2118

• General Information about alprazolam

Drug Name	alprazolam
Pubchem Compound ID	2118 MC
ATC Classification system	N05BA12 MC
Molecular Weight	308.764920
Molecular Formula	C17H13ClN4 MC
2D Structure	



- Compound Descriptors
- Compound IDs
- MeSH term List for alprazolam
- Synonym list for alprazolam
- Pharmacological action of alprazolam
- Protein Interactions
- External references for alprazolam

› Synonym list for alprazolam

› Pharmacological action of alprazolam

▼ Protein Interactions

Genes interacting with the molecule according to MATADOR: Manually Annotated Targets and Drugs Online Resource

Show entries Search:

Gene	Official Full Name	Interaction type	STRING Reference
CH25H	cholesterol 25-hydroxylase	DIRECT	ENSP00000260706
COQ6	coenzyme Q6 homolog, monooxygenase (S. cerevisiae)	DIRECT	ENSP00000333946
CYP11A1	cytochrome P450, family 11, subfamily A, polypeptide 1	DIRECT	ENSP00000268053
CYP11B1	cytochrome P450, family 11, subfamily B, polypeptide 1	DIRECT	ENSP00000292427
CYP11B2	cytochrome P450, family 11, subfamily B, polypeptide 2	DIRECT	ENSP00000325822
CYP17A1	cytochrome P450, family 17, subfamily A, polypeptide 1	DIRECT	ENSP00000278017
CYP19A1	cytochrome P450, family 19, subfamily A, polypeptide 1	DIRECT	ENSP00000260433
CYP1A1	cytochrome P450, family 1, subfamily A, polypeptide 1	DIRECT	ENSP00000268062
CYP1A2	cytochrome P450, family 1, subfamily A, polypeptide 2	DIRECT	ENSP00000342007
CYP1B1	cytochrome P450, family 1, subfamily B, polypeptide 1	DIRECT	ENSP00000260630

Showing 1 to 10 of 71 entries ◀ ▶

› External references for alprazolam

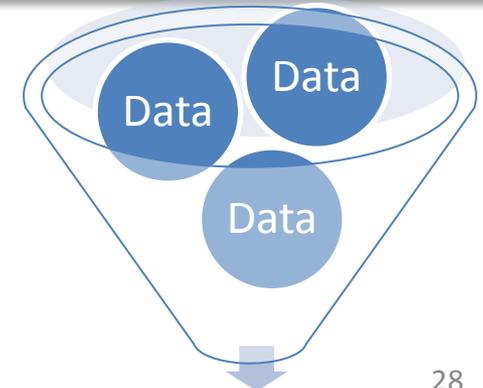
Scenario prossimo futuro

E' lecito prevedere una crescita sempre maggiore di dati ad elevata *dimensionalità*.

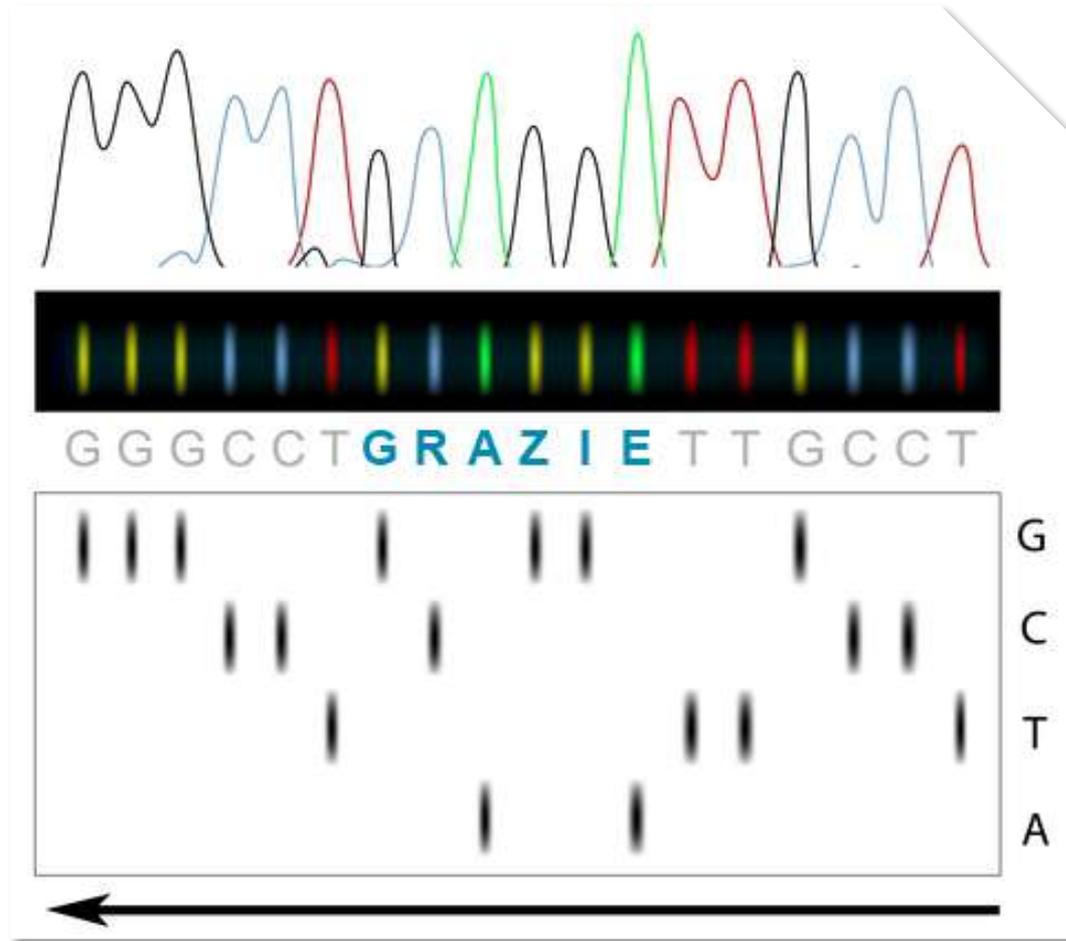
- ✓ Un numero sempre maggiore di banche dati e servizi è assolutamente atteso
- ✓ **Ridondanza** dei dati in un numero sempre maggiore di repositories pubblici
- ✓ Numero sempre crescente di accession identificativi delle *entries* nei database pubblici.
- ✓ Aumento delle attività di **Data Mining** su grandi volumi di dati collezionati su banche dati differenti con scopi e finalità diverse renderà sempre più difficile orientarsi e ottenere una sintesi di ciò che realmente si sta cercando

Sfide da raccogliere

- ✓ Integrazione di questa mole di informazioni costantemente in crescita; **riorganizzazione** in un contesto che sia quanto più possibile vicino alla natura dell'uomo di recepirle.
- ✓ Creazione di **servizi di integrazione** delle numerose fonti – intento del progetto **Biocloud search enGene**
- ✓ Trasformazione del dato in informazione
- ✓ **Trasformazione delle informazioni in conoscenza**



Conoscenza



Un ringraziamento particolare a:



REGIONE AUTONOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA