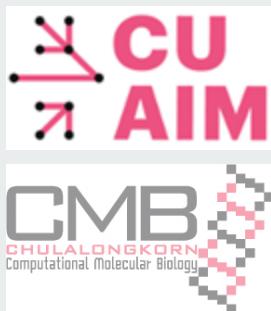


---

# 3000788 Intro to Comp Molec Biol

## Lecture 21: Online resources for biomedical research

Fall 2025



**Sira Sriswasdi, PhD**

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

# Today's agenda

---

- Why should we use online resources?
- Different types of online resources
- Examples of online resources



boston.iti.cs.cmu.edu

Answer to your research question may already be found in online databases and prediction tools



# Why should we use online sources?

# Why should we use online resources?

---

- Interpret your experimental results
  - Identified a mutation in a patient
    - What are potential molecular impacts?
    - Is it likely to be pathogenic?
    - How often is it found in the population?
  - Identified a set of differentially expressed genes
    - Which functions / pathways are enriched? How are they interacting?
    - Are they also observed in other perturbation studies?
    - Do some of them share transcription factor?
  - Identified a protein target for modulating a disease
    - What is the 3D structure for this protein? Can some drug bind to it?

# Why should we use online resources?

---

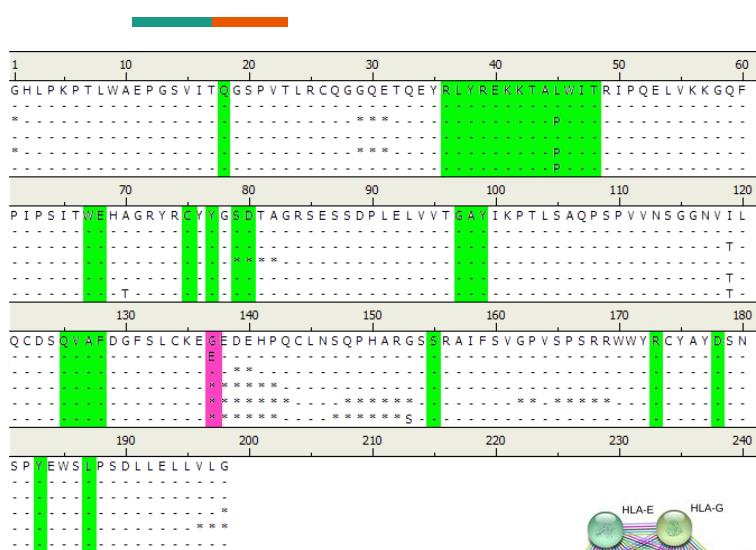
- Generalizing and validating your hypothesis
  - Identified three molecular subtypes in a local cohort
    - Are similar subtypes found in different populations?
    - Do subtypes show similar biomarkers across populations?
    - Are they associated with similar clinical outcomes?
  - Found certain drug to be effective for a disease
    - Has this drug been reportedly used to treat similar disease?

# Why should we use online resources?

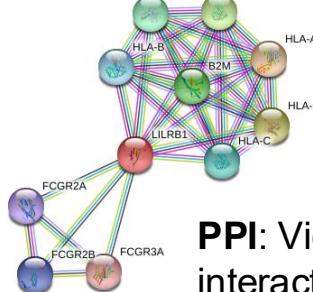
---

- Applying new analysis to existing data
  - Focusing on different factors
    - No paper analyzes every aspect of their data
    - Specific cell types
    - Specific pathways, e.g., surface proteins for T cell engineering
  - New techniques or interpretation
    - New assumption leads to new technique choice and findings
  - Increased sample size by aggregating data from multiple cohorts
    - More batch effects, BUT...
    - Opportunity to find generalizable patterns that transcend confounding effects

# Example 1: Interpreting the effects of a mutation

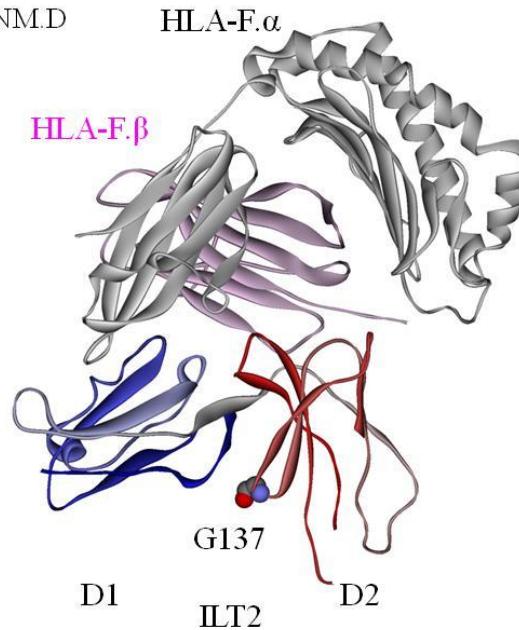


**Sequence:** Look up variant information on ClinVar and compare to commonly found mutations

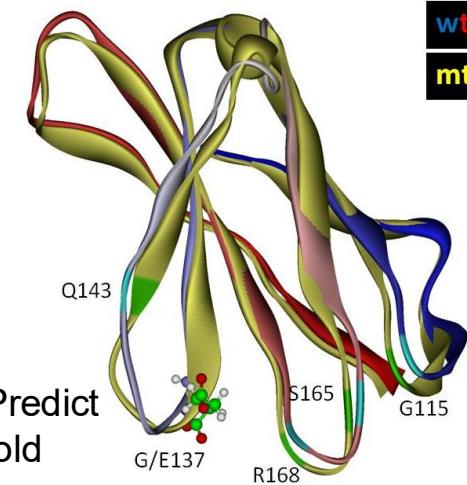


**PPI:** View potential interactors in STRING

5KNM.D



**Structure:** Highlight mutated residue on structure from PDB.



**Structure:** Predict with AlphaFold

# Example 1: Interpreting the effects of a mutation

## Accurate proteome-wide missense variant effect prediction with AlphaMissense

JUN CHENG  , GUIDO NOVATI, JOSHUA PAN, CLARE BYCROFT  , AKVILĖ ŽEMGULYTĖ, TAYLOR APPLEBAUM  , ALEXANDER PRITZEL, LAI HONG WONG,

MICHAL ZIELINSKI  , [...] , AND ŽIGA AVSEC 

+6 authors

[Authors Info & Affiliations](#)

SCIENCE • 19 Sep 2023 • Vol 381, Issue 6664 • DOI: 10.1126/science.adg7492



Available for download  
(known variant only)



- Available as API from  
Google cloud

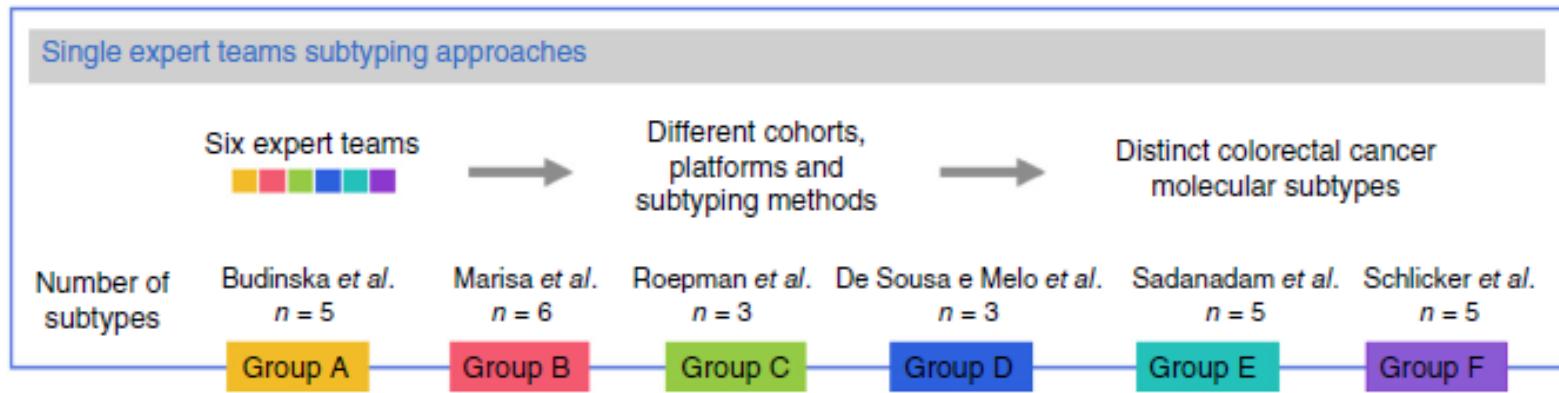
**AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model**

 Žiga Avsec,  Natasha Latysheva,  Jun Cheng,  Guido Novati,  Kyle R. Taylor,  Tom Ward,  Clare Bycroft,  Lauren Nicolaisen,  Eirini Arvaniti,  Joshua Pan,  Raina Thomas,  Vincent Dutordoir,  Matteo Perino,  Soham De,  Alexander Karollus,  Adam Gayoso,  Toby Sargeant,  Anne Mottram,  Lai Hong Wong,  Pavol Drotár,  Adam Kosiorek,  Andrew Senior,  Richard Tanburn,  Taylor Applebaum,  Souradeep Basu,  Demis Hassabis,  Pushmeet Kohli

doi: <https://doi.org/10.1101/2025.06.25.661532>

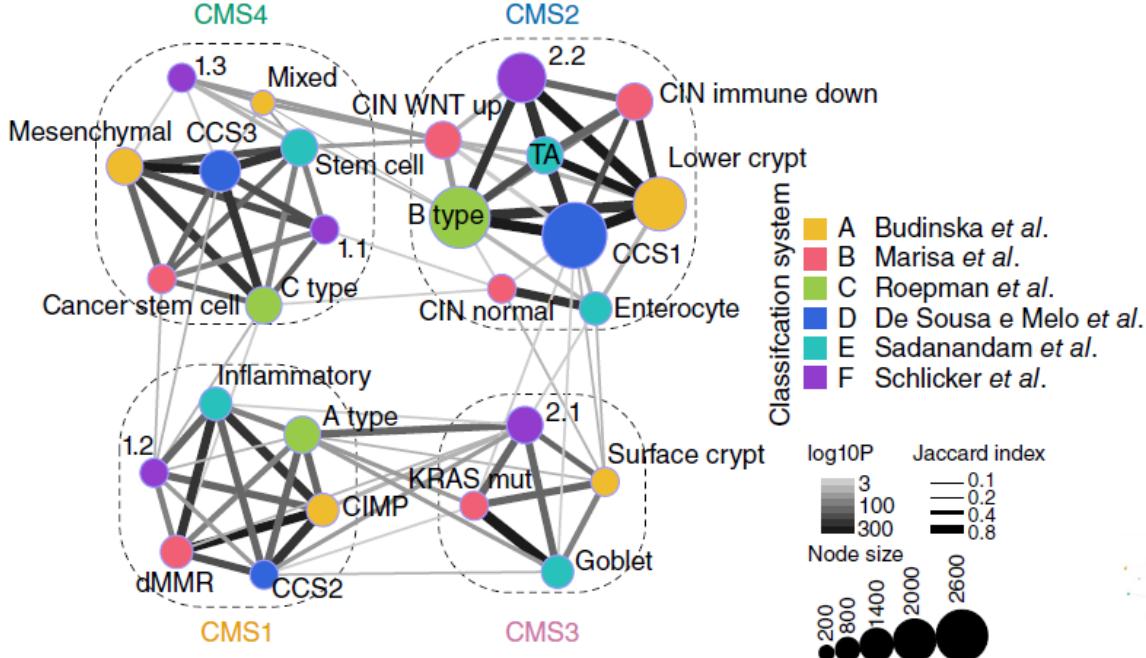
This article is a preprint and has not been certified by peer review [what does this mean?].

## Example 2: Consensus cancer molecular subtyping



- Different clinicians use different systems to characterize cancer patients
- But there should be a common molecular basis that applies to cancer patients from any demographics!

## Example 2: Consensus cancer molecular subtyping



- Apply 6 subtyping schemes on a group of 4,000 patients
- Identify subtypes often assigned to the same patients
  - 1.1 / CCS3 / Stem cell / Mesenchymal / C type
- Consensus!



# Knowledgebases

# Classical, foundational databases

- NCBI/GenBank
- Ensembl
- ENCODE
- Uniprot
- GeneCard
- InterPro
- OMIM

UniProtKB - O75473 (LGR5\_HUMAN)

Display [Help video](#)

Entry      [BLAST](#) [Align](#) [Format](#) [Add to basket](#) [History](#)

Protein | Leucine-rich repeat-containing G-protein coupled receptor 5  
Gene | LGR5  
Organism | Homo sapiens (Human)  
Status | Reviewed - Annotation score: 5/5 - Experimental evidence at protein level<sup>i</sup>

Function<sup>i</sup>

Receptor for R-spondins that potentiates the canonical Wnt signaling pathway and acts as a stem cell marker of the intestinal epithelium by binding to R-spondins (RSPO1, RSPO2, RSPO3 or RSPO4), associates with phosphorylated LRP6 and frizzled receptors that are activated by triggering the canonical Wnt signaling pathway to increase expression of target genes. In contrast to classical G-protein coupled receptors heterotrimeric G-proteins to transduce the signal. Involved in the development and/or maintenance of the adult intestinal stem cells during

5 Publications

Miscellaneous

LGR5 is used as a marker of adult tissue stem cells in the intestine, stomach, hair follicle, and mammary epithelium. 1 Publication

GO - Molecular function<sup>i</sup>

- G protein-coupled peptide receptor activity
- G protein-coupled receptor activity
- protein-hormone receptor activity
- transmembrane signaling receptor activity

Complete GO annotation on QuickGO ...

GO - Biological process<sup>i</sup>

# Some common accession systems

---

ID system	Accession
Gene Symbol	LGR5
HGNC gene	4504
Entrez (NCBI)	8549
RefSeq transcript	NM_00367
Ensembl	ENSG00000139292
Uniprot protein	O75473
OMIM	606667

# ENCODE: database for epigenetics profiles



- Processed ChIP-seq, ATAC-seq, etc.

# GeneCard: finding gene aliases

---

## Aliases for LGR5 Gene

### Aliases for LGR5 Gene

GeneCards Symbol: *LGR5* <sup>2</sup>

Leucine Rich Repeat Containing G Protein-Coupled Receptor 5 <sup>2 3 5</sup>

GPR49 <sup>3 4 5</sup>

GPR67 <sup>3 4 5</sup>

HG38 <sup>2 3 5</sup>

FEX <sup>2 3 5</sup>

Leucine-Rich Repeat-Containing G-Protein Coupled Receptor 5 <sup>3 4</sup>

### External Ids for LGR5 Gene

HGNC: 4504 NCBI Entrez Gene: 8549 Ensembl: ENSG00000139292 OMIM®: 606667 UniProtKB/Swiss-Prot: O75473

### Previous HGNC Symbols for LGR5 Gene

GPR67, GPR49

- The gene you want to find may be reported with different names, depending on the database used during bioinformatics analysis

# JASPAR: database of DNA binding motifs

Detailed information of matrix profile **MA0161.2**

Home > Matrix > MA0161.2

Profile summary	
Name:	NFIC
Matrix ID:	MA0161.2
Class:	SMAD/NF-1 DNA-binding domain factors
Family:	Nuclear factor 1
Collection:	CORE
Taxon:	Vertebrates
Species:	Homo sapiens
Data Type:	ChIP-seq
Validation:	12101405
Uniprot ID:	P08651
Source:	29126285
Comment:	

Sequence logo 



Frequency matrix      


	A [	2615	3896	412	154	352	310	223	264	10158	2626	334-
C [	2647	2137	10542	303	333	129	121	10725	397	2951	271-	
G [	2478	2086	329	369	177	11087	10973	88	478	3186	276-	
T [	3867	3488	324	10781	10745	81	290	530	574	2844	279.	

# Biomart: bulk gene/protein annotation download

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

 Search all species

**Dataset**  
Human genes (GRCh38.p13)

**Filters**  
[None selected]

**Attributes**  
Gene stable ID  
Gene stable ID version  
Transcript stable ID  
Transcript stable ID version  
PDB ID

**Dataset**  
[None Selected]

Structures       Variant (Somatic)  
 Homologues (Max select 6 orthologues)       Sequences

GENE:

EXTERNAL:

**GO**

GO term accession       GO term evidence code  
 GO term name       GO domain  
 GO term definition

**GOSlim GOA**

GOSlim GOA Accession(s)       GOSlim GOA Description

**External References (max 3)**

BioGRID Int Information from external database sources.  STRING: Protein-protein interact...  NCBI gene (formerly Entrezgene) accession  
 Datasets ID  NCBI gene (formerly Entrezgene) ID  
 CCDS ID  PDB ID  
 ChEMBL ID  Reactome ID  
 DataBase of Aberrant 3' Splice Sites name  Reactome gene ID  
 DataBase of Aberrant 3' Splice Sites ID  Reactome transcript ID  
 DataBase of Aberrant 5' Splice Sites name  RefSeq mRNA ID  
 DataBase of Aberrant 5' Splice Sites ID  RefSeq mRNA predicted ID

# Biomart: bulk gene/protein annotation download

Dataset 11465 / 69292 Genes Human genes (GRCh38.p13) Filters Chromosome/scaffold: 1 , 4 , 5 Attributes Gene stable ID Gene stable ID version Transcript stable ID Transcript stable ID version PDB ID

Export all results to File TSV  Unique results only  Go

Email notification to

View 10 rows as HTML  Unique results only

Gene stable ID	Gene stable ID version	Transcript stable ID	Transcript stable ID version	PDB ID
<a href="#">ENSG00000160072</a>	<a href="#">ENSG00000160072.20</a>	<a href="#">ENST00000673477</a>	<a href="#">ENST00000673477.1</a>	
<a href="#">ENSG00000160072</a>	<a href="#">ENSG00000160072.20</a>	<a href="#">ENST00000308647</a>	<a href="#">ENST00000308647.8</a>	
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST00000511072</a>	<a href="#">ENST00000511072.5</a>	
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST00000378391</a>	<a href="#">ENST00000378391.6</a>	<a href="#">2N1I</a>
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST00000378391</a>	<a href="#">ENST00000378391.6</a>	<a href="#">6BW4</a>
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST00000514189</a>	<a href="#">ENST00000514189.5</a>	
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST00000270722</a>	<a href="#">ENST00000270722.10</a>	<a href="#">2N1I</a>
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST00000270722</a>	<a href="#">ENST00000270722.10</a>	<a href="#">6BW4</a>
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST00000463591</a>	<a href="#">ENST00000463591.1</a>	
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST00000509860</a>	<a href="#">ENST00000509860.1</a>	

- Useful as input for bioinformatics tools

# MSigDB: Curated gene sets

---

## Human MSigDB Collections



The 33196 gene sets in the Human Molecular Signatures Database (MSigDB) are divided into 9 major collections, and several sub-collections. See the table below for a brief description of each, and the [Human MSigDB Collections: Details and Acknowledgments](#) page for more detailed descriptions. See also the [MSigDB Release Notes](#).

Click on the "browse gene sets" links in the table below to view the gene sets in a collection. Or download the gene sets in a collection by clicking on the links below the "Download Files" headings. For a description of the GMT file format see the [Data Formats](#) in the Documentation section. The gene sets can be downloaded as NCBI (Entrez) Gene Identifiers or HUGO (HGNC) Gene Symbols. There are also JSON bundles containing the HUGO (HGNC) Gene Symbols along with some useful metadata. An XML file containing all the Human MSigDB gene sets is available as well.

### H: hallmark gene sets (browse 50 gene sets)

Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying overlaps between gene sets in other MSigDB collections and retaining genes that display coordinate expression. [details](#)

[Download GMT Files](#)  
[Gene Symbols](#)  
[NCBI \(Entrez\) Gene IDs](#)  
[JSON bundle](#)

### C1: positional gene sets (browse 299 gene sets)

Gene sets corresponding to human chromosome cytogenetic bands. [details](#)

[Download GMT Files](#)  
[Gene Symbols](#)  
[NCBI \(Entrez\) Gene IDs](#)  
[JSON bundle](#)

### C2: curated gene sets (browse 6449 gene sets)

Gene sets in this collection are curated from various sources, including online pathway databases and the biomedical literature. Many sets are also contributed by individual domain experts. The gene set page for each gene set lists its source. The C2 collection is divided into the following two sub-collections: Chemical and genetic perturbations (CGP) and Canonical pathways (CP). [details](#)

[Download GMT Files](#)  
[Gene Symbols](#)  
[NCBI \(Entrez\) Gene IDs](#)  
[JSON bundle](#)

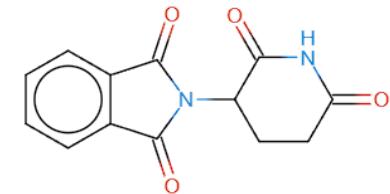
- Function-based / disease-based / perturbation-based / publication-based / genomic location-based gene sets
- Useful for creating gene panel for enrichment test and picking gene markers for visualization

# Open Targets: drug-centric database



## THALIDOMIDE

ChEMBL: [CHEMBL468](#) | DrugBank: [DB01041](#) | ChEBI: [74947](#) | DailyMed: [thalidomide](#)



Indication	Therapeutic Area	Gene	rsID	Star Allele	Genotype ID?	Variant Consequence	Drug Response Phenotype
immune system disease	immune system	SULT1C4	rs1402467	N/A	2_108378352_C_G,G	VEP Missense variant ProtVar	increased chance of response to tre
multiple myeloma	4 areas	CHST3	rs1871450	N/A	10_72012256_G_A,A	VEP 3 prime utr variant	increased risk of toxicity?
diffuse large B-cell lymphoma	4 areas				Phase III	2 entries	
colorectal neoplasm	2 areas				Phase III	ClinicalTrials.gov	
fallopian tube cancer	2 areas				Phase III	ClinicalTrials.gov	
vascular malformation	2 areas				Phase III	ClinicalTrials.gov	
peritoneum cancer	2 areas				Phase III	ClinicalTrials.gov	
malignant epithelial tumor of ovary	3 areas				Phase III	ClinicalTrials.gov	
hepatocellular carcinoma	3 areas				Phase III	7 entries	
extranodal nasal NK/T cell lymphoma	4 areas				Phase III	ClinicalTrials.gov	



# Data repository

# Sequence Read Archive (SRA): sequencing data

454 sequencing of Human HapMap individual NA18505 genomic paired-end library (SRR000001)

Metadata Analysis Reads Data access

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR000001	471.0k	129.5Mbp	312.5M	41.3%	2008-04-04	public

Quality graph ([bigger](#))

This run has 4 reads per spot:

L=4, 100%       $\bar{L}=187, \sigma=95.9, 100\%$       L=44, 50%       $\bar{L}=123, \sigma=65.5, 50\%$

[Legend](#)

Also, European Nucleotide Archive (ENA) and Japan's DDBJ

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout	Action
SRX000007	SID2748	LS454	WGS	GENOMIC	RANDOM	PAIRED	BLAST

Biosample	Sample Description	Organism	Links
<a href="#">SAMN00001583</a> (SRS000100)	Human HapMap individual Coriell catalog ID NA18505	<a href="#">Homo sapiens</a>	<ul style="list-style-type: none"><li><a href="#">dbSNP Batch ID 1061891</a></li><li><a href="#">Individual record in dbSNP</a></li></ul>

Bioproject	SRA Study	Title
<a href="#">PRJNA33627</a>	<a href="#">SRP000001</a>	Paired-end mapping reveals extensive structural variation in the human genome

Show abstract

# PRIDE / MassIVE / ProteomeXchange

[\*\*<< Full experiment listing\*\*](#)

**PXD027487**

PXD027487 is an original dataset announced via ProteomeXchange.

Dataset Summary	
Title	Interactome of GIGYF2 and EIF4E2 with and without SMG1i treatment
Description	Translation of messenger RNAs (mRNAs) with premature translation termination codons produces truncated proteins with potentially deleterious effects. This is prevented by nonsense-mediated mRNA decay (NMD) of these mRNAs. NMD is triggered by ribosomes terminating upstream of a splice site marked by an exon-junction complex (EJC), but also acts on many mRNAs lacking a splice junction after their termination codon. We developed a genome-wide CRISPR flow cytometry screen to identify regulators of mRNAs with premature termination codons in K562 cells. This screen recovered essentially all core NMD factors and suggested a role for EJC factors in degradation of PTCs without downstream splicing. Among the strongest hits were the translational repressors GIGYF2 and EIF4E2. GIGYF2 and EIF4E2 mediate translational repression but not mRNA decay of a subset of NMD targets and interact with NMD factors genetically and physically. Our results suggest a model wherein recognition of a stop codon as premature can lead to its translational repression through GIGYF2 and EIF4E2.
HostingRepository	PRIDE
AnnounceDate	2021-10-12
AnnouncementXML	<a href="#">Submission_2021-10-12_00:39:40.790.xml</a>



- Repository for raw mass spectrometry data (mostly proteomics)
- Search for already identified peptides/proteins

# PepQuery: database for cancer proteomics

PepQueryDB: a database of indexed MS/MS spectra for PepQuery search

48

Datasets



25405

Total raw MS files

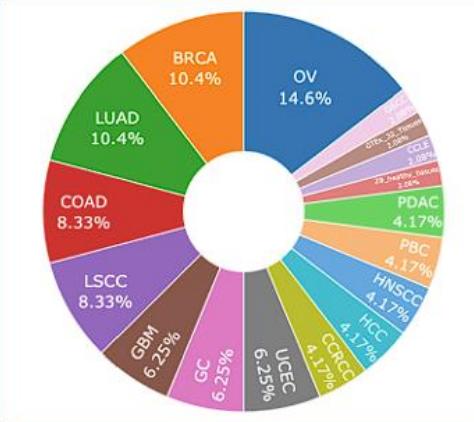


1,012,696,284

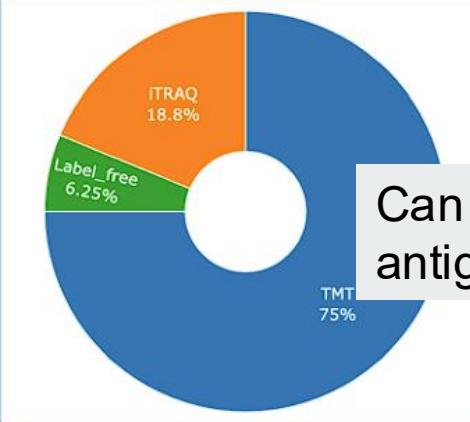
Total MS/MS spectra



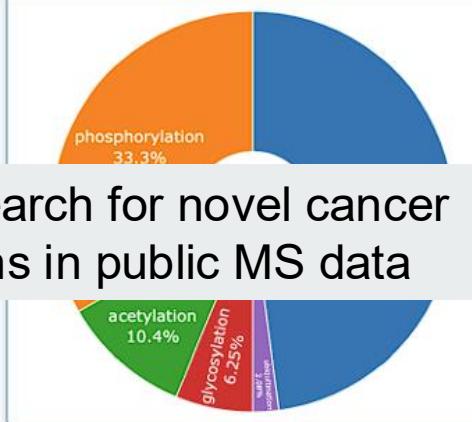
Sample type



Experimental type

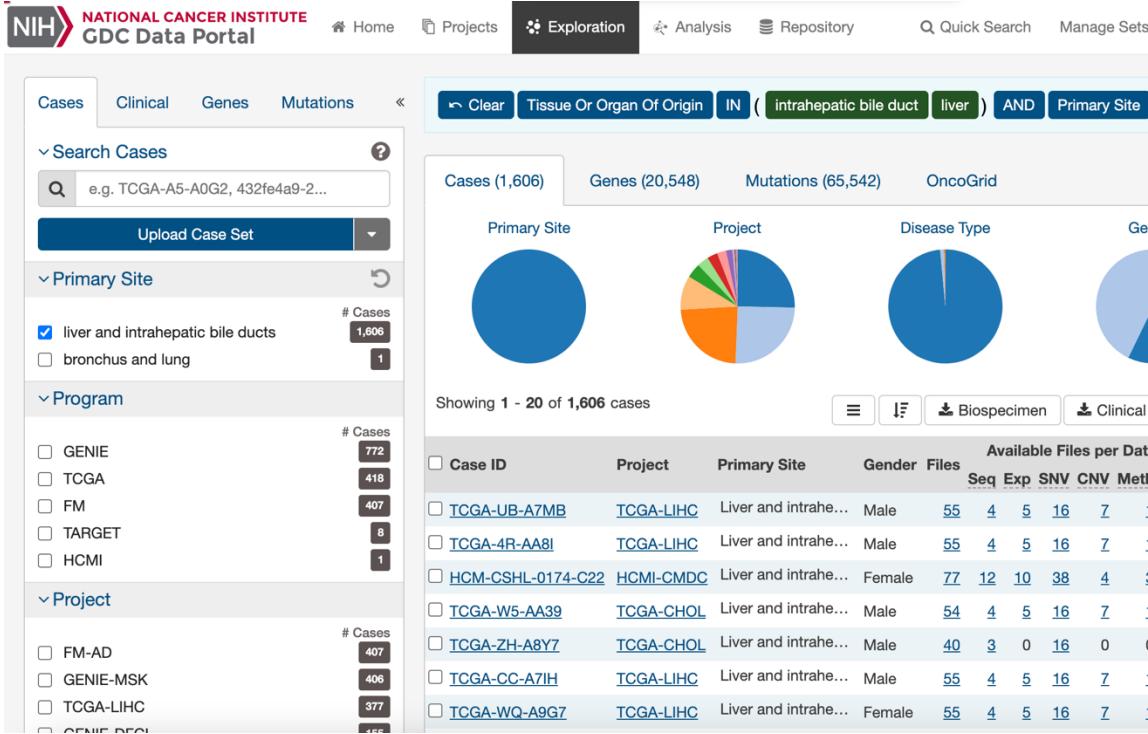


PTM type



Can search for novel cancer antigens in public MS data

# Genomic Data Commons: cancer multiomics



- Cancer multi-omics
  - Exome
  - RNA-seq
  - Methylation
- Include clinical and demographic data
- Link to pathology and radiology images / proteomics data in some cases

# Protein Data Bank: database of 3D structures

RCSB PDB Deposit Search Visualize Analyze Download Learn More Documentation Careers MyPDB ▾

Refinements ? 

SCIENTIFIC NAME OF SOURCE ORGANISM

- Homo sapiens (13)
- Xenopus tropicalis (11)
- Mus musculus (10)
- Danio rerio (2)
- unidentified (1)

TAXONOMY

- Eukaryota (32)
- unclassified sequences (1)

EXPERIMENTAL METHOD

- X-RAY DIFFRACTION (32)
- SOLUTION NMR (1)

POLYMER ENTITY TYPE

- Protein (33)

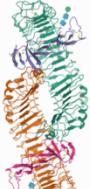
REFINEMENT RESOLUTION (Å)

- 1.5 - 2.0 (2)
- 2.0 - 2.5 (11)
- 2.5 - 3.0 (9)
- 3.0 - 3.5 (8)
- 4.0 - 4.5 (1)
- > 4.5 (1)

Summary Gallery Compact -- Tabular Report -- Score Download Files  All  Selected

Display 25 per page

Displaying 1 to 25 of 33 Structures Page 1 of 2 ← Previous Next →

  
[3D View](#)

**4BSU**

[Structure of the ectodomain of LGR5 in complex with R-spondin-1 \(Fu1Fu2\) in C2 crystal form](#)

Peng, W.C., de Lau, W., Forneris, F., Granneman, J.C.M., Huch, M., Clevers, H., Gros, P.

(2013) Cell Rep 3: 1885

**Released** 2013-06-26  
**Method** X-RAY DIFFRACTION 3.2 Å  
**Organisms** Homo sapiens  
**Macromolecule** LEUCINE-RICH REPEAT-CONTAINING G-PROTEIN COUPLED RECEPTOR 5 (protein)  
R-SPONDIN-1 (protein)  
**Unique Ligands** NAG  
**Unique branched monosaccharides** BMA, NAG



**4BST**

[Structure of the ectodomain of LGR5 in complex with R-spondin-1 \(Fu1Fu2\) in P6122 crystal form](#)

Peng, W.C., de Lau, W., Forneris, F., Granneman, J.C.M., Huch, M., Clevers, H., Gros, P.

(2013) Cell Rep 3: 1885

**Released** 2013-06-19  
**Method** X-RAY DIFFRACTION 4.3 Å

# Gene Expression Omnibus / ArrayExpress

Study type	see all
<input checked="" type="checkbox"/> chip-chip by tiling array	706
<input checked="" type="checkbox"/> chip-seq	1,467
<input checked="" type="checkbox"/> comparative genomic hybridization by array	883
<input checked="" type="checkbox"/> genotyping by array	254
<input checked="" type="checkbox"/> methylation profiling by array	617
<input checked="" type="checkbox"/> other	222
<input checked="" type="checkbox"/> rna-seq of coding rna	1,437
<input checked="" type="checkbox"/> rna-seq of non coding rna	241
<input checked="" type="checkbox"/> transcription profiling by array	9,339
<input checked="" type="checkbox"/> unknown experiment type	218

You query contains a term which has too many synonyms and more specific phrases in EFO. The results shown below do not include those expanded terms.

## Search results for breast cancer metastasis

1 – 20 of 1,417+ results

Sort by: Relevance ▾ ▾ ^

E-MTAB-4801 • 9 January 2019 • 3 links • 3 files

### Variation in RNA expression in a panel of 30 breast cancer cell lines

... type comparison design EFO\_0001745 cell line cell line BT20 BT474 BT549 ... 27 other values Homo sapiens female mammary gland invasive ductal carcinoma breast ductal adenocarcinoma metaplastic breast carcinoma squamous cell breast carcinoma, acantholytic variant breast adenocarcinoma breast...

E-MTAB-8807 • 4 April 2020 • 1 link • 2 files

### Estrogen receptor beta inhibits cholesterol biosynthesis through overexpression of mir-181a-5p in Triple Negative Breast Cancer

... carcinoma squamous cell breast carcinoma, acantholytic variant not specified invasive breast ductal carcinoma invasive lobular carcinoma adenocarcinoma atypical carcinoma carcinoma ER beta negative ER beta positive HCC1806 null null null epithelial cell squamous cell breast carcinoma, acantholytic...

# Google Dataset / FigShare



## Dataset Search

breast cancer metastasis

1,494,669 results found

- Whole-exome sequencing of **breast cancer metastasis** and corresponding blood samples
- Genomic Evolution of **Breast Cancer Metastasis** and Relapse
- A clinical decision support system learned from data to personalize treatment recommendations towards preventing **breast cancer metastasis**
- Cooperativity between EMT and non-EMT cells promotes **breast cancer metastasis** via paracrine GLI activation
- Primary tumor grafts as advanced models for breast cancer that authentically reflect tumor histopathology, growth, metastasis, and patient outcomes (copy number)
- Impramine Blue: A potent therapeutic regimen that suppresses breast cancer growth and metastasis

figshare

data

The screenshot shows the figshare search interface. At the top right is the figshare logo with a colorful circular icon. Below it is a search bar containing the query 'breast cancer metastasis'. To the right of the search bar is a blue button with a magnifying glass icon. Underneath the search bar, the text '1,494,669 results found' is displayed. The results are presented in a grid format. The first result is an image of the Royal College of Surgeons in Ireland (RCSI) crest. Below the image, the title 'Investigating novel mechanisms of metastasis in ...' is visible, along with a brief abstract and author information. The second result is another image of the RCSI crest, with the title 'Stratification of radiosensitive brain metastases based o...' and a similar abstract. The third result is an image of the RCSI crest, with the title 'Identification of Molecular Mediators of Endocrine ...' and a similar abstract.

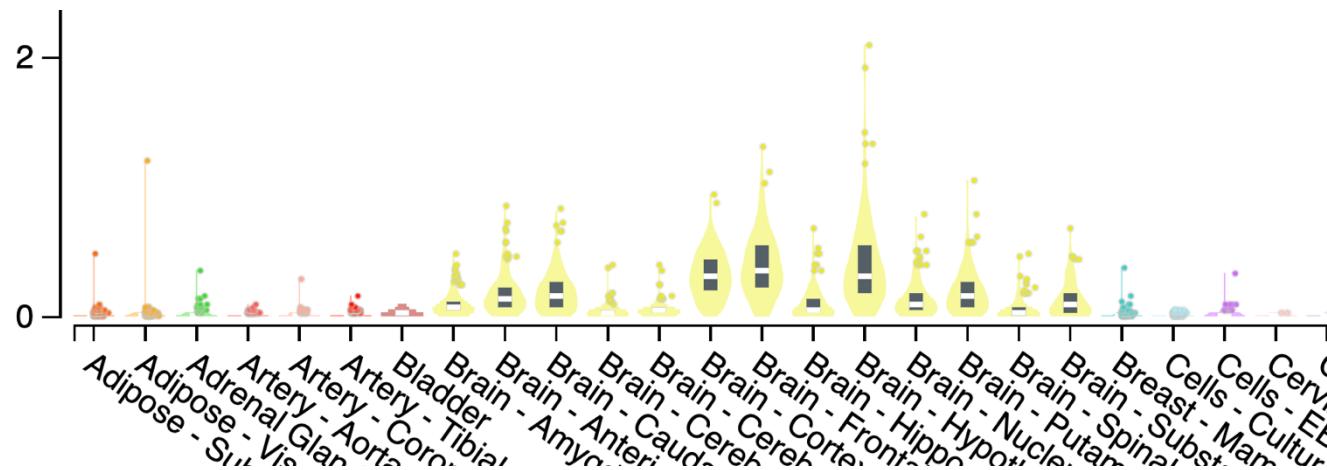
- Google Dataset compiles links from major public data repositories
- Figshare is a free repository that scientists often use for non-omics data



# Atlas of human data

# GTEx: database for tissue transcriptomics markers

Gene Symbol	Gencode ID	Entrez Gene ID	Location	Gene Description
LIN28B	ENSG00000187772.7	389421	chr6:104936616-105083332:+	lin-28 homolog B [Source:HGNC Symbol;Acc:HGNC:32207]



- Gene expression profiles for normal tissues (human & primates)

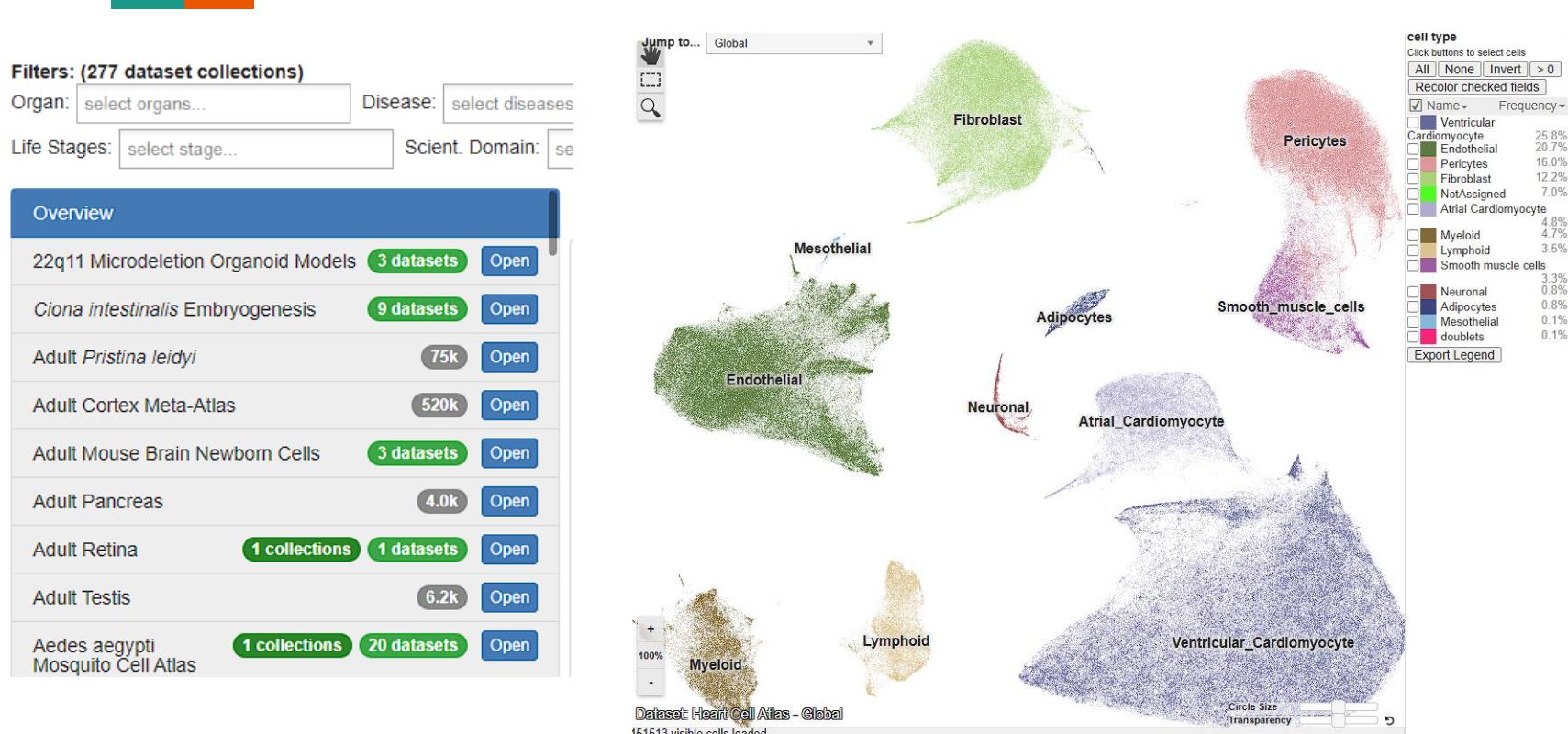
# Human Cell Atlas: cell type-specific omics data

The screenshot shows the Human Cell Atlas Data Portal interface. At the top, there is a navigation bar with links for Datasets, HCA BioNetworks, Guides, Metadata, Contribute, APIs, Updates, and Follow HCA. Below the navigation bar, the page title is "Integrated human endoderm-derived organoids cell atlas (HEOCA) v1.0". There are two tabs: "Atlas Overview" (selected) and "Source Datasets". A section titled "Component Atlases" lists three entries:

Atlas Name	Tissue	Disease	Cells	Explore	Download
human endoderm organoid cell atlas	21 tissues	normal	806.6k	CZ CELLxGENE	
human intestine organoid cell atlas	4 tissues	normal	353.1k	CZ CELLxGENE	
human lung organoid cell atlas	5 tissues	normal	221.4k	CZ CELLxGENE	

The screenshot shows the "Heart Network" page under the "HCA Biological Network Atlases". The page has a header with the "Heart Network" logo and navigation links for "Network Overview" and "HCA Heart Datasets" (which is currently selected). Below the header, there is a section with the text: "The HCA datasets below focus on the network's tissues and are". Underneath this, there is a "Project Title" field containing the text "A Cellular Atlas of Pitx2-Dependent Cardiac Development.", followed by two additional project titles: "A cell and transcriptome atlas of human arterial vasculature" and "A human cell atlas of fetal gene expression.".

# UCSC Cell Browser: visualizer for public single-cell data



# Human Protein Atlas: body-wide biomarkers

ACE2



RNA  
TISSUE

RNA  
BRAIN

RNA  
SINGLE CELL

RNA  
TISSUE CELL

RNA  
PATHOLOGY

RNA  
IMMUNE

MS  
BLOOD

RNA  
SUBCELL

RNA  
CELL LINE



## PROTEIN SUMMARY

RNA DATA

## GENE/PROTEIN

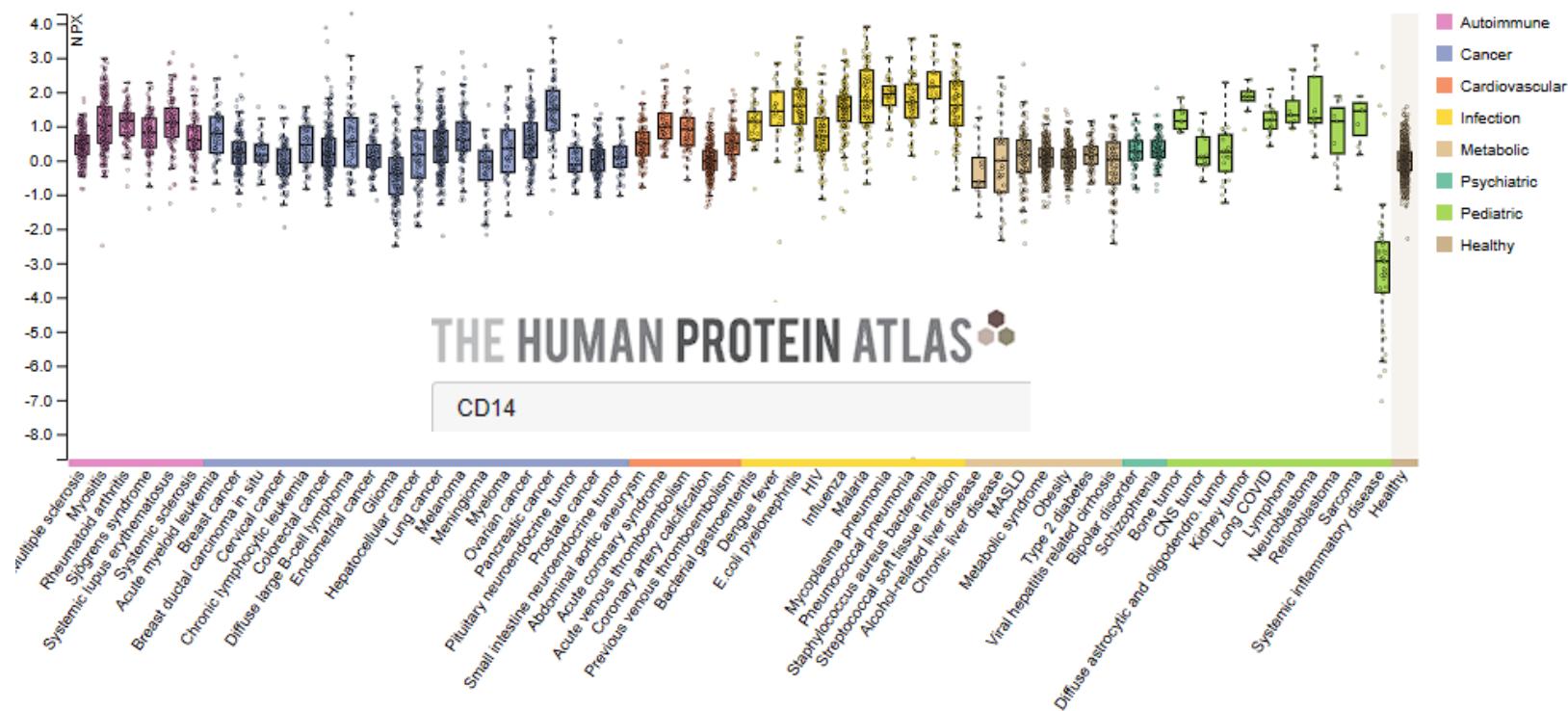
## ANTIBODIES AND VALIDATION



### HUMAN PROTEIN ATLAS SUMMARY<sup>i</sup>

Protein <sup>i</sup>	Angiotensin I converting enzyme 2
Gene name <sup>i</sup>	ACE2
Tissue specificity <sup>i</sup>	Tissue enhanced (gallbladder, intestine, kidney)
Tissue expression cluster <sup>i</sup>	Intestine & Kidney - Transmembrane transport (mainly)
Single cell type specificity <sup>i</sup>	Cell type enriched (Proximal enterocytes)
Single cell type expression cluster <sup>i</sup>	Enterocytes - Digestion (mainly)
Immune cell specificity <sup>i</sup>	Not detected in immune cells
Brain specificity <sup>i</sup>	Not detected in human brain
Cancer prognostic summary	Prognostic marker in renal cancer (favorable) and liver cancer (favorable)
Predicted location <sup>i</sup>	Membrane, Secreted (different isoforms)
Extracellular location <sup>i</sup>	Secreted to blood

# Human Protein Atlas: disease biomarkers

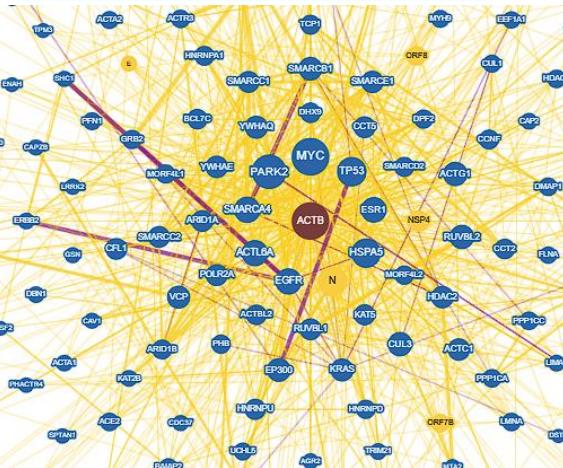




# Exploration of interaction networks

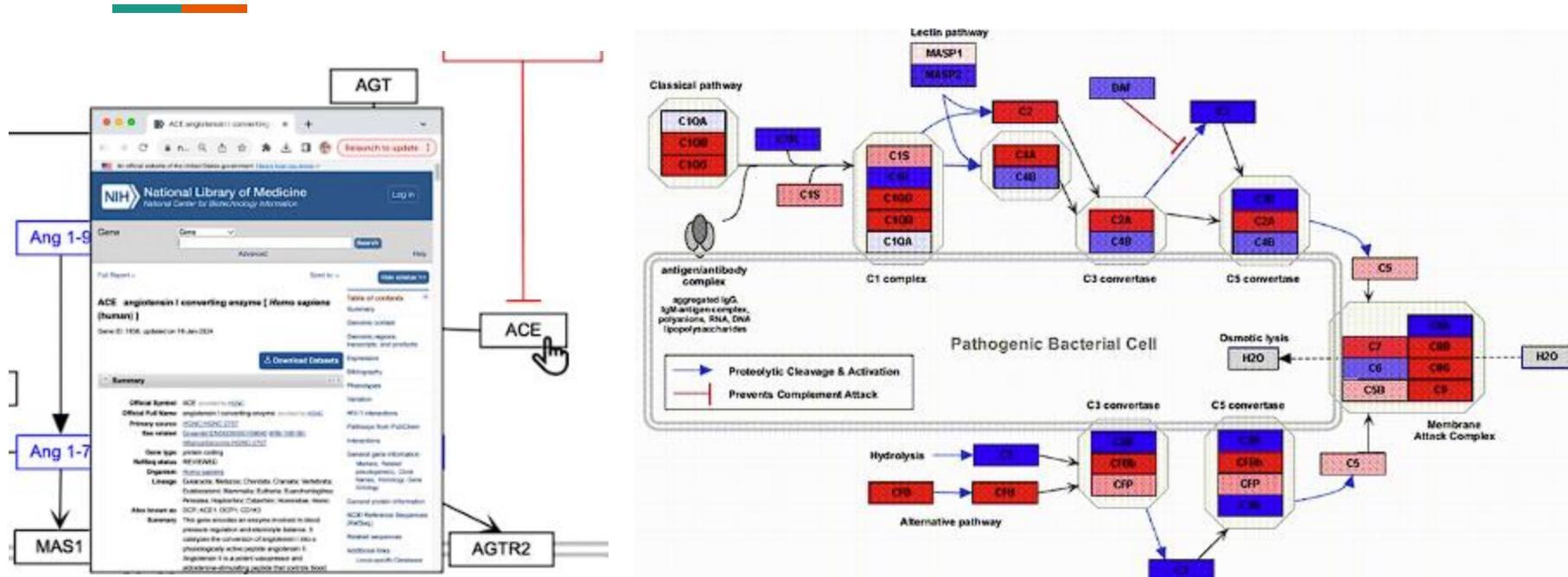
# BioGrid: database for protein-protein interactions

ACTB
actin, beta
1213 unique interactors
1683 raw interactions
31 post-translational modifications



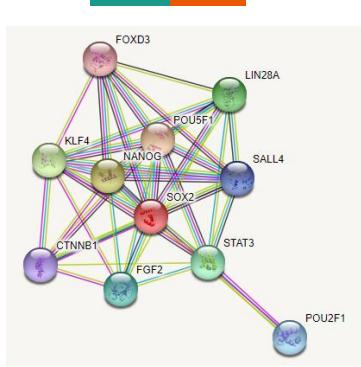
Interactor	Role	Organism	Experimental Evidence Code	Dataset	Throughput	HTP Score
2510003E04RIK	BAIT	<i>M. musculus</i>	Affinity Capture-MS	Hein MY (2015)	High	-
A2M	BAIT	<i>H. sapiens</i>	Two-hybrid	Soler-Lopez M (2011)	Low	-
AAR2	BAIT	<i>H. sapiens</i>	Affinity Capture-MS	Malinova A (2017)	High	-
AASDHPPPT	BAIT	<i>H. sapiens</i>	Cross-Linking-MS (XL-MS)	Wheat A (2021)	High	-
ABCG8	BAIT	<i>H. sapiens</i>	Affinity Capture-MS	Hutlin EL (2021/pre-pub)	High	0.4281
ABI1	HIT	<i>H. sapiens</i>	Affinity Capture-MS	Cho NH (2022)	High	-
ABI2	HIT	<i>H. sapiens</i>	Affinity Capture-MS	Cho NH (2022)	High	-
ABLIM1	BAIT	<i>H. sapiens</i>	Reconstituted Complex	Roof DJ (1997)	Low	-
ABLIM1	HIT	<i>H. sapiens</i>	Proximity Label-MS	Go CD (2021)	High	22.93

# WikiPathway: integrating pathway with expression data

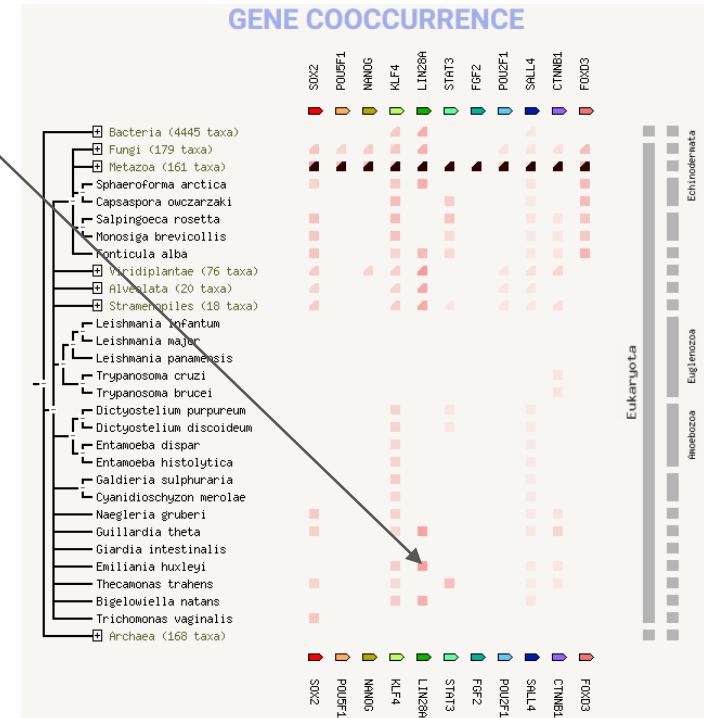


- Visualization of gene/protein activity on top of interaction networks

# STRING: multi-evidence protein-protein interaction



LIN28A	25	PEDAARRAAD-EPQLLHGAGICKWFNVRMGGFGFLSMTARAGVALDPVVDV P A R AD EP +G CKWF+V+ GFGF+ + + D+ EOD22827 19
LIN28A	73	FVHQSKLHMEGFRSLKEGEAVEFTFKKSAGK--LESIRVTGPGGVFCIG FVHQ+ + +GFRSL EGEA+EF + AK L++I VTGPGG F G EOD22827 58
LIN28A	120	SERRP + R P APREP EOD22827 107



- Combine experimental evidence with literature mining (co-occurrence) and evolution (co-appear in same taxa)
- Link to external 3D structure and protein expression data



# Analysis tools

# miRNA target

 **TargetScanHuman**  
Prediction of microRNA targets

Release 7.1: June 2016      Agarwal et al., 2015

Search for predicted microRNA targets in mammals

[Go to TargetScanMouse]  
[Go to TargetScanWorm]  
[Go to TargetScanFly]  
[Go to TargetScanFish]

1. Select a species

AND

2. Enter a human gene symbol (e.g. "Hmga2")   
or an Ensembl gene (ENSG00000149948) or transcript (ENST00000403681) ID

AND/OR

3. Do one of the following:

- Select a broadly conserved\* microRNA family
- Select a conserved\* microRNA family
- Select a poorly conserved but confidently annotated microRNA family
- Select another miRBase annotation  
Other miRBase annotations

Note that most of these families are star miRNAs or RNA fragments misannotated as miRNAs.

Enter a microRNA name (e.g. "miR-9-5p")

 miRBase

Home | Search | Browse | Help | Download | Blog | Submit

Latest miRBase blog posts

High confidence miRNAs set available for miRBase 21  
By sam (July 3, 2014)  
As mentioned previously, we briefly held off from releasing the set of "high confidence" miRNAs for miRBase 21, because of a last-gasp bug. Those data are now available, tagged with the label "high confidence" on the entry pages, and for download on the FTP site. The total number of miRNAs labelled "high confidence" has increased [...]

miRBase 21 finally arrives  
By sam (June 26, 2014)  
Apologies for the longer-than-usual wait, miRBase 21 is now available on the website, and all data available for download on the FTP site. As usual, the release notes describe the major changes. Of particular note this time, the Genome Reference Consortium have released a new human genome assembly, GRCh38. We have therefore remapped the human [...]

## miRBase: the microRNA database



Target Search  
Target Mining  
Custom Prediction  
FuncMir Collection  
Data Download  
Statistics  
Help | FAQ  
Comments  
Citation | Policy

Choose one of the following search options:

Search by miRNA name  
Human  Go Clear

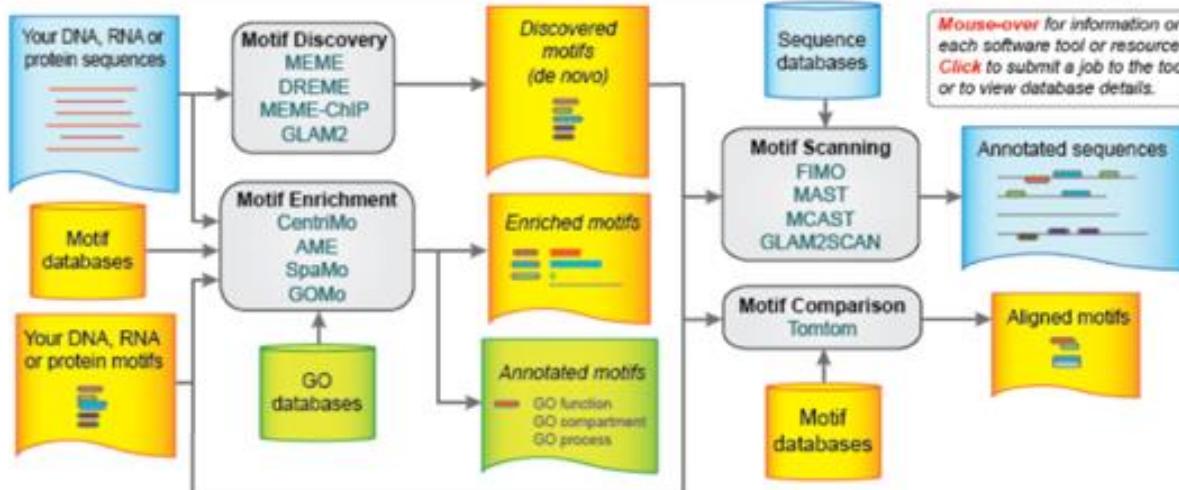
Search by gene target  
Human  Gene Symbol  Go Clear

miRDB is an online database for miRNA target prediction and functional annotations. A bioinformatics tool, MirTarget, which was developed by analyzing thousands of miRNA sequencing experiments. Common features associated with miRNA target binding have targets with machine learning methods. miRDB hosts predicted miRNA targets in five sections. As a recent update, users may provide their own sequences for customized target prediction, computational analyses and literature mining, functionally active miRNAs in humans as well as associated functional annotations, are presented in the FuncMir Collection in miRDB.

# MEME Suite: One-stop service for DNA motif analysis

## The MEME Suite

Motif-based sequence analysis tools



- Explore transcription factor's roles
- Find binding motifs near differentially expressed genes
- Detect new regulatory signals

# SWISS-MODEL: Homology modeling



## SWISS-MODEL

[Modelling](#)[Repository](#)

### Start a New Modelling Project

Target

Target MAAHKGAEEHHHKAAEHHQAAKHHHAAAEEHHHEKGEHEQAAHHADTAYAHHKHAEEHAAQAAKHDAAHHAPKPH 73

Sequence(s):

(Format must be FASTA,

Clustal,

plain string, or a valid

UniProtKB AC)

Add Hetero Target

Reset

Project Title:

Class example

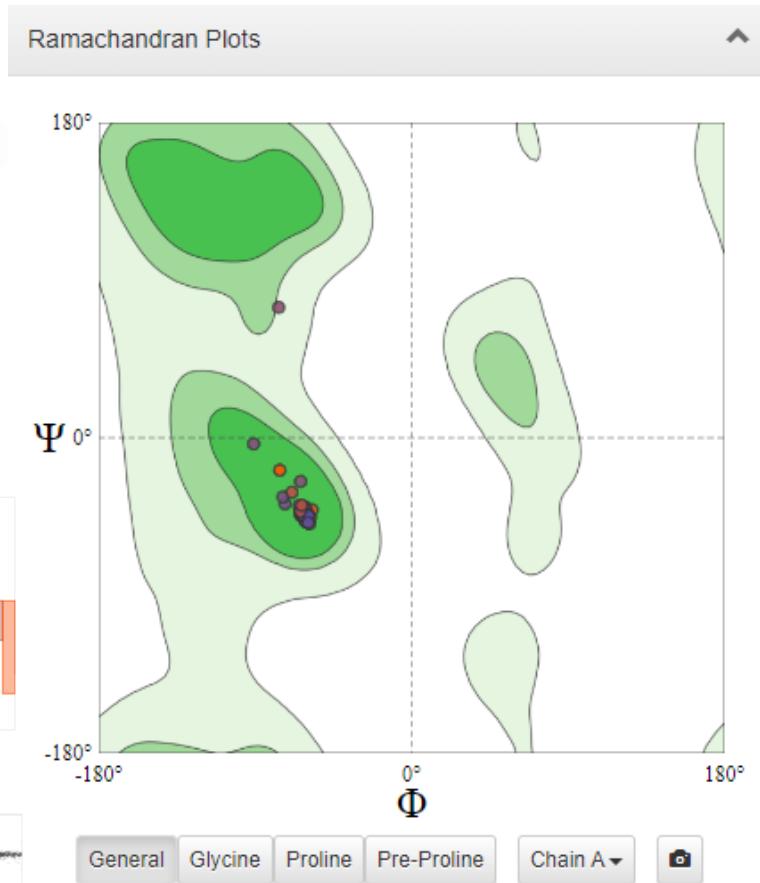
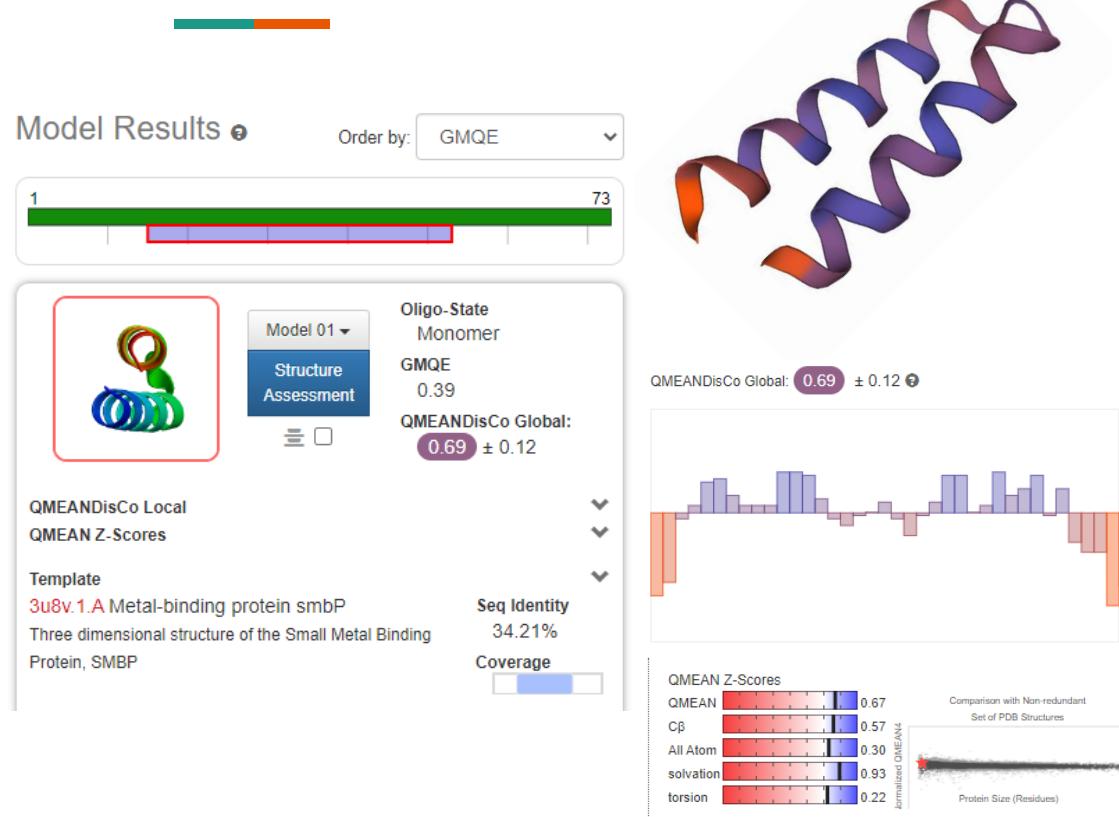
Email:

Optional

Search For Templates

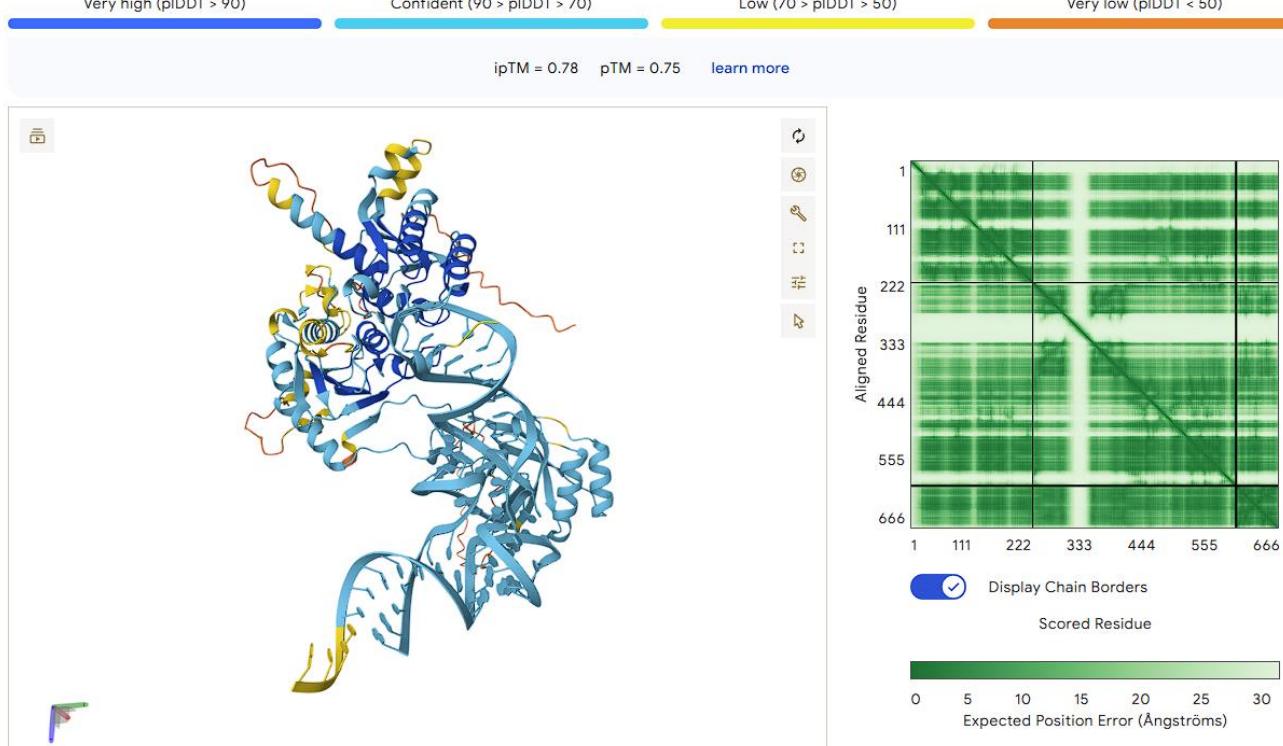
Build Model

# Model quality check



# AlphaFold v3: multi-mer structure prediction

# AlphaFold v3: multi-mer structure prediction



# Foldseek: protein structural similarity search

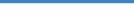
Foldseek Search  GITHUB SÖDING LAB S

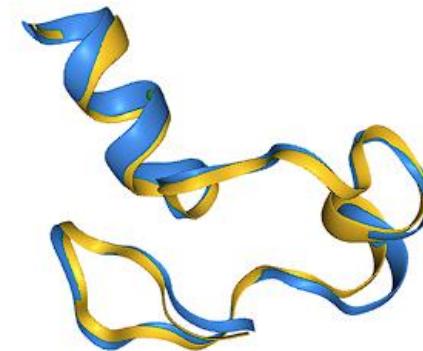
Results for job: jLU-XgYLj2cTLg\_bEG2z2krQqhqy\_FtTfZcNQg

TM-Score: 0.8494 RMSD: 0.73

ALL DATABASES BFVD (79) AFDB-PROTEOME (23) AFDB-SWISSPROT (19) AFDB50 (1000) BFMD (2) CATH50 (11) GMG

**BFVD 79 hits**

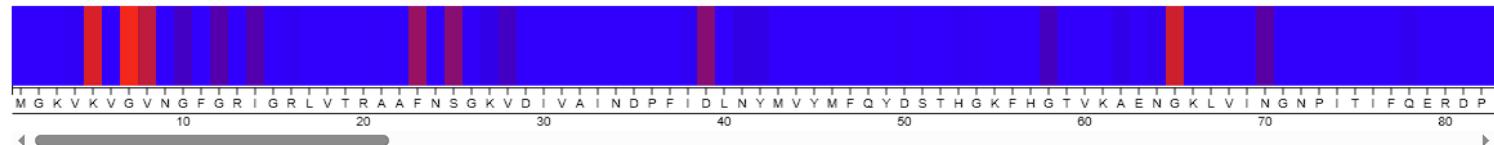
Target	Scientific Name	Prob.	Seq. Id.	E-Value	Position in query
<a href="#">AOA2H4JEY9</a>	<a href="#">uncultured Caudovirales phage</a>	0.99	35.2	2.95e-1	
<a href="#">AOA6J5L1I0</a>	<a href="#">uncultured Caudovirales phage</a>	0.84	44.8	1.06e+0	
<a href="#">AOA889IQ95</a>	<a href="#">Burkholderia phage BCSR129</a>	0.84	40	3.18e-1	
<a href="#">AOA6J7XTM1</a>	<a href="#">uncultured Caudovirales phage</a>	0.84	32.1	9.83e-1	
<a href="#">AOA7G8LLJ9</a>	<a href="#">unclassified Nymphadoravirus</a>	0.66	25	9.83e-1	
<a href="#">AOA2I6SCC4</a>	<a href="#">White spot syndrome virus</a>	0.54	25.8	3.04e+0	
<a href="#">AOA0A0RV67</a>	<a href="#">Bacillus phage Moonbeam</a>	0.54	21.2	1.79e+0	



- Discover proteins in nature with desire 3D structures

# DeepFRI: predicting protein functions

## Sequence View



Chain: seq

## Sequence-Based Molecular Function - GO Term Predictions

Name	Go Term	Score	Action
small molecule binding	GO:0036094	0.63	
organic cyclic compound binding	GO:0097159	0.63	
heterocyclic compound binding	GO:1901363	0.62	
nucleotide binding	GO:0000166	0.52	
nucleoside phosphate binding	GO:1901265	0.52	

# GEO2R: online differential expression for microarray

The screenshot shows the NCBI GEO Accession Display page for series GSE33113. The top navigation bar includes links for HOME, SEARCH, SITE MAP, GEO Publications, FAQ, MIAME, and Email GEO. The main content area displays the series information for GSE33113, including its status as public on Nov 05, 2011, and its title as AMC colon cancer AJCCII. It also lists the organism as Homo sapiens and the experiment type as Expression profiling by array. Below this, a 'Relations' section shows a BioProject entry for PRJNA156585. At the bottom left is a button labeled 'Analyze with GEO2R'. On the right side, there are links for 'Reviewer access' and 'Sign Out'.

Scope:  Format:  Amount:  GEO accession:

**Series GSE33113**

Status: Public on Nov 05, 2011  
Title: AMC colon cancer AJCCII  
Organism: [Homo sapiens](#)  
Experiment type: Expression profiling by array

**Relations**

BioProject: [PRJNA156585](#)

Analyze with GEO2R

Query DataSets for GSE33113

- Published microarray data on GEO can be analyzed online with GEO2R
- R code can also be exported

# GenePattern / Galaxy: online bioinformatics platforms



## Features

### Powerful genomics tools in a user-friendly interface



GenePattern provides hundreds of analytical tools for the analysis of gene expression (RNA-seq and microarray), sequence variation and copy number, proteomic, flow cytometry, and network analysis. These tools are all available through a Web interface with no programming experience required.

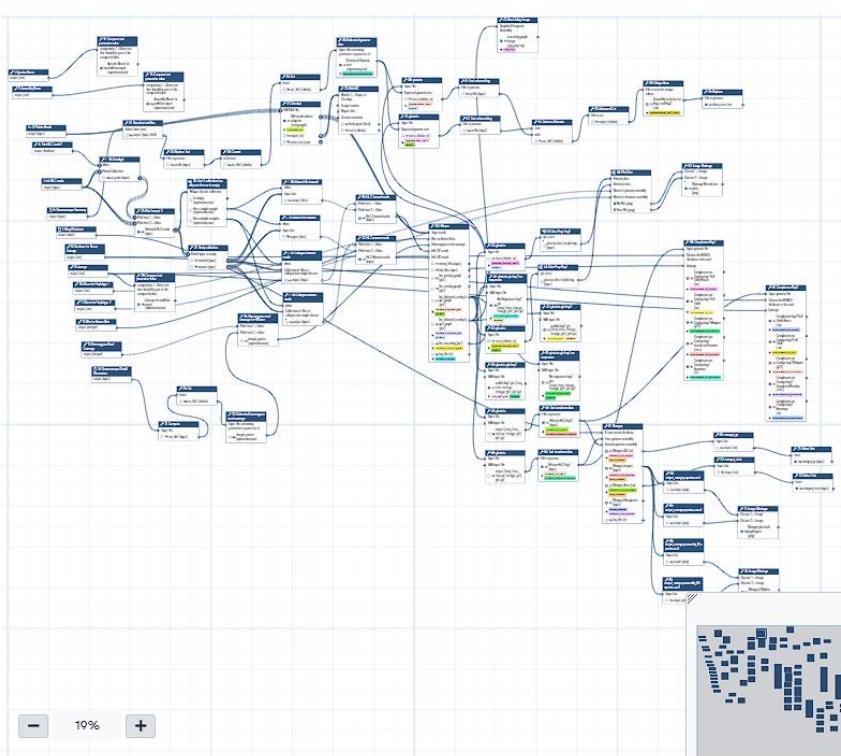
### GenePattern Notebook



The GenePattern Notebook environment extends the Jupyter Notebook system, allowing researchers to create documents that interleave formatted text, graphics and other multimedia, executable code, and GenePattern analyses, creating a single "research narrative" that puts scientific discussion and analyses in the same place.

A screenshot of the Galaxy web interface. The top navigation bar says 'Galaxy'. On the left, there's a sidebar with links: 'Upload', 'Tools' (which is currently selected and highlighted in grey), 'Workflows', 'Visualization', 'Histories', and 'Pages'. The main content area has a search bar at the top labeled 'All Tools' and 'search tools'. Below the search bar, there are several categories listed: 'Get Data', 'Send Data', 'Collection Operations', 'GENERAL TEXT TOOLS' (which is also highlighted in grey), 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Datamash', 'GENOMIC FILE MANIPULATION' (highlighted in grey), and 'FASTA/FASTQ'.

# Public bioinformatics workflow on Galaxy



## Genome Assembly from Hifi reads with HiC phasing ★ - VGP4 (release v0.4) 🖊



Assemble Genome using PacBio HiFi and HiC data from the same individual ...

Show more ▾

VGP

Reviewed

about 6 hours ago



### Creators



Galaxy



VGP



Delphine Lariviere

### Description

Assemble Genome using PacBio HiFi and HiC data from the same individual for phasing. Prerequisite: Run k-mer profiling workflow (VGP1). This workflow uses HiFiasm for contigging, and generates assembly statistics, BUSCO reports, Merqury histograms, and the genome assembly contigs in fasta and GFA format.

# WebGestalt / gProfiler: tools for functional enrichment



## WEB-based GEne SeT AnaLysis Toolkit

*Translating gene lists into biological insights...*

[Manual](#) | [API](#) | [Citation](#) | [User Forum](#) | [GOView](#) | [WebGestalt](#)

**Basic parameters**

**Method of Interest** Over-Representation Analysis Gene Set Enrichment Analysis  
Network Topology-based Analysis

**Organism of Interest** Homo sapiens  
Common Organisms: Homo sapiens Mus musculus Rattus norvegicus

**Functional Database** geneontology  
+ Biological Process noRedundant

gene Add List +

**Analyte Type** Gene/Protein Metabolite PTM Other

**Upload ID List** Click to upload Reset

**Input ID List** OR  
FBXL16 CAB39L

- Single-list
- Gene/GO:BP
- Metabolite/W
- Phosphosite/K
- Gene/FunMap



Query Upload query Uploa

Input is whitespace-separated list of c

X:1000:1000000  
rs17396340  
GO:0005005  
ENSG00000156103  
NLRP1

# Summary

---

- Knowledgebase and atlas explains biological entities and link you to related resources
- Repository is where you can obtain raw data to reanalyze
- Bioinformatics and AI prediction tools generate & strengthen hypothesis
- **Caution:** Online analysis can appear deceptively simple, please make sure to study the limitation of the tools and record the parameters used

# Any question?

---

- See you next time