



# 3000788 Intro to Comp Molec Biol

## Lecture 22: Online databases and tools

November 2, 2023



**Sira Sriswasdi, PhD**

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

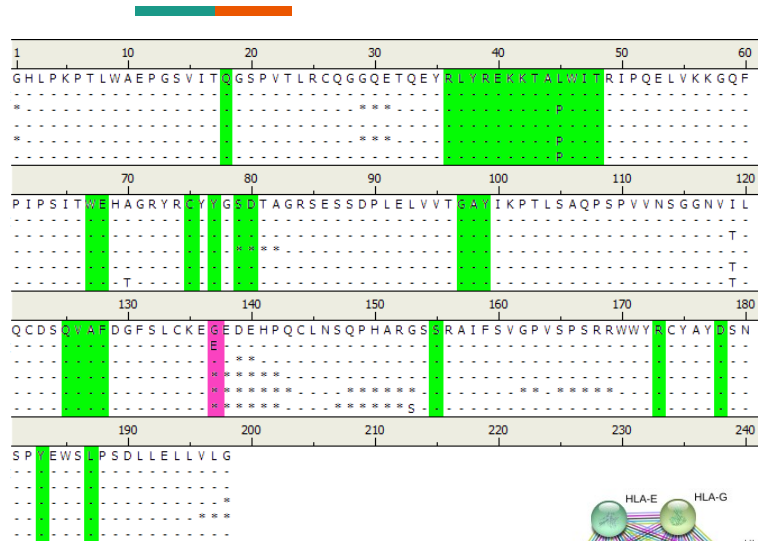
# Some motivations

- Quick answers to quick questions
  - Impact of a newly identified mutation?
  - Where does a TF bind on genome?
- Free hypothesis generation
  - What genes are consistently up-regulated in a certain disease?
- Combine data from multiple modalities
  - Omics
  - Clinical
  - Imaging

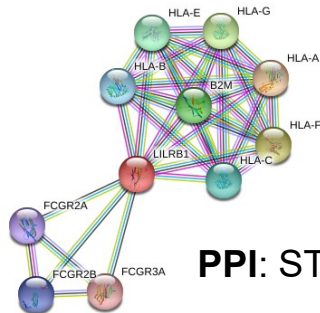


Image from [boston.lti.cs.cmu.edu](http://boston.lti.cs.cmu.edu)

# Example: G137E mutation on ILT2



**Sequence:** Look up variant on ClinVar and alignment to reported sequences

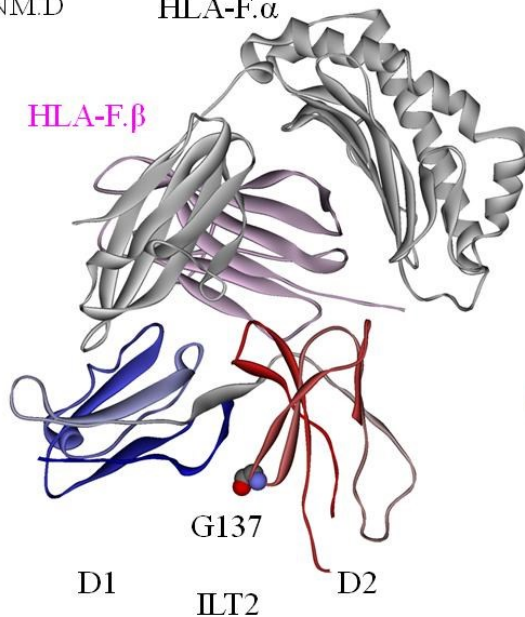


**PPI: STRING**

5KNM.D

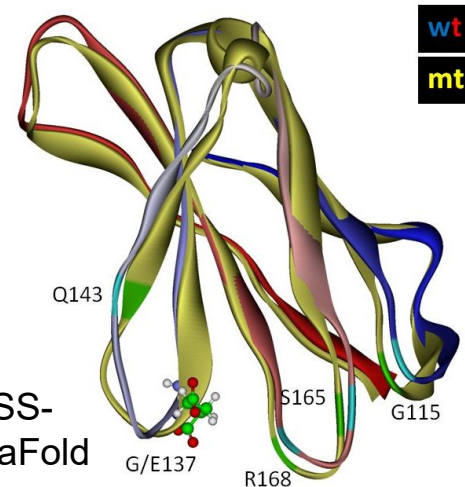
HLA-F.α

HLA-F.β

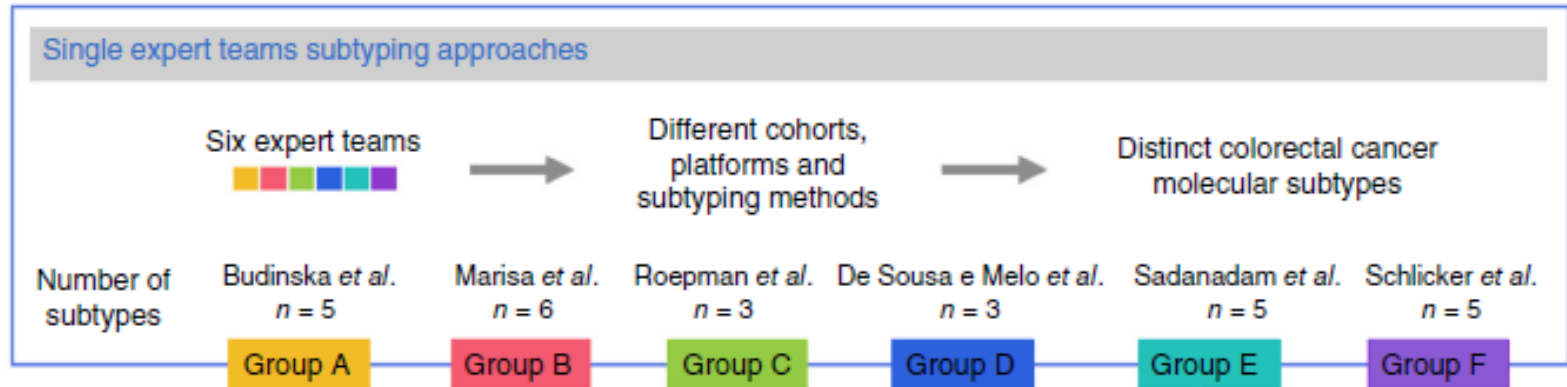


**Structure:** Highlight mutated residue on structure from PDB.

**Modeling:** SWISS-MODEL or AlphaFold



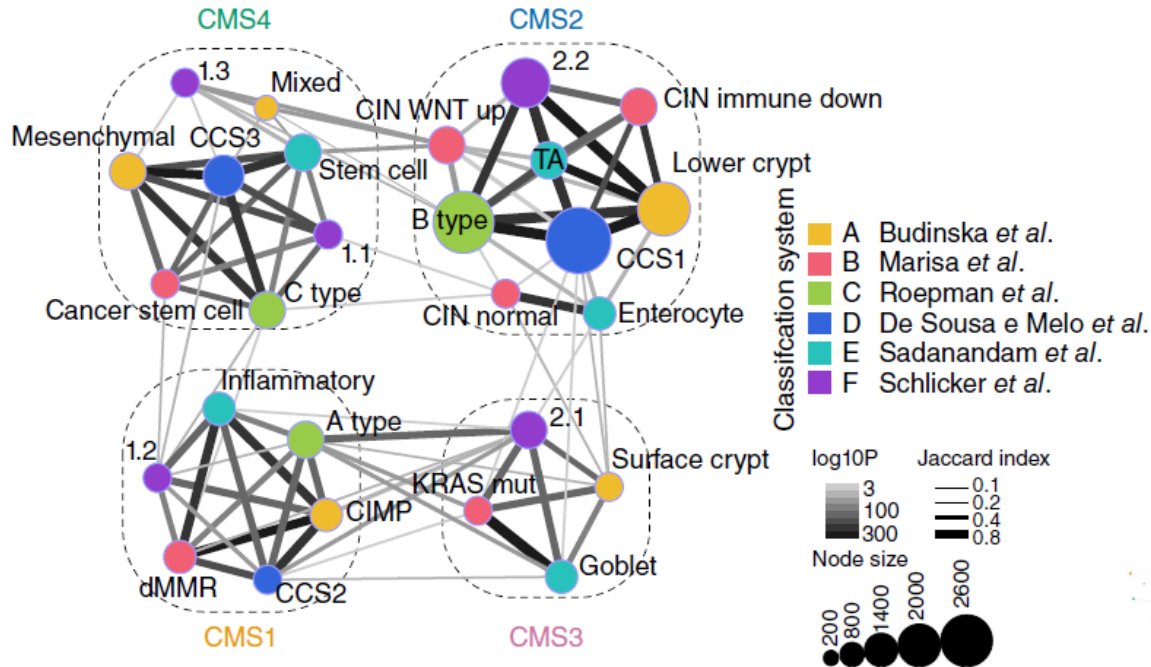
# Example: Consensus cancer subtyping



Guinney *et al.* Nat Medicine 21: 1350-1356 (2015)

- Different clinicians characterize cancer patients differently
- But there should be a common molecular basis!

# Example: Consensus cancer subtyping



- Compare clinician's decisions on a common group of patients
- Identify 4 major groups: Consensus Molecular Subtype (CMS)



# Knowledgebase

# Knowledgebase

- NCBI/GenBank
- Ensembl
- Uniprot
- GeneCard
- ENCODE
- Human Protein Atlas



## UniProtKB - O75473 (LGR5\_HUMAN)

### Display

[Help video](#)[BLAST](#)[Align](#)[Format](#)[Add to basket](#)[History](#)[Entry](#)[Publications](#)[Feature viewer](#)[Feature table](#)**Protein****Leucine-rich repeat-containing G-protein coupled receptor 5****Gene****LGR5****Organism***Homo sapiens (Human)***Status**Reviewed - Annotation score: ●●●●●● - Experimental evidence at protein level<sup>i</sup>

None

☒ Function☒ Names & Taxonomy☒ Subcellular location☒ Pathology & Biotech☒ PTM / Processing☒ Expression☒ Interaction☒ Structure☒ Family & Domains☒ Sequences (3)☒ Similar proteins☒ Cross-references

### 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: binding to R-spondins (RSPO1, RSPO2, RSPO3 or RSPO4), associates with phosphorylated LRP6 and frizzled receptors that are activated 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 [Source: GO\\_Central](#)
- G protein-coupled receptor activity [Source: ProtInc](#)
- protein-hormone receptor activity [Source: InterPro](#)
- transmembrane signaling receptor activity [Source: UniProtKB](#)

[Complete GO annotation on QuickGO ...](#)

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

# Accession IDs




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



# g:Profiler: Accession ID converter

**g:Profiler**


[News](#) [Archives](#) [Beta](#) [API](#) [R client](#) [FAQ](#) [Docs](#) [Contact](#) [Cite g:Profiler](#) [Services using g:P](#) [List of organisms](#) 

**g:GOST**  
Functional profiling


**g:Convert**  
Gene ID conversion

**g:Orth**  
Orthology search

**g:SNPense**  
SNP id to gene name


**Query** 


LGR5




Run query


Export to CSV

Show query URL 


Show short link 

**Options**


Organism: 



Homo sapiens (Human) 

Target namespace

ENSG 

Numeric IDs treated as



initial alias	converted alias 	name 	description	namespace
LGR5	ENSG00000139292	LGR5	leucine rich repeat containing G protein-coupled receptor 5 [Source:HGNC Symbol;Acc...	ENTREZGENE, HGNC, UNIPROT_G

# ENCODE: Search for TF binding sites

## Choose analysis





ENCODE4 v1.6.1 GRCh38 

## Filter files

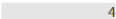


### File format

bigWig  6  
bigBed narrowPeak  4

### Output type

signal p-value  3  
fold change over control  3  
IDR thresholded peaks  3  
conservative IDR thresholded peaks  1


### Replicates

1, 2  4  
1  3  
2  3

Search for a gene


Enter gene name here

Sort by:  Replicates

 Output type

 Legend

 Reset coordinates

chr1 10.5346Mbp to 12.2343Mbp 

GRCh38

GENCODE V29

dbSNP (153)

representative DNase  
hypersensitivity sites

cCRE, all

ENCF696XCZ (rep 1, 2)

HepG2

NFIC

ChIP-seq

batch: ENCBS114ENC

conservative IDR thresholded  
peaks

- Processed ChIP-seq peaks for specific transcription factor

# GeneCard: 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>

G-Protein Coupled Receptor HG38 <sup>3 4</sup>

G-Protein Coupled Receptor 49 <sup>3 4</sup>

G-Protein Coupled Receptor 67 <sup>3 4</sup>

Orphan G Protein-Coupled Receptor HG38 <sup>3</sup>

G Protein-Coupled Receptor 49 <sup>2</sup>

GRP49 <sup>3</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

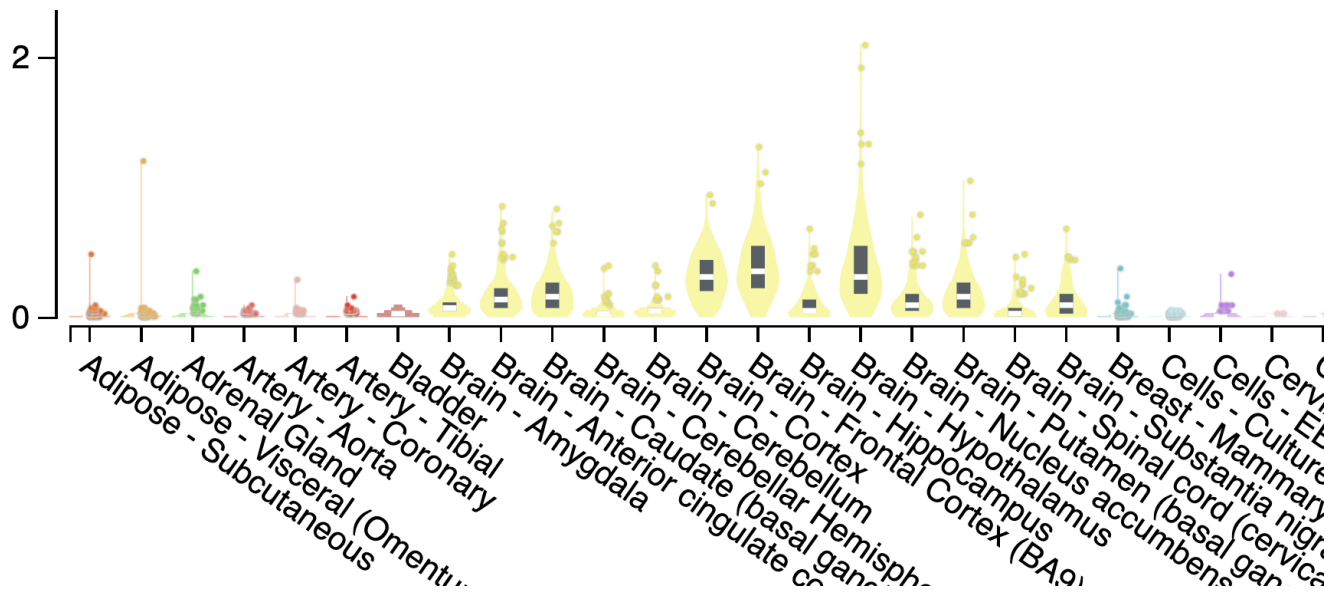
### Previous GeneCards Identifiers for LGR5 Gene

GC12P070121, GC12P071833, GC12P068883

Search aliases for LGR5 gene in PubMed and other databases

# GTEx: Tissue-specific expression

Gene Symbol	Gencode ID	Entrez Gene ID	Location	Gene Description
LIN28B	ENSG00000187772.7	<a href="#">389421</a>	chr6:104936616-105083332:+	lin-28 homolog B [Source:HGNC Symbol;Acc:HGNC:32207]



# Human Protein Atlas: Protein localization

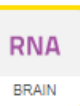
ACE2



SUMMARY



TISSUE



RNA

BRAIN



RNA

SINGLE CELL



RNA

TISSUE CELL

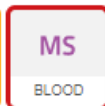


PATHOLOGY



RNA

IMMUNE



MS

BLOOD



RNA

SUBCELL



RNA

CELL LINE



METABOLIC

## 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

# JASPAR: 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

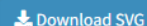
**Validation:** [12101405](#)

**Uniprot ID:** [P08651](#)

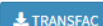
**Source:** [29126285](#)

**Comment:**

### Sequence logo

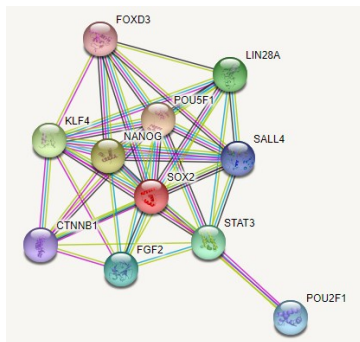


### Frequency matrix

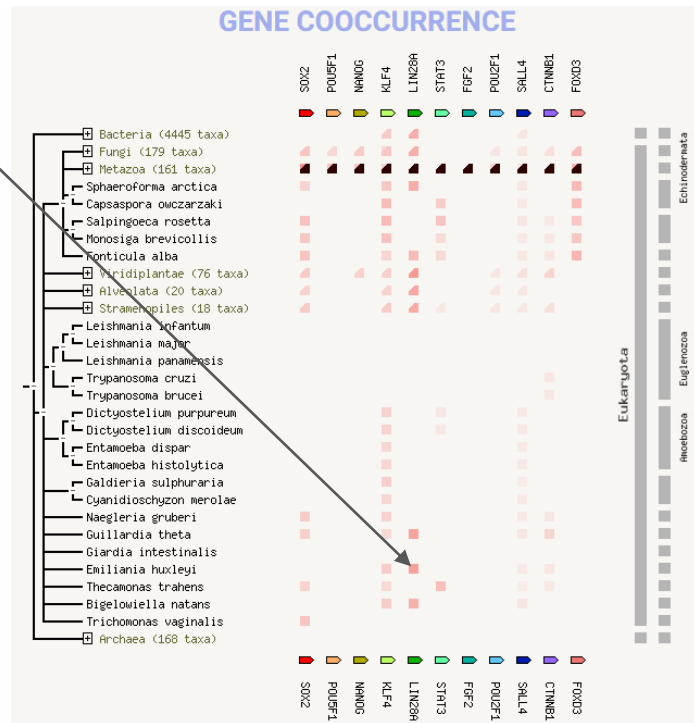


A [	2615	3896	412	154	352	310	223	264	10158	2626	334
C [	2647	2137	10542	303	333	129	121	10725	397	2951	271
G [	2478	2086	329	369	177	11087	10973	88	478	3186	276
T [	3867	3488	324	10781	10745	81	290	530	574	2844	279

# STRING: Protein-protein interaction



LIN28A	25	PEDAARAAD-EPQLLHGAGICKWFNVRMGFGFLSMTARAGVALDPPVDV P A R A D E P +G CKWF+V+ GFGF+ + + D+ PNTATRVADPEPA---PSGKCKWFDVQKGFGFIDVE-----NQEQDL
LIN28A	73	FVHQSKLHMEGFRSLKEGEAVEFTFKKSAKG--LESIRVTGPGGVFCIG FVHQ+ + +GFRSL EGEA+EF + AK L++I VTGPGG F G FVHQTDIAKAGFRSLAEGEALFQVSRDAKTNKLKAEVTPGGDFVEG
LIN28A	120	SERRP + R P
EOD22827	107	APREP



Viewers

Legend

Settings

Analysis

Table

More

Less

**Network**  
 Summary view: shows current interactions. Nodes can be moved; popups provide information on nodes & edges.

**Experiments**  
 Co-purification, co-crystallization, Yeast2Hybrid, Genetic Interactions, etc ... as imported from primary sources.

**Databases**  
 Known metabolic pathways, protein complexes, signal transduction pathways, etc ... from curated databases.

**Textmining**  
 Automated, unsupervised textmining - searching for proteins that are frequently mentioned together.


**Cooccurrence** currently showing  
 Gene families whose occurrence patterns across genomes show similarities.

**Coexpression**  
 Proteins whose genes are observed to be correlated in expression, across a large number of experiments.

**Neighborhood**  
 Groups of genes that are frequently observed in each other's genomic neighborhood.

**Fusion**  
 Genes that are sometimes fused into single open reading frames.

# Biomart: Gene/Protein data mapping download

 [BLAST/BLAT](#) | [VEP](#) | [Tools](#) | [BioMart](#) | [Downloads](#) | [Help & Docs](#) | [Blog](#) Search all species

[New](#) | [Count](#) | [Results](#)

[URL](#) | [XML](#) | [Perl](#) | [Help](#)

**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  
☐ Homologues (Max select 6 orthologues)

☐ Variant (Somatic)  
☐ Sequences

☒ GENE:

☐ EXTERNAL:

**GO**  
☐ GO term accession  
☐ GO term name  
☐ GO term definition

☐ GO term evidence code  
☐ GO domain

**GOSlim GOA**  
☐ GOSlim GOA Accession(s)

☐ GOSlim GOA Description

**External References (max 3)**  
☐ BioGRID Int. Information from external database sources.  
☐ Datasets ID  
☐ CCDS ID  
☐ ChEMBL ID  
☐ DataBase of Aberrant 3' Splice Sites name  
☐ DataBase of Aberrant 3' Splice Sites ID  
☐ DataBase of Aberrant 5' Splice Sites name  
☐ DataBase of Aberrant 5' Splice Sites ID

☐ STRING: Protein-protein interactions  
☐ NCBI gene (formerly Entrezgene) accession  
☐ NCBI gene (formerly Entrezgene) ID  
☒ PDB ID  
☐ Reactome ID  
☐ Reactome gene ID  
☐ Reactome transcript ID  
☐ RefSeq mRNA ID  
☐ RefSeq mRNA predicted ID



# Biomart: Gene/Protein data mapping download

[New](#) [Count](#) [Results](#) [★ URL](#) [XML](#) [Perl](#) [Help](#)

**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

**Dataset**  
[None Selected]

**Please restrict your query using criteria below**  
(If filter values are truncated in any lists, hover over the list item to see the full text)

▣ REGION:

☒ Chromosome/scaffold

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
..

# Biomart: Gene/Protein data mapping 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

## Dataset

[None Selected]

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="#">ENST000000673477</a>	<a href="#">ENST000000673477.1</a>	
<a href="#">ENSG00000160072</a>	<a href="#">ENSG00000160072.20</a>	<a href="#">ENST000000308647</a>	<a href="#">ENST000000308647.8</a>	
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST000000511072</a>	<a href="#">ENST000000511072.5</a>	
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST000000378391</a>	<a href="#">ENST000000378391.6</a>	<a href="#">2N1I</a>
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST000000378391</a>	<a href="#">ENST000000378391.6</a>	<a href="#">6BW4</a>
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST000000514189</a>	<a href="#">ENST000000514189.5</a>	
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST000000270722</a>	<a href="#">ENST000000270722.10</a>	<a href="#">2N1I</a>
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST000000270722</a>	<a href="#">ENST000000270722.10</a>	<a href="#">6BW4</a>
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST000000463591</a>	<a href="#">ENST000000463591.1</a>	
<a href="#">ENSG00000142611</a>	<a href="#">ENSG00000142611.17</a>	<a href="#">ENST000000509860</a>	<a href="#">ENST000000509860.1</a>	



# Data repository

# Data repository

## 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%      L=187,  $\sigma$ =95.9, 100%      L=44, 50%      L=123,  $\sigma$ =65.5, 50%

Legend

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout	Action
<a href="#">SRX000007</a>	SID2748	LS454	WGS	GENOMIC	RANDOM	PAIRED	<a href="#">BLAST</a>

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](#)



<< Full experiment listing

PXD027487

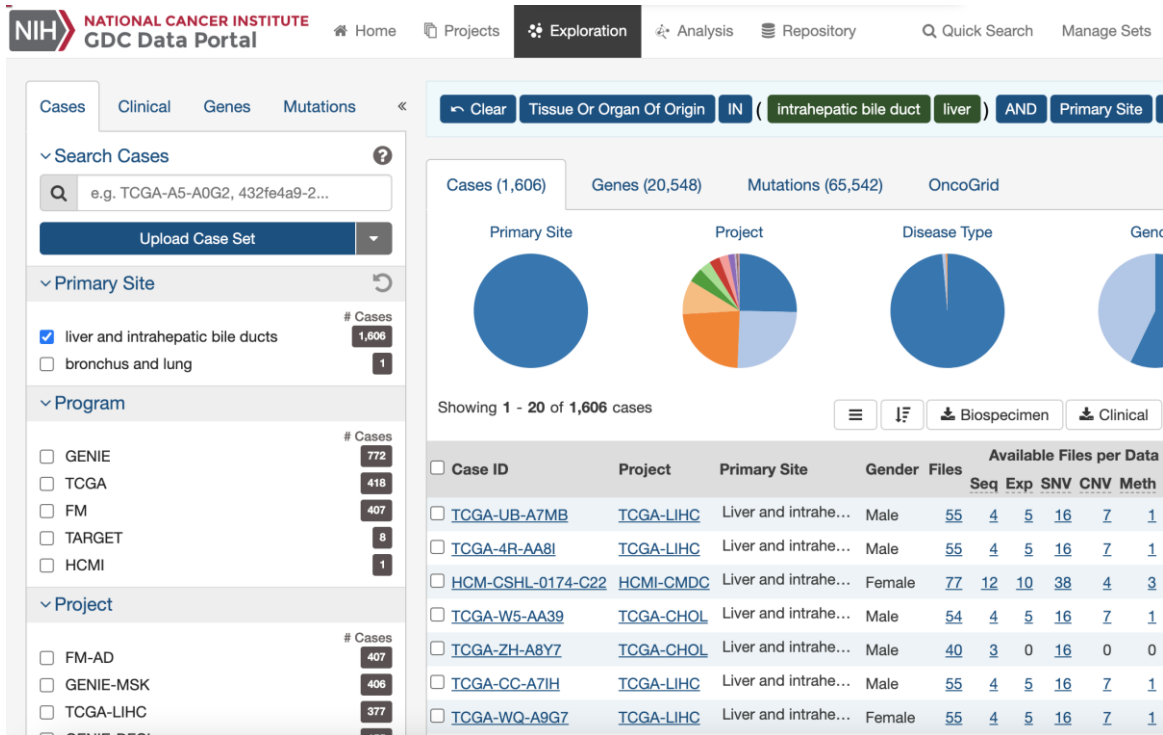
PXD027487 is an original dataset announced via ProteomeXchange.

### Dataset Summary

Title	Interactome of GIGYF2 and EIF4E2 with and without SMG11 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>
DigitalObjectIdentifier	
ReviewLevel	Peer-reviewed dataset
DatasetOrigin	Original dataset

- Reanalysis with a common pipeline

# Genomic Data Commons



- Cancer multi-omics
  - Exome
  - RNA-seq
  - Methylation
- Clinical and demographic data
- Some with histopathological images and radiographic images

# PDB: 3D structures

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

Refinements

Summary Gallery Compact -- Tabular Report -- ↓ Score Download Files ☒ All ☐ Selected

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

**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**

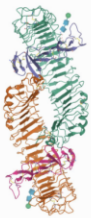

- ☐ X-RAY DIFFRACTION (32)
- ☐ SOLUTION NMR (1)

**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)

**4BSU**   Download File View File ☒

**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

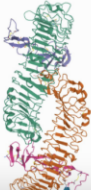

**Method** X-RAY DIFFRACTION 3.2 Å

**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**   Download File View File ☒

**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

# AlphaFold: Predicted 3D structures

## 3D viewer

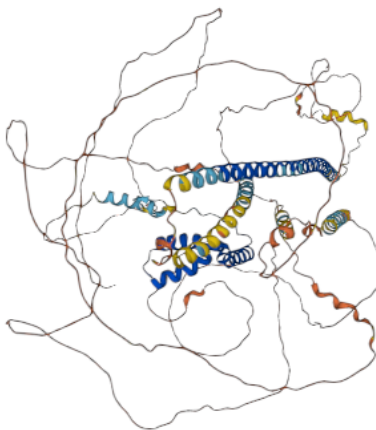
### Model Confidence:

-  Very high (pLDDT > 90)
-  Confident (90 > pLDDT > 70)
-  Low (70 > pLDDT > 50)
-  Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

Sequence of AF-A0A0G2JT... Chain 1: Transcriptio... A

```
1  MSSKQATSPFACTVDGEETMTQDLTSREKEEGSDQHPASHLPLHPIMHNKPHSEELPTLVSTIQDADWDSDVLSSQORMESENNKLCSLYSFRNTSTSPHKPDEGSREREIMNSVTFGTPERRK
11  21  31  41  51  61  71  81  91  101  111  121
2  GSLADVVDTLKQKKLEEMTRTEQEDSSCKEKLKSKDWKEKMERLNTSELLGEIKGTPESLAEKERQLSTMITQLISLREQLLAHDEQKKLAASQIEKQKQMDLARQQEQIARQQQQLLQQ
131  141  151  161  171  181  191  201  211  221  231  241
3  HKINLLQQQIQVQGHMFPPLMIPIFPHDQRTLAARAAQQGFLFPPGITYKPGDNYFVQFIPSTMAAARASGLSPQLQKQGHVSHPQINPRLKGISDRLGRNLDPEYHGGGHSYNHKQIEQLYAA
251  261  271  281  291  301  311  321  331  341  351  361  371
```



# AlphaFold: Predicted 3D structures

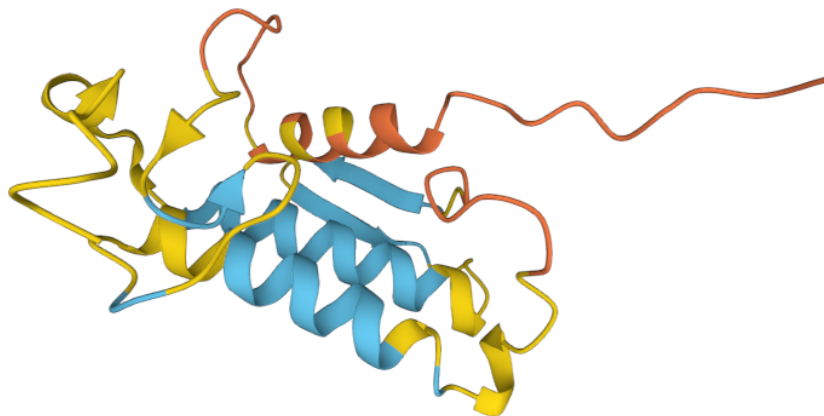
## 3D viewer

### Model Confidence:

- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

Sequence of AF-A0A1S5QI... Chain 1: Phospholipa... A  
1 11 21 31 41 51 61 71 81 91 101 111 121  
GAAAHSSIPRAVWQFRNMIKCTIPGSDPLKDYNNYGCYCGLGSGTFVDDLDRCQTHDHCYSQAKKLESCRFLLIDNPYNTNYSYSCSGSEITCSAKNNKCEEFCNCDREAAICFSKVPYNKEHK  
121  
NLDITGKFC





# Gene Expression Omnibus / ArrayExpress

▼ Study type

TOP 10

see all

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

- Search for omics datasets

# Google Dataset / FigShare

## Dataset Search

breast cancer metastasis



- 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)
- Imipramine Blue: A potent therapeutic regimen that suppresses breast cancer growth and metastasis

1,494,669 results found



Investigating novel mechanisms of metastasis in ...



Stratification of radiosensitive brain metastases based o...



Identification of Molecular Mediators of Endocrine ...

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

# 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.

<b>H: hallmark gene sets</b> (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. <a href="#">details</a>	<a href="#">Download GMT Files</a> <a href="#">Gene Symbols</a> <a href="#">NCBI (Entrez) Gene IDs</a>  <a href="#">JSON bundle</a>
<b>C1: positional gene sets</b> (browse 299 gene sets)	Gene sets corresponding to human chromosome cytogenetic bands. <a href="#">details</a>	<a href="#">Download GMT Files</a> <a href="#">Gene Symbols</a> <a href="#">NCBI (Entrez) Gene IDs</a>  <a href="#">JSON bundle</a>
<b>C2: curated gene sets</b> (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). <a href="#">details</a>	<a href="#">Download GMT Files</a> <a href="#">Gene Symbols</a> <a href="#">NCBI (Entrez) Gene IDs</a>  <a href="#">JSON bundle</a>

- Function-based / Disease-based / Publication-based / Genomic location-based
- For gene panel, enrichment test, etc.



# Analysis tools

# miRNA target



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 miRNA 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

**Search by gene target**  
Human  Gene Symbol

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

# SWISS-MODEL: Homology modeling



BIOZENTRUM

University of Basel  
The Center for Molecular Life Sciences

SWISS-MODEL

Modelling

Repository

## Start a New Modelling Project ⓘ

Target

⚙️ Target MDAHKGAEEHHHKAEEHHEQAQKHHHAAAEHHEKGEHEQAQHHADTAYAHKKHAEHHAQAQKHDAEEHHAPKPH 73

Sequence(s):

(Format must be FASTA,

Clustal,

plain string, or a valid

UniProtKB AC)

Add Hetero Target

↺ Reset

Project Title:

Class example

Email:

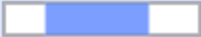
Optional

Search For Templates

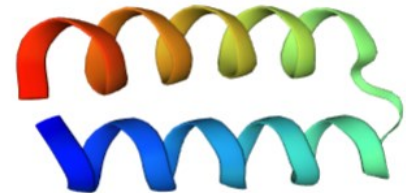
Build Model

# Template search

## Template Results

Templates		More ▼					
Sort	Coverage	GMQE	QSQE	Identity	Method	Oligo State	Ligands
✓	3u8v.1.A Metal-binding protein smbP <i>Three dimensional structure of the Small Metal Binding Protein, SMBP</i>						
▼		0.39	-	34.21	X-ray, 1.9Å	monomer ✓	1 x NI <sup>Ⓢ</sup>

- Identify known 3D structures with similar amino acid sequences



# Model quality check

## Model Results

Order by: GMQE



Model 01  
Structure  
Assessment

Oligo-State  
Monomer

GMQE  
0.39

QMEANDisCo Global:  
0.69 ± 0.12

QMEANDisCo Local  
QMEAN Z-Scores

Template

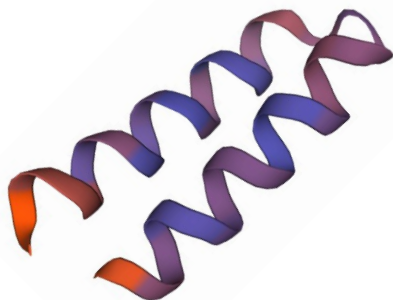
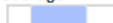
3u8v.1.A Metal-binding protein smbP

Three dimensional structure of the Small Metal Binding  
Protein, SMBP

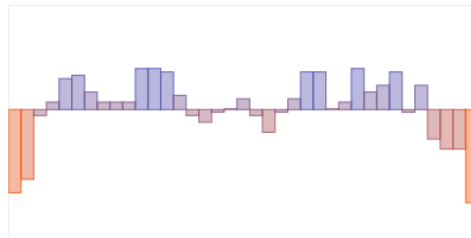
Seq Identity

34.21%

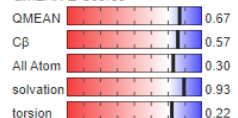
Coverage



QMEANDisCo Global: 0.69 ± 0.12



QMEAN Z-Scores

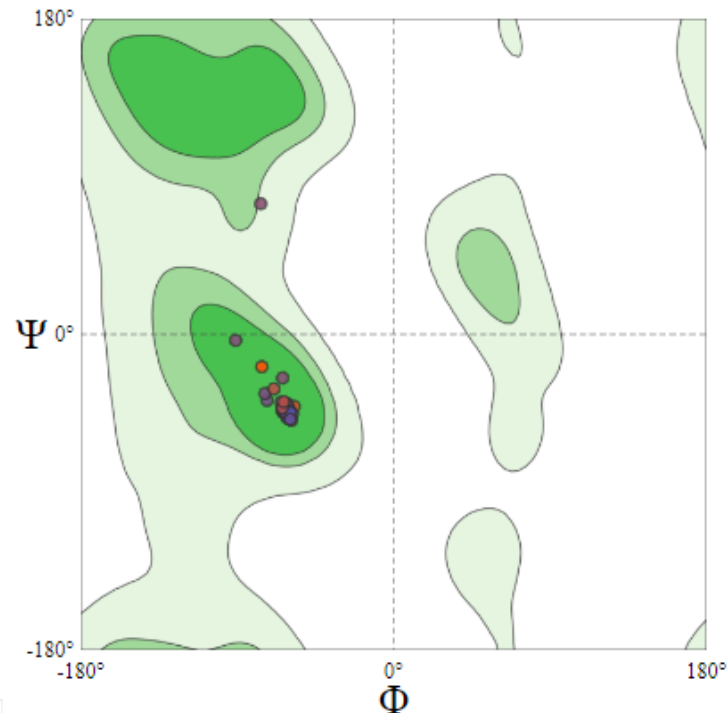


Comparison with Non-redundant  
Set of PDB Structures

normalized QMEAN

Protein Size (Residues)

## Ramachandran Plots



General

Glycine

Proline

Pre-Proline

Chain A



- Overview



# AI-assisted protein folding (ColabFold)

## 3. Enter the amino acid sequence(s) to fold

Enter the amino acid sequence(s) to fold:

- If you enter only a single sequence, the monomer model will be used.
- If you enter multiple sequences, the multimer model will be used.

sequence\_1: " MAAHKGAEH HHKAAEHHEQA AKHHHAAA EHKGEHEQA

sequence\_2: " Insert text here

<https://github.com/sokrypton/ColabFold>

## 5. Run AlphaFold and download prediction



Once this cell has been executed, a zip-archive w

In case you are having issues with the relaxation violations.

run\_relax: ☒

Relaxation is faster with a GPU, but we have four reverting to using without GPU.

relax\_use\_gpu: ☐

[Show code](#)

- Accessible, simplified Python interface

# ColabFold on-screen outputs

Getting MSA for sequence 1

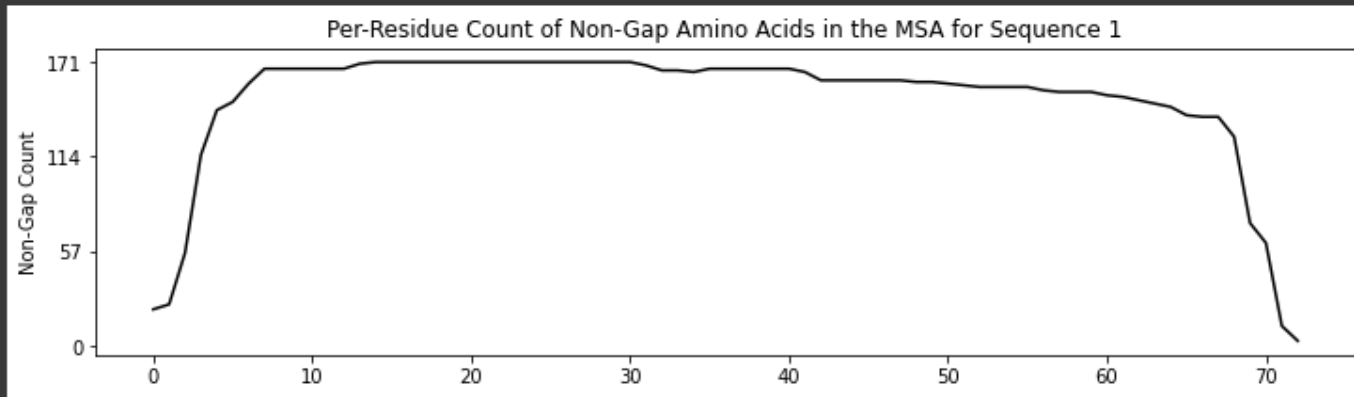
Searching mgnify: 100%  147/147 [elapsed: 25:21 remaining: 00:00]

58 unique sequences found in uniref90 for sequence 1

110 unique sequences found in smallbfd for sequence 1

9 unique sequences found in mgnify for sequence 1

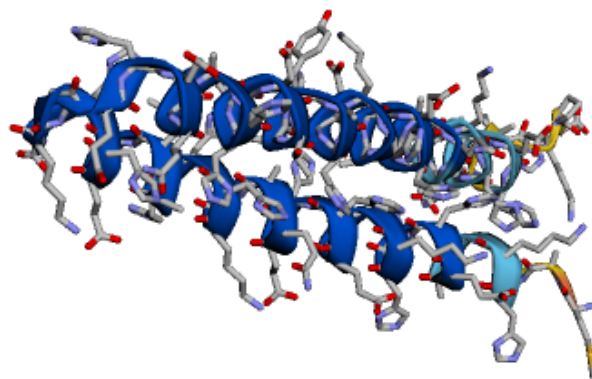
173 unique sequences found in total for sequence 1



# AlphaFold output in PDB format

selected\_prediction.pdb - Notepad

File	Edit	Format	View	Help							
ATOM	1	N	MET	A	1	21.997	-7.681	-15.856	1.00	47.20	N
ATOM	2	H	MET	A	1	22.579	-7.136	-15.236	1.00	47.20	H
ATOM	3	H2	MET	A	1	22.491	-8.529	-16.097	1.00	47.20	H
ATOM	4	H3	MET	A	1	21.827	-7.147	-16.697	1.00	47.20	H
ATOM	5	CA	MET	A	1	20.726	-8.011	-15.175	1.00	47.20	C
ATOM	6	HA	MET	A	1	20.096	-8.607	-15.835	1.00	47.20	H
ATOM	7	C	MET	A	1	19.955	-6.731	-14.844	1.00	47.20	C
ATOM	8	CB	MET	A	1	20.994	-8.837	-13.903	1.00	47.20	C
ATOM	9	HB2	MET	A	1	21.674	-8.300	-13.243	1.00	47.20	H
ATOM	10	HB3	MET	A	1	20.050	-8.988	-13.379	1.00	47.20	H
ATOM	11	O	MET	A	1	19.729	-6.438	-13.681	1.00	47.20	O
ATOM	12	CG	MET	A	1	21.574	-10.218	-14.209	1.00	47.20	C
ATOM	13	HG2	MET	A	1	20.933	-10.709	-14.941	1.00	47.20	H
ATOM	14	HG3	MET	A	1	22.577	-10.123	-14.623	1.00	47.20	H
ATOM	15	SD	MET	A	1	21.645	-11.258	-12.738	1.00	47.20	S
ATOM	16	CE	MET	A	1	21.921	-12.878	-13.501	1.00	47.20	C
ATOM	17	HE1	MET	A	1	22.851	-12.867	-14.069	1.00	47.20	H
ATOM	18	HE2	MET	A	1	21.089	-13.121	-14.162	1.00	47.20	H
ATOM	19	HE3	MET	A	1	21.987	-13.637	-12.721	1.00	47.20	H
ATOM	20	N	ALA	A	2	19.598	-5.931	-15.854	1.00	54.34	N
ATOM	21	H	ALA	A	2	19.735	-6.197	-16.818	1.00	54.34	H
ATOM	22	CA	ALA	A	2	18.910	-4.655	-15.629	1.00	54.34	C
ATOM	23	HA	ALA	A	2	19.241	-4.225	-14.683	1.00	54.34	H
ATOM	24	C	ALA	A	2	17.385	-4.836	-15.515	1.00	54.34	C
ATOM	25	CB	ALA	A	2	19.332	-3.685	-16.734	1.00	54.34	C
ATOM	26	HB1	ALA	A	2	18.880	-2.710	-16.551	1.00	54.34	H
ATOM	27	HB2	ALA	A	2	18.993	-4.039	-17.707	1.00	54.34	H
ATOM	28	HB3	ALA	A	2	20.415	-3.561	-16.744	1.00	54.34	H
ATOM	29	O	ALA	A	2	16.797	-4.300	-14.587	1.00	54.34	O
ATOM	30	N	ALA	A	3	16.797	-5.715	-16.340	1.00	57.44	N
ATOM	31	H	ALA	A	3	17.335	-6.061	-17.121	1.00	57.44	H
ATOM	32	CA	ALA	A	3	15.353	-5.979	-16.393	1.00	57.44	C
ATOM	33	HA	ALA	A	3	14.852	-5.081	-16.755	1.00	57.44	H
ATOM	34	C	ALA	A	3	14.700	-6.342	-15.044	1.00	57.44	C



## Model Confidence

- Very low (pLDDT < 50)
- Low (70 > pLDDT > 50)
- Confident (90 > pLDDT > 70)
- Very high (pLDDT > 90)

# GenePattern / Galaxy



## 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.

#### Blog > GP updates

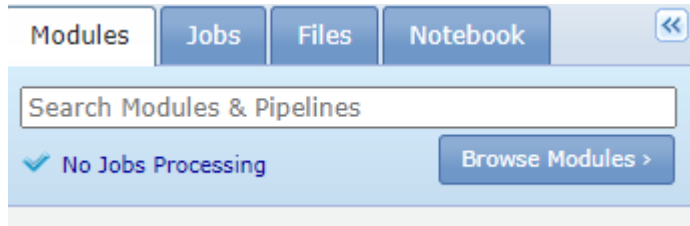
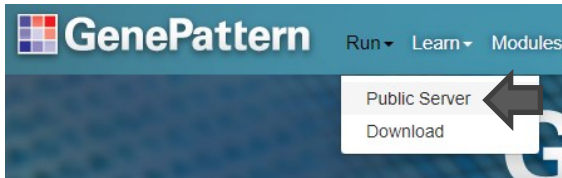
- GenePattern Coverage and Support December 23, 2021 - January 2, 2022
- End of Support for GParc
- End of support for modules using the deprecated GenePattern patch mechanism.
- End of support for Cufflinks suite of modules
- End of support for MAGeCK

[view more >](#)

#### Papers > Related Publications

- Comparability and reproducibility of biomedical data
- TopCluster: a multiple gene list feature analyzer for

# Content



## Browse Modules & Pipelines

clustering

cnv analysis

conumee

data format conversion

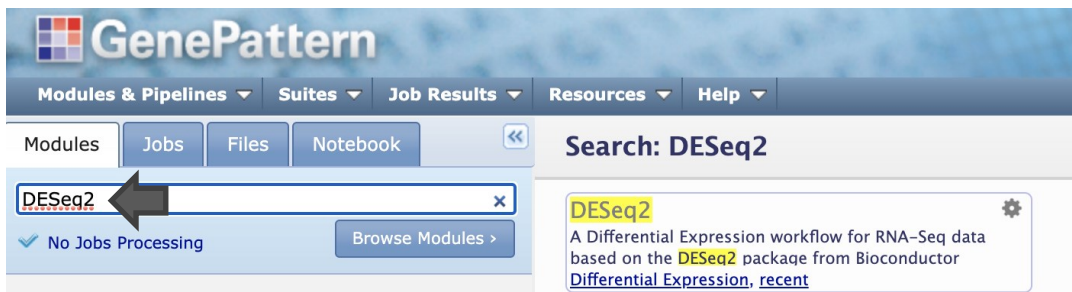
differential expression

dimension reduction

flow cytometry

- Modules = standard bioinformatics tools
- Notebook = Python environment (similar to Google Colab) with specific bioinformatics library

# DESeq2 on GenePattern



- Input expression data (RNA-seq read count) and sample label
- .gct and .cls are text files

**input file\***

▼Hide Files...(Selected 1 files)

htseq\_count.gct

A GCT file containing raw RNA-Seq counts, such as is produced by MergeHTSeqCounts

**cls file\***

▼Hide Files...(Selected 1 files)

sample\_info.cls

A categorical CLS file specifying the phenotype classes for the samples in the GCT file. This should contain exactly two classes with the control specified first.

**confounding variable cls file**

Upload File... Add Path or URL... Drag Files Here

2GB file upload limit using the Upload File... button. For files > 2GB upload from the Files tab.


A categorical CLS file specifying an additional confounding variable, mapped to the input file samples. Use this for a two-factor comparison.

#1.2					
60488	12				
Name	Description	sample1	sample2	sample3	sample4
ENSG00000242268.2	ENSG00000242268.2	2	10	0	2
ENSG00000270112.3	ENSG00000270112.3	8	6	0	0
ENSG00000167578.15	ENSG00000167578.15	102	633	468	1200
ENSG00000273842.1	ENSG00000273842.1	0	0	0	0
ENSG00000078237.5	ENSG00000078237.5	370	364	1220	692

```
12 2 1
# alive dead
1 1 0 0 1 0 0 0 1 0 1 1
```


# DESeq2 output


## 471244. DESeq2


Source:  GenePattern production (new)

submitted: Oct 25 01:06:41 AM, completed: Oct 25 01:10:38 AM, size: 30 MB

[Show details](#)

 **Comments (0)**

 **Tags (0)**

 **input.file:** [expression.gct](#)

 **cls.file:** [label.cls](#)

 [class\\_example.Positive.vs.Negative.DESeq2\\_results\\_report.txt](#) (670.0 KB) (Last modified: 2022-10-25 01:10:26.0)

 [class\\_example.Positive.vs.Negative.QC.DispEsts.png](#) (78.0 KB) (Last modified: 2022-10-25 01:10:26.0)

 [class\\_example.Positive.vs.Negative.QC.MAplot.png](#) (9.0 KB) (Last modified: 2022-10-25 01:10:26.0)


 [class\\_example.Positive.vs.Negative.mean\\_values\\_by\\_class.txt](#) (230.0 KB) (Last modified: 2022-10-25 01:10:26.0)


 [class\\_example.Positive.vs.Negative.normalized\\_counts.gct](#) (14.5 MB) (Last modified: 2022-10-25 01:10:26.0)

 [class\\_example.Positive.vs.Negative.normalized\\_counts.txt](#) (14.5 MB) (Last modified: 2022-10-25 01:10:26.0)

 [class\\_example.Positive.vs.Negative....0\\_downregulated\\_genes\\_report.txt](#) (3.0 KB) (Last modified: 2022-10-25 01:10:26.0)

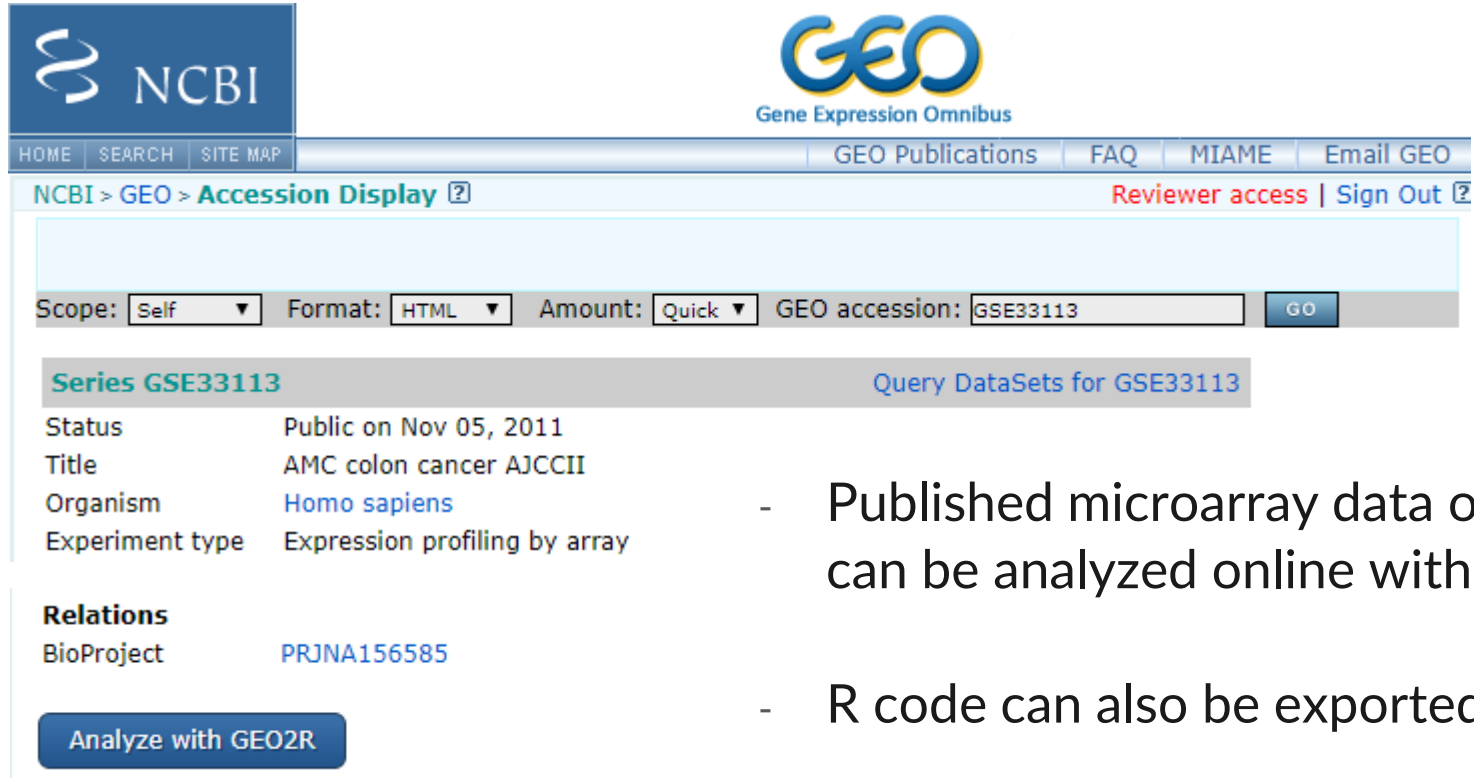
 [class\\_example.Positive.vs.Negative.top20\\_upregulated\\_genes\\_report.txt](#) (3.0 KB) (Last modified: 2022-10-25 01:10:26.0)

 [stderr.txt](#) (1.0 KB) (Last modified: Tue Oct 25 01:10:37 UTC 2022)

 [stdout.txt](#) (4.0 KB) (Last modified: Tue Oct 25 01:10:37 UTC 2022)

 [gp\\_execution\\_log.txt](#) (1.0 KB) (Last modified: Tue Oct 25 01:10:38 UTC 2022)

# GEO2R: Differential expression



The screenshot displays the NCBI GEO2R web interface. At the top, the NCBI logo is on the left and the GEO logo (Gene Expression Omnibus) is on the right. Below the logos is a navigation bar with links: HOME, SEARCH, SITE MAP, GEO Publications, FAQ, MIAME, and Email GEO. The main header shows the breadcrumb path: NCBI > GEO > Accession Display. To the right of this path are links for Reviewer access and Sign Out. Below the header is a search bar with the following fields: Scope (set to Self), Format (set to HTML), Amount (set to Quick), GEO accession (set to GSE33113), and a GO button. Below the search bar, the series GSE33113 is highlighted, with a link to Query DataSets for GSE33113. The series details are listed below: Status (Public on Nov 05, 2011), Title (AMC colon cancer AJCCII), Organism (Homo sapiens), and Experiment type (Expression profiling by array). Below the series details is a section for Relations, showing a BioProject link to PRJNA156585. At the bottom left, there is a button labeled Analyze with GEO2R.

NCBI > GEO > Accession Display [?](#) [Reviewer access](#) | [Sign Out](#) [?](#)

Scope:  Format:  Amount:  GEO accession:

**Series GSE33113** [Query DataSets for GSE33113](#)

Status	Public on Nov 05, 2011
Title	AMC colon cancer AJCCII
Organism	<a href="#">Homo sapiens</a>
Experiment type	Expression profiling by array

**Relations**

BioProject	<a href="#">PRJNA156585</a>
------------	-----------------------------

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



# Demo time!



- ColabFold interface
- STRING
- GEO2R
- Biomart

# Any question?



- See you on Nov 6
- Look out for instruction on how to set up Python on the course website