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

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?

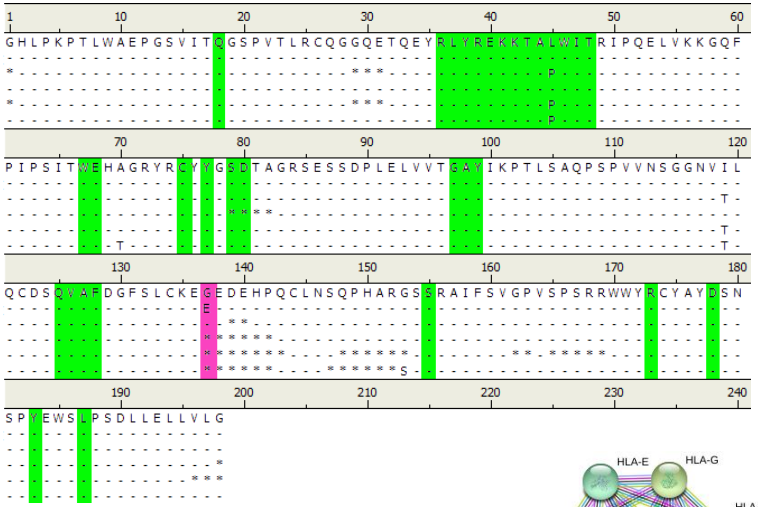
Why should we use online resources?

- 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?

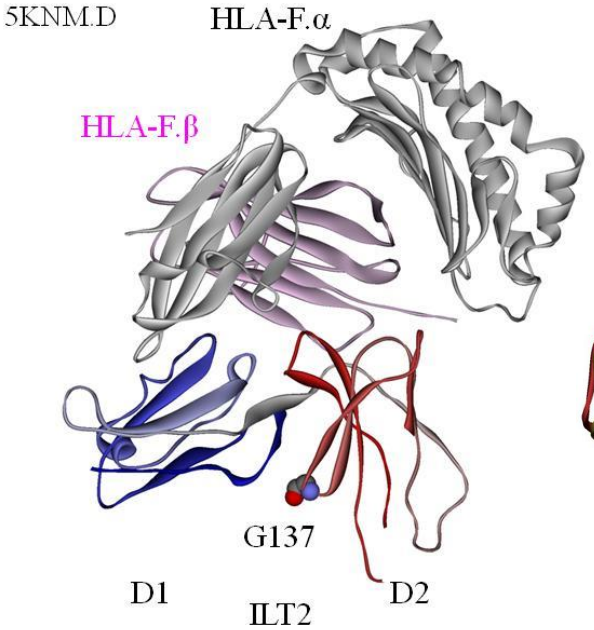
Why should we use online resources?

- 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

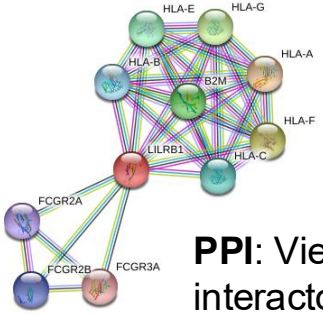
Example 1: Intepreting the effects of a mutation



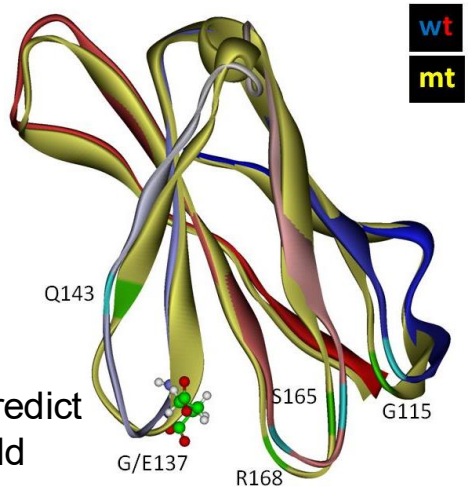
Sequence: Look up variant information on ClinVar and compared to commonly found mutations



Structure: Highlight mutated residue on structure from PDB.



PPI: View potential interactors in STRING



Structure: Predict with AlphaFold

Example 1: Intepreting 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

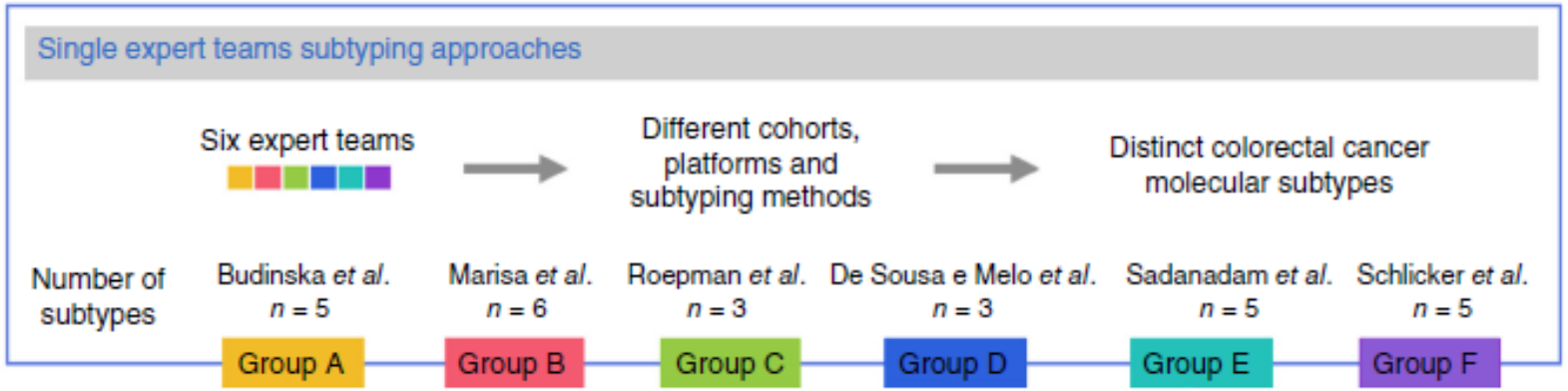
- Available for download (known variant only)
- Available as API from Google cloud

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

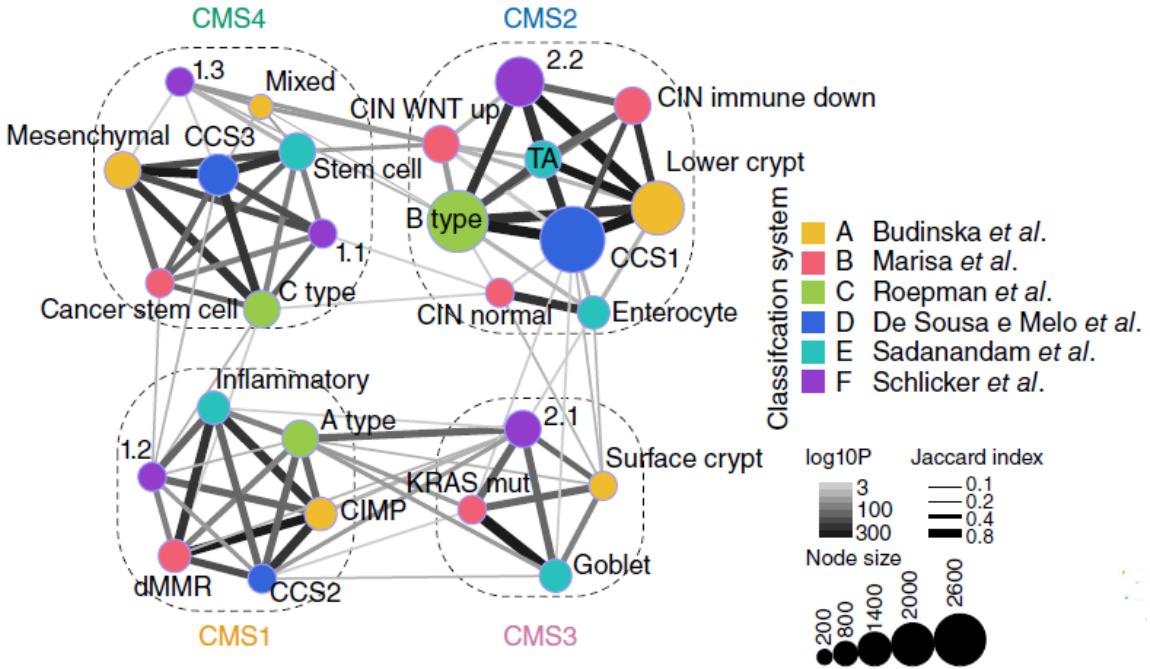
Example 2: Consensus cancer molecular subtyping



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

- 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



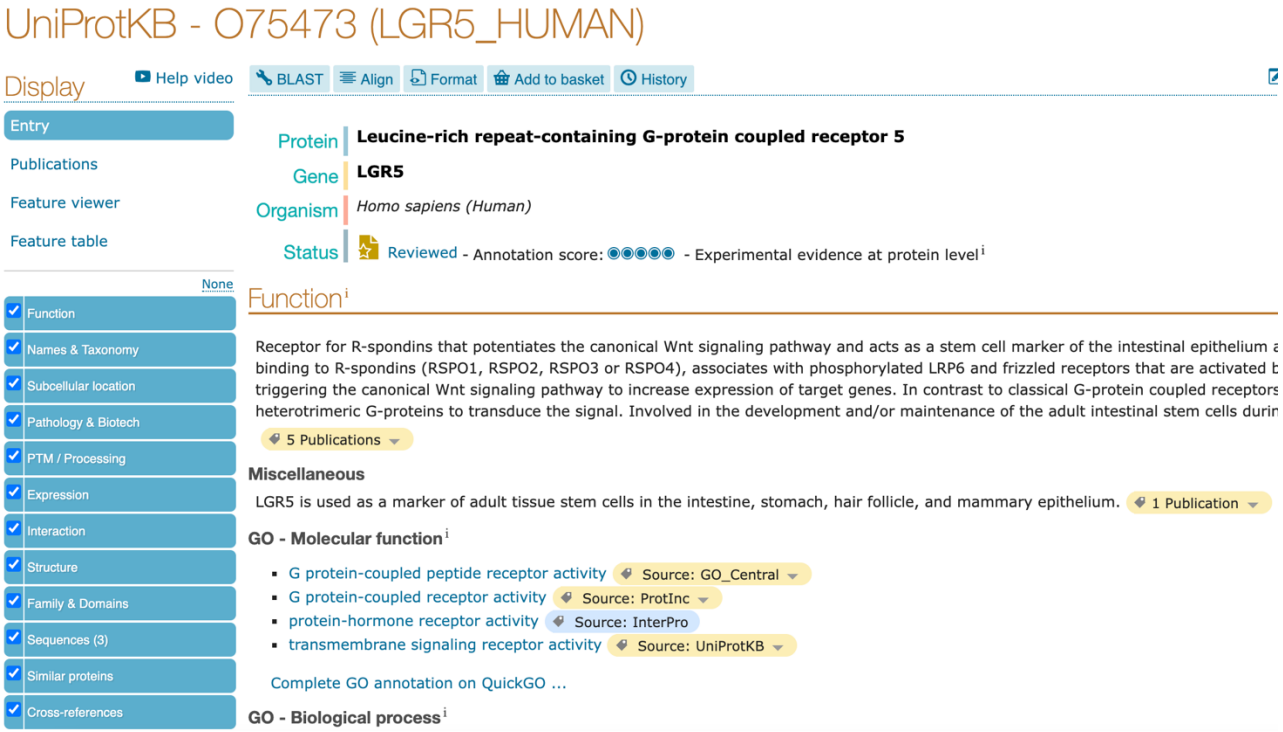
- Apply 6 subtyping schemes on a group of 4,000 patients
- Identify subtypes often assigned to the same patients
  - **1.1** / **CCS3** / **Stem cell** / **Mesenchymal** / **C type**
  - Consensus!

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

Knowledgebases

Classical, foundational databases

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



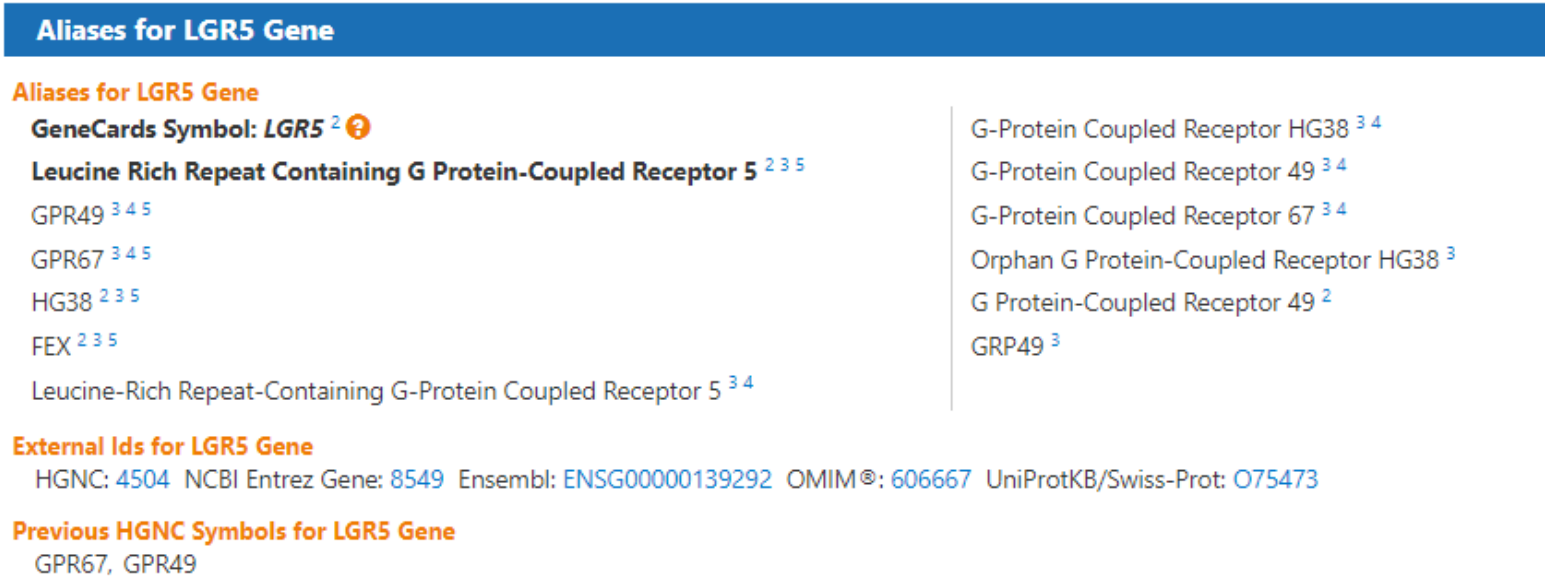
Some common accession systems

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* [2]

Leucine Rich Repeat Containing G Protein-Coupled Receptor 5 [2 3 5]

GPR49 [3 4 5]

GPR67 [3 4 5]

HG38 [2 3 5]

FEX [2 3 5]

Leucine-Rich Repeat-Containing G-Protein Coupled Receptor 5 [3 4]

G-Protein Coupled Receptor HG38 [3 4]

G-Protein Coupled Receptor 49 [3 4]

G-Protein Coupled Receptor 67 [3 4]

Orphan G Protein-Coupled Receptor HG38 [3]

G Protein-Coupled Receptor 49 [2]

GRP49 [3]

External Ids for LGR5 Gene

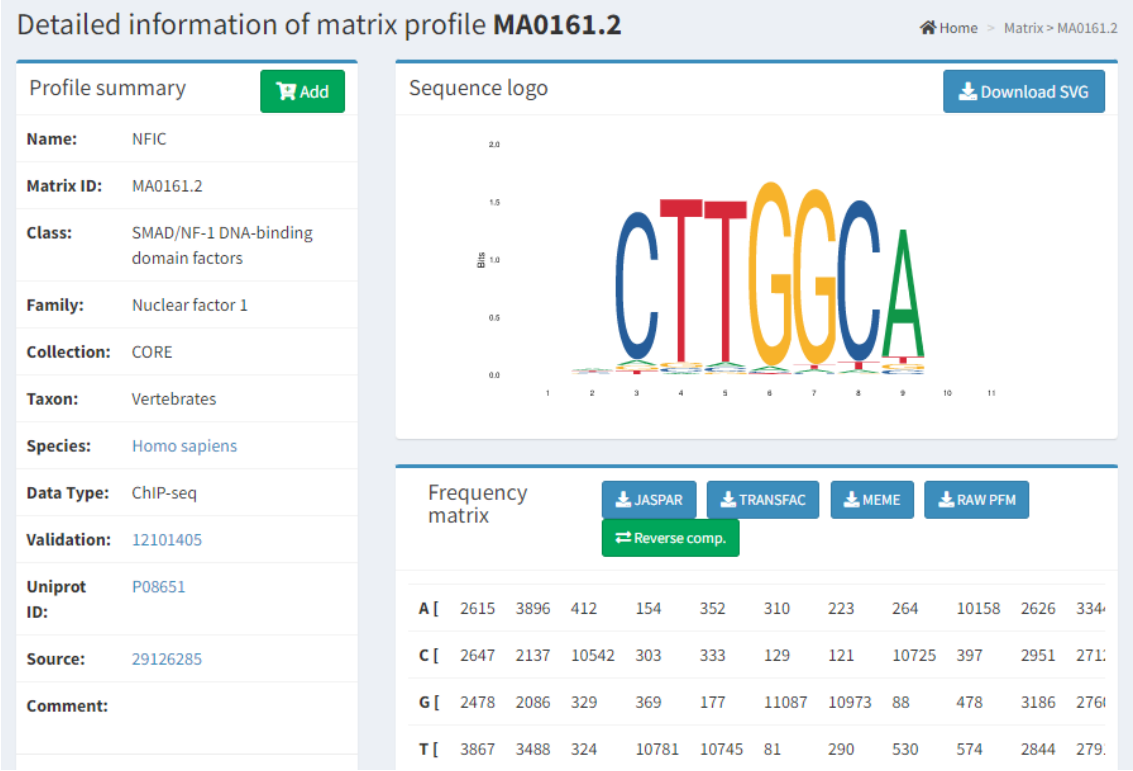
HGNC: 4504 NCBI Entrez Gene: 8549 Ensembl: ENSG00000139292 OMIM®: 606667 UniProtKB/Swiss-Prot: O75473

Previous HGNC Symbols for LGR5 Gene

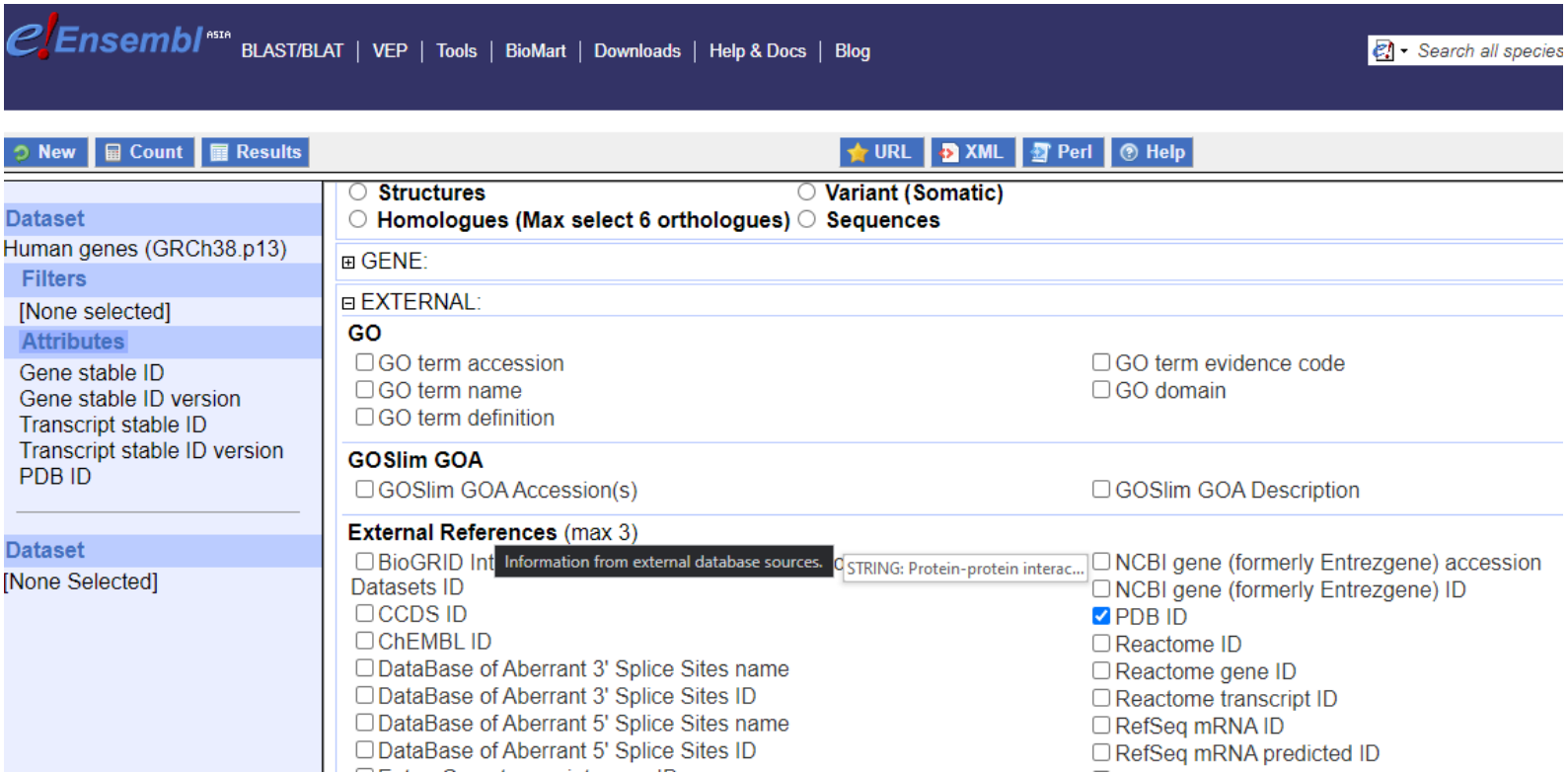
GPR67, GPR49

- 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



Biomart: bulk gene/protein annotation download



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

Dataset

[None Selected]

Export all results to File TSV Unique results only Go

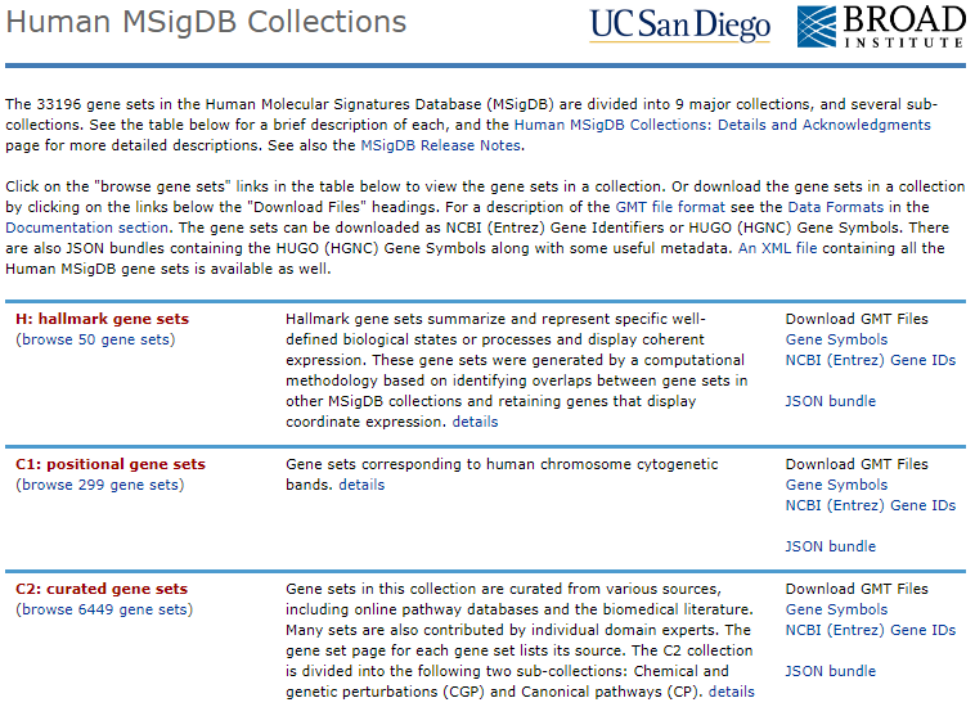
Email notification to

View 10 rows as HTML Unique results only

Gene stable ID	Gene stable ID version	Transcript stable ID	Transcript stable ID version	PDB ID
ENSG00000160072	ENSG00000160072.20	ENST00000673477	ENST00000673477.1	
ENSG00000160072	ENSG00000160072.20	ENST00000308647	ENST00000308647.8	
ENSG00000142611	ENSG00000142611.17	ENST00000511072	ENST00000511072.5	
ENSG00000142611	ENSG00000142611.17	ENST00000378391	ENST00000378391.6	2N1I
ENSG00000142611	ENSG00000142611.17	ENST00000378391	ENST00000378391.6	6BW4
ENSG00000142611	ENSG00000142611.17	ENST00000514189	ENST00000514189.5	
ENSG00000142611	ENSG00000142611.17	ENST00000270722	ENST00000270722.10	2N1I
ENSG00000142611	ENSG00000142611.17	ENST00000270722	ENST00000270722.10	6BW4
ENSG00000142611	ENSG00000142611.17	ENST00000463591	ENST00000463591.1	
ENSG00000142611	ENSG00000142611.17	ENST00000509860	ENST00000509860.1	

- Useful as input for bioinformatics tools

MSigDB: Curated gene sets



Human MSigDB Collections

UC San Diego BROAD INSTITUTE

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.

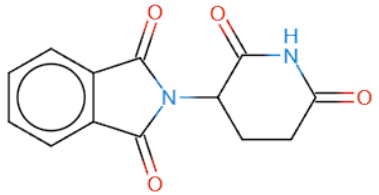
H: hallmark gene sets (browse 50 gene sets)	Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying overlaps between gene sets in other MSigDB collections 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



Gene	rsID	Star Allele	Genotype ID[?]	Variant Consequence	Drug Response Phenotype
SULT1C4	rs1402467	N/A	2_108378352_C_G,G	VEP Missense variant ProtVar	increased chance of response to tre
CHST3	rs1871450	N/A	10_72012256_G_A,A	VEP 3 prime utr variant	increased risk of toxicity[?]
CHST3	rs1871450	N/A	10_72012256_G_A,G	VEP 3 prime utr variant	increased risk of toxicity[?]

Indication	Therapeutic Ar			
immune system disease	immune syster			
multiple myeloma	4 areas			
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%	¯L=187, σ=95.9, 100%	L=44, 50%	¯L=123, σ=65.5, 50%

Legend

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

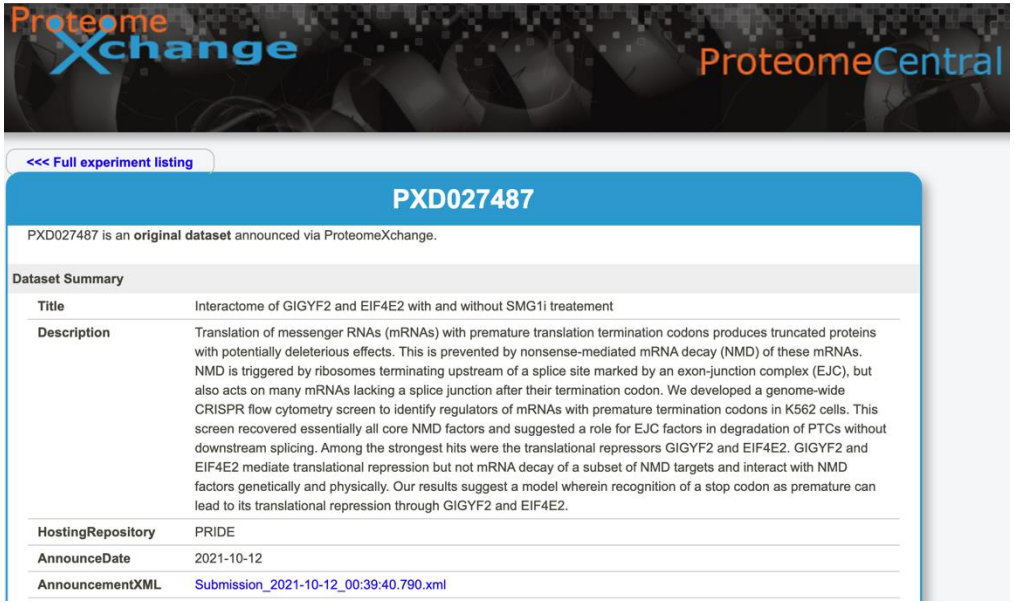
Biosample	Sample Description	Organism	Links
SAMN00001583 (SRS000100)	Human HapMap individual Coriell catalog ID NA18505	Homo sapiens	• dbSNP Batch ID 1061891 • Individual record in dbSNP

Bioproject	SRA Study	Title
PRJNA33627	SRP000001	Paired-end mapping reveals extensive structural variation in the human genome

Show abstract

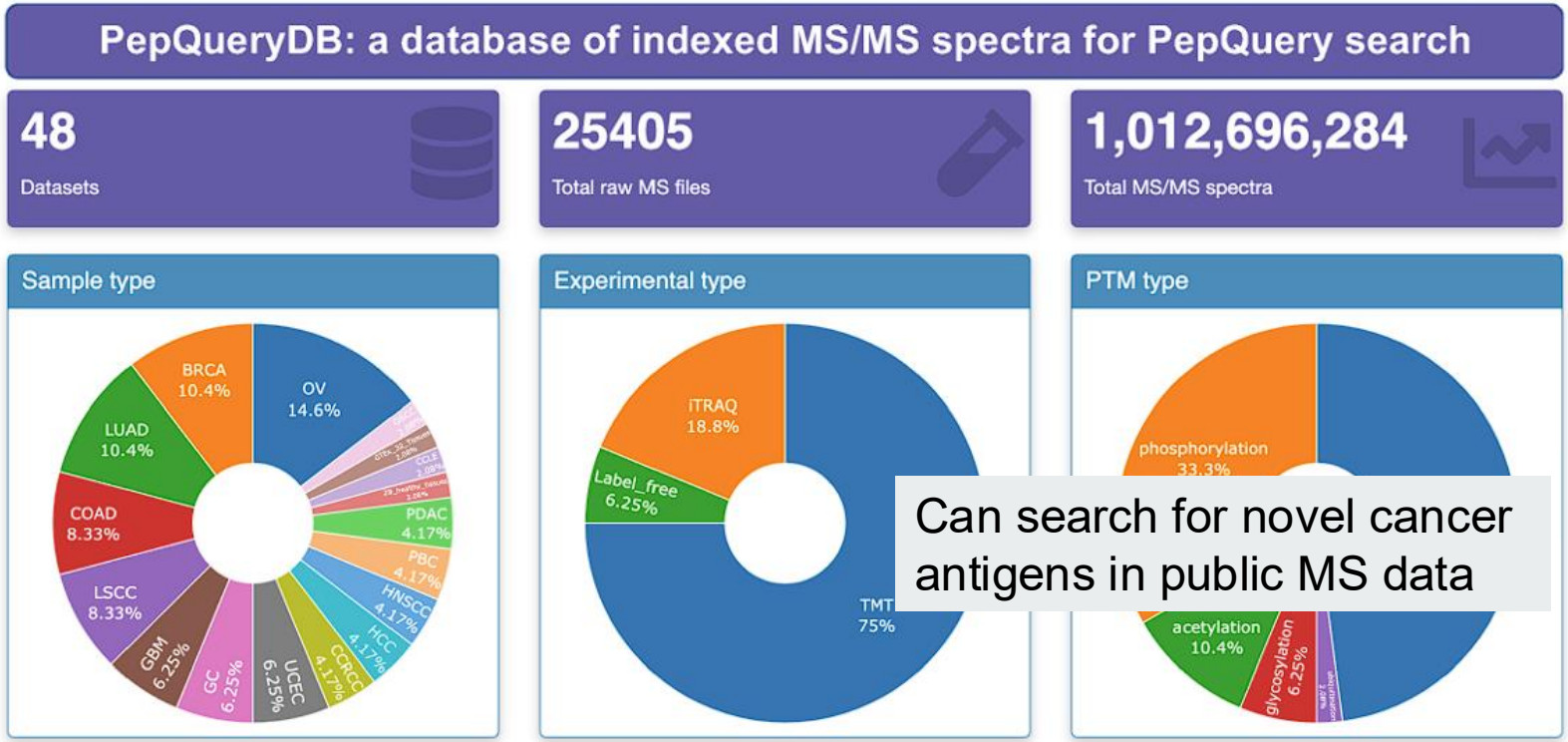
Also, European Nucleotide Archive (ENA) and Japan's DDBJ

PRIDE / MassIVE / ProteomeXchange



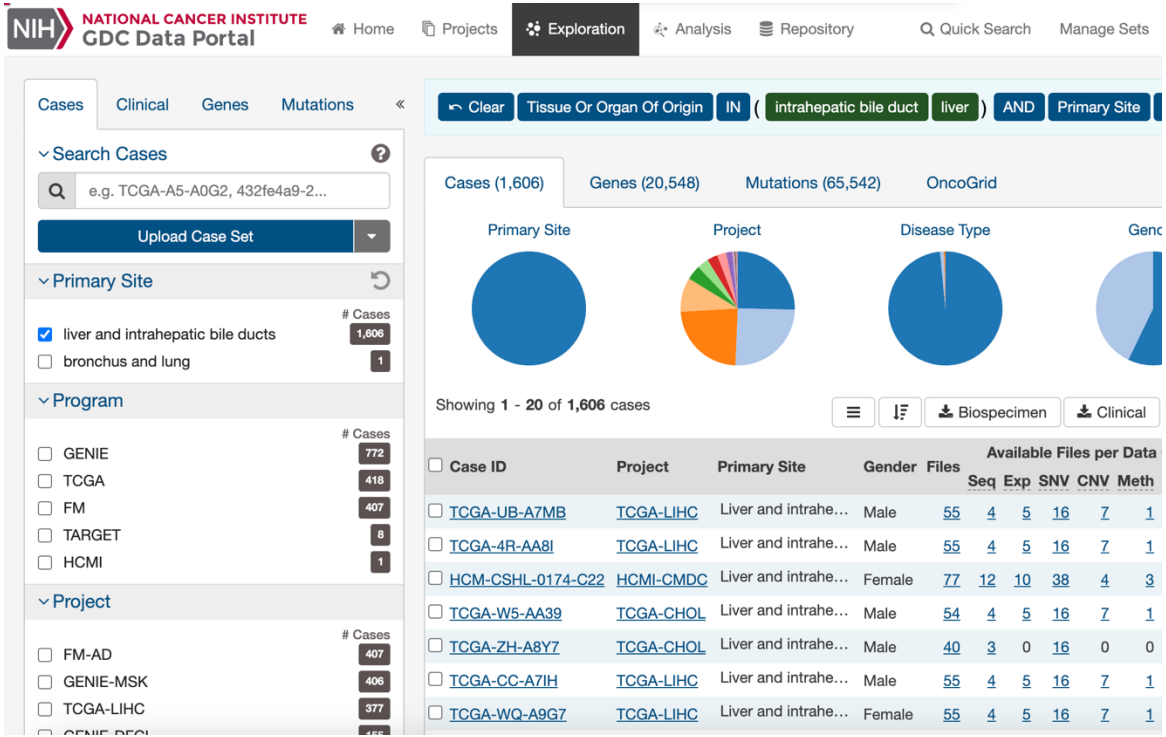
- Repository for raw mass spectrometry data (mostly proteomics)
- Search for already identified peptides/proteins

PepQuery: database for cancer proteomics



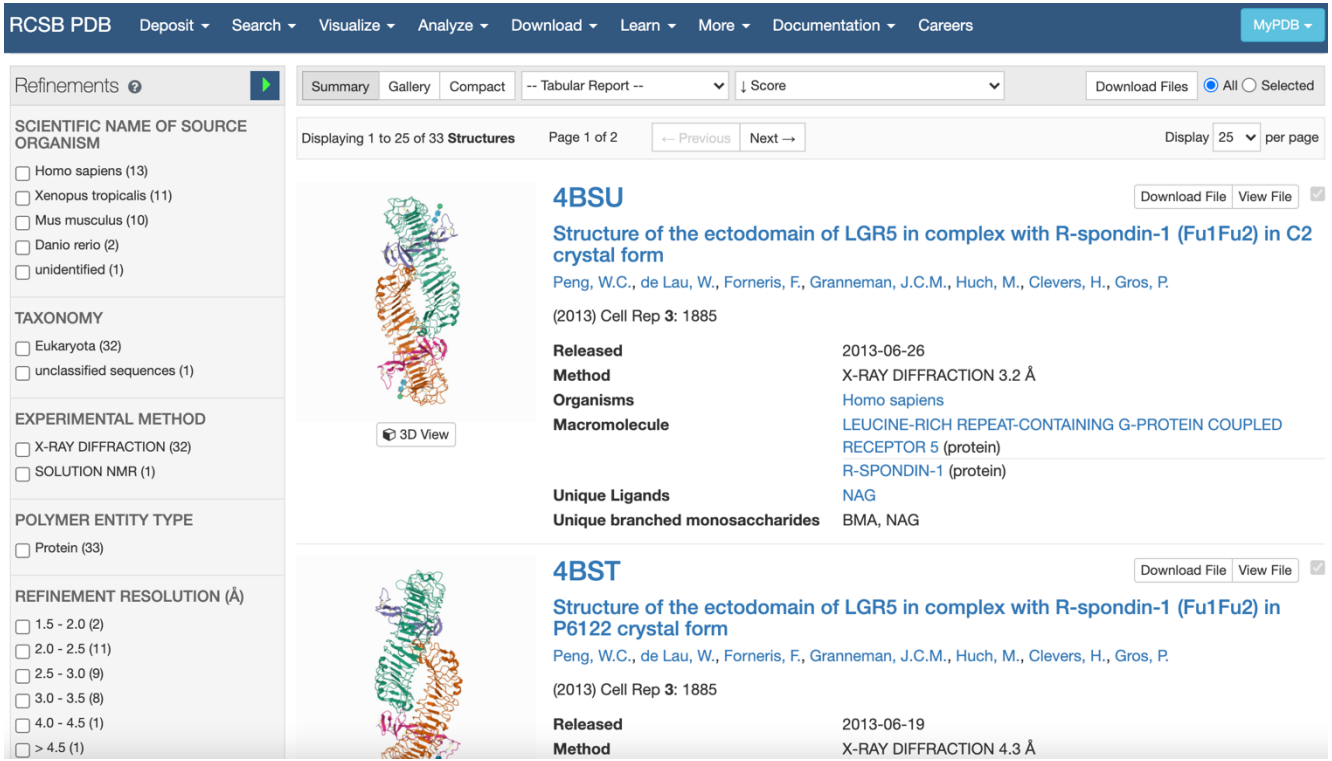
Can search for novel cancer antigens in public MS data

Genomic Data Commons: cancer multiomics

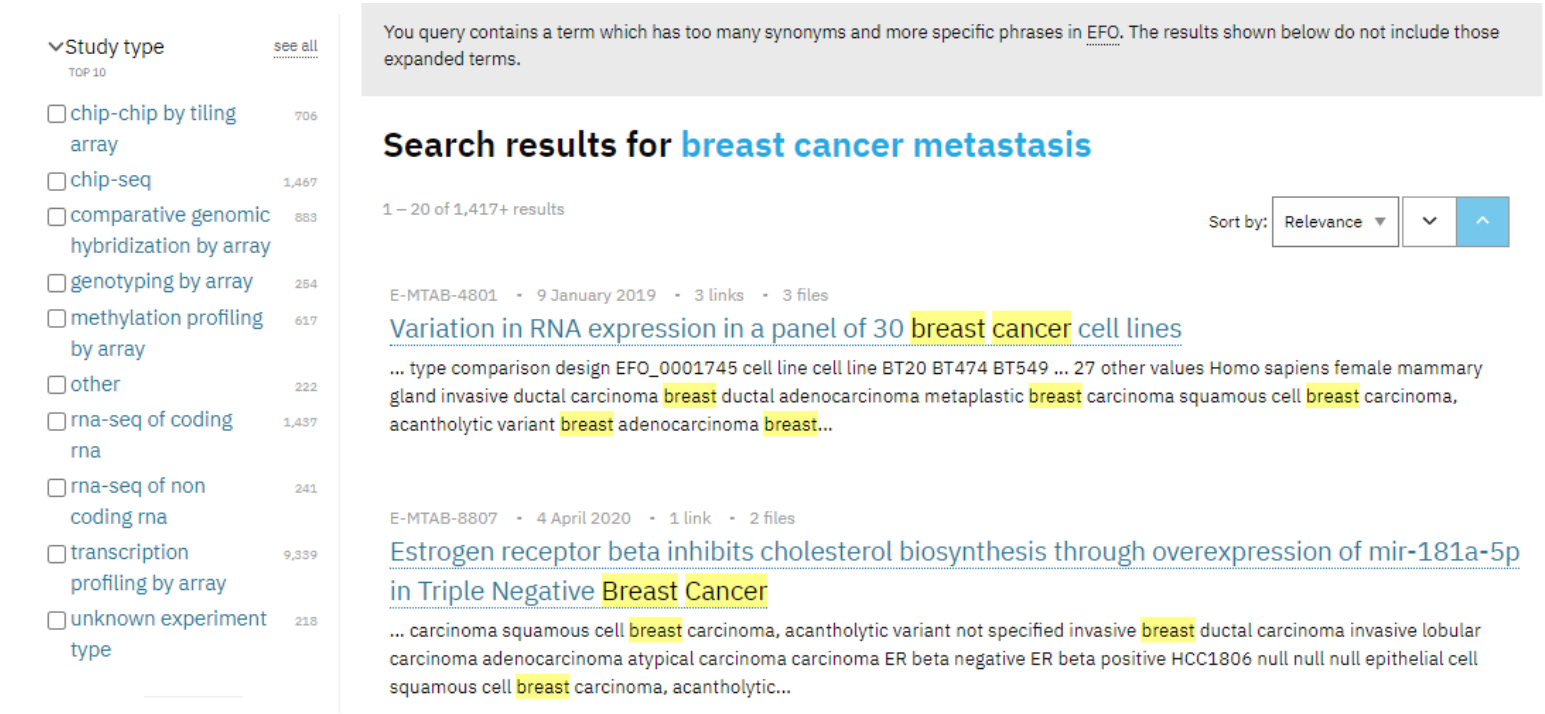


- Cancer multi-omics
  - Exome
  - RNA-seq
  - Methylation
- Include clinical and demographic data
- Link to pathology and radiology images / proteomics data in some cases

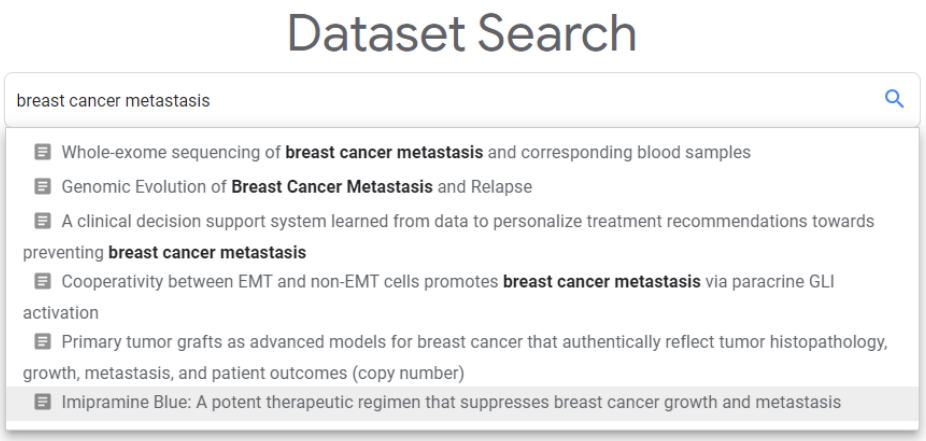
Protein Data Bank: database of 3D structures



Gene Expression Omnibus / ArrayExpress



Google Dataset / FigShare

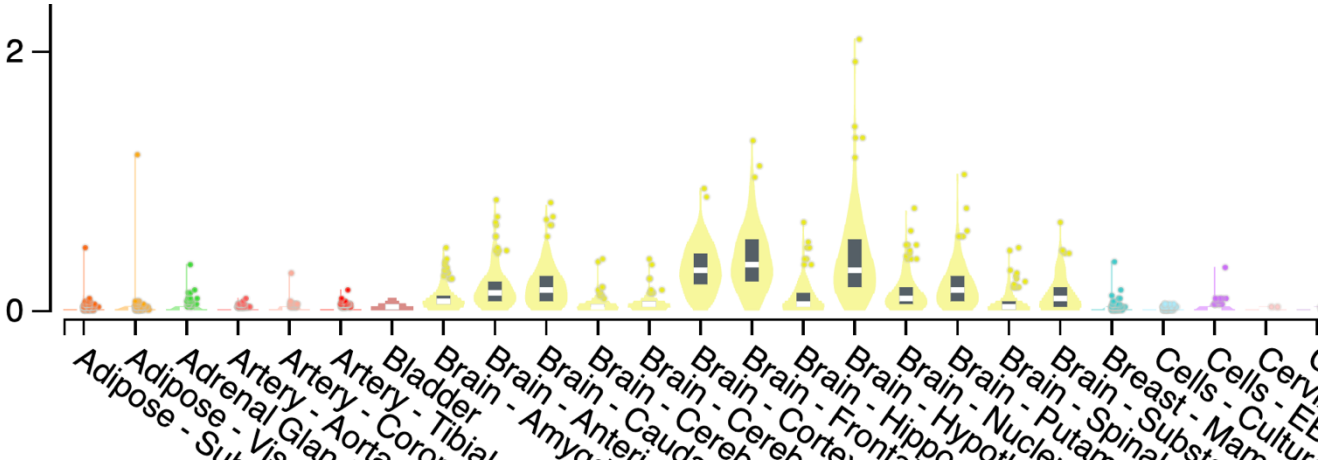


- Google Dataset compiles links from major public data repositories
- Figshare is a free repository that scientists often use for non-omics data

Atlas of human data

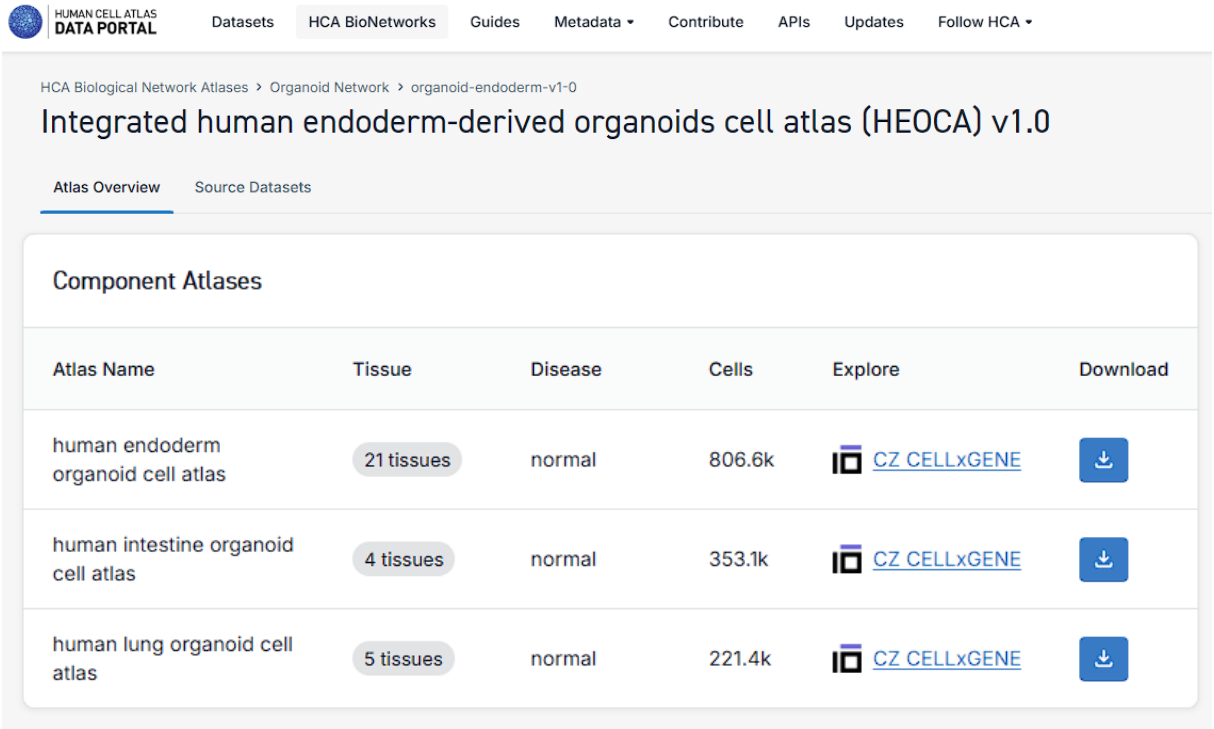
GTEx: database for tissue transcriptomics markers

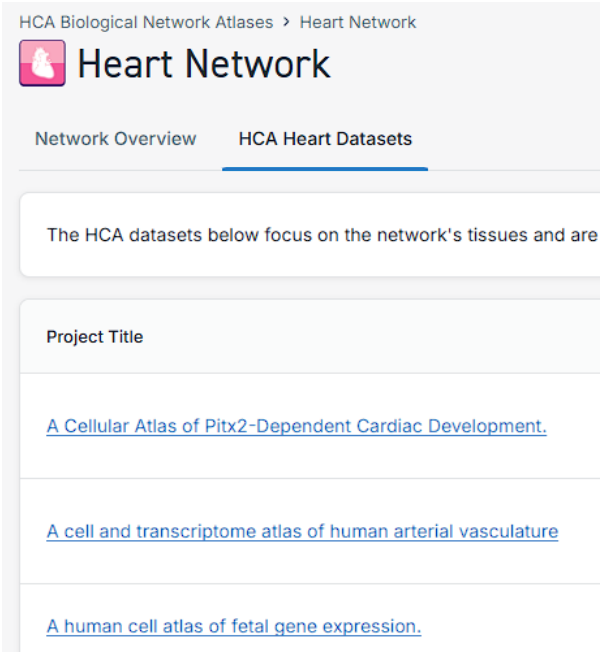
Gene Symbol	Gencode ID	Entrez Gene ID	Location	Gene Description
LIN28B	ENSG00000187772.7	389421	chr6:104936616-105083332:+	lin-28 homolog B [Source:HGNC Symbol;Acc:HGNC:32207]



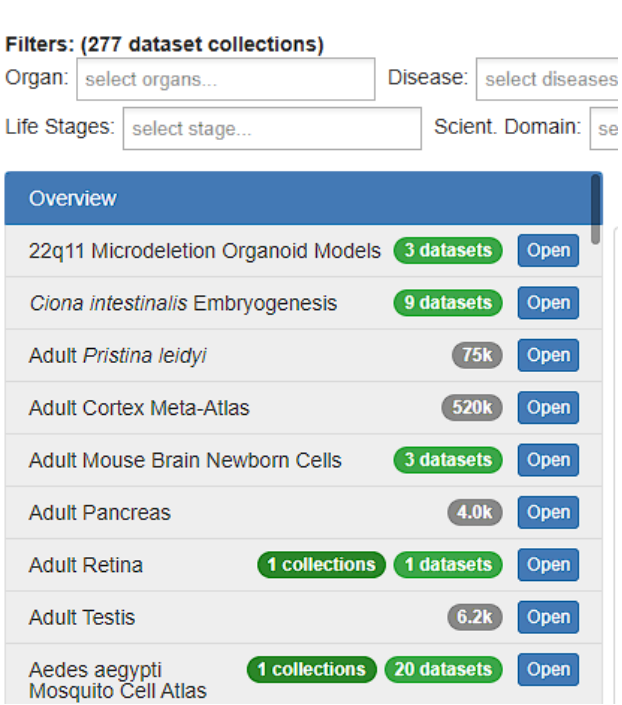
- Gene expression profiles for normal tissues (human & primates)

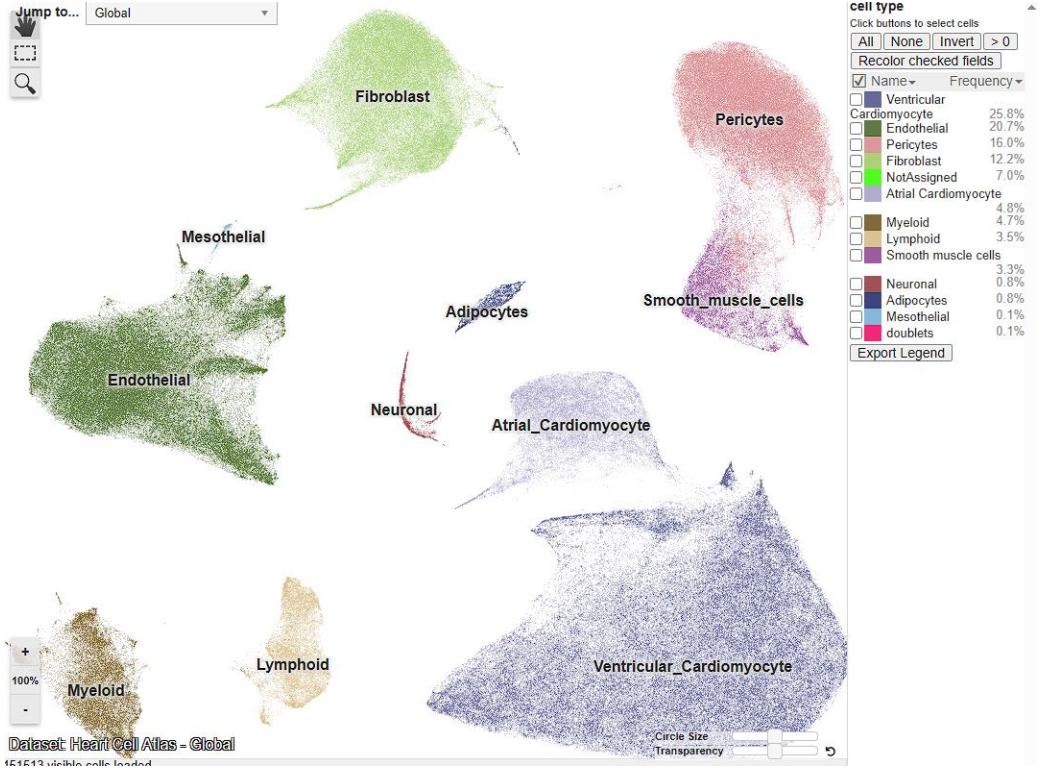
Human Cell Atlas: cell type-specific omics data





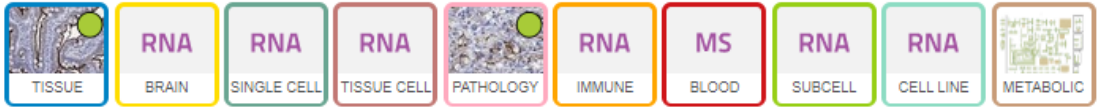
UCSC Cell Browser: visualizer for public single-cell data





Human Protein Atlas: body-wide biomarkers

ACE2



PROTEIN SUMMARY

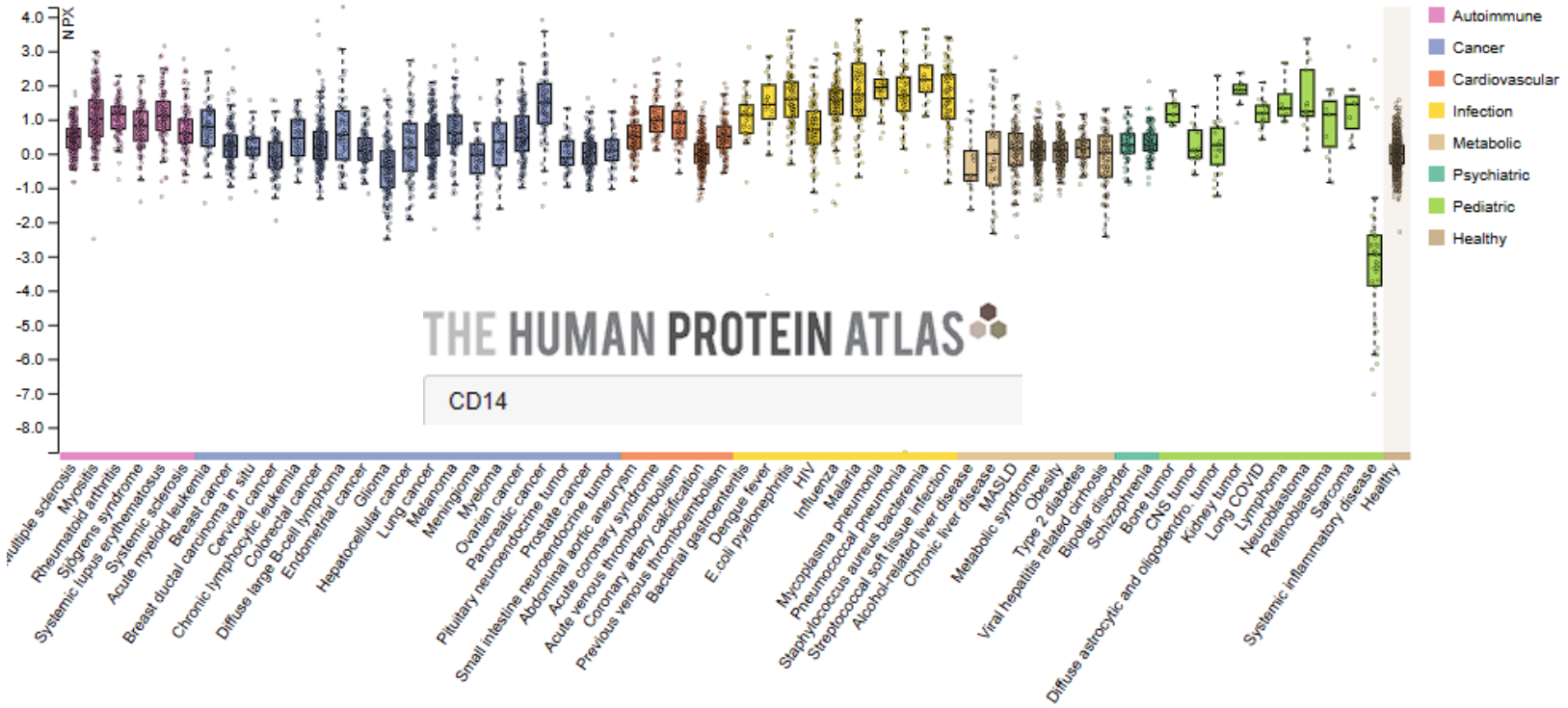
RNA DATA

GENE/PROTEIN

ANTIBODIES AND VALIDATION

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

Human Protein Atlas: disease biomarkers



Exploration of interaction networks

BioGrid: database for protein-protein interactions

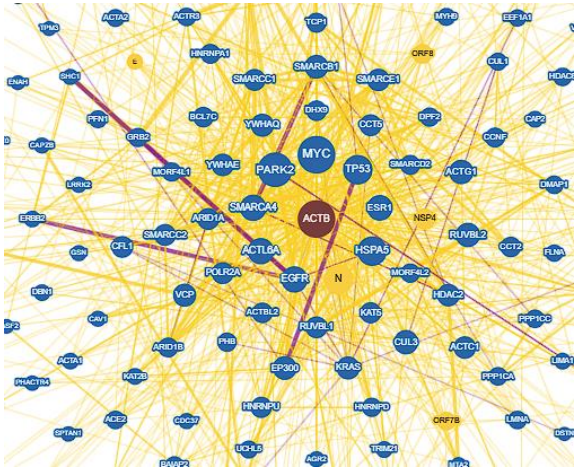
ACTB

actin, beta

1213 unique interactors

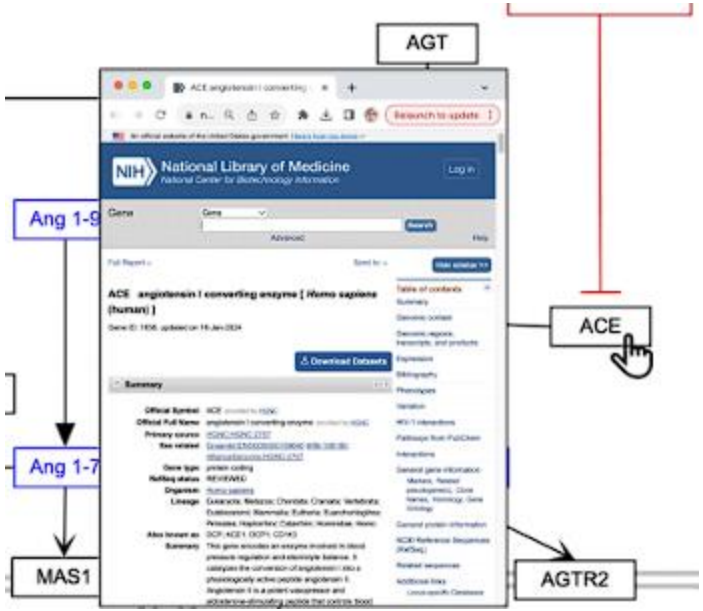
1683 raw interactions

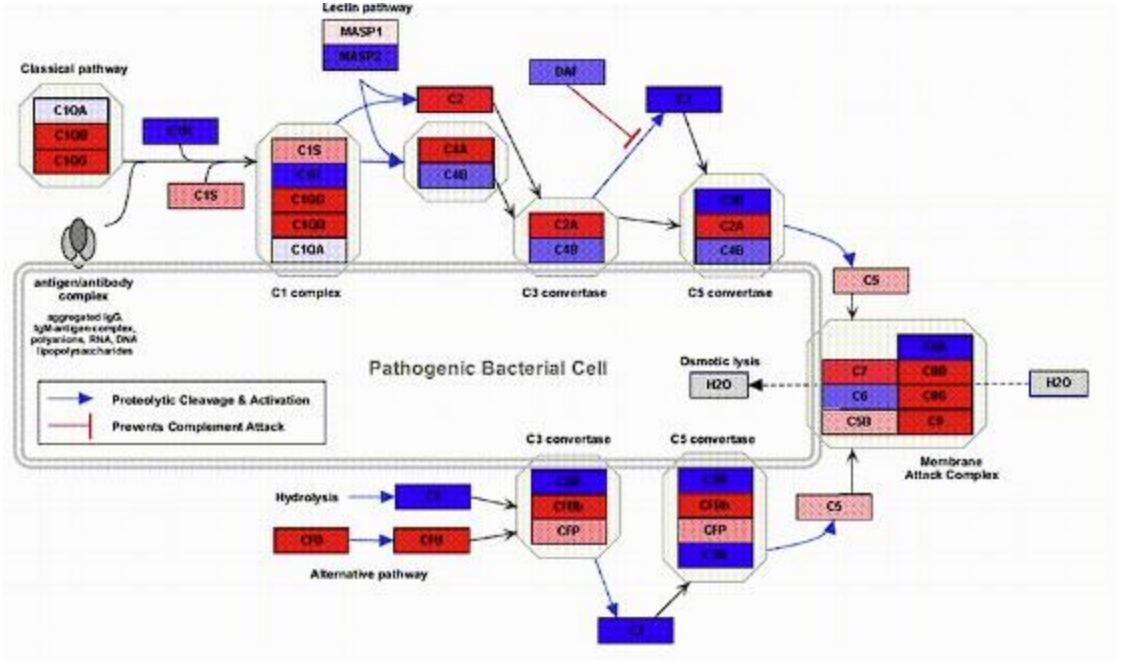
31 post-translational modifications



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

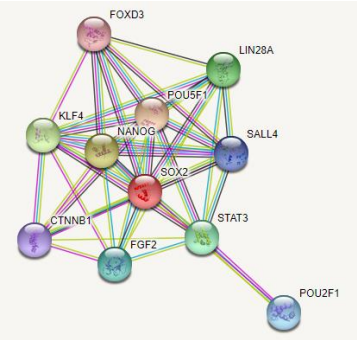
WikiPathway: integrating pathway with expression data





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

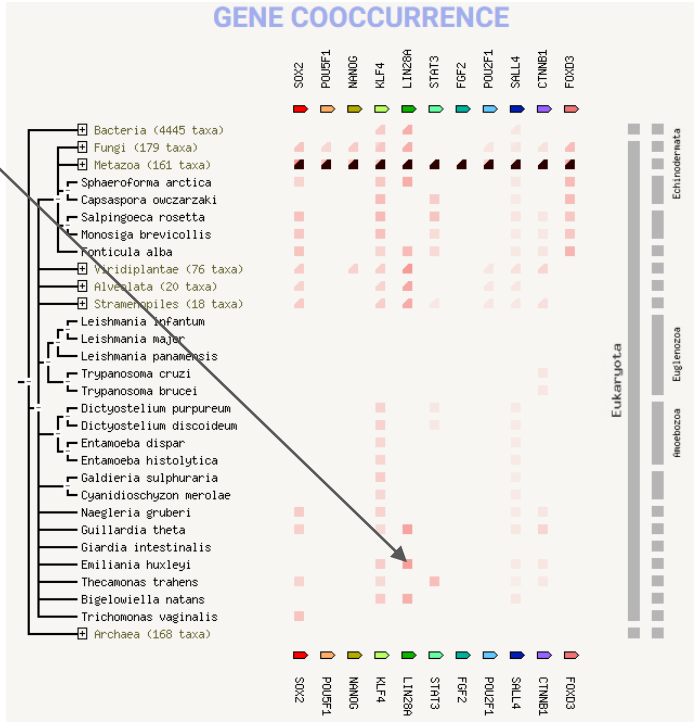
STRING: multi-evidence protein-protein interaction




```
LIN28A      25      PEDAARAAD-EPQLLHGAGICKWFNVRMGFGFLSMTARAGVALDPPVDV
                    P  A R AD EP     +G CKWF+V+ GFGF+ +        +   D+
EOD22827    19      PNTATRVADPEPA---PSGKCKWFDVQKGFGFIDVE-------NQEQDL

LIN28A      73      FVHQSKLHMEGFRSLKEGEAVEFTFKKSAKG--LESIRVTGPGGVFCIG
                    FVHQ+ +  +GFRSL EGEA+EF   + AK   L++I VTGPGG F  G
EOD22827    58      FVHQTDIKAKGFRSLAEGEALEFKVSRDAKTNKLKAIEVTGPGGDFVEG

LIN28A      120     SERRP
                    + R P
EOD22827    107     APREP
```

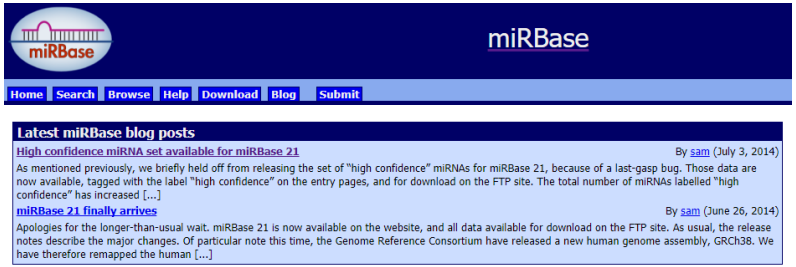


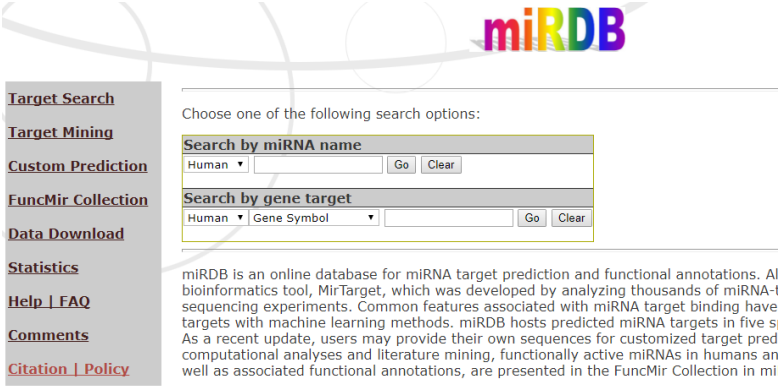
- Combine experimental evidence with literature mining (co-occurrence) and evolution (co-appear in same taxa)

- Link to external 3D structure and protein expression data

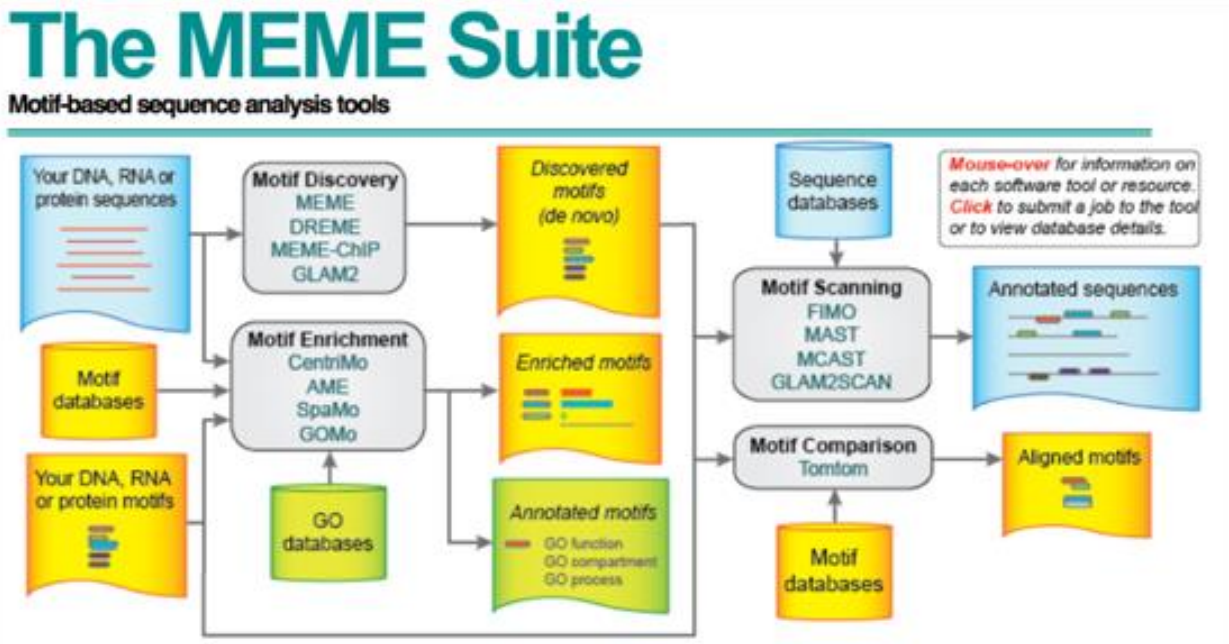
Analysis tools

miRNA target



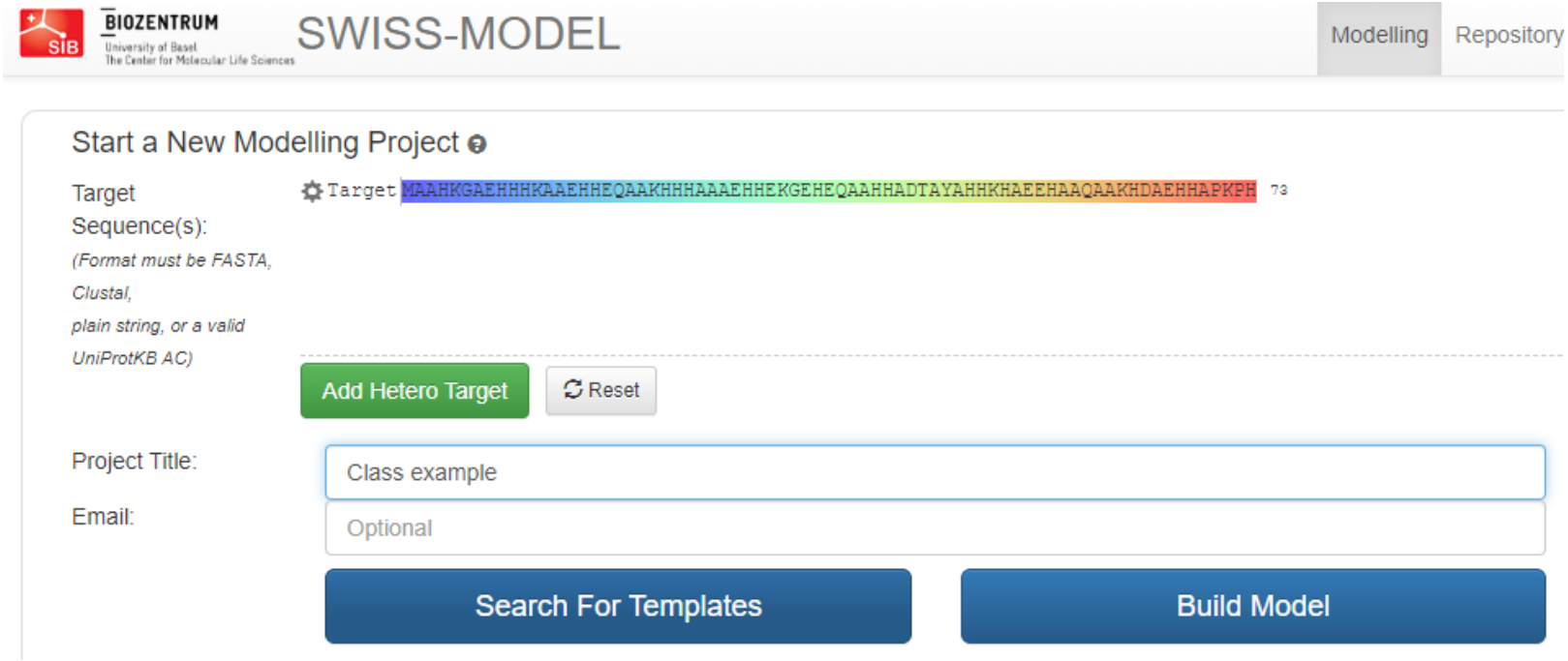


MEME Suite: One-stop service for DNA motif analysis

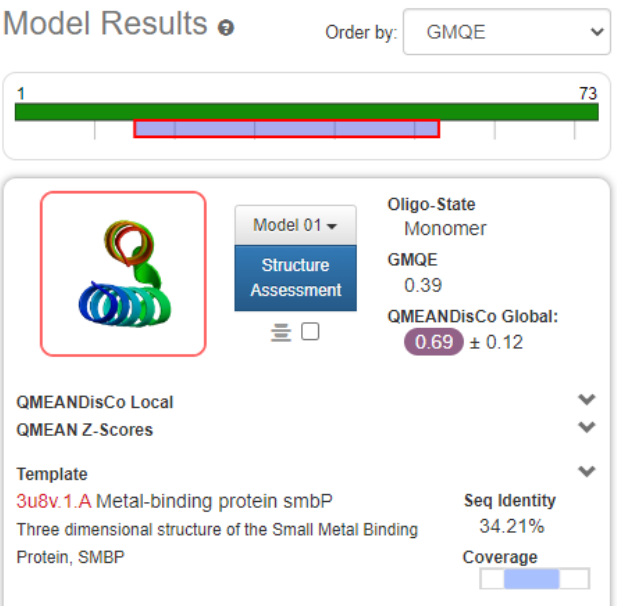


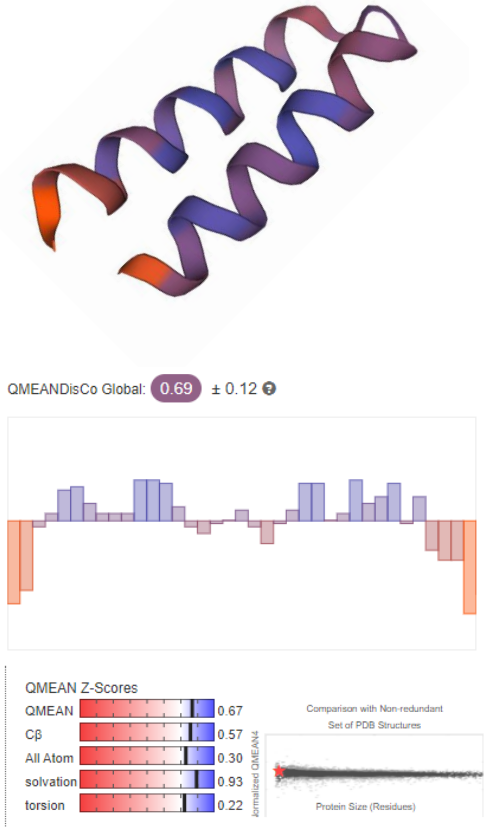
- Explore transcription factor's roles
- Find binding motifs near differentially expressed genes
- Detect new regulatory signals

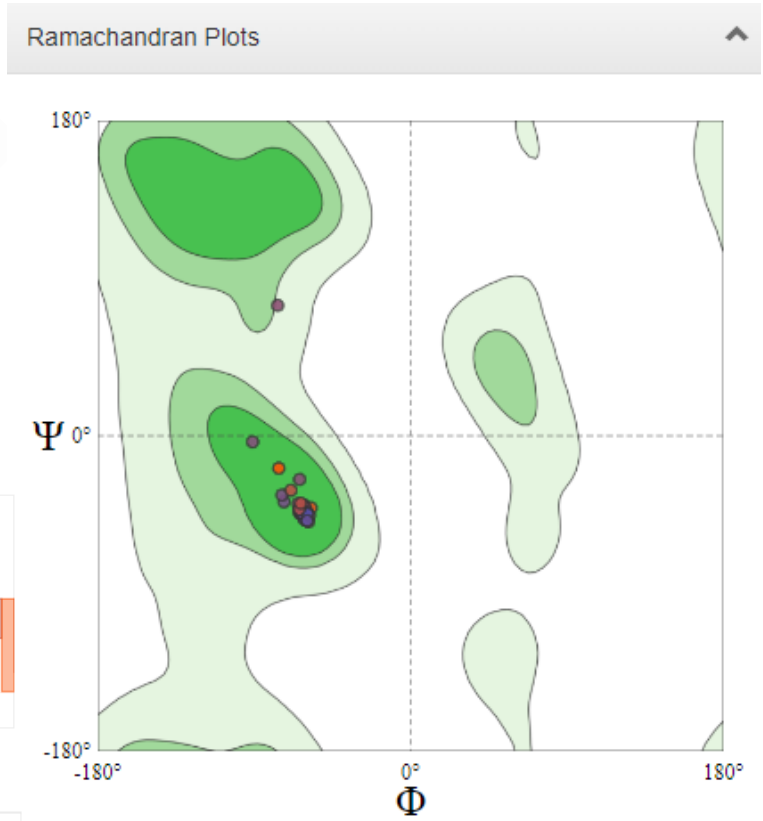
SWISS-MODEL: Homology modeling



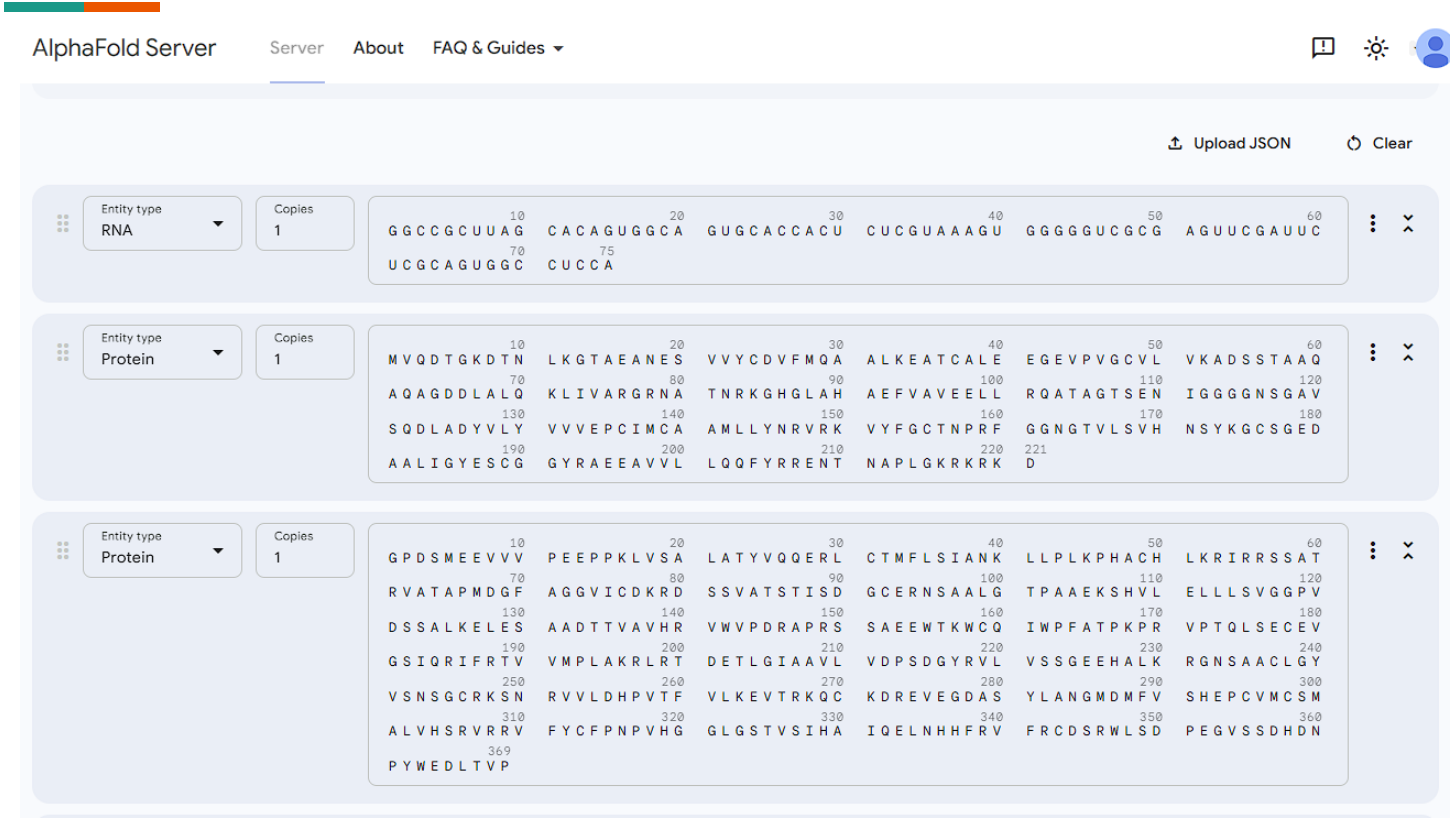
Model quality check



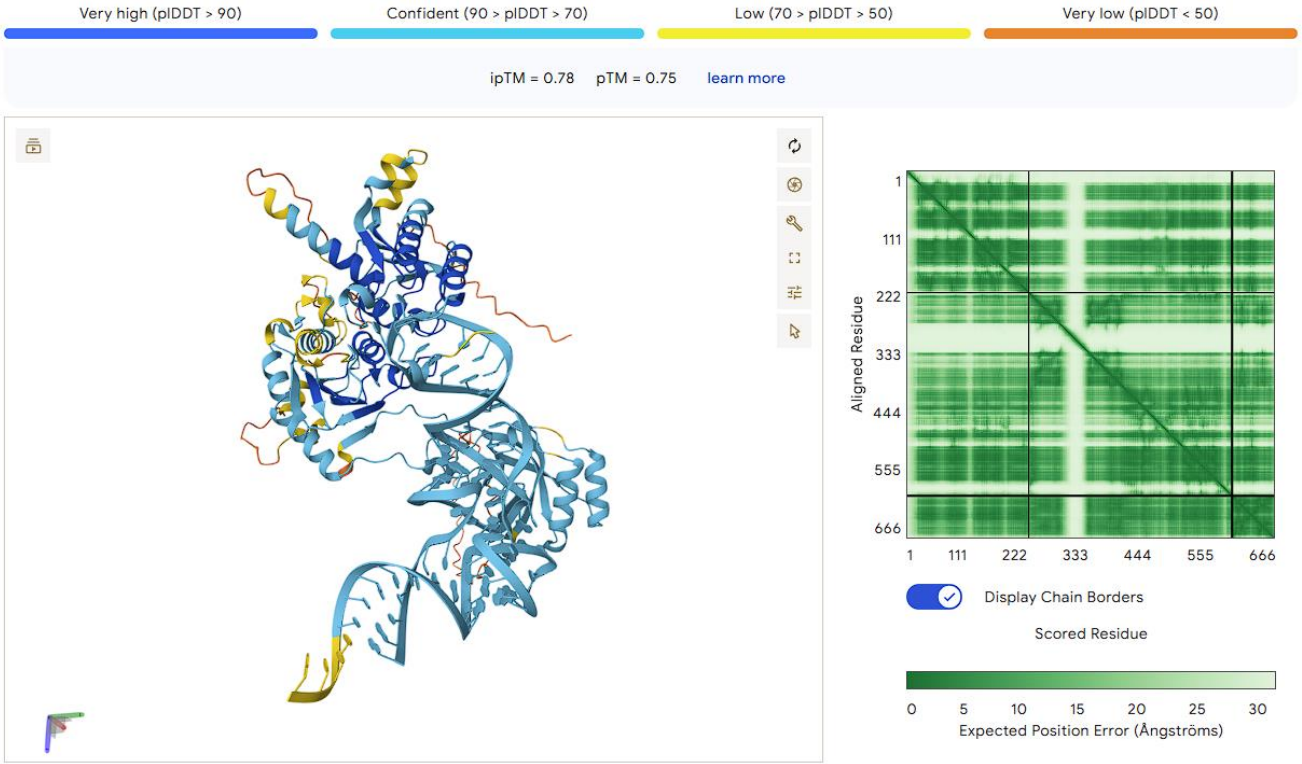




AlphaFold v3: multi-mer structure prediction



AlphaFold v3: multi-mer structure prediction



Foldseek: protein structural similarity search

Foldseek Search

GITHUB SÖDING LAB S

Results for job: jLU-XgYLj2cTLg_bEG2z2krQqhqy_FtTfZcNQg

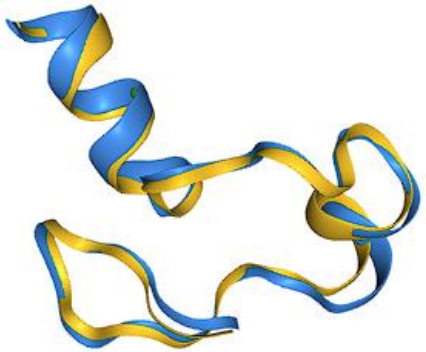
ALL DATABASES | BFVD (79) | AFDB-PROTEOME (23) | AFDB-SWISSPROT (19) | AFDB50 (1000) | BFMD (2) | CATH50 (11) | GMG

BFVD 79 hits

Target	Scientific Name	Prob.	Seq. Id.	E-Value	Position in query
A0A2H4JEY9	uncultured Caudovirales phage	0.99	35.2	2.95e-1	
A0A6J5L1I0	uncultured Caudovirales phage	0.84	44.8	1.06e+0	
A0A889IQ95	Burkholderia phage BCSR129	0.84	40	3.18e-1	
A0A6J7XTM1	uncultured Caudovirales phage	0.84	32.1	9.83e-1	
A0A7G8LLJ9	unclassified Nymphadoravirus	0.66	25	9.83e-1	
A0A2I6SCC4	White spot syndrome virus	0.54	25.8	3.04e+0	
A0A0A0RV67	Bacillus phage Moonbeam	0.54	21.2	1.79e+0	

TM-Score: 0.8494

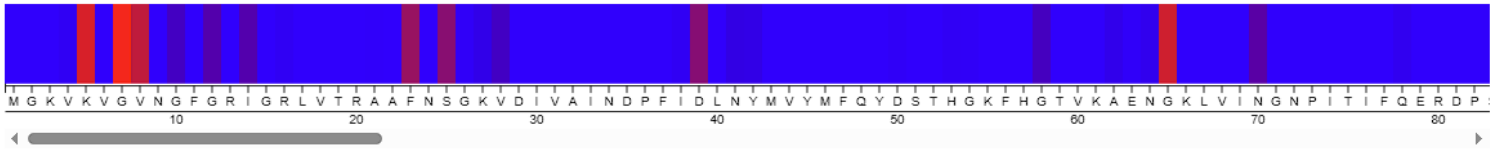
RMSD: 0.73



- Discover proteins in nature with desire 3D structures

DeepFRI: predicting protein functions

Sequence View

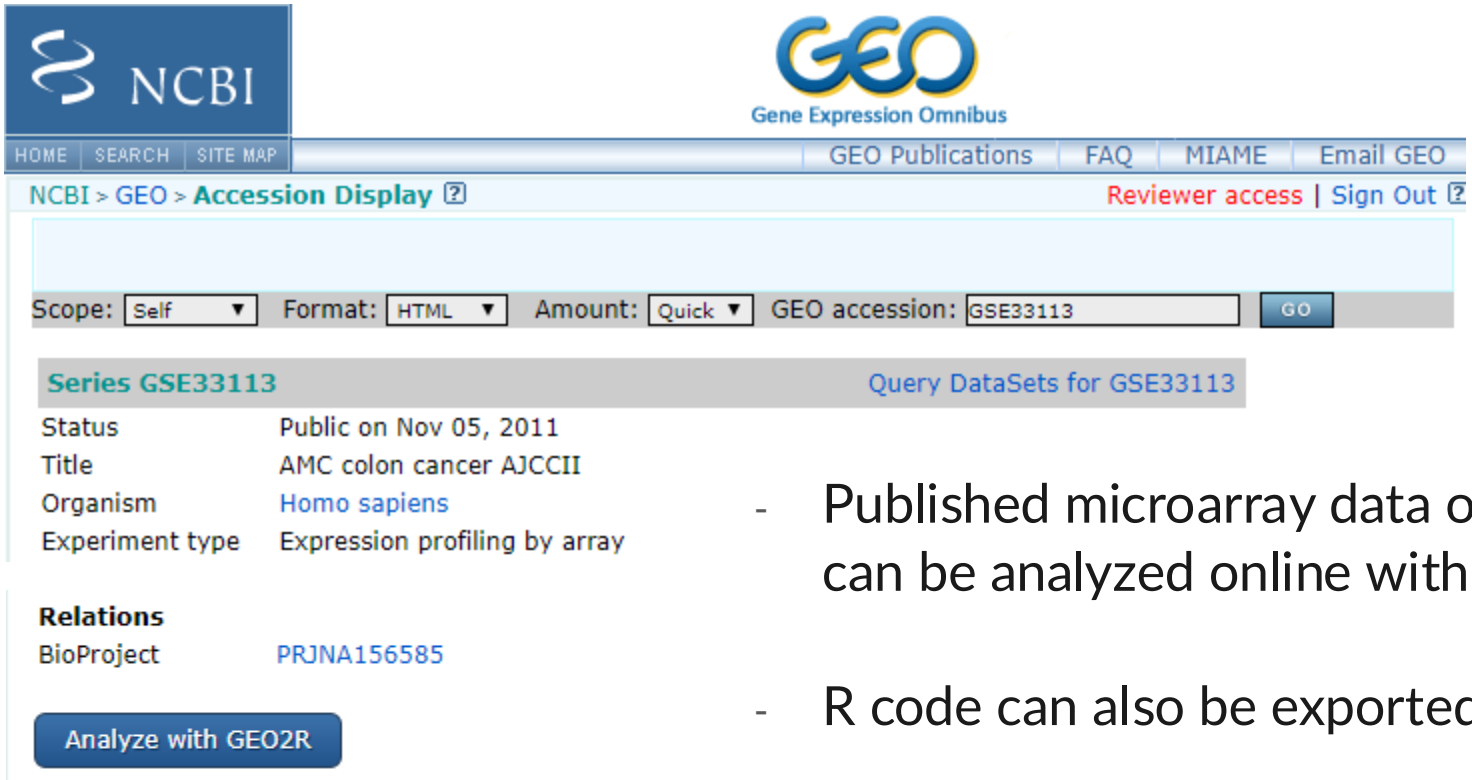


Chain: seq

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



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

GenePattern / Galaxy: online bioinformatics platforms

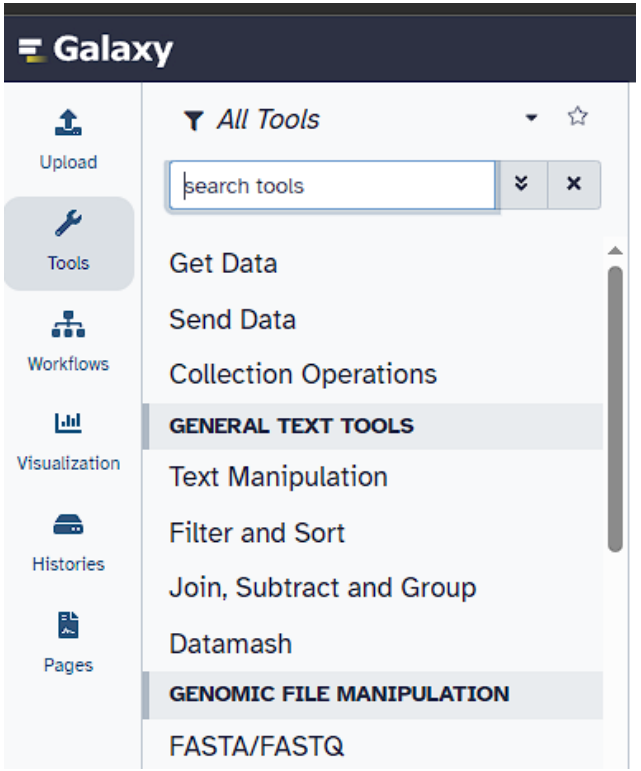
Features

Powerful genomics tools in a user-friendly interface

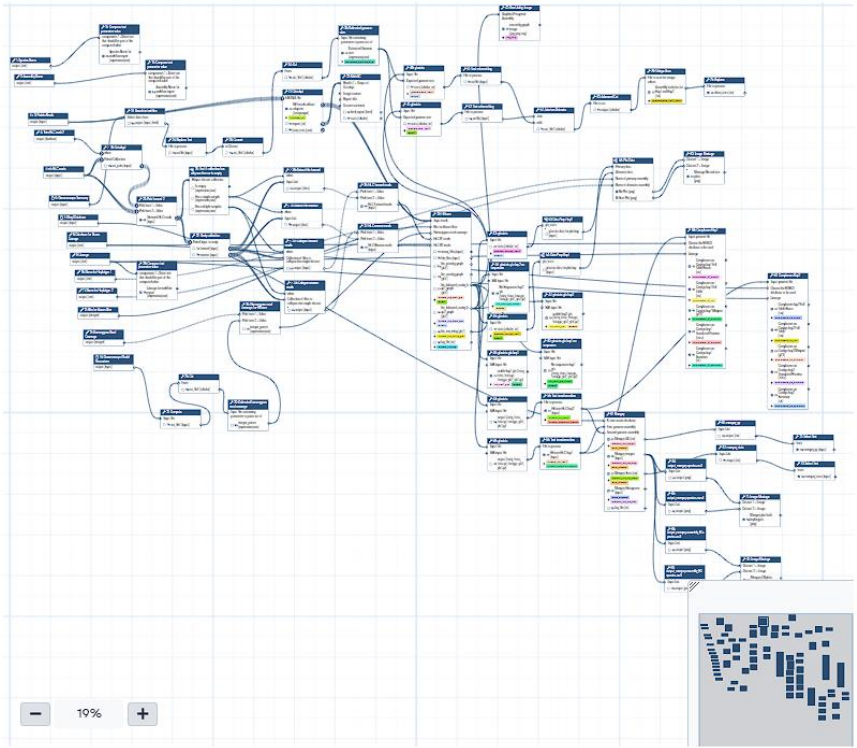
GenePattern provides hundreds of analytical tools for the analysis of gene expression (RNA-seq and microarray), sequence variation and copy number, proteomic, flow cytometry, and network analysis. These tools are all available through a Web interface with no programming experience required.

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.



Public bioinformatics workflow on Galaxy



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

iwc

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

Show more

VGP Reviewed

about 6 hours ago

Import

Creators

Galaxy

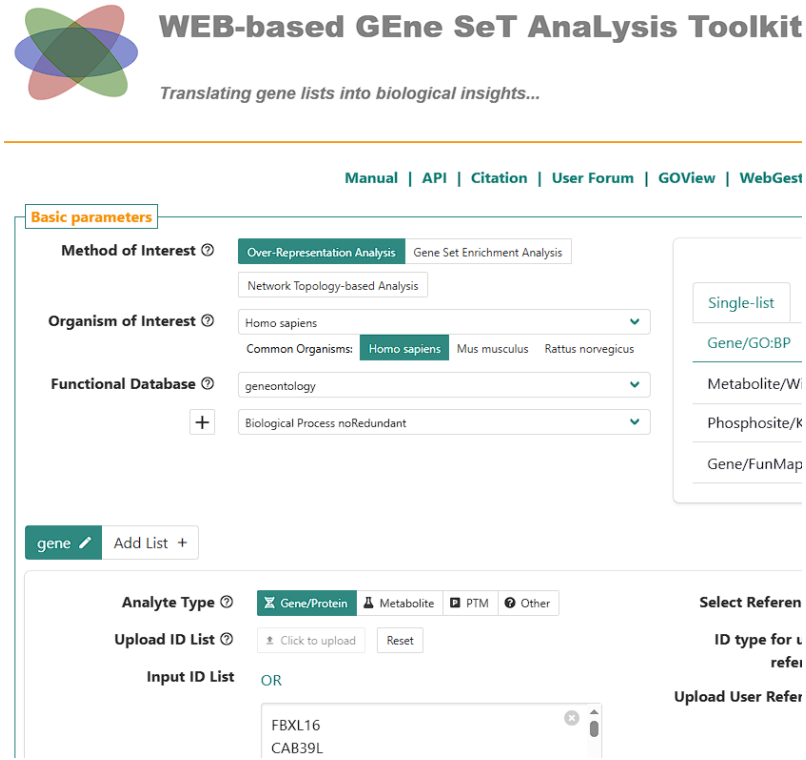
VGP

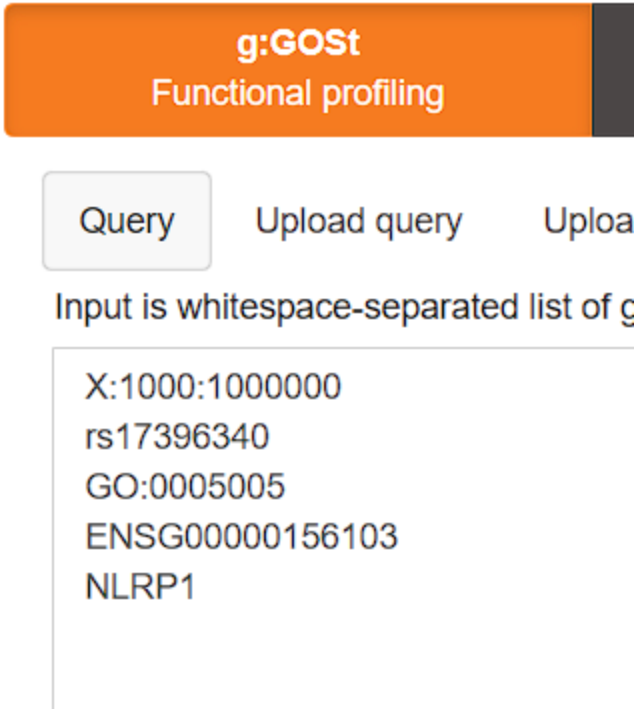
Delphine Lariviere

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





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