Biodatabases

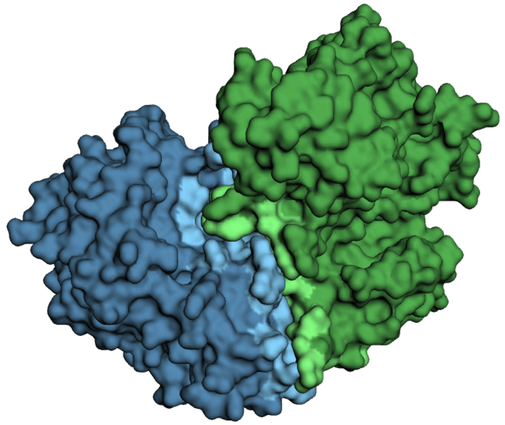
Accessing the internet's biology knowledge bases

Juan Felipe Beltrán, Ph.D.
Cornell University Department of Biomedical Engineering
juanfelipe@cornell.edu
@offbyjuan

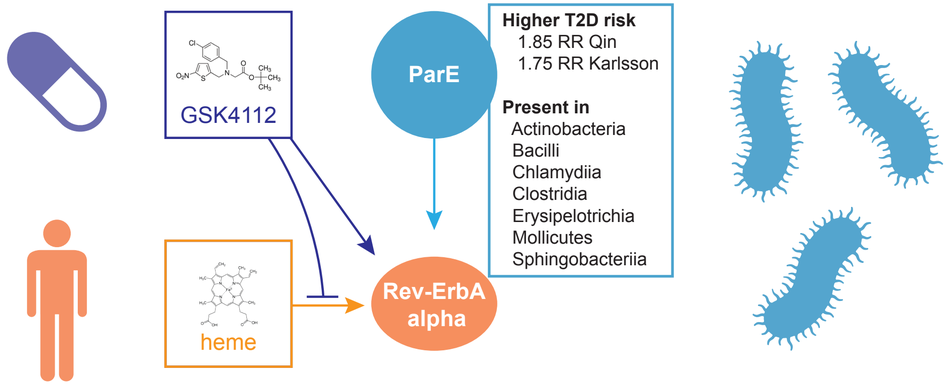
NYU Cardiovascular Research Center

About Me

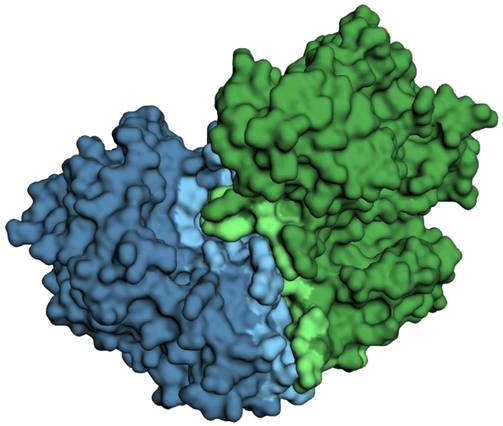
Can we predict **where** proteins bind?



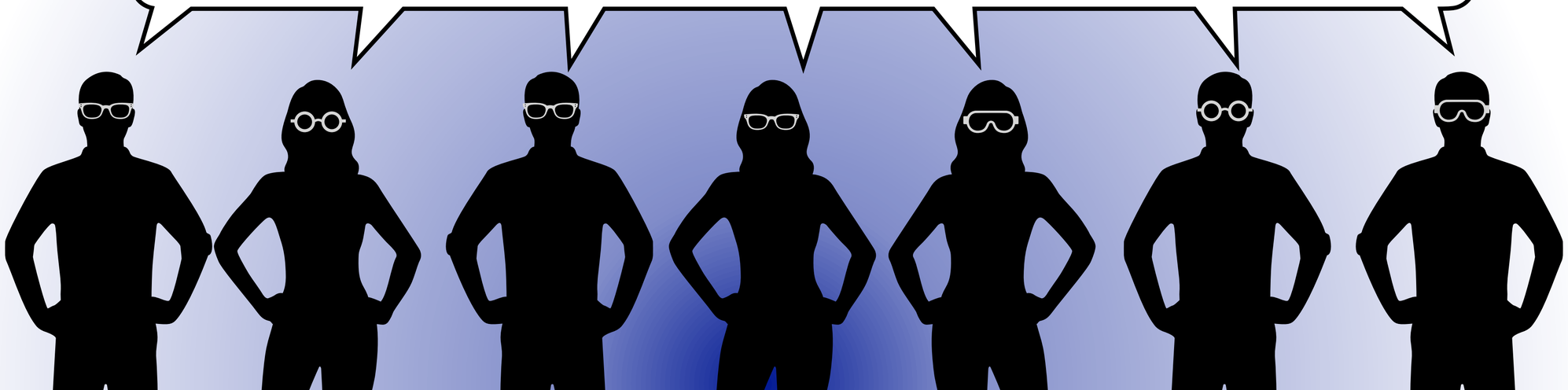
How do **bacteria** modulate human health?



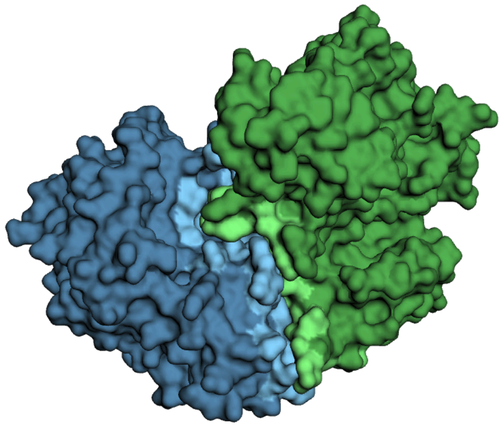
Can we predict **whether** proteins bind?



YEAH, PROBABLY!



How can we know if proteins bind?



Transcript coexpression

Structural docking!

Check their domains?

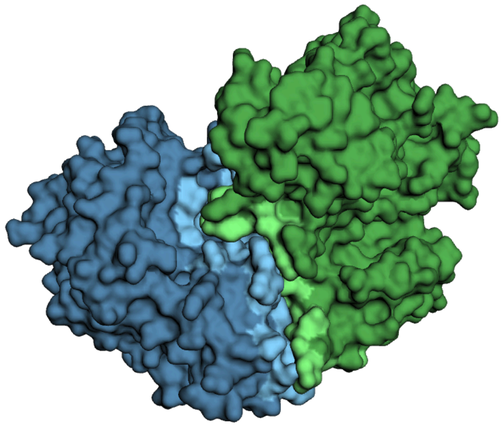
Homology modeling

Coevolution.

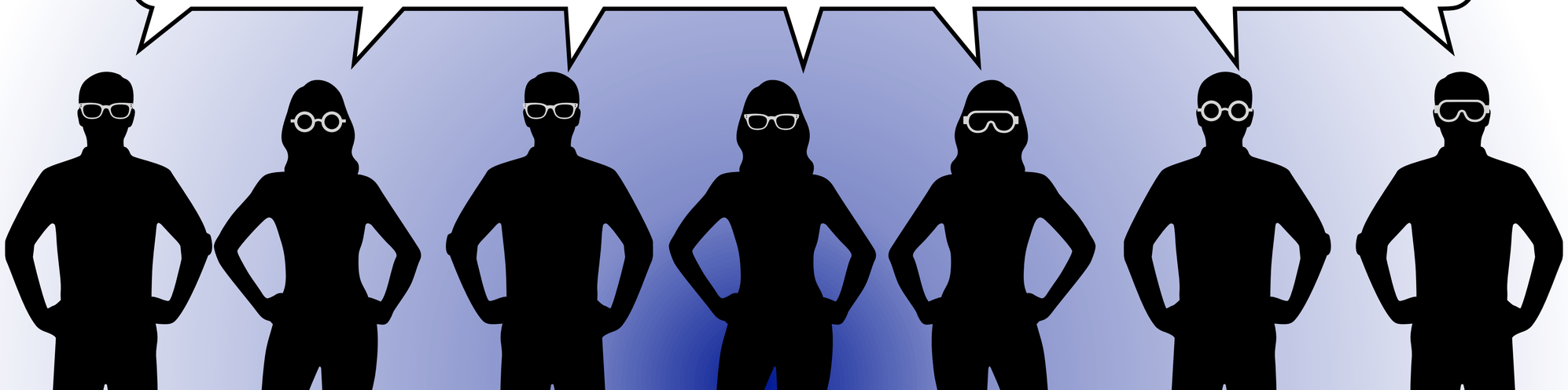
Past experiments

Matching diseases!

Go! Predict which proteins bind!



… where can I find the data?



where can I find the data?

Species

Pathways

Functions

Diseases

DNA

RNA

Proteins

Sequences

Mutations

Interactors

Complexes

Tissues

Cell Types

Organelles

Polymers

Probes &
Drugs

Pathogens

Commensals

Motifs

Domains

Structures

UniProt.org

A pretty great “North Star”

UniProtKB

UniProt Knowledgebase

Swiss-Prot (562,253)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (180,690,447)

Automatically annotated and not reviewed.

Records that await full manual annotation.

UniRef

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.

UniParc

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

Proteomes

A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

Supporting data

Literature citations

Cross-ref. databases

Taxonomy

Diseases

XXX

Subcellular locations

Keywords

UniProt.org

A pretty great “North Star”

Getting started

Text search
Our basic text search allows you to search all the resources available

BLAST
Find regions of similarity between your sequences

Sequence alignments
Align two or more protein sequences using the Clustal Omega program

Retrieve/ID mapping
Batch search with UniProt IDs or convert them to another type of database ID (or vice versa)

Peptide search
Find sequences that exactly match a query peptide sequence

YouTube

Amino Acid OR Nucleotide!

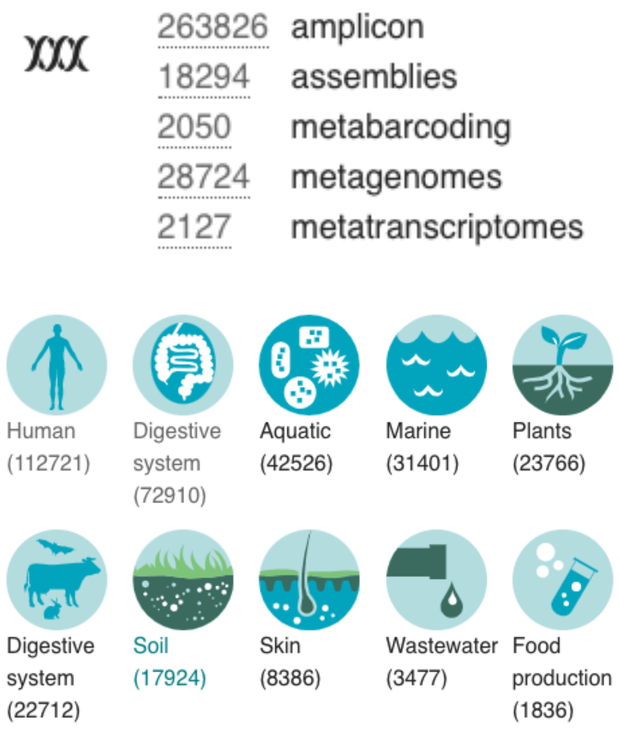
UniProt is a great way to learn about other resources!

Let’s take a tour of TP53 in the UniProt interface

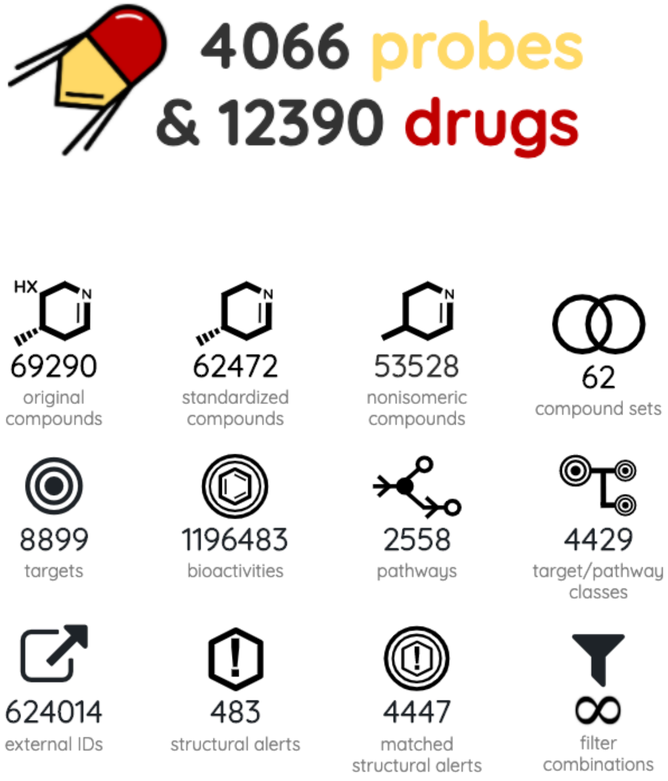
https://www.uniprot.org/uniprot/?query=tp53&sort=score

(Get ready to take notes!)

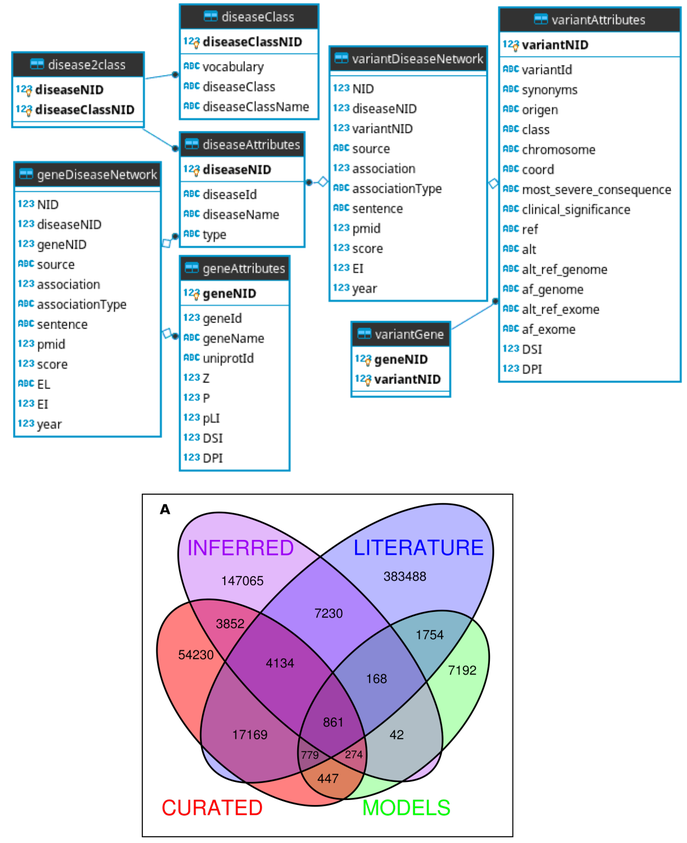
Not found in UniProt, but quite useful



MGnify
(https://www.ebi.ac.uk/metagenomics/)



Probes & Drugs
(https://www.probes-drugs.org/home/)



DisGeNet
(https://www.ebi.ac.uk/metagenomics/)

Microbiomes

Small Molecules

Disease Associations

“I will remember none of this.”

- My Audience, 2020

3 skills to survive in the biodatabases landscape

Finding identifiers

Climbing ontologies

Tracking data

Databases are very specific about their nouns!

Databases are very specific about their nouns!

“TP53”

Could mean anything

“7157”

A gene found in a **full genome**

Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	P53; BCC7; LFS1; BMFS5; TRP53

TP53

Gene: TP53
Title: tumor protein p53
Location: complement(7,668,402..7,687,550)
Length: 19,149 nt

Links & Tools
View GeneID: 7157 (TP53)
View HGNC: 11998
View MIM: 191170

Databases are very specific about their nouns!

"TP53"

Could mean anything

"7157"

A gene found in a **full genome**

Accession prefix	Molecule type	Comment
AC_	Genomic	Complete genomic molecule, usually alternate assembly
NC_	Genomic	Complete genomic molecule, usually reference assembly
NG_	Genomic	Incomplete genomic region
NT_	Genomic	Contig or scaffold, clone-based or WGS[a]
NW_	Genomic	Contig or scaffold, primarily WGS[a]
NZ_[b]	Genomic	Complete genomes and unfinished WGS data
NM_	mRNA	Protein-coding transcripts (usually curated)
NR_	RNA	Non-protein-coding transcripts
XM_[c]	mRNA	Predicted model protein-coding transcript
XR_[c]	RNA	Predicted model non-protein-coding transcript
AP_	Protein	Annotated on AC_ alternate assembly
NP_	Protein	Associated with an NM_ or NC_ accession
YP_[c]	Protein	Annotated on genomic molecules without an instantiated transcript record
XP_[c]	Protein	Predicted model, associated with an XM_ accession
WP_	Protein	Non-redundant across multiple strains and species

[a] Whole Genome Shotgun sequence data.

[b] An ordered collection of WGS sequence for a genome.

[c] Computed.

Suffixes on identifiers are used for variation

“TP53”

Could mean anything

“P04637”

A protein in UniProt

“P04637-1”

“P04637-2”

“P04637-3”

Isoforms in UniProt

Don’t know what a noun means?
Google it, follow links, go hunting!

There are even identifiers for **verbs!**

Accession	GO:0009306
Name	protein secretion
Ontology	biological_process
Synonyms	glycoprotein secretion, protein secretion during cell fate commitment, protein secretion resulting in cell fate commitment
Alternate IDs	GO:0045731, GO:0045166
Definition	The controlled release of proteins from a cell. *Source:* GOC:ai
Comment	None
History	See term history for GO:0009306 at QuickGO
Subset	goslim_drosophila
Related	Link to all **genes and gene products** annotated to protein secretion.
	Link to all direct and indirect **annotations** to protein secretion.
	Link to all direct and indirect **annotations download** (limited to first 10,000) for protein secretion.

chromatin immunoprecipitation assay

http://purl.obolibrary.org/obo/MI_0402

Chromatin immunoprecipitation (ChIP) is a powerful approach that allows one to define the interaction of factors with specific chromosomal sites in living cells. An antibody against a protein suspected of binding a given cis-element is used to immunoprecipitate fragmented chromatin fragments. Cells or tissue may first be briefly treated with an agent such formaldehyde to crosslink proteins to DNA. Nucleic acids are then identified by sequencing, for example polymerase chain reaction analysis of the immunoprecipitate with primers flanking the cis-element or next-generation sequencing techniques [http://www.ncbi.nlm.nih.gov/pubmed/12054902]

3 skills to survive in the biodatabases landscape

Finding identifiers

Climbing ontologies

Tracking data

chromatin immunoprecipitation assay

http://purl.obolibrary.org/obo/MI_0402

Chromatin immunoprecipitation (ChIP) is a powerful approach that allows one to define the interaction of factors with specific chromosomal sites in living cells. An antibody against a protein suspected of binding a given cis-element is used to immunoprecipitate fragmented chromatin fragments. Cells or tissue may first be briefly treated with an agent such formaldehyde to crosslink proteins to DNA. Nucleic acids are then identified by sequencing, for example polymerase chain reaction analysis of the immunoprecipitate with primers flanking the cis-element or next-generation sequencing techniques [http://www.ncbi.nlm.nih.gov/pubmed/12054902]

Search MI

chromatin immunoprecipitation assay

http://purl.obolibrary.org/obo/MI_0402

Chromatin immunoprecipitation (ChIP) is a powerful approach that allows one to define the interaction of factors with specific chromosomal sites in living cells. An antibody against a protein suspected of binding a given cis-element is used to immunoprecipitate fragmented chromatin fragments. Cells or tissue may first be briefly treated with an agent such formaldehyde to crosslink proteins to DNA. Nucleic acids are then identified by sequencing, for example polymerase chain reaction analysis of the immunoprecipitate with primers flanking the cis-element or next-generation sequencing techniques [http://www.ncbi.nlm.nih.gov/pubmed/12054902]

Search MI

modified chromatin immunoprecipitation

http://purl.obolibrary.org/obo/MI_1028

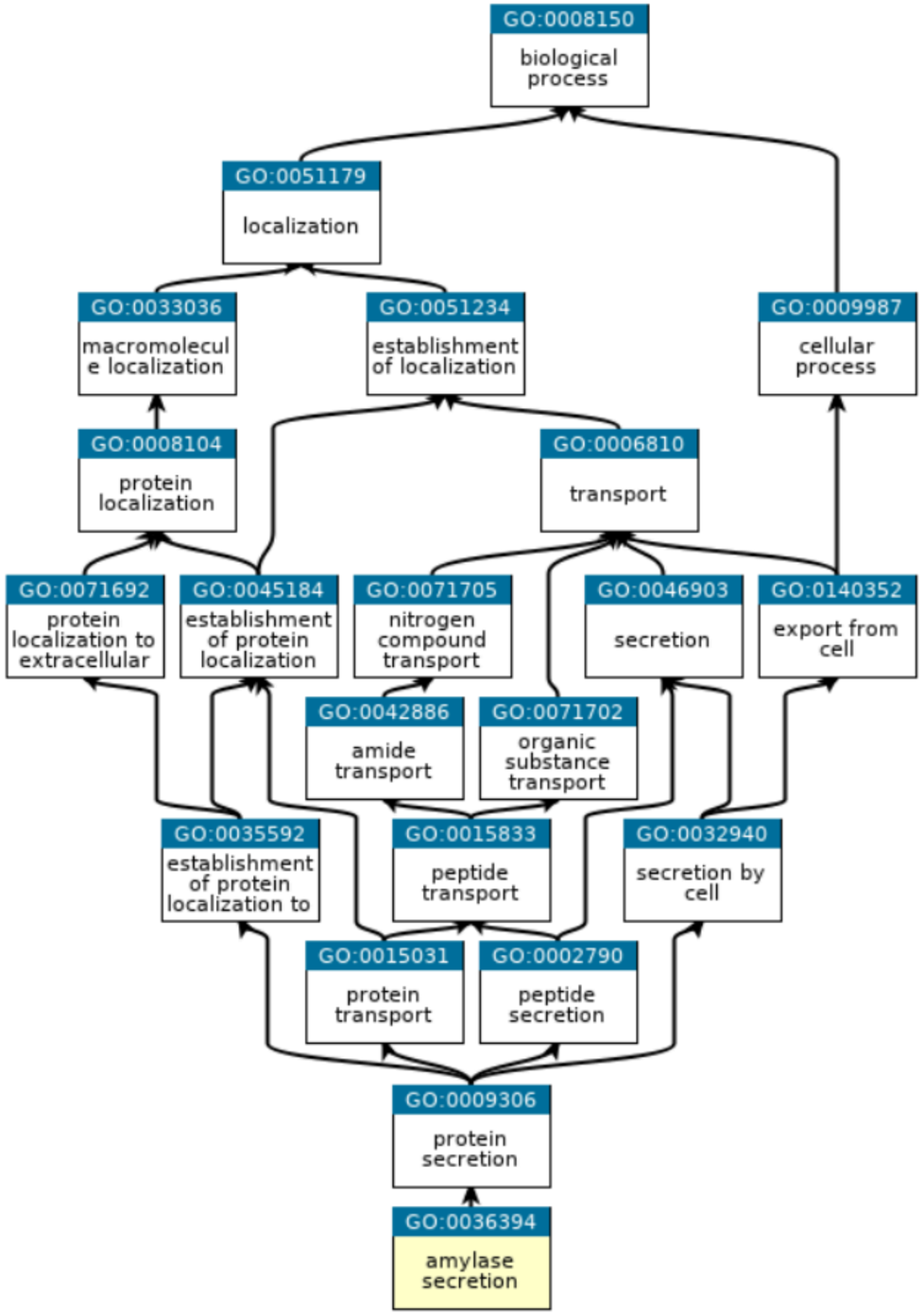
Chromatin-bound protein networks isolated using magnetic beads coated with antibodies. [http://www.ncbi.nlm.nih.gov/pubmed/19106085]

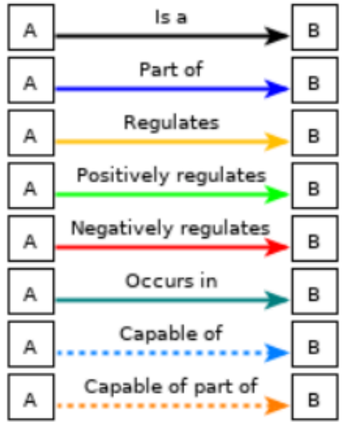
Accession	GO:0009306
Name	protein secretion
Ontology	biological_process
Synonyms	glycoprotein secretion, protein secretion during cell fate commitment, protein secretion resulting in cell fate commitment
Alternate IDs	GO:0045731, GO:0045166
Definition	The controlled release of proteins from a cell. *Source:* GOC:ai
Comment	None
History	See term history for GO:0009306 at QuickGO
Subset	goslim_drosophila
Related	Link to all **genes and gene products** annotated to protein secretion.
	Link to all direct and indirect **annotations** to protein secretion.
	Link to all direct and indirect **annotations download** (limited to first 10,000) for protein secretion.

Accession	GO:0009306
Name	protein secretion
Ontology	biological_process
Synonyms	glycoprotein secretion, protein secretion during cell fate commitment, protein secretion resulting in cell fate commitment
Alternate IDs	GO:0045731, GO:0045166
Definition	The controlled release of proteins from a cell. *Source:* GOC:ai
Comment	None
History	See term history for GO:0009306 at QuickGO
Subset	goslim_drosophila
Related	Link to all **genes and gene products** annotated to protein secretion. Link to all direct and indirect **annotations** to protein secretion. Link to all direct and indirect **annotations download** (limited to first 10,000) for protein secretion.

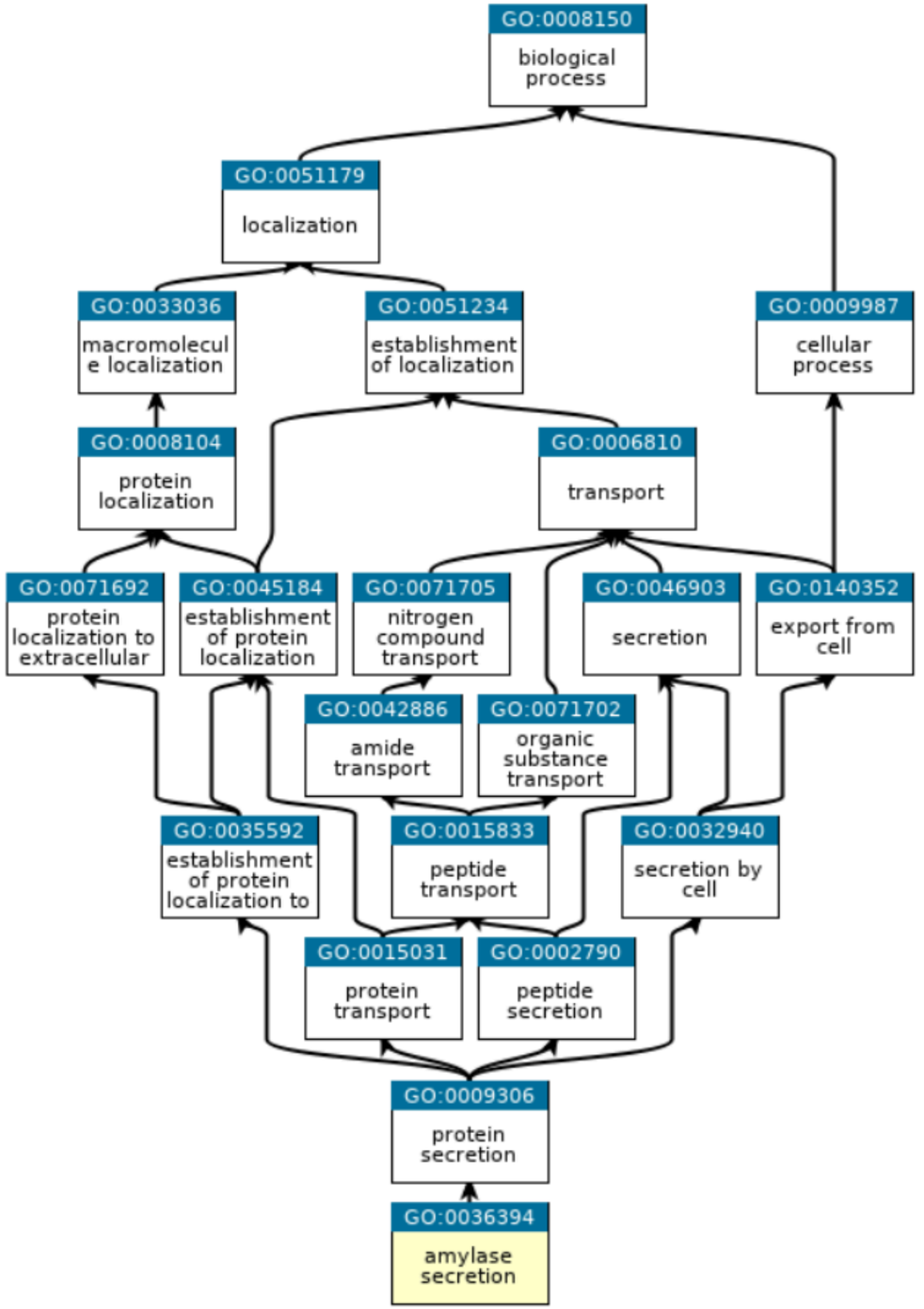
Accession	GO:0036394
Name	amylase secretion
Ontology	biological_process
Synonyms	amylase release
Alternate IDs	None
Definition	The controlled release of amylase from a cell. *Source:* PMID:19028687, GOC:jc
Comment	None
History	See term history for GO:0036394 at QuickGO
Subset	None
Related	Link to all **genes and gene products** annotated to amylase secretion. Link to all direct and indirect **annotations** to amylase secretion. Link to all direct and indirect **annotations download** (limited to first 10,000) for amylase secretion.

more general



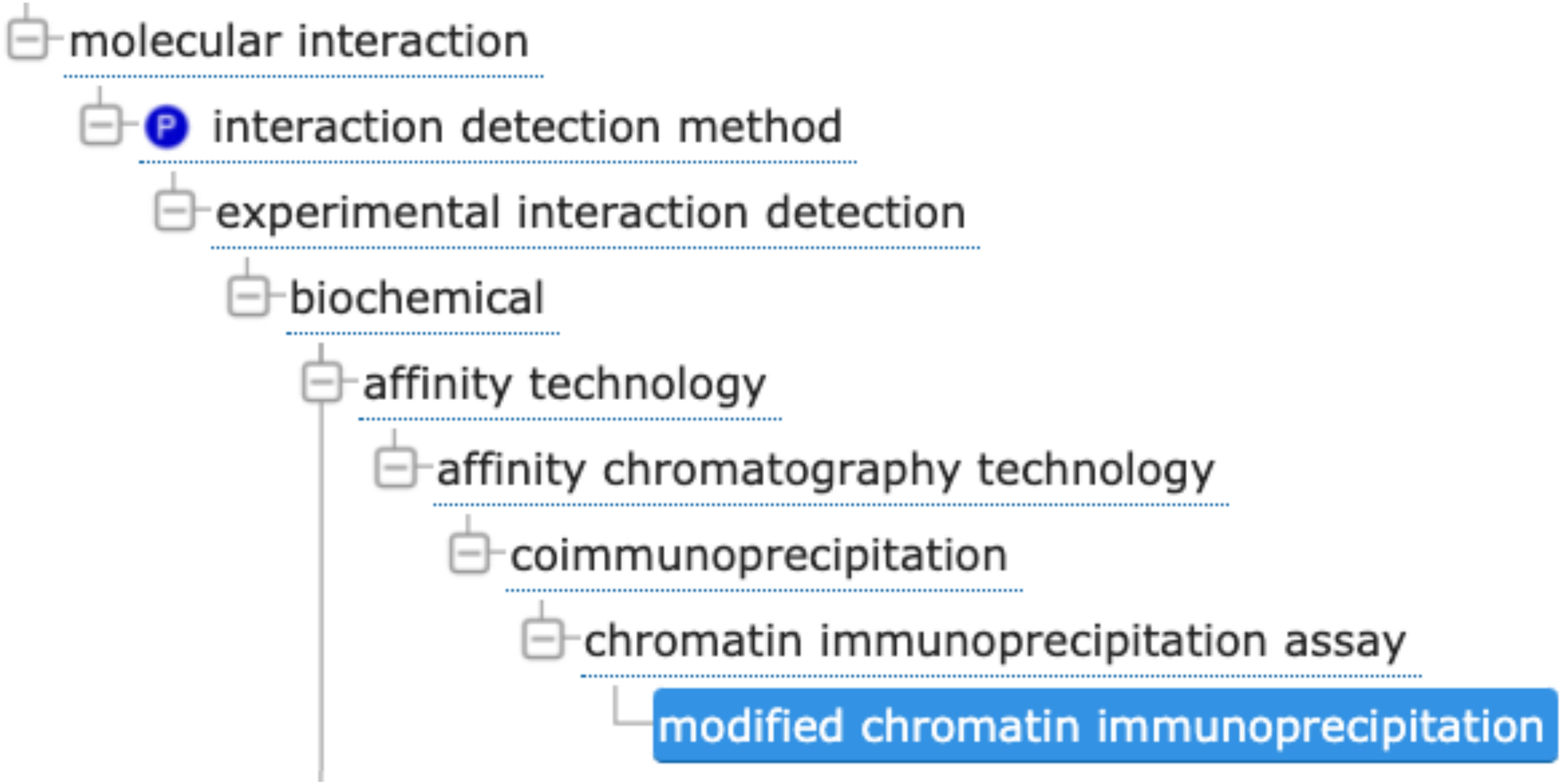


more specific



- GO:0008150 biological_process
- GO:0051179 localization
- GO:0006810 transport
  - GO:0051234 establishment of localization
  - GO:0033036 macromolecule localization
  - GO:0071705 nitrogen compound transport
    - GO:0042886 amide transport
    - GO:0009987 cellular process
    - GO:0071702 organic substance transport
    - GO:0008104 protein localization
      - GO:0045184 establishment of protein localization
      - GO:0140352 export from cell
      - GO:0015833 peptide transport
      - GO:0071692 protein localization to extracellular region
      - GO:0046903 secretion
        - GO:0035592 establishment of protein localization to extracellular region
        - GO:0002790 peptide secretion
        - GO:0015031 protein transport
        - GO:0032940 secretion by cell
          - GO:0009306 protein secretion
            - **GO:0036394 amylase secretion**
              - GO:0036395 pancreatic amylase secretion

http://geneontology.org/



https://www.ebi.ac.uk/ols/ontologies

https://www.ebi.ac.uk/ols/ontologies/mi

https://www.ebi.ac.uk/ols/search?q=chip+seq

Search results for *chip seq*

Previous *Showing **1** to **10** of **665** results* Next

ChIP-seq topic:3169

http://edamontology.org/topic_3169

The analysis of protein-DNA interactions where chromatin immunoprecipitation (ChIP) is used in combination with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins.

Ontology: Bioinformatics operations, data types, formats, identifiers and topics EDAM

ChIP-Seq NCIT:C106049

http://purl.obolibrary.org/obo/NCIT_C106049

A molecular genetic technique that combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to map the binding sites of DNA-associated proteins in a sample of cells. First, crosslinked protein-DNA complexes are isolated using ChIP. Next, the crosslinks are broken, the proteins are removed and the purified DNA is modified with adaptor oligonucleotides to facilitate massively parallel DNA sequencing. Following sequencing, the DNA sequences that are obtained can be mappe…

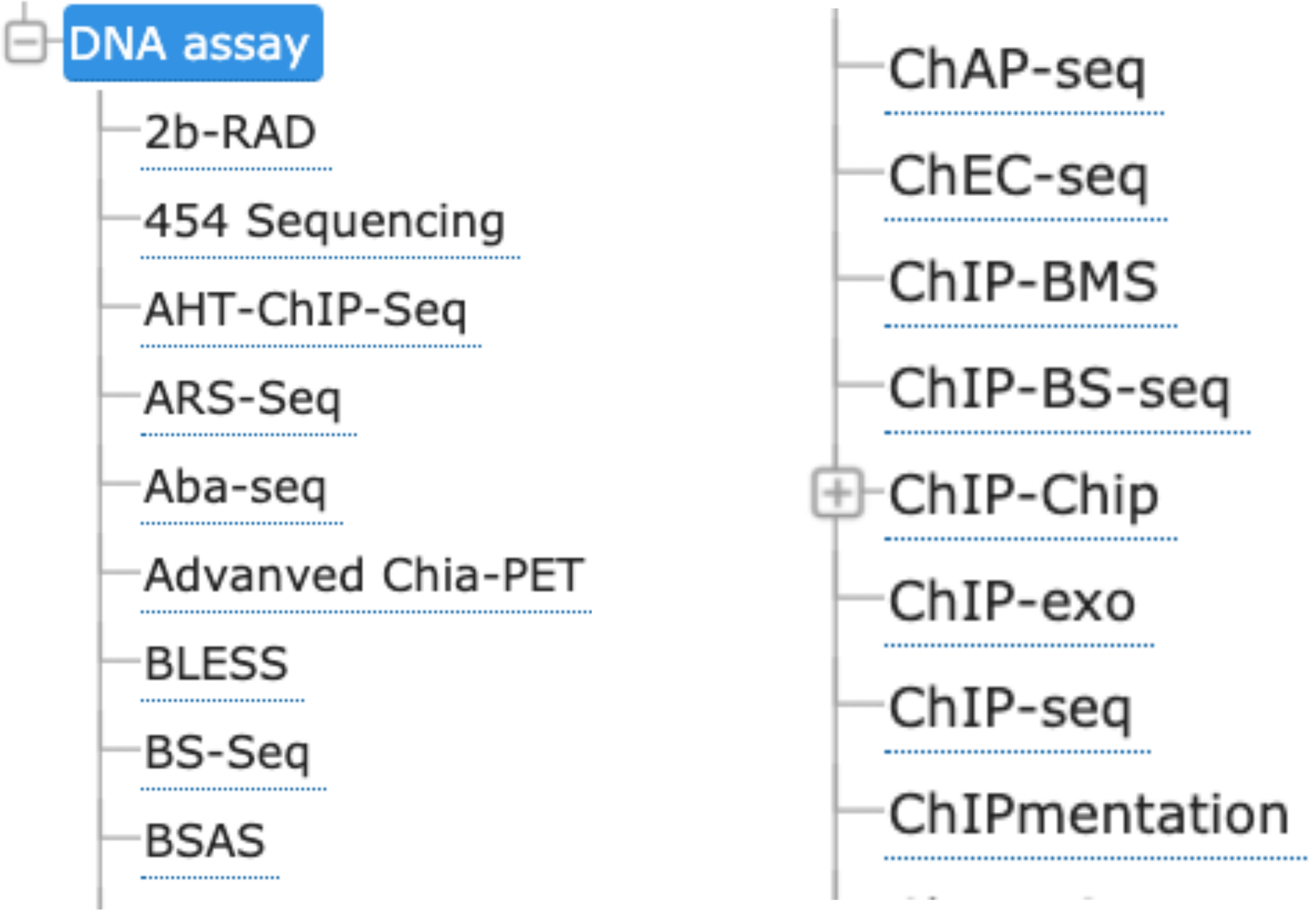
Ontology: NCI Thesaurus OBO Edition NCIT

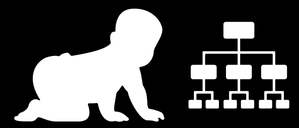
ChIP-seq EFO:0002692

http://www.ebi.ac.uk/efo/EFO_0002692

ChIP-seq is an assay in which chromatin immunoprecipitation with high throughput sequencing is used to identify the cistrome of DNA-associated proteins.

Ontology: Experimental Factor Ontology EFO


DNA assay
2b-RAD
454 Sequencing
AHT-ChIP-Seq
ARS-Seq
Aba-seq
Advanved Chia-PET
BLESS
BS-Seq
BSAS
ChAP-seq
ChEC-seq
ChIP-BMS
ChIP-BS-seq
ChIP-Chip
ChIP-exo
ChIP-seq
ChIPmentation



Species

Experiments

Sequences

Mutations

Interactors

Pathways

Functions

Diseases

DNA

RNA

Proteins

Complexes

Tissues

Cell Types

Organelles

Polymers

Probes &
Drugs

Pathogens

Commensals

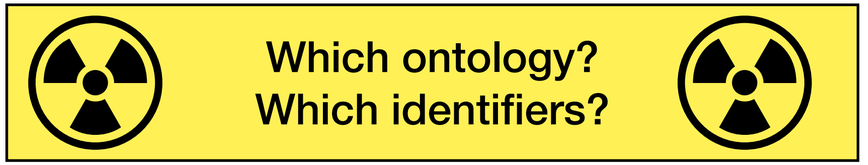
Motifs

Domains

Structures

Ontologies are useful to keep data tidy **and** to learn,
but you probably know them from “enrichment analysis”

I HAVE	I WANT	
A gene	That gene's function	Find that gene's function annotation (try UniProt!)
A function	The genes that do it	Find the function's gene annotation (try geneontology.org)
A group of genes	The functions they most uniquely relate to	Run Gene Ontology Analysis (IPA / PANTHER / EnrichR)



3 skills to survive in the biodatabases landscape

Finding identifiers

Climbing ontologies

Tracking data

From the EBI IntAct Website (molecular interaction database)

[https://www.ebi.ac.uk/intact/]

Molecule 'A'	Links 'A'	Molecule 'B'	Links 'B'	Interaction Detection Method	Publication Identifier	Interaction AC	Source Database
TP53	P04637	p26663-pro_0000037536	P26663-PRO_0000037536	fluorescence microscopy	9188558	EBI-6902998	HPIDb

UniProt ID

UniProt ID with suffix

PSI-MI Ontology ID

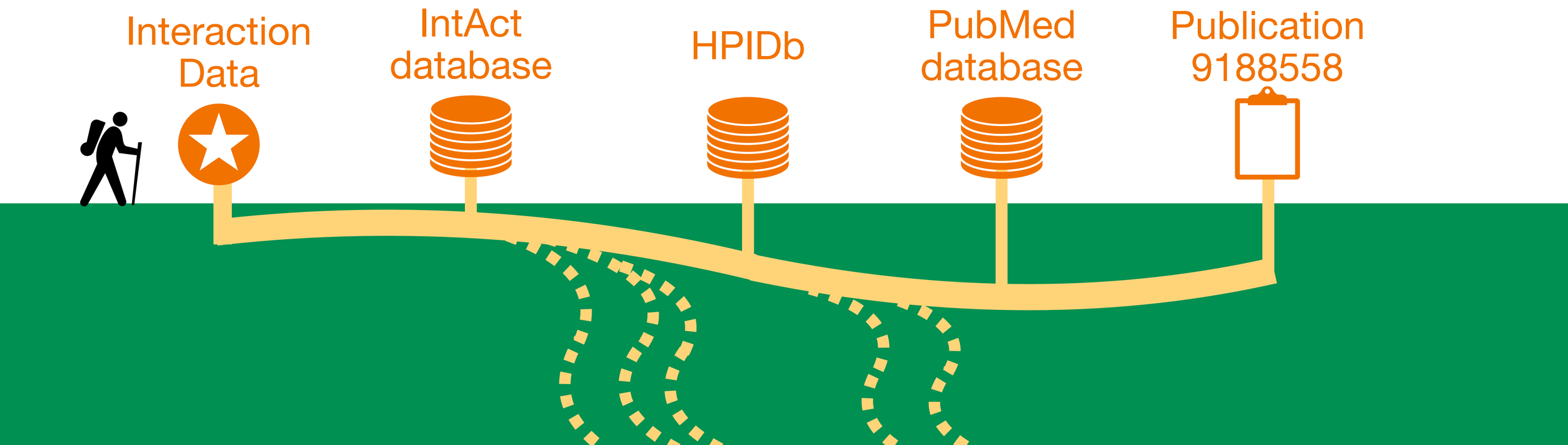
Pubmed ID

IntAct ID

From the EBI IntAct Website (molecular interaction database)

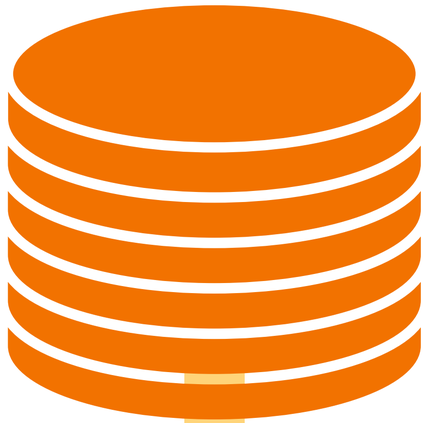
[https://www.ebi.ac.uk/intact/]

Molecule 'A'	Links 'A'	Molecule 'B'	Links 'B'	Interaction Detection Method	Publication Identifier	Interaction AC	Source Database
TP53	P04637	p26663-pro_0000037536	P26663-PRO_0000037536	fluorescence microscopy	9188558	EBI-6902998	HPIDb



IntAct database

WARNING:
Reviewer #3 doesn't know the answer to this, but might have preferences nonetheless.



- What does this database **use**?
- What does this database **produce**?
- When was this database **last updated**?
- **Is it still maintained**?

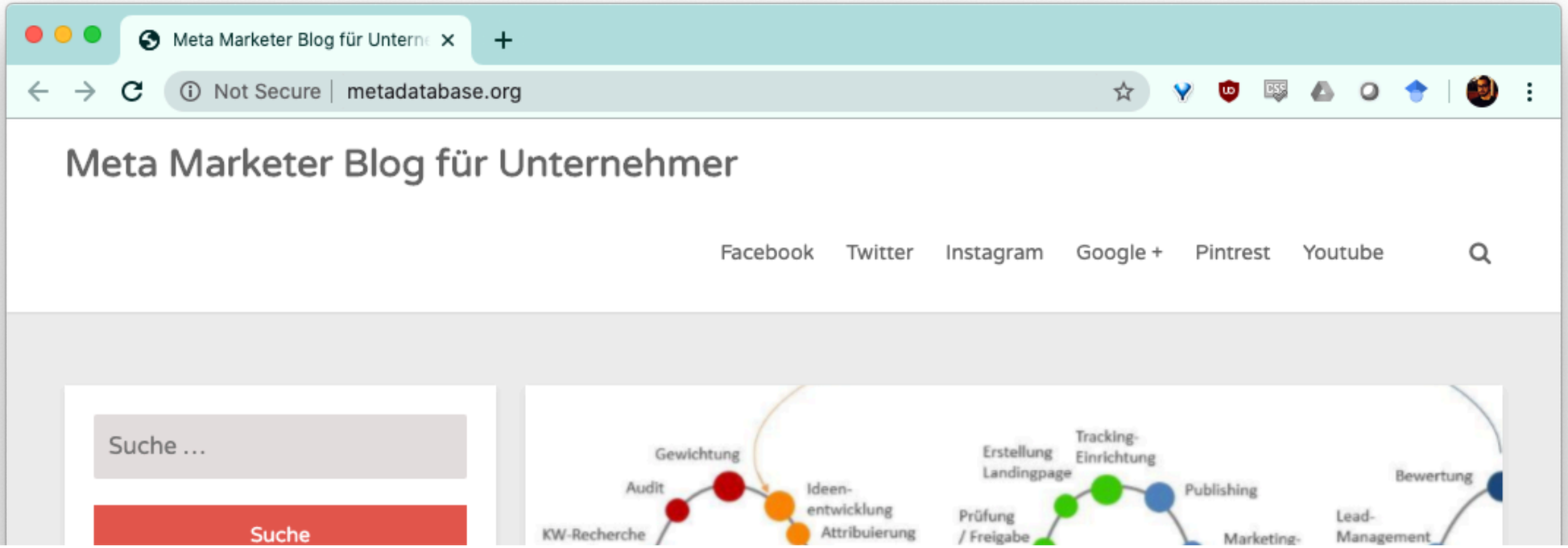
Interaction Data

Nucleic Acids Research

Nucleic Acids Res. 2012 Jan; 40(Database issue): D1250–D1254.
Published online 2011 Dec 1. doi: 10.1093/nar/gkr1099

PMCID: PMC3245051
PMID: 22139927

MetaBase—the wiki-database of biological databases



Finding identifiers

Climbing ontologies

Tracking data

Finding identifiers

Climbing ontologies

Tracking data

And keep a log!

Thank you!

Juan Felipe Beltrán, Ph.D.
Cornell University Department of Biomedical Engineering
juanfelipe@cornell.edu
@offbyjuan

NYU Cardiovascular Research Center