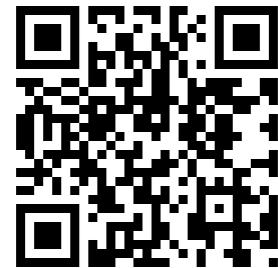


Prof. Dr. Boas Pucker

Pathway Databases

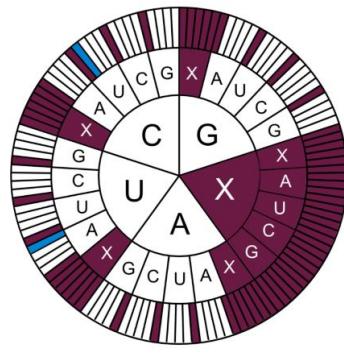
Availability of slides

- All materials are freely available (CC BY) - after the lectures:
 - eCampus: WBIO-A-08
 - GitHub: <https://github.com/bpucker/teaching>
- Questions: Feel free to ask at any time
- Feedback, comments, or questions: [pucker\[a\]uni-bonn.de](mailto:pucker[a]uni-bonn.de)

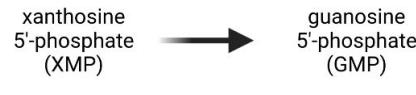


My figures and content can be re-used in accordance with CC BY 4.0, but this might not apply to all images/logos. Some figure were constructed using bioRender.com.

Why do we need pathway databases?



Expanding the genetic code
with an additional DNA base (X)



guanosine monophosphate synthase (GMPS); EC6.3.5.2

<https://www.genome.jp/pathway/ec00230+6.3.5.2>

Identification of candidate enzymes in a **database**



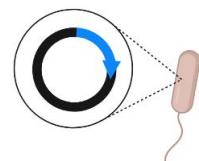
Croton tiglium is naturally producing a precursor of this unnatural DNA base



RNA extraction, sequencing,
and construction of a
transcriptome assembly



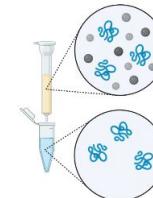
Cloning of GMPS coding sequence into a plasmid



Transformation into *Escherichia coli*



Heterologous expression of *GMPS*

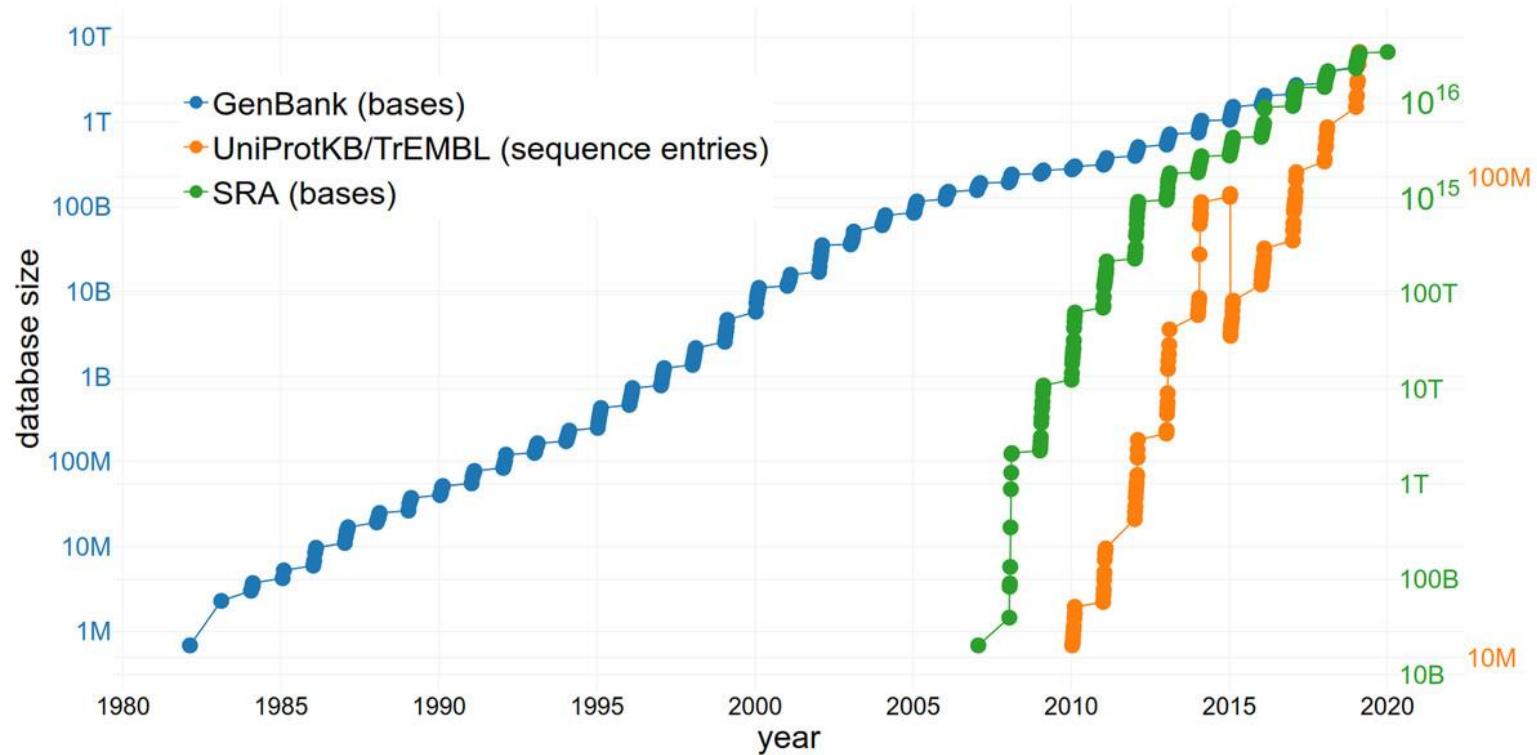


Protein purification



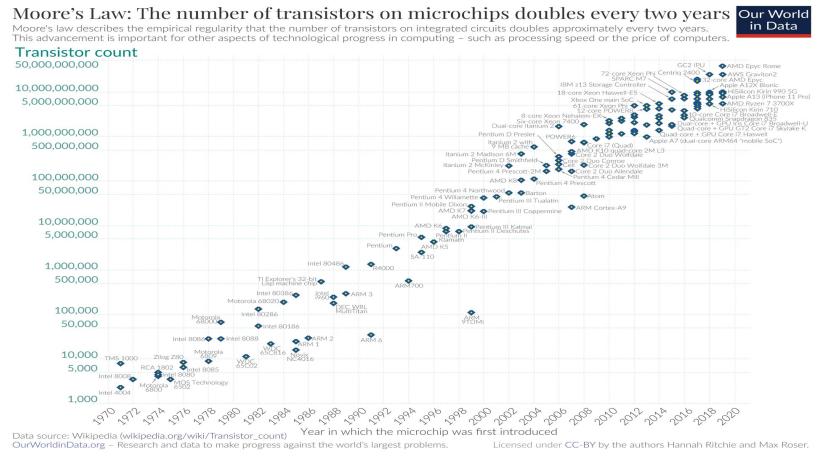
Enzyme assay

Exponential growth of databases



Generation of large data sets

- Sequencing technologies are outpacing Moore's law
 - Moore's law = number of transistors on microchips doubles every two years
 - Sequencing is becoming widely available (democratization of sequencing)
 - Proteomics and metabolomics are developing rapidly



Types of pathway databases

- Reactions (KEGG, BioCyc, MetaCyc, Reactome)
 - Enzymes (BRENDA)
 - Transcription factors (JASPAR)
 - Metabolites (MetaboLights)
 - Species specific databases (TAIR, Phytozome, PLAZA)

How to access data?

- Open Data = Initiative to make research data publicly available
- FAIR data = Findable Accessible Interoperable Reuseable data
 - Findable = registered in a central database
 - Accessible = available through a repository (not just from authors)
 - Interoperable = data format allows automatic processing
 - Reusable = sufficient metadata to use data sets in other studies
- ORCID is used as a universal login solution across many platforms
 - ORCID = Open Researcher and Contributor ID
 - SSO = Single Sign-On (user stays logged in across websites)

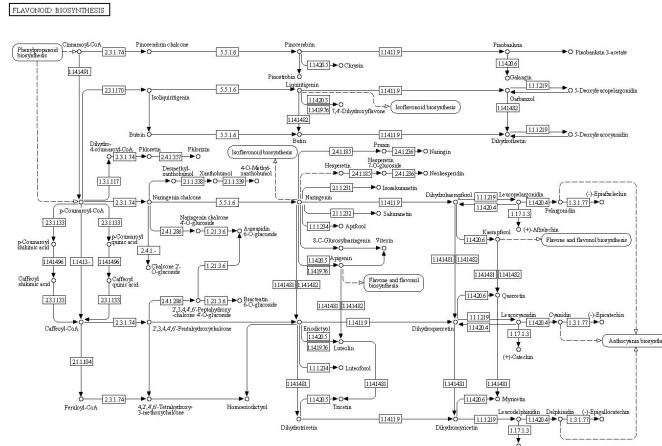


<https://www.go-fair.org/>



<https://orcid.org/>

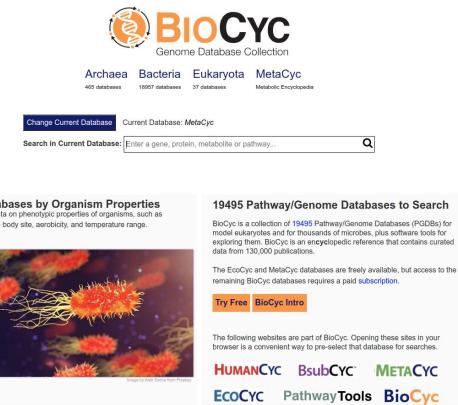
- Maps of pathways showing the individual reactions with catalyzing enzymes
- Information about genomes and genes
- Chemical details about enzymes, substrates, and products
- KEGG is financed through a subscription model (for FTP download), but website is freely accessible



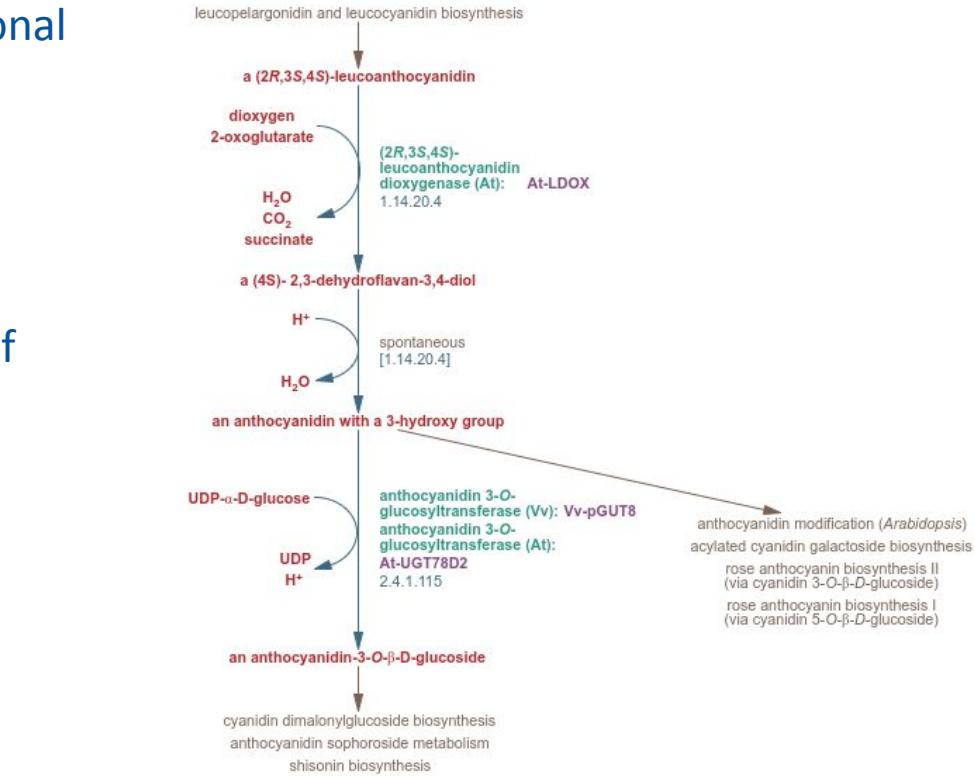
map00941 Flavonoid biosynthesis

BioCyc: Collection of databases

- Integration of databases and tools for omics
- Curation is involved to increase the quality
- Strong focus on microorganism:
 - EcoCyc = **Escherichia coli Encyclopedia**
 - BsubCyc = **Bacillus subtilis Encyclopedia**
 - YeastCyc = **Yeast Encyclopedia**
- BUT:
 - Search for metabolites
 - Search for pathways
 - Search for reactions



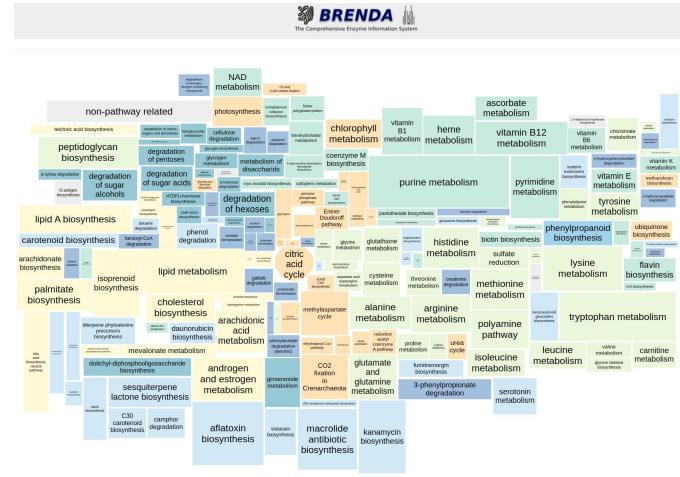
- Integrates genomic data with functional annotation
- Visualization of pathway databases
- Shows intermediates and enzymes of biosynthesis pathways
- MetaFlux: flux-balance analysis



Enzyme Commission (EC) numbers

- Four digit system to classify enzymes:
 - General type of reaction
 - Subclass
 - Sub-subclass
 - Serial number (within sub-subclass)
- General types:
 - EC1 = Oxidoreductases
 - EC2 = Transferases
 - EC3 = Hydrolases
 - EC4 = Lyases
 - EC5 = Isomerases
 - EC6 = Ligases

- Enzyme database hosted at TU Braunschweig (BRICS)
- Text and structure-based queries
- Visualization of pathways
- Manual curation of datasets
- Many details about enzyme properties (substrates, kinetics, mutants, ...)



- DFR (dihydroflavonol 4-reductase) is a central enzyme in the anthocyanin biosynthesis
- BRENDA provides information about reaction mechanism and species-specific parameters
- Enzymatic properties of multiple species allow quick comparison

Information on EC 1.1.1.219 - dihydroflavonol 4-reductase and Organism(s) *Arabidopsis thaliana*

for references in articles please use BRENDA:EC1.1.1.219

EC Tree

- └ 1 Oxidoreductases
 - └ 1 Acting on the CH-OH group of donors
 - └ 1.1.1 With NAD⁺ or NADP⁺ as acceptor
 - └ 1.1.1.219 dihydroflavonol 4-reductase

IUBMB Comments

This plant enzyme, involved in the biosynthesis of anthocyanidins, is known to act on (+)-dihydrokaempferol, (+)-taxifolin, and (+)-dihydromyricetin, although some enzymes may act only on a subset of these compounds. Each dihydroflavonol is reduced to the corresponding cis-flavan-3,4-diol. NAD⁺ can act instead of NADP⁺, but more slowly.

Specify your search results

Mark a special word or phrase in this record:

Search Reference ID:

Search UniProt Accession:

Select one or more organisms in this record:

All organisms
 Allium cepa
 Anthurium andraeanum
 Arabidopsis thaliana
 Brassica napus
 Brassica napus

This record set is specific for:

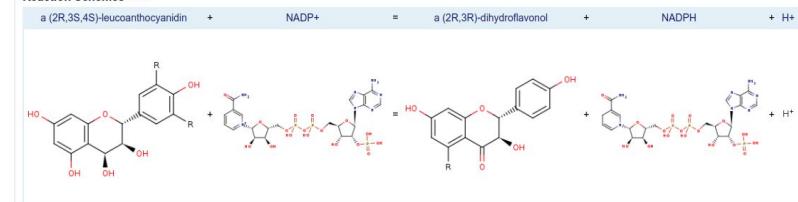
Arabidopsis thaliana

Word Map hide



The taxonomic range for the selected organisms is: *Arabidopsis thaliana*
 The expected taxonomic range for this enzyme is: Bacteria, Eukaryota, Archaea

Reaction Schemes hide



Synonyms

Bo9g058630, BrDFR1, BrDFR10, BrDFR11, BrDFR12, BrDFR2, BrDFR3, BrDFR4, BrDFR5, BrDFR6, BrDFR7, more

BRENDA - example (2)

- Specific substrates and products are listed
- K_M and V_{max} values are included
- References point to publications and other databases

Go to Substrate/Product Search

| SUBSTRATE ▾ | PRODUCT ▾ | REACTION DIAGRAM | ORGANISM ▾ | UNIPROT ▾ | COMMENTARY (Substrate) ▾ | X LITERATURE (Substrate) ▾ | COMMENTARY (Product) ▾ | X LITERATURE (Product) ▾ | Reversibility |
|--|--------------------------|------------------|----------------------|-----------|------------------------------|----------------------------|------------------------|--------------------------|-----------------|
| dihydrokaempferol + NADPH + H ⁺ | leucocyanidin + NADP+ | ◇ | Arabidopsis thaliana | P51102 | - | 702000 | - | - | ? reversible |
| dihydromyricetin + NADPH + H ⁺ | leucodelphinidin + NADP+ | ◇ | Arabidopsis thaliana | P51102 | - | 702000 | - | - | ? irreversible |
| dihydroquercetin + NADPH + H ⁺ | leucocyanidin + NADP+ | ◇ | Arabidopsis thaliana | P51102 | - | 702000 | - | - | ? not specified |
| additional information | 2 | - | Arabidopsis thaliana | P51102 | does not catalyze naringenin | 702000 | - | - | ? |

Go to Collector Search

| COFACTOR ▾ | ORGANISM ▾ | UNIPROT ▾ | COMMENTARY ▾ | X LITERATURE ▾ | IMAGE ▾ |
|------------|----------------------|-----------|--------------|----------------|---------|
| NADPH | Arabidopsis thaliana | P51102 | - | 702000 | ◇ |

Go to Specific Activity Search

| SPECIFIC ACTIVITY ▾ | UNIPROT ▾ | COMMENTARY ▾ | X LITERATURE ▾ |
|---------------------|----------------------|--------------|-------------------------------------|
| 0.00032 | Arabidopsis thaliana | P51102 | with dihydrokaempferol as substrate |
| 0.0006 | Arabidopsis thaliana | P51102 | with eriodictyol as substrate |
| 0.001 | Arabidopsis thaliana | P51102 | with dihydromyricetin as substrate |
| 0.0013 | Arabidopsis thaliana | P51102 | with dihydroquercetin as substrate |

Go to Organism Search

| ORGANISM ▾ | COMMENTARY ▾ | X LITERATURE ▾ | UNIPROT ▾ | SEQUENCE DB ▾ | SOURCE ▾ |
|----------------------|--------------|----------------|-----------|---------------|----------|
| Arabidopsis thaliana | - | 702000 | P51102 | UniProt | BRENDA |

Go to General Information Search

| GENERAL INFORMATION ▾ | ORGANISM ▾ | UNIPROT | COMMENTARY ▾ | X LITERATURE ▾ |
|------------------------|----------------------|---------|--|----------------|
| physiological function | Arabidopsis thaliana | P51102 | DFR plays a key role in determining intensity and pigment coloration because its specificity and activities dictate the type and amount of the colorless leucoanthocyanidins | 702000 |

Go to Sequence and Transmembrane Helices Search

| UNIPROT ▾ | ENTRY NAME ▾ | ORGANISM ▾ | NO. OF AA ▾ | NO. OF TRANSM. HELICES ▾ | MOLECULAR WEIGHT(Da) ▾ | SOURCE ▾ | SEQUENCE ▾ | LOCALIZATION PREDICTION ▾ | X LITERATURE ▾ |
|-----------|--------------|----------------------|-------------|--------------------------|------------------------|------------|---------------|---------------------------------|----------------|
| P51102 | DFRA_ARATH | Arabidopsis thaliana | 382 | 0 | 42775 | Swiss-Prot | Show Sequence | other Location (Reliability: 5) | 702000 |

Go to Cloned (Commentary) Search

| CLONED (Commentary) ▾ | ORGANISM ▾ | UNIPROT ▾ | LITERATURE ▾ |
|--|----------------------|-----------|--------------|
| into pT7Blue-2TOPo and heterologously expressed in Escherichia coli TOP10F strain; DFR cDNA cloned into pRSF-HFT and inserted into Escherichia coli BL21Star to create E-coli strain | Arabidopsis thaliana | P51102 | 702000 |

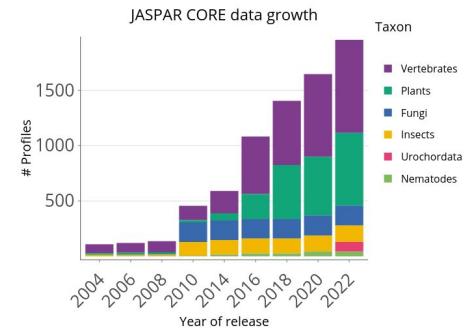
Select item on the left to see more content.

ExternalLinks (specific for EC Number 1.1.1.219)

- ExplorEnz (official portal for IUBMB Enzyme Nomenclature)
- Enzyme Nomenclature Database
- KEGG
- MetaCyc
- SABIO-RK
- NCBI: PubMed, Protein, Nucleotide, Structure, Gene, OMIM
- Enzyme Nomenclature (alternative site)
- UnProt
- PDB
- PROSITE Database of protein families and domains
- InterPro (database of protein families, domains and functional sites)

JASPAR: THE transcription factor database

- Contains motifs of transcription factor binding sites
- Additional details about transcription factors
- Most comprehensive database
- Regularly updated



Profile summary Add

| | |
|-------------|-----------------------------|
| Name: | MYB124 |
| Matrix ID: | MA1426.1 |
| Class: | Tryptophan cluster factors |
| Family: | Myb |
| Collection: | CORE |
| Taxon: | Plants |
| Species: | <i>Arabidopsis thaliana</i> |
| Data Type: | PBM |
| Validation: | 20675570 |
| Uniprot ID: | Q94FL6 |
| Source: | 31133749 |
| Comment: | |

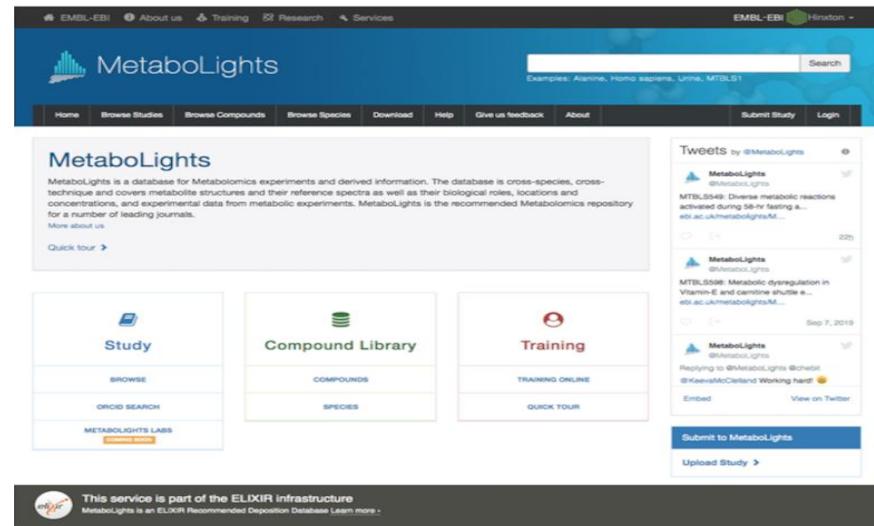
Sequence logo Download SVG



Frequency matrix

| | A | C | G | T | |
|-----|-----|-----|-----|-----|---------------|
| A [| 324 | 200 | 546 | 40 |] |
| C [| 230 | 284 | 266 | 799 | 0 |
| G [| 221 | 183 | 178 | 124 | 963 |
| T [| 223 | 330 | 8 | 36 | 0 |
| | | | | | Reverse comp. |
| | 1 | 0 | 999 | 0 | 718 |
| | 61 | 0 | 680 | 119 | 159 |
| | 80 | 0 | 318 | 100 | 197 |
| | 271 | 0 | 318 | 100 | 309 |

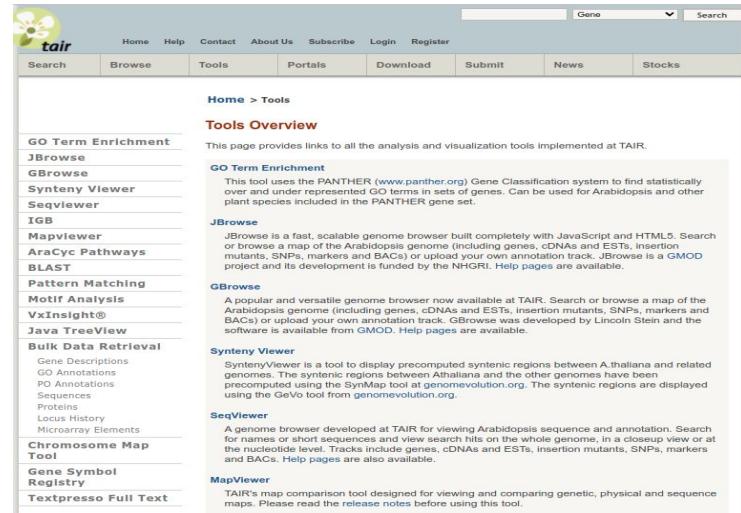
- Database of metabolomics experiments
- Hosted at European Bioinformatics Institute (UK)
- Collaboration with journals to enforce data submission



The screenshot shows the homepage of the MetaboLights database. At the top, there is a navigation bar with links to EMBL-EBI, About us, Training, Research, Services, Home, Browse Studies, Browse Compounds, Browse Species, Download, Help, Give us feedback, About, Submit Study, and Login. Below the navigation bar is the MetaboLights logo and a search bar. The main content area features a brief introduction to the database, a "Quick tour" link, and three large buttons for "Study", "Compound Library", and "Training". The "Study" button has sub-links for "BROWSE", "ORCID SEARCH", and "METABOLIGHTS LABS". The "Compound Library" button has sub-links for "COMPOUNDS" and "SPECIES". The "Training" button has sub-links for "TRAINING ONLINE" and "QUICK TOUR". On the right side of the page, there is a sidebar titled "Tweets by @Metabolights" showing recent tweets from the official account. At the bottom, there is a footer note stating "This service is part of the ELIXIR infrastructure" and "MetaboLights is an ELIXIR Recommended Deposition Database Learn more".

Species specific databases: TAIR

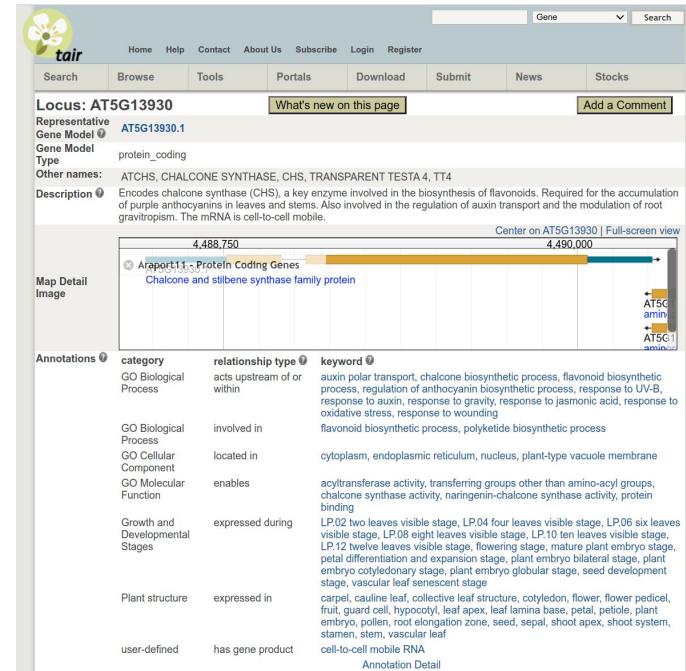
- TAIR is located at Phoenix Bioinformatics
- Institutional subscriptions provide funding for maintenance
- Full access requires subscription, but limited number of visits is free
- Collection of data sets and tools (e.g. BLAST)
- Current version: TAIR10 (=2010)
- Substitution by Araport11 was started, but is no longer funded



The screenshot shows the TAIR Tools page. At the top, there's a navigation bar with links for Home, Help, Contact, About Us, Subscribe, Login, and Register. Below that is a search bar with dropdown menus for Gene and Search. A sidebar on the left lists various tools: GO Term Enrichment, JBrowse, GBrowse, Synteny Viewer, Seqviewer, IGB, Mapviewer, AraCyc Pathways, BLAST, Pattern Matching, Motif Analysis, VxInsight®, Java TreeView, Bulk Data Retrieval, Gene Descriptions, GO Annotations, PO Annotations, Sequences, Proteins, Locus History, Microarray Elements, Chromosome Map Tool, Gene Symbol Registry, and Textpresso Full Text. The main content area shows a "Tools Overview" section with a brief description of the GO Term Enrichment tool, which uses the PANTHER system to find statistically overrepresented GO terms in gene sets. It also describes JBrowse, GBrowse, Synteny Viewer, Seqviewer, and MapViewer, along with their respective descriptions and links.

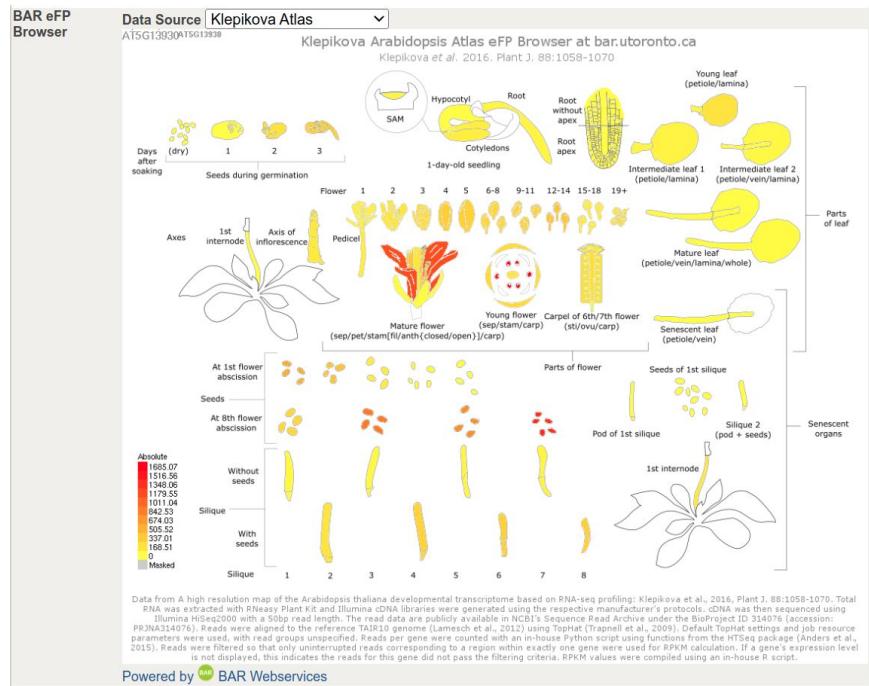
TAIR10 - functional annotation

- TAIR10 provides information about function of genes
- References to other database for additional information
- Lists of publications about a given gene
- *Arabidopsis thaliana* gene functions are transferred to genes of other species



TAIR10 - eFP Browser

- Gene expression analysis is integrated in TAIR
 - eFP browser provides expression data
 - eFP = electronic Fluorescent Pictograph
 - Microarray and RNA-Seq data sets are basis of gene expression analysis



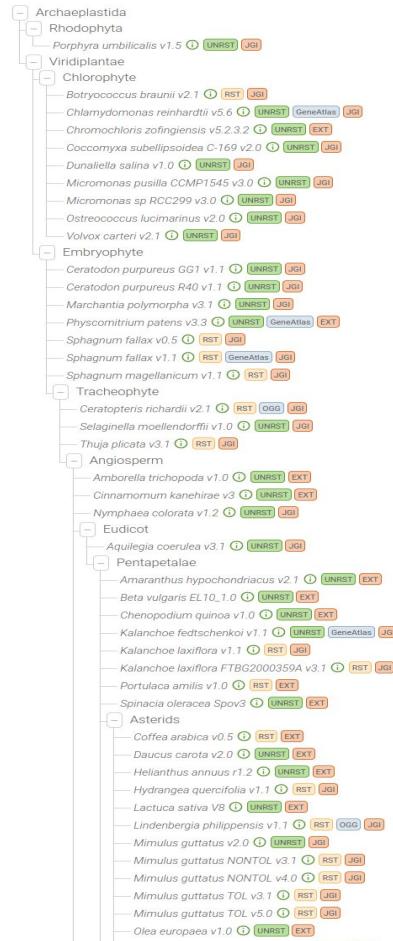
TAIR10 - references

- Aggregates all references reporting about a gene
- Connection of knowledge
- AGI = Arabidopsis Gene Identifier
- AGIs mentioned in publication are used to assign studies to genes

| Publication | author/title | source | associated loci | date |
|---|---|--|-----------------|------|
| Dahhan, D.A., Reynolds, G.D., Car... Proteomic characterization of isolated Arabidopsis callose-coated vesicles reveals many highly conserved and plant-specific components | THE PLANT CELL | AT2G46830 AT3G01780 AT3G51240 AT5G18930 AT5G81380 | | 2022 |
| Verma, D., Neugart, B.K., Saito, A.K... A blue-light-specific phosphatase, MAP kinase phosphatase1, positively regulates blue light-mediated seedling development in Arabidopsis thaliana | PLANTA | AT1G29930 AT1G67099 AT2G04550 AT2G43790 AT2G46340 AT3G06110 AT3G23610 AT3G55270 AT4G36730 AT5G11260 AT5G13930 AT5G23720 AT5G24120 AT5G40440 | | 2021 |
| Kislevy, K.V., Sunun, A.R., Aleya... External dsRNA Downregulates Anthocyanin Biosynthesis-Related Genes and Affects Anthocyanin Accumulation in Arabidopsis thaliana | INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES | AT1G71030 AT1G77450 AT5G13930 | | 2021 |
| Markus, C., Pećina, A., Merot, A... Insights into the Role of Transcriptional Gene Silencing in Response to Chemical treatments in Arabidopsis thaliana | INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES | AT2G36490 AT5G13930 | | 2021 |
| Zhang, X., He, Y., Li, L., Liu, H., Hon... Involvement of the R2R3-MYB transcription factor AtR2R3-MYB21 and its homologs in regulating the stilbenoid flavonols accumulation in Arabidopsis | JOURNAL OF EXPERIMENTAL BOTANY | AT1G30135 AT2G47460 AT3G01530 AT3G27810 AT3G62610 AT5G08640 AT5G13930 AT5G40348 AT5G40350 AT5G49330 FL51 | | 2021 |
| Yu, Z.G., Zhou, X.T., Lin, W., He,... Photoprotection of Arabidopsis leaves under short-term high light treatment: the antioxidant capacity is more important than anthocyanin accumulation effect | PLANT PHYSIOLOGY AND BIOCHEMISTRY | AT1G17610 AT4G16860 AT4G22880 AT5G13930 | | 2021 |
| Chapman, J.M., Muday, G.K... Flavonols modulate lateral root emergence by scavenging reactive oxygen species in Arabidopsis thaliana | JOURNAL OF BIOLOGICAL CHEMISTRY | AT5G07990 AT5G08640 AT5G13930 | | 2020 |
| Gaynorba, S.B., Muday, G.K... Flavonols regulate root hair development by modulating reactive oxygen species in the root epidermis | DEVELOPMENT | AT5G13930 AT5G51060 | | 2020 |
| Weng, C.Y., Zhu, M.H., Liu, Z.Q., Z... Integrated bioinformatics analyses identified SCL3-induced regulatory network in Arabidopsis thaliana roots | BIOTECHNOLOGY LETTERS | AT1G50420 AT3G51240 AT3G55120 AT5G05270 AT5G13930 | | 2020 |
| Naake, T., Maeda, H.A., Proost, S.,... Kingdom-wide analysis of the evolution of the plant type III polyketide synthases and their family | PLANT PHYSIOLOGY | AT1G02050 AT4G34850 AT5G13930 | | 2020 |
| Hanamouchi, N., Yinkel, B.S., Li, J.J., Li... Modulation of Arabidopsis Flavonol Biosynthesis Genes by Cyst and Root-Knot Nematodes | PLANTS (BASEL) | AT2G47460 AT5G08640 AT5G13930 | | 2020 |
| Chen, J., Wu, J., Chen, M., Cui, M., Kong,... Chitin synthase genes involved in vegetative growth, asexual reproduction and pathogenesis of Phytophthora capsici and <i>P. sojae</i> | ENVIRONMENTAL MICROBIOLOGY | AT2G33580 AT3G21630 AT5G13930 | | 2019 |
| Nakabayashi, T., Nakabayashi, T., Nakabayashi, T... Formation of Flavonoid Metabolites: Functional Significance of Protein-Protein Interactions and Impact on Flavonoid Chemosynthesis | FRONT PLANT SCI | AT5G13930 | | 2019 |
| Rai, N., Neugart, S., Yan, Y., Wang,... How do cryptochromes and UV-B interact in natural and simulated sunlight? | JOURNAL OF EXPERIMENTAL BOTANY | AT5G13930 AT5G63860 | | 2019 |

View Complete List (15 of 96 displayed)

- Operated by the Joint Genome Institute (JGI), USA
- Collection of plant genome sequences and the corresponding annotations (best annotation source!)
- Data sets are available for download
- V13 contains 261 genome assemblies
- References of all data sets are clearly presented



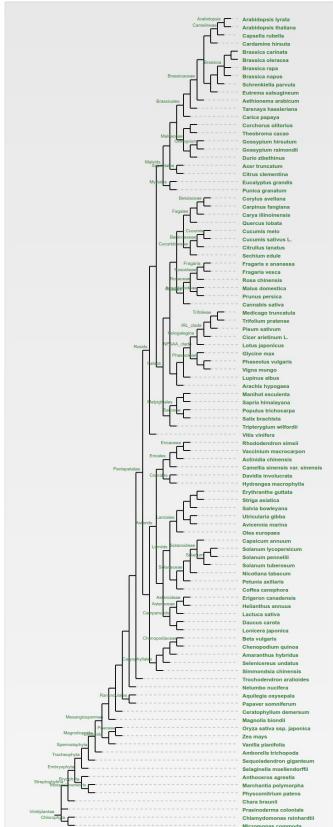
- Hosted in Belgium (Department of Plant Biotechnology and Bioinformatics); partly funded by ELIXIR
- Collection of plant genome sequences and corresponding annotations
- Focus on comparative genomics
- PLAZA v5 contains 134 genome sequences and their annotations

News

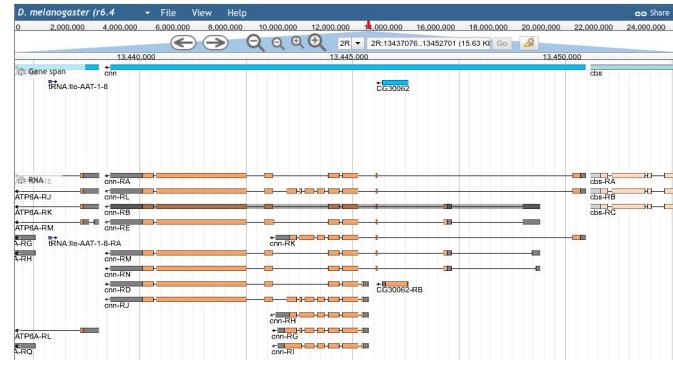
• Public release PLAZA 3.0 (2014-11)

Dicots PLAZA 5.0 summary

- Integration of structural and functional annotation of 100 dicots
- Includes c. 254,318 genes, of which 88.6% are protein coding. These protein coding genes are clustered in 69,306 multi-gene gene families (47.5% multi-species gene families).
- More information on the data content can be found in the data overview



- Website of the Drosophila research community
- BLAST service
- JBrowse = Web service to inspect the genome sequence
- Gene expression data; Data downloads
- Information about meetings, job opportunities, and courses



- Website of the *Cenorhabditis elegans* community
- Tools and resources are similar to FlyBase:
 - Option to run BLAST
 - Check of primers (e-PCR)
 - References to publications
 - Community news

e-PCR Search

This search uses Greg Schuler's e-PCR [program](#) to search for STSs in the current *C. elegans* genome.

You may also consider alternative tools to e-PCR, such as [UCSC In-Silico PCR](#) [and Primer-BLAST](#).

Enter a list of primer pairs to find using the following format:

| Name | Left Oligo | Right Oligo | Length |
|---------|--------------------------|-------------------------|--------|
| assay_1 | CGATAAACAACTCAACGGCATAAT | TTTGAACACTGATATAGAGGGCA | 1188 |
| assay_2 | AAGGTTATTATGCGGTGAAAT | AGCACTTGGAGCTTGATGAAATC | 2191 |
| assay_3 | AGATTGGAACGATAACGCAGATA | TTTGCCAATTTCGATTTTTTTT | 1603 |

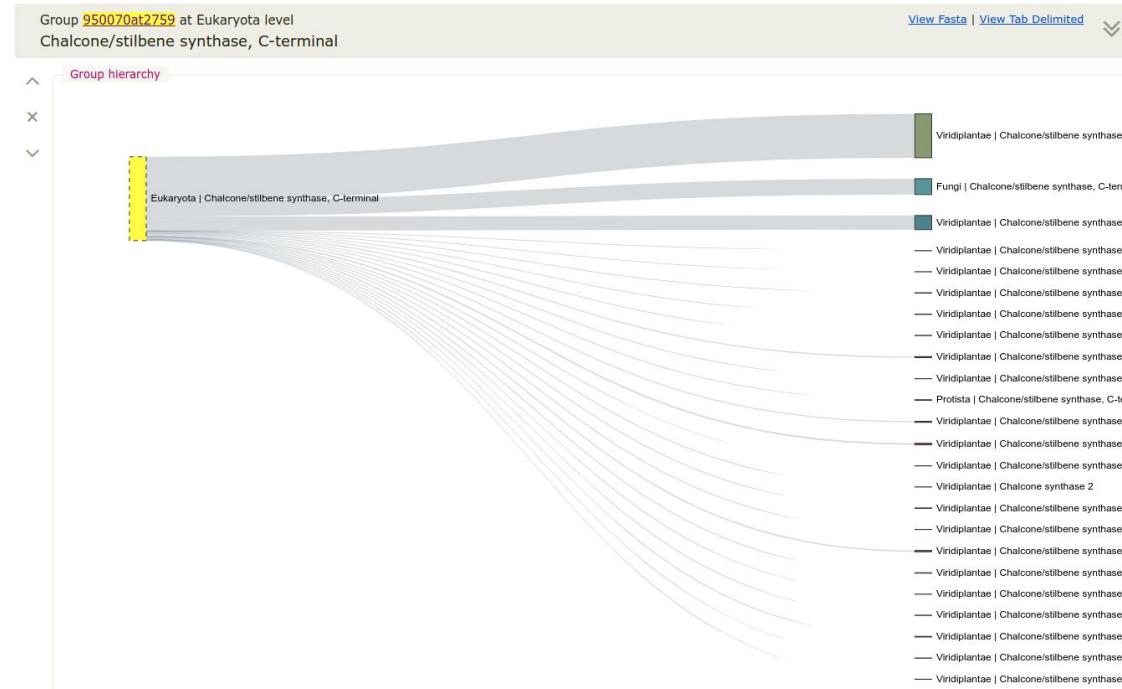
Alternatively, select [List of assay names](#) and enter a list of PCR product names already in WormBase to see what genes they intersect.

e-PCR Search

This is a List of primer pairs List of assay names
Allow product length difference of: bp Allow oligo mismatches of: bp
Results in: Text-only format HTML with hyperlinks

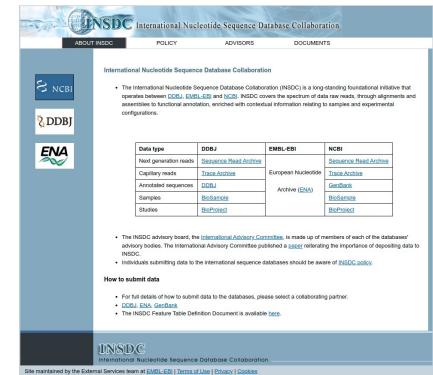
This may take several minutes to run. Hit the stop or back buttons to abort.

- Database of similar genes in other species



INSDC - a collaborative effort

- INSDC = International Nucleotide Sequence Database Collaboration
- Three databases are mirrored i.e. constantly updated
 - NCBI = National Center for Biotechnology Information (USA)
 - ENA = European Nucleotide Archive (Europe)
 - DDBJ = DNA DataBase of Japan (Japan)
- Users can search in any of the three databases and get the same results



The screenshot shows the homepage of the INSDC website. At the top, there's a navigation bar with links for 'ABOUT INSDC', 'POLICY', 'ADVISORY', and 'DOCUMENTS'. Below the navigation is a banner for 'International Nucleotide Sequence Database Collaboration' featuring logos for NCBI, DDBJ, and ENA. The main content area has a heading 'International Nucleotide Sequence Database Collaboration' and a bulleted list:

- The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing international initiative that operates between NCBI, EMBL-EBI and JGC. INSDC covers the spectrum of data from reads, through fragments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations.

Below this is a table comparing the three databases:

| Data type | NCBI | EMBL-EBI | JGC |
|-----------------------|------------------------|-----------------------------------|------------------------|
| Next generation reads | Sequence Read Archives | Sequence Read Archives | Sequence Read Archives |
| Captivity reads | Trace Archive | European Nucleotide Archive (ENA) | Trace Archive |
| Annotated sequences | DDBJ | GenBank | GenBank |
| Samples | Bioprojects | Bioprojects | Bioprojects |
| Studies | BioStudies | bioRxiv | BioStudies |

Below the table, there's a section titled 'How to submit data' with instructions:

- For full details on how to submit data to the databases, please select a collaborating partner: [DDBJ](#), [EBI](#), [GenBank](#).
- The INSDC Feature Table Definition Document is available [here](#).

At the bottom, it says 'Site maintained by the External Services team at [EMBL-EBI](#) | [Terms of Use](#) | [Privacy](#) | [Cookies](#)'.

- NCBI operates a large number of databases and services
- Famous service: BLAST
- Important databases:
 - PubMed: scientific publications
 - GenBank: individual sequences
 - Gene Expression Omnibus (GEO): gene expression data sets
 - Sequence Read Archive (SRA): sequencing data

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit

Deposit data or manuscripts into NCBI databases



Download

Transfer NCBI data to your computer



Learn

Find help documents, attend a class or watch a tutorial



Develop

Use NCBI APIs and code libraries to build applications



Analyze

Identify an NCBI tool for your data analysis task



Research

Explore NCBI research and collaborative projects

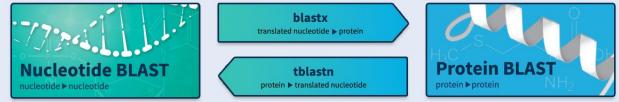


- BLAST = Basic Local Alignment Search Tool
- Probably the most famous website of the NCBI
- Comparison of sequences against a large database
- Numerous variants of the initial BLASTn were developed

Basic Local Alignment Search Tool
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS
BLAST+ 2.13.0 is here!
Starting with this release, we are including the blastn_vdb and blastn_vdb executables in the BLAST+ distribution.
Thu, 17 Mar 2022 12:00:00 EST [More BLAST news...](#)

Web BLAST



BLAST Genomes
Enter organism common name, scientific name, or tax id [Search](#)
Human Mouse Rat Microbes

Standalone and API BLAST

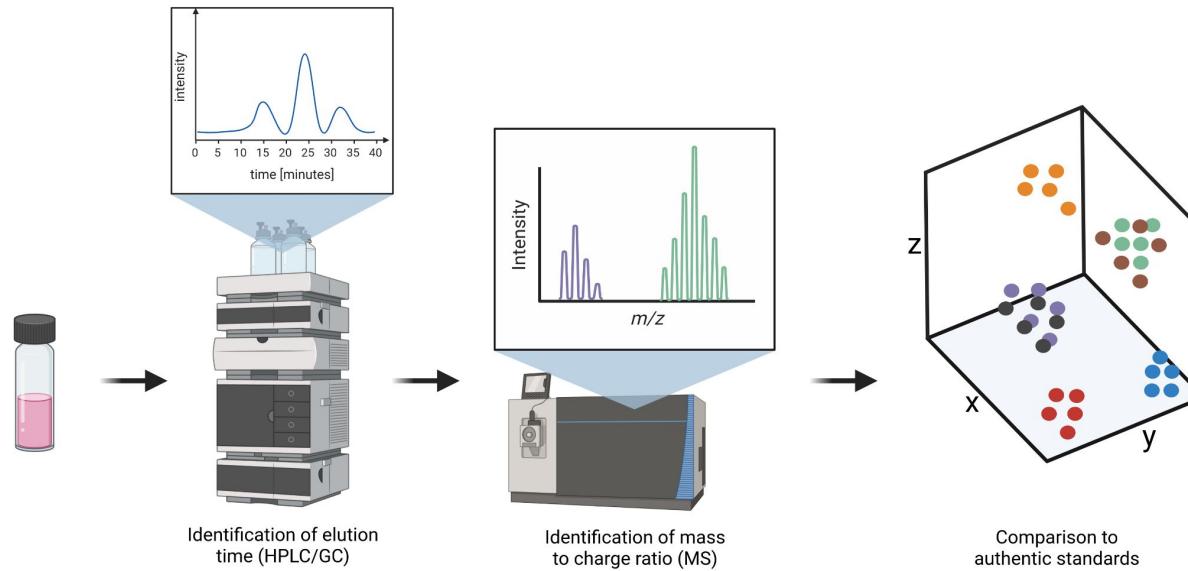


Specialized searches

| | | | |
|--|--|---|--|
| SmartBLAST | Primer-BLAST | Global Align | CD-search |
| Find proteins highly similar to your query | Design primers specific to your PCR template | Compare two sequences across their entire span (Needleman-Wunsch) | Find conserved domains in your sequence |
| IgBLAST | VecScreen | CDART | Multiple Alignment |
| Search immunoglobulins and T cell receptor sequences | Search sequences for vector contamination | Find sequences with similar conserved domain architecture | Align sequences using domain and protein constraints |
| MOLE-BLAST | | | |
| Establish taxonomy for uncultured or environmental sequences | | | |

Identification of metabolites

- Comparison of HPLC/GC-MS results
- Authentic standards required for identification of metabolite identity



- Search for studies based on metabolites or species
 - Example: find all studies associated with tyrosine
- Search for information about individual compounds
 - Delphinidin 3-O- β -D-glucoside
- Data sets are available for download

Download data from MetaboLights

 [MetaboLights ISAcreator software download.](#) To make it easy for new users, please download and just unzip our pre-packaged ISAcreator with plugin and configurations.

 [Experiments.](#) All public MetaboLights experiments can be downloaded from our public [ftp archive](#). Please find zip archives under the "studies" folder. Each public study can be found in the corresponding MTBLSt-id folder. Complete experiments can be opened with [ISAcreator](#) or you can extract the archives using your normal unzip program.

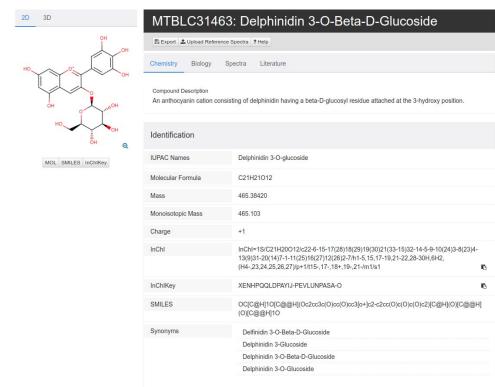
Status: Public | Release Date: 2016-09-21
MTBL533: Discovery of A-type procyanidin dimers in yellow raspberries by untargeted metabolomics and correlation based data analysis

Elisabete Carvalho, Pietro Franceschi

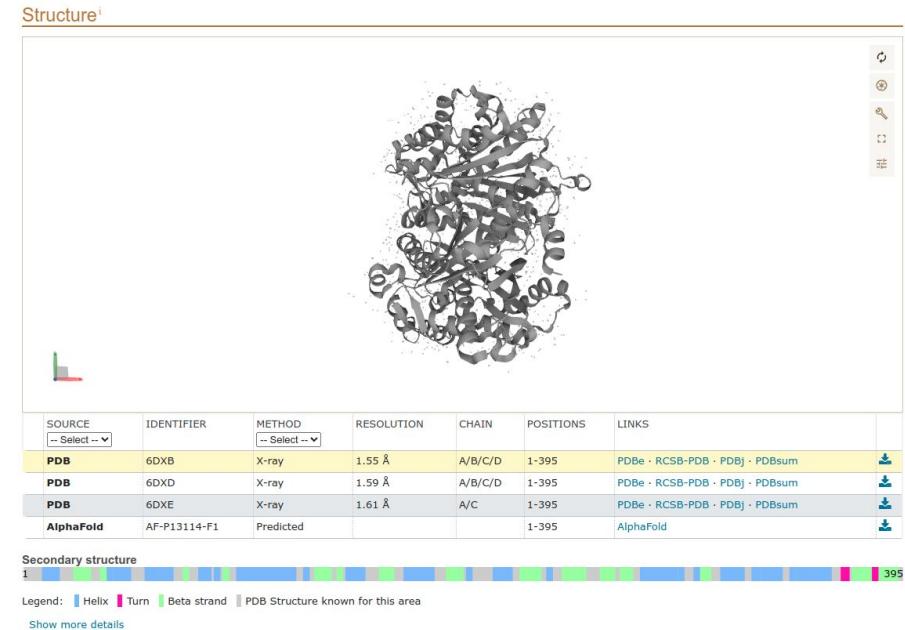
Introduction:
 Raspberries are becoming increasingly popular due to their reported health beneficial properties. Despite the presence of only trace amounts of anthocyanins, yellow varieties seems to show similar or better effects in comparison to conventional raspberries.
Objectives:
 The aim of this work is to characterize the metabolic differences between red and yellow berries, focusing on the compounds showing a higher concentration in yellow varieties.

Methods:
 The metabolomic profile of 13 red and 12 yellow raspberries (of different varieties, locations and collection dates) was determined by UPLC-TOF-MS and a novel correlation based approach was implemented to extract the pseudospectra of the most relevant biomarkers from high energy LC-MS data.

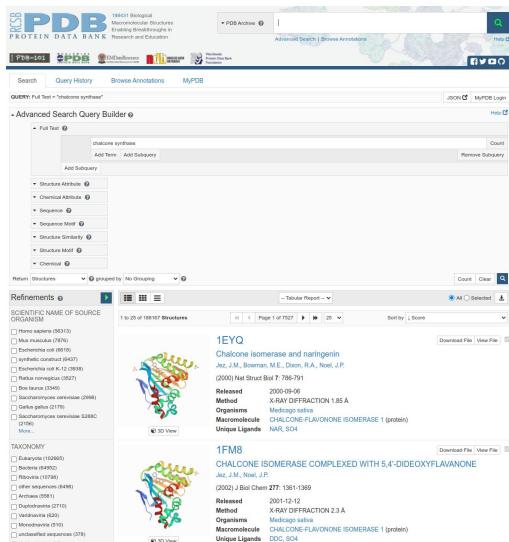
Results:
 Among the metabolites showing higher concentration in yellow raspberries it was possible to identify a series of compounds showing a pseudospectrum similar to that of A-type procyanidin polymers. The annotation of this group of compounds was confirmed by specific MS experiments and performing standard injections.
Conclusion:
 In berries lacking anthocyanins the polyphenol metabolism might be shifted to the formation of a novel class of A-type procyanidin polymers.



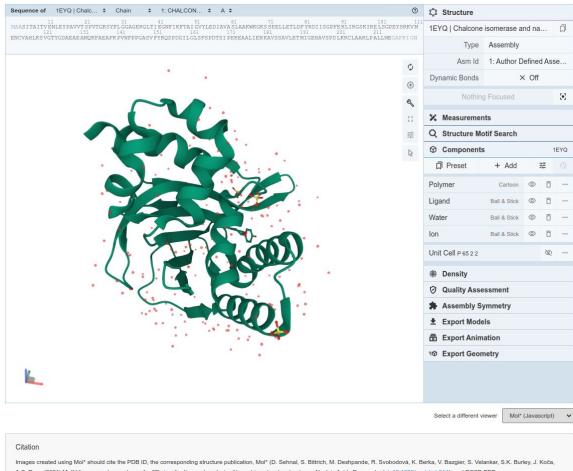
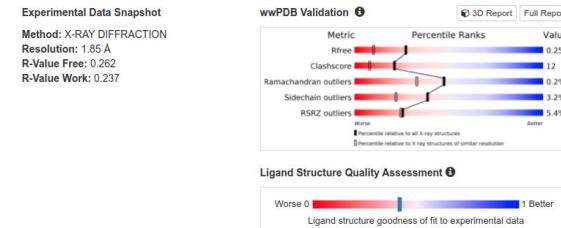
- Curated database of protein sequences
- Provides additional information retrieved from other databases
- 2D and 3D structure of protein
- Sequence available for download



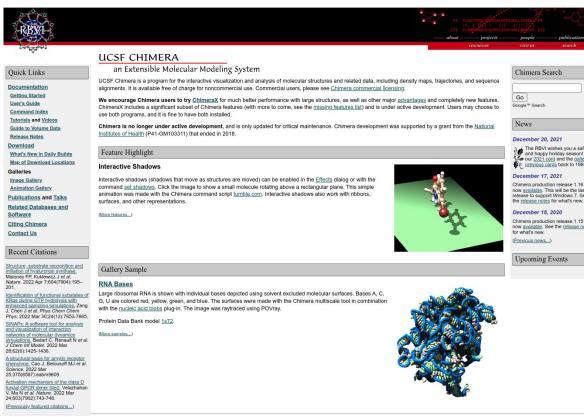
- Large collection of protein structures
- Download of protein structure files possible
- PDB viewer (DeepView) allows interactive inspection of protein structures



The screenshot shows the RCSB PDB homepage with a search bar containing "chalcone isomerase". Below the search bar, a "Query Full Text" section is expanded, showing search terms like "chalcone isomerase", "Add Term", and "Add Subquery". On the left, there's a sidebar for "Refinements" with sections for "SCIENTIFIC NAME SOURCE ORGANISM", "TAXONOMY", and "3D View". The main search results table shows 1 to 25 of 16817 structures, with the entry "1EYQ Chalcone isomerase and trimerin" highlighted.



- Downloaded PDB files can be displayed in external tools
- Chimera: <https://www.cgl.ucsf.edu/chimera/>
- PyMol: <https://pymol.org/2/>



UCSF CHIMERA
an Extensible Molecular Modeling System

Users can download the software, evaluate and analyze macromolecular structures and related data, including density maps, ligandposes, and sequence alignments. It is available free of charge for noncommercial use. Commercial users, please see [Chimera commercial license](#).

We encourage Chimera users to try ChimeraX for much better performance with large structures, as well as other major advances and completely new features. ChimeraX is a free, open-source molecular visualization toolkit—just now in beta, see the [ChimeraX beta site](#) and to order active development. Users may choose to use both programs, and it is fine to have both installed.

Chimera is no longer under active development, and is only updated for critical maintenance. Chimera development was supported by a grant from the National Institutes of Health (P41 GM031171) that ended in 2016.

Quick Links

- Documentation
- Getting Started
- Users' Guide
- Command Line
- Training & Tutorials
- Guides to Science
- Data Sources
- Downloads
- What's New in Daily Builds
- Mac OS X License
- Galleries
- Topics Gallery
- Annotations
- Publications and Tables
- Database Databases and Software
- Using Chimera
- Contact Us

Recent Citations

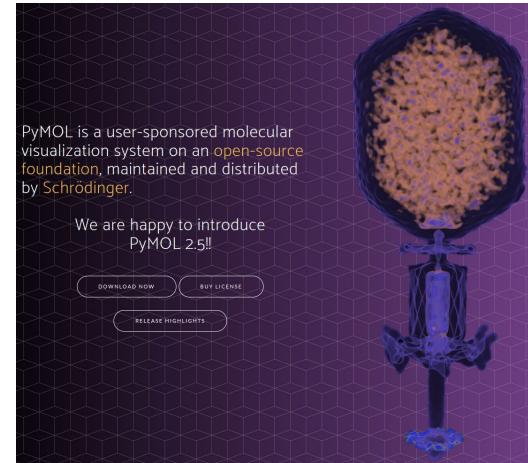
Gallery Sample

RNA Bases

The RNA base is shown with individual bonds depicted using colored oxidized molecular surfaces. Bases A, C, G, U are colored red, yellow, green, and blue. The surfaces were made with the Chimera mSurface tool in combination with the [povray](#) and [blends](#) plug-in. The image was raytraced using POV-ray.

[Protein Data Bank model 1z72](#)

[View details...](#)



AlphaFold DB: Protein Structure Database

- AlphaFold is an Artificial Intelligence tool developed by DeepMind
 - Leading tool in the prediction of protein structures
- Corresponding database is maintained by EMBL-EBI
- Database contains predicted protein structures

Chalcone synthase
AlphaFold structure prediction

Download [PDB file](#) | [mmCIF file](#) | [Predicted aligned error](#)

NEW Feedback on structure [Looks great](#) | [Could be improved](#)

Information

Protein Chalcone synthase
Gene Unknown
Source organism Hypericum androsaenum (Tutsan) [Go to search of](#)
UniProt Q9FUB7 [Go to UniProt if](#)
Experimental structures None available in the PDB
Biological function The primary product of this enzyme is 4,2',4'-tetrahydroxychalcone (also termed naringenin-chalcone or chalcone) which undergoes enzyme-catalyzed or spontaneous isomerization into naringenin. [Go to UniProt if](#)

3D viewer 

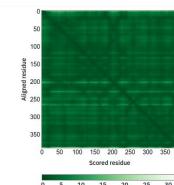
Model Confidence:
█ Very high ($\text{pLDOT} > 90$)
█ Confident ($90 > \text{pLDOT} > 50$)
█ Low ($70 > \text{pLDOT} > 50$)
█ Very low ($\text{pLDOT} < 50$)

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

Views

Predicted aligned error

Sequence Features (coming soon)



Aligned residue
Scored residue

Predicted aligned error

The colour at position (x, y) indicates AlphaFold's expected position error at residue x , when the predicted and true structures are aligned on residue y .

This is useful for assessing inter-domain accuracy - see the tutorial below.

Predicted aligned error tutorial

Identification of enzymes - InterProScan5

- Screen of protein sequences against collection of protein signatures
- Allows the assignment of functional annotation terms
- Available as web service, but also as stand alone tool

InterProScan 5 Sequence Search
 This form allows you to scan your sequence for matches against the InterPro collection of protein signature databases.
 This web form is for internal debugging purposes only and is **not supported**. To submit jobs to InterProScan 5, please visit the InterPro Sequence Search or the InterProScan 5 Web Service.

Please Note
 This web form is for internal debugging purposes and will be retired in April 2016. If you have any questions/concerns please contact us via the feedback link above.

STEP 1 - Enter your input sequence
 Enter or paste a PROTEIN sequence in any supported format:
 uniprot:KPYH_HUMAN

Or, upload a file: (no file chosen) Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Select the applications to run

| | | | | |
|--|---|--|-----------------------------------|--------------------------------------|
| <input checked="" type="checkbox"/> TIGRFAM | <input type="checkbox"/> SFLD | <input type="checkbox"/> Prokka | <input type="checkbox"/> SignalP | <input type="checkbox"/> SignalP_EUK |
| <input type="checkbox"/> SignalP_GRAM_POSITIVE | <input checked="" type="checkbox"/> SignP_GRAM_NEGATIVE | <input type="checkbox"/> SUPERFAMILY | <input type="checkbox"/> PANTHER | <input type="checkbox"/> GenoID |
| <input type="checkbox"/> Hmmp | <input type="checkbox"/> ProSiteProfiles | <input type="checkbox"/> ProSitePatterns | <input type="checkbox"/> Coils | <input type="checkbox"/> SMART |
| <input type="checkbox"/> COO | <input type="checkbox"/> PRINTS | <input type="checkbox"/> Pfam | <input type="checkbox"/> MotifBle | <input type="checkbox"/> TMHMM |

STEP 3 - Submit your job
 Be notified by email (Tick this box if you want to be notified by email when the results are available)

Results for job iprscan5-I20220320-102557-0980-87120812-p2m

Tool Output Submission Details

| <input type="button" value="Download in XML format"/> | <input type="button" value="Download in TSV format"/> | <input type="button" value="Download in GFF3 format"/> | <input type="button" value="Download in SVG format"/> | <input type="button" value="Download HTML tarball file"/> | <input type="button" value="Download in JSON format"/> |
|---|---|--|--|---|--|
| sp P13114 CHSY_ARATH | ba46b0f06bfb1c91d161f0f91310f43 | 395 | Gene3D G3DSA:3.40.47.10 | - 1 241 4.8E-101 T 20-03-2022 | |
| sp P13114 CHSY_ARATH | ba46b0f06bfb1c91d161f0f91310f43 | 395 | SUPERFAMILY SSF53901 | Thiolase-like 241 393 2.98E-51 T 20-03-2020 | |
| sp P13114 CHSY_ARATH | ba46b0f06bfb1c91d161f0f91310f43 | 395 | SUPERFAMILY SSF53901 | Thiolase-like 10 237 2.38E-78 T 20-03-20 | |
| sp P13114 CHSY_ARATH | ba46b0f06bfb1c91d161f0f91310f43 | 395 | Pfam PF02797 Chalcone and stilbene synthases, C-terminal domain | 244 394 0.0 T 394 1.5E-71 | |
| sp P13114 CHSY_ARATH | ba46b0f06bfb1c91d161f0f91310f43 | 395 | PANTHER PTHR11877:SF81 BNAA02G30320D PROTEIN | 6 394 0.0 T 20-03-2022 20-03-2022 | |
| sp P13114 CHSY_ARATH | ba46b0f06bfb1c91d161f0f91310f43 | 395 | Gene3D G3DSA:3.40.47.10 | - 242 395 9.7E-62 T 20-03-2022 IPR01603 | |
| sp P13114 CHSY_ARATH | ba46b0f06bfb1c91d161f0f91310f43 | 395 | PANTHER PTHR11877 HYDROXYMETHYLGLUTARYL-COA SYNTHASE | 6 394 0.0 T | |
| sp P13114 CHSY_ARATH | ba46b0f06bfb1c91d161f0f91310f43 | 395 | CDD cd00831 CHS_like | 21 390 0.0 T 20-03-2022 - | |
| sp P13114 CHSY_ARATH | ba46b0f06bfb1c91d161f0f91310f43 | 395 | PIRSF PIRSF000451 PKS_III | 7 394 0.0 T 20-03-2022 IPR011141 | |
| sp P13114 CHSY_ARATH | ba46b0f06bfb1c91d161f0f91310f43 | 395 | Pfam PF00195 Chalcone and stilbene synthases, N-terminal domain | 10 233 2.9E-119 | |
| sp P13114 CHSY_ARATH | ba46b0f06bfb1c91d161f0f91310f43 | 395 | ProSitePatterns PS00441 Chalcone and stilbene synthases active site. | 161 177 - | |

Gene Ontology (GO) and AmiGO

- Defined statements about the function of a gene (controlled vocabulary)
- Hierarchical structure
 - Example: ‘metabolic process’ > ‘biosynthetic process’ > ... > ‘chalcone synthase’
- Supported by the Alliance of Genome Resources
- Connected with various other databases e.g. TAIR, FlyBase, Reactome, UniProt
- Machine readable to allow automatic processing
- Tools: Blast2GO and AmiGO
 - analyze the function of a sequence (web service and standalone)

- I GO:0008150 biological_process
- I GO:0008152 metabolic process
- I GO:0009058 biosynthetic process
- I GO:0071704 organic substance metabolic process
 - I GO:0009812 flavonoid metabolic process
- I GO:1901576 organic substance biosynthetic process
 - ▼ GO:0009813 flavonoid biosynthetic process
 - I GO:0009718 anthocyanin-containing compound biosynthetic process
 - I GO:0051551 aurone biosynthetic process
 - I GO:0033485 cyanidin 3-O-glucoside biosynthetic process
 - I GO:0033486 delphinidin 3-O-glucoside biosynthetic process
 - I GO:0051553 flavone biosynthetic process
 - I GO:0009716 flavonoid phytoalexin biosynthetic process
 - I GO:0051557 leucoanthocyanidin biosynthetic process
 - R GO:0009964 negative regulation of flavonoid biosynthetic process
 - I GO:0033487 pelargonidin 3-O-glucoside biosynthetic process
 - R GO:0009963 positive regulation of flavonoid biosynthetic process
 - R GO:0009962 regulation of flavonoid biosynthetic process

- AmiGO allows text or sequence searches
- Options to explore annotations in detail
- Database covers various kingdoms (not restricted to plants)

| Total term(s): 6; showing: 1-6 Results count <input type="button" value="10"/> | | «First | <Prev | Next> | Last» | Custom DL (up to 100000) | Bookmark |
|--|--|------------------------|--------------------------|--------------------------|---|--|--------------------------|
| <input type="checkbox"/> Term | Definition | Ontology source | | Ontology ID space | Synonyms | Alt ID | |
| <input type="checkbox"/> flavonoid glucuronidation | The modification of a flavonoid by the conjugation of glucuronic acid. The resultant flavonoid glucu more... | biological_process | | GO | flavonoid glucuronide biosynthesis flavonoid glucuronide biosynthetic process more... | | |
| <input type="checkbox"/> flavonoid phytoalexin biosynthetic process | The chemical reactions and pathways resulting in the formation of flavonoid phytoalexins, a group of more... | biological_process | | GO | flavonoid phytoalexin anabolism flavonoid phytoalexin biosynthesis flavonoid phytoalexin formation flavonoid phytoalexin synthesis | | |
| <input type="checkbox"/> flavonoid biosynthetic process | The chemical reactions and pathways resulting in the formation of flavonoids, a group of phenolic derivatives containing a flavan skeleton. | biological_process | | GO | flavonoid anabolism flavonoid biosynthesis flavonoid formation flavonoid synthesis | | |
| <input type="checkbox"/> regulation of flavonoid biosynthetic process | Any process that modulates the frequency, rate or extent of the chemical reactions and pathways resulting in the formation of flavonoids. | biological_process | | GO | regulation of flavonoid anabolism regulation of flavonoid biosynthesis more... | | |
| <input type="checkbox"/> negative regulation of flavonoid biosynthetic process | Any process that stops, prevents, or reduces the frequency, rate or extent of the chemical reactions more... | biological_process | | GO | down regulation of flavonoid biosynthetic process down-regulation of flavonoid biosynthetic process more... | | |
| <input type="checkbox"/> positive regulation of flavonoid biosynthetic process | Any process that activates or increases the frequency, rate or extent of the chemical reactions and pathways resulting in the formation of flavonoids. | biological_process | | GO | positive regulation of flavonoid anabolism positive regulation of flavonoid biosynthesis more... | | |

DOI: [10.5281/zenodo.6363634](https://doi.org/10.5281/zenodo.6363634)

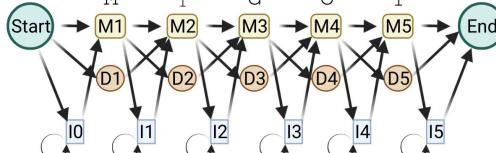
Last file loaded on 2022-03-11, see [full details](#)
AmiGO 2 version: 2.5.17 (amigo-production)

Pfam: Protein family database

- Assignment of protein functions based on Hidden Markov Models (HMMs)
- Sequences are screened based on HMM profile

Sequences are aligned
 ↓
 Modelling of insertions/deletions
 ↓
 Amino acid frequency at each alignment position is encoded
 ↓
 HMM profile generated

| | | |
|------|-------------|---------------|
| seq1 | MTGC - Y | 2 = deletion |
| seq2 | MSGC - F | 5 = insertion |
| seq3 | MTGC - Y | |
| seq4 | M - GCAY | |
| | 1 2 3 4 5 6 | |



[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#) | [ABOUT](#)

Pfam 35.0 (November 2021, 19632 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

QUICK LINKS [YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...](#)
SEQUENCE SEARCH Analyze your protein sequence for Pfam matches
VIEW A PFAM ENTRY View Pfam annotation and alignments
VIEW A CLAN See groups of related entries
VIEW A SEQUENCE Look at the domain organisation of a protein sequence
VIEW A STRUCTURE Find the domains on a PDB structure
KEYWORD SEARCH Query Pfam by keywords
JUMP TO Enter any accession or ID to jump to the page for a Pfam entry or clan, UniProt structure, etc.
 Or view the [help](#) pages for more information

Recent Pfam blog posts
[Pfam 35.0 is released!](#) (posted 19 November 2021)

Pfam 35.0 contains a total of 19,632 families and clans. Since the last release, we have built 460 new families, killed 7 families and created 12 new clans. UniProt Reference Proteomes has increased by 7% since Pfam 34.0, and now contains 61 million sequences. Of the sequences that are in UniProt Reference Proteomes, 75.2% have [...]

[AlphaFolding the Protein Universe!](#) (posted 22 July 2021)

Hot on the tail of our inclusion of the Baker group's trRosetta structural models we are excited to announce the inclusion of models from AlphaFold 2.0 generated by DeepMind and stored in the AlphaFold Database (AlphaFold DB). AlphaFold 2.0's performance in the CASP14 competition was spectacular, producing near experimental quality structure models. The new AlphaFold [...]

[Google Research Team bring Deep Learning to Pfam!](#) (posted 24 March 2021)

We are delighted to announce the first fruits of a collaboration between the Pfam team and a Google Research team led by Dr Lucy Colwell, with Maxwell Bileshi and David Belanger. In 2019, Colwell's team published a preprint describing a new deep learning method that was trained on Pfam data, and which improves upon the [...]

Citing Pfam

If you find Pfam useful, please consider [citing](#) the reference that describes this work:

[Pfam: The protein families database in 2021](#): J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson, R.D. Finn, A. Bateman. *Nucleic Acids Research* (2020) doi: 10.1093/nar/gkaa913

- Online tool for the annotation of all protein sequences in a submitted FASTA file
- FASTA format:
 - Header line starts with ‘>’ followed by the sequence name
 - Header is followed by unrestricted number of sequence lines

```
>TRINITY_DN100013_c0_g1_i1
MIGPMPGMEGKMLPLGASPVGLEVVLVLSVVASTPILSDTSLCTPSFYHFLLLLGPITSISNLIVRPFFLTLSSITVIYGRLCFFAFDFYVY
>TRINITY_DN100016_c0_g1_i1
MPKKSSNIQQNNQNRIKRKGRPPKKKKYVQQIDSSDEDILSVRHASTRPRIISIRRNEIPMRPEIHIPPTSIPSFNNPNNIHQSSSDRQMPP
>TRINITY_DN100019_c0_g1_i1
MPPKAADKKPAAKAPATASKAPEKKDAGKTAATGDKKKRTKARKETYSSYIYKVLKQVHPDTGISNRAMSILNSFVNDFERVATEASKLA
>TRINITY_DN10001_c0_g1_i2
MAKVGNPIVDIETDGSVNEPESSEKNIEVSSSSTQAPESTNTTELLVNEKKAFSLATPAVRRVAREHNIDINNIKGTKNGRITKEDILNYV
>TRINITY_DN100025_c0_g1_i1
MVENQDGCFFKPGWKEFVRNSDLEGGDFLVFNLDKISYQVVIFDGTCCPKDLCFPSIMNPIFQHLRNKIFLSKKEEIKLKGNRKVHSVNEN
>TRINITY_DN10002_c0_g1_i1
MSDELNQVFQRCREQKRPVFVAFVTAGYPDSEETVDILLGLEAGGADIIELGIPFTDPMVDGKTIQDANNVALENKIDISKCLSYVSESRAK
```

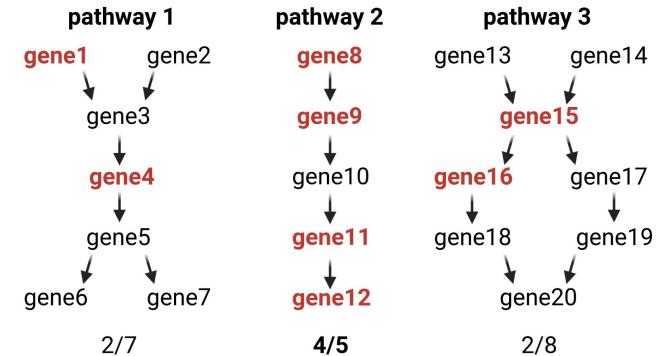
Concept of enrichment analyses

- Omics analyses can identify a large set of differentially expressed genes
- Making sense of large gene sets is challenging
- Identification of a common function is necessary to reduce complexity
- Enrichment analyses can help to identify shared functions or pathways

Differentially expressed genes (DEGs)

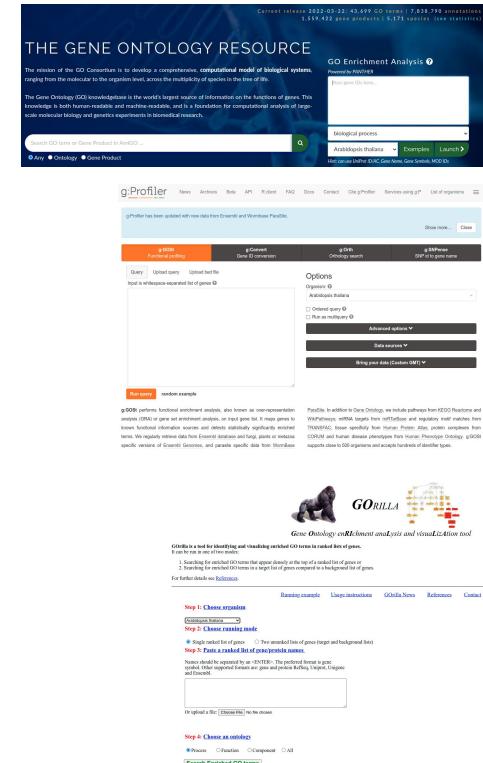
gene1
gene4
gene8
gene9
gene11
gene12
gene15
gene16
gene22
...

↓
Identification of enriched pathways



Enrichment analyses - GO

- Functional enrichment analyses can be based on GO terms
- Assignment of GO terms to genes of the species of interest
- Tools for GO enrichment analysis:
 - PANTHER: <http://geneontology.org/>
 - g:Profiler: <http://biit.cs.ut.ee/gprofiler/gost>
 - GOrilla: <http://cbl-gorilla.cs.technion.ac.il/>



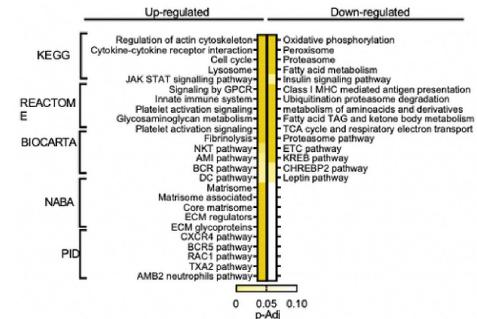
The screenshot shows the Gene Ontology Resource homepage. At the top, it displays current statistics: 10,522,022 GO terms, 1,549,422 gene products, 51,172 species, and 7,838,795 annotations. Below this, there's a search bar and a navigation menu with links to Home, Help, and Log In.

The main content area features several analysis tools:

- g:Profiler:** A tool for functional enrichment analysis. It has tabs for "GOKE" (GOKE enrichment), "Gene ID converter", "GO DB", "Ontology search", and "g:MPbase". It includes options for "Organism", "Organism database", "Convert every", "Run as regulatory", and "Advanced options".
- GOrilla:** A tool for identifying enriched GO terms in ranked lists of genes. It has tabs for "GOrilla", "GOrilla browser", "GOrilla API", and "GOrilla help". It includes sections for "What is GOrilla?", "How does GOrilla work?", and "For further details see References".
- PANTHER:** A tool for protein classification and annotation. It has tabs for "PANTHER", "PANTHER browser", "PANTHER API", and "PANTHER help". It includes sections for "What is PANTHER?", "How does PANTHER work?", and "For further details see References".

Enrichment analyses - KEGG

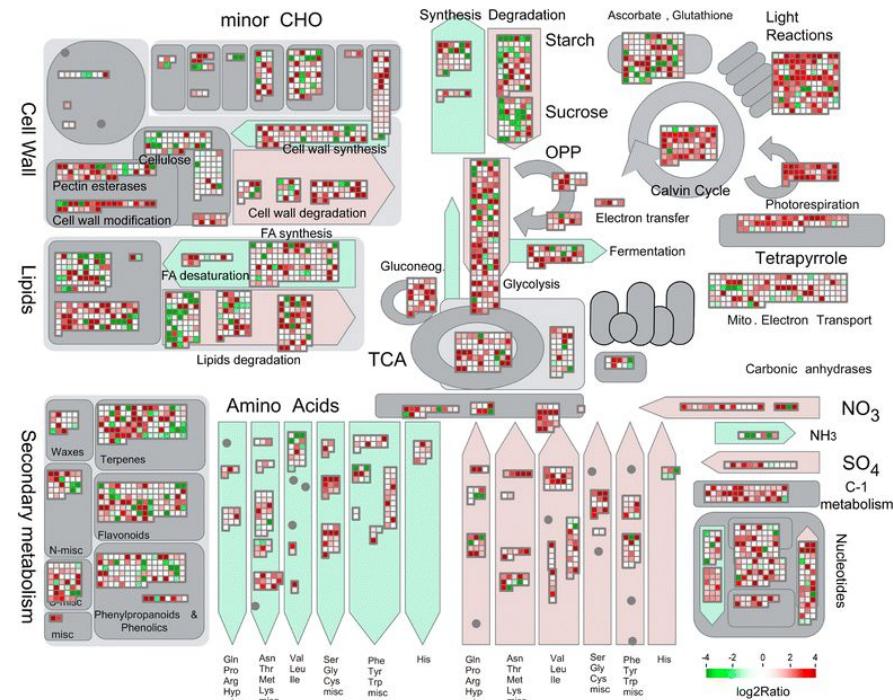
- Identification of enriched KEGG pathways among a gene set
- Available tools:
 - KEGG Mapper:
<https://www.genome.jp/kegg/kegg1b.html>
 - pathfindeR: <https://doi.org/10.3389/fgene.2019.00858>
- Names of pathways is not always informative
- KEGG pathways can be too superficial and are only a first step in analyses



Enrichment analyses - MapMan

- Identification of enriched pathways
- Visualization of up- and down-regulated pathways:
 - Red = up-regulation
 - Green = down-regulation
 - White = neutral (no change)
- Overview of the metabolism of a cell
- MapMan:

<https://mapman.gabipd.org/mapman>



Summary

- Types of databases (reactions, enzymes, sequences, metabolites, species-specific)
- Species-specific databases
- Finding information about enzymes, transcription factors, and other genes
- Finding information about protein structures
- NCBI & INSDC
- Enrichment analyses

Time for questions!

Questions

1. What is the meaning of FAIR data?
2. How would you get insights into the function of an unknown sequence?
3. How can you identify the enzymes involved in the flavonol biosynthesis?
4. Where can you find information about the substrate affinity of the chalcone synthase?
5. What is TAIR?
6. Where can you find information about the expression of *MYB12* in different plant organs?
7. Where can you get the genome sequence and corresponding annotation of *Marchantia polymorpha*?
8. Which databases are connected through INSDC?
9. Where can you find the results of metabolomic studies?
10. Where can you find information about the structure of CHS?
11. Which tools can be applied to annotate large sets of (poly)peptide sequences?
12. Which tools can be applied for pathway enrichment analyses?
13. What are the two line types in a FASTA file?