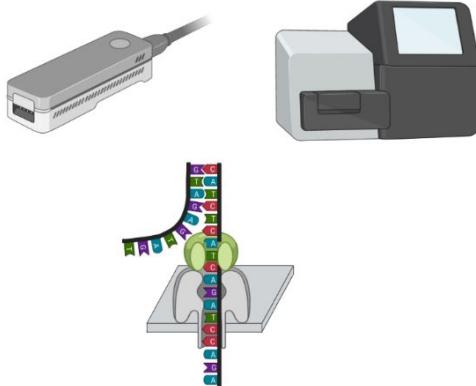
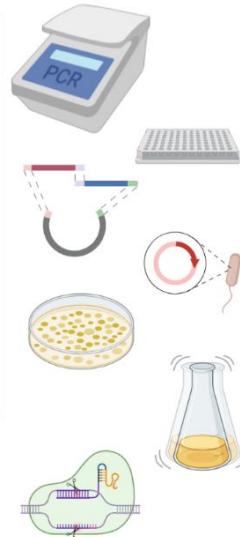
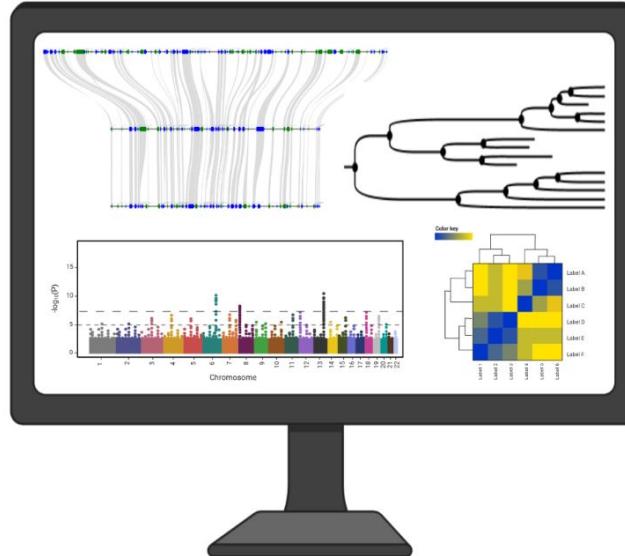




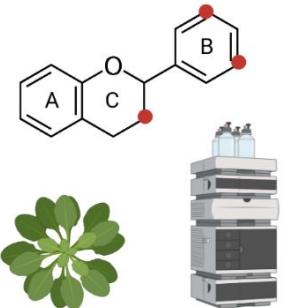
Technische
Universität
Braunschweig



Plant Biotechnology
and Bioinformatics



species biosynthesis proteins analysis different conditions
biosynthesis species activities bellman variants H2R3-MYB
within genes functional variants Col-0 Col-0 variant
dissolve site data divergent variants non-canonical
sequences IGR single reference multiple protein annotation level identified
sites synthesis amino acid evolutionary pathway
plants reference genes accessions identification
pigments model genome key against canonical free
Keyvernia genes evolution systems biology long Canophyllales
plants sequencing for conserved Arabidopsis
flavonoid conservation read transcription synthesis evolution
accessions identification sequence MYB introns residues RNA-Seq



Pathway Databases

Prof. Dr. Boas Pucker (Plant Biotechnology and Bioinformatics)

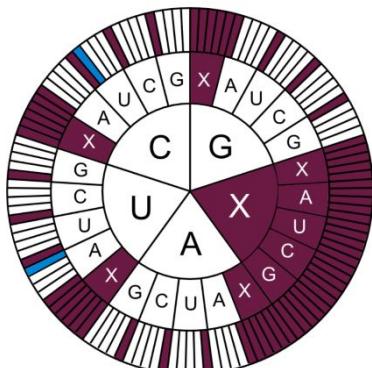
Availability of slides

- All materials are freely available (CC BY) - after the lectures:
 - StudIP: [Lecture: Grundlagen der Biochemie und Bioinformatik der Pflanzen \(Bio-MB 09\)](#)
 - GitHub: <https://github.com/bpucker/teaching>
- Questions: Feel free to ask at any time
- Feedback, comments, or questions: b.pucker[a]tu-bs.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)



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

Identification of candidate
enzymes in a **database**



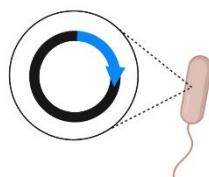
Croton tiglium is naturally
producing a precursor of this
unnatural DNA base



RNA extraction, sequencing,
and construction of a
transcriptome assembly



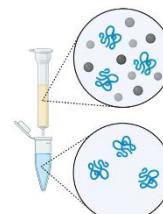
Cloning of GMPS coding
sequence into a plasmid



Transformation into
Escherichia coli



Heterologous
expression of GMPS



Protein purification



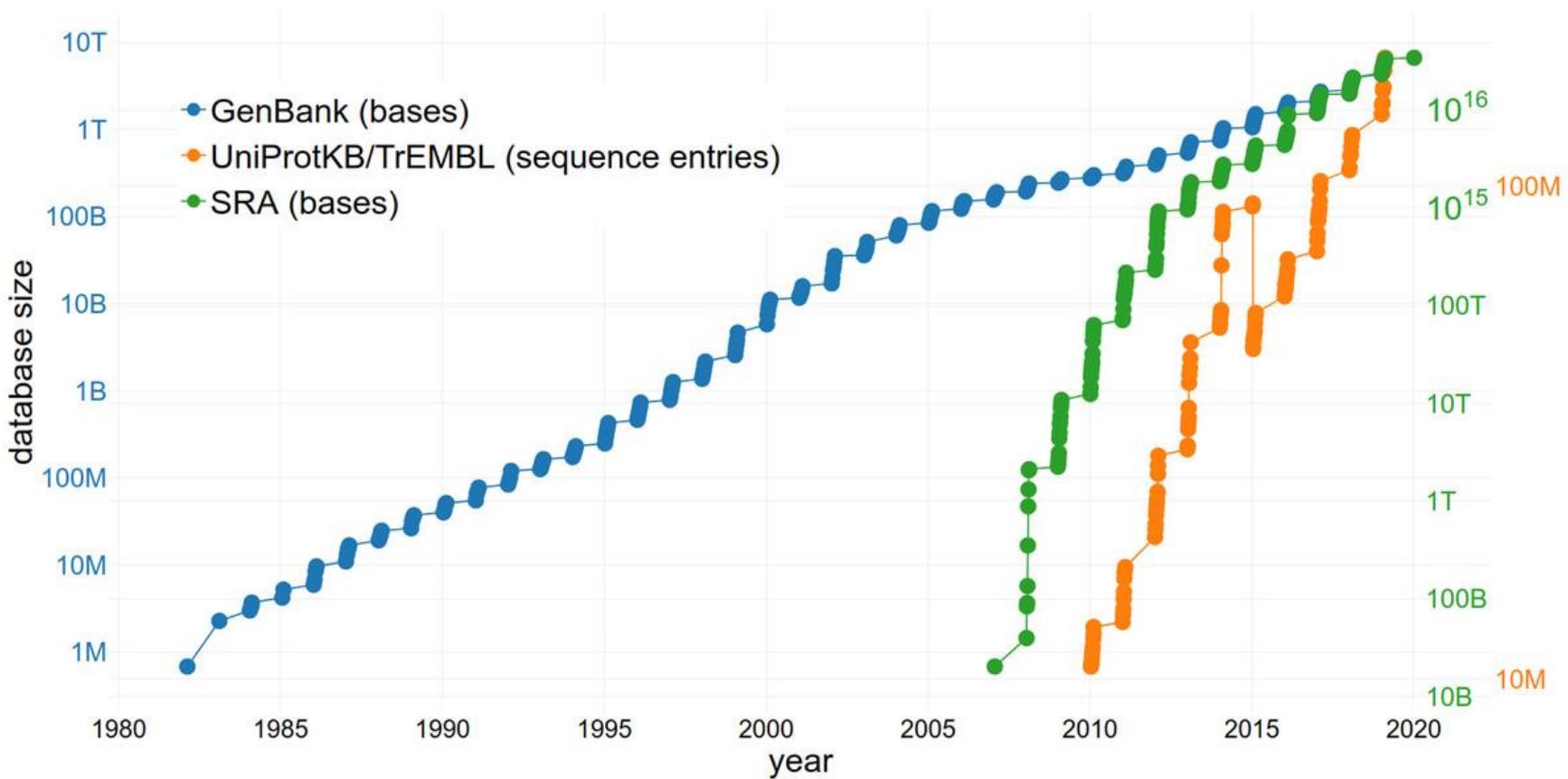
Enzyme assay

<http://2017.igem.org/Team:Bielefeld-CeBiTec>
Haak et al., 2018: 10.3389/fmolb.2018.00062



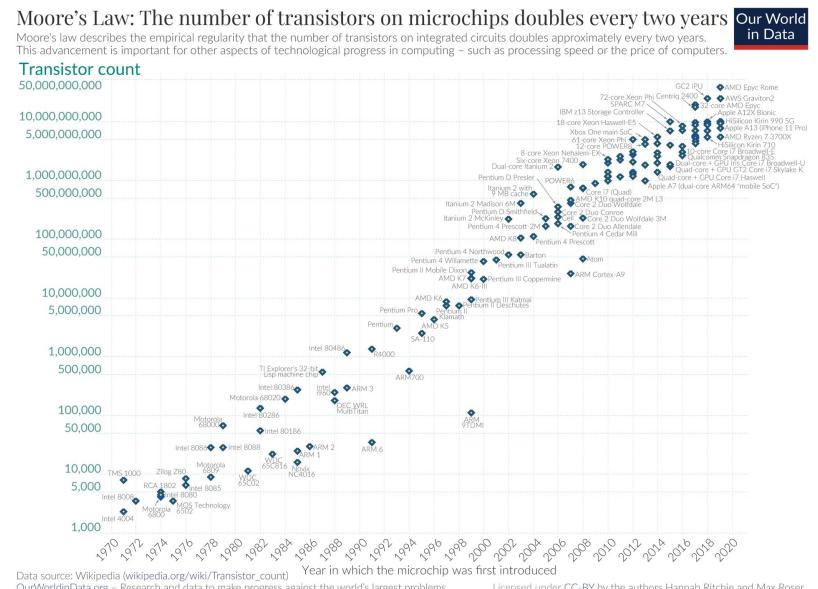
Technische
Universität
Braunschweig

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)

The KEGG homepage features a sidebar with links to various KEGG databases like KEGG Pathway, KEGG Orthology, and KEGG Genomes. The main content area includes a search bar, a 'Search BRENDA' button, and a 'Browse JASPAR CORE' section with images of a fly, plant, and nematode.

- Enzymes (BRENDA)

The BRENDA homepage has sections for 'Task-based queries', 'Structure-based queries', 'Explorer', 'Prediction', and 'Supporting & External'. It also features a 'Visualization' section with links to Enzyme Word Maps, Organism Word Maps, and Functional Parameter Statistics.

- Transcription factors (JASPAR)



- Metabolites (MetaboLights)

The MetaboLights homepage includes a search bar, a 'Quick tour' link, and sections for 'Study', 'Compound Library', and 'Species'. It also features a 'Metabolite ID' search interface and a 'Metabolite ID' download page for Arabidopsis.

- Species specific databases (TAIR, Phytozome, PLAZA)

<https://www.genome.jp/kegg/>
<https://www.brenda-enzymes.org/>
<https://www.ebi.ac.uk/metabolights/>
<https://www.arabidopsis.org/>

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)



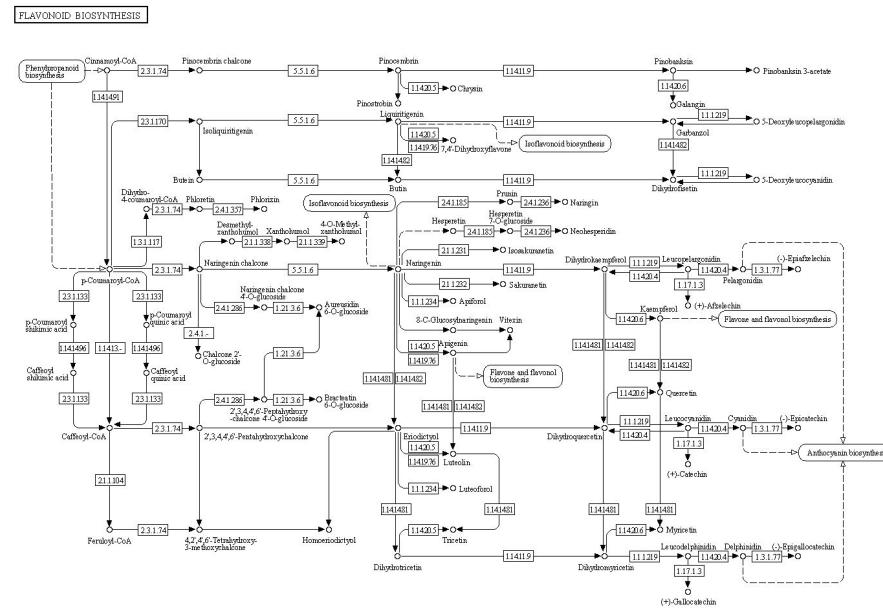
<https://www.go-fair.org/>



<https://orcid.org/>

KEGG: Kyoto Encyclopedia of Genes and Genomes

- 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

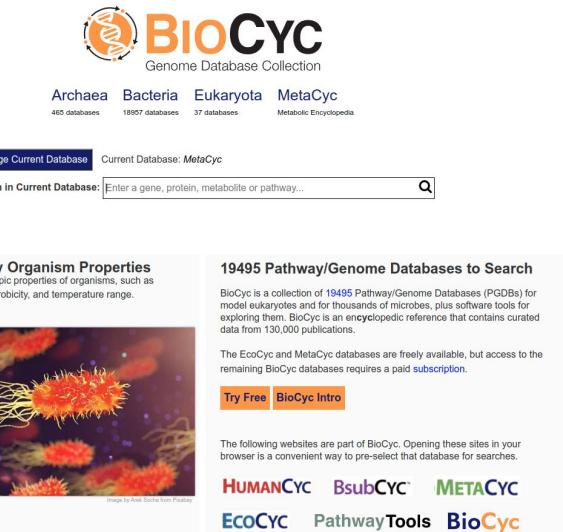


map00941 Flavonoid biosynthesis

Kanehisa & Goto, 2000: <https://doi.org/10.1093/nar/28.1.27>
Kanehisa et al., 2022: <https://doi.org/10.1002/pro.4172>
<https://www.genome.jp/kegg/>

BioCyc: Collection of databases

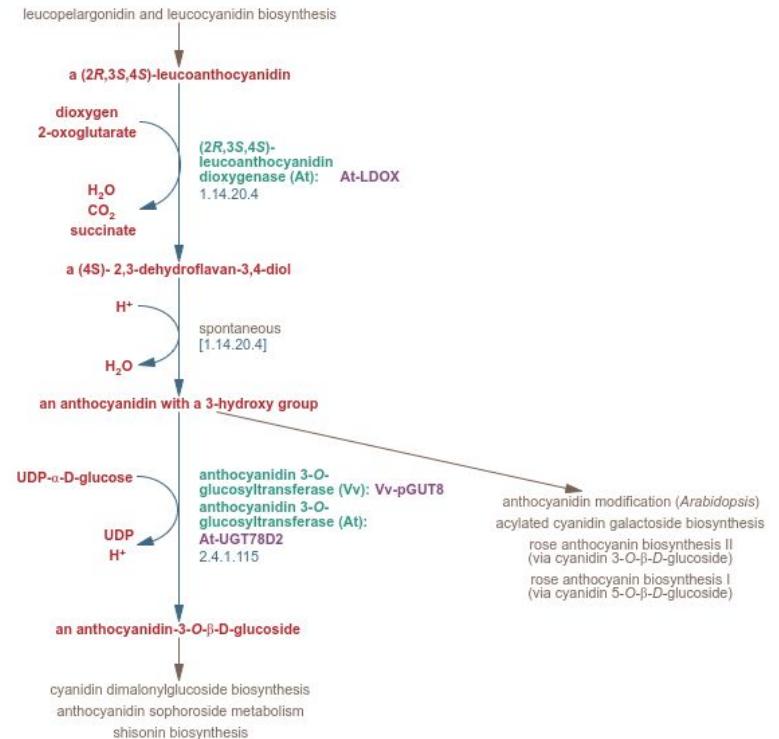
- Integration of databases and tools for omics
- Curation is involved to increase the quality
- Strong focus on microorganism:
 - EcoCyc = **Escherichia coli Encyclopedia**
 - BsubCyc = **Bacillus subtilis Encyclopedia**
 - YeastCyc = **Yeast Encyclopedia**
- BUT:
 - Search for metabolites
 - Search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



The screenshot shows the BioCyc homepage. At the top right is the BioCyc logo with the text "Genome Database Collection". Below it are links for Archaea (465 databases), Bacteria (18907 databases), Eukaryota (37 databases), and MetaCyc. A search bar at the top right contains the placeholder "Enter a gene, protein, metabolite or pathway..." with a magnifying glass icon. Below the search bar is a section titled "Search Databases by Organism Properties" with a sub-section for "19495 Pathway/Genome Databases to Search". This section includes a paragraph about BioCyc's coverage of various organisms, a "Learn More" link, and a small image of bacterial cells. To the right of this are links for "Try Free" and "BioCyc Intro". At the bottom are links for HUMANCyc, BsubCyc, METACYC, EcoCyc, PathwayTools, and BioCyc.

MetaCyc: Metabolite Encyclopedia

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



<https://metacyc.org/>

Karp et al., 2015: <https://doi.org/10.48550/arXiv.1510.03964>

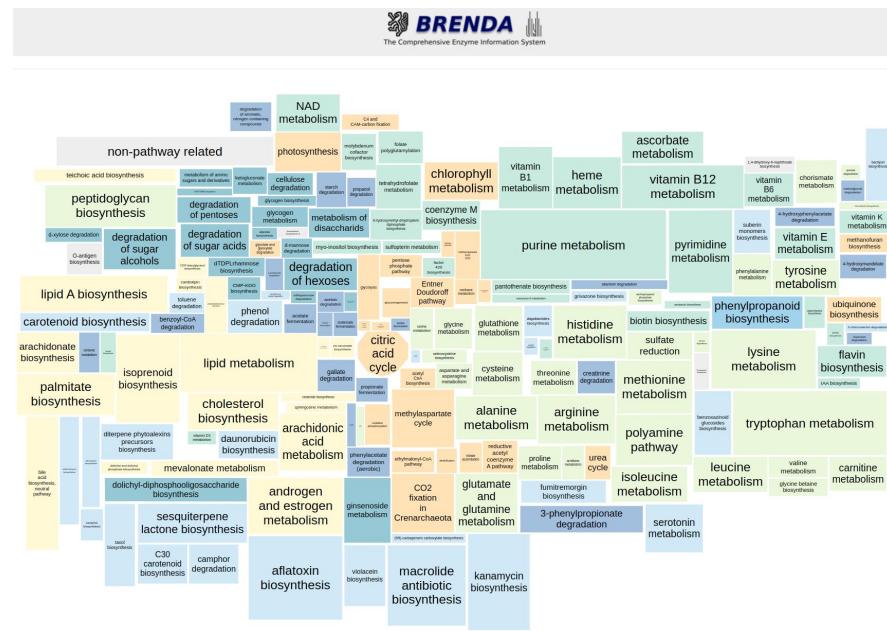
Figure: <https://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5125>

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

BRENDA: BRAunschweig ENzyme DAtabase

- 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, ...)



BRENDA - example (1)

- 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

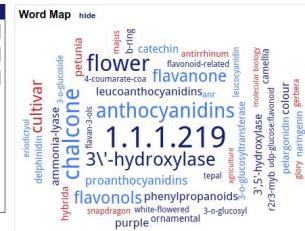
L-1 Oxidoreductases
L-1.1 Acting on the CH-OH group of donors
L-1.1.1 With NAD⁺ or NADP⁺ as acceptor
L-1.1.1.219 dihydroflavonol 4-reductase

IUBMB Comments

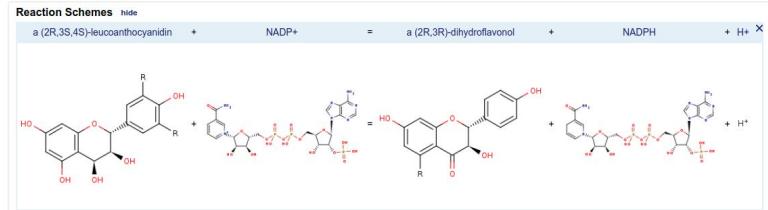
This plant enzyme, involved in the biosynthesis of anthocyanidins, is known to act on (+)-dihydrokaempferol, (+)-taxifolin, and (+)-dihydromyricetin, although some enzymes may act only on a subset of these compounds. Each dihydroflavonol is reduced to the corresponding cis-flavan-3-diol. NAD⁺ can act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

This record set is specific for:
Arabidopsis thaliana



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



Synonyms
Boag058630, BrDFR1, BrDFR10, BrDFR11, BrDFR12, BrDFR2, BrDFR3, BrDFR4, BrDFR5, BrDFR6, more

<https://www.brenda-enzymes.org/>

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 r=reversible i=inversible ? not specified ▲▼
dihydrokempferol + NADPH + H+	leucopelargonidin + NADP+	?	Arabidopsis thaliana	P51102	-	702000	-	-	?
dihydromyricetin + NADPH + H+	leucodelphinidin + NADP+	?	Arabidopsis thaliana	P51102	-	702000	-	-	?
dihydroquercetin + NADPH + H+	leucocyanidin + NADP+	?	Arabidopsis thaliana	P51102	-	702000	-	-	?
additional information	?	-	Arabidopsis thaliana	P51102	does not catalyze naringenin	702000	-	-	?

[Go to Cofactor Search](#)

COPARTNER ▲▼	ORGANISM ▲▼	UNIPROT	COMMENTARY ▲▼ X	LITERATURE ▲▼	IMAGE
NADPH	Arabidopsis thaliana P51102	-	702000	?	

[Go to Specific Activity Search](#)

SPECIFIC ACTIVITY [μmol/min/mg] ▲▼	ORGANISM ▲▼	UNIPROT	COMMENTARY ▲▼ X	LITERATURE ▲▼
0.00002	Arabidopsis thaliana	P51102	with dihydrokempferol as substrate	702000
0.0006	Arabidopsis thaliana	P51102	with eriodictol as substrate	702000
0.001	Arabidopsis thaliana	P51102	with dihydromyricetin as substrate	702000
0.0013	Arabidopsis thaliana	P51102	with dihydroquercitin as substrate	702000

[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 ▲▼	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 AA Sequence and Transmembrane Helices Search](#)

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

[Go to Cloned \(commentary\) Search](#)

CLONED (Commentary) ▲▼	ORGANISM ▲▼	UNIPROT	LITERATURE ▲▼
into pTrcHis2-TOPO and heterologously expressed in Escherichia coli TOP10F' strain. DFR cDNA cloned into pRSF-FHT and inserted into Escherichia coli BL21Star to create E-color strain	Arabidopsis thaliana	P51102	702000

[Select Items on the left to see more content.](#)

[Go to References Search](#)

REF. ▲▼	AUTHORS ▲▼	TITLE ▲▼	JOURNAL ▲▼	VOL ▲▼	PAGES ▲▼	YEAR ▲▼	ORGANISM (UNIPROT) ▲▼	PUBMED ID ▲▼	SOURCE
702000	Leonard, E.; Yan, Y.; Chemler, J.; Matern, U.; Martens, S.; Koffas, M.	Characterization of dihydroflavonol 4-reductases for recombinant plant pigment biosynthesis applications	BioCat., Biotransform.	26	243-251	2008	Fragaria x ananassa (Q22617), Ipomoea nil (Q24607), Arabidopsis thaliana (P51102), Rosa hybrid cultivar (Q41158), Lilium sp. (Q6UAQ7), Anthurium andraeanum (Q84L22)	-	BRENDA

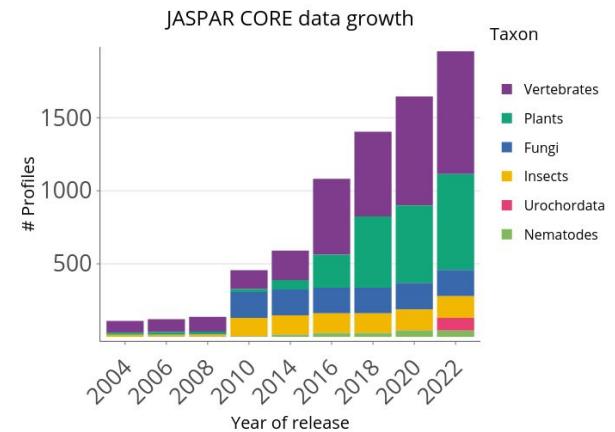
[EXTERNAL LINKS \(specific for EC-Number 1.1.1.219\)](#)

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

<https://www.brenda-enzymes.org/>

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: *Arabidopsis thaliana*
Data Type: PBM
Validation: 20675570
Uniprot ID: Q94FL6
Source: 31133749
Comment:

Sequence logo Download SVG

Frequency matrix JASPAR TRANSFAC MEME RAW PFM Reverse comp.

	A	C	G	T	
A [324	200	546	40	36
C [230	284	266	799	0
G [221	183	178	124	963
T [223	330	8	36	0
					1 61 80 271]
					0 999 0 718 610 222]
					680 119 159 197]
					318 100 149 309]

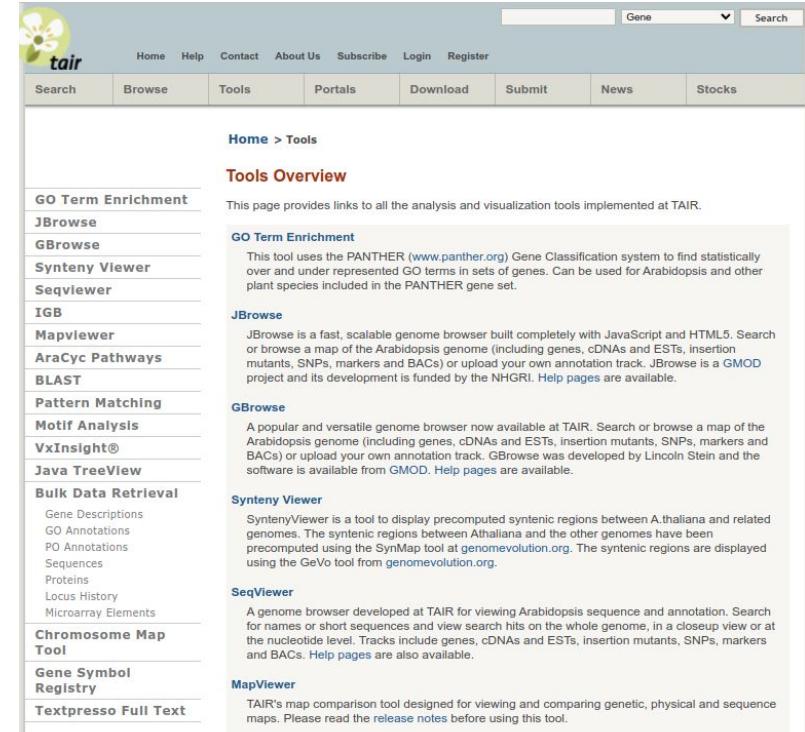
MetaboLights

- 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, a search bar, and user login options. Below the header, the MetaboLights logo is displayed, followed by a brief introduction about the database being a repository for metabolomics experiments and structures. A "Quick tour" link is available. The main content area features three large cards: "Study" (with icons for BROWSE, ORCID SEARCH, and METABOLIGHTS LABS), "Compound Library" (with icons for COMPOUNDS and SPECIES), and "Training" (with icons for TRAINING ONLINE and QUICK TOUR). To the right, a sidebar displays tweets from the official MetaboLights Twitter account (@Metabolights) and a "Submit Study" button.

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

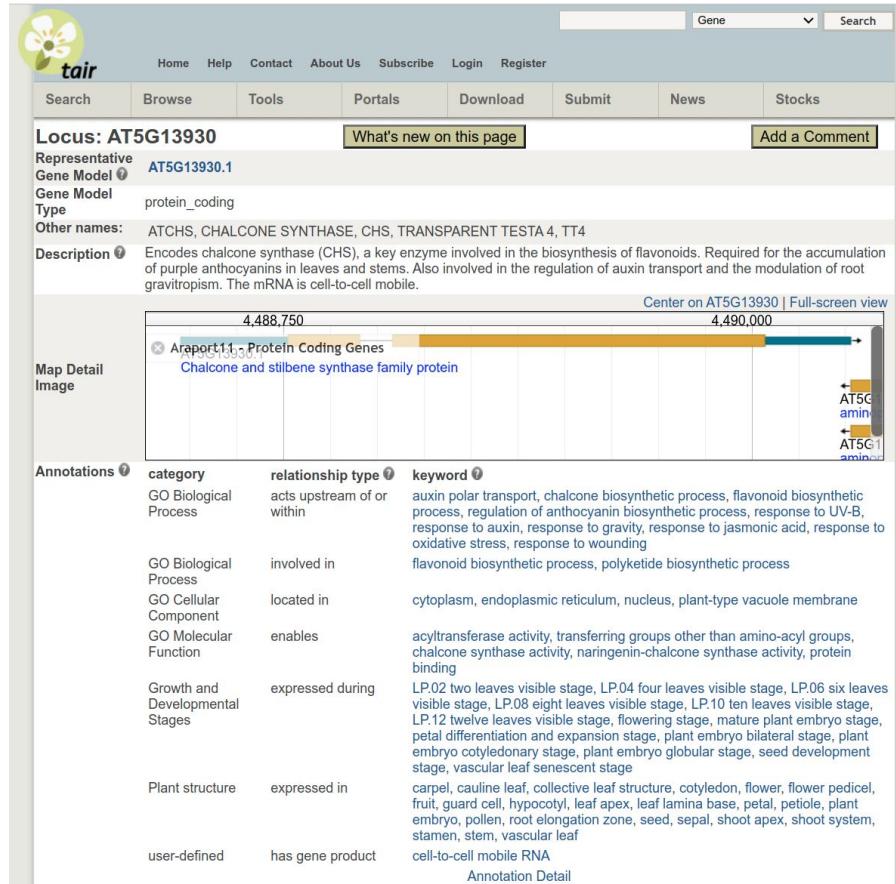


The screenshot shows the TAIR Tools Overview page. At the top, there is a navigation bar with links for Home, Help, Contact, About Us, Subscribe, Login, and Register. Below the navigation bar is a search bar with a dropdown menu set to 'Gene' and a 'Search' button. The main content area has a breadcrumb trail 'Home > Tools'. A section titled 'Tools Overview' states: 'This page provides links to all the analysis and visualization tools implemented at TAIR.' Below this, there are several tool categories listed on the left with their descriptions on the right:

- GO Term Enrichment**: This tool uses the PANTHER (www.panther.org) Gene Classification system to find statistically over and under represented GO terms in sets of genes. Can be used for Arabidopsis and other plant species included in the PANTHER gene set.
- JBrowse**: JBrowse is a fast, scalable genome browser built completely with JavaScript and HTML5. Search or browse a map of the Arabidopsis genome (including genes, cDNAs and ESTs, insertion mutants, SNPs, markers and BACs) or upload your own annotation track. JBrowse is a GMOD project and its development is funded by the NHGRI. Help pages are available.
- GBrowse**: GBrowse is a popular and versatile genome browser now available at TAIR. Search or browse a map of the Arabidopsis genome (including genes, cDNAs and ESTs, insertion mutants, SNPs, markers and BACs) or upload your own annotation track. GBrowse was developed by Lincoln Stein and the software is available from GMOD. Help pages are available.
- Synteny Viewer**: SyntenyViewer is a tool to display precomputed synteny regions between *A.thaliana* and related genomes. The synteny regions between *A.thaliana* and the other genomes have been precomputed using the SynMap tool at genomevolution.org. The synteny regions are displayed using the GeVo tool from genomevolution.org.
- SeqViewer**: A genome browser developed at TAIR for viewing Arabidopsis sequence and annotation. Search for names or short sequences and view search hits on the whole genome, in a closeup view or at the nucleotide level. Tracks include genes, cDNAs and ESTs, insertion mutants, SNPs, markers and BACs. Help pages are also available.
- MapViewer**: TAIR's map comparison tool designed for viewing and comparing genetic, physical and sequence maps. Please read the release notes before using this tool.
- Textpresso Full Text**

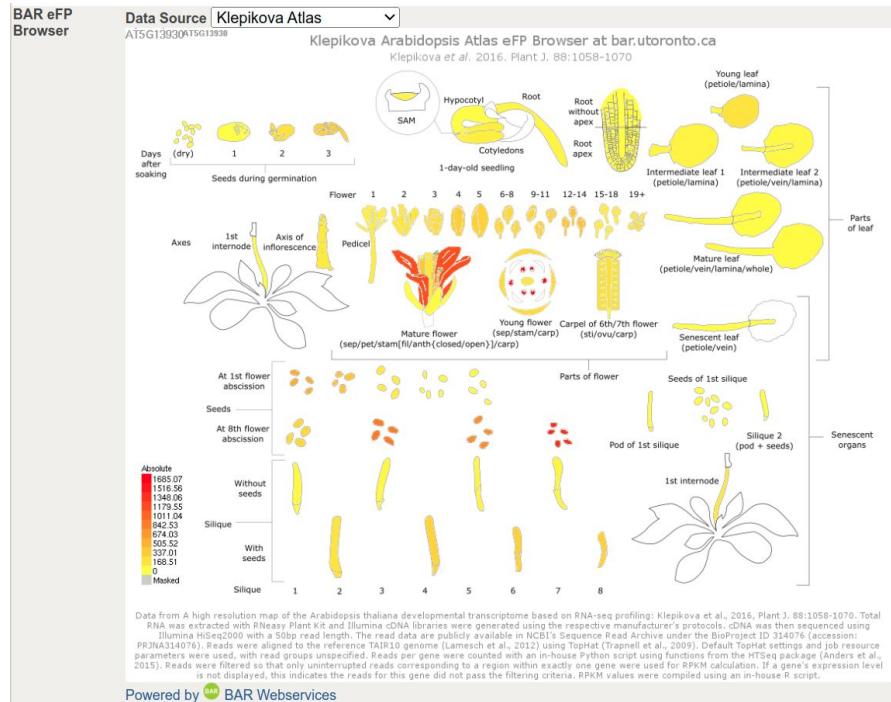
TAIR10 - functional annotation

- 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

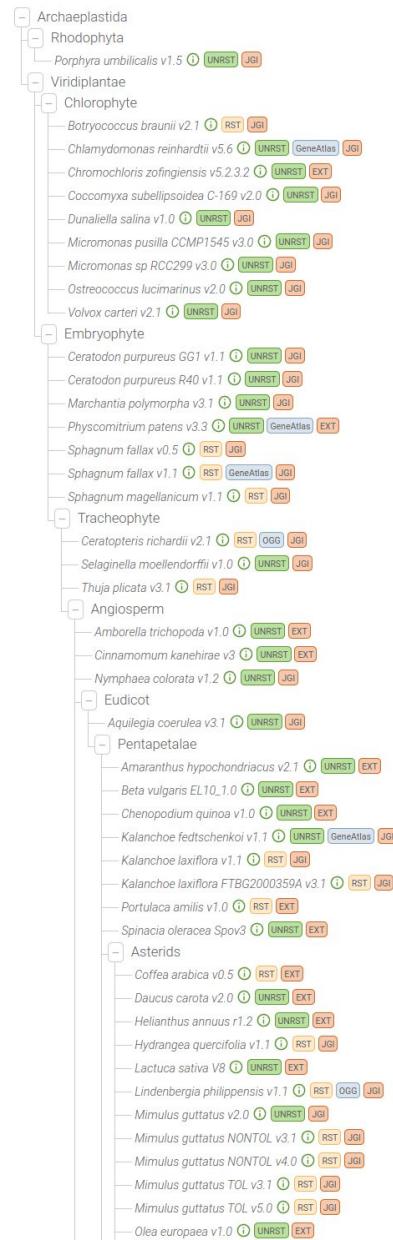
Publication	author/title	source	associated loci	date
Dahhan, D.A., Reynolds, G.D., Cár... Proteomic characterization of isolated <i>Arabidopsis</i> clathrin-coated vesicles reveals evolutionarily conserved and plant-specific components	THE PLANT CELL	AT2G46830 AT3G01780 AT3G51240 AT5G13930 AT5G61380		2022
Verma, D., Bhagat, P.K., Sinha, A.K... A dual-specificity phosphatase, MAP kinase phosphatase 1, positively regulates blue light-mediated seedling development in <i>Arabidopsis</i>	PLANTA	AT1G28930 AT1G67090 AT2G04550 AT2G43790 AT2G46340 AT3G09110 AT3G23610 AT3G55270 AT4G39730 AT5G11260 AT5G13930 AT5G23720 AT5G24120 AT5G40440		2021
Kiselev, K.V., Suprun, A.R., Aleyno... External dsRNA Downregulates Anthocyanin Biosynthesis-Related Genes and Affects Anthocyanin Accumulation in <i>Arabidopsis thaliana</i>	INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES	AT1G71030 AT1G77450 AT5G13930		2021
Markus, C., Pecinka, A., Merotto, A... Insights into the Role of Transcriptional Gene Silencing in Response to Herbicide-Treatments in <i>Arabidopsis thaliana</i>	INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES	AT2G36490 AT5G13930		2021
Zhang, X., He, Y., Li, L., Liu, H., Hon... Involvement of the R2R3-MYB transcription factors MYB21 and its homologs in regulating the stamen flavonols accumulation in <i>Arabidopsis</i>	JOURNAL OF EXPERIMENTAL BOTANY	AT1G30135 AT2G47460 AT3G01530 AT3G27810 AT3G62610 AT5G08640 AT5G13930 AT5G40348 AT5G40350 AT5G49330 FLS1		2021
Yu, Z.C., Zheng, X.T., Lin, W., He,... Photoprotection of <i>Arabidopsis</i> leaves under short-term high light treatment: The antioxidant capacity is more important than the anthocyanin shielding 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 <i>Arabidopsis thaliana</i>	JOURNAL OF BIOLOGICAL CHEMISTRY	AT5G07990 AT5G08640 AT5G13930		2020
Gayomba, S.R., Muday, G.K... Flavonols regulate root hair development by modulating accumulation of 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 <i>Arabidopsis thaliana</i> roots	BIOTECHNOLOGY LETTERS	AT1G50420 AT3G51240 AT3G55120 AT5G05270 AT5G13930		2020
Naaike, T., Maeda, H.A., Proost, S.... Kingdom-wide analysis of the evolution of the plant type III polyketide synthase superfamily	PLANT PHYSIOLOGY	AT1G02050 AT4G34850 AT5G13930		2020
Hamamouch, N., Winkel, B.S.J., Li... Modulation of <i>Arabidopsis</i> Flavonol Biosynthesis Genes by Cyst and Root-Knot Nematodes	PLANTS (BASEL)	AT2G47460 AT5G08640 AT5G13930		2020
Cheng, W., Lin, M., Qiu, M., Kong,... Chitin synthase is involved in vegetative growth, asexual reproduction, and pathogenesis of <i>Phytophthora capsici</i> and <i>P. sojae</i>	ENVIRONMENTAL MICROBIOLOGY	AT2G33580 AT3G21630 AT5G13930		2019
Nakayama, T., Takahashi, S., Waki, T... Formation of Flavonoid Metabolites: Functional Significance of Protein-Protein Interactions and Impact on Flavonoid Chemodiversity	FRONT PLANT SCI	AT5G13930		2019
Rai, N., Neugart, S., Yan, Y., Wang,... How do cryptochromes and UVR8 interact in natural and simulated sunlight?	JOURNAL OF EXPERIMENTAL BOTANY	AT5G13930 AT5G63860		2019

View Complete List (15 of 96 displayed)

<https://www.arabidopsis.org/>

Phytozome

- 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



<https://phytozome-next.jgi.doe.gov/>
Goodstein et al., 2012: 10.1093/nar/gkr944

PLAZA

- 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

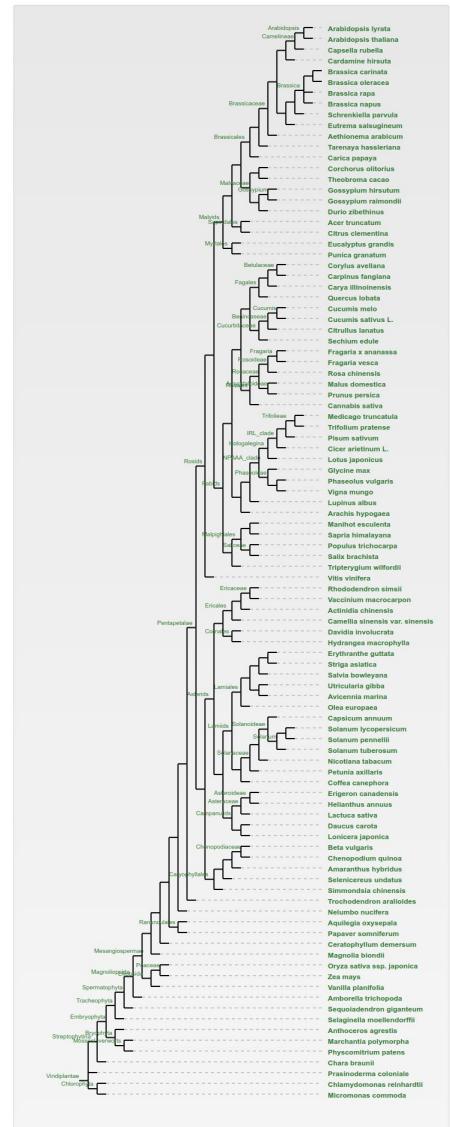
<https://bioinformatics.psb.ugent.be/plaza/>
Van Bel et al., 2021: 10.1093/nar/gkab1024

News

• Public release PLAZA 5.0 2021-09-13

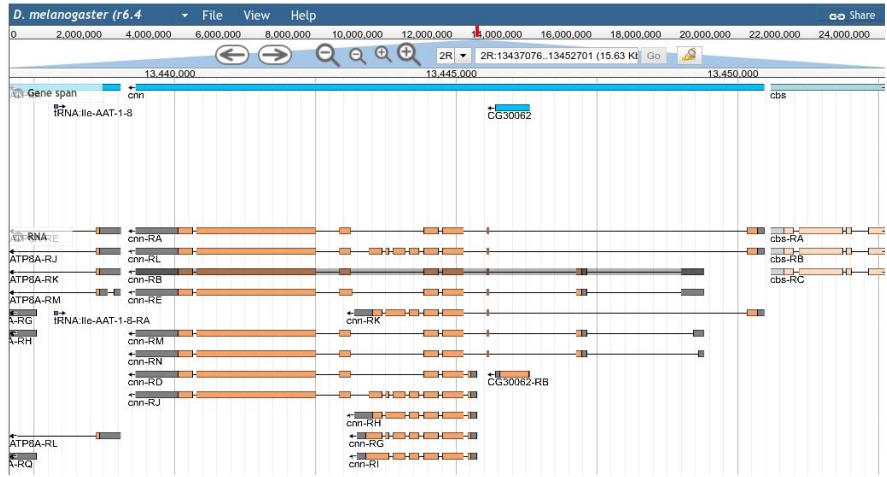
Dicots PLAZA 5.0 summary

- Integration of structural and functional annotation of 100 species.
- Includes 4,234,318 genes, of which 86.6% are protein-coding genes. These coding genes are clustered in 68,306 multi-gene gene families (47.5% multi-species gene families).
- More information on the data content can be found in the data overview



FlyBase

- 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



<https://flybase.org/>
<https://doi.org/10.1093/nar/26.1.85>

WormBase

- 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	CGATAAACATCAACGGCATAAT	TTTGAAACTGATATAGAGGGGCA	1188
assay_2	AAGGTTATTATGCGGTGGAAT	AGCCTTGTAGCTTGATGAAATC	2191
assay_3	AGATTGGAACGATAACGCAGATA	TTTGCCAATTGCAATTATTTT	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

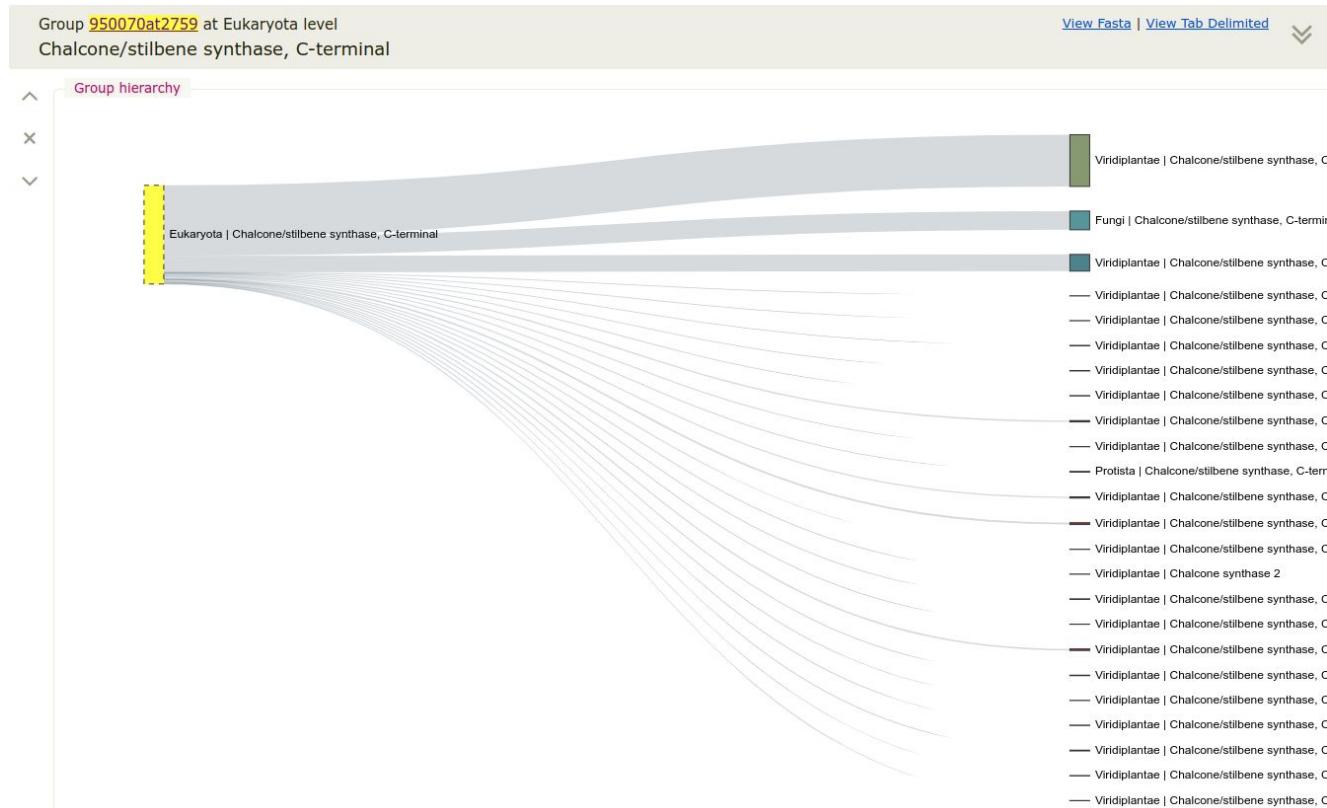
Allow product length difference of: bp Allow oligo mismatches of: bp

Results in: Text-only format HTML with hyperlinks

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

OrthoDB

- 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 INSDC homepage with a light blue header featuring the INSDC logo and the text "International Nucleotide Sequence Database Collaboration". Below the header are navigation links for "ABOUT INSDC", "POLICY", "ADVISORS", and "DOCUMENTS". The main content area has a dark blue sidebar on the left with logos for NCBI, DDBJ, and ENA. The main content area contains text about the collaboration and a table comparing data types across the three databases. At the bottom, there's information on how to submit data and a footer with the INSDC logo and copyright details.

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	Sequence Read Archive	Sequence Read Archive
Capillary reads	Trace Archive	Trace Archive	Trace Archive
Annotated sequences	DDBJ	GenBank	GenBank
Samples	BioSample	BioSample	BioSample
Studies	BioProject	BioProject	BioProject

<https://www.insdc.org/>

NCBI databases

- 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



NCBI - BLAST

- 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 tblasn_vdb executables in the BLAST+ distribution.

Thu, 17 Mar 2022 12:00:00 EST [More BLAST news...](#)

Web BLAST

Nucleotide BLAST
Nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

BLAST Genomes
Enter organism common name, scientific name, or tax id **Search**
Human Mouse Rat Microbes

Standalone and API BLAST

[Download BLAST](#)
Get BLAST databases and executables

[Use BLAST API](#)
Call BLAST from your application

[Use BLAST in the cloud](#)
Start an instance at a cloud provider

Specialized searches

SmartBLAST
Find proteins highly similar to your query

Primer-BLAST
Design primers specific to your PCR template

Global Align
Compare two sequences across their entire span (Needleman-Wunsch)

CD-search
Find conserved domains in your sequence

IgBLAST
Search immunoglobulins and T cell receptor sequences

VecScreen
Search sequences for vector contamination

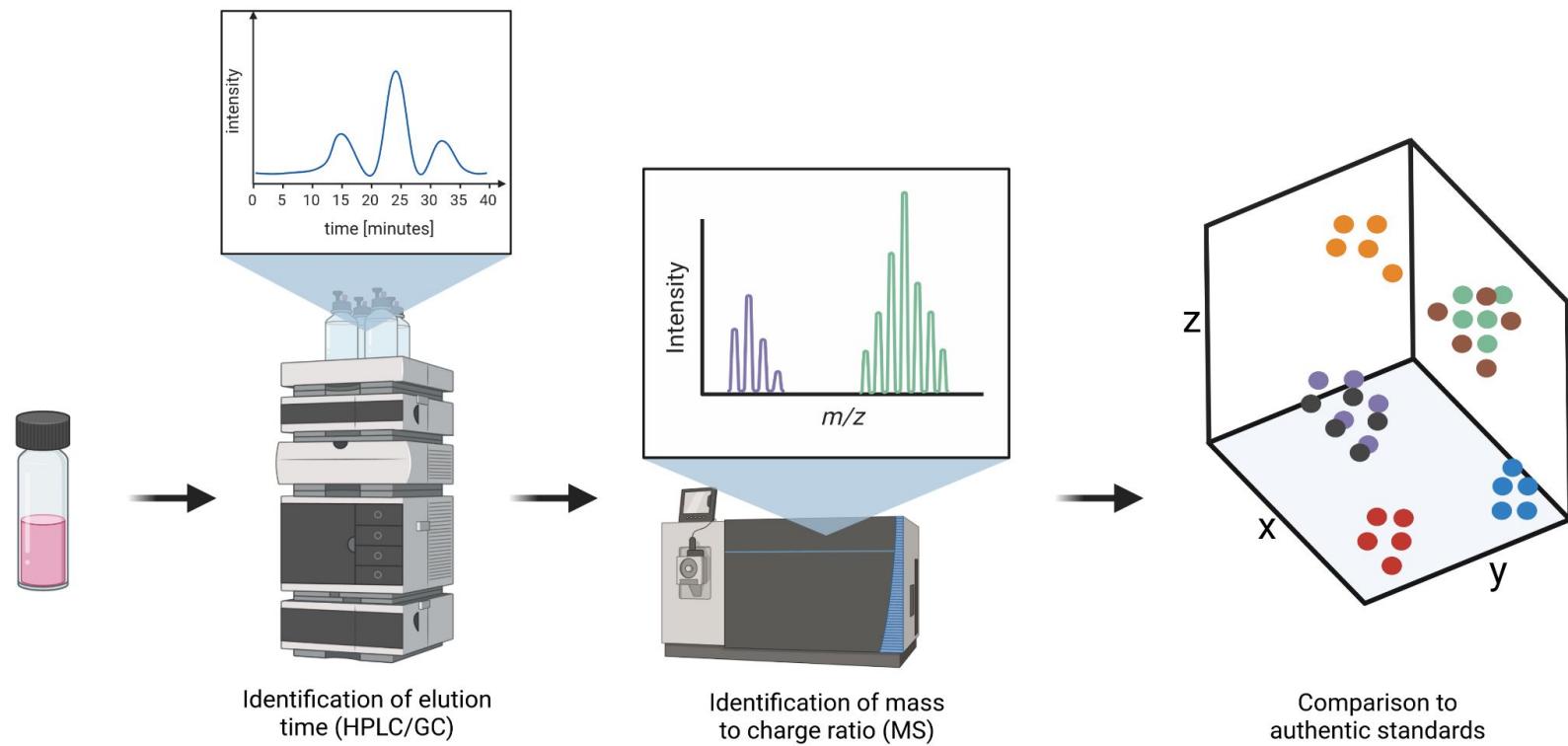
CDART
Find sequences with similar conserved domain architecture

Multiple Alignment
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



MetaboLights

- 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 2016-08-23

MTBLS333: 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, focussing on the compounds showing a higher concentration in yellow varieties.

Methods:

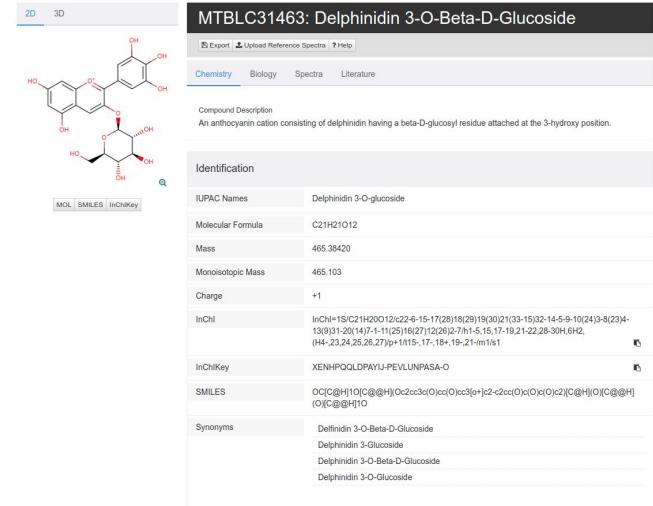
The metabolomic profile of 13 red and 12 yellow raspberries (of different varieties, locations and collection dates) was determined by UPLC-TOF-MS and a novel correlation based approach was implemented to extract the pseudospectra of the most relevant biomarkers from high energy LC-MS runs.

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/MS experiments and performing standard injections.

Conclusion:

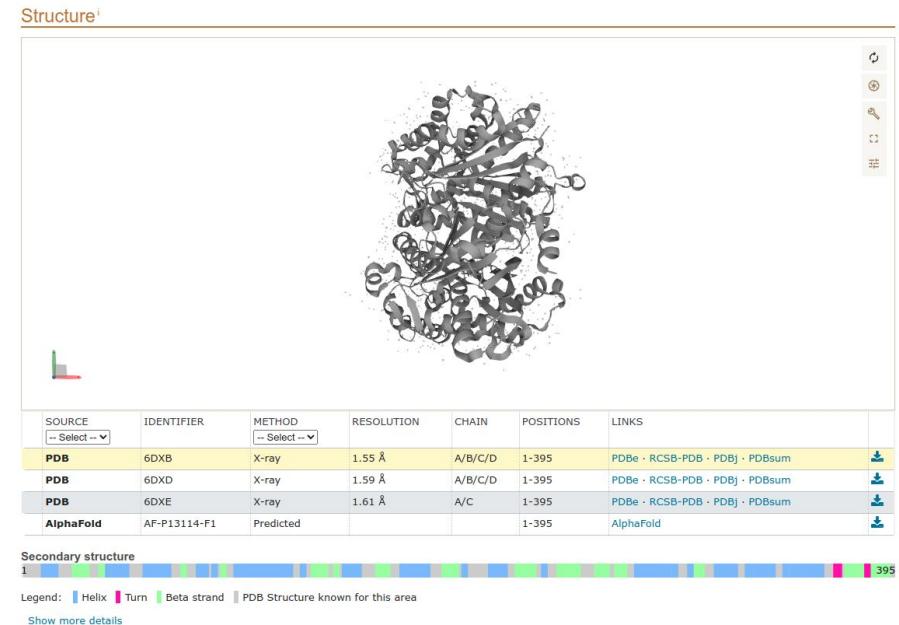
In berries lacking anthocyanins the polyphenol metabolism might be shifted to the formation of a novel class of A-type procyanidin polymers.



<https://www.ebi.ac.uk/metabolights/>

Identification of enzymes - UniProt/SwissProt

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



PDB: Protein Data Bank

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

PDB PROTEIN DATA BANK

100,000 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

POB Archive | Advanced Search | Browse Annotations

Search Query History Browse Annotations MyPDB

Full Text "xanone synthase"

JSON MyPDB Login Help

Advanced Search Query Builder

Full Text (1)

Structure Attribute (1)

Chemical Attribute (1)

Sequence (1)

Sequence Motif (1)

Structure Similarity (1)

Structure Motif (1)

Chemical (1)

Add Subquery Count Remove Subquery

Return Structures grouped by No Grouping

Refinements (1)

SCIENTIFIC NAME OF SOURCE ORGANISM (1)

Homo sapiens (5013)

Mus musculus (7576)

Escherichia coli (16)

Arabidopsis thaliana (12389)

Rattus norvegicus (352)

Batrach taurina (3549)

Securidotea cernuana (2995)

Gelius galii (2179)

Securidotea cernuana S28BC (2156)

More...

TAXONOMY

Eukaryota (10265)

Bacteria (46452)

Archaea (15798)

Viruses (1648)

Archaea (5081)

Diplobionta (2710)

Vardovidae (620)

Mitochondria (1015)

Prokaryotic sequences (378)

Nastastidae (48)

100,000 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

POB Archive | Advanced Search | Browse Annotations

Search Query History Browse Annotations MyPDB

Full Text "xanone synthase"

JSON MyPDB Login Help

Advanced Search Query Builder

Full Text (1)

Structure Attribute (1)

Chemical Attribute (1)

Sequence (1)

Sequence Motif (1)

Structure Similarity (1)

Structure Motif (1)

Chemical (1)

Add Subquery Count Remove Subquery

Return Structures grouped by No Grouping

Refinements (1)

SCIENTIFIC NAME OF SOURCE ORGANISM (1)

Homo sapiens (5013)

Mus musculus (7576)

Escherichia coli (16)

Arabidopsis thaliana (12389)

Rattus norvegicus (352)

Batrach taurina (3549)

Securidotea cernuana (2995)

Gelius galii (2179)

Securidotea cernuana S28BC (2156)

More...

TAXONOMY

Eukaryota (10265)

Bacteria (46452)

Archaea (15798)

Viruses (1648)

Archaea (5081)

Diplobionta (2710)

Vardovidae (620)

Mitochondria (1015)

Prokaryotic sequences (378)

Nastastidae (48)

1EYQ

Chalcone isomerase and naringenin
Jez, J.M., Bowman, M.E., Rosta, A.R., Noel, J.P.
(2000) Nat Struct Biol 7: 786-791

Released 2000-09-05

Method X-RAY DIFFRACTION 1.8 Å

Organism *Medicago sativa*

Macromolecule CHALCONE-FLAVONE ISOMERASE 1 (protein)

Unique Ligands NAR, SOD

Download File | View File

1FMB

CHALCONE ISOMERASE COMPLEXED WITH 5,4-DIDEOXYFLAVANONE
Jez, J.M., Noel, J.P.
(2002) J Biol Chem 277: 1361-1369

Released 2001-12-12

Method X-RAY DIFFRACTION 2.3 Å

Organism *Medicago sativa*

Macromolecule CHALCONE-FLAVONE ISOMERASE 1 (protein)

Unique Ligands DDC, SOD

Download File | View File

1FMB

CHALCONE ISOMERASE COMPLEXED WITH 5,4-DIDEOXYFLAVANONE
Jez, J.M., Noel, J.P.
(2002) J Biol Chem 277: 1361-1369

Released 2001-12-12

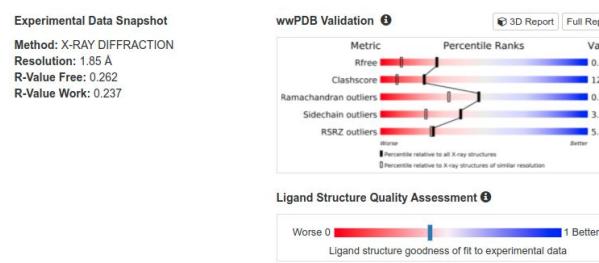
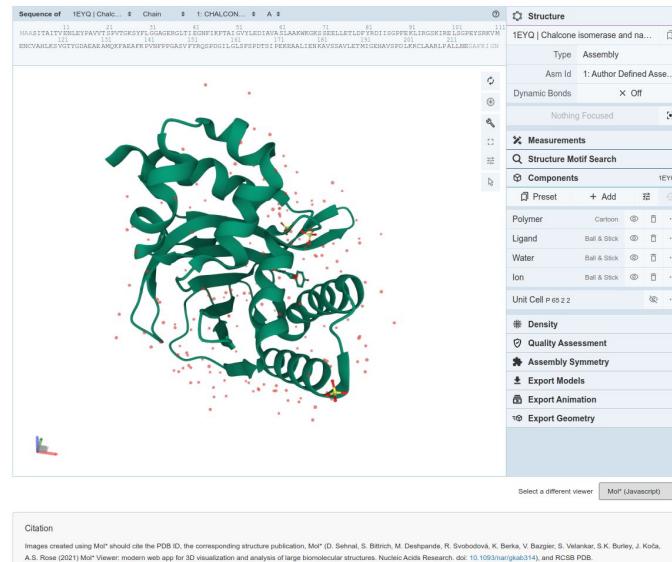
Method X-RAY DIFFRACTION 2.3 Å

Organism *Medicago sativa*

Macromolecule CHALCONE-FLAVONE ISOMERASE 1 (protein)

Unique Ligands DDC, SOD

Download File | View File



This is version 1.2 of the entry. See complete [history](#)

<https://www.rcsb.org/>
https://spdbv.unil.ch/main_tut.html
Zardecki et al., 2022: 10.1002/pro.4200

Chimera & PyMol

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



[about](#) [projects](#) [people](#) [publications](#)

[resources](#) [visit us](#) [search](#)

UCSF CHIMERA

an Extensible Molecular Modeling System

UCSF Chimera is a program for the interactive visualization and analysis of molecular structures and related data, including density maps, trajectories, and sequence alignments. It is available free of charge for noncommercial use. Commercial users, please see [Chimera commercial licensing](#).

We encourage Chimera users to try [ChimeraX](#) for much better performance with large structures, as well as other major [advantages](#) and completely new features. ChimeraX includes a significant subset of Chimera features (with more to come, see the [missing features list](#)) and is under 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](#) (R41-GM033311) that ended in 2018.

What's New in Daily Builds

[Main Download Locations](#)

Get Started

[User's Guide](#)

[Command Index](#)

[Tutorials and Videos](#)

[Guide to Volume Data](#)

[Release Notes](#)

Download

[What's New in Daily Builds](#)

[Main Download Locations](#)

Get Started

[User's Guide](#)

[Command Index](#)

[Tutorials and Videos](#)

[Guide to Volume Data](#)

[Release Notes](#)

Documentation

[Getting Started](#)

[User's Guide](#)

[Command Index](#)

[Tutorials and Videos](#)

[Guide to Volume Data](#)

[Release Notes](#)

Chimera Search

[Go](#)

[Google+ Search](#)

News

December 20, 2021

The RBVI wishes you a safe and happy holiday season and the gathering of friends and family. [rbvi.usnews.org](#) back to 1984

December 17, 2021

Chimera production release 1.16 is now ready. This release includes the [release notes](#) for what's new. See the [release notes](#) for what's new.

[\(Previous news...\)](#)

December 18, 2020

Chimera production release 1.15 is now ready. This release includes the [release notes](#) for what's new. See the [release notes](#) for what's new.

[\(Previous news...\)](#)

Upcoming Events

Feature Highlight

Interactive Shadows

Interactive shadows (shadows that move as structures are moved) can be enabled in the [Effects](#) dialog or with the command `set shadow`. Click the image to show a small molecule rotating above a rectangular plane. This simple animation was made with the Chimera command script [tumble.com](#). Interactive shadows also work with ribbons, surfaces, and other representations.

[\(More features...\)](#)

Gallery Sample

RNA Bases

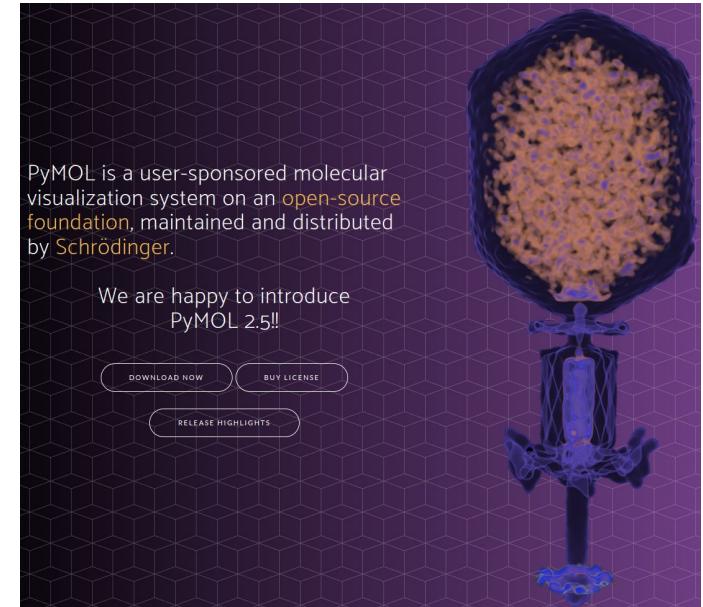
Large ribosomal RNA is shown with individual bases depicted using solvent excluded molecular surfaces. Bases A, C, G, U are colored red, yellow, green, and blue. The surfaces were made with the Chimera multiscale tool in combination with the [nucleic acid tools](#) plug-in. The image was raytraced using PovRay.

Protein Data Bank model 1a7z.

[\(More samples...\)](#)

Activation mechanism of the class D fungal GPCR *Dimer Star* Venkatahannan, S., et al. *Nature*, 2022 Mar 24;603(7492):743-748.

[\(Previously featured citations...\)](#)



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 androsaemum (Tutsan) [go to search ↗](#)
UniProt Q9FBU7 [go to UniProt ↗](#)
Experimental structures None available in the PDB
Biological function The primary product of this enzyme is 4,2',4',6'-tetrahydroxylchalcone (also termed naringenin-chalcone or chalcone) which undergoes enzymecatalyzed or spontaneous isomerization into naringenin. [go to UniProt ↗](#)

3D viewer

Model Confidence:
Very high (pLDOT > 90)
Confident (90 > pLDOT > 70)
Low (70 > pLDOT > 50)
Very low (pLDOT < 50)

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

Sequence of AF-Q9FBU7.F1 1 Chalcone s... A 4

Views

Predicted aligned error

Sequence Features (coming soon)

Aligned residue

Scored residue

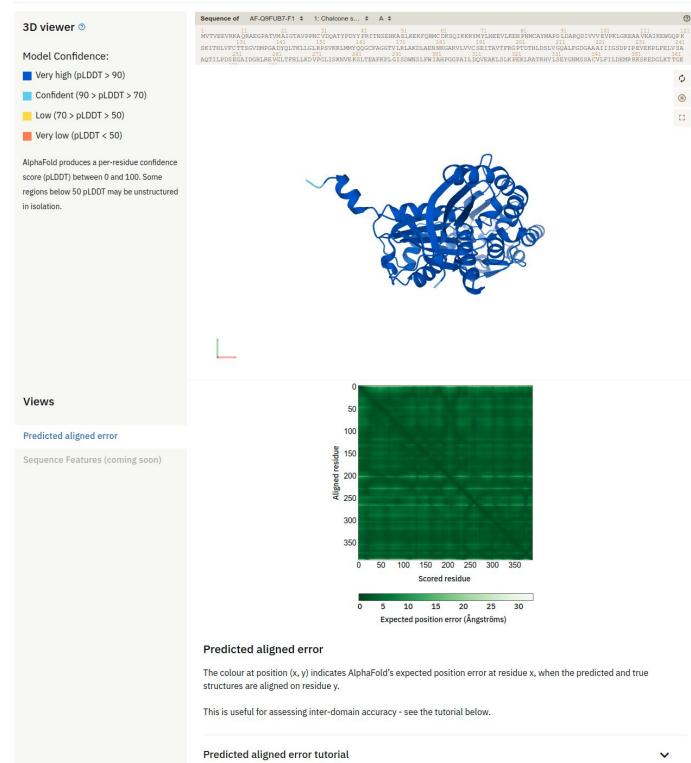
Expected position error (Ångströms)

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



<https://alphafold.ebi.ac.uk/>

Jumper et al., 2021: 10.1038/s41586-021-03819-2



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 form is for debugging purposes only and is **not supported**. To submit jobs to InterProScan 5, please visit the [InterPro Sequence Search](#) or the [InterProScan 5 Web services](#).

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:KPYMF_HUMAN

Or, upload a file Choose File | No file chosen Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Select the applications to run

TIGRFAM SFLD Phobius SignalP SignalP_EUK
 SignalP_GRAM_POSITIVE SignalP_GRAM_NEGATIVE SUPERFAMILY PANTHER Gene3D
 Hmmp ProSiteProfiles ProSitePatterns Coils SMART
 CDD PRINTS Pfam MobiDBlite
 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

sp P13114 CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	Gene3D	G3DSA:3.40.47.10	-	1	241	4.8E-101	T	20-03-2022	
sp P13114 CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	SUPERFAMILY	SSF53901	Thiolase-like	241	393	2.98E-51	T	20-03-20	
sp P13114 CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	SUPERFAMILY	SSF53901	Thiolase-like	10	237	2.38E-78	T	20-03-20	
sp P13114 CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	Pfam	PF02797	Chalcone and stilbene synthases, C-terminal domain	244	394	1.5E-71			
sp P13114 CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	PANTHER	PTHR11877:SF81	BNAA02G30320D PROTEIN	6	394	0.0	T	20-03-2022	
sp P13114 CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	Gene3D	G3DSA:3.40.47.10	-	242	395	9.7E-62	T	20-03-2022	IPR01603
sp P13114 CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	PANTHER	PTHR11877	HYDROXYMETHYLGLUTARYL-COA SYNTHASE	6	394	0.0	T		
sp P13114 CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	CDD	cd00831 CHS_like	21	390	0.0	T	20-03-2022	-	-
sp P13114 CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	PIRSF	PIRSF000451	PKS_III	7	394	0.0	T	20-03-2022	IPR011141
sp P13114 CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	Pfam	PF00195	Chalcone and stilbene synthases, N-terminal domain	10	233	2.9E-119			
sp P13114 CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	ProSitePatterns	PS00441	Chalcone and stilbene synthases active site.	161	177	-			

http://www.ebi.ac.uk/Tools/services/web/toolform.ebi?tool=iprscan5&sequence=uniprot:KPYMF_HUMAN
Jones et al., 2014: 10.1093/bioinformatics/btu031



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
```

<http://amigo.geneontology.org/amigo/term/GO:0009813>
The Gene Ontology Consortium, 2007: 10.1093/nar/gkm883

AmiGO

- 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»	<input type="button" value="Custom DL (up to 100000)"/>	Bookmark
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)

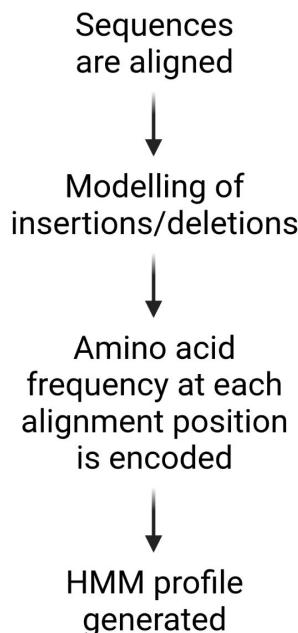
Pfam: Protein family database

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

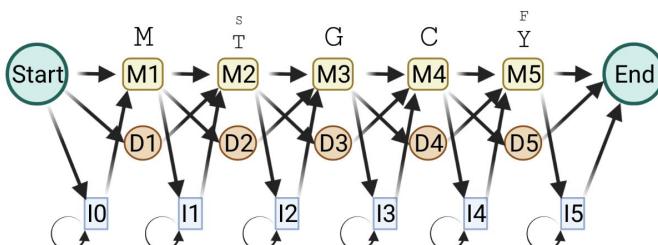
Pfam 35.0 (November 2021, 19632 entries)

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

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



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



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](#) [Go](#) [Example](#)

Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB 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 Bileschi 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. Paladini, S. Raj, L.J. Richardson, R.D. Finn, A. Bateman

[Nucleic Acids Research](#) (2020) doi: 10.1093/nar/gkaa913

Pfam is part of the ELIXIR infrastructure
Pfam is an ELIXIR service [Read more](#)

Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk
European Molecular Biology Laboratory

<http://pfam.xfam.org/>

Mistry et al., 2020: 10.1093/nar/gkaa913

Mercator - functional annotation of protein sequences

- 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
MIGPMPGMEGKLMPLPGASPVGLEVVLVASTPILSDTSLCTPSFYHFLLLGPITSISNLIVRPFLLSSITVIYGRLCFFAFDFYVY
>TRINITY_DN100016_c0_g1_i1
MPKKSSNIQQNNQNRRIKRKGRRPKKKYVQQIDSSDEDILSVRHASTRPRIISIRRNEIPMRPEIHIPPTSIPSFNNPNNIHQSSSDRQMPP
>TRINITY_DN100019_c0_g1_i1
MPPKAADKKPAAKAPATASKAPEKKDAGKKTAAATGDKKRTKARKETYSSYIYKVLKQVHPDTGISNRAMSILNSFVNDFERVATEASKLA
>TRINITY_DN10001_c0_g1_i2
MAKVGNPVIDETDGSVNEPESSEKNIEVSSSTQAPESTNTTELLVNEKKAFSLATPAVRRVAREHNIDINNIKGTGKNGRITKEDILNYV
>TRINITY_DN100025_c0_g1_i1
MVENQDGCFKPGWKEFVRNSDLEGGDFLVNLVDKISYQVVIFDGTCACPDKLCPFSIMNPIFIQHLRNKIFLSKKEEIKLKGNRKVHSVNEN
>TRINITY_DN10002_c0_g1_i1
MSDELNQVFQRCREQKRPVVFVAFVTAGYPDSEETVDILLGLEAGGADIIIELGIPFTDPMVDGKTIQDANNVALENKIDISKCLSYVSESRAK
```

<https://plabipd.de/portal/mercator-sequence-annotation>
Lohse et al., 2014: 10.1111/pce.12231
Haak et al., 2018: 10.3389/fmolb.2018.00062

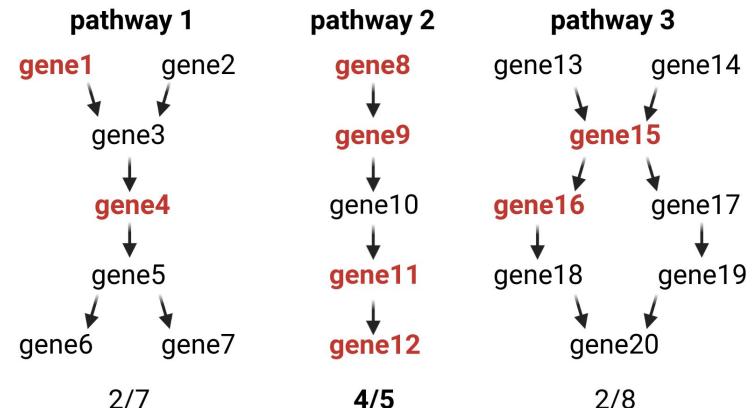
Concept of enrichment analyses

- 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

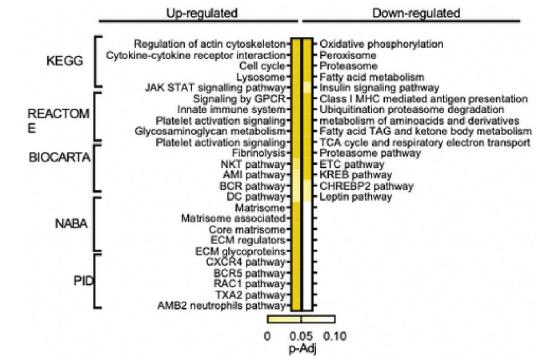
- Functional enrichment analyses can be based on GO terms
- Assignment of GO terms to genes of the species of interest
- Tools for GO enrichment analysis:
 - PANTHER: <http://geneontology.org/>
 - g:Profiler: <http://biit.cs.ut.ee/gprofiler/gost>
 - GOrilla: <http://cbl-gorilla.cs.technion.ac.il/>

g:GOB performs functional enrichment analysis, also known as over-representation analysis (ORA) or gene set enrichment analysis (GSEA). It maps genes to known functional categories and assesses statistically significant enriched terms. We regularly retrieve data from Ensembl database and parasite or metazoa specific versions of Ensembl Genomes, and parasite specific data from Wormbase Parallels. In addition to Gene Ontology, we include pathways from KEGG Reaction and Interaction, metabolic pathways from MetaCyc, and secondary metabolite matches from TR�SPLAC. Disease specificities from Human Disease Atlas encompass both CORUM and human disease phenotypes from Human Phenotype Ontology. g:GOB supports close to 500 organisms and accepts hundreds of identifier types.



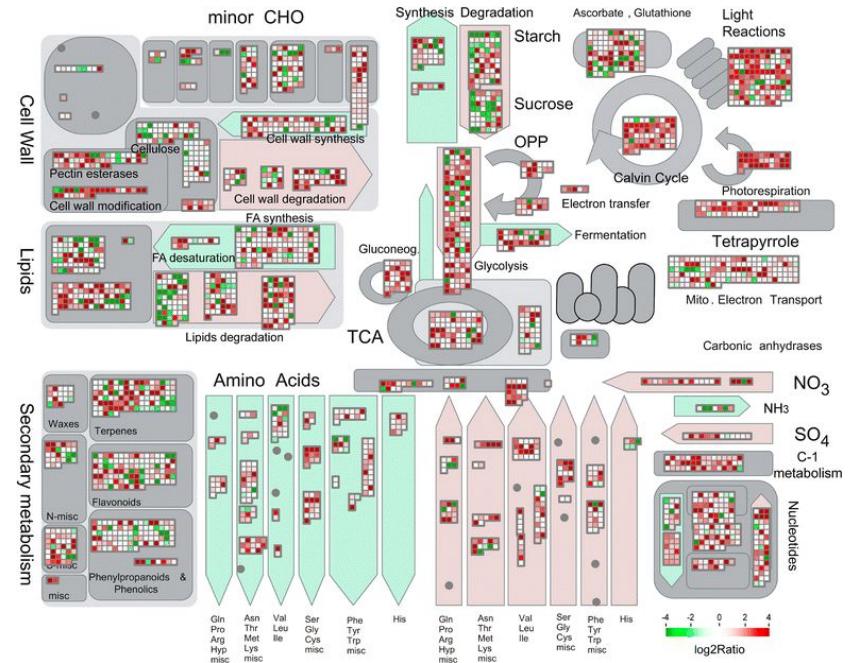
Enrichment analyses - KEGG

- 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?