DATA INTEGRATION AND LINKING OF GENOMIC, METABOLOMIC AND TRANSCRIPTOMIC DATA USING MAPMAN

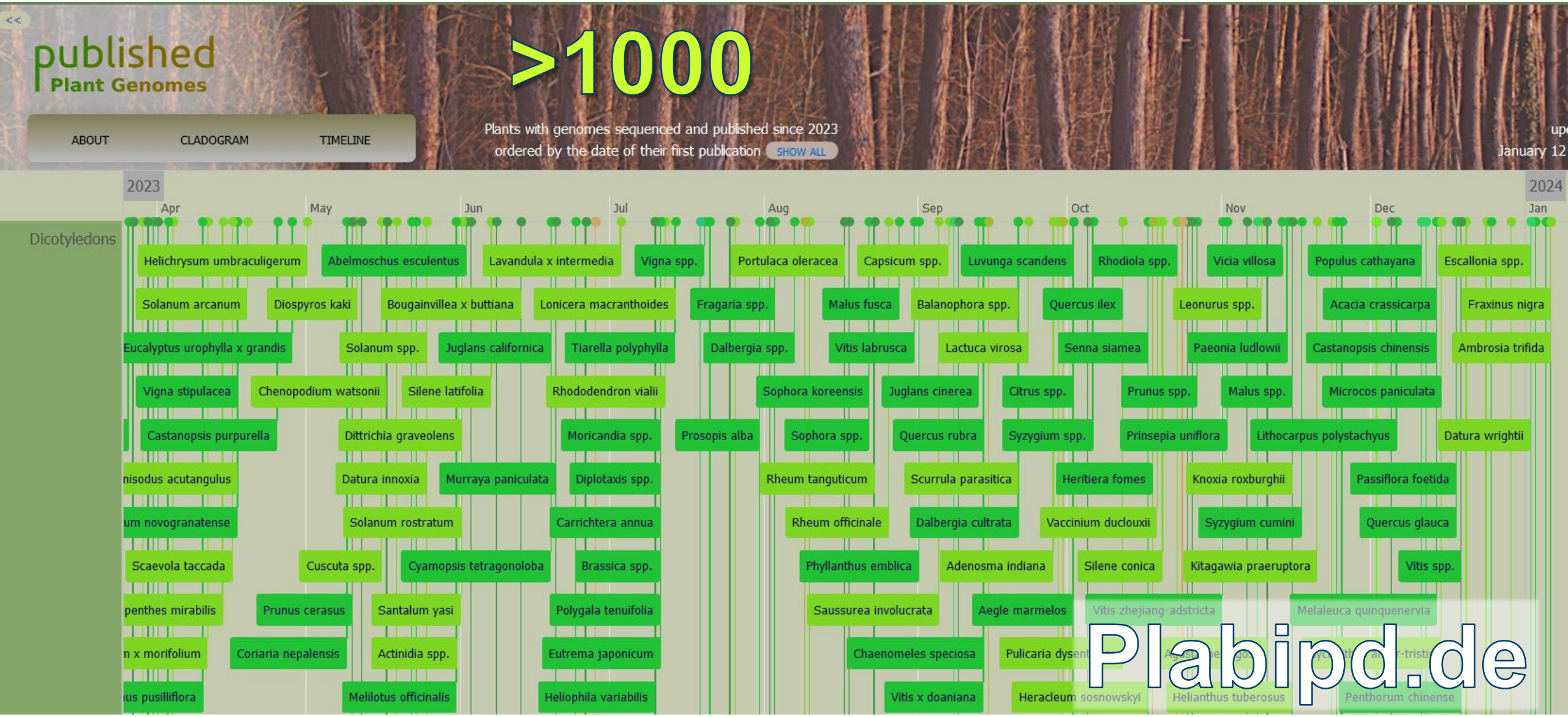
14. FEB. 2024 | BJÖRN USADEL

IBG-4, FORSCHUNGSZENTRUM JUELICH, GERMANY

JÜLICH
Forschungszentrum
Member of the Helmholtz Association

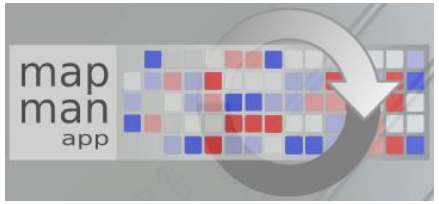
Forschungszentrum Jülich
1.7 km² research campus 11 institutes and over 80 institute departments, 7,100 people

HHU Düsseldorf
Plant Excellence cluster

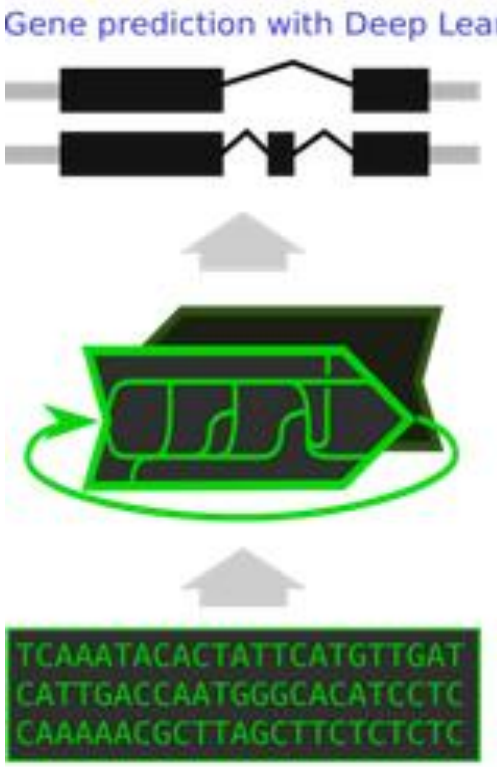

published
Plant Genomes
ABOUT
CLADOGRAM
TIMELINE
>1000
Plants with genomes sequenced and published since 2023
ordered by the date of their first publication
SHOW ALL
January 12
2023
2024
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Dicotyledons
Helichrysum umbraculigerum
Abelmoschus esculentus
Lavandula x intermedia
Vigna spp.
Portulaca oleracea
Capsicum spp.
Luvunga scandens
Rhodiola spp.
Vicia villosa
Populus cathayana
Escallonia spp.
Solanum arcanum
Diospyros kaki
Bougainvillea x buttiana
Lonicera macranthoides
Fragaria spp.
Malus fusca
Balanophora spp.
Quercus ilex
Leonurus spp.
Acacia crassicarpa
Fraxinus nigra
Eucalyptus urophylla x grandis
Solanum spp.
Juglans californica
Tiarella polyphylla
Dalbergia spp.
Vitis labrusca
Lactuca virosa
Senna siamea
Paeonia ludlowii
Castanopsis chinensis
Ambrosia trifida
Vigna stipulacea
Chenopodium watsonii
Silene latifolia
Rhododendron vialii
Sophora koreensis
Juglans cinerea
Citrus spp.
Prunus spp.
Malus spp.
Microcos paniculata
Castanopsis purpurella
Dittrichia graveolens
Moricandia spp.
Prosopis alba
Sophora spp.
Quercus rubra
Syzygium spp.
Prinsepia uniflora
Lithocarpus polystachyus
Datura wrightii
nisodus acutangulus
Datura innoxia
Murraya paniculata
Diplotaxis spp.
Rheum tanguticum
Scurrula parasitica
Heritiera fomes
Knoxia roxburghii
Passiflora foetida
um novogranatense
Solanum rostratum
Carrichtera annua
Rheum officinale
Dalbergia cultrata
Vaccinium duclouxii
Syzygium cumini
Quercus glauca
Scaevola taccada
Cuscuta spp.
Cyamopsis tetragonoloba
Brassica spp.
Phyllanthus emblica
Adenosma indiana
Silene conica
Kitagawia praeruptora
Vitis spp.
penthes mirabilis
Prunus cerasus
Santalum yasi
Polygala tenuifolia
Saussurea involucrata
Aegle marmelos
Vitis zhejiang-adstricta
Melaleuca quinquenervia
n x morifolium
Coriaria nepalensis
Actinidia spp.
Eutrema japonicum
Chaenomeles speciosa
Pulicaria dysent
us pusilliflora
Melilotus officinalis
Heliophila variabilis
Vitis x doaniana
Heracleum sosnowskyi
Helianthus tuberosus
Penthorum chinense
PlabipD.de

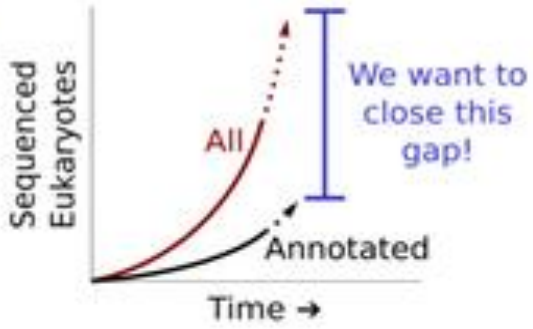
GWAS, GENOMES AND THEN?

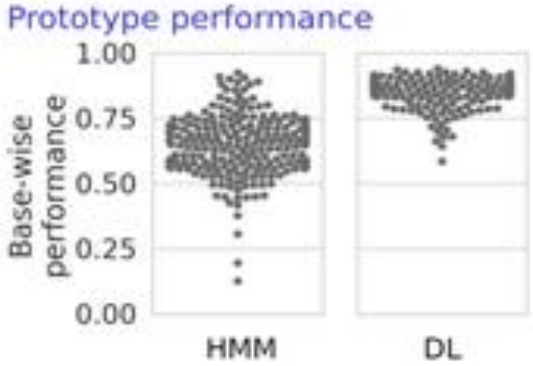
- Gene Finding with **Helixer**
- Functional Gene Prediction with **Mercator4**
- Visualisation of Omics Data with **MapMan**



FINDING GENES IS STILL HARD



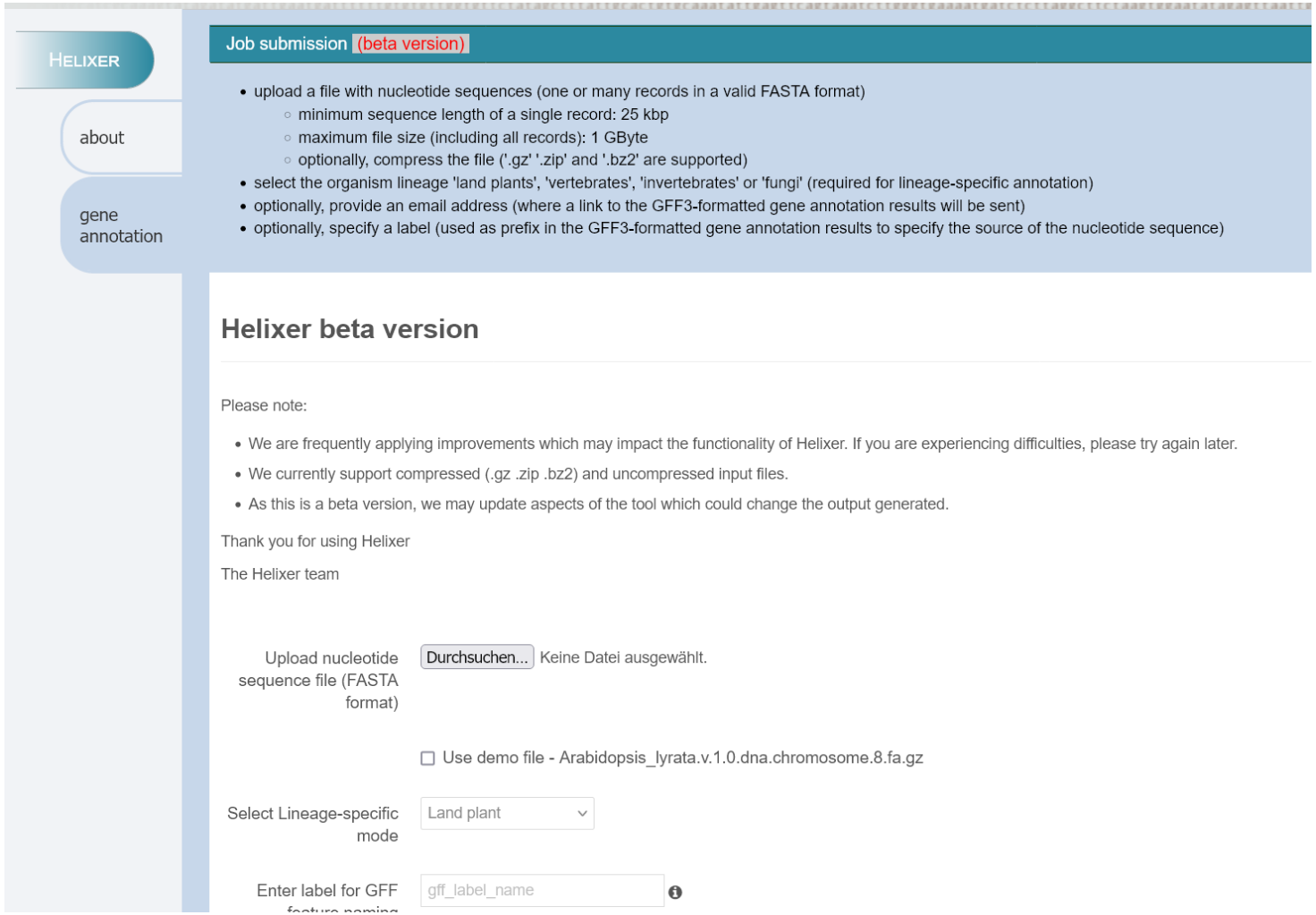




Who has not had to deal with V1, V2 etc of gene annotation? Gene finding is <u>hard.</u>

Helixer uses AI to help – and often outperforms simple “full length” RNA sequencing endeavors

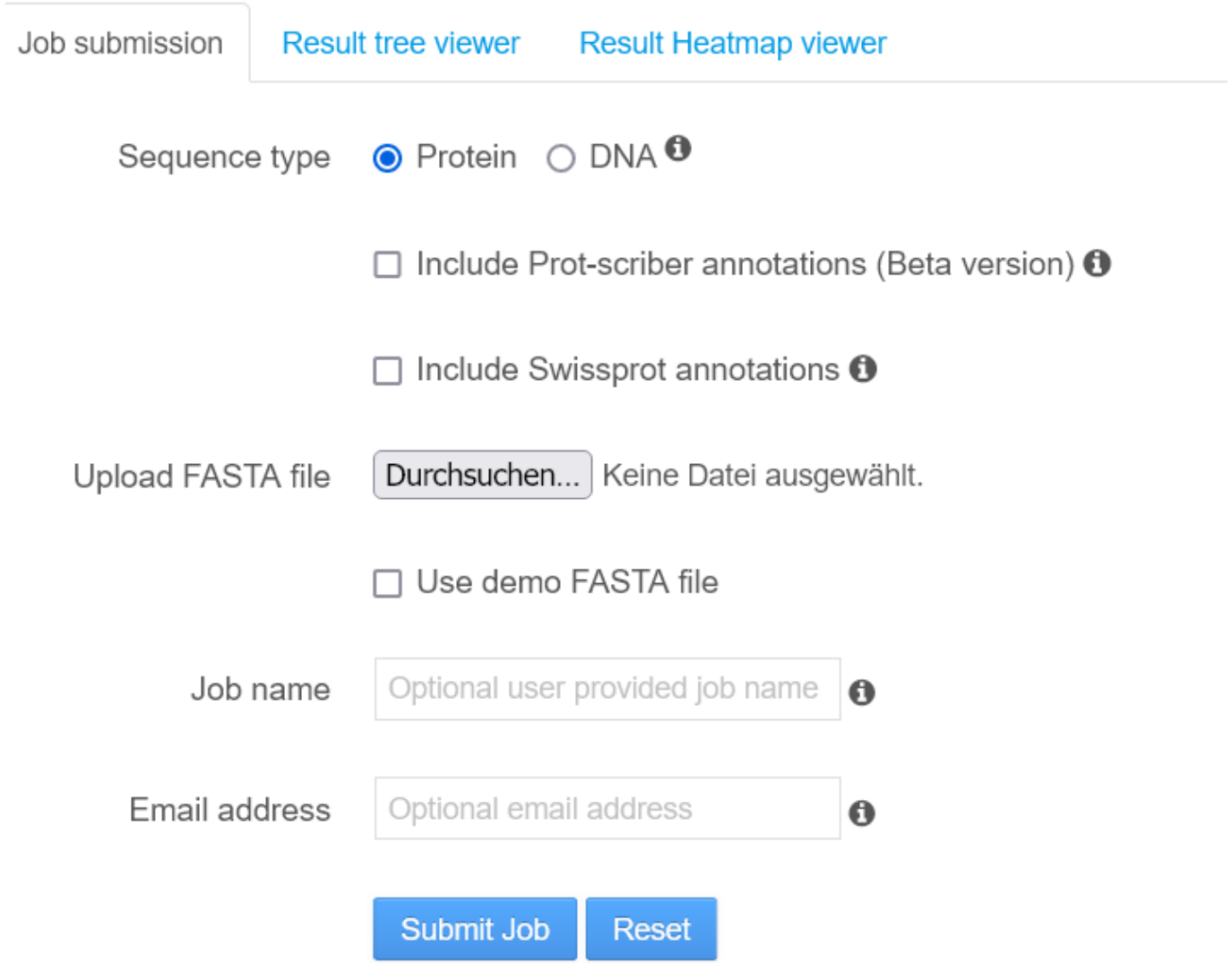
We use both but <u>nothing</u> is perfect. We can still find and improve genes in Arabidopsis now



Meant as a bioinformatics downloadable tool

but “small” genomes ca 1 GBase i.e. most berry crops can be done online

WHAT DO ALL THE GENES/PROTEINS DO?



COMPARATIVE VIEW OF GENE NUMBERS

- 9 | Secondary metabolism
  - 9.1 | terpenoids
    - mevalonate (MVA) pathway
      - 3 2 2 2 acetyl-CoA C-acyltransferase *(ACAT1/2)
      - 3 1 1 1 3-hydroxy-3-methylglutaryl-CoA synthase *(HMGS)
      - 3 2 3 1 3-hydroxy-3-methylglutaryl-CoA reductase *(HMGR)
      - 0 1 1 1 mevalonate kinase *(MVK)
      - 2 1 1 1 phosphomevalonate kinase *(PMK)
      - 1 2 1 1 mevalonate diphosphate decarboxylase *(MVD1/2)
      - 1 2 2 1 isopentenyl diphosphate isomerase *(IDI1/2)
    - methylerythritol phosphate (MEP) pathway
      - 1-deoxy-d-xylulose 5-phosphate import
        - 1 1 1 1 D-xylulose kinase
        - 1 1 1 0 D-xylulose 5-phosphate transporter
      - 7 3 4 2 1-deoxy-D-xylulose 5-phosphate synthase *(DXS)
      - 2 1 1 2 1-deoxy-D-xylulose 5-phosphate reductase *(DXR)
      - 0 1 1 1 4-diphosphocytidyl-2-C-methyl-D-erythritol synthase
      - 1 1 1 1 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase
      - 1 1 1 1 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase

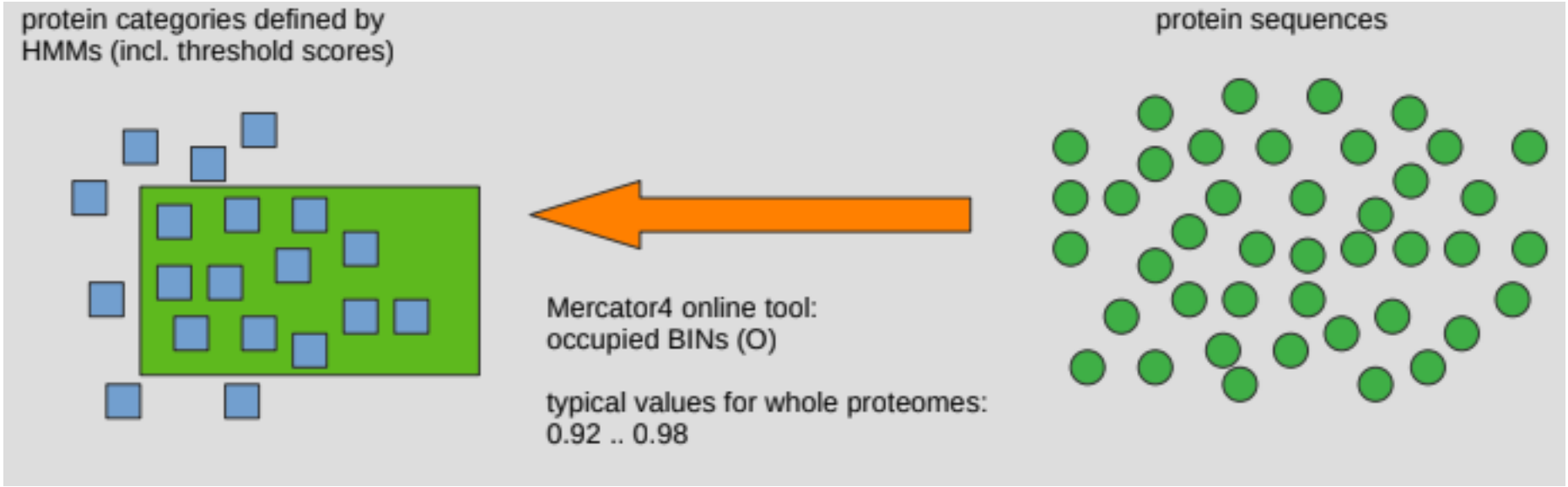
CLASSES FOR ALL LAND PLANTS

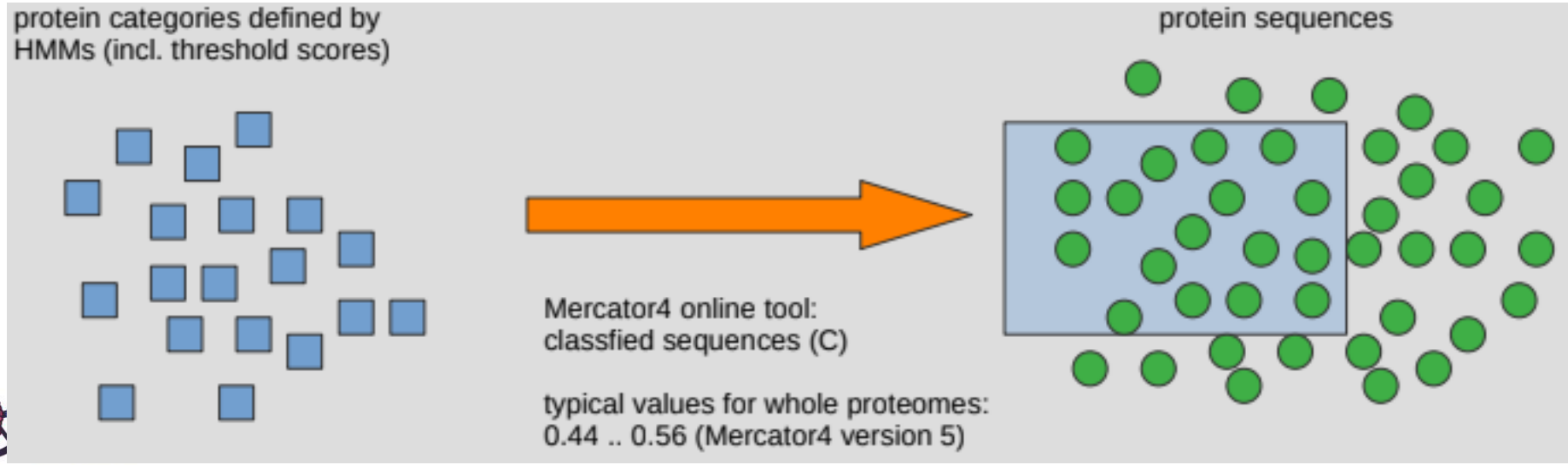
Hierarchical framework consisting of 31-top-level categories (also called *BINs*).

- 1 Photosynthesis
- 2 Cellular respiration
- 3 Carbohydrate metabolism
- 4 Amino acid metabolism
- 5 Lipid metabolism
- 6 Nucleotide metabolism
- 7 Coenzyme metabolism
- 8 Polyamine metabolism
- 9 Secondary metabolism
- 10 Redox homeostasis
- 11 Phytohormone action
- 12 Chromatin organisation
- 13 Cell division
- 14 DNA damage response
- 15 RNA biosynthesis
- 16 RNA processing
- 17 Protein biosynthesis
- 18 Protein modification
- 19 Protein homeostasis
- 20 Cytoskeleton organisation
- 21 Cell wall organisation
- 22 Vesicle trafficking
- 23 Protein translocation
- 24 Solute transport
- 25 Nutrient uptake
- 26 External stimuli response
- 27 Multi-process regulation
- 28 Plant reproduction
- 30 Clade-specific metabolism

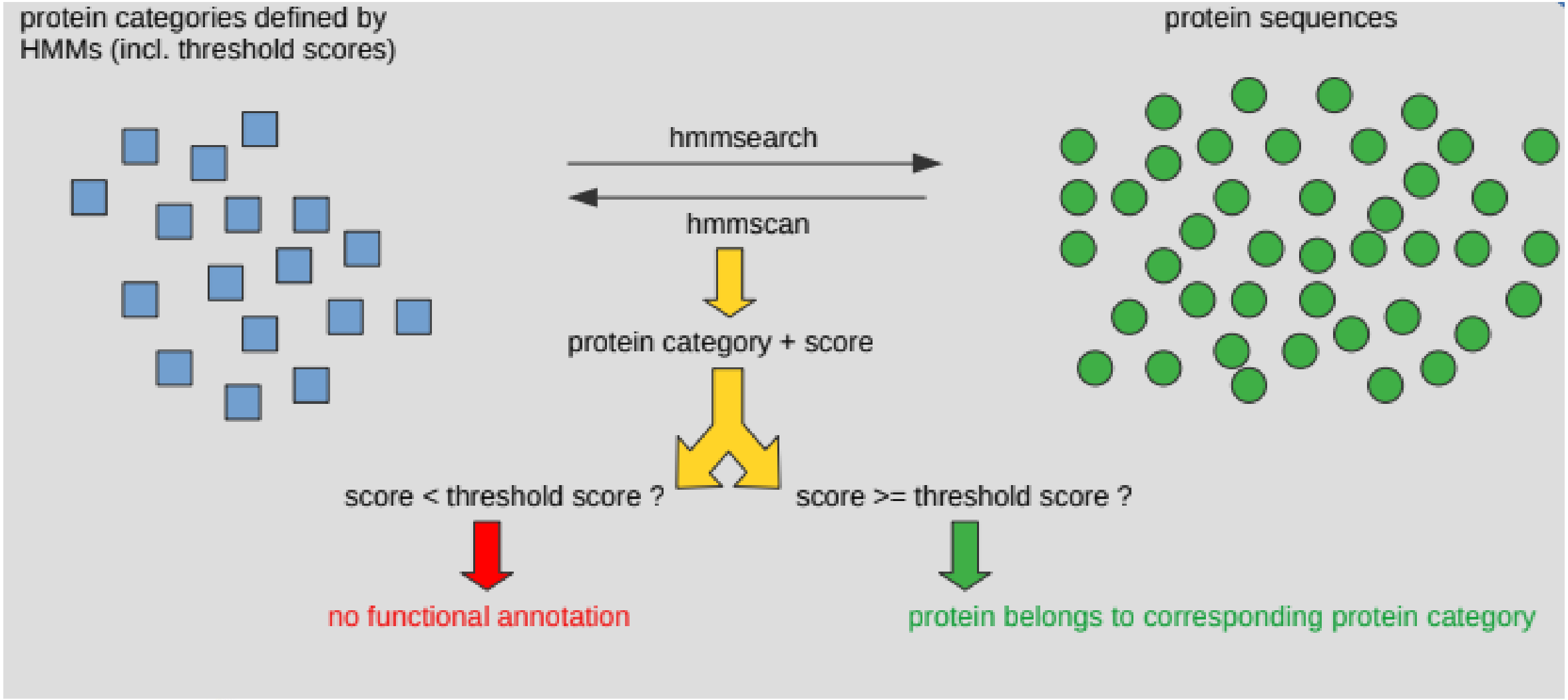
- 50 Enzyme classification

QUECK QUALITY OF PROTEOME





FUNCTIONAL ANNOTATION OF PLANT PROTEOMES

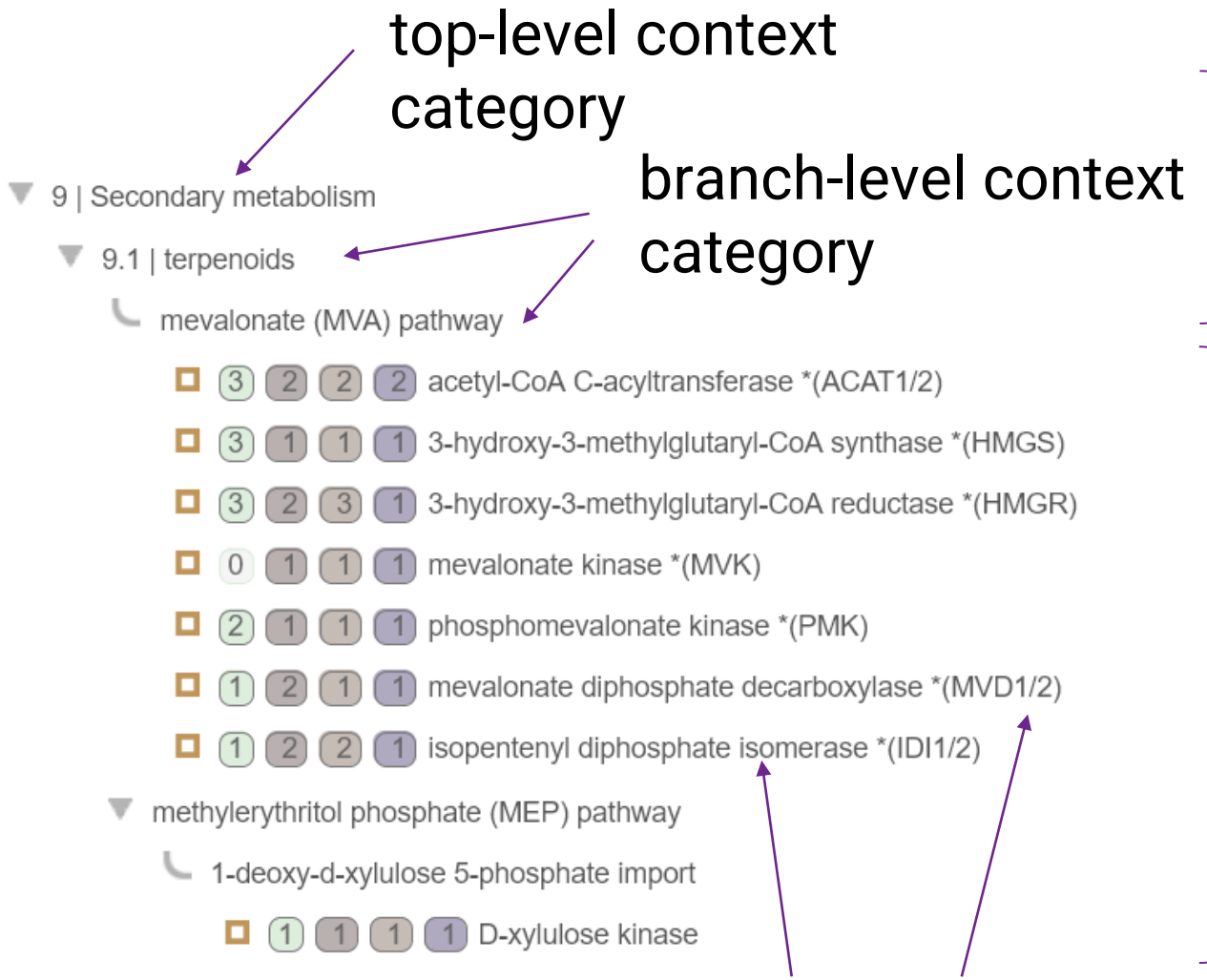


hmmsearch (protein HMM query *vs* protein database) &
hmmscan (protein query *vs* protein HMM database)

BIN STRUCTURE

Each child node is more specialised than its parent *BIN*. Protein sequences are assigned to leaf level.

Soon… Enzyme classes linked to metabolic reactions



VERSION HISTORY

Annual release cycle (late summer -Autumn)

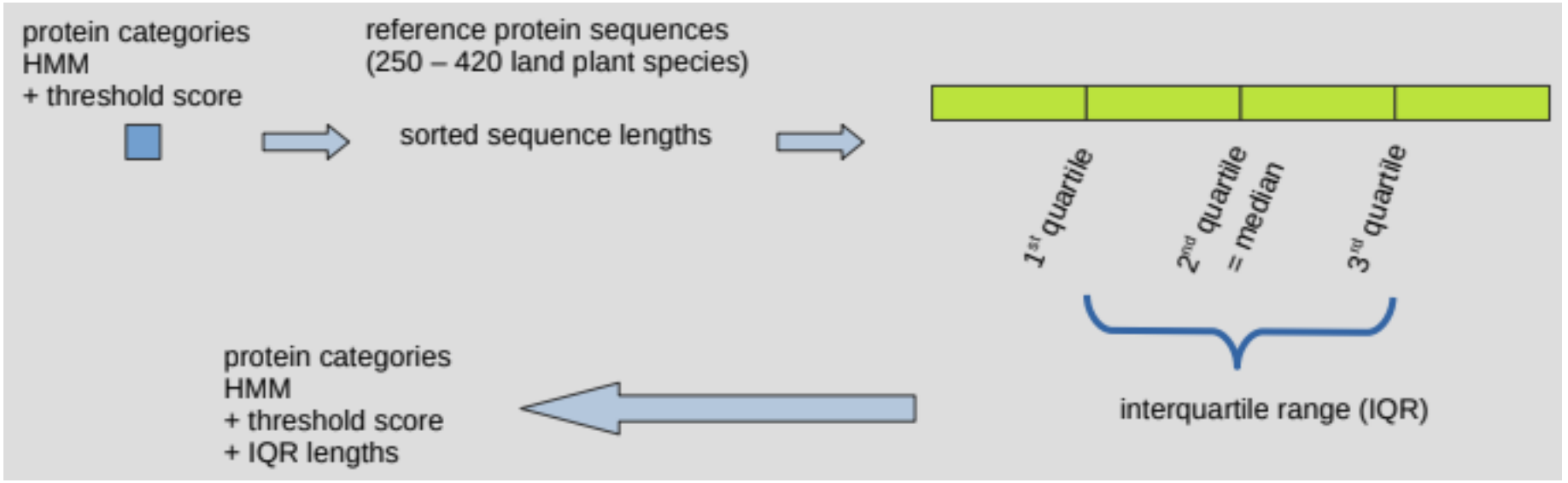
version	0.6 (2017)	1.0 (2018)	2.0 (2019)	3.0 (2020)	4.0 (2021)	5.0 (2022)	6.0 (2023)
protein categories + context nodes	3395 + 1068	4145 + 1339	4500 + 1491	4869 + 1608	5251 + 1702	5783 + 1817	~ 6180 + 1940

VERSION HISTORY

Annual release cycle (late summer -Autumn)

version	0.6 (2017)	1.0 (2018)	2.0 (2019)	3.0 (2020)	4.0 (2021)	5.0 (2022)	6.0 (2023)
protein categories + context nodes	3395 + 1068	4145 + 1339	4500 + 1491	4869 + 1608	5251 + 1702	5783 + 1817	~ 6180 + 1940
Classified		47.88	49.73	52.49	54.42	56.08	57.83
Annotated		72.0	72.76	73.53	74.60	74.93	94.54

FRAGMENTED SEQUENCES SINCE VERSION 6



example values:

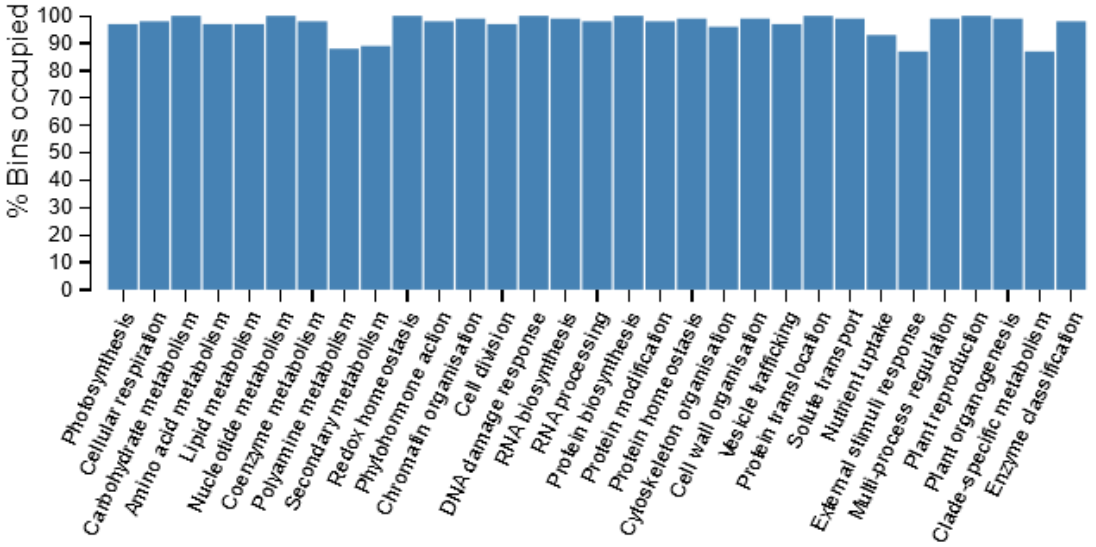
species	below IQR	within IQR	above IQR
Arabidopsis thaliana	0.150	0.728	0.122
Glycine max	0.138	0.707	0.155
Trifolium pratense	0.126	0.721	0.152
Solanum pennellii	0.146	0.719	0.135
Cuscuta campestris	0.202	0.594	0.204
Cuscuta pentagona (from transcriptome)	0.447	0.436	0.117
Peperomia fraseri (from transcriptome)	0.656	0.327	0.017

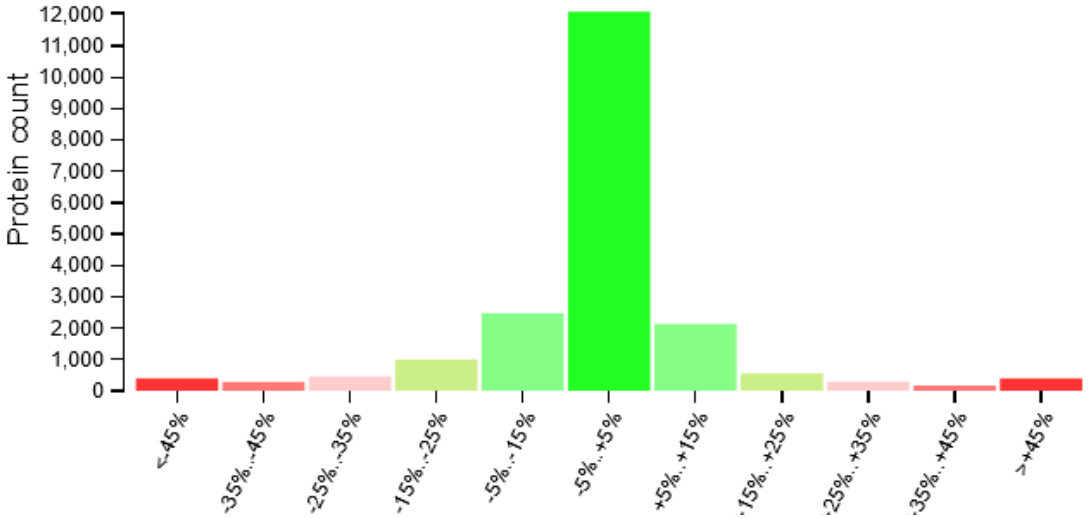
fragmented proteome

Job name:	pep6
Job ID:	GFA-76bd35bbcdb46999f2fc89f9b6b4b212
Number of sequences:	27416
Sequence type:	Protein
Prot-scriber	False
Swissprot	False
Submitted to cluster:	13.2.2024, 21:37:07
Job status:	FINISHED
Last update:	13.2.2024, 21:47:08

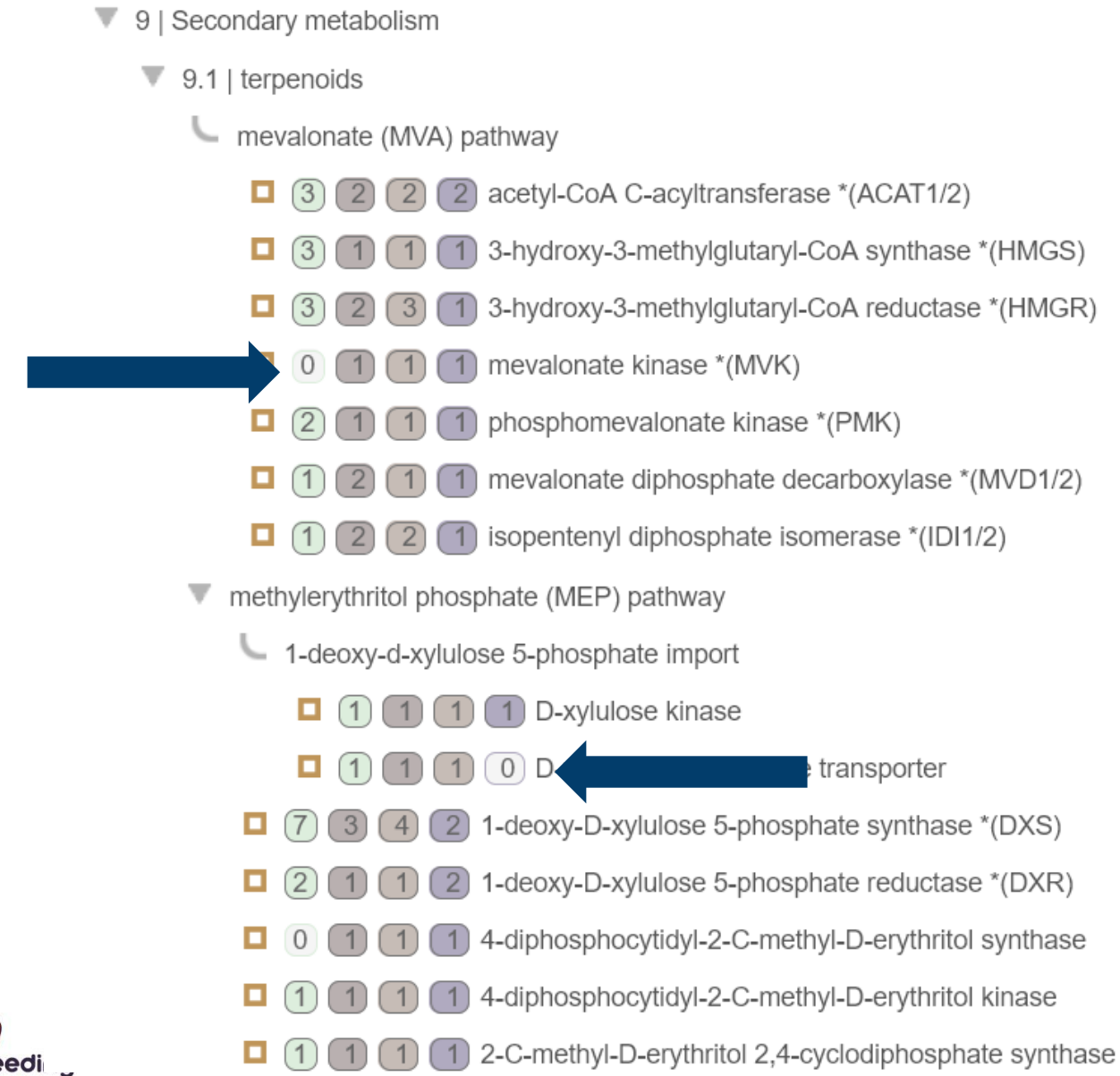
Submitted sequences (S):	27416
Annotated sequences (A):	15854
Classified sequences (C):	15854
Occupied bins (O):	6115
Bins available (B):	6223
Summary ℹ:	S:27416,A:57.83%,C:57.83%,O:98.26%,B: 6223

Mapping file for MapMan:	MapMan mapping file
Annotated FASTA file:	Mercator4 annotated FASTA file





PRESENCE ABSENCE AND COPY NUMBERS



Search by position on chromosome

chromosome (available position range)

Chr1 (1 .. 23207424)

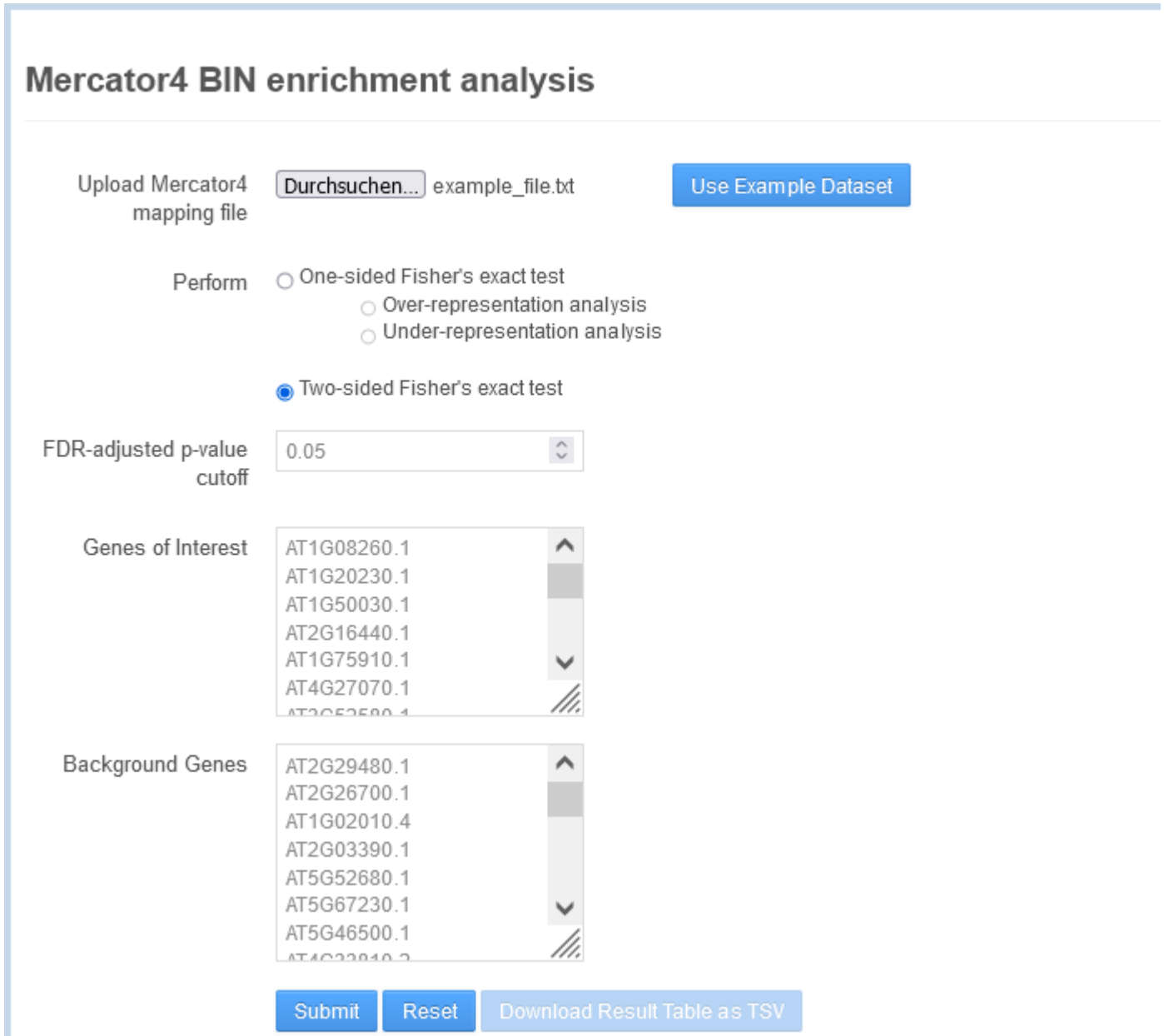
1

100000

search reset

19 protein result(s)

	Plant species Sequence details	Description Functional context (based on MAPMAN4)	Orthologous proteins
1	Camelina sativa Csa01g001010	putative component of lumen subcomplex of chloroplast NDH *Photosynthesis: photophosphorylation*	*protein family*
2	Camelina sativa Csa01g001020	(yeast MED11)-like component of Mediator transcriptional regulatory complex *RNA biosynthesis: RNA polymerase II-dependent transcription*	*protein family*
3	Camelina sativa Csa01g001030	alpha-dioxygenase *Lipid metabolism: lipid degradation*	*protein family*
4	Camelina sativa Csa01g001040	putative ribonuclease H1 involved in R-loop homeostasis *RNA biosynthesis: RNA polymerase II-dependent transcription*	*protein family*
5	Camelina sativa Csa01g001050	protein of unknown function	*protein family*
6	Camelina sativa Csa01g001060	G-type subunit of vacuolar H(+)-ATPase peripheral V1 subcomplex *Solute transport: primary active transport*	*protein family*
7	Camelina sativa Csa01g001070	putative PIG-N-type protein of GPI biosynthetic pathway *Protein modification: lipidation*	*protein family*
8	Camelina sativa Csa01g001080	putative plastidial CRS/CFM-type RNA intron splicing factor *RNA processing: organelle machinery*	*protein family*
9	Camelina sativa Csa01g001090	putative plastidial CRS/CFM-type RNA intron splicing factor	*protein family*


Mercator4 BIN enrichment analysis
Upload Mercator4 mapping file
Durchsuchen...
example_file.txt
Use Example Dataset
Perform
One-sided Fisher's exact test
Over-representation analysis
Under-representation analysis
Two-sided Fisher's exact test
FDR-adjusted p-value cutoff
0.05
Genes of Interest
AT1G08260.1
AT1G20230.1
AT1G50030.1
AT2G16440.1
AT1G75910.1
AT4G27070.1
Background Genes
AT2G29480.1
AT2G26700.1
AT1G02010.4
AT2G03390.1
AT5G52680.1
AT5G67230.1
AT5G46500.1
Submit
Reset
Download Result Table as TSV

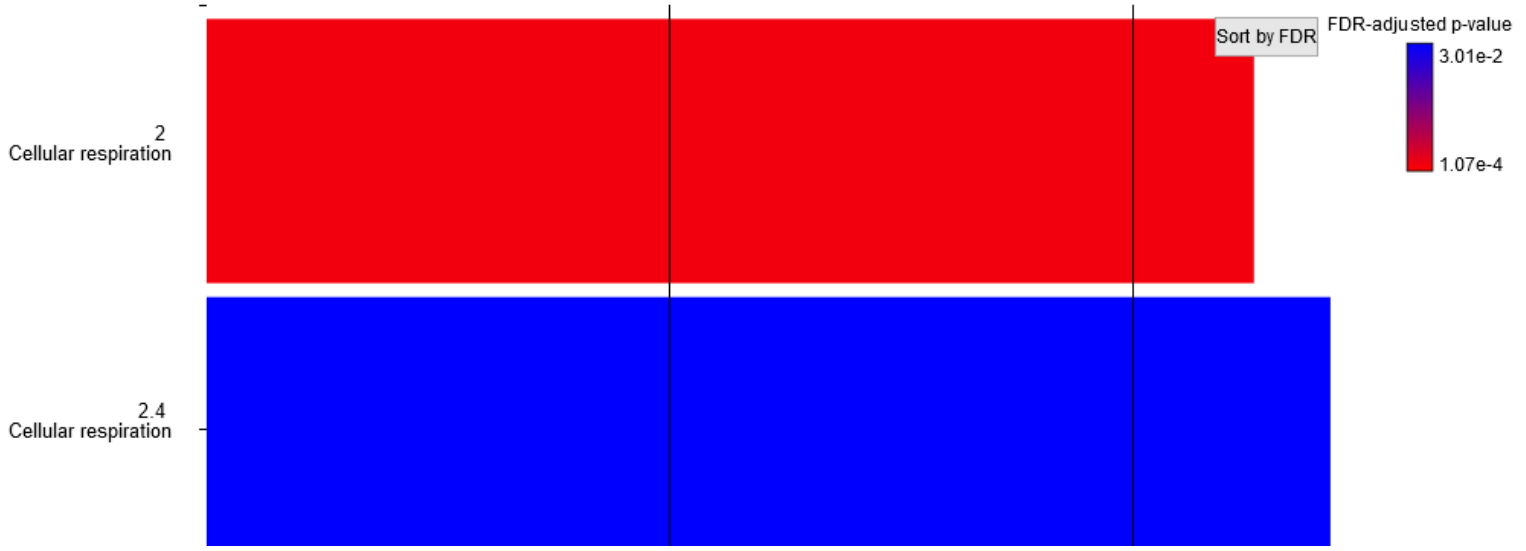
Result Table | MapMan BIN Chart | Enrichment Visualization

Toggle column: Genes of Interest (List)

Search:

MapMan4 category number	Context of Protein Function	#Genes of Interest IN MapMan4 category	#Genes of Interest NOT IN MapMan4 category	#Background Genes IN MapMan4 category	#Background Genes NOT IN MapMan4 category	Enrichment Factor	p-value	FDR-adjusted p-value
2	Cellular respiration	18	184	29	1441	4.52	0.00000270400822086635 97	0.001616996916078083
2.4	Cellular respiration.oxidative phosphorylation	12	190	18	1452	4.85	0.00008388516891239334	0.03009799860576673
35	No Mercator4 annotation	40	162	569	901	0.51	5.988643766125384e-8	0.00010743626916428939
35.1	No Mercator4 annotation.other annotation available	40	162	505	965	0.58	0.000020617733553228634	0.009247053498623042
50	Enzyme classification	62	140	219	1251	2.06	1.8643729410822603e-7	0.00016723425281507876

Showing 1 to 5 of 5 entries



MAPMAN VISUALISATIONS

Protein function annotation

Mapman4 category (BIN)		protein name	protein description
BINCODE	NAME	IDENTIFIER	DESCRIPTION
"1.3.1"	"Photosynthesis. photorespiration. phosphoglycolate phosphatase"	"ab0g12345"	"phosphoglycolate phosphatase"
"2.3.1"	"Cellular respiration. tricarboxylic acid cycle. citrate synthase"	"ab0g23456"	"citrate synthase"
"18.9.1"	"Protein modification. hydroxylation. prolyl hydroxylase"	"ab0g34567"	"prolyl hydroxylase"

example

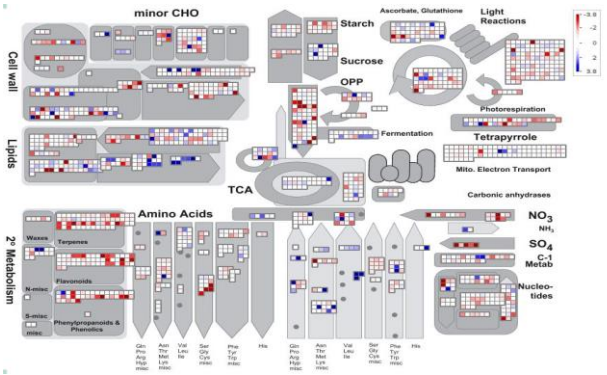
Run Mercator4 on genome assembly

Gene expression data

	condition 1		condition 2	
	expression change log2 fold	significance	expression change log2 fold	significance
example	"cond1 vs. normal"	"cond1 vs. normal"	"cond2 vs. normal"	"cond2 vs. normal"
"ab0g12345"	-0.25	0	0.85	1
"ab0g23456"	0.63	1	1.25	0
"ab0g34567"	0.46	0	1.62	1

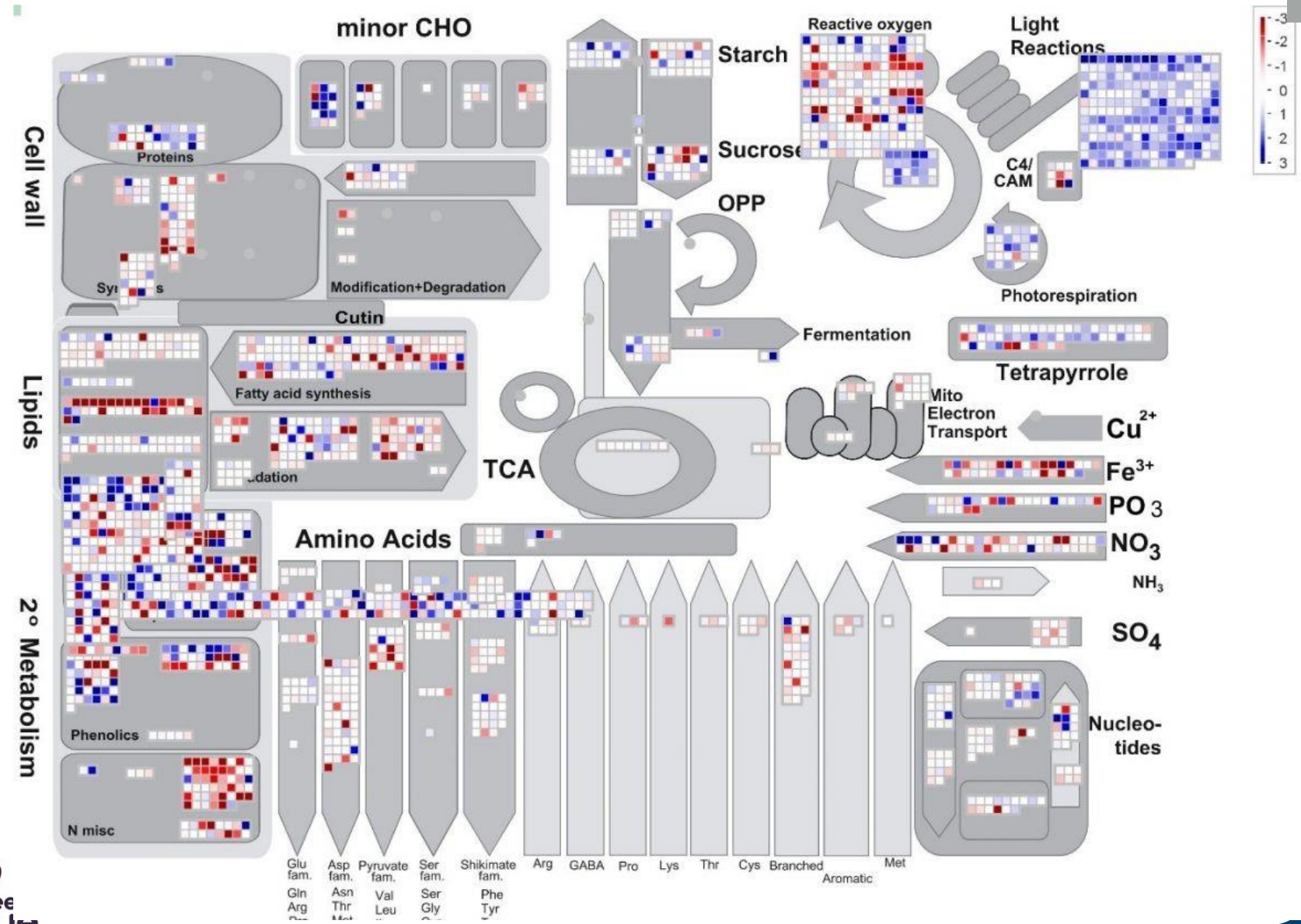
Differential gene expression (snRNA, RNA-Seq, Microarray data), Metabolites, proteins

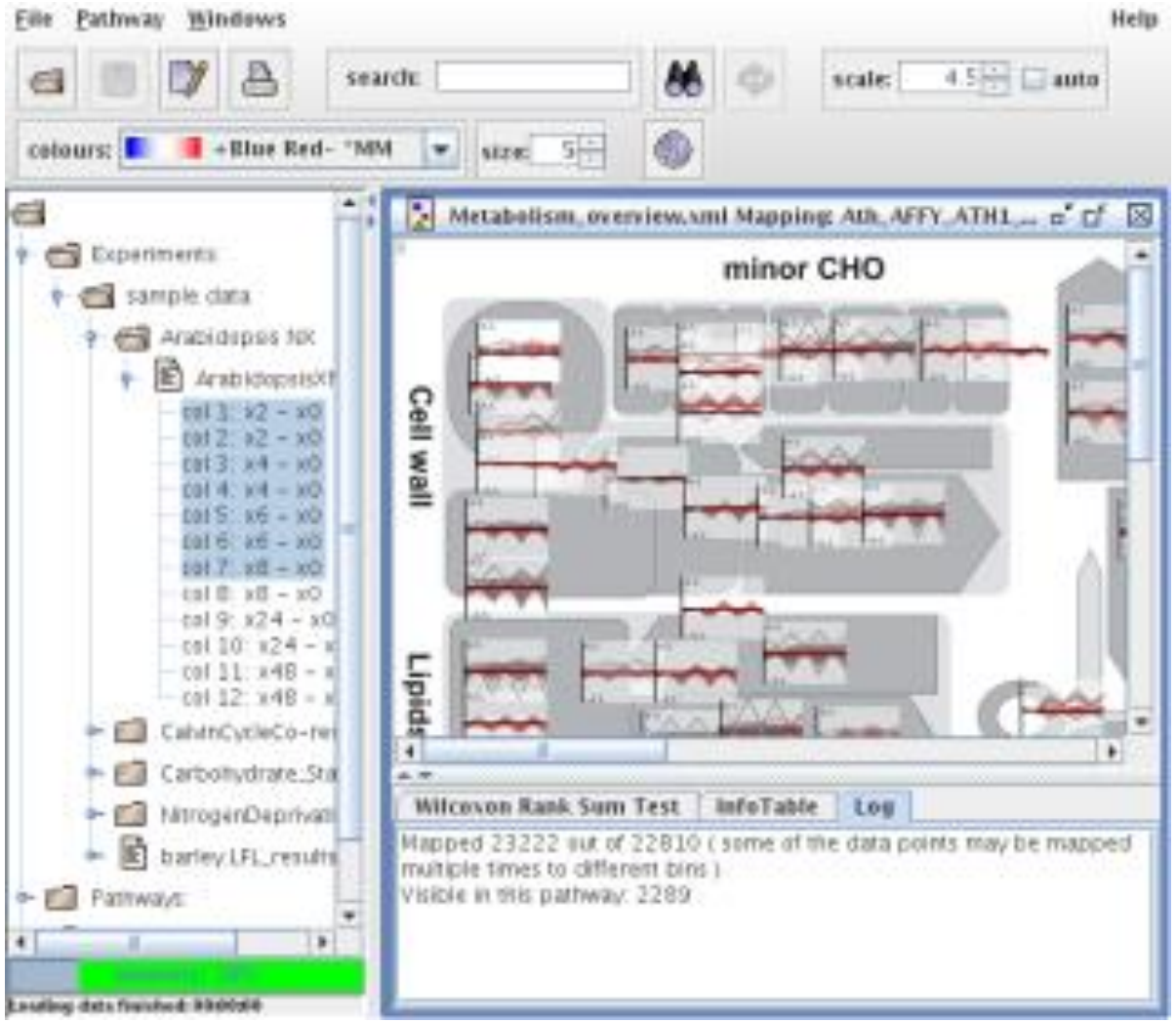
Pathway diagram



Pathway you are interested in, available on the plabipd website

Important Note: Identifiers must be identical


minor CHO
Starch
Sucrose
Reactive oxygen
Light Reactions
C4/ CAM
OPP
Photorespiration
Fermentation
Tetrapyrrole
Mito Electron Transport
Cu2+
Fe3+
PO 3
NO3
NH3
SO4
Nucleo-tides
Cell wall
Proteins
Modification+Degradation
Cutin
Lipids
Fatty acid synthesis
TCA
2° Metabolism
Phenolics
N misc
Amino Acids
Glu fam.
Gln Arg
Asp fam.
Asn Thr
Pyruvate fam.
Val Leu
Ser fam.
Ser Gly
Shikimate fam.
Phe Tyr
Arg
GABA
Pro
Lys
Thr
Cys
Branched
Aromatic
Met


File Pathway Windows Help
search:
scale: 4.5 auto
colours: +Blue Red- *MM
size: 5
Experiments
sample data
Arabidopsis tax
Pathways
Metabolism_overview.xml Mapping Ath_AFFY_ATH1_...
minor CHO
Cell wall
Lipids
Wilcoxon Rank Sum Test
InfoTable
Log
Mapped 23222 out of 22810 (some of the data points may be mapped multiple times to different bins)
Visible in this pathway: 2289

- Dr. Alisandra Denton CEPLAS Düsseldorf (HELIXER)
- Dr. Rainer Scwacke, Sebastian Beier, Marie Bolger