EUROPEAN SCIENCE FOUNDATION

SETTING SCIENCE AGENDAS FOR EUROPE

Standing Committee for Life, Earth and Environmental Sciences (LESC)

European Networking Summer School (ENSS)
Plant Genomics & Bioinformatics

28.7.2009

Supported by:

Austria Fonds zur Förderung der wissenschaftlichen Forschung (FWF)

Belgium: Fonds voor Wetenschappelijk Onderzoek (FWO)

Finland: Academy of Finland - Research Council fo Biosciences and Environment

Ireland: Irish Research Council for Science Engineering and Technology (IRCSET)

Italy: Consiglio Nazionale delle Ricerche (CNR) - Dipartimento Agroalimentare

Netherlands: Nederlandse Wetenschappelijk voon Onderzoek (NWO)

Norway: The Research Council of Norway

Poland: The Polish Academy of Science

Romania: Ministry of Education and Research

United Kingdom: Biotechnology and Biological Sciences Research Council (BBSRC)

EUROPEAN
SCIENCE
FOUNDATION
SETTING SCIENCE AGENDAS FOR EUROPE

AIMS

- Support plant genome research networks based by training of young investigators

- Summer courses with theoretic and practical training

- Access to technologies, resources, skills and know-how

ENSS 2009

Plant Bioinformatics, Systems and Synthetic Biology

27-31 July 2009
University of Nottingham, UK
Natalio Krasnogor, Jaume Bacardit, Malcolm Bennett

ENSS 2010

Plant Epigenetics

September 2010
Leibniz Institute of Plant Genetics and Crop
Plant Research (IPK) in Gatersleben, Germany
Michael Florian Mette

Comparative and Functional Genomics

Comparative genomics involves the use of computer programs to line up multiple genomes/genes for the identification of similarities

Functional genomics is the understanding of the function of genes and other parts of the genome

What is needed to do comparative and functional genomics?

model organism

Why are model organisms important?

Criteria for a good model organism?

Relationship of the model to important crop plants?

How many genes are the same?

Why using knock out/down mutants?

How will they help us determine gene function?

What will you hear?

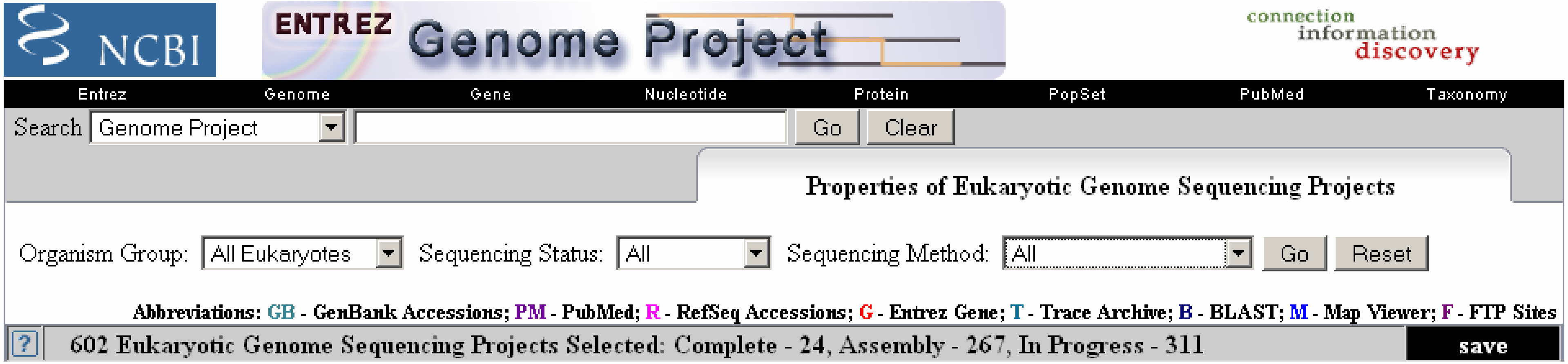
Background on annotating gene function using comparative genomic tools

Example to show how these tools can be employed to get a glimps on the function of a yet unknown gene

Example for the use comparing genomes/genes from individuals between populations to determine their function

All Starts With Genome Sequencing Projects

How many plant genomes have been sequenced?


NCBI
ENTREZ Genome Project
connection information discovery
Entrez Genome Gene Nucleotide Protein PopSet PubMed Taxonomy
Go Clear
Go Reset
save
?

Search: Genome Project

Properties of Eukaryotic Genome Sequencing Projects

Organism Group: All Eukaryotes Sequencing Status: All Sequencing Method: All

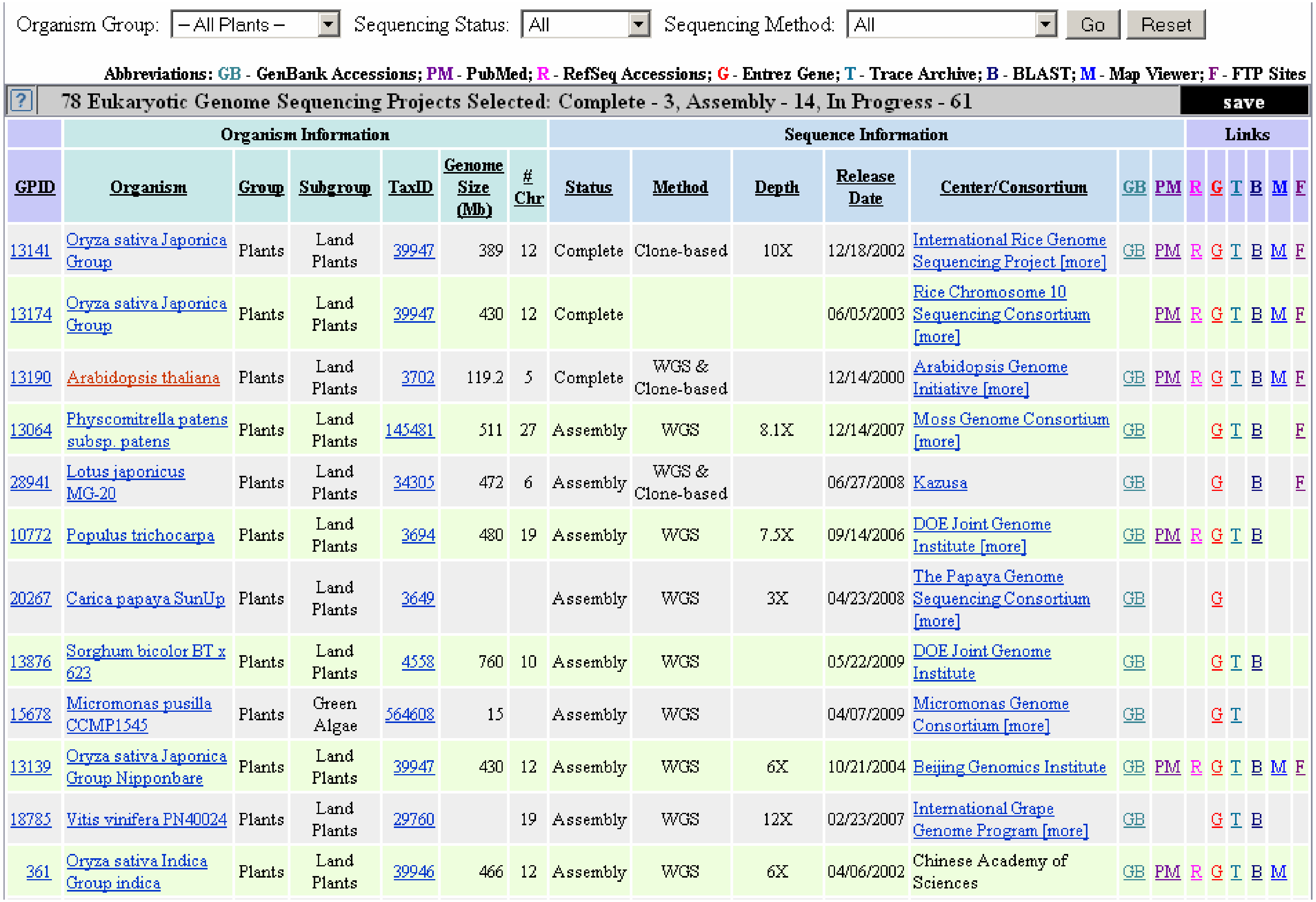
Abbreviations: GB - GenBank Accessions; PM - PubMed; R - RefSeq Accessions; G - Entrez Gene; T - Trace Archive; B - BLAST; M - Map Viewer; F - FTP Sites

602 Eukaryotic Genome Sequencing Projects Selected: Complete - 24, Assembly - 267, In Progress - 311

http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi

http://www.ensembl.org/info/about/species.html

Plant Genome Sequencing Projects



Organism Group: – All Plants – ▾ Sequencing Status: All ▾ Sequencing Method: All ▾ Go Reset

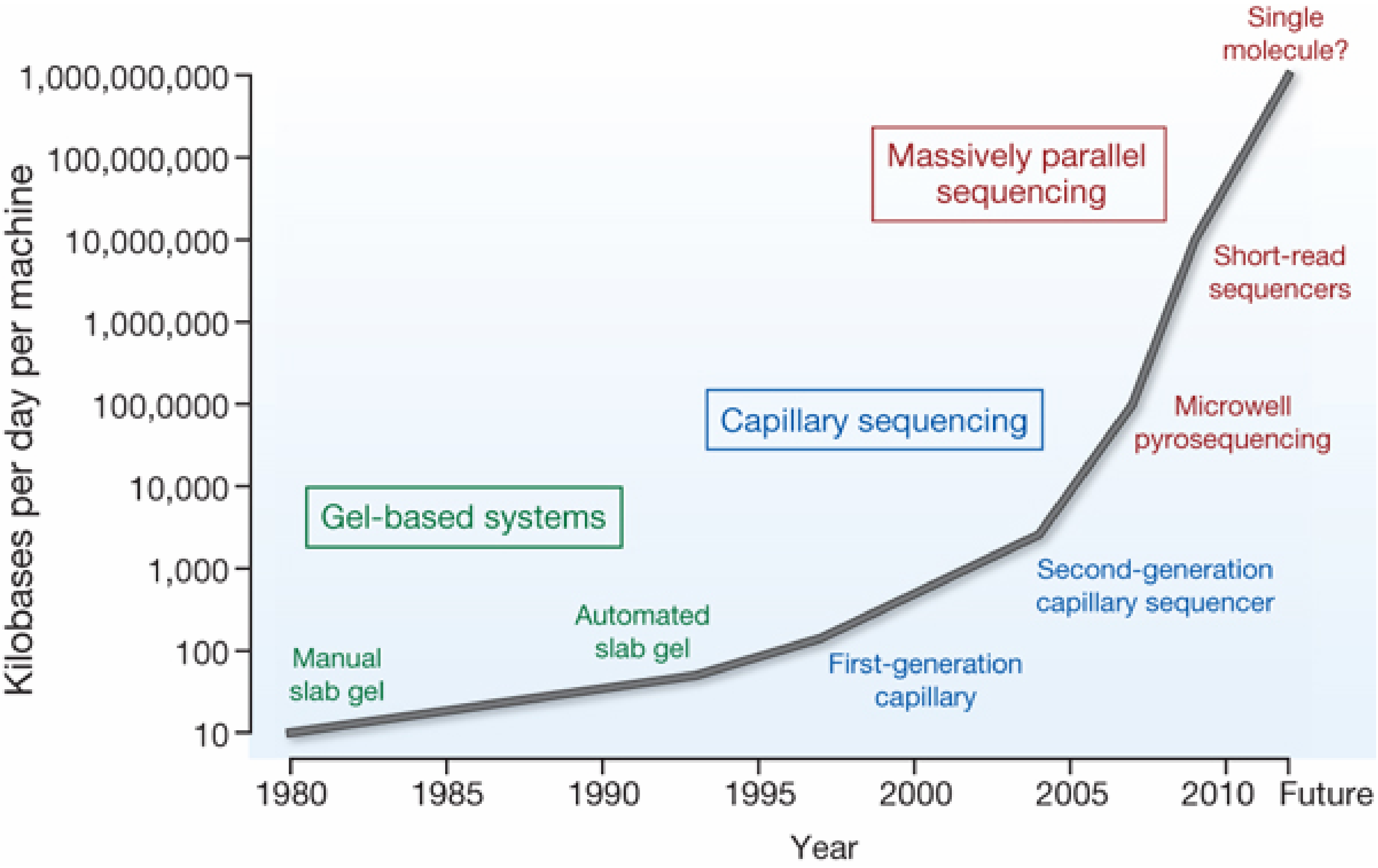
Abbreviations: GB - GenBank Accessions; PM - PubMed; R - RefSeq Accessions; G - Entrez Gene; T - Trace Archive; B - BLAST; M - Map Viewer; F - FTP Sites

? 78 Eukaryotic Genome Sequencing Projects Selected: Complete - 3, Assembly - 14, In Progress - 61 save

	Organism Information						Sequence Information					Links							
GPID	Organism	Group	Subgroup	TaxID	Genome Size (Mb)	# Chr	Status	Method	Depth	Release Date	Center/Consortium	GB	PM	R	G	T	B	M	F
13141	Oryza sativa Japonica Group	Plants	Land Plants	39947	389	12	Complete	Clone-based	10X	12/18/2002	International Rice Genome Sequencing Project [more]	GB	PM	R	G	T	B	M	F
13174	Oryza sativa Japonica Group	Plants	Land Plants	39947	430	12	Complete			06/05/2003	Rice Chromosome 10 Sequencing Consortium [more]		PM	R	G	T	B	M	F
13190	Arabidopsis thaliana	Plants	Land Plants	3702	119.2	5	Complete	WGS & Clone-based		12/14/2000	Arabidopsis Genome Initiative [more]	GB	PM	R	G	T	B	M	F
13064	Physcomitrella patens subsp. patens	Plants	Land Plants	145481	511	27	Assembly	WGS	8.1X	12/14/2007	Moss Genome Consortium [more]	GB			G	T	B		F
28941	Lotus japonicus MG-20	Plants	Land Plants	34305	472	6	Assembly	WGS & Clone-based		06/27/2008	Kazusa	GB			G		B		F
10772	Populus trichocarpa	Plants	Land Plants	3694	480	19	Assembly	WGS	7.5X	09/14/2006	DOE Joint Genome Institute [more]	GB	PM	R	G	T	B		
20267	Carica papaya SunUp	Plants	Land Plants	3649			Assembly	WGS	3X	04/23/2008	The Papaya Genome Sequencing Consortium [more]	GB			G				
13876	Sorghum bicolor BTx 623	Plants	Land Plants	4558	760	10	Assembly	WGS		05/22/2009	DOE Joint Genome Institute	GB			G	T	B		
15678	Micromonas pusilla CCMP1545	Plants	Green Algae	564608	15		Assembly	WGS		04/07/2009	Micromonas Genome Consortium [more]	GB			G	T			
13139	Oryza sativa Japonica Group Nipponbare	Plants	Land Plants	39947	430	12	Assembly	WGS	6X	10/21/2004	Beijing Genomics Institute	GB	PM	R	G	T	B	M	F
18785	Vitis vinifera PN40024	Plants	Land Plants	29760		19	Assembly	WGS	12X	02/23/2007	International Grape Genome Program [more]	GB			G	T	B		
361	Oryza sativa Indica Group indica	Plants	Land Plants	39946	466	12	Assembly	WGS	6X	04/06/2002	Chinese Academy of Sciences	GB	PM	R	G	T	B	M	

28.7.2009

Improvements in the rate of DNA sequencing over the past 30 years and into the future



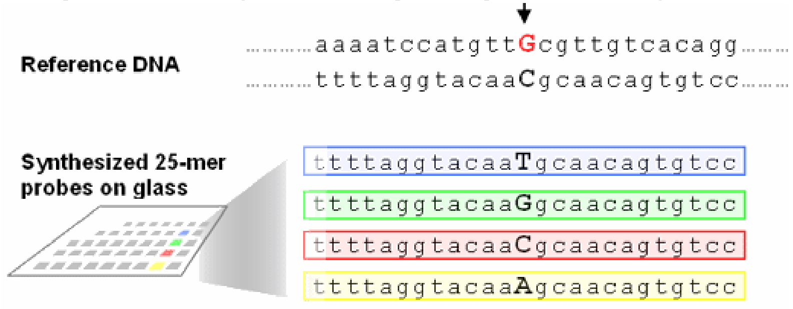
Stratton et al., 2009

28.7.2009

Quote from Joe Ecker IARC2009

Capillary sequencing – 500 people – 7 years – 70.000.000 $

Perlegen sequencing – 50 people – 1 year – 70.000 $

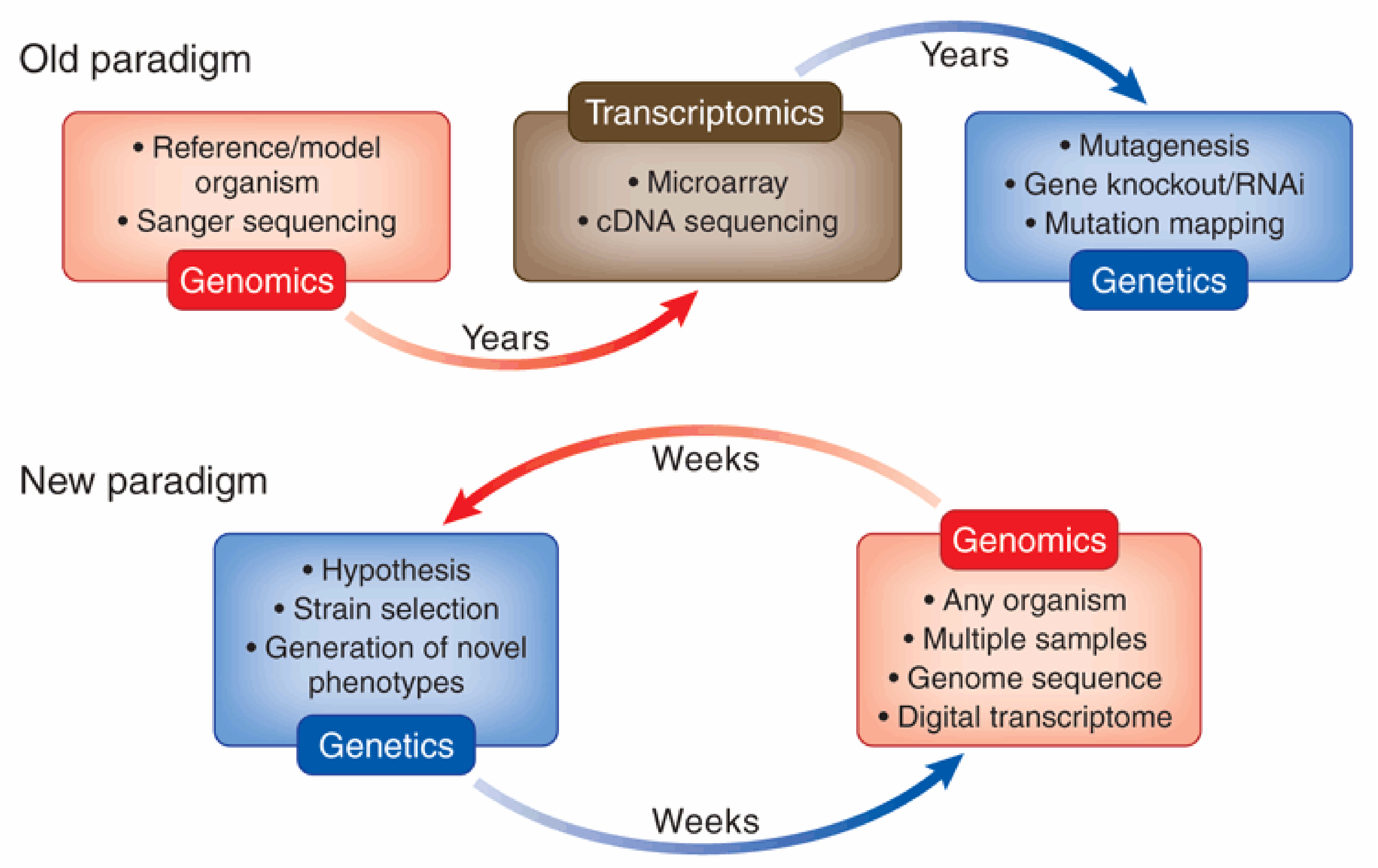

Reference DNA
……….aaaatccatgttGcgttgtcacagg……..
……….ttttaggtacaaCgcaacagtgtcc……..
Synthesized 25-mer probes on glass
ttttaggtacaaTgcaacagtgtcc
ttttaggtacaaGgcaacagtgtcc
ttttaggtacaaCgcaacagtgtcc
ttttaggtacaaAgcaacagtgtcc

	454 GS FLX*	AB SOLiD	Illumina GAII
Chemistry	Pyrosequencing	Ligation based	Reversible terminators
	Standard	Fragment	Fragment
Run Time	7 hours	3-6.5 days	3 days

Next generation sequencing – 2 people – 7 days – 7.000 $ - 50 x coverage

28.7.2009

Paradigm Change



Old paradigm

- Reference/model organism
- Sanger sequencing

Genomics

Years

Transcriptomics

- Microarray
- cDNA sequencing

Years

- Mutagenesis
- Gene knockout/RNAi
- Mutation mapping

Genetics

New paradigm

Weeks

- Hypothesis
- Strain selection
- Generation of novel phenotypes

Genetics

Genomics

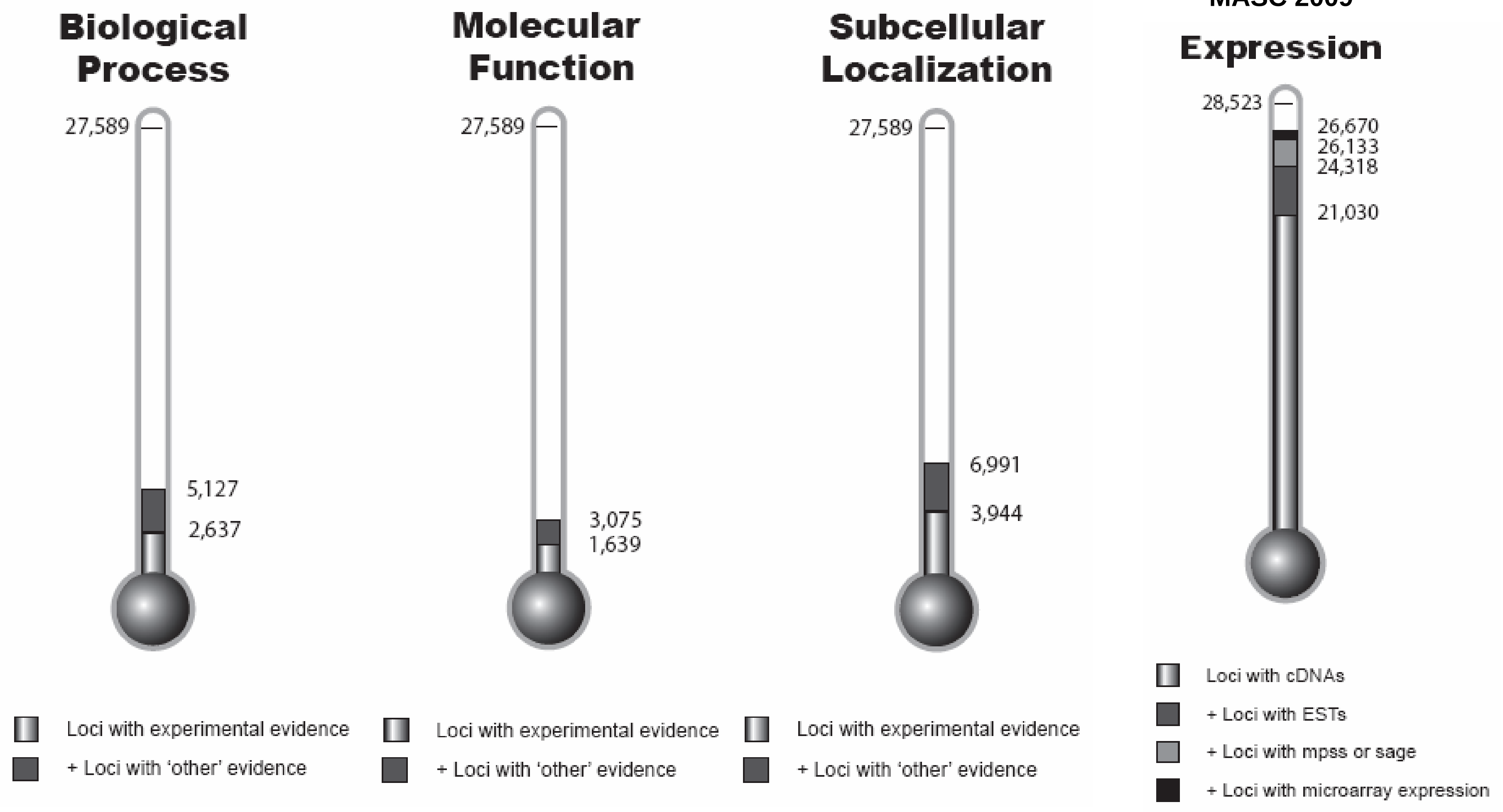
- Any organism
- Multiple samples
- Genome sequence
- Digital transcriptome

Weeks

28.7.2009

After genome sequencing still many questions remain – example Arabidopsis

MASC 2009



MASC 2007

28.7.2009

Sequenced genomes – the basis to address questions on

Function of all genes

Functional redundancies/diversification of gene families

Role of single nucleotide polymorphisms (SNPs, natural variation)

Role of alternative splicing variants

Role of noncoding regions and repeats in the genome

When? – Regulation (transcriptional, post-transcriptional, post-translational,..)

Where? - Localization (organs, tissues, cellular, sub-cellular)

Interacting partners - Networks

Biological role(s)

Functional Genomic Tools

Sequences genome, full-length cDNA clones

Gene knock-outs, knock-downs (T-DNA, transposon, amiRNA, tilling, gene targeting, collection of natural variants, ...)

Methods for studying functions of nonprotein-coding sequences

Comprehensive analysis of gene expression (microarray, deep sequencing, cell sorting, laser dissection, reporter constructs, …)

Large-scale protein analyses (proteomics, protein arrays, large scale Y2H, interactomes-networks, 3D structures)

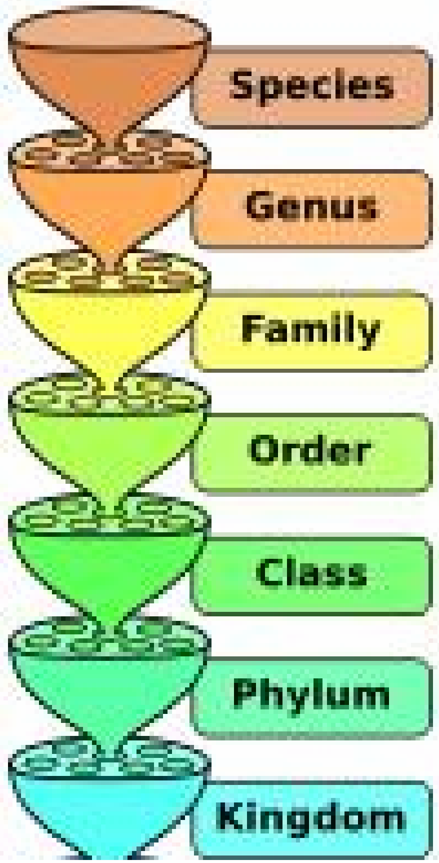
Metabolomics

-omics

Comparative genomics between species

comparison of genomes from different taxa

FUNGI
PLANTAE
ANIMALIA
PROTISTA
MONERA



28.7.2009

Comparative genomics within a species

comparison of genomes from different individuals between populations that might be differentially adapted to particular environments ….

What to compare?

on the structural level:

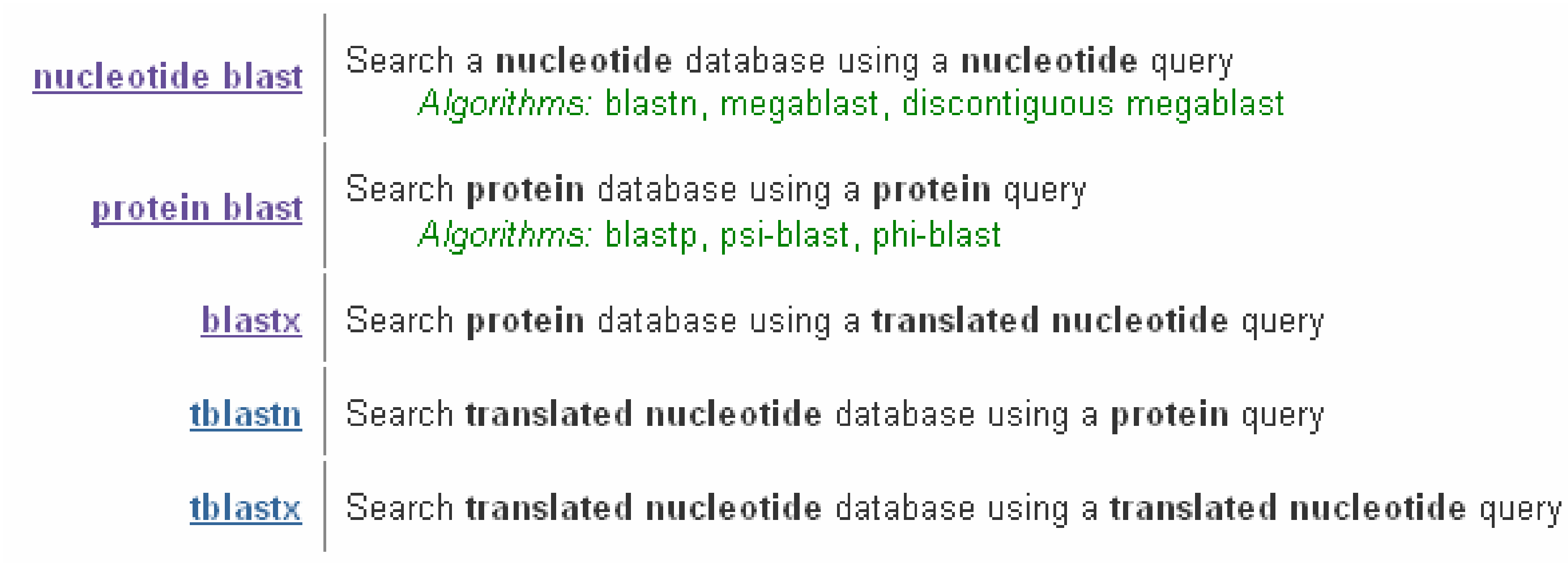
sequence similarity (nucleic acid, protein, domains)
gene location (synteny)
gene structure (length, number of exons)
amount of noncoding DNA
highly conserved regions (fundamental/essential genes)
highly/less polymorphic regions (indication of adaptation, selection)

on the functional level:

expression pattern
epigenetic regulation
post-transcriptional
translational regulation
subcellular localization
interactions
post-translational regulation/modification

How to compare?

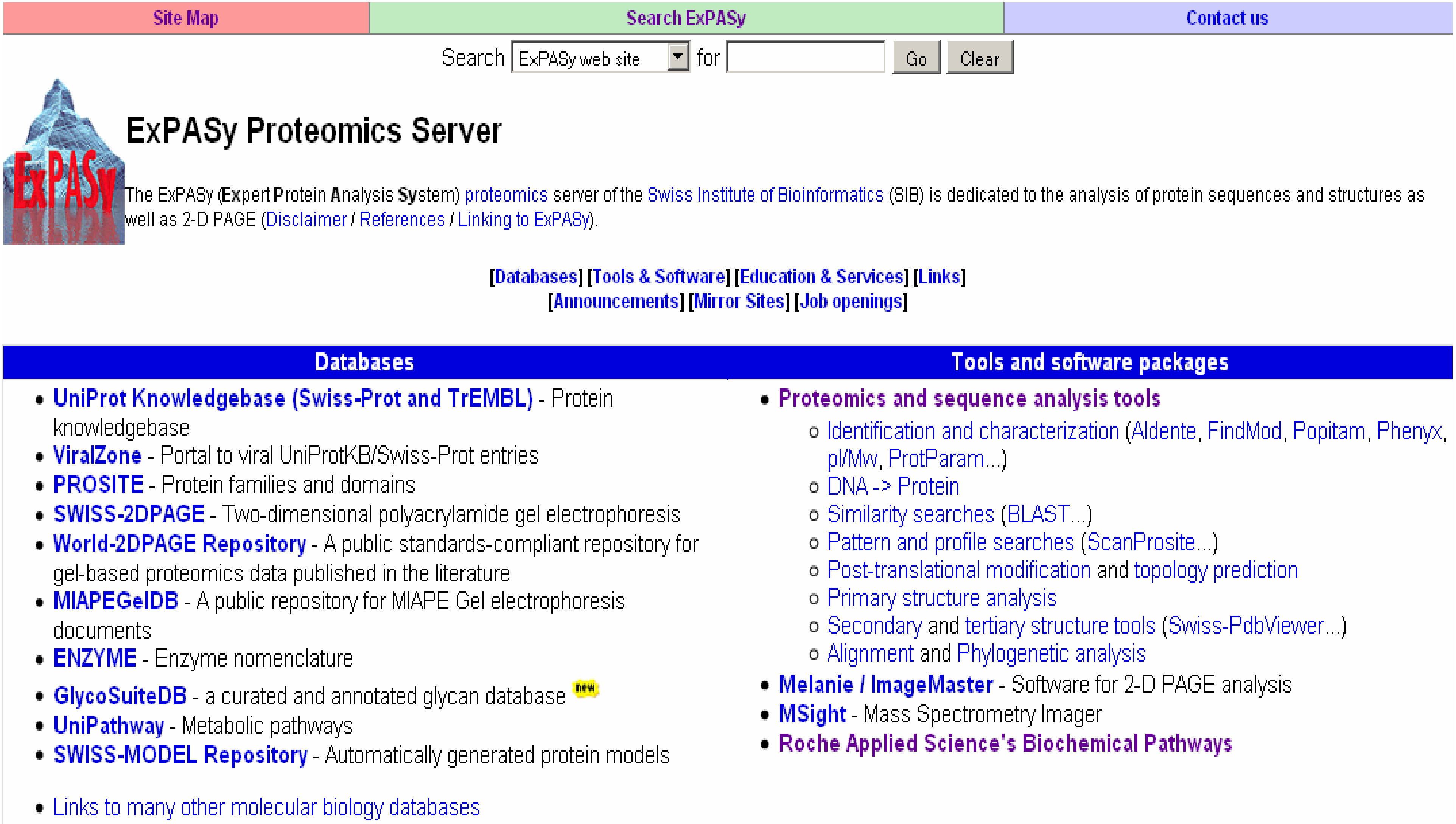
Search tools for homologies: BLAST
FASTA



nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

How to compare?

http://www.expasy.org/



ExPASy Proteomics Server

The ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE (Disclaimer / References / Linking to ExPASy).

[Databases] [Tools & Software] [Education & Services] [Links]
[Announcements] [Mirror Sites] [Job openings]

Databases	Tools and software packages

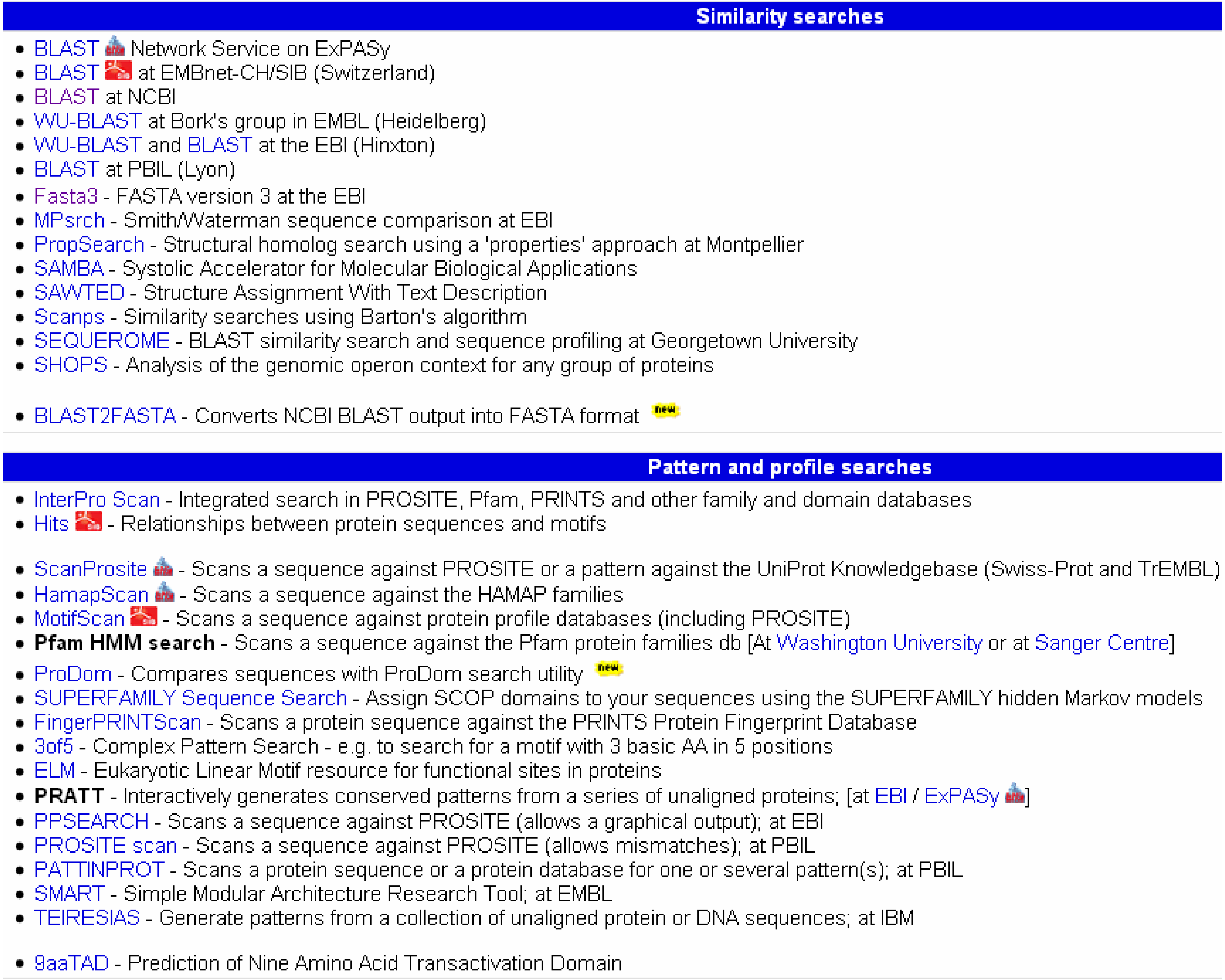
Databases

- UniProt Knowledgebase (Swiss-Prot and TrEMBL) - Protein knowledgebase
- ViralZone - Portal to viral UniProtKB/Swiss-Prot entries
- PROSITE - Protein families and domains
- SWISS-2DPAGE - Two-dimensional polyacrylamide gel electrophoresis
- World-2DPAGE Repository - A public standards-compliant repository for gel-based proteomics data published in the literature
- MIAPEGelDB - A public repository for MIAPE Gel electrophoresis documents
- ENZYME - Enzyme nomenclature
- GlycoSuiteDB - a curated and annotated glycan database new
- UniPathway - Metabolic pathways
- SWISS-MODEL Repository - Automatically generated protein models

- Links to many other molecular biology databases

Tools and software packages

- Proteomics and sequence analysis tools
  - Identification and characterization (Aldente, FindMod, Popitam, Phenyx, pI/Mw, ProtParam...)
  - DNA -> Protein
  - Similarity searches (BLAST...)
  - Pattern and profile searches (ScanProsite...)
  - Post-translational modification and topology prediction
  - Primary structure analysis
  - Secondary and tertiary structure tools (Swiss-PdbViewer...)
  - Alignment and Phylogenetic analysis
- Melanie / ImageMaster - Software for 2-D PAGE analysis
- MSight - Mass Spectrometry Imager
- Roche Applied Science's Biochemical Pathways



Similarity searches

- BLAST Network Service on ExPASy
- BLAST at EMBnet-CH/SIB (Switzerland)
- BLAST at NCBI
- WU-BLAST at Bork's group in EMBL (Heidelberg)
- WU-BLAST and BLAST at the EBI (Hinxton)
- BLAST at PBIL (Lyon)
- Fasta3 - FASTA version 3 at the EBI
- MPsrch - Smith/Waterman sequence comparison at EBI
- PropSearch - Structural homolog search using a 'properties' approach at Montpellier
- SAMBA - Systolic Accelerator for Molecular Biological Applications
- SAWTED - Structure Assignment With Text Description
- Scanps - Similarity searches using Barton's algorithm
- SEQUEROME - BLAST similarity search and sequence profiling at Georgetown University
- SHOPS - Analysis of the genomic operon context for any group of proteins

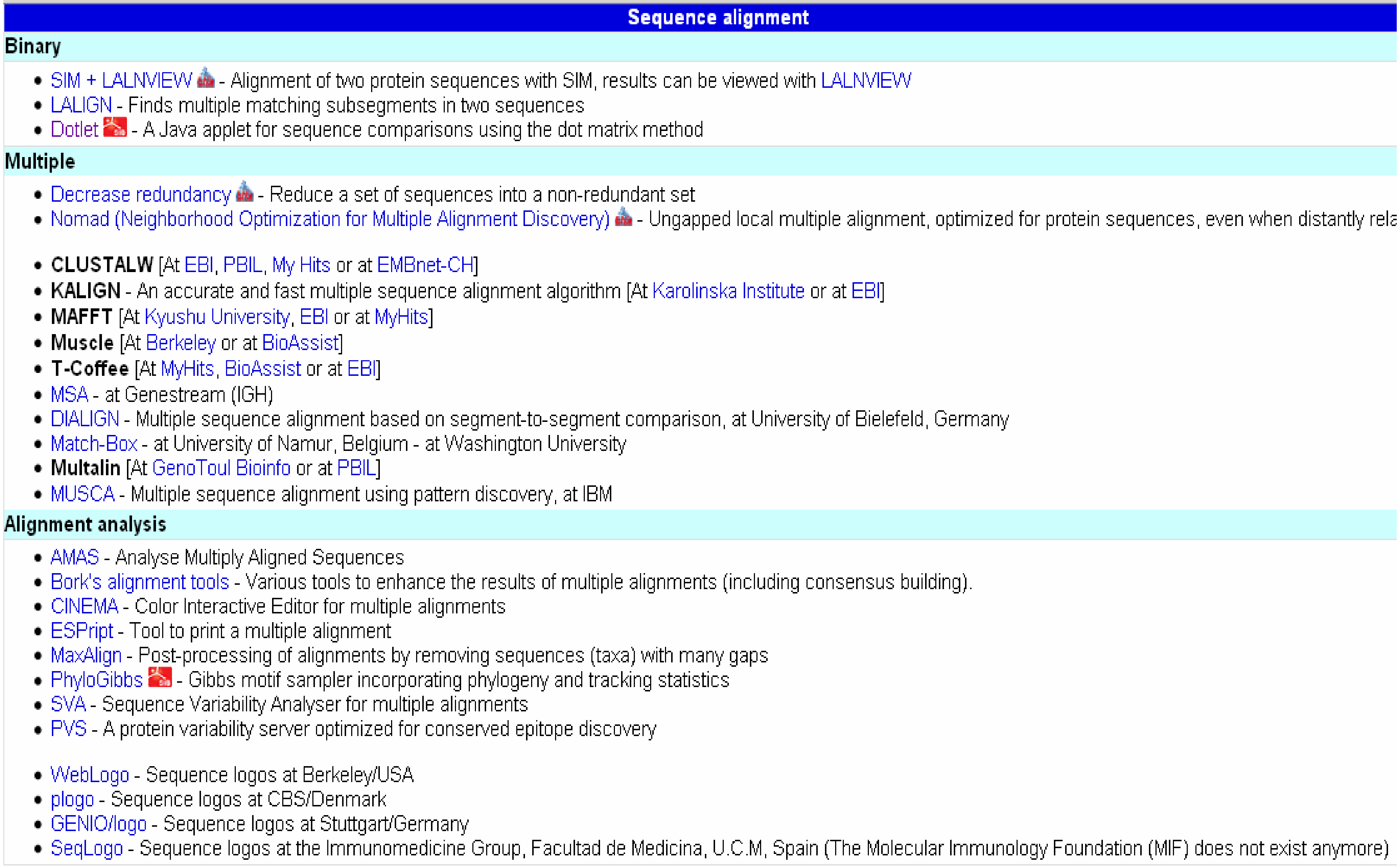
- BLAST2FASTA - Converts NCBI BLAST output into FASTA format new

Pattern and profile searches

- InterPro Scan - Integrated search in PROSITE, Pfam, PRINTS and other family and domain databases
- Hits - Relationships between protein sequences and motifs

- ScanProsite - Scans a sequence against PROSITE or a pattern against the UniProt Knowledgebase (Swiss-Prot and TrEMBL)
- HamapScan - Scans a sequence against the HAMAP families
- MotifScan - Scans a sequence against protein profile databases (including PROSITE)
- **Pfam HMM search** - Scans a sequence against the Pfam protein families db [At Washington University or at Sanger Centre]
- ProDom - Compares sequences with ProDom search utility new
- SUPERFAMILY Sequence Search - Assign SCOP domains to your sequences using the SUPERFAMILY hidden Markov models
- FingerPRINTScan - Scans a protein sequence against the PRINTS Protein Fingerprint Database
- 3of5 - Complex Pattern Search - e.g. to search for a motif with 3 basic AA in 5 positions
- ELM - Eukaryotic Linear Motif resource for functional sites in proteins
- **PRATT** - Interactively generates conserved patterns from a series of unaligned proteins; [at EBI / ExPASy]
- PPSEARCH - Scans a sequence against PROSITE (allows a graphical output); at EBI
- PROSITE scan - Scans a sequence against PROSITE (allows mismatches); at PBIL
- PATTINPROT - Scans a protein sequence or a protein database for one or several pattern(s); at PBIL
- SMART - Simple Modular Architecture Research Tool; at EMBL
- TEIRESIAS - Generate patterns from a collection of unaligned protein or DNA sequences; at IBM

- 9aaTAD - Prediction of Nine Amino Acid Transactivation Domain



Sequence alignment

Binary

- SIM + LALNVIEW - Alignment of two protein sequences with SIM, results can be viewed with LALNVIEW
- LALIGN - Finds multiple matching subsegments in two sequences
- Dotlet - A Java applet for sequence comparisons using the dot matrix method

Multiple

- Decrease redundancy - Reduce a set of sequences into a non-redundant set
- Nomad (Neighborhood Optimization for Multiple Alignment Discovery) - Ungapped local multiple alignment, optimized for protein sequences, even when distantly rela

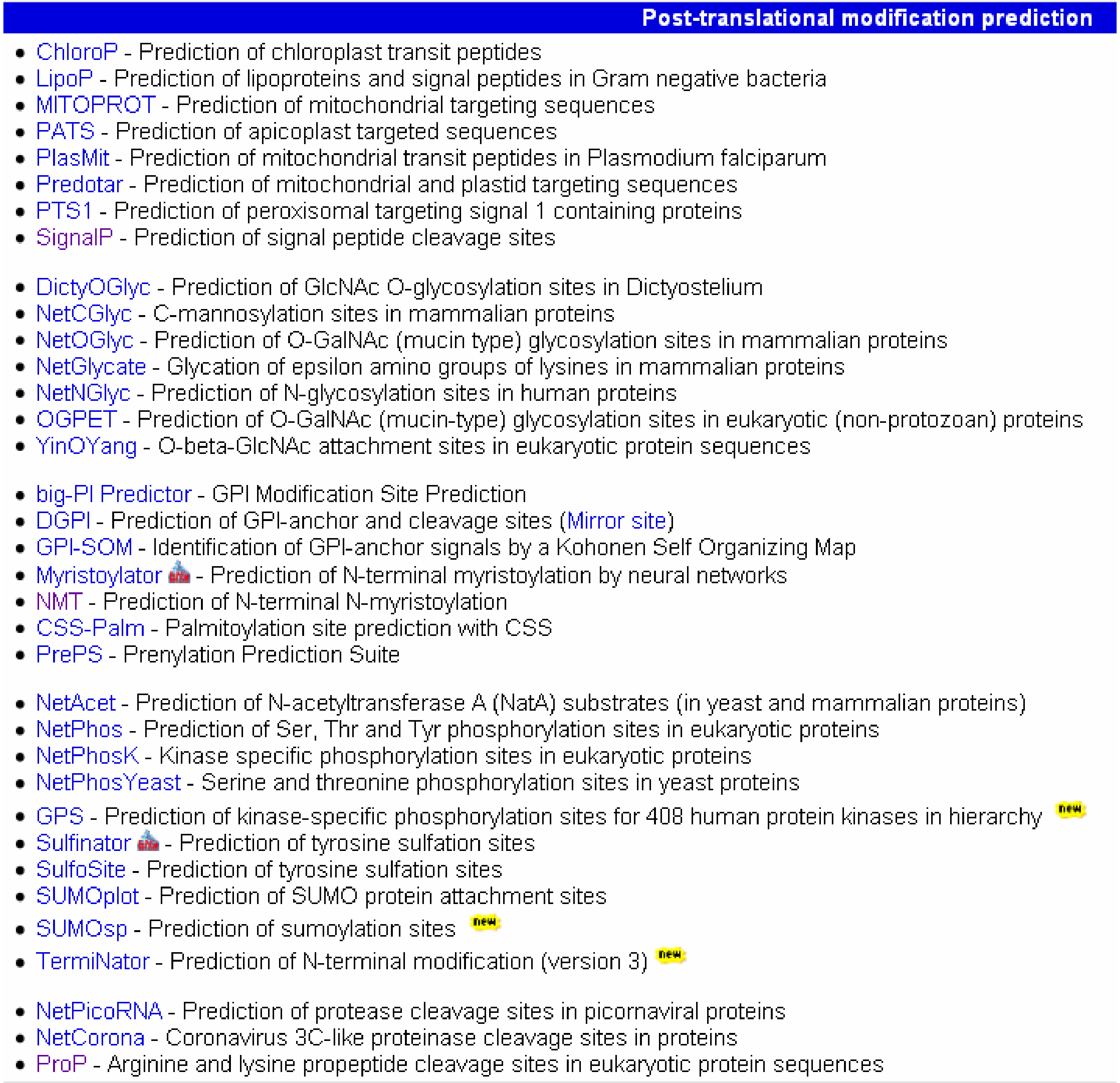
- **CLUSTALW** [At EBI, PBIL, My Hits or at EMBnet-CH]
- **KALIGN** - An accurate and fast multiple sequence alignment algorithm [At Karolinska Institute or at EBI]
- **MAFFT** [At Kyushu University, EBI or at MyHits]
- **Muscle** [At Berkeley or at BioAssist]
- **T-Coffee** [At MyHits, BioAssist or at EBI]
- MSA - at Genestream (IGH)
- DIALIGN - Multiple sequence alignment based on segment-to-segment comparison, at University of Bielefeld, Germany
- Match-Box - at University of Namur, Belgium - at Washington University
- **Multalin** [At GenoToul Bioinfo or at PBIL]
- MUSCA - Multiple sequence alignment using pattern discovery, at IBM

Alignment analysis

- AMAS - Analyse Multiply Aligned Sequences
- Bork's alignment tools - Various tools to enhance the results of multiple alignments (including consensus building).
- CINEMA - Color Interactive Editor for multiple alignments
- ESPript - Tool to print a multiple alignment
- MaxAlign - Post-processing of alignments by removing sequences (taxa) with many gaps
- PhyloGibbs - Gibbs motif sampler incorporating phylogeny and tracking statistics
- SVA - Sequence Variability Analyser for multiple alignments
- PVS - A protein variability server optimized for conserved epitope discovery

- WebLogo - Sequence logos at Berkeley/USA
- plogo - Sequence logos at CBS/Denmark
- GENIO/logo - Sequence logos at Stuttgart/Germany
- SeqLogo - Sequence logos at the Immunomedicine Group, Facultad de Medicina, U.C.M, Spain (The Molecular Immunology Foundation (MIF) does not exist anymore)

28.7.2009



Post-translational modification prediction

- ChloroP - Prediction of chloroplast transit peptides
- LipoP - Prediction of lipoproteins and signal peptides in Gram negative bacteria
- MITOPROT - Prediction of mitochondrial targeting sequences
- PATS - Prediction of apicoplast targeted sequences
- PlasMit - Prediction of mitochondrial transit peptides in Plasmodium falciparum
- Predotar - Prediction of mitochondrial and plastid targeting sequences
- PTS1 - Prediction of peroxisomal targeting signal 1 containing proteins
- SignalP - Prediction of signal peptide cleavage sites

- DictyOGlyc - Prediction of GlcNAc O-glycosylation sites in Dictyostelium
- NetCGlyc - C-mannosylation sites in mammalian proteins
- NetOGlyc - Prediction of O-GalNAc (mucin type) glycosylation sites in mammalian proteins
- NetGlycate - Glycation of epsilon amino groups of lysines in mammalian proteins
- NetNGlyc - Prediction of N-glycosylation sites in human proteins
- OGPET - Prediction of O-GalNAc (mucin-type) glycosylation sites in eukaryotic (non-protozoan) proteins
- YinOYang - O-beta-GlcNAc attachment sites in eukaryotic protein sequences

- big-PI Predictor - GPI Modification Site Prediction
- DGPI - Prediction of GPI-anchor and cleavage sites (Mirror site)
- GPI-SOM - Identification of GPI-anchor signals by a Kohonen Self Organizing Map
- Myristoylator - Prediction of N-terminal myristoylation by neural networks
- NMT - Prediction of N-terminal N-myristoylation
- CSS-Palm - Palmitoylation site prediction with CSS
- PrePS - Prenylation Prediction Suite

- NetAcet - Prediction of N-acetyltransferase A (NatA) substrates (in yeast and mammalian proteins)
- NetPhos - Prediction of Ser, Thr and Tyr phosphorylation sites in eukaryotic proteins
- NetPhosK - Kinase specific phosphorylation sites in eukaryotic proteins
- NetPhosYeast - Serine and threonine phosphorylation sites in yeast proteins
- GPS - Prediction of kinase-specific phosphorylation sites for 408 human protein kinases in hierarchy new
- Sulfinator - Prediction of tyrosine sulfation sites
- SulfoSite - Prediction of tyrosine sulfation sites
- SUMOplot - Prediction of SUMO protein attachment sites
- SUMOsp - Prediction of sumoylation sites new
- TermiNator - Prediction of N-terminal modification (version 3) new

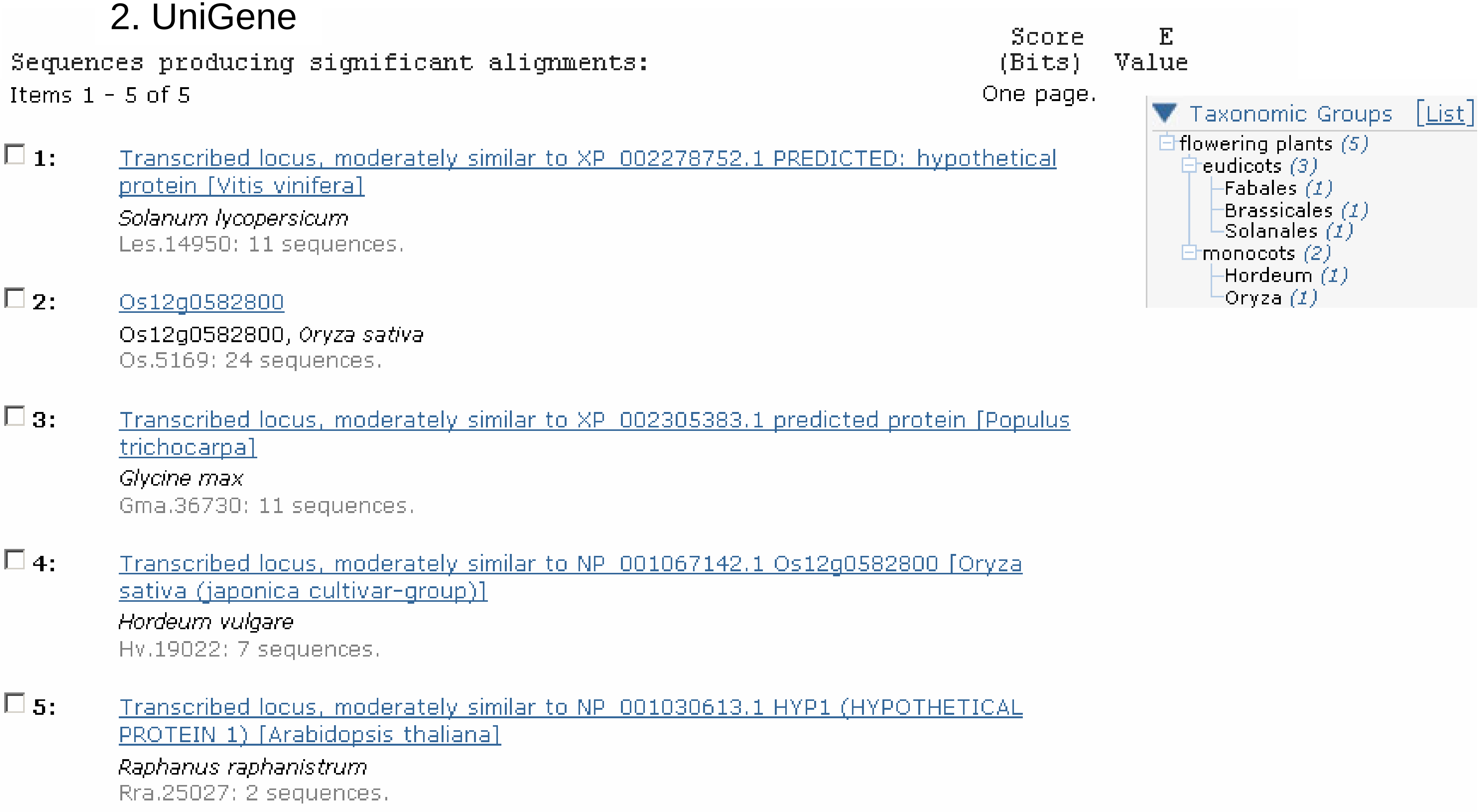
- NetPicoRNA - Prediction of protease cleavage sites in picornaviral proteins
- NetCorona - Coronavirus 3C-like proteinase cleavage sites in proteins
- ProP - Arginine and lysine propeptide cleavage sites in eukaryotic protein sequences

28.7.2009

Example

Proteome analysis in Poplar result a peptide of MILSALLTSVGINLGLC

2. UniGene



Sequences producing significant alignments: Score (Bits) E Value

Items 1 - 5 of 5 One page.

□ **1:** Transcribed locus, moderately similar to XP_002278752.1 PREDICTED: hypothetical protein [Vitis vinifera]
Solanum lycopersicum
Les.14950: 11 sequences.

□ **2:** Os12g0582800
Os12g0582800, *Oryza sativa*
Os.5169: 24 sequences.

□ **3:** Transcribed locus, moderately similar to XP_002305383.1 predicted protein [Populus trichocarpa]
Glycine max
Gma.36730: 11 sequences.

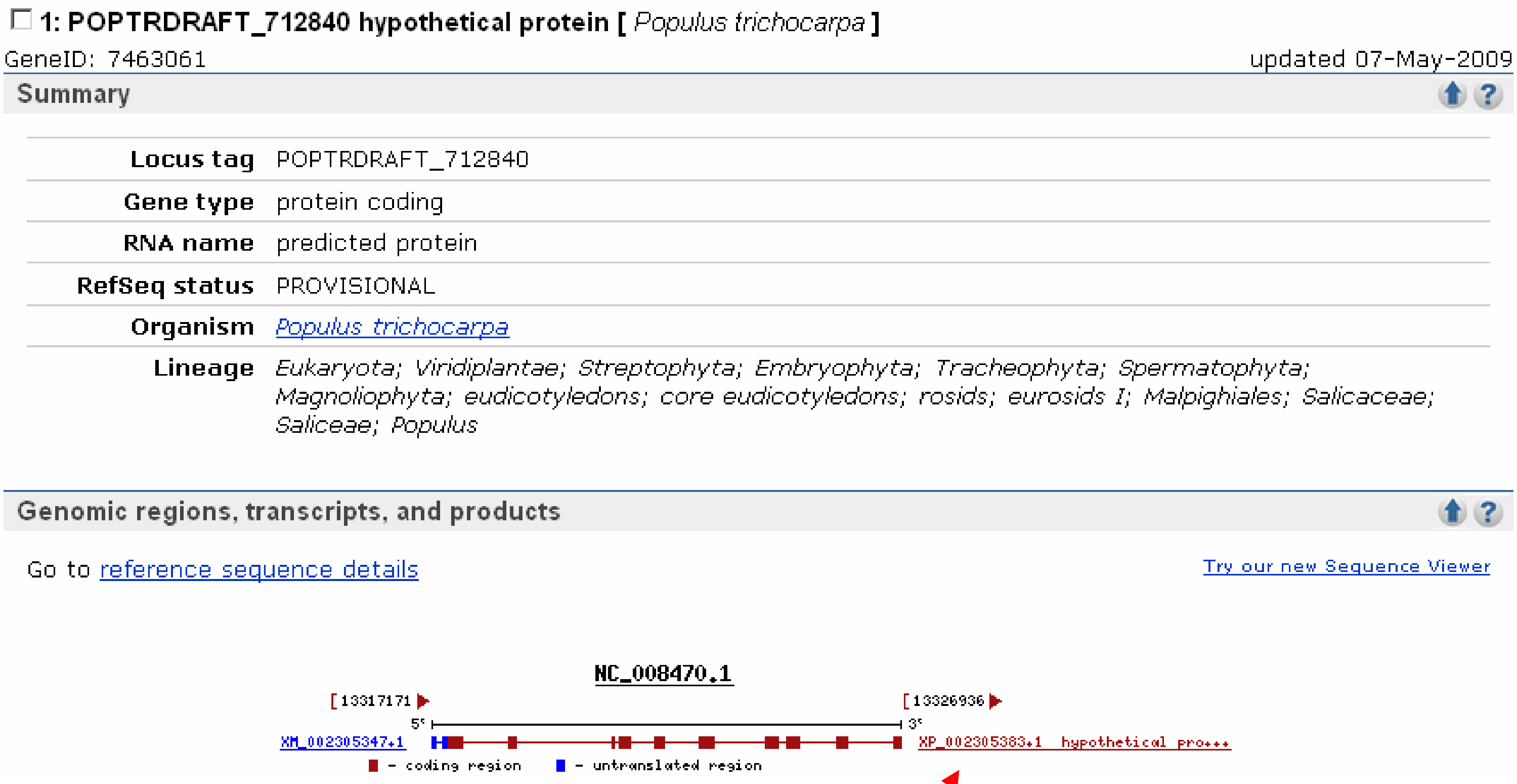
□ **4:** Transcribed locus, moderately similar to NP_001067142.1 Os12g0582800 [Oryza sativa (japonica cultivar-group)]
Hordeum vulgare
Hv.19022: 7 sequences.

□ **5:** Transcribed locus, moderately similar to NP_001030613.1 HYP1 (HYPOTHETICAL PROTEIN 1) [Arabidopsis thaliana]
Raphanus raphanistrum
Rra.25027: 2 sequences.

Taxonomic Groups [List]

- flowering plants *(5)*
  - eudicots *(3)*
    - Fabales *(1)*
    - Brassicales *(1)*
    - Solanales *(1)*
  - monocots *(2)*
    - Hordeum *(1)*
    - Oryza *(1)*

G – Entrez Gene



□ **1: POPTRDRAFT_712840 hypothetical protein [** *Populus trichocarpa* **]**

GeneID: 7463061 updated 07-May-2009

Summary

Locus tag	POPTRDRAFT_712840
Gene type	protein coding
RNA name	predicted protein
RefSeq status	PROVISIONAL
Organism	*Populus trichocarpa*
Lineage	*Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids; eurosids I; Malpighiales; Salicaceae; Saliceae; Populus*

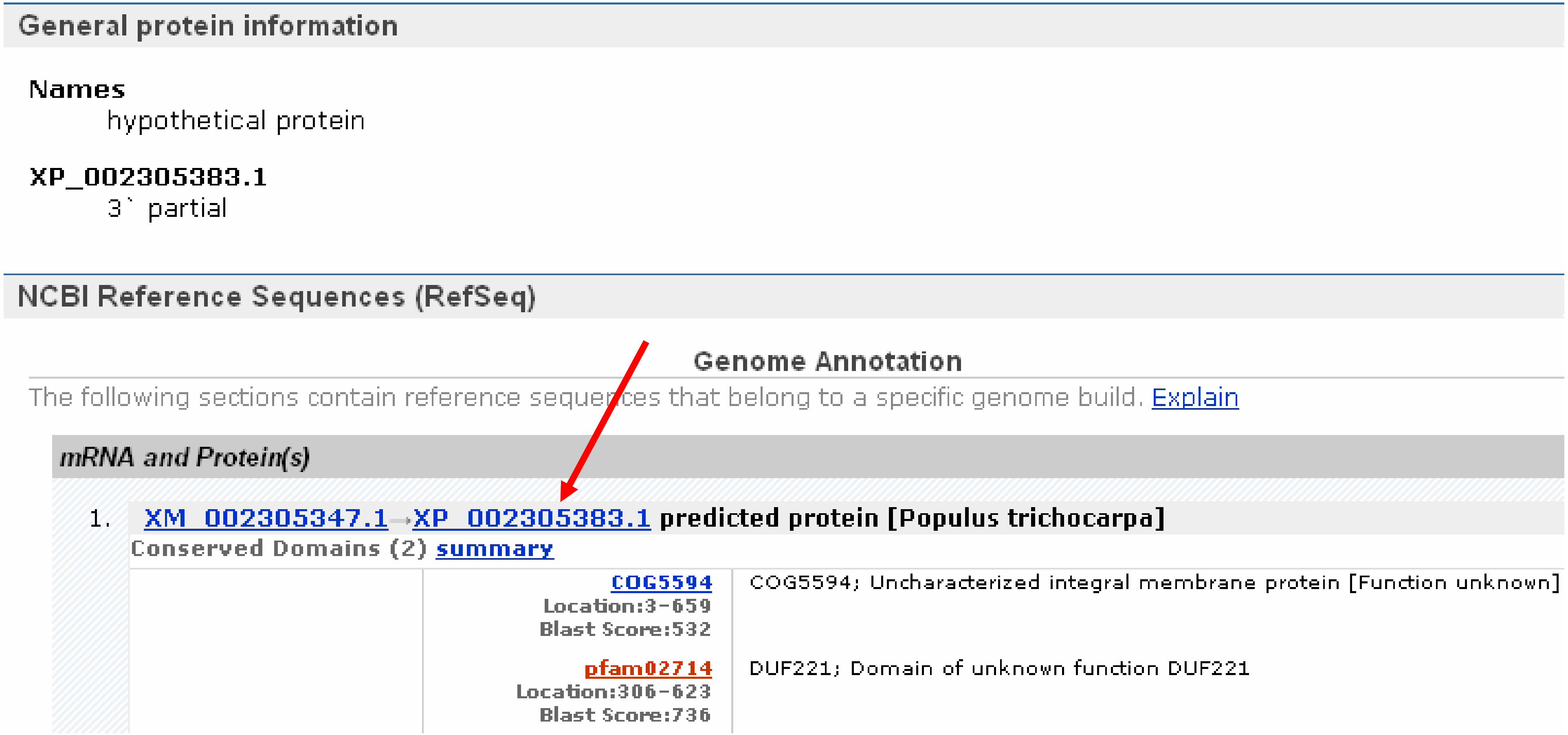
Genomic regions, transcripts, and products

Go to reference sequence details

Try our new Sequence Viewer

28.7.2009

G – Entrez Gene



General protein information

Names

hypothetical protein

XP_002305383.1

3` partial

NCBI Reference Sequences (RefSeq)

Genome Annotation

The following sections contain reference sequences that belong to a specific genome build. Explain

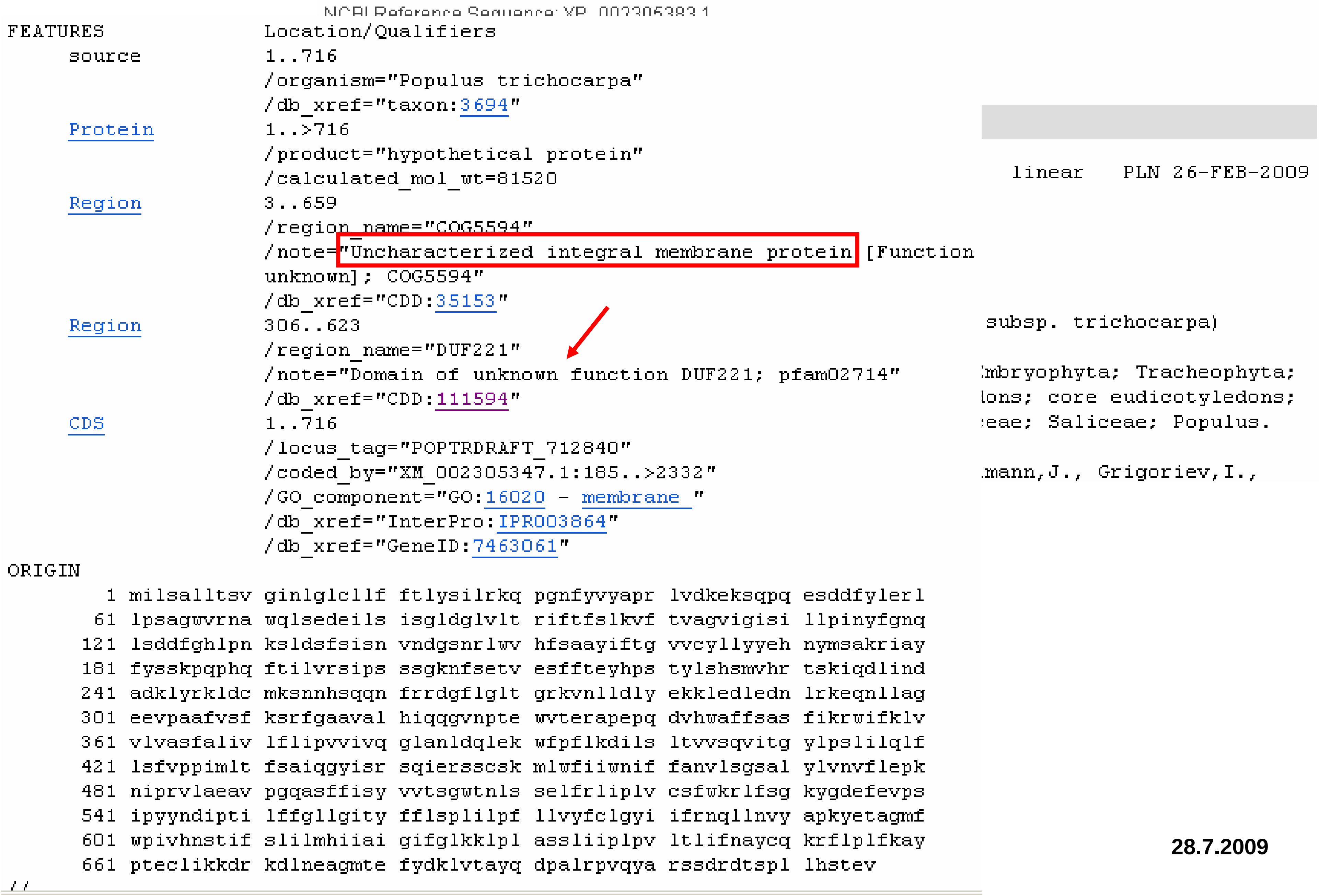
mRNA and Protein(s)

1. XM_002305347.1→XP_002305383.1 **predicted protein [Populus trichocarpa]**

Conserved Domains (2) summary

	COG5594 Location:3-659 Blast Score:532	COG5594; Uncharacterized integral membrane protein [Function unknown]
	pfam02714 Location:306-623 Blast Score:736	DUF221; Domain of unknown function DUF221

Domains – Function - Localization?



NCBI Reference Sequence: XP_002306383.1

```
FEATURES             Location/Qualifiers
     source          1..716
                     /organism="Populus trichocarpa"
                     /db_xref="taxon:3694"
     Protein         1..>716
                     /product="hypothetical protein"
                     /calculated_mol_wt=81520
     Region          3..659
                     /region_name="COG5594"
                     /note="Uncharacterized integral membrane protein [Function
                     unknown]; COG5594"
                     /db_xref="CDD:35153"
     Region          306..623
                     /region_name="DUF221"
                     /note="Domain of unknown function DUF221; pfam02714"
                     /db_xref="CDD:111594"
     CDS             1..716
                     /locus_tag="POPTRDRAFT_712840"
                     /coded_by="XM_002305347.1:185..>2332"
                     /GO_component="GO:16020 - membrane "
                     /db_xref="InterPro:IPR003864"
                     /db_xref="GeneID:7463061"
ORIGIN
        1 milsalltsv ginlglcllf ftlysilrkq pgnfyvyapr lvdkeksqpq esddfylerl
       61 lpsagwvrna wqlsedeils isgldglvlt riftfslkvf tvagvigisi llpinyfgnq
      121 lsddfghlpn ksldsfsisn vndgsnrlwv hfsaayiftg vvcyllyyeh nymsakriay
      181 fysskpqphq ftilvrsips ssgknfsetv esffteyhps tylshsmvhr tskiqdlind
      241 adklyrkldc mksnnhsqqn frrdgflglt grkvnlldly ekkledledn lrkeqnllag
      301 eevpaafvsf ksrfgaaval hiqqgvnpte wvterapepq dvhwaffsas fikrwifklv
      361 vlvasfaliv lflipvvivq glanldqlek wfpflkdils ltvvsqvitg ylpslilqlf
      421 lsfvppimlt fsaiqgyisr sqiersscsk mlwfiiwnif fanvlsgsal ylvnvflepk
      481 niprvlaeav pgqasffisy vvtsgwtnls selfrliplv csfwkrlfsg kygdefevps
      541 ipyyndipti lffgllgity fflsplilpf llvyfclgyi ifrnqllnvy apkyetagmf
      601 wpivhnstif slilmhiiai gifglkklpl assliiplpv ltlifnaycq krflplfkay
      661 pteclikkdr kdlneagmte fydklvtayq dpalrpvqya rssdrdtspl lhstev
//
```

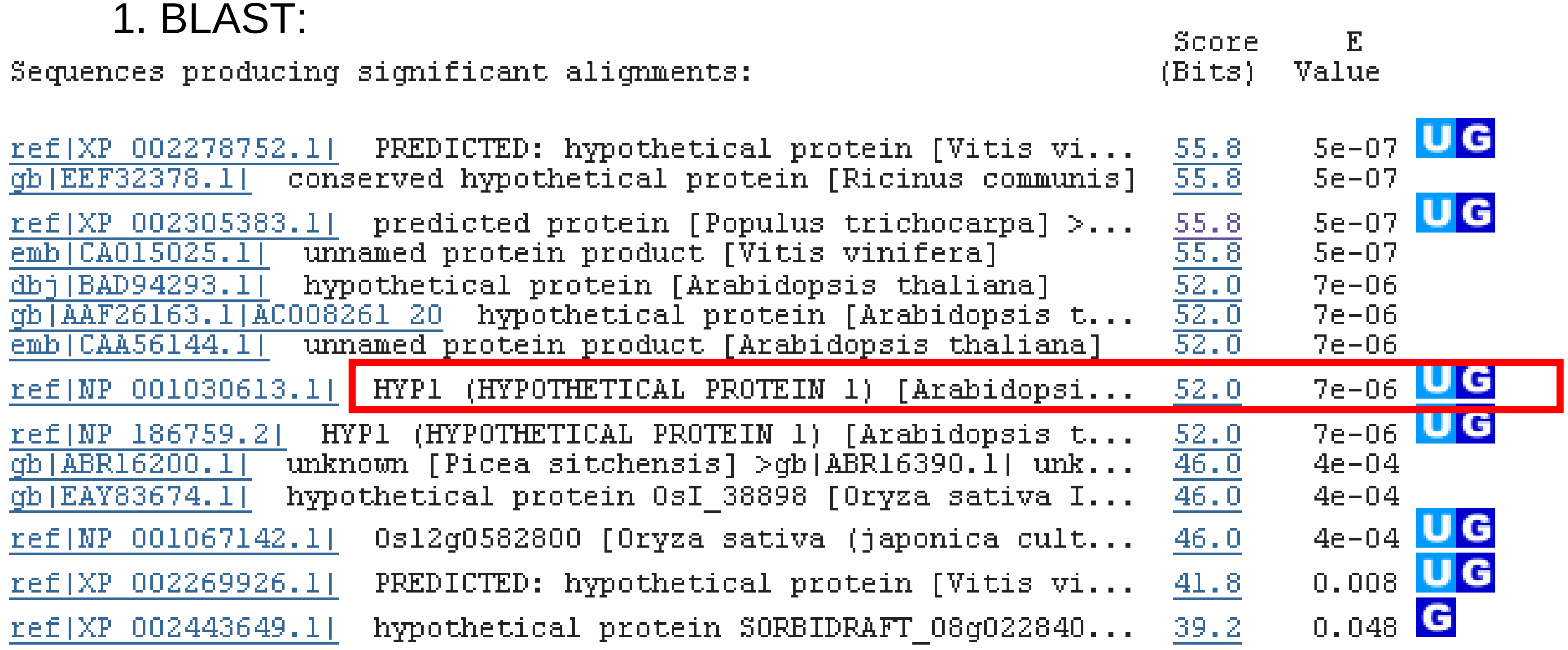
```
 linear   PLN 26-FEB-2009
subsp. trichocarpa)
:mbryophyta; Tracheophyta;
lons; core eudicotyledons;
:eae; Saliceae; Populus.
.mann,J., Grigoriev,I.,
```

28.7.2009

Example

Proteome analysis in Poplar result a peptide of MILSALLTSVGINLGLC

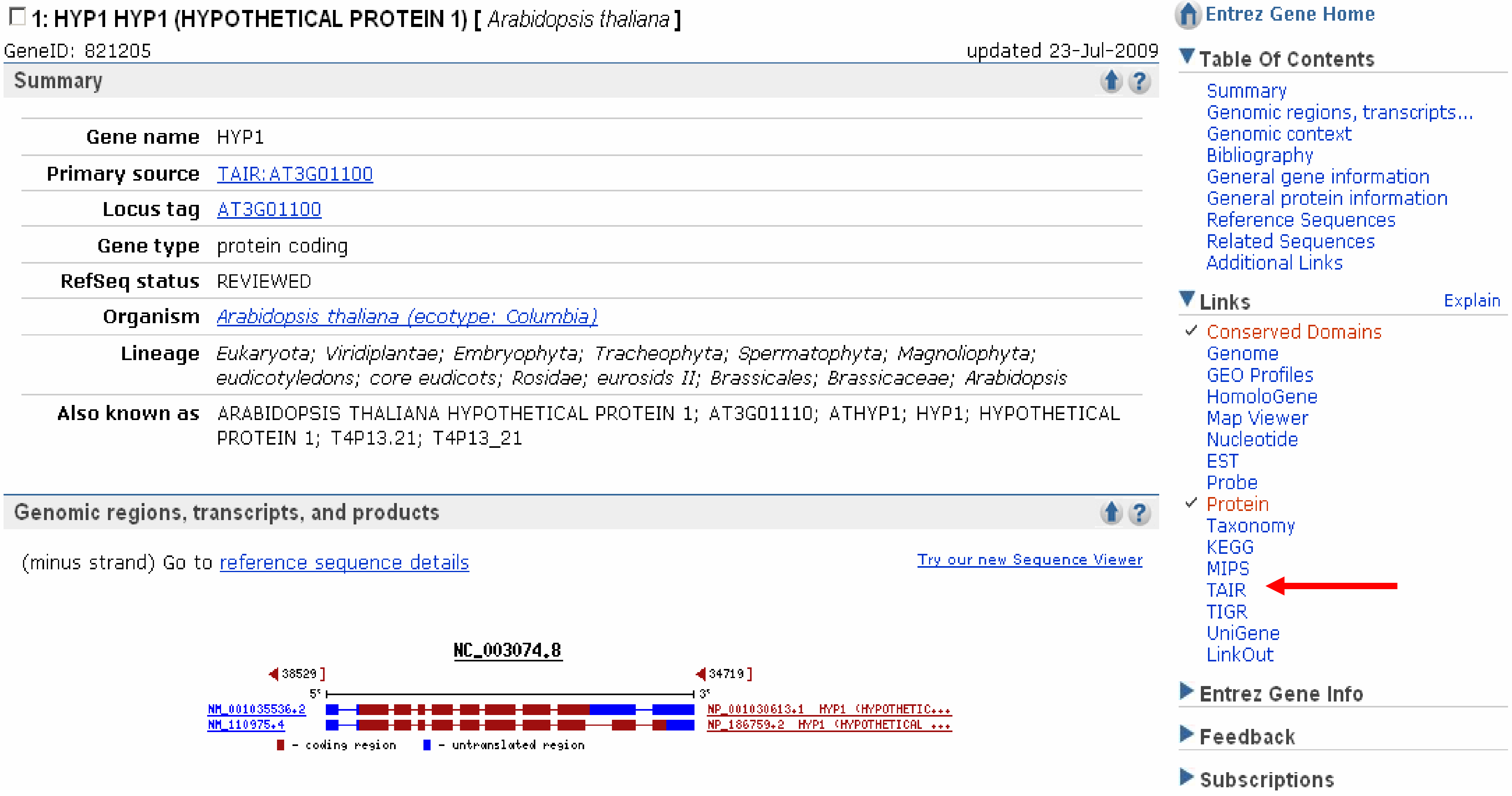
1. BLAST:



```
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

ref|XP_002278752.1|  PREDICTED: hypothetical protein [Vitis vi...    55.8    5e-07  U G
gb|EEF32378.1|  conserved hypothetical protein [Ricinus communis]    55.8    5e-07
ref|XP_002305383.1|  predicted protein [Populus trichocarpa] >...    55.8    5e-07  U G
emb|CAO15025.1|  unnamed protein product [Vitis vinifera]            55.8    5e-07
dbj|BAD94293.1|  hypothetical protein [Arabidopsis thaliana]         52.0    7e-06
gb|AAF26163.1|AC008261_20  hypothetical protein [Arabidopsis t...    52.0    7e-06
emb|CAA56144.1|  unnamed protein product [Arabidopsis thaliana]      52.0    7e-06
ref|NP_001030613.1|  HYP1 (HYPOTHETICAL PROTEIN 1) [Arabidopsi...    52.0    7e-06  U G
ref|NP_186759.2|  HYP1 (HYPOTHETICAL PROTEIN 1) [Arabidopsis t...    52.0    7e-06  U G
gb|ABR16200.1|  unknown [Picea sitchensis] >gb|ABR16390.1| unk...    46.0    4e-04
gb|EAY83674.1|  hypothetical protein OsI_38898 [Oryza sativa I...    46.0    4e-04
ref|NP_001067142.1|  Os12g0582800 [Oryza sativa (japonica cult...    46.0    4e-04  U G
ref|XP_002269926.1|  PREDICTED: hypothetical protein [Vitis vi...    41.8    0.008  U G
ref|XP_002443649.1|  hypothetical protein SORBIDRAFT_08g022840...    39.2    0.048  G
```

G – Entrez Gene



□ **1: HYP1 HYP1 (HYPOTHETICAL PROTEIN 1) [*Arabidopsis thaliana*]**

GeneID: 821205 updated 23-Jul-2009

Summary

Gene name	HYP1
Primary source	TAIR:AT3G01100
Locus tag	AT3G01100
Gene type	protein coding
RefSeq status	REVIEWED
Organism	*Arabidopsis thaliana (ecotype: Columbia)*
Lineage	*Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; Rosidae; eurosids II; Brassicales; Brassicaceae; Arabidopsis*
Also known as	ARABIDOPSIS THALIANA HYPOTHETICAL PROTEIN 1; AT3G01110; ATHYP1; HYP1; HYPOTHETICAL PROTEIN 1; T4P13.21; T4P13_21

Genomic regions, transcripts, and products

(minus strand) Go to reference sequence details

Try our new Sequence Viewer

Entrez Gene Home

Table Of Contents

- Summary
- Genomic regions, transcripts...
- Genomic context
- Bibliography
- General gene information
- General protein information
- Reference Sequences
- Related Sequences
- Additional Links

Links Explain

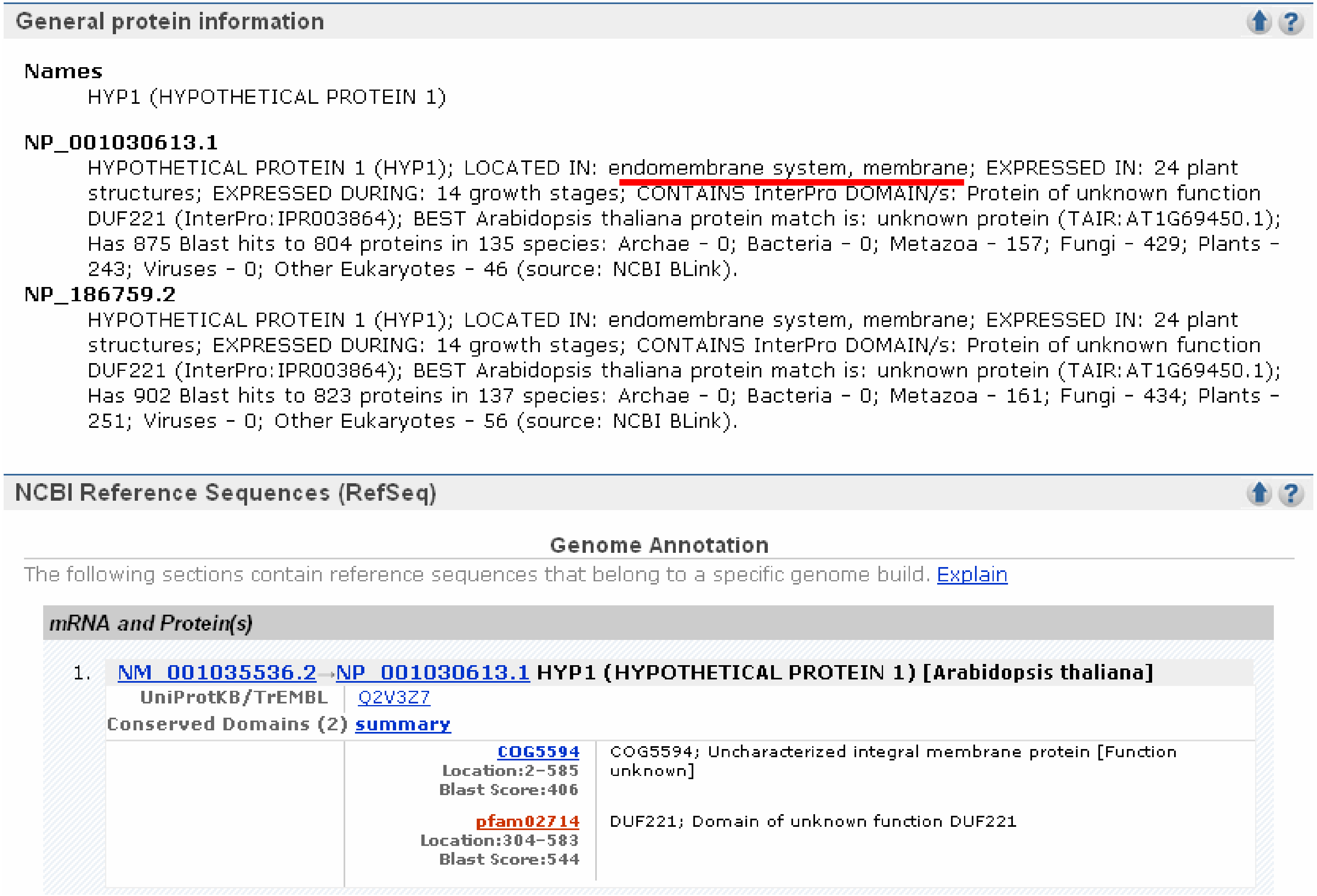
- ✓ Conserved Domains
- Genome
- GEO Profiles
- HomoloGene
- Map Viewer
- Nucleotide
- EST
- Probe
- ✓ Protein
- Taxonomy
- KEGG
- MIPS
- TAIR ←
- TIGR
- UniGene
- LinkOut

Entrez Gene Info

Feedback

Subscriptions

G – Entrez Gene



General protein information

Names

HYP1 (HYPOTHETICAL PROTEIN 1)

NP_001030613.1

HYPOTHETICAL PROTEIN 1 (HYP1); LOCATED IN: endomembrane system, membrane; EXPRESSED IN: 24 plant structures; EXPRESSED DURING: 14 growth stages; CONTAINS InterPro DOMAIN/s: Protein of unknown function DUF221 (InterPro:IPR003864); BEST Arabidopsis thaliana protein match is: unknown protein (TAIR:AT1G69450.1); Has 875 Blast hits to 804 proteins in 135 species: Archae - 0; Bacteria - 0; Metazoa - 157; Fungi - 429; Plants - 243; Viruses - 0; Other Eukaryotes - 46 (source: NCBI BLink).

NP_186759.2

HYPOTHETICAL PROTEIN 1 (HYP1); LOCATED IN: endomembrane system, membrane; EXPRESSED IN: 24 plant structures; EXPRESSED DURING: 14 growth stages; CONTAINS InterPro DOMAIN/s: Protein of unknown function DUF221 (InterPro:IPR003864); BEST Arabidopsis thaliana protein match is: unknown protein (TAIR:AT1G69450.1); Has 902 Blast hits to 823 proteins in 137 species: Archae - 0; Bacteria - 0; Metazoa - 161; Fungi - 434; Plants - 251; Viruses - 0; Other Eukaryotes - 56 (source: NCBI BLink).

NCBI Reference Sequences (RefSeq)

Genome Annotation

The following sections contain reference sequences that belong to a specific genome build. Explain

mRNA and Protein(s)

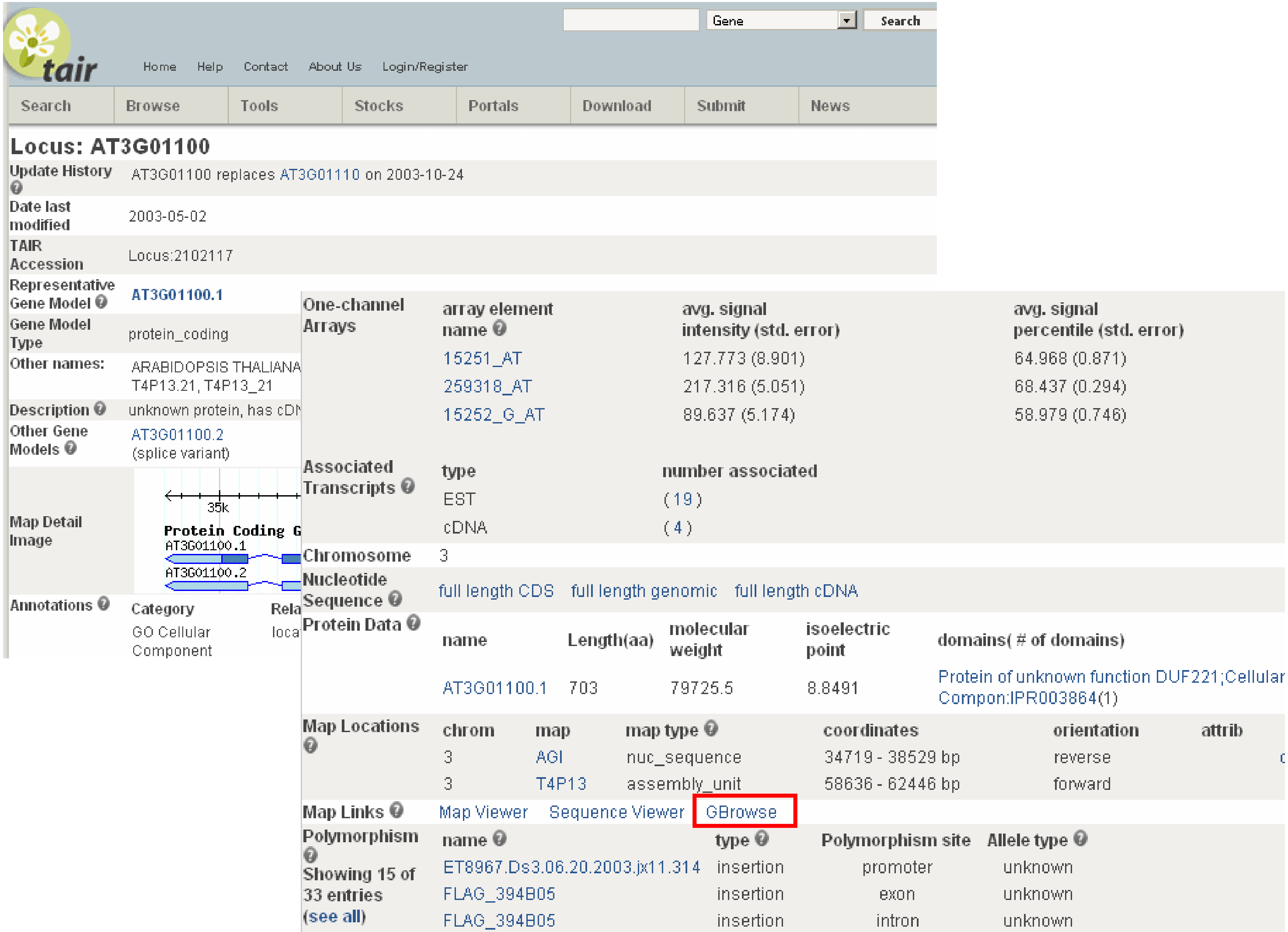
1. NM_001035536.2 → NP_001030613.1 **HYP1 (HYPOTHETICAL PROTEIN 1) [Arabidopsis thaliana]**

UniProtKB/TrEMBL Q2V3Z7

Conserved Domains (2) **summary**

COG5594 Location:2-585 Blast Score:406	COG5594; Uncharacterized integral membrane protein [Function unknown]
pfam02714 Location:304-583 Blast Score:544	DUF221; Domain of unknown function DUF221

28.7.2009



Locus: AT3G01100

Update History	AT3G01100 replaces AT3G01110 on 2003-10-24
Date last modified	2003-05-02
TAIR Accession	Locus:2102117
Representative Gene Model	AT3G01100.1
Gene Model Type	protein_coding
Other names:	ARABIDOPSIS THALIANA T4P13.21, T4P13_21
Description	unknown protein, has cDN
Other Gene Models	AT3G01100.2 (splice variant)
Map Detail Image	
Annotations	Category: GO Cellular Component; Rela: loca

One-channel Arrays

array element name	avg. signal intensity (std. error)	avg. signal percentile (std. error)
15251_AT	127.773 (8.901)	64.968 (0.871)
259318_AT	217.316 (5.051)	68.437 (0.294)
15252_G_AT	89.637 (5.174)	58.979 (0.746)

Associated Transcripts

type	number associated
EST	(19)
cDNA	(4)

Chromosome 3

Nucleotide Sequence full length CDS full length genomic full length cDNA

Protein Data

name	Length(aa)	molecular weight	isoelectric point	domains(# of domains)
AT3G01100.1	703	79725.5	8.8491	Protein of unknown function DUF221;Cellular Compon:IPR003864(1)

Map Locations

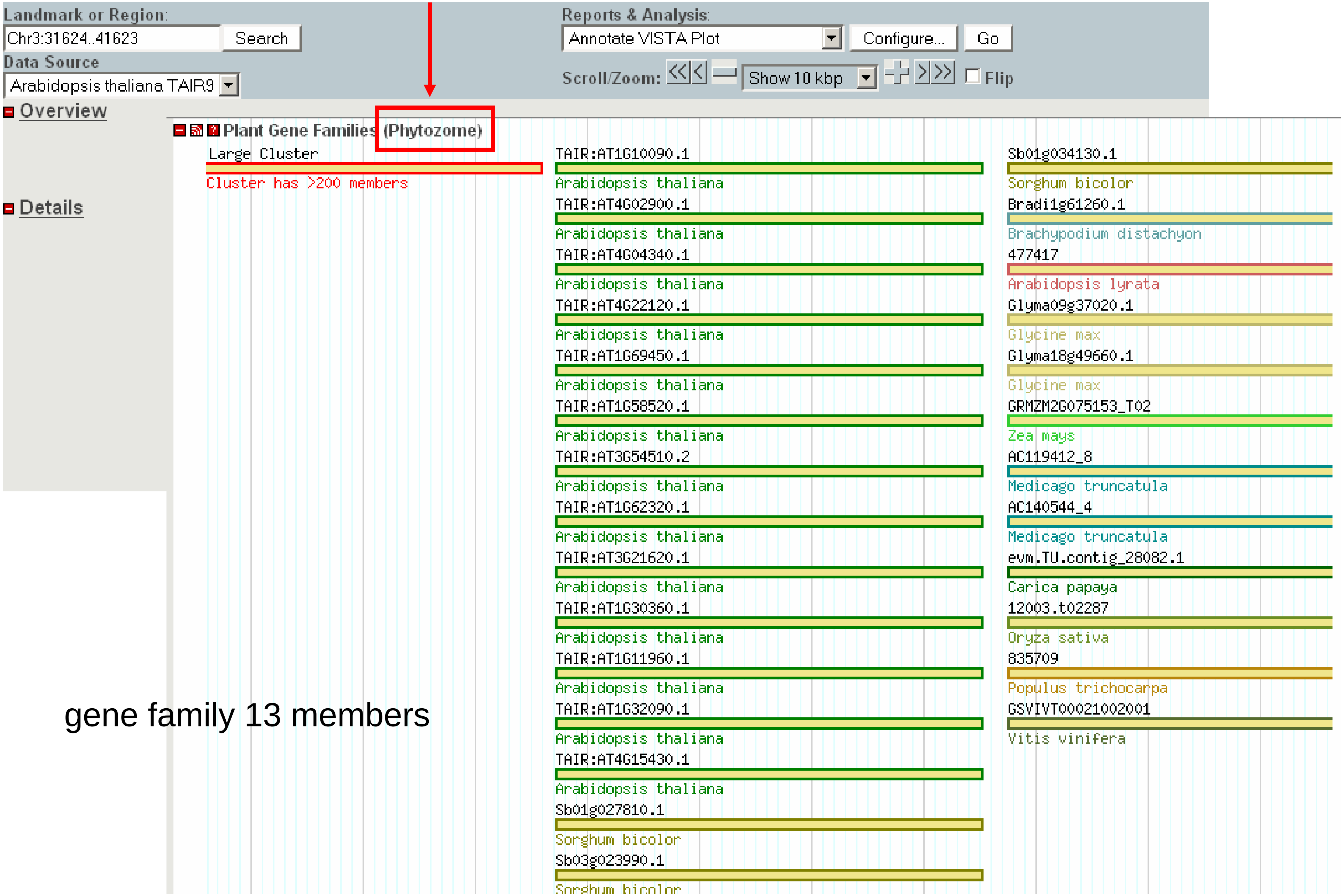
chrom	map	map type	coordinates	orientation	attrib
3	AGI	nuc_sequence	34719 - 38529 bp	reverse	c
3	T4P13	assembly_unit	58636 - 62446 bp	forward	

Map Links Map Viewer Sequence Viewer GBrowse

Polymorphism
Showing 15 of 33 entries (see all)

name	type	Polymorphism site	Allele type
ET8967.Ds3.06.20.2003.jx11.314	insertion	promoter	unknown
FLAG_394B05	insertion	exon	unknown
FLAG_394B05	insertion	intron	unknown

GBrowser



Landmark or Region:

Chr3:31624..41623 Search

Data Source

Arabidopsis thaliana TAIR9

Reports & Analysis:

Annotate VISTA Plot Configure... Go

Scroll/Zoom: Show 10 kbp □ Flip

Overview

Details

Plant Gene Families (Phytozome)

Large Cluster

Cluster has >200 members

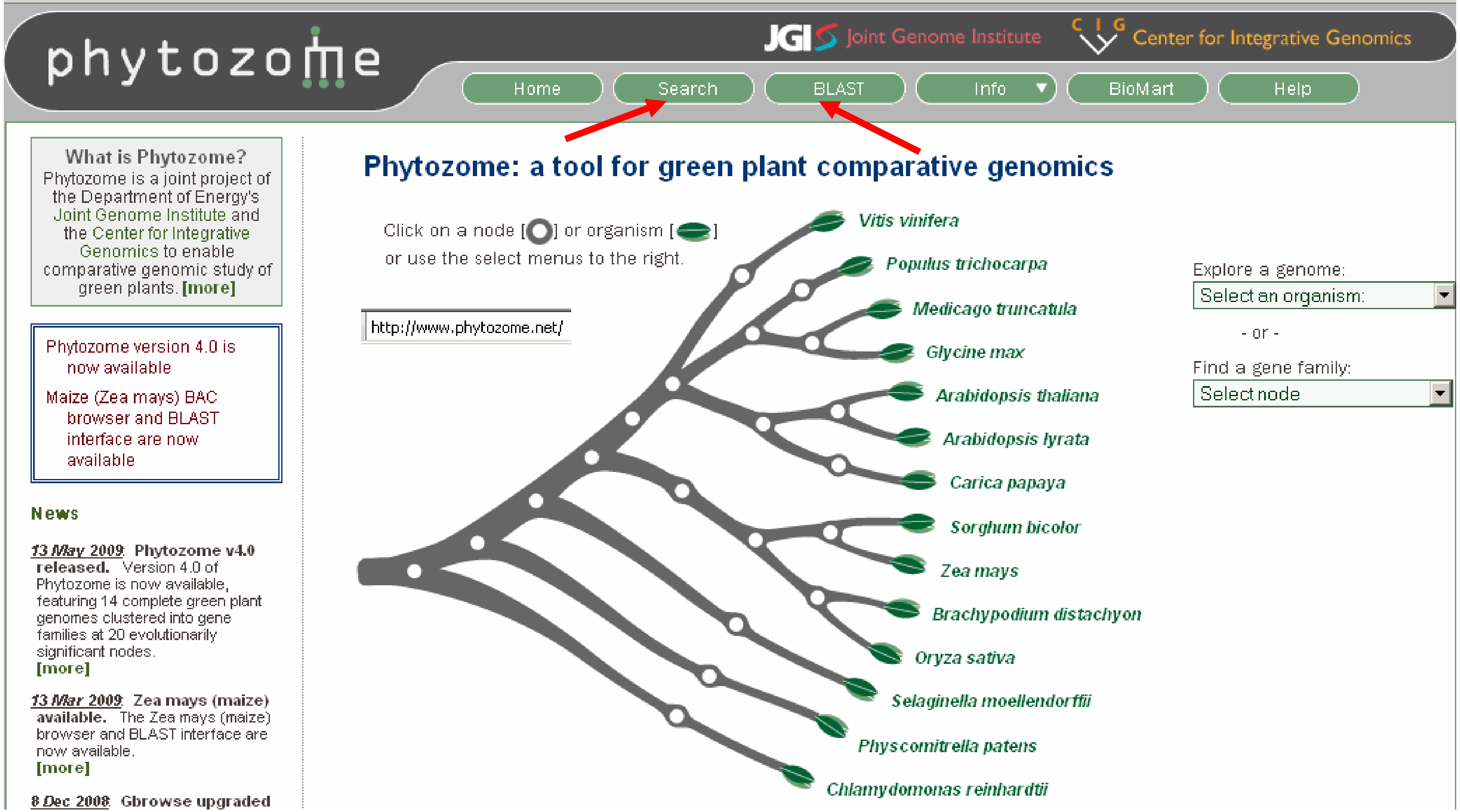
TAIR:AT1G10090.1
Arabidopsis thaliana
TAIR:AT4G02900.1
Arabidopsis thaliana
TAIR:AT4G04340.1
Arabidopsis thaliana
TAIR:AT4G22120.1
Arabidopsis thaliana
TAIR:AT1G69450.1
Arabidopsis thaliana
TAIR:AT1G58520.1
Arabidopsis thaliana
TAIR:AT3G54510.2
Arabidopsis thaliana
TAIR:AT1G62320.1
Arabidopsis thaliana
TAIR:AT3G21620.1
Arabidopsis thaliana
TAIR:AT1G30360.1
Arabidopsis thaliana
TAIR:AT1G11960.1
Arabidopsis thaliana
TAIR:AT1G32090.1
Arabidopsis thaliana
TAIR:AT4G15430.1
Arabidopsis thaliana
Sb01g027810.1
Sorghum bicolor
Sb03g023990.1
Sorghum bicolor

Sb01g034130.1
Sorghum bicolor
Bradi1g61260.1
Brachypodium distachyon
477417
Arabidopsis lyrata
Glyma09g37020.1
Glycine max
Glyma18g49660.1
Glycine max
GRMZM2G075153_T02
Zea mays
AC119412_8
Medicago truncatula
AC140544_4
Medicago truncatula
evm.TU.contig_28082.1
Carica papaya
12003.t02287
Oryza sativa
835709
Populus trichocarpa
GSVIVT00021002001
Vitis vinifera

gene family 13 members

Synteny Search

Can the annotation of one member of the gene family in any plant species guide to the function?



What is Phytozome?
Phytozome is a joint project of the Department of Energy's Joint Genome Institute and the Center for Integrative Genomics to enable comparative genomic study of green plants. **[more]**

Phytozome version 4.0 is now available

Maize (Zea mays) BAC browser and BLAST interface are now available

News

13 May 2009 **Phytozome v4.0 released.** Version 4.0 of Phytozome is now available, featuring 14 complete green plant genomes clustered into gene families at 20 evolutionarily significant nodes.
[more]

13 Mar 2009 **Zea mays (maize) available.** The Zea mays (maize) browser and BLAST interface are now available.
[more]

8 Dec 2008 **Gbrowse upgraded**

Phytozome: a tool for green plant comparative genomics

Click on a node [○] or organism [●] or use the select menus to the right.

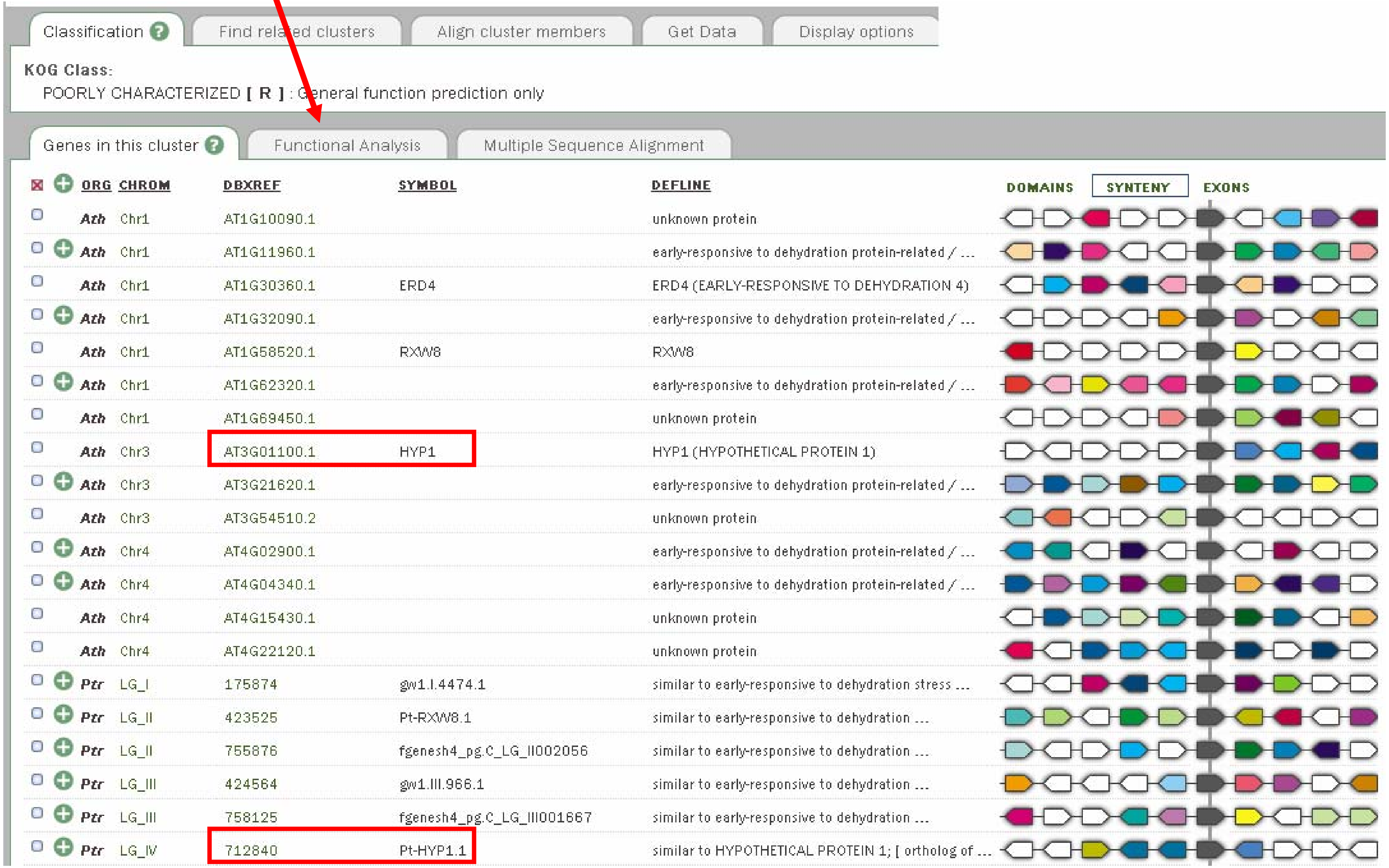
Explore a genome:
Select an organism: ▼

- or -

Find a gene family:
Select node ▼

28.7.2009

Search for HYP1



Classification ? | Find related clusters | Align cluster members | Get Data | Display options

KOG Class:

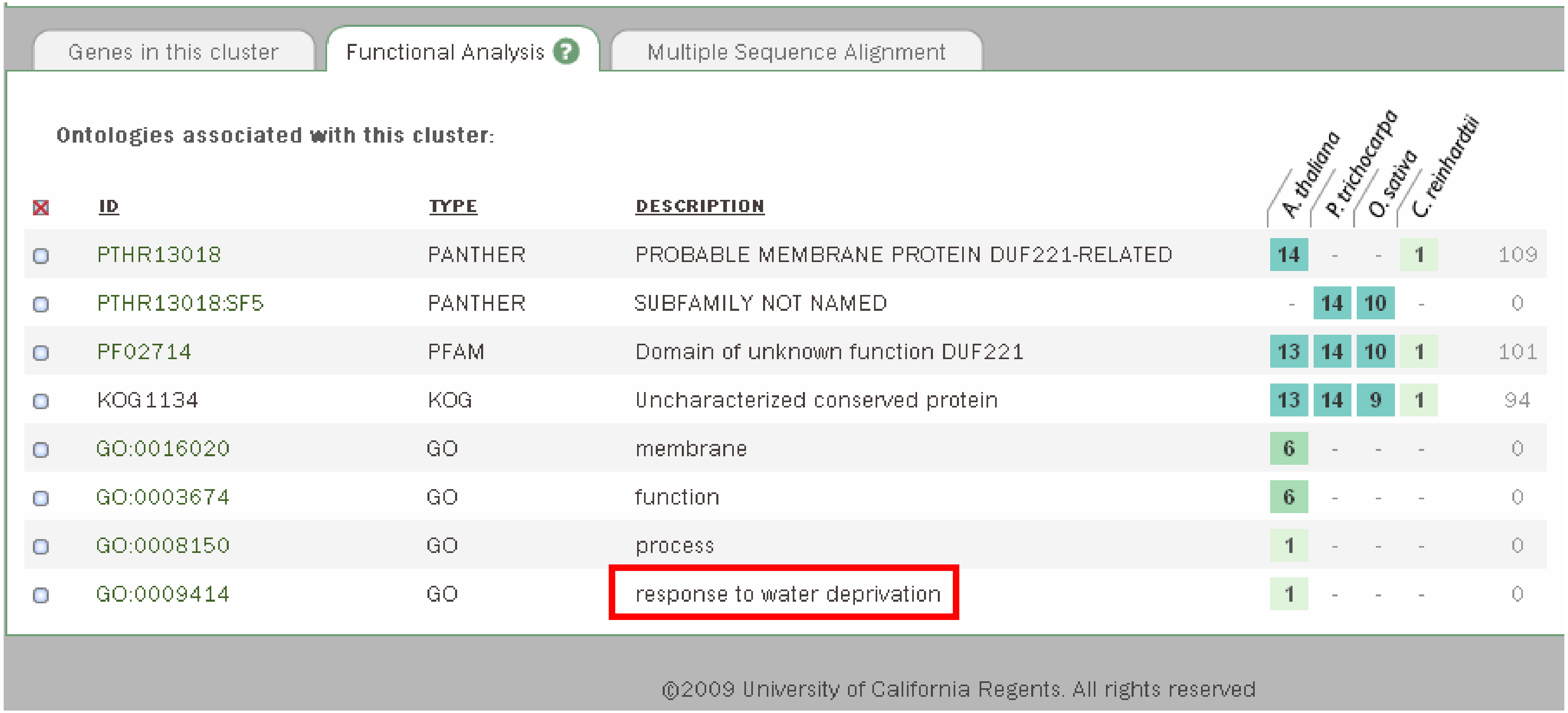
POORLY CHARACTERIZED **[R]** : General function prediction only

Genes in this cluster ? | Functional Analysis | Multiple Sequence Alignment

☒	⊕	ORG	CHROM	DBXREF	SYMBOL	DEFLINE	DOMAINS SYNTENY EXONS
○		*Ath*	Chr1	AT1G10090.1		unknown protein	
○	⊕	*Ath*	Chr1	AT1G11960.1		early-responsive to dehydration protein-related / ...	
○		*Ath*	Chr1	AT1G30360.1	ERD4	ERD4 (EARLY-RESPONSIVE TO DEHYDRATION 4)	
○	⊕	*Ath*	Chr1	AT1G32090.1		early-responsive to dehydration protein-related / ...	
○		*Ath*	Chr1	AT1G58520.1	RXW8	RXW8	
○	⊕	*Ath*	Chr1	AT1G62320.1		early-responsive to dehydration protein-related / ...	
○		*Ath*	Chr1	AT1G69450.1		unknown protein	
○		*Ath*	Chr3	AT3G01100.1	HYP1	HYP1 (HYPOTHETICAL PROTEIN 1)	
○	⊕	*Ath*	Chr3	AT3G21620.1		early-responsive to dehydration protein-related / ...	
○		*Ath*	Chr3	AT3G54510.2		unknown protein	
○	⊕	*Ath*	Chr4	AT4G02900.1		early-responsive to dehydration protein-related / ...	
○	⊕	*Ath*	Chr4	AT4G04340.1		early-responsive to dehydration protein-related / ...	
○		*Ath*	Chr4	AT4G15430.1		unknown protein	
○		*Ath*	Chr4	AT4G22120.1		unknown protein	
○	⊕	*Ptr*	LG_I	175874	gw1.I.4474.1	similar to early-responsive to dehydration stress ...	
○	⊕	*Ptr*	LG_II	423525	Pt-RXW8.1	similar to early-responsive to dehydration ...	
○	⊕	*Ptr*	LG_II	755876	fgenesh4_pg.C_LG_II002056	similar to early-responsive to dehydration ...	
○	⊕	*Ptr*	LG_III	424564	gw1.III.966.1	similar to early-responsive to dehydration ...	
○	⊕	*Ptr*	LG_III	758125	fgenesh4_pg.C_LG_III001667	similar to early-responsive to dehydration ...	
○	⊕	*Ptr*	LG_IV	712840	Pt-HYP1.1	similar to HYPOTHETICAL PROTEIN 1; [ortholog of ...	

28.7.2009

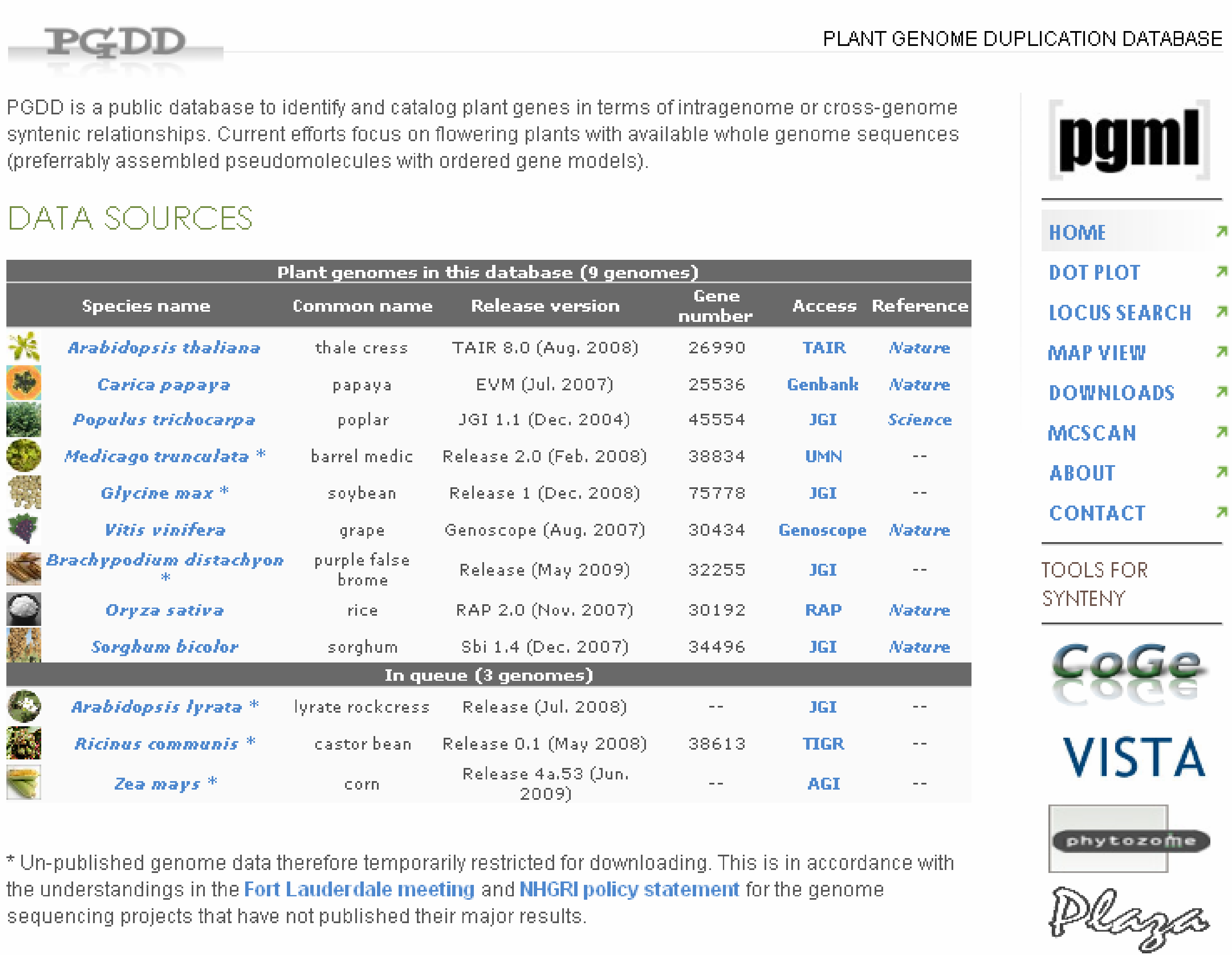
Common Function?



Ontologies associated with this cluster:

☒	ID	TYPE	DESCRIPTION	A. thaliana	P. trichocarpa	O. sativa	C. reinhardtii	
☐	PTHR13018	PANTHER	PROBABLE MEMBRANE PROTEIN DUF221-RELATED	14	-	-	1	109
☐	PTHR13018:SF5	PANTHER	SUBFAMILY NOT NAMED	-	14	10	-	0
☐	PF02714	PFAM	Domain of unknown function DUF221	13	14	10	1	101
☐	KOG1134	KOG	Uncharacterized conserved protein	13	14	9	1	94
☐	GO:0016020	GO	membrane	6	-	-	-	0
☐	GO:0003674	GO	function	6	-	-	-	0
☐	GO:0008150	GO	process	1	-	-	-	0
☐	GO:0009414	GO	response to water deprivation	1	-	-	-	0

Other Synteny Tools



PLANT GENOME DUPLICATION DATABASE

PGDD is a public database to identify and catalog plant genes in terms of intragenome or cross-genome syntenic relationships. Current efforts focus on flowering plants with available whole genome sequences (preferrably assembled pseudomolecules with ordered gene models).

DATA SOURCES

Plant genomes in this database (9 genomes)					
Species name	Common name	Release version	Gene number	Access	Reference
Arabidopsis thaliana	thale cress	TAIR 8.0 (Aug. 2008)	26990	TAIR	*Nature*
Carica papaya	papaya	EVM (Jul. 2007)	25536	Genbank	*Nature*
Populus trichocarpa	poplar	JGI 1.1 (Dec. 2004)	45554	JGI	*Science*
Medicago truncatula *	barrel medic	Release 2.0 (Feb. 2008)	38834	UMN	--
Glycine max *	soybean	Release 1 (Dec. 2008)	75778	JGI	--
Vitis vinifera	grape	Genoscope (Aug. 2007)	30434	Genoscope	*Nature*
Brachypodium distachyon *	purple false brome	Release (May 2009)	32255	JGI	--
Oryza sativa	rice	RAP 2.0 (Nov. 2007)	30192	RAP	*Nature*
Sorghum bicolor	sorghum	Sbi 1.4 (Dec. 2007)	34496	JGI	*Nature*
In queue (3 genomes)					
Arabidopsis lyrata *	lyrate rockcress	Release (Jul. 2008)	--	JGI	--
Ricinus communis *	castor bean	Release 0.1 (May 2008)	38613	TIGR	--
Zea mays *	corn	Release 4a.53 (Jun. 2009)	--	AGI	--

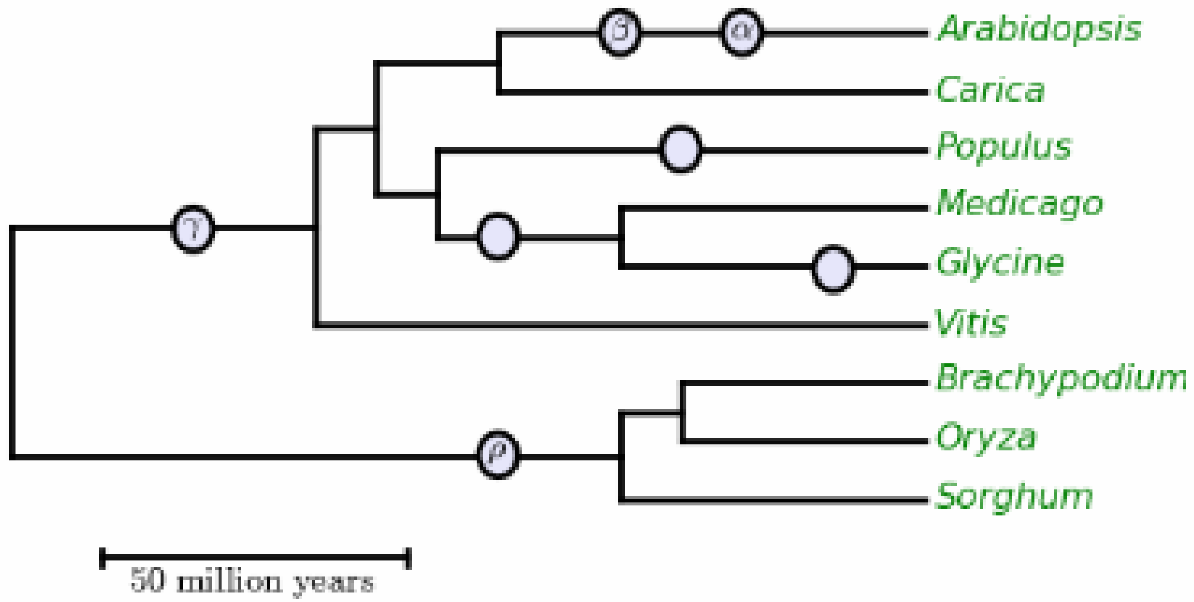
* Un-published genome data therefore temporarily restricted for downloading. This is in accordance with the understandings in the Fort Lauderdale meeting and NHGRI policy statement for the genome sequencing projects that have not published their major results.

http://chibba.agtec.uga.edu/duplication/

Plant Genome Duplication Database (PGDD)

The duplication history of major angiosperm taxa

lavender circles represent inferred polyploidy events, drawn roughly to scale

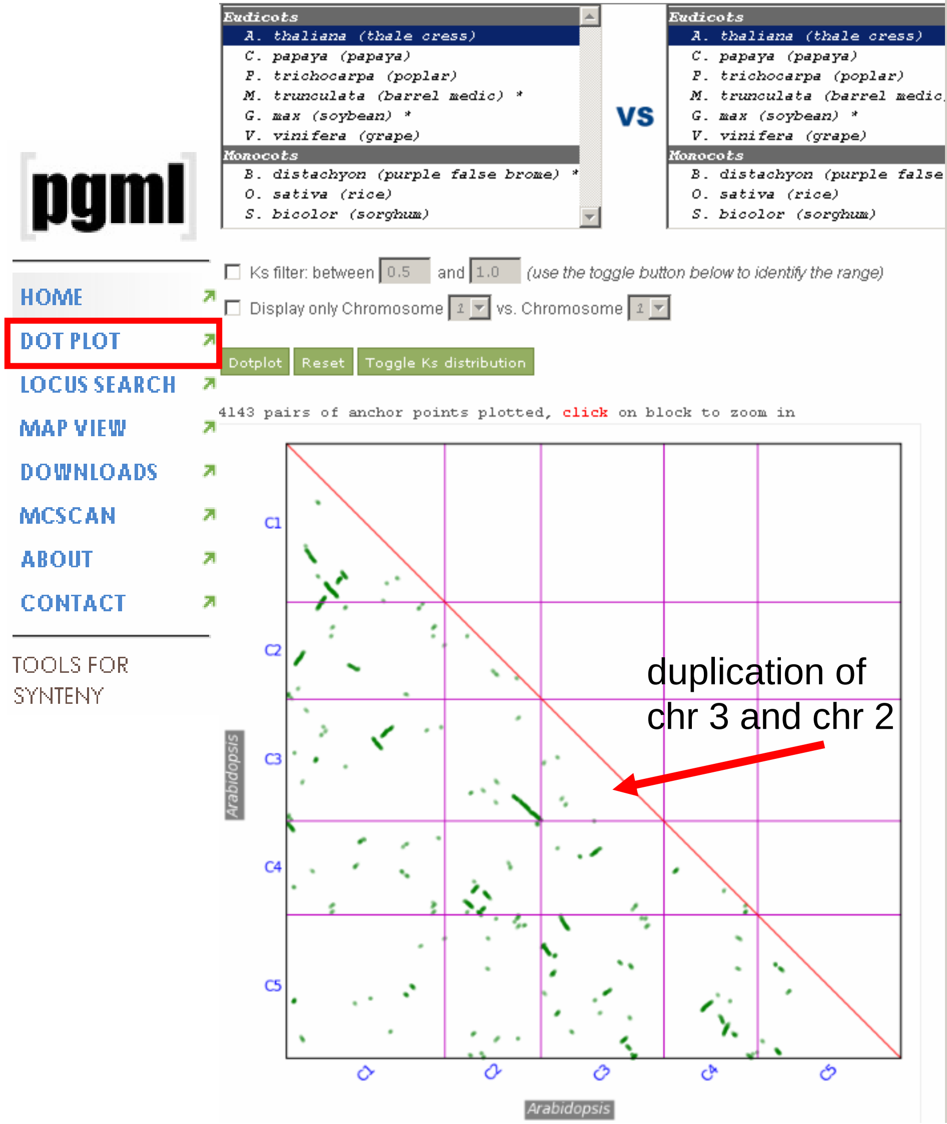


http://chibba.agtec.uga.edu/duplication/

28.7.2009

Intra-genome Dotblot analysis at PGDD

non-synonymous substitution (Ka)
synonyoumous substitution (Ks)

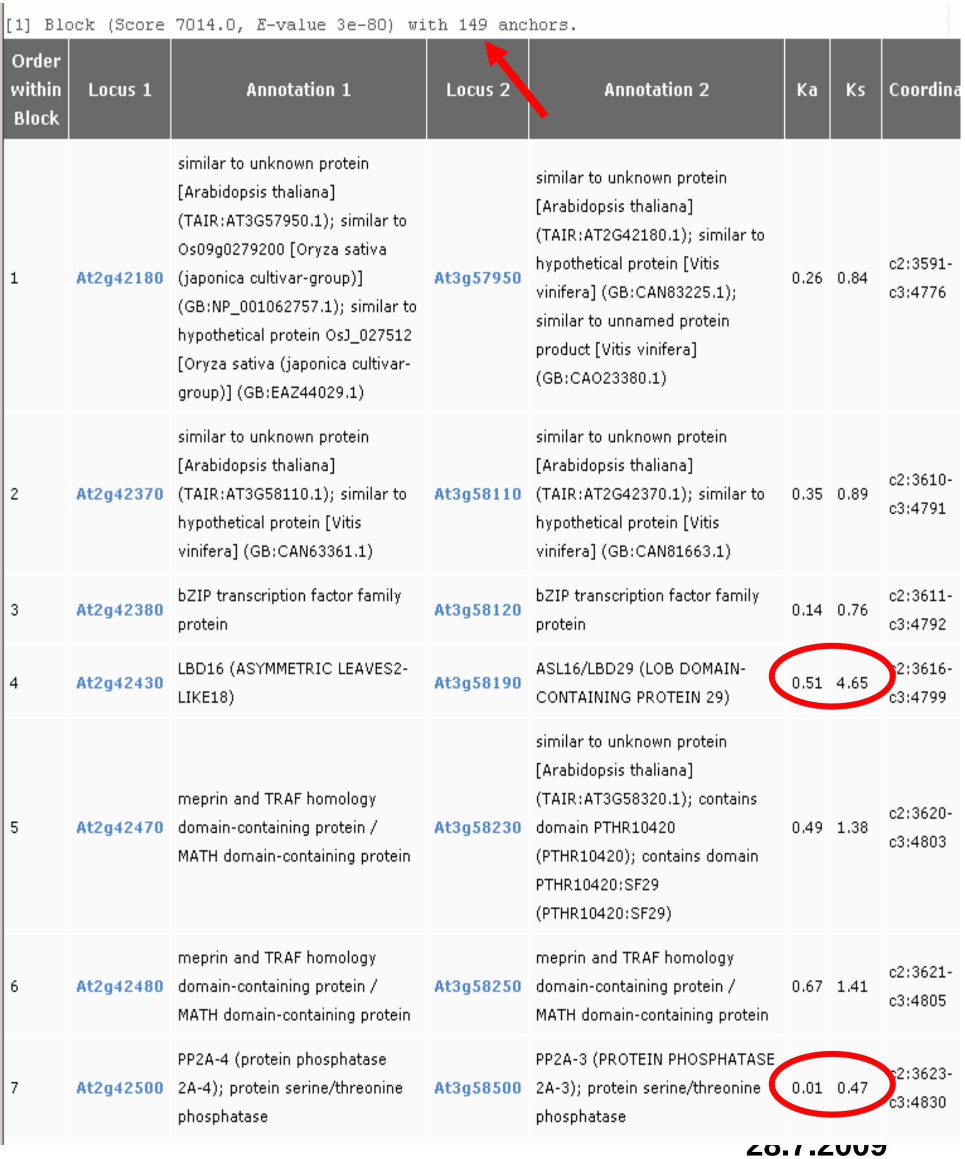


☐ Ks filter: between 0.5 and 1.0 (use the toggle button below to identify the range)

☐ Display only Chromosome 1 vs. Chromosome 1

4143 pairs of anchor points plotted, click on block to zoom in

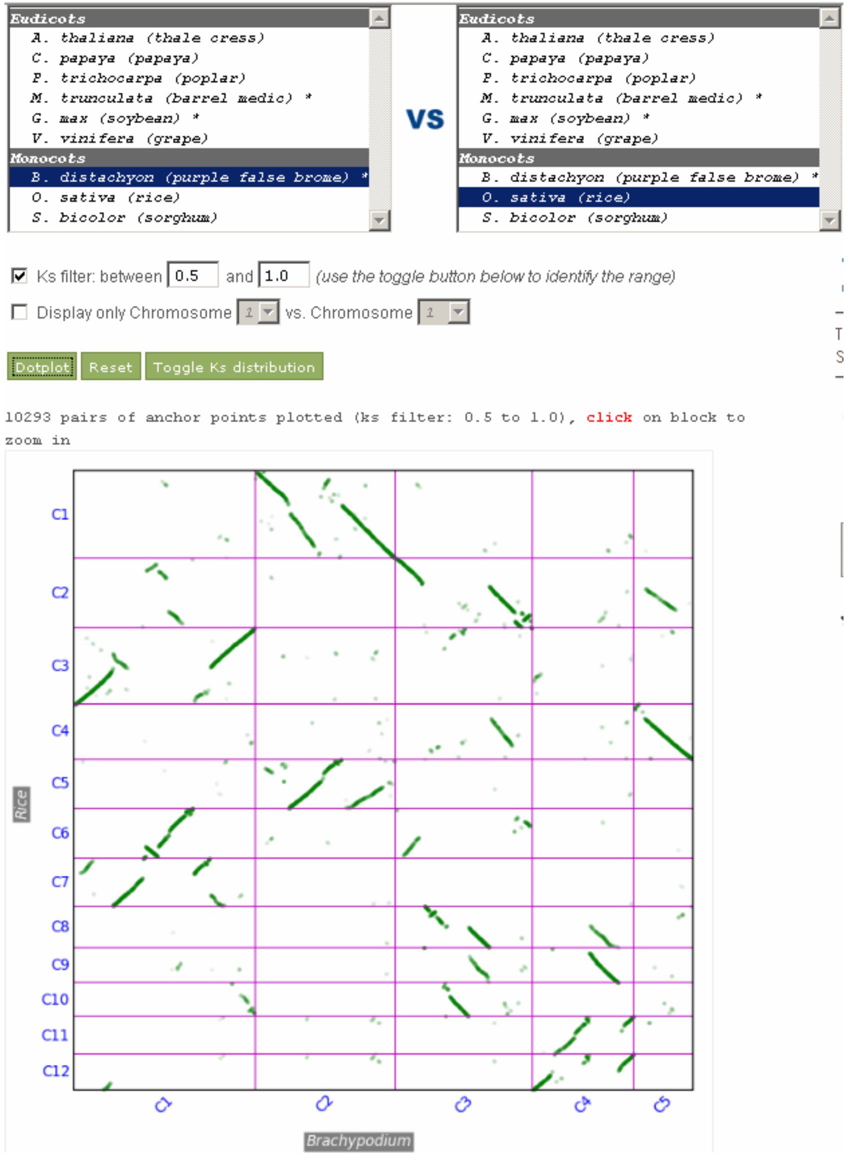
duplication of chr 3 and chr 2



[1] Block (Score 7014.0, E-value 3e-80) with 149 anchors.

Order within Block	Locus 1	Annotation 1	Locus 2	Annotation 2	Ka	Ks	Coordina
1	At2g42180	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G57950.1); similar to Os09g0279200 [Oryza sativa (japonica cultivar-group)] (GB:NP_001062757.1); similar to hypothetical protein OsJ_027512 [Oryza sativa (japonica cultivar-group)] (GB:EAZ44029.1)	At3g57950	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G42180.1); similar to hypothetical protein [Vitis vinifera] (GB:CAN83225.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO23380.1)	0.26	0.84	c2:3591-c3:4776
2	At2g42370	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G58110.1); similar to hypothetical protein [Vitis vinifera] (GB:CAN63361.1)	At3g58110	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G42370.1); similar to hypothetical protein [Vitis vinifera] (GB:CAN81663.1)	0.35	0.89	c2:3610-c3:4791
3	At2g42380	bZIP transcription factor family protein	At3g58120	bZIP transcription factor family protein	0.14	0.76	c2:3611-c3:4792
4	At2g42430	LBD16 (ASYMMETRIC LEAVES2-LIKE18)	At3g58190	ASL16/LBD29 (LOB DOMAIN-CONTAINING PROTEIN 29)	0.51	4.65	c2:3616-c3:4799
5	At2g42470	meprin and TRAF homology domain-containing protein / MATH domain-containing protein	At3g58230	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G58320.1); contains domain PTHR10420 (PTHR10420); contains domain PTHR10420:SF29 (PTHR10420:SF29)	0.49	1.38	c2:3620-c3:4803
6	At2g42480	meprin and TRAF homology domain-containing protein / MATH domain-containing protein	At3g58250	meprin and TRAF homology domain-containing protein / MATH domain-containing protein	0.67	1.41	c2:3621-c3:4805
7	At2g42500	PP2A-4 (protein phosphatase 2A-4); protein serine/threonine phosphatase	At3g58500	PP2A-3 (PROTEIN PHOSPHATASE 2A-3); protein serine/threonine phosphatase	0.01	0.47	c2:3623-c3:4830

Cross-genome Dotblot analysis at PGDD



Eudicots
- A. thaliana (thale cress)
- C. papaya (papaya)
- P. trichocarpa (poplar)
- M. trunculata (barrel medic) *
- G. max (soybean) *
- V. vinifera (grape)

Monocots
- B. distachyon (purple false brome) *
- O. sativa (rice)
- S. bicolor (sorghum)

VS

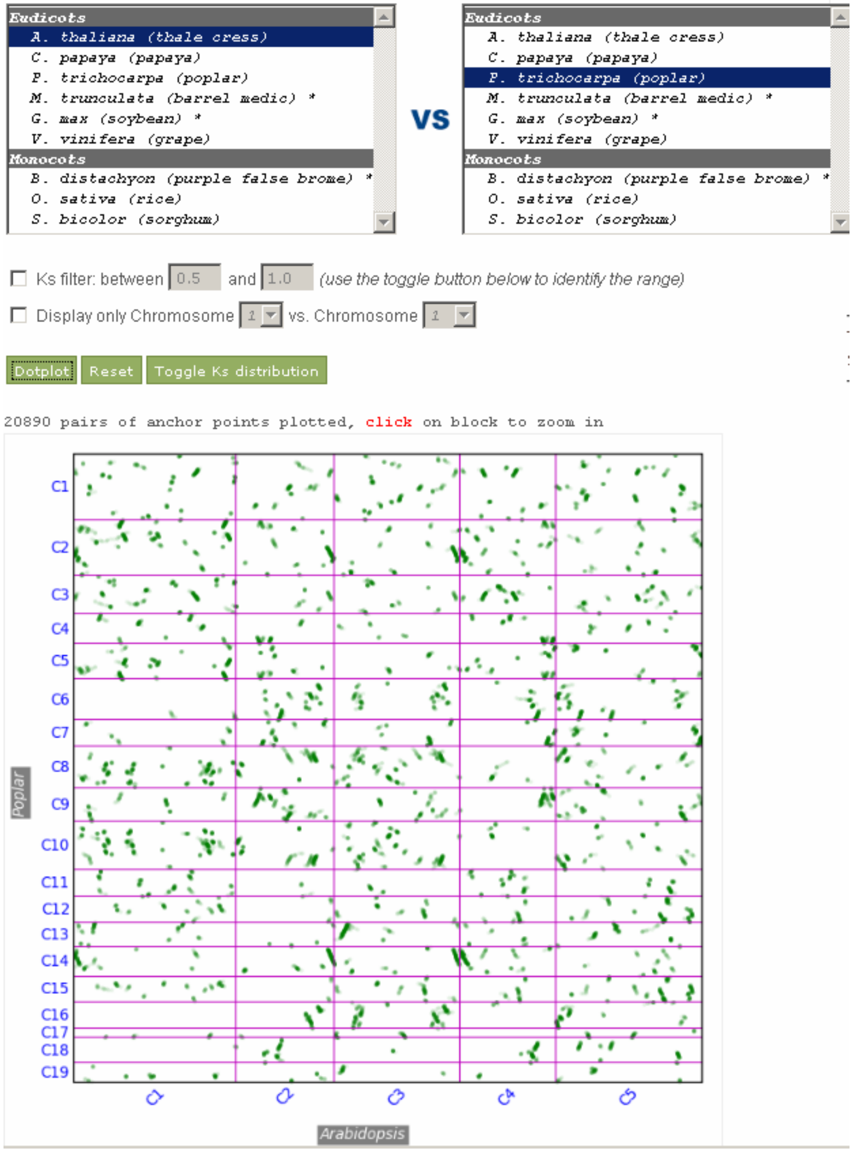
Eudicots
- A. thaliana (thale cress)
- C. papaya (papaya)
- P. trichocarpa (poplar)
- M. trunculata (barrel medic) *
- G. max (soybean) *
- V. vinifera (grape)

Monocots
- B. distachyon (purple false brome) *
- O. sativa (rice)
- S. bicolor (sorghum)

☑ Ks filter: between 0.5 and 1.0 (use the toggle button below to identify the range)

□ Display only Chromosome 1 vs. Chromosome 1

10293 pairs of anchor points plotted (ks filter: 0.5 to 1.0), click on block to zoom in



Eudicots
- A. thaliana (thale cress)
- C. papaya (papaya)
- P. trichocarpa (poplar)
- M. trunculata (barrel medic) *
- G. max (soybean) *
- V. vinifera (grape)

Monocots
- B. distachyon (purple false brome) *
- O. sativa (rice)
- S. bicolor (sorghum)

VS

Eudicots
- A. thaliana (thale cress)
- C. papaya (papaya)
- P. trichocarpa (poplar)
- M. trunculata (barrel medic) *
- G. max (soybean) *
- V. vinifera (grape)

Monocots
- B. distachyon (purple false brome) *
- O. sativa (rice)
- S. bicolor (sorghum)

□ Ks filter: between 0.5 and 1.0 (use the toggle button below to identify the range)

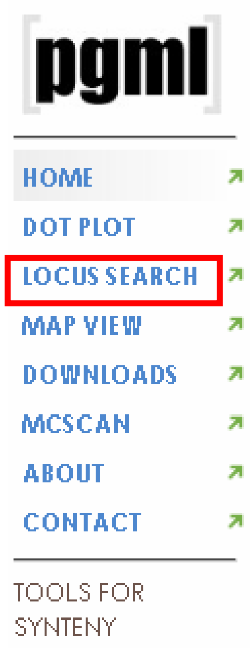
□ Display only Chromosome 1 vs. Chromosome 1

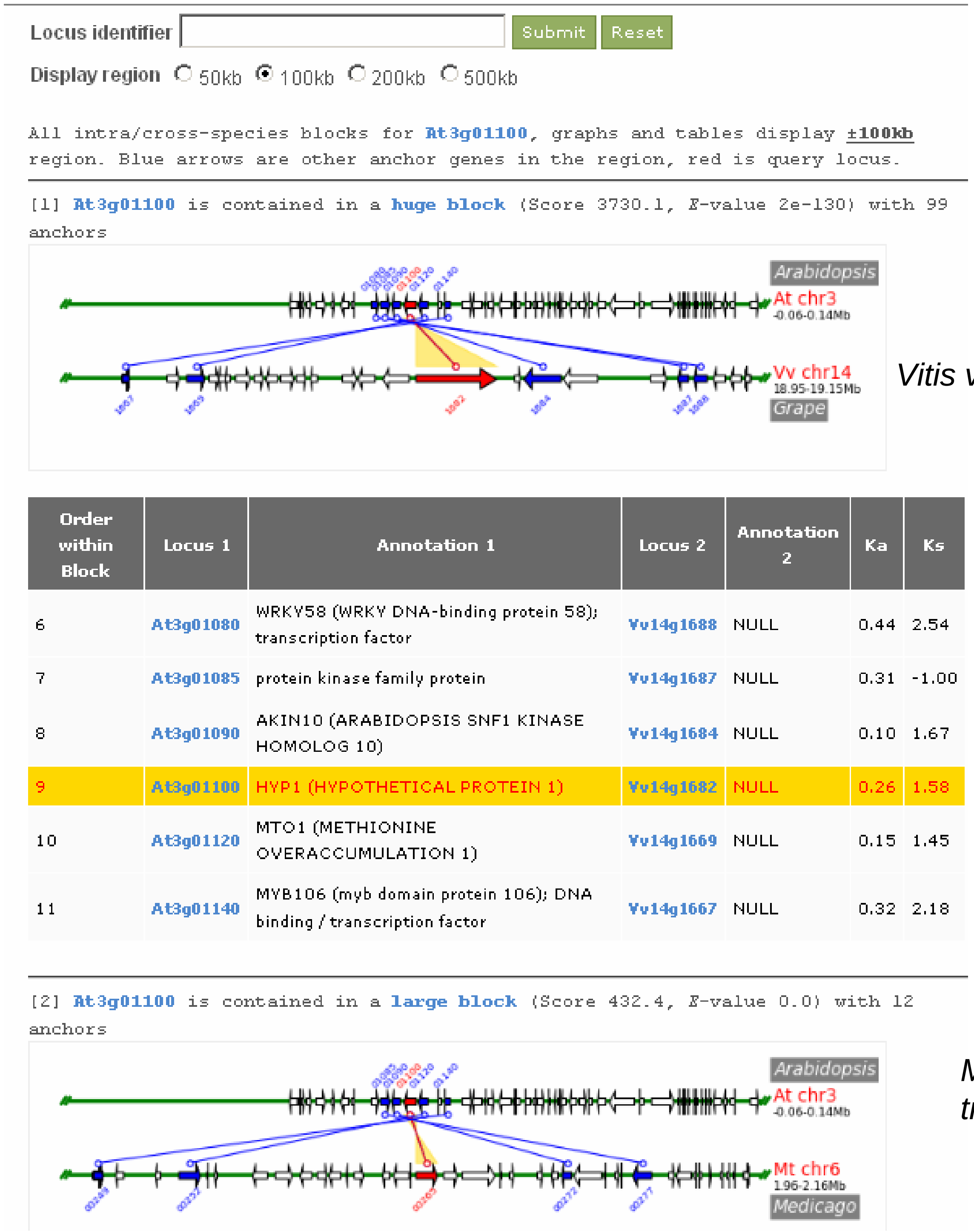
20890 pairs of anchor points plotted, click on block to zoom in

28.7.2009

PGDD

microsynteny





Locus identifier [ ] Submit Reset

Display region ○ 50kb ● 100kb ○ 200kb ○ 500kb

All intra/cross-species blocks for At3g01100, graphs and tables display ±100kb region. Blue arrows are other anchor genes in the region, red is query locus.

[1] At3g01100 is contained in a huge block (Score 3730.1, *E*-value 2e-130) with 99 anchors

Order within Block	Locus 1	Annotation 1	Locus 2	Annotation 2	Ka	Ks
6	At3g01080	WRKY58 (WRKY DNA-binding protein 58); transcription factor	Vv14g1688	NULL	0.44	2.54
7	At3g01085	protein kinase family protein	Vv14g1687	NULL	0.31	-1.00
8	At3g01090	AKIN10 (ARABIDOPSIS SNF1 KINASE HOMOLOG 10)	Vv14g1684	NULL	0.10	1.67
9	At3g01100	HYP1 (HYPOTHETICAL PROTEIN 1)	Vv14g1682	NULL	0.26	1.58
10	At3g01120	MTO1 (METHIONINE OVERACCUMULATION 1)	Vv14g1669	NULL	0.15	1.45
11	At3g01140	MYB106 (myb domain protein 106); DNA binding / transcription factor	Vv14g1667	NULL	0.32	2.18

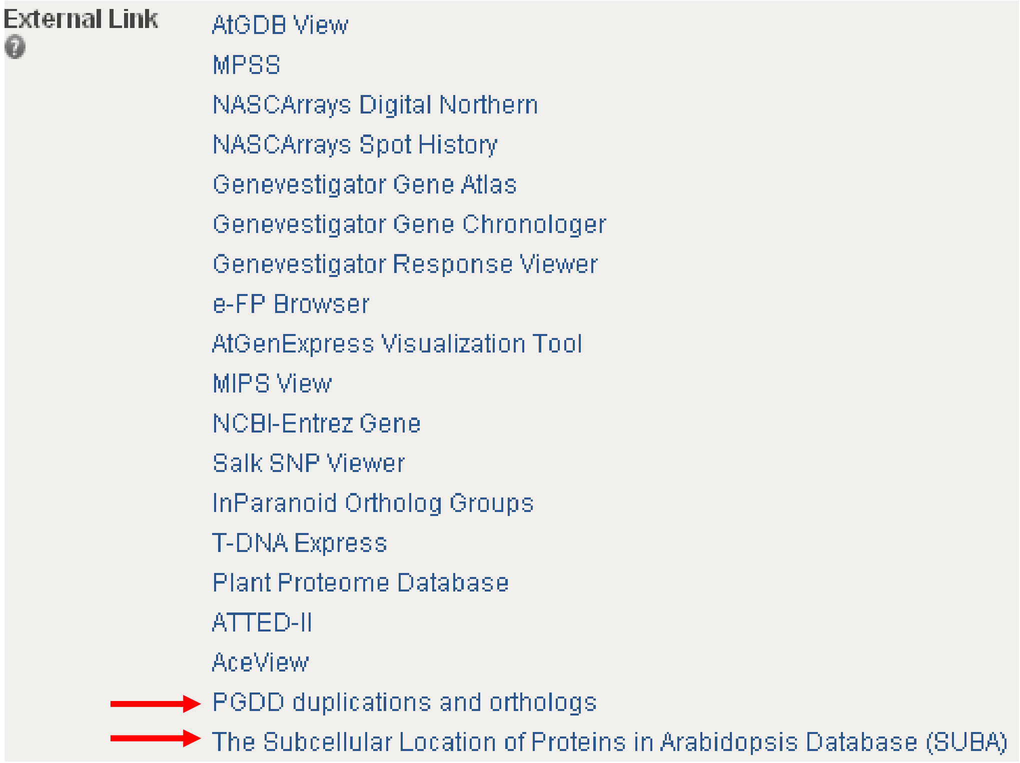
[2] At3g01100 is contained in a large block (Score 432.4, *E*-value 0.0) with 12 anchors

Vitis vinifera

Medicago trunculata

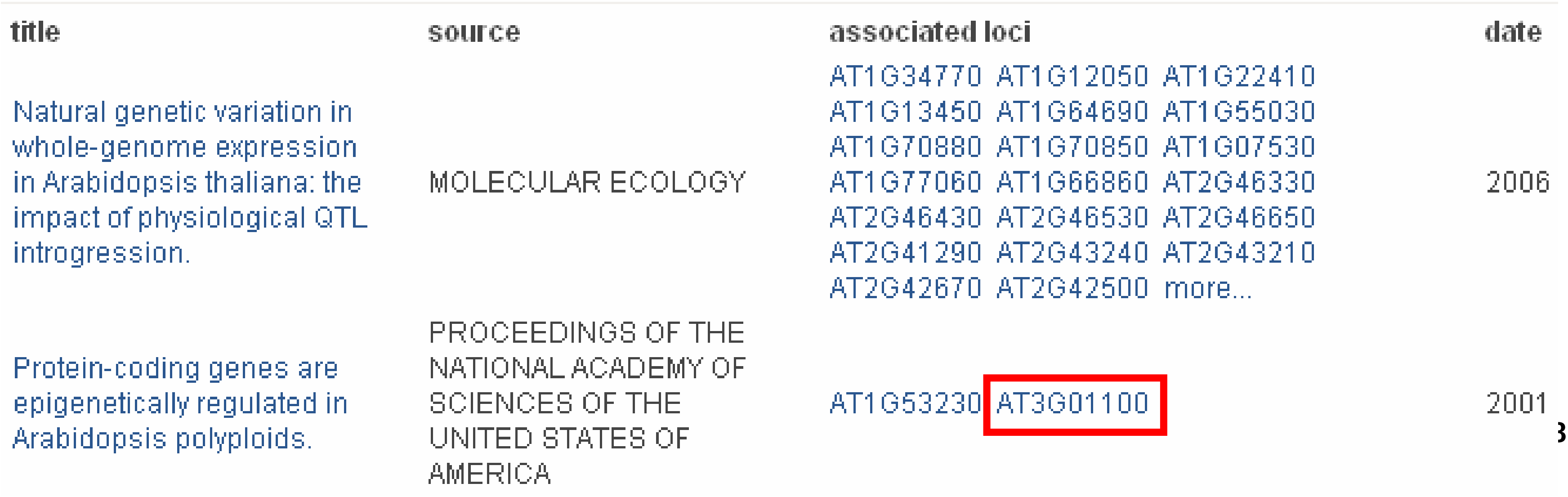
28.7.2009

Back to TAIR



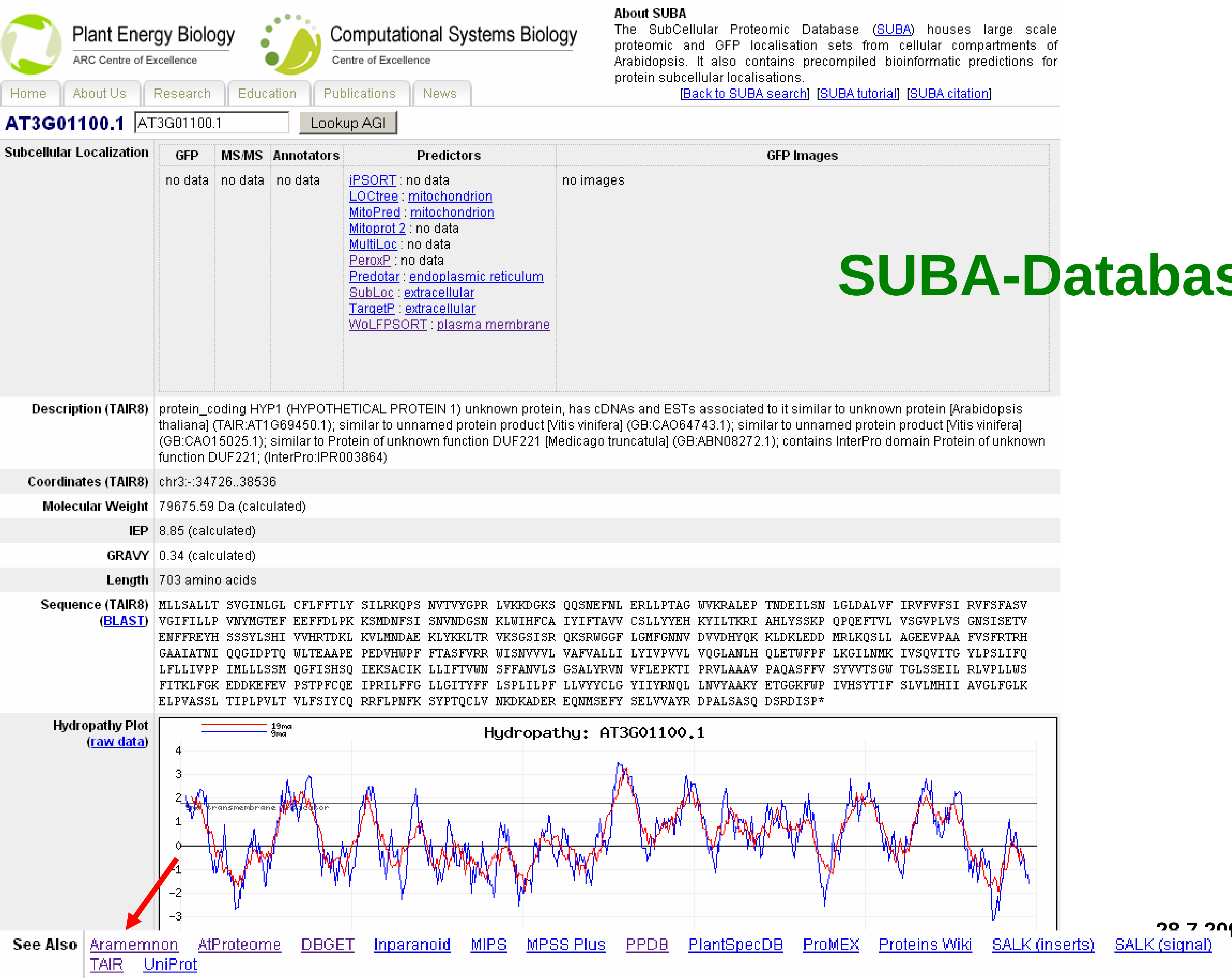
External Link ?

- AtGDB View
- MPSS
- NASCArrays Digital Northern
- NASCArrays Spot History
- Genevestigator Gene Atlas
- Genevestigator Gene Chronologer
- Genevestigator Response Viewer
- e-FP Browser
- AtGenExpress Visualization Tool
- MIPS View
- NCBI-Entrez Gene
- Salk SNP Viewer
- InParanoid Ortholog Groups
- T-DNA Express
- Plant Proteome Database
- ATTED-II
- AceView
- → PGDD duplications and orthologs
- → The Subcellular Location of Proteins in Arabidopsis Database (SUBA)



title	source	associated loci	date
Natural genetic variation in whole-genome expression in Arabidopsis thaliana: the impact of physiological QTL introgression.	MOLECULAR ECOLOGY	AT1G34770 AT1G12050 AT1G22410 AT1G13450 AT1G64690 AT1G55030 AT1G70880 AT1G70850 AT1G07530 AT1G77060 AT1G66860 AT2G46330 AT2G46430 AT2G46530 AT2G46650 AT2G41290 AT2G43240 AT2G43210 AT2G42670 AT2G42500 more...	2006
Protein-coding genes are epigenetically regulated in Arabidopsis polyploids.	PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA	AT1G53230 AT3G01100	2001

SUBA-Database



About SUBA

The SubCellular Proteomic Database (SUBA) houses large scale proteomic and GFP localisation sets from cellular compartments of Arabidopsis. It also contains precompiled bioinformatic predictions for protein subcellular localisations.

[Back to SUBA search] [SUBA tutorial] [SUBA citation]

AT3G01100.1 AT3G01100.1

Subcellular Localization

GFP	MS/MS	Annotators	Predictors	GFP Images
no data	no data	no data	iPSORT : no data LOCtree : mitochondrion MitoPred : mitochondrion Mitoprot 2 : no data MultiLoc : no data PeroxP : no data Predotar : endoplasmic reticulum SubLoc : extracellular TargetP : extracellular WoLFPSORT : plasma membrane	no images

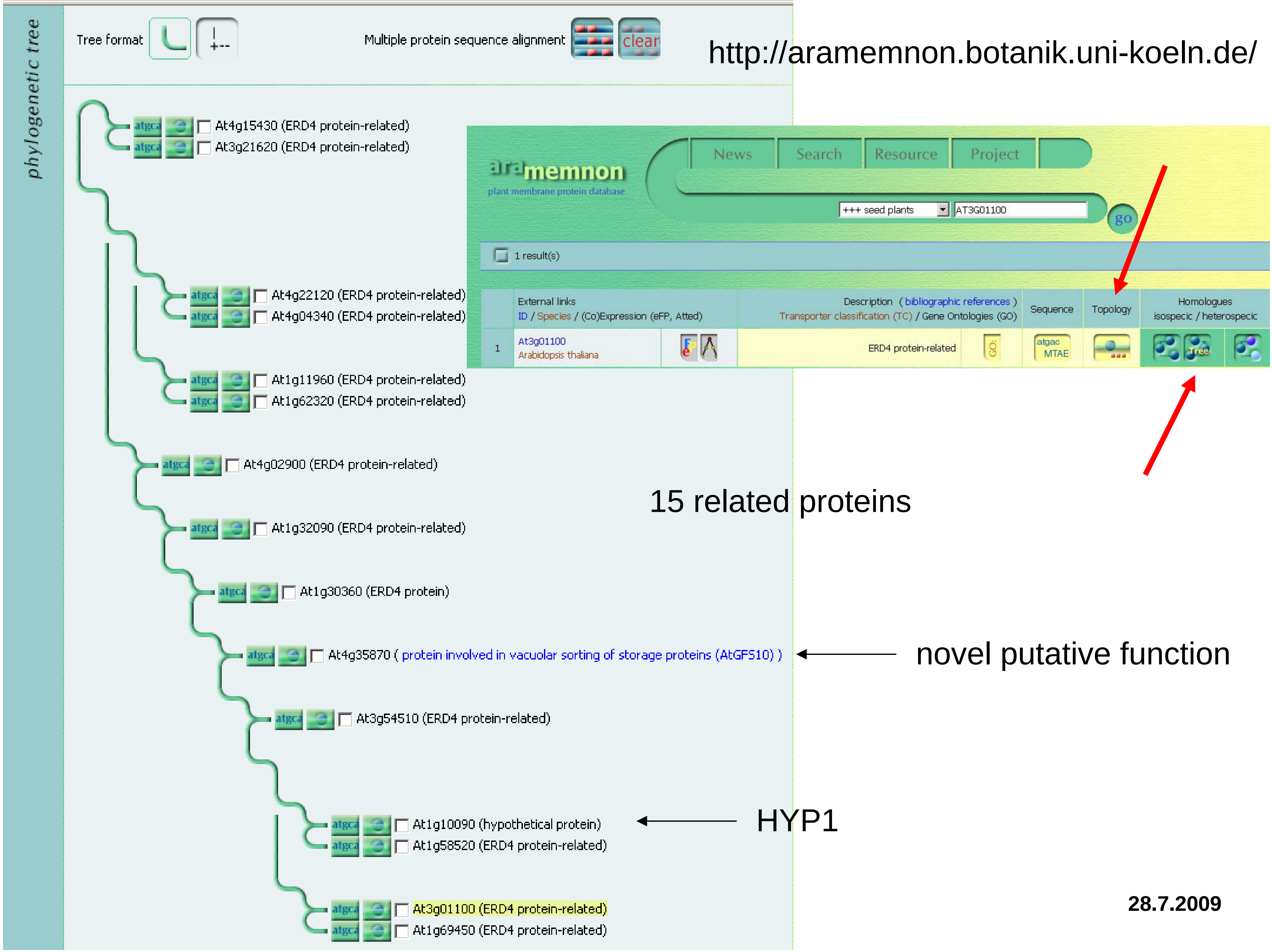
Description (TAIR8)	protein_coding HYP1 (HYPOTHETICAL PROTEIN 1) unknown protein, has cDNAs and ESTs associated to it similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G69450.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO64743.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO15025.1); similar to Protein of unknown function DUF221 [Medicago truncatula] (GB:ABN08272.1); contains InterPro domain Protein of unknown function DUF221; (InterPro:IPR003864)
Coordinates (TAIR8)	chr3:-:34726..38536
Molecular Weight	79675.59 Da (calculated)
IEP	8.85 (calculated)
GRAVY	0.34 (calculated)
Length	703 amino acids

Sequence (TAIR8) (BLAST)

```
MLLSALLT SVGINLGL CFLFFTLY SILRKQPS NVTVYGPR LVKKDGKS QQSNEFNL ERLLPTAG WVKRALEP TNDEILSN LGLDALVF IRVFVFSI RVFSFASV
VGIFILLP VNYMGTEF EEFFDLPK KSMDNFSI SNVNDGSN KLWIHFCA IYIFTAVV CSLLYYEH KYILTKRI AHLYSSKP QPQEFTVL VSGVPLVS GNSISETV
ENFFREYH SSSYLSHI VVHRTDKL KVLMNDAE KLYKKLTR VKSGSISR QKSRWGGF LGMFGNNV DVVDHYQK KLDKLEDD MRLKQSLL AGEEVPAA FVSFRTRH
GAAIATNI QQGIDPTQ WLTEAAPE PEDVHWPF FTASFVRR WISNVVVL VAFVALLI LYIVPVVL VQGLANLH QLETWFPF LKGILNMK IVSQVITG YLPSLIFQ
LFLLIVPP IMLLLSSM QGFISHSQ IEKSACIK LLIFTVWN SFFANVLS GSALYRVN VFLEPKTI PRVLAAAV PAQASFFV SYVVTSGW TGLSSEIL RLVPLLWS
FITKLFGK EDDKEFEV PSTPFCQE IPRILFFG LLGITYFF LSPLILPF LLVYYCLG YIIYRNQL LNVYAAKY ETGGKFWP IVHSYTIF SLVLMHII AVGLFGLK
ELPVASSL TIPLPVLT VLFSIYCQ RRFLPNFK SYPTQCLV NKDKADER EQNMSEFY SELVVAYR DPALSASQ DSRDISP*
```

Hydropathy Plot (raw data)

See Also Aramemnon AtProteome DBGET Inparanoid MIPS MPSS Plus PPDB PlantSpecDB ProMEX Proteins Wiki SALK (inserts) SALK (signal) TAIR UniProt http://aramemnon.botanik.uni-koeln.de/



□ At4g15430 (ERD4 protein-related)
□ At3g21620 (ERD4 protein-related)
□ At4g22120 (ERD4 protein-related)
□ At4g04340 (ERD4 protein-related)
□ At1g11960 (ERD4 protein-related)
□ At1g62320 (ERD4 protein-related)
□ At4g02900 (ERD4 protein-related)
□ At1g32090 (ERD4 protein-related)
□ At1g30360 (ERD4 protein)
□ At4g35870 (protein involved in vacuolar sorting of storage proteins (AtGFS10))
□ At3g54510 (ERD4 protein-related)
□ At1g10090 (hypothetical protein)
□ At1g58520 (ERD4 protein-related)
□ At3g01100 (ERD4 protein-related)
□ At1g69450 (ERD4 protein-related)

□ 1 result(s)

	External links ID / Species / (Co)Expression (eFP, Atted)		Description (bibliographic references) Transporter classification (TC) / Gene Ontologies (GO)		Sequence	Topology	Homologues isospecic / heterospecic
1	At3g01100 Arabidopsis thaliana	eFP	ERD4 protein-related	GO:	atgac MTAE		Tree

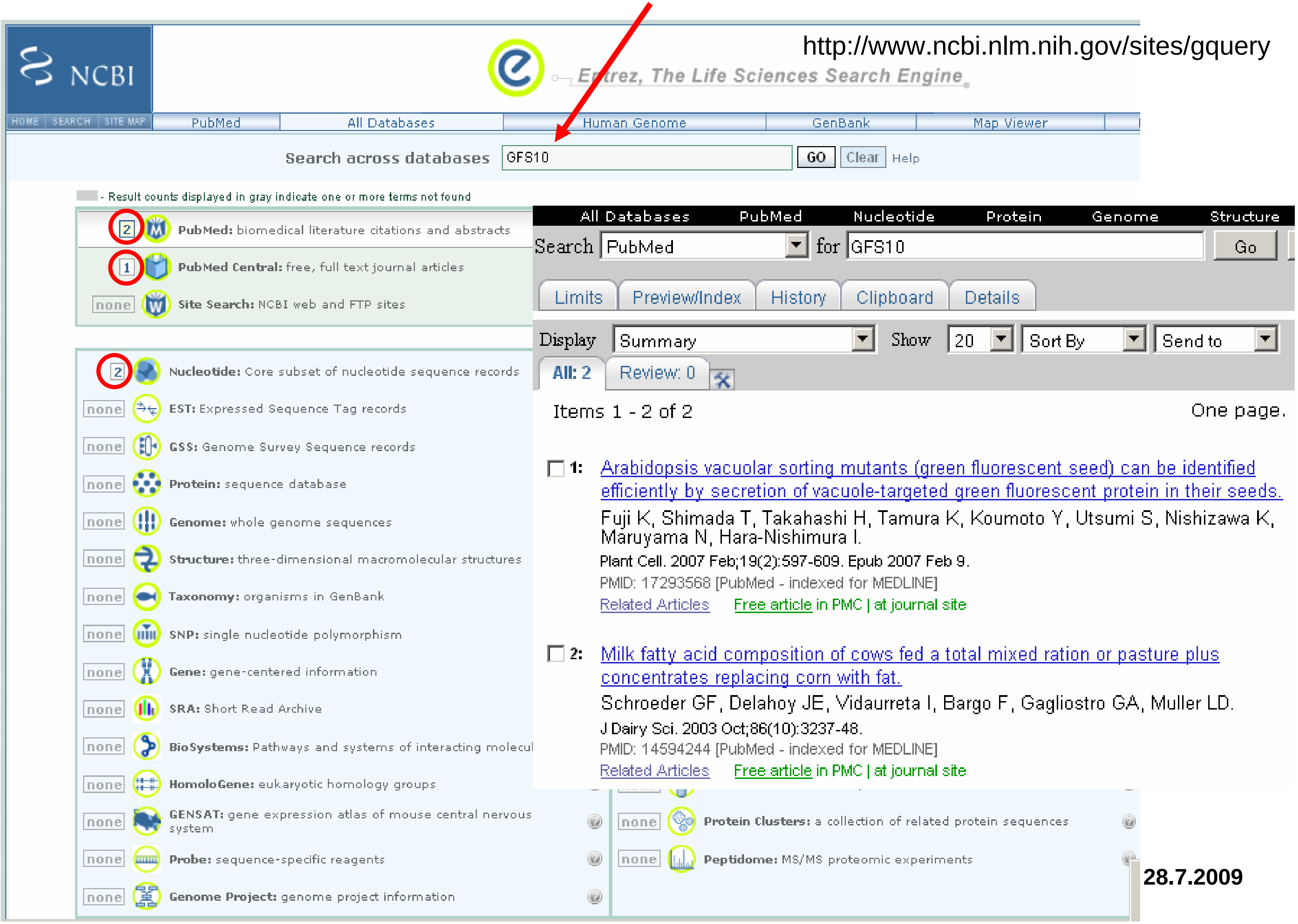
15 related proteins

novel putative function

HYP1

28.7.2009

What is AtGFS10?



http://www.ncbi.nlm.nih.gov/sites/gquery

Entrez, The Life Sciences Search Engine

Search across databases GFS10 GO Clear Help

- Result counts displayed in gray indicate one or more terms not found

2	**PubMed:** biomedical literature citations and abstracts
1	**PubMed Central:** free, full text journal articles
none	**Site Search:** NCBI web and FTP sites

2	**Nucleotide:** Core subset of nucleotide sequence records
none	**EST:** Expressed Sequence Tag records
none	**GSS:** Genome Survey Sequence records
none	**Protein:** sequence database
none	**Genome:** whole genome sequences
none	**Structure:** three-dimensional macromolecular structures
none	**Taxonomy:** organisms in GenBank
none	**SNP:** single nucleotide polymorphism
none	**Gene:** gene-centered information
none	**SRA:** Short Read Archive
none	**BioSystems:** Pathways and systems of interacting molecul
none	**HomoloGene:** eukaryotic homology groups
none	**GENSAT:** gene expression atlas of mouse central nervous system
none	**Probe:** sequence-specific reagents
none	**Genome Project:** genome project information

none	**Protein Clusters:** a collection of related protein sequences
none	**Peptidome:** MS/MS proteomic experiments

Search PubMed for GFS10 Go

Display Summary Show 20 Sort By Send to

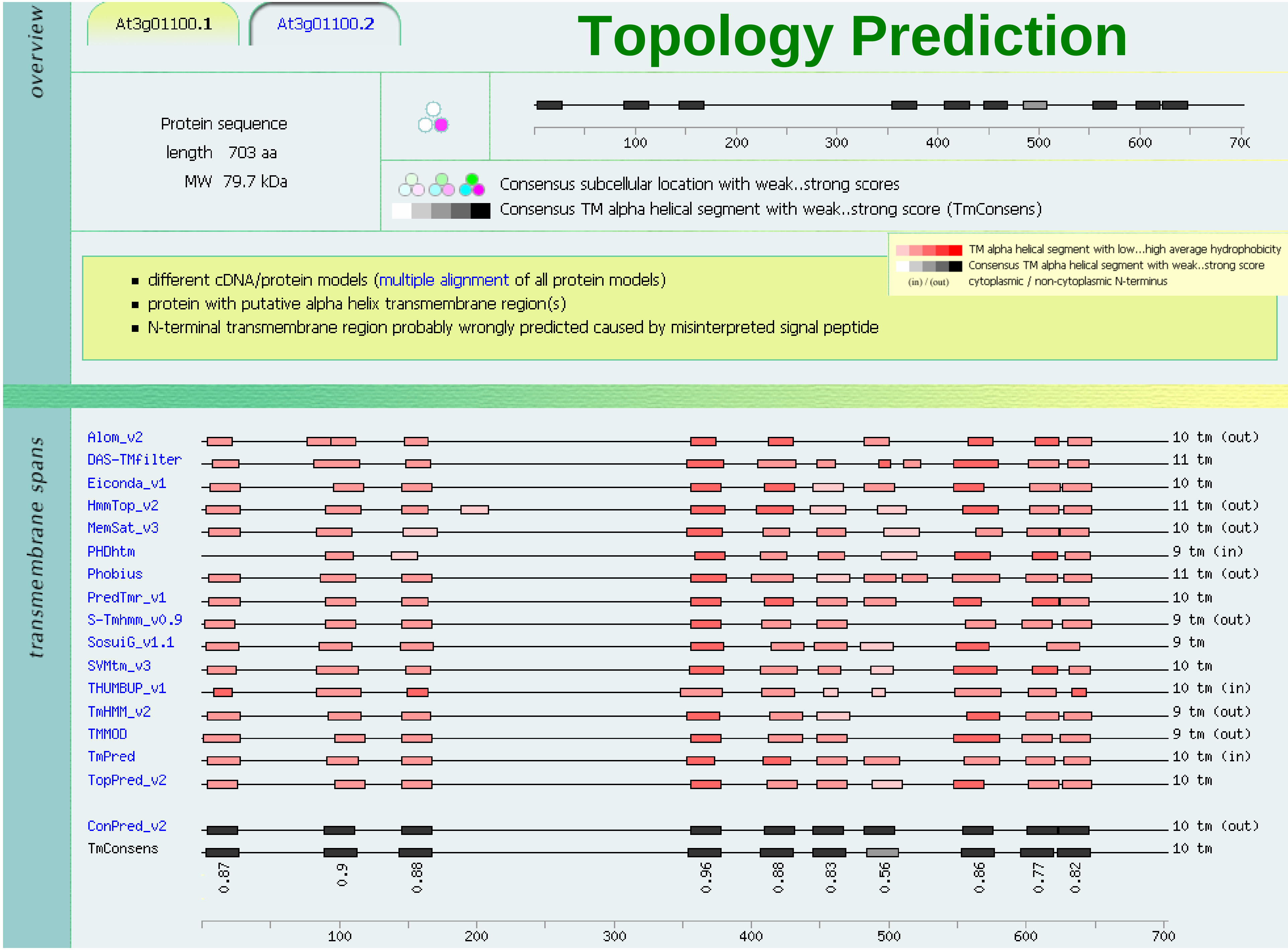
All: 2 Review: 0

Items 1 - 2 of 2 One page.

□ 1: Arabidopsis vacuolar sorting mutants (green fluorescent seed) can be identified efficiently by secretion of vacuole-targeted green fluorescent protein in their seeds.
Fuji K, Shimada T, Takahashi H, Tamura K, Koumoto Y, Utsumi S, Nishizawa K, Maruyama N, Hara-Nishimura I.
Plant Cell. 2007 Feb;19(2):597-609. Epub 2007 Feb 9.
PMID: 17293568 [PubMed - indexed for MEDLINE]
Related Articles Free article in PMC | at journal site

□ 2: Milk fatty acid composition of cows fed a total mixed ration or pasture plus concentrates replacing corn with fat.
Schroeder GF, Delahoy JE, Vidaurreta I, Bargo F, Gagliostro GA, Muller LD.
J Dairy Sci. 2003 Oct;86(10):3237-48.
PMID: 14594244 [PubMed - indexed for MEDLINE]
Related Articles Free article in PMC | at journal site

28.7.2009



Topology Prediction

Protein sequence
length 703 aa
MW 79.7 kDa

Consensus subcellular location with weak..strong scores

Consensus TM alpha helical segment with weak..strong score (TmConsens)

TM alpha helical segment with low...high average hydrophobicity

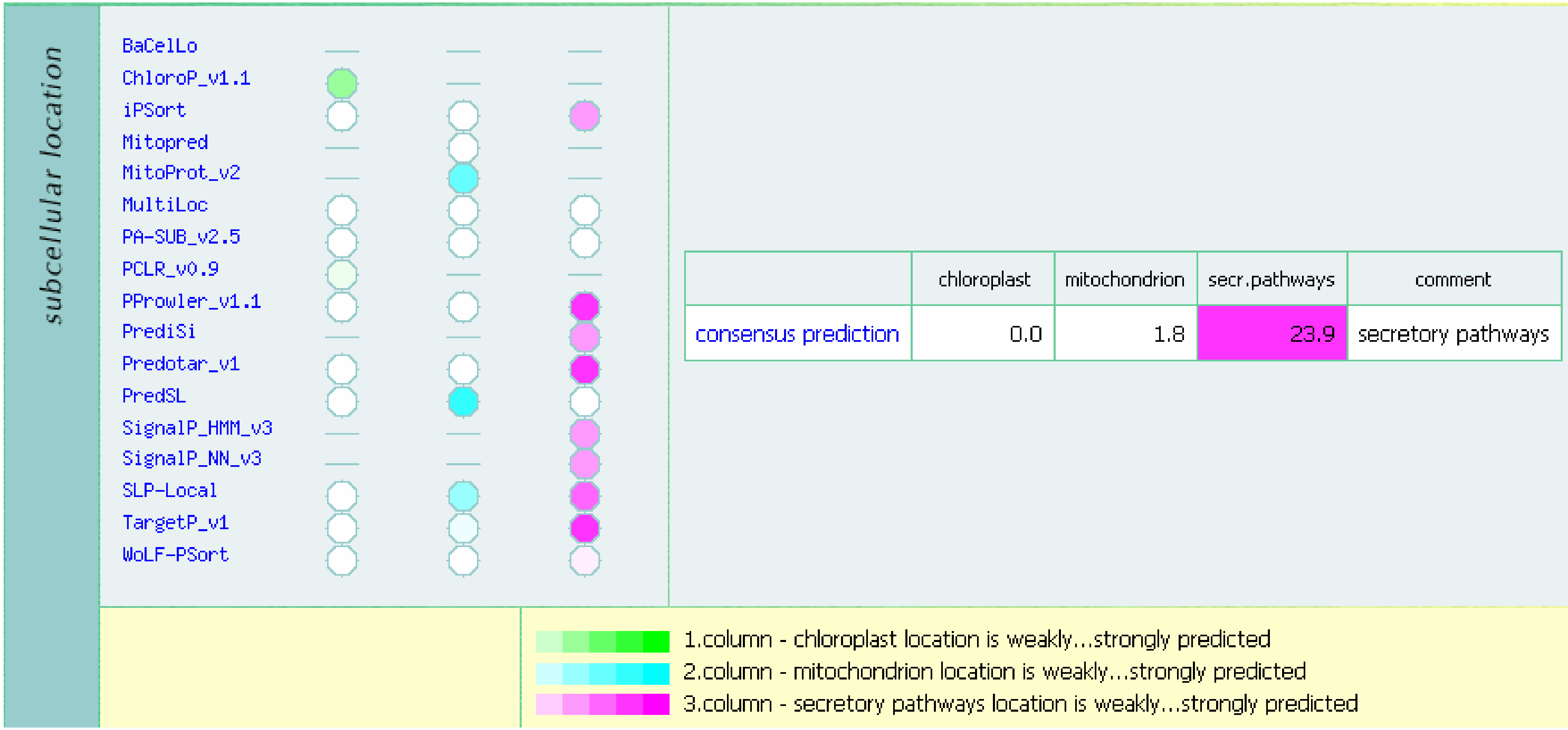
Consensus TM alpha helical segment with weak..strong score

(in) / (out) cytoplasmic / non-cytoplasmic N-terminus

- different cDNA/protein models (multiple alignment of all protein models)
- protein with putative alpha helix transmembrane region(s)
- N-terminal transmembrane region probably wrongly predicted caused by misinterpreted signal peptide

Alom_v2	10 tm (out)
DAS-TMfilter	11 tm
Eiconda_v1	10 tm
HmmTop_v2	11 tm (out)
MemSat_v3	10 tm (out)
PHDhtm	9 tm (in)
Phobius	11 tm (out)
PredTmr_v1	10 tm
S-Tmhmm_v0.9	9 tm (out)
SosuiG_v1.1	9 tm
SVMtm_v3	10 tm
THUMBUP_v1	10 tm (in)
TmHMM_v2	9 tm (out)
TMMOD	9 tm (out)
TmPred	10 tm (in)
TopPred_v2	10 tm
ConPred_v2	10 tm (out)
TmConsens	10 tm

Strongly predicted to be in the secretroy pathway



subcellular location

BaCelLo	—	—	—
ChloroP_v1.1	●	—	—
iPSort	○	○	●
Mitopred	—	○	—
MitoProt_v2	—	●	—
MultiLoc	○	○	○
PA-SUB_v2.5	○	○	○
PCLR_v0.9	●	—	—
PProwler_v1.1	○	○	●
PrediSi	—	—	●
Predotar_v1	○	○	●
PredSL	○	●	○
SignalP_HMM_v3	—	—	●
SignalP_NN_v3	—	—	●
SLP-Local	○	●	●
TargetP_v1	○	●	●
WoLF-PSort	○	○	●

	chloroplast	mitochondrion	secr.pathways	comment
consensus prediction	0.0	1.8	23.9	secretory pathways

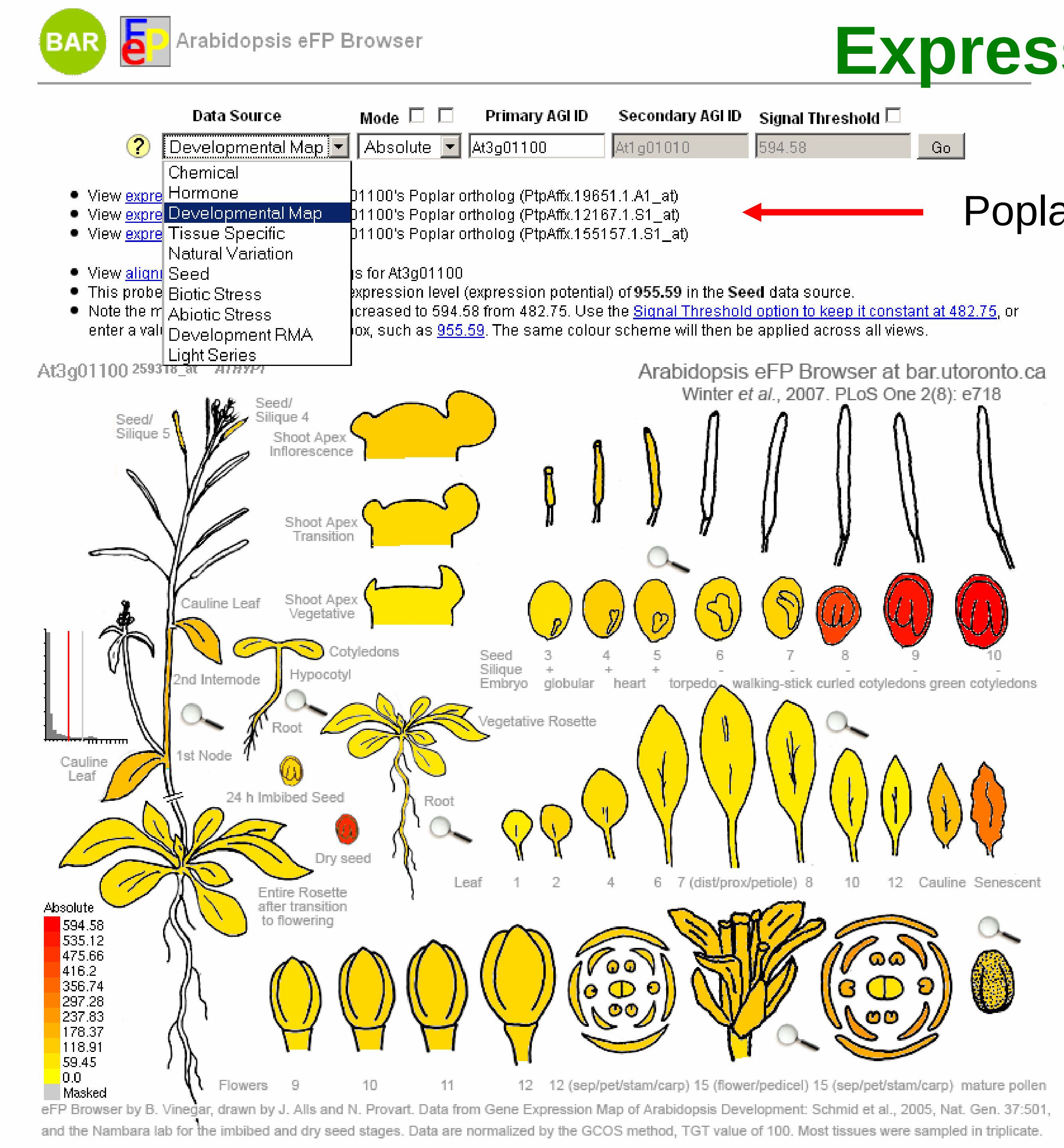
1.column - chloroplast location is weakly...strongly predicted

2.column - mitochondrion location is weakly...strongly predicted

3.column - secretory pathways location is weakly...strongly predicted

28.7.2009

Expression Analyses



Data Source: Developmental Map
Mode ☐ ☐: Absolute
Primary AGI ID: At3g01100
Secondary AGI ID: At1g01010
Signal Threshold ☐: 594.58

- View expre 01100's Poplar ortholog (PtpAffx.19651.1.A1_at)
- View expre 01100's Poplar ortholog (PtpAffx.12167.1.S1_at)
- View expre 01100's Poplar ortholog (PtpAffx.155157.1.S1_at)

- View aligni s for At3g01100
- This probe xpression level (expression potential) of **955.59** in the **Seed** data source.
- Note the m creased to 594.58 from 482.75. Use the Signal Threshold option to keep it constant at 482.75, or enter a valu ox, such as 955.59. The same colour scheme will then be applied across all views.

At3g01100 259318_at AtHYP1

Arabidopsis eFP Browser at bar.utoronto.ca
Winter *et al*., 2007. PLoS One 2(8): e718

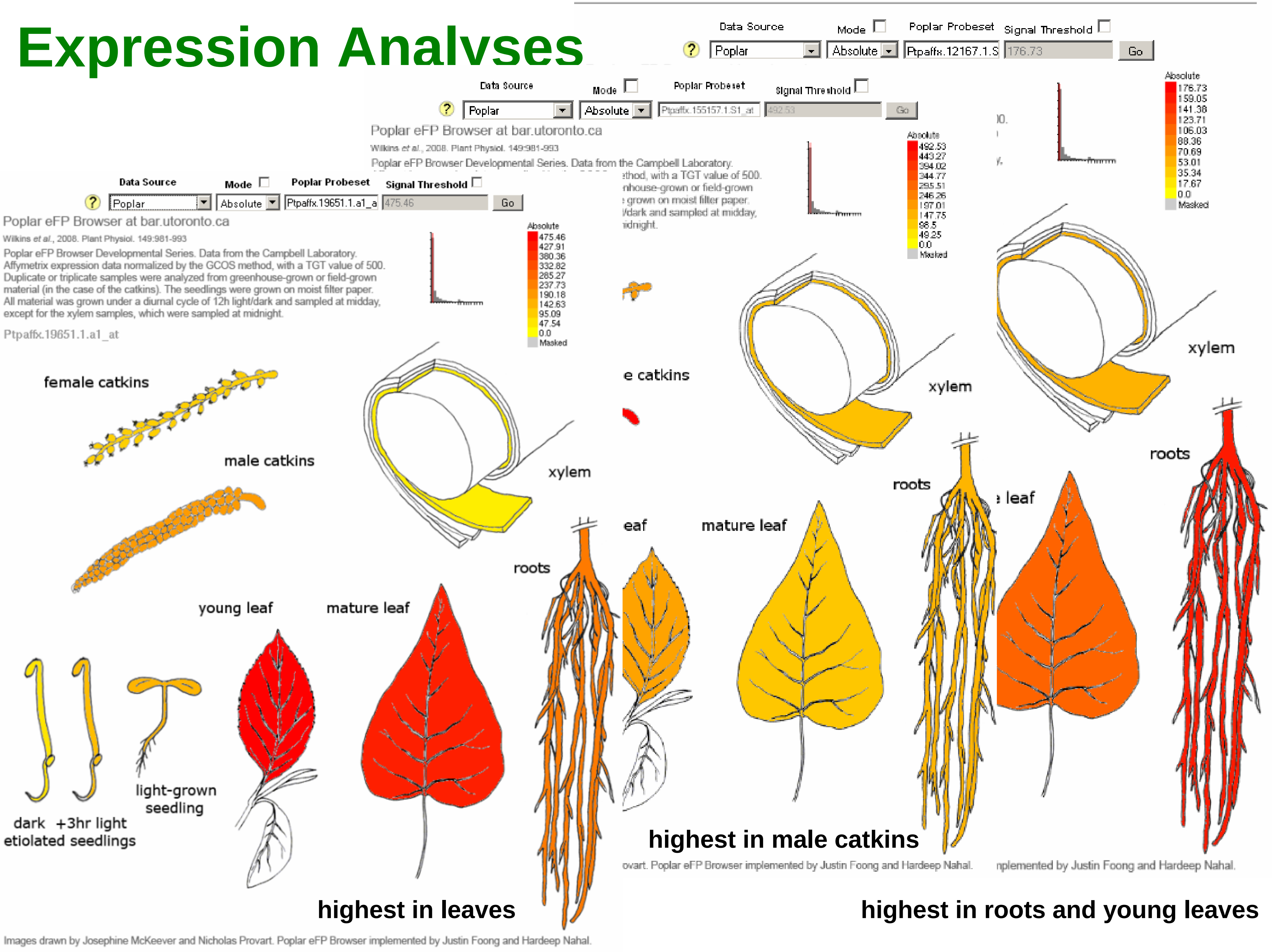
eFP Browser by B. Vinegar, drawn by J. Alls and N. Provart. Data from Gene Expression Map of Arabidopsis Development: Schmid et al., 2005, Nat. Gen. 37:501, and the Nambara lab for the imbibed and dry seed stages. Data are normalized by the GCOS method, TGT value of 100. Most tissues were sampled in triplicate.

Poplar homologs

http://bbc.botany.utoronto.ca/efp/cgi-bin/efpWeb.cgi

28.7.2009

Expression Analvses



Data Source | Mode □ | Poplar Probeset | Signal Threshold □

? Poplar | Absolute | Ptpaffx.19651.1.a1_a | 475.46 | Go

Poplar eFP Browser at bar.utoronto.ca

Wilkins *et al.*, 2008. Plant Physiol. 149:981-993

Poplar eFP Browser Developmental Series. Data from the Campbell Laboratory. Affymetrix expression data normalized by the GCOS method, with a TGT value of 500. Duplicate or triplicate samples were analyzed from greenhouse-grown or field-grown material (in the case of the catkins). The seedlings were grown on moist filter paper. All material was grown under a diurnal cycle of 12h light/dark and sampled at midday, except for the xylem samples, which were sampled at midnight.

Ptpaffx.19651.1.a1_at

Images drawn by Josephine McKeever and Nicholas Provart. Poplar eFP Browser implemented by Justin Foong and Hardeep Nahal.

highest in leaves

Data Source | Mode □ | Poplar Probeset | Signal Threshold □

? Poplar | Absolute | Ptpaffx.155157.1.S1_at | 492.53 | Go

Poplar eFP Browser at bar.utoronto.ca

Wilkins *et al.*, 2008. Plant Physiol. 149:981-993

Poplar eFP Browser Developmental Series. Data from the Campbell Laboratory.
ethod, with a TGT value of 500.
nhouse-grown or field-grown
grown on moist filter paper.
/dark and sampled at midday,
idnight.

ovart. Poplar eFP Browser implemented by Justin Foong and Hardeep Nahal.

highest in male catkins

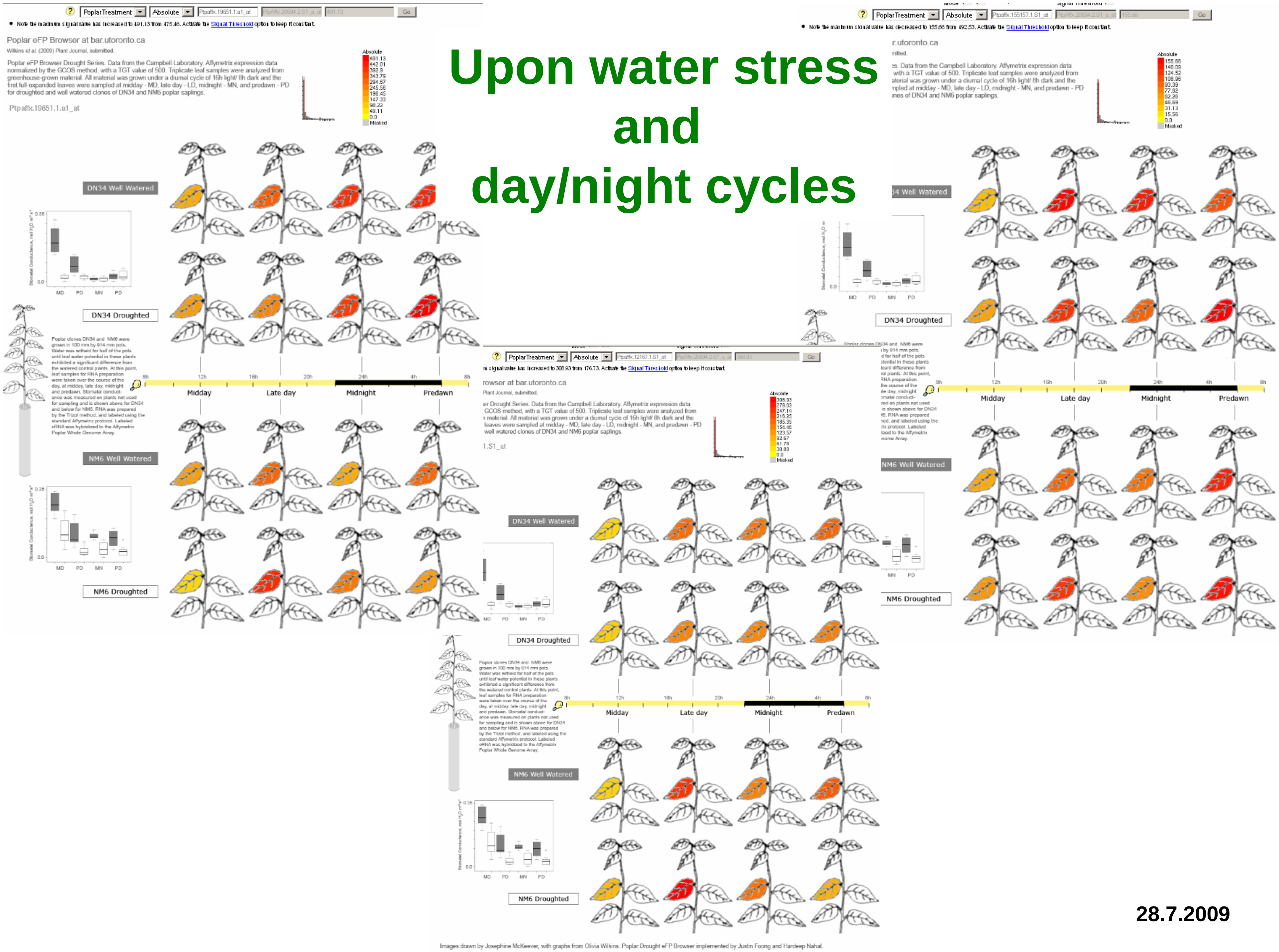
Data Source | Mode □ | Poplar Probeset | Signal Threshold □

? Poplar | Absolute | Ptpaffx.12167.1.S | 176.73 | Go

nplemented by Justin Foong and Hardeep Nahal.

highest in roots and young leaves

Upon water stress
and
day/night cycles



PoplarTreatment | Absolute | Ptpaffx.19651.1.a1_at | PtpAffx.20694.2.S1_a_at | 491.13 | Go

- Note the maximum signal value has increased to 491.13 from 475.46. Activate the Signal Threshold option to keep it constant.

Poplar eFP Browser at bar.utoronto.ca

Wilkins et al. (2009) Plant Journal, submitted.

Poplar eFP Browser Drought Series. Data from the Campbell Laboratory. Affymetrix expression data normalized by the GCOS method, with a TGT value of 500. Triplicate leaf samples were analyzed from greenhouse-grown material. All material was grown under a diurnal cycle of 16h light/ 8h dark and the first full-expanded leaves were sampled at midday - MD, late day - LD, midnight - MN, and predawn - PD for droughted and well watered clones of DN34 and NM6 poplar saplings.

Ptpaffx.19651.1.a1_at

Poplar clones DN34 and NM6 were grown in 100 mm by 914 mm pots. Water was witheld for half of the pots until leaf water potential in these plants exhibited a significant difference from the watered control plants. At this point, leaf samples for RNA preparation were taken over the course of the day, at midday, late day, midnight and predawn. Stomatal conductance was measured on plants not used for sampling and is shown above for DN34 and below for NM6. RNA was prepared by the Trizol method, and labeled using the standard Affymetrix protocol. Labeled cRNA was hybridized to the Affymetrix Poplar Whole Genome Array.

PoplarTreatment | Absolute | Ptpaffx.155157.1.S1_at | PtpAffx.20694.2.S1_a_at | 155.66 | Go

- Note the maximum signal value has decreased to 155.66 from 492.53. Activate the Signal Threshold option to keep it constant.

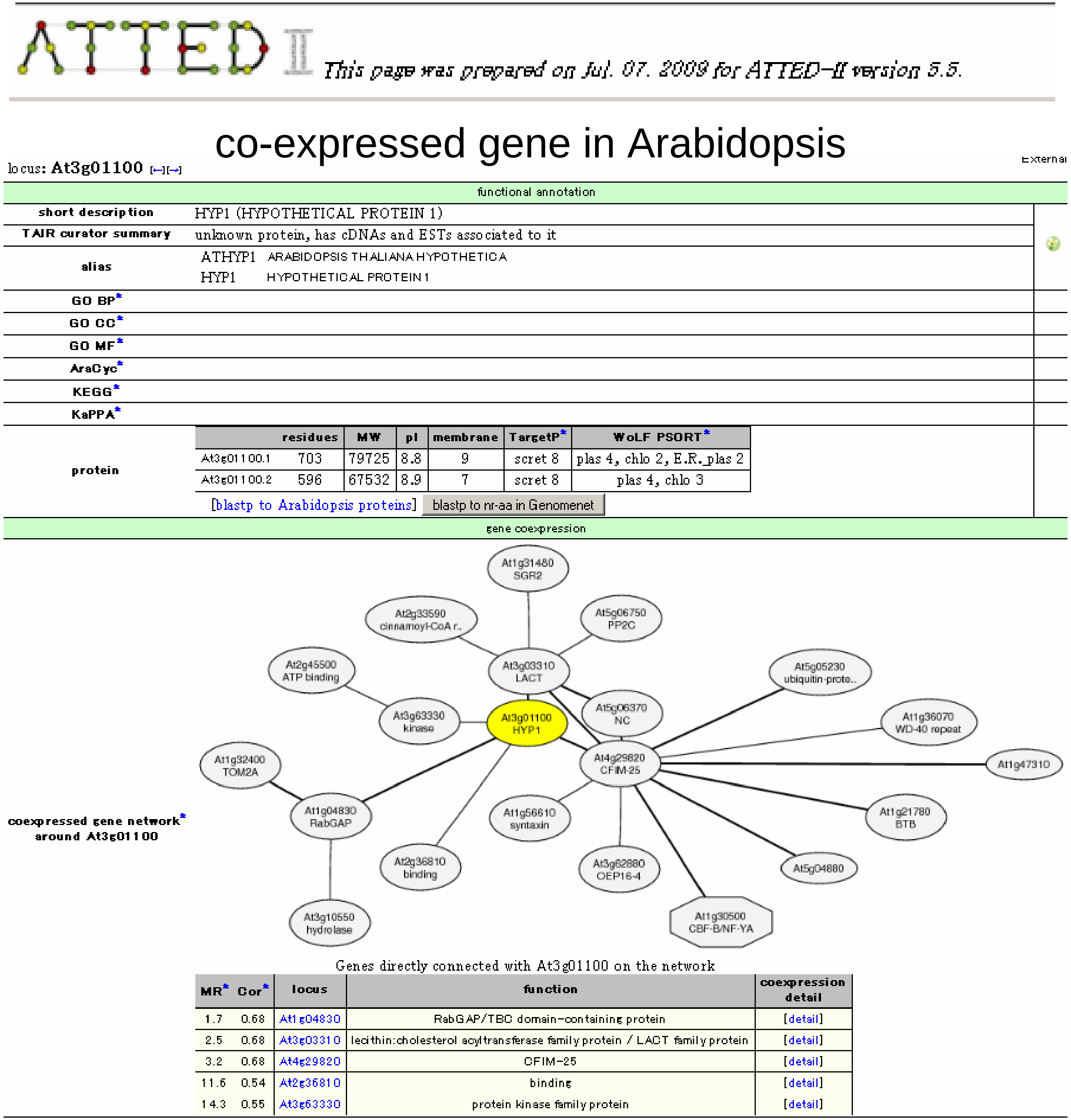
PoplarTreatment | Absolute | Ptpaffx.12167.1.S1_at | PtpAffx.20694.2.S1_a_at | 308.93 | Go

- [...]m signal value has increased to 308.93 from 176.73. Activate the Signal Threshold option to keep it constant.

Images drawn by Josephine McKeever, with graphs from Olivia Wilkins. Poplar Drought eFP Browser implemented by Justin Foong and Hardeep Nahal.

28.7.2009

co-expressed gene in Arabidopsis



ATTED II
This page was prepared on Jul. 07. 2009 for ATTED-II version 5.5.
External

locus: At3g01100

functional annotation

short description	HYP1 (HYPOTHETICAL PROTEIN 1)
TAIR curator summary	unknown protein, has cDNAs and ESTs associated to it
alias	ATHYP1 ARABIDOPSIS THALIANA HYPOTHETICA HYP1 HYPOTHETICAL PROTEIN 1
GO BP*	
GO CC*	
GO MF*	
AraCyc*	
KEGG*	
KaPPA*	

protein

	residues	MW	pI	membrane	TargetP*	WoLF PSORT*
At3g01100.1	703	79725	8.8	9	scret 8	plas 4, chlo 2, E.R._plas 2
At3g01100.2	596	67532	8.9	7	scret 8	plas 4, chlo 3

[blastp to Arabidopsis proteins] blastp to nr-aa in Genomenet

gene coexpression

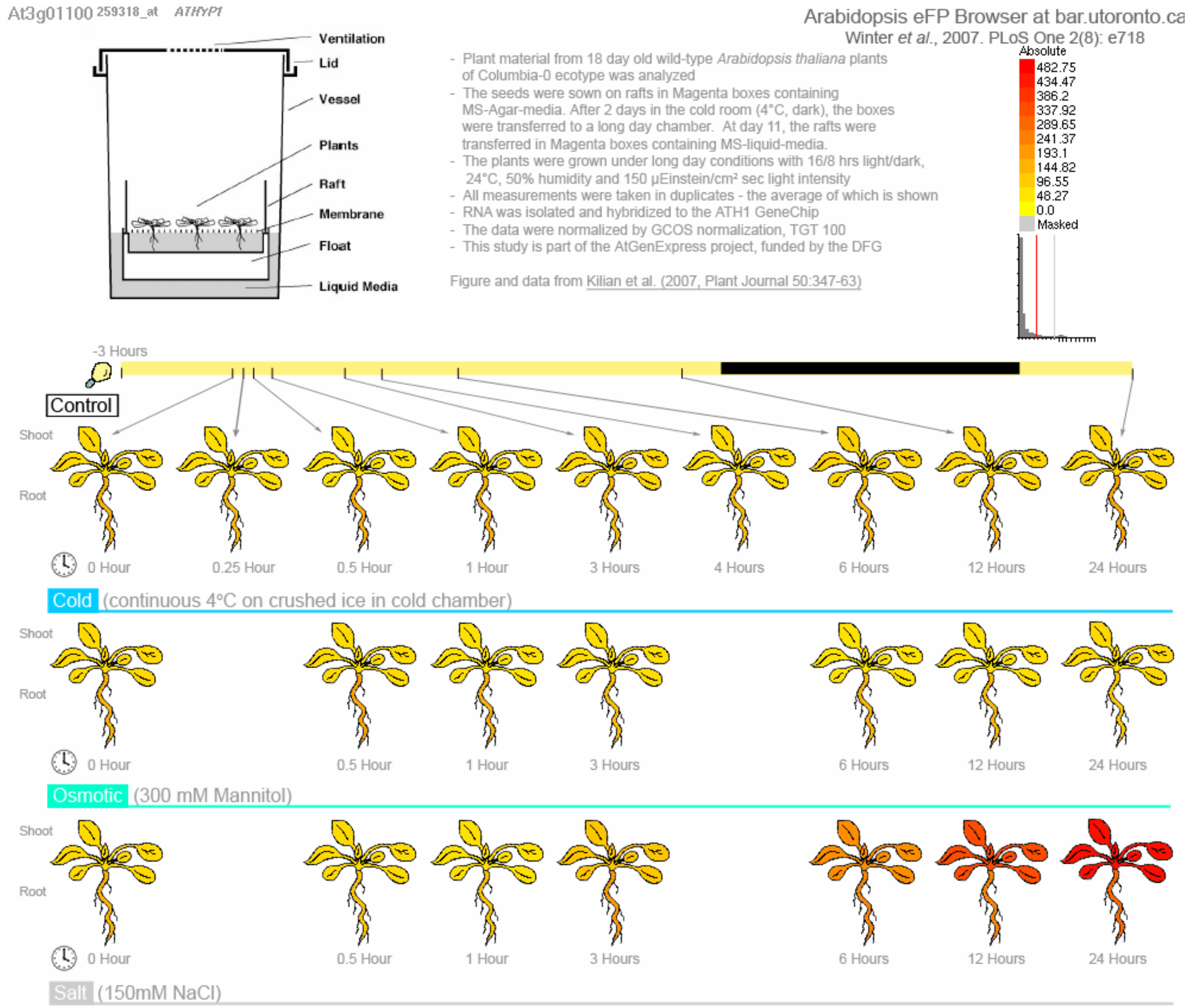
coexpressed gene network* around At3g01100

Genes directly connected with At3g01100 on the network

MR*	Cor*	locus	function	coexpression detail
1.7	0.68	At1g04830	RabGAP/TBC domain-containing protein	[detail]
2.5	0.68	At3g03310	lecithin:cholesterol acyltransferase family protein / LACT family protein	[detail]
3.2	0.68	At4g29820	CFIM-25	[detail]
11.6	0.54	At2g36810	binding	[detail]
14.3	0.55	At3g63330	protein kinase family protein	[detail]

28.7.2009

Expression of the Arabidopsis homolog upon osmotic stress



- Plant material from 18 day old wild-type *Arabidopsis thaliana* plants of Columbia-0 ecotype was analyzed
- The seeds were sown on rafts in Magenta boxes containing MS-Agar-media. After 2 days in the cold room (4°C, dark), the boxes were transferred to a long day chamber. At day 11, the rafts were transferred in Magenta boxes containing MS-liquid-media.
- The plants were grown under long day conditions with 16/8 hrs light/dark, 24°C, 50% humidity and 150 µEinstein/cm² sec light intensity
- All measurements were taken in duplicates - the average of which is shown
- RNA was isolated and hybridized to the ATH1 GeneChip
- The data were normalized by GCOS normalization, TGT 100
- This study is part of the AtGenExpress project, funded by the DFG

Figure and data from Kilian et al. (2007, Plant Journal 50:347-63)

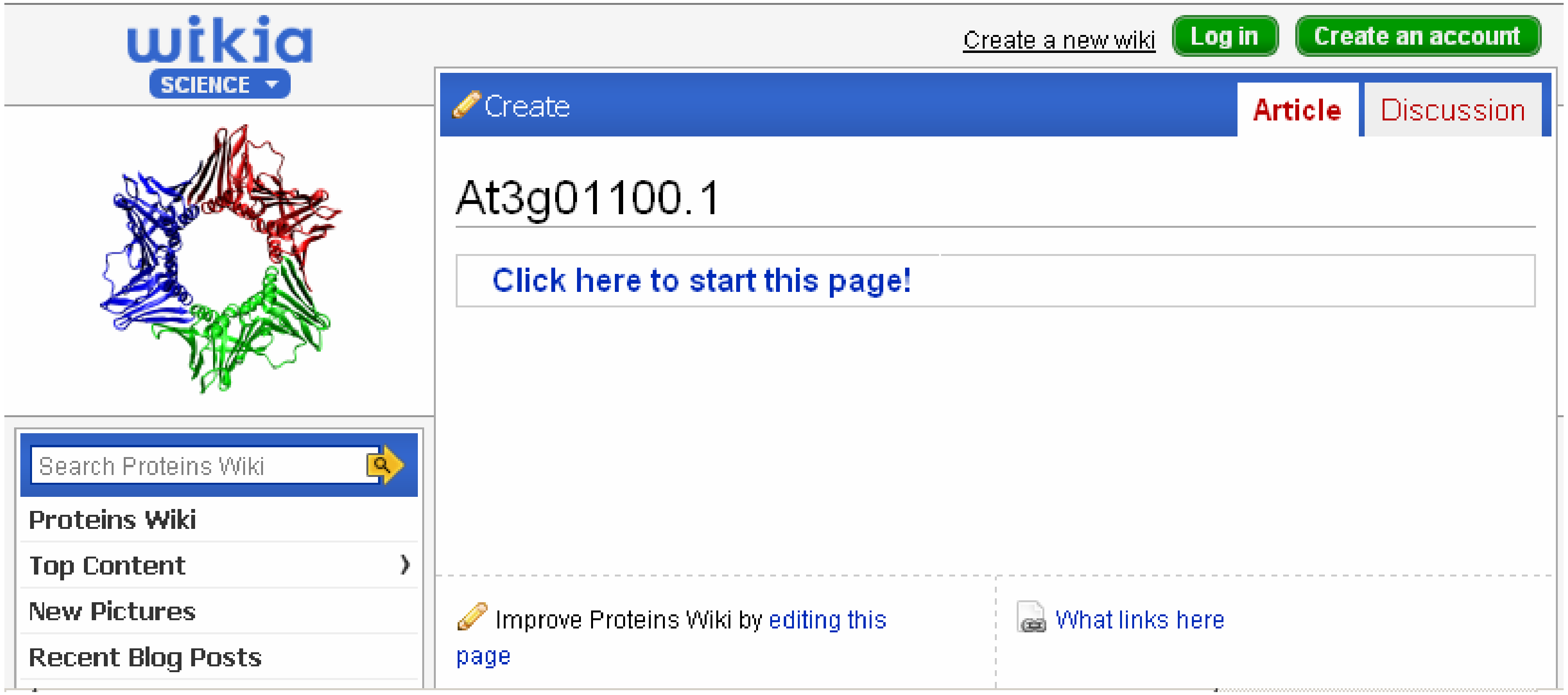
Cold (continuous 4°C on crushed ice in cold chamber)

Osmotic (300 mM Mannitol)

Salt (150mM NaCl)

28.7.2009

Still no entry in Proteins Wiki



Create a new wiki | Log in | Create an account

Create | Article | Discussion

At3g01100.1

Click here to start this page!

Improve Proteins Wiki by editing this page

What links here

http://proteins.wikia.com/wiki/

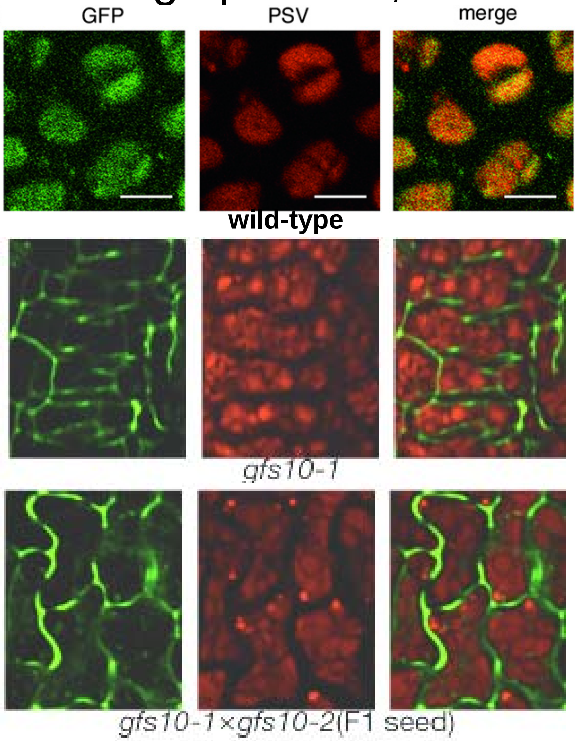
Summary Peptide/Protein Annotation MILSALLTSVGINLGLC

- belongs to gene family –
  - members HYP1 (hypothetical protein 1)
  - RXW8 (name of cDNA)
  - ERD4 (early responsive to dehydration)
  - AtGFS10 (protein involved in vacuolar sorting fo storage proteins, green fluorescent seed, *gfs* mutant)

Quote: „no closely related homologs of GFS10 in the Arabidopsis

but topology similar – Aramemnon and 39% amino acid h

- integral membrane protein
- plasma membrane, secretory pathway
- induced during water stress



secrete vacuole-targeted GFP out of the seed cells

28.7.2009

What next?

Design experiments for functional characterization

Poplar gene with homolog in Arabidopsis (81%)

First functional tests in Arabidopsis: knock outs

over-/ectopic-/ inducible expression

in vivo localization – XFP, immuno

interaction with other proteins

My Favourite Site
SIGnAL T-DNA Express

Arabidopsis Tiling Array Transcriptome

Arabidopsis Gene ORFeome Collection

Salk Insertion Sequence Database

Salk Homozygote T-DNA Collection

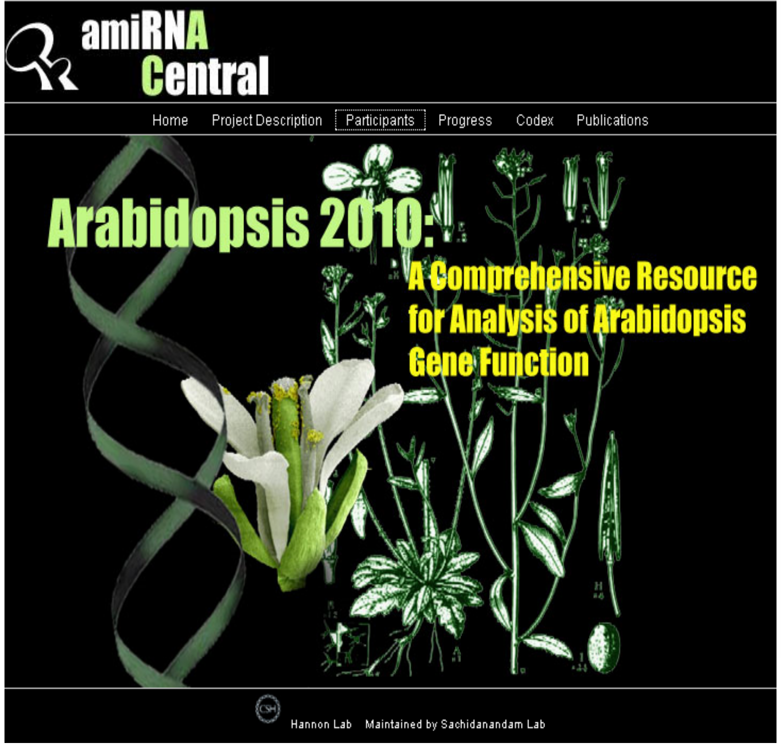
High Resolution Arabidopsis Methylome

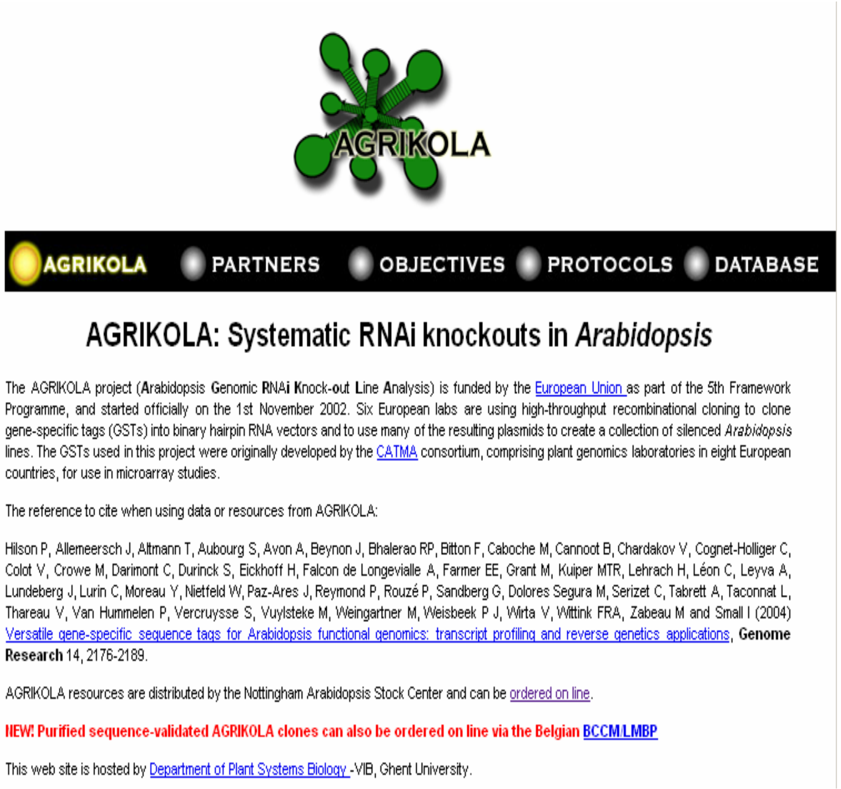
Perlegen Arabidopsis Resequencing

Rice Functional Genomics Database

Single Feature Polymorphism Database

High Resolution Arabidopsis Exosome





AGRIKOLA: Systematic RNAi knockouts in *Arabidopsis*

The AGRIKOLA project (Arabidopsis Genomic RNAi Knock-out Line Analysis) is funded by the European Union as part of the 5th Framework Programme, and started officially on the 1st November 2002. Six European labs are using high-throughput recombinational cloning to clone gene-specific tags (GSTs) into binary hairpin RNA vectors and to use many of the resulting plasmids to create a collection of silenced *Arabidopsis* lines. The GSTs used in this project were originally developed by the CATMA consortium, comprising plant genomics laboratories in eight European countries, for use in microarray studies.

The reference to cite when using data or resources from AGRIKOLA:

Hilson P, Allemeersch J, Altmann T, Aubourg S, Avon A, Beynon J, Bhalerao RP, Bitton F, Caboche M, Cannoot B, Chardakov V, Cognet-Holliger C, Colot V, Crowe M, Darimont C, Durinck S, Eickhoff H, Falcon de Longevialle A, Farmer EE, Grant M, Kuiper MTR, Lehrach H, Léon C, Leyva A, Lundeberg J, Lurin C, Moreau Y, Nietfeld W, Paz-Ares J, Reymond P, Rouzé P, Sandberg G, Dolores Segura M, Serizet C, Tabrett A, Taconnat L, Thareau V, Van Hummelen P, Vercruysse S, Vuylsteke M, Weingartner M, Weisbeek P J, Wirta V, Wittink FRA, Zabeau M and Small I (2004) Versatile gene-specific sequence tags for Arabidopsis functional genomics: transcript profiling and reverse genetics applications, **Genome Research** 14, 2176-2189.

AGRIKOLA resources are distributed by the Nottingham Arabidopsis Stock Center and can be ordered on line.

NEW: Purified sequence-validated AGRIKOLA clones can also be ordered on line via the Belgian BCCM/LMBP

This web site is hosted by Department of Plant Systems Biology -VIB, Ghent University.

28.7.2009

QTL-Analysis or Association Mapping

28.7.2009

Understanding functional consequences of natural variation: trichome patterning in Arabidopsis

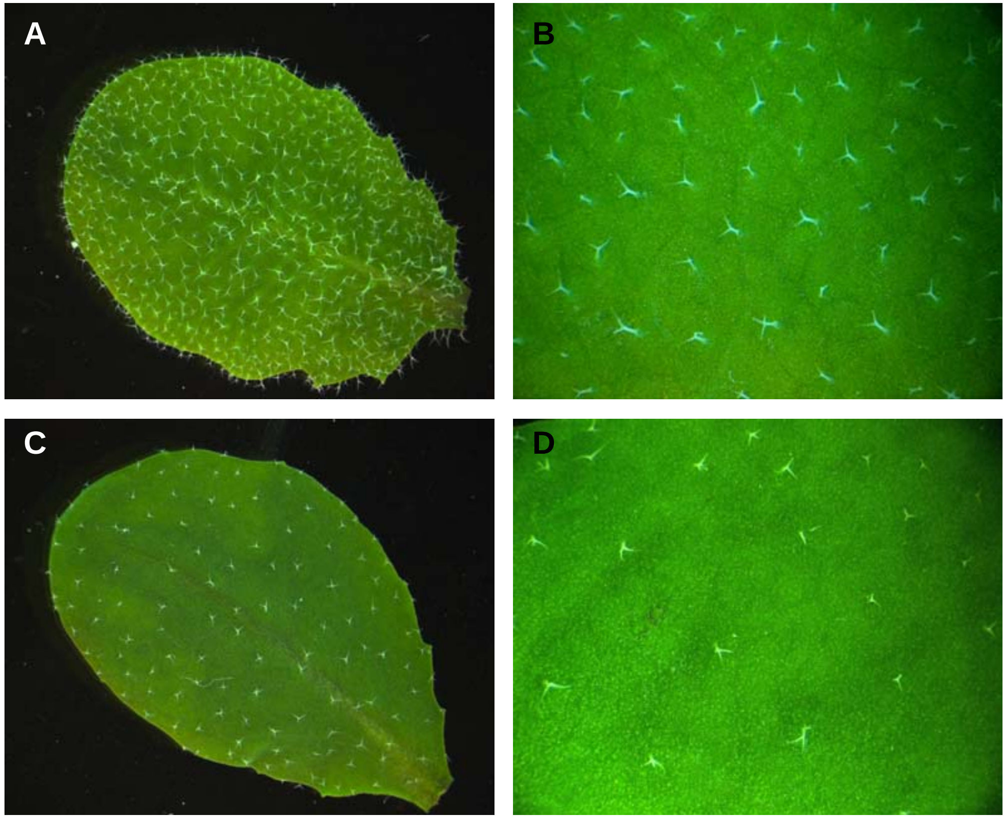
Example for the comparison of genomes/gene and their function from individuals between populations

Julia Hilscher, Christian Schlötterer, Marie-Theres Hauser

28.7.2009

Trichomes in Arabidopsis

- Single cell structure
- Present on leaves, stem, petioles, sepals
- 32C->polyploid
- Model for cell fate specification



28.7.2009

Trichome function

- **Arabidopsis**
  - Protection against herbivory

    Mauricio & Rausher (1997), Handley, Ekbom & Ågren (2005)

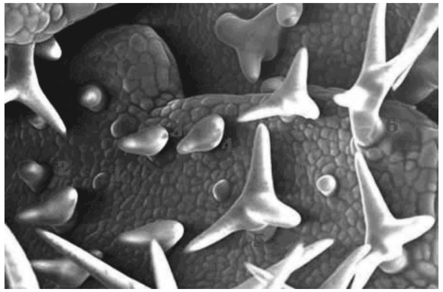
- **A. lyrata**
  - Protection against herbivory

    Kivimäki, Kärkkäinen, Gaudeul, Løe & Ågren (2007)

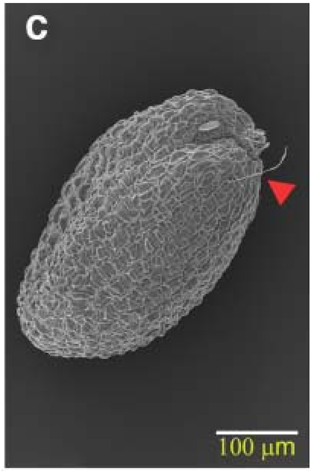
- **Other plants**
  - Decrease of water loss
  - Increased light reflection
  - Freezing tolerance
  - Ca^{++} homeostasis
  - Heavy metal storage
  - Metabolite production and storage
  - Cotton fiber development

Cross-species function of Trichome regulators

Trichome development of Arabidopsis

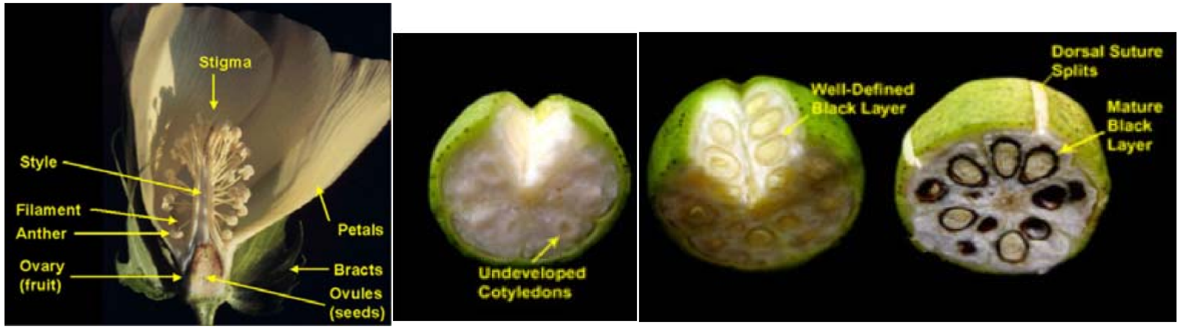


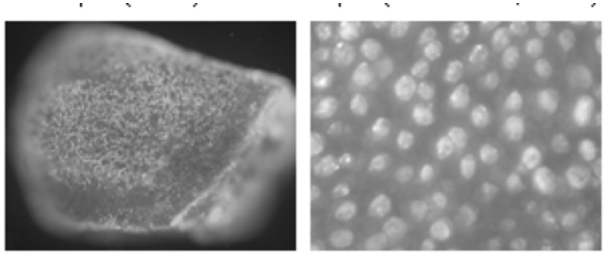
35S::GaMYB2 in Arabidopsis



„hairs on seeds“

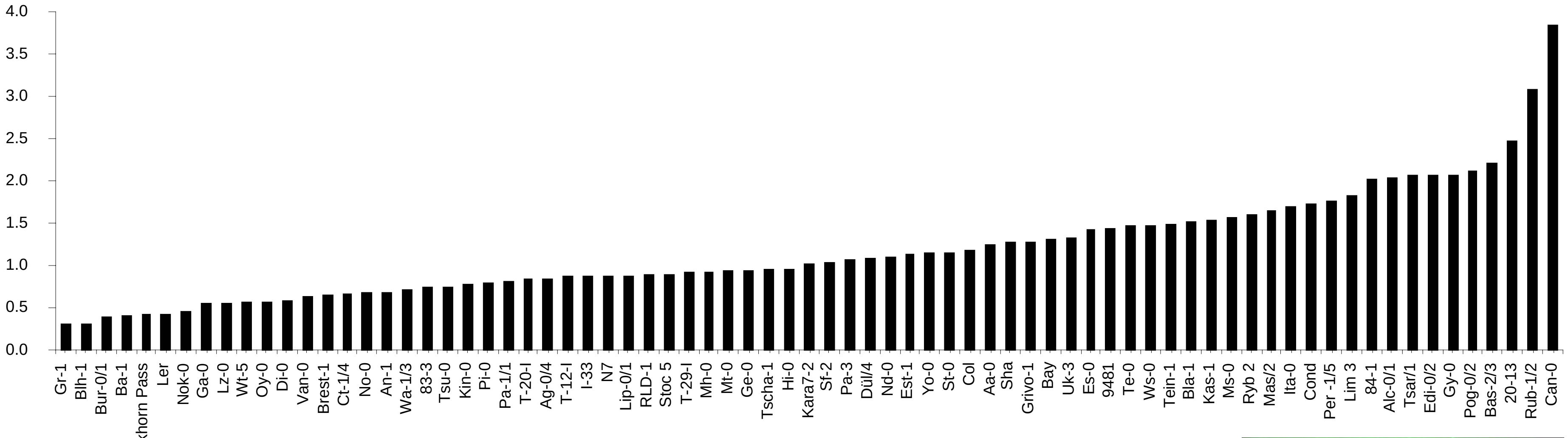
Cotton fiber development

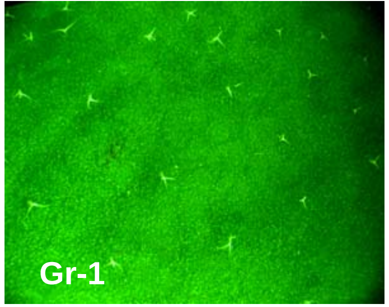


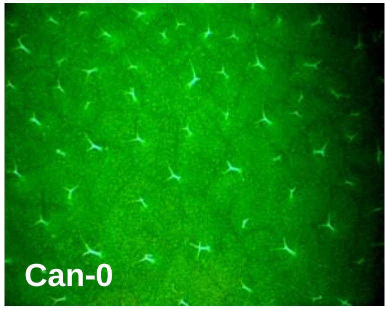


0 dpa ovule (20X) 0 dpa (100X)

Trichome density differs in natural populations



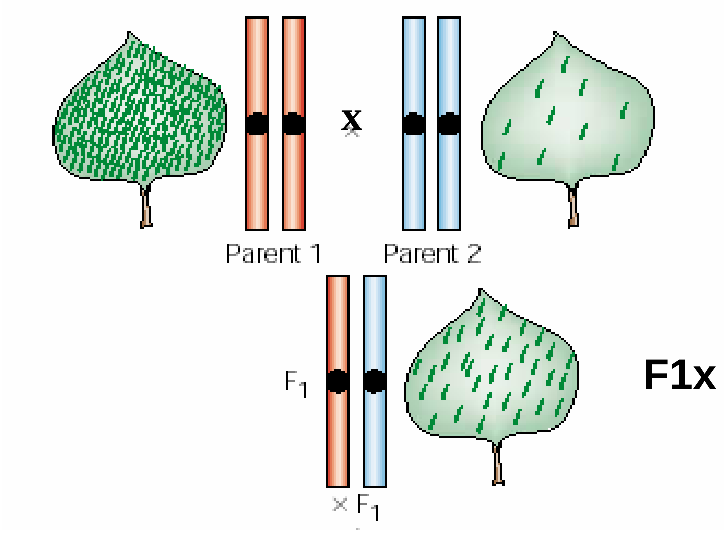




1.5 mm

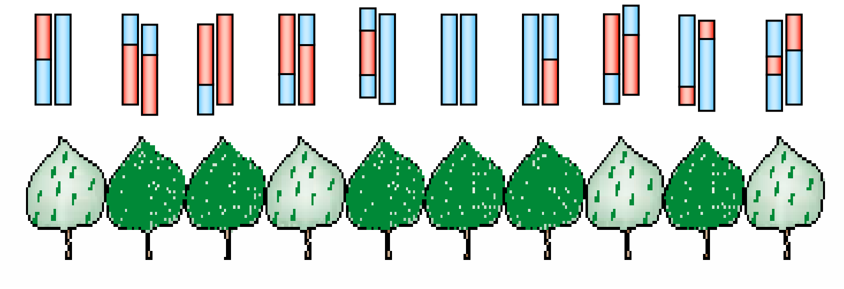
28.7.2009

QTL mapping

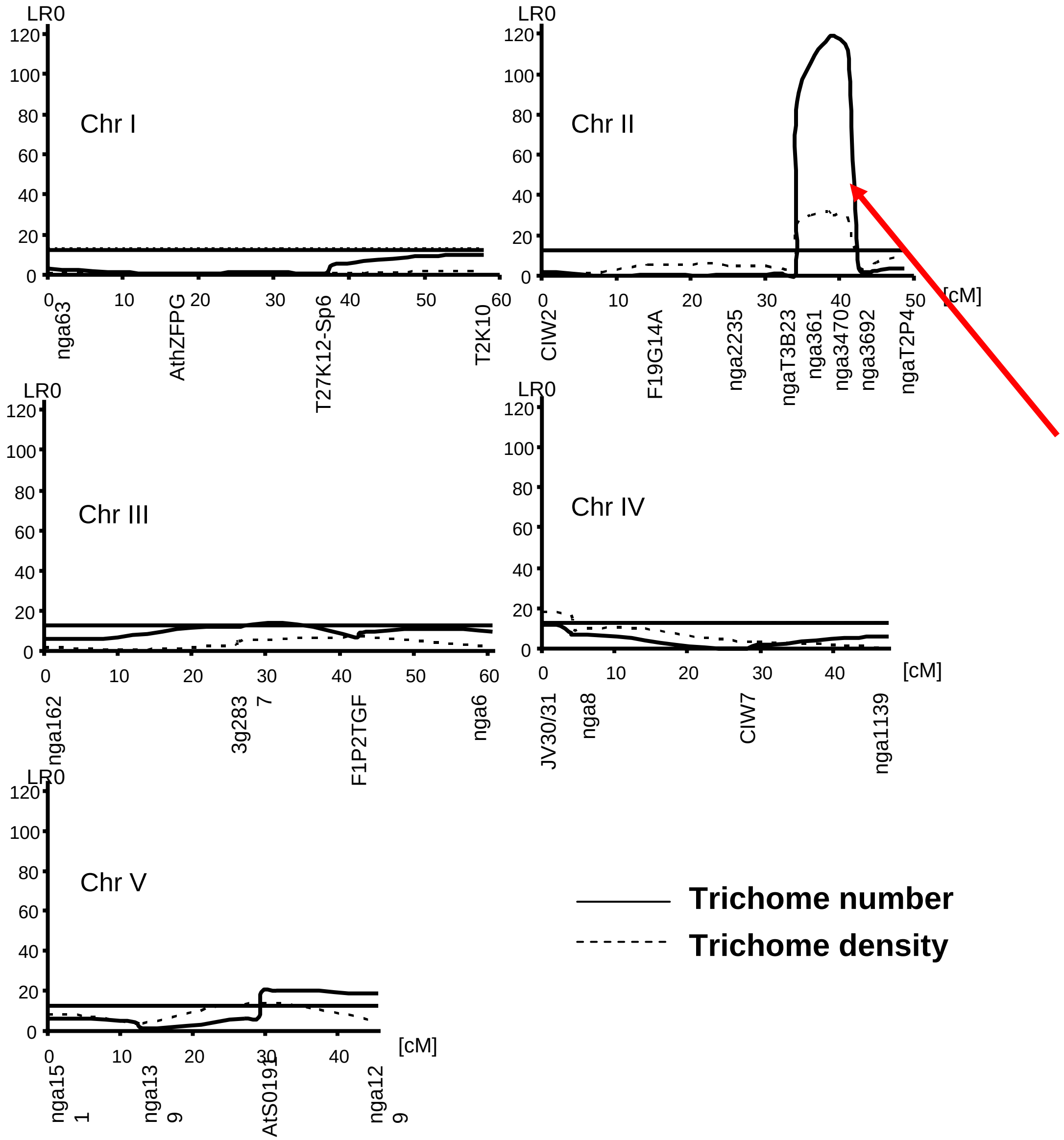


F1x F1

F2



Composite Interval Mapping

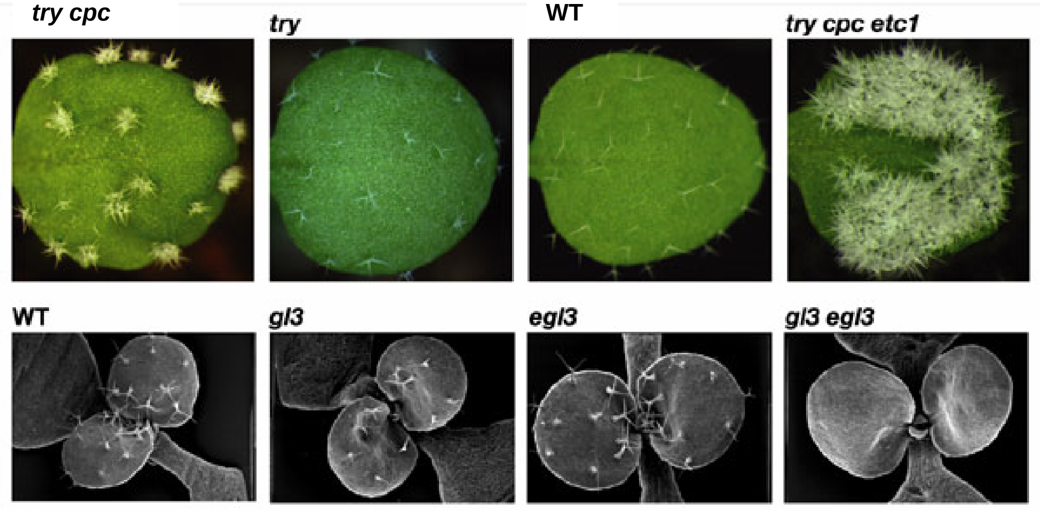


- 266 F2 individuals
- 24 microsatellites
- A single QTL on chromosome 2 explains 33% of the variation in trichome number

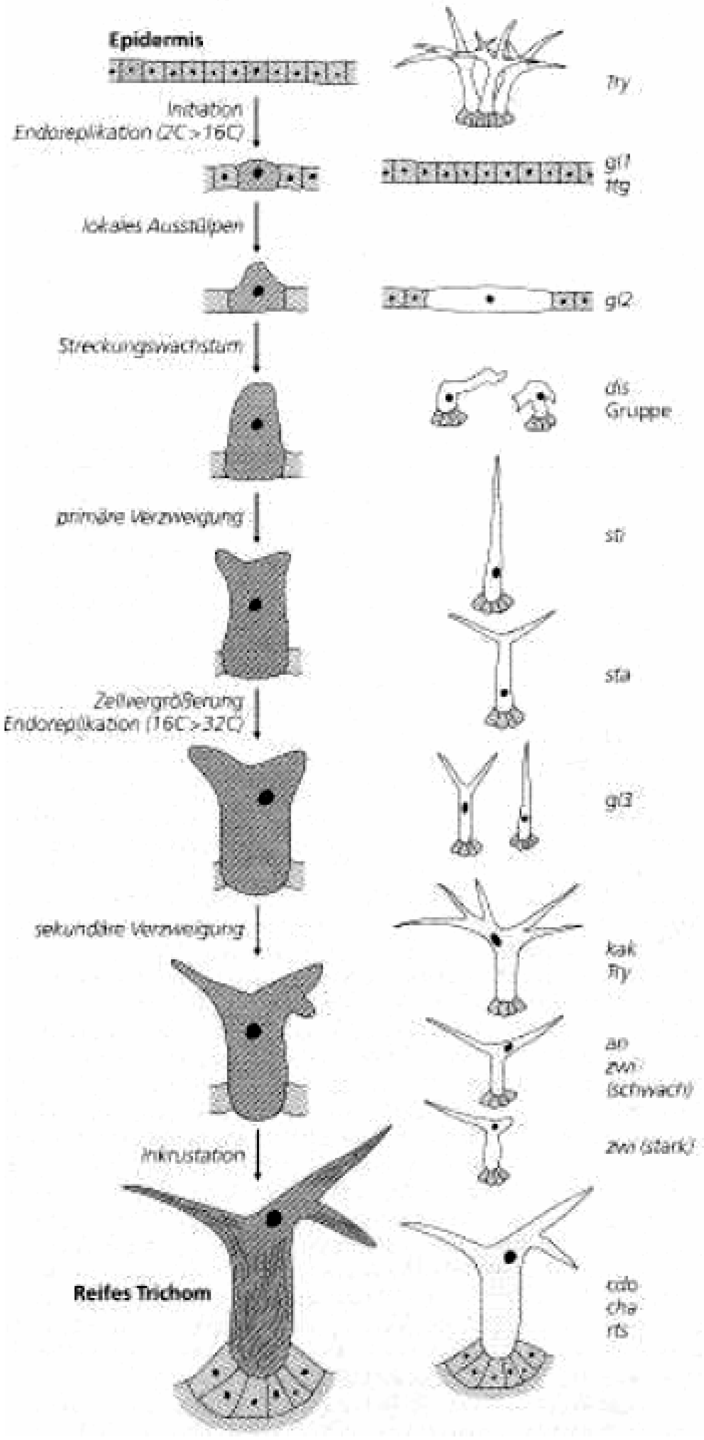
28.7.2009

Components of trichome development

Patterning involves positive and negative regulators with redundant functions



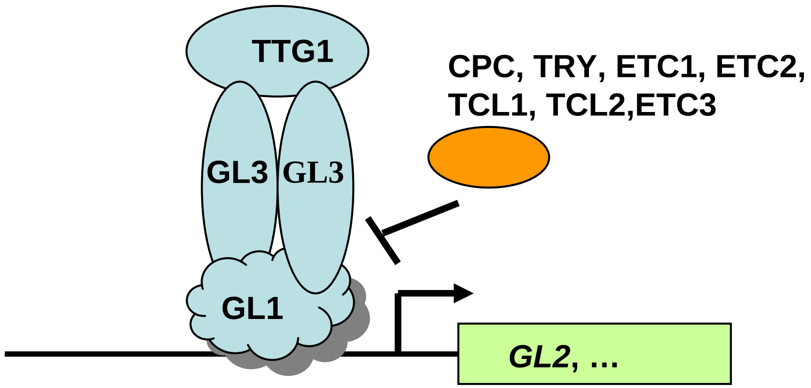
Nature Reviews | Molecular Cell Biology



Model of epidermal patterning

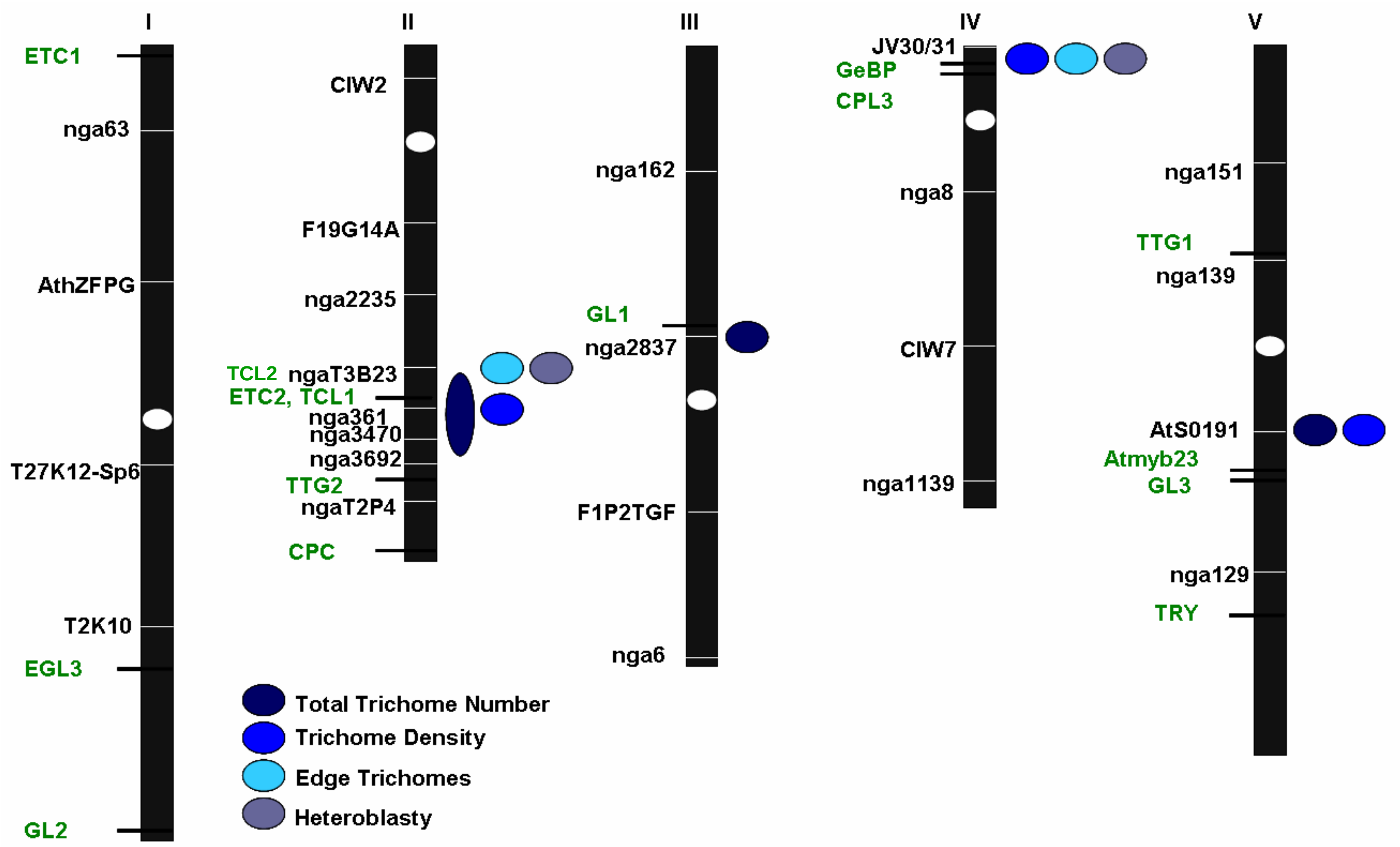
Phenotypic class of mutants	Gene	Action
no trichomes “glabrous”	*GL1*	Positive regulators
	TTG1	
decreased trichome #	*GL3/ EGL3*	
	GL2	
nests of trichomes	*TRY*	Negative regulators single-repeat R3 MYB genes, act non-cell autonomous
increased trichome #	*CPC*	

All of the genes or their paralogs are also involved in root hair patterning



modified from Larkin et al., 2003

Composite Interval Mapping

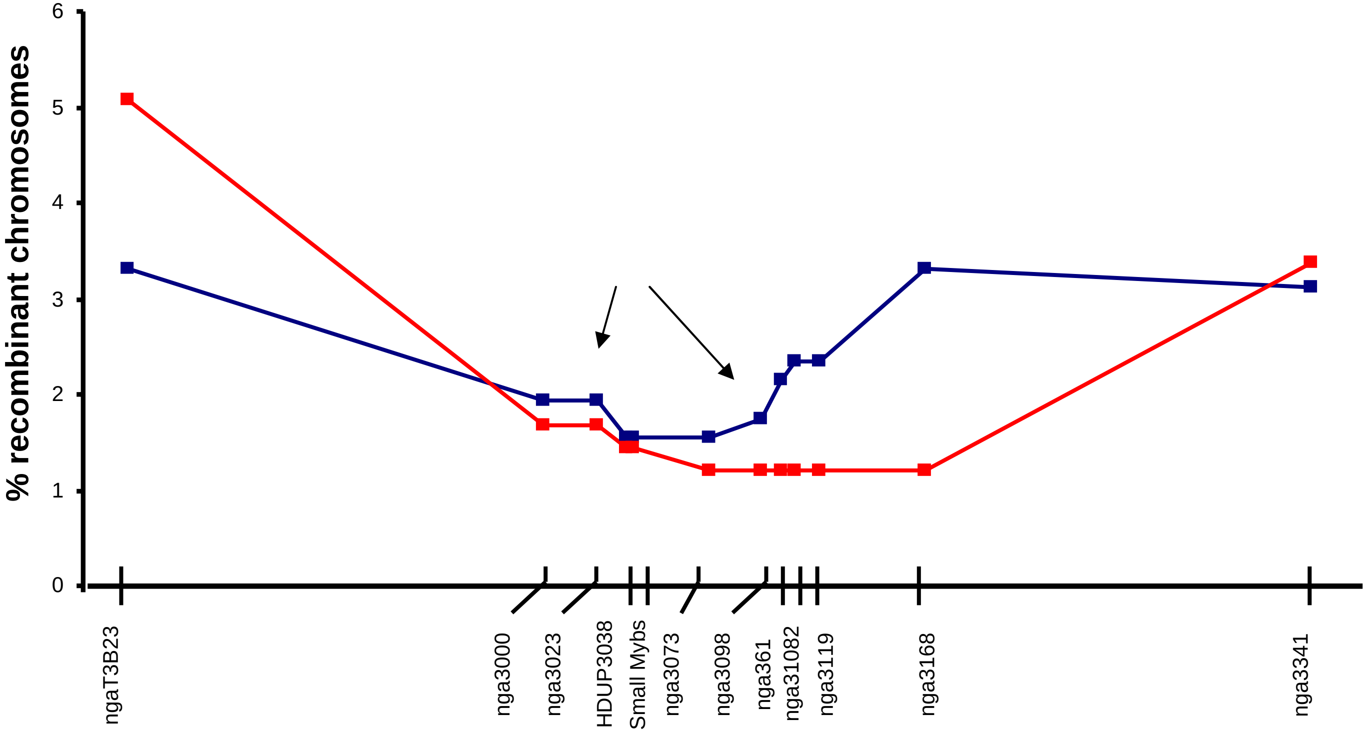


28.7.2009

Fine mapping on chromosome II

- Selective genotyping
  - 465 F2 with extreme phenotype
  - 10 additional markers
- 66 recombinants in the initial interval

-> expected mapping resolution of 40kb

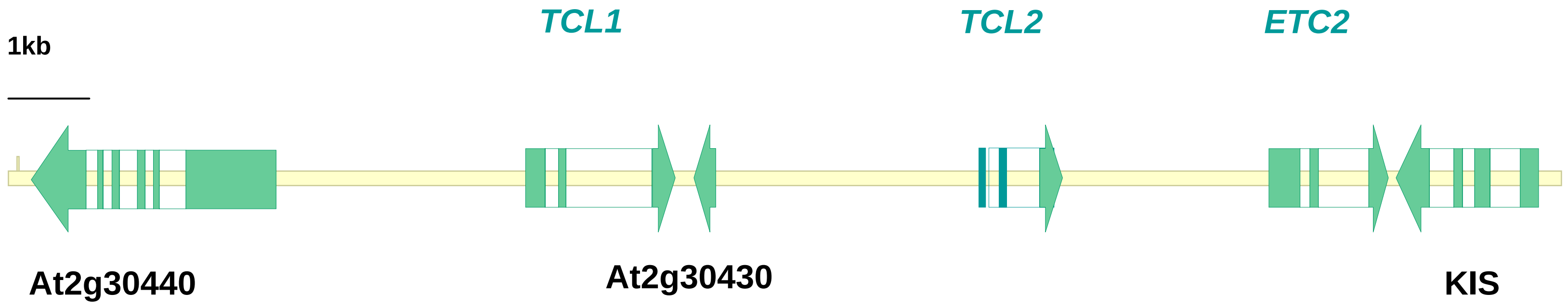


- 288 kb mapping interval with remainig 87 genes

28.7.2009

3 candidate genes

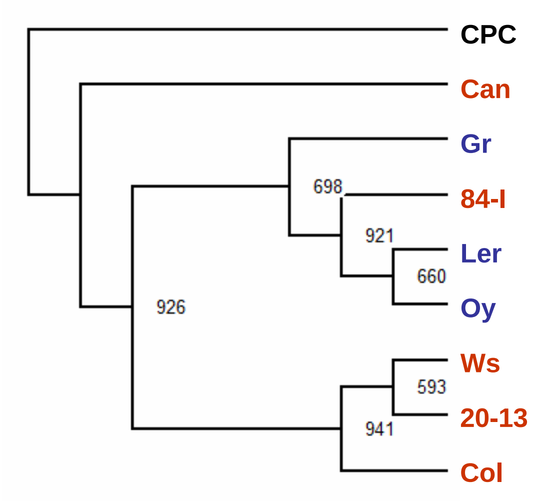
Three single-repeat R3 MYB genes are located in the fine-mapping interval



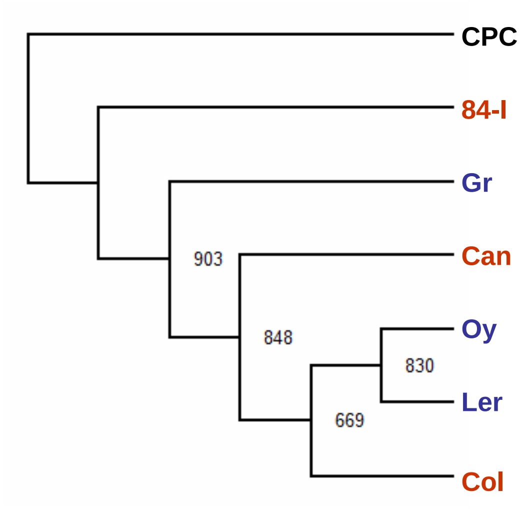
28.7.2009

Association with trichome number

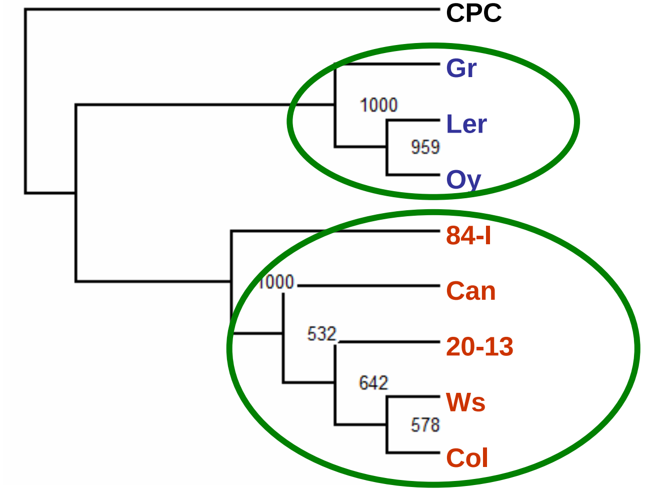
Small MYB_A



Small MYB_B



Small MYB_C



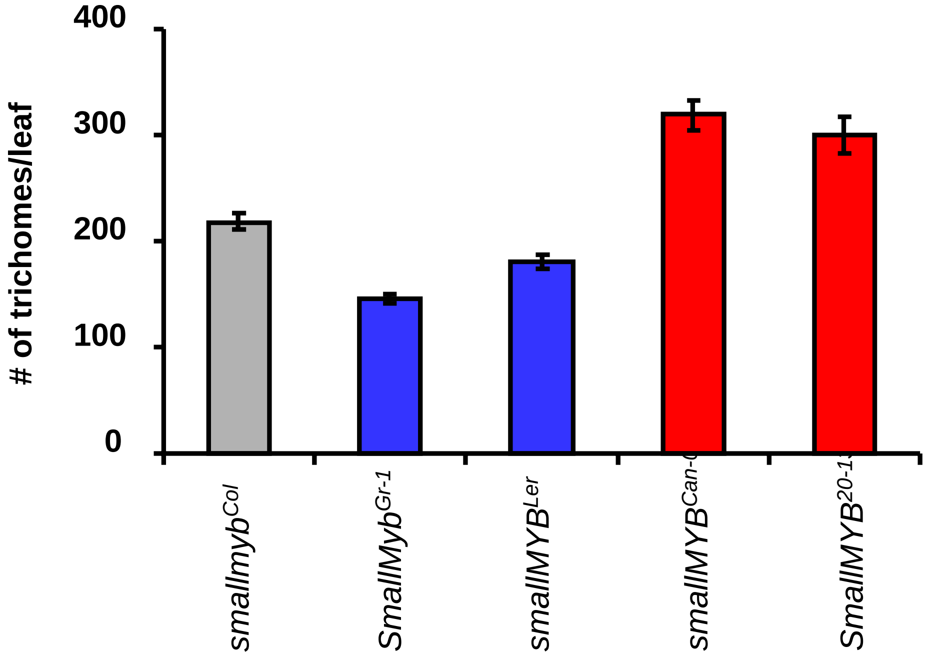
low trichome number accessions

high trichome number accessions

28.7.2009

Complementation test

mutants are complemented with the wildtype alleles from high and low trichome number accessions



F1 cross to

low trichome number accession

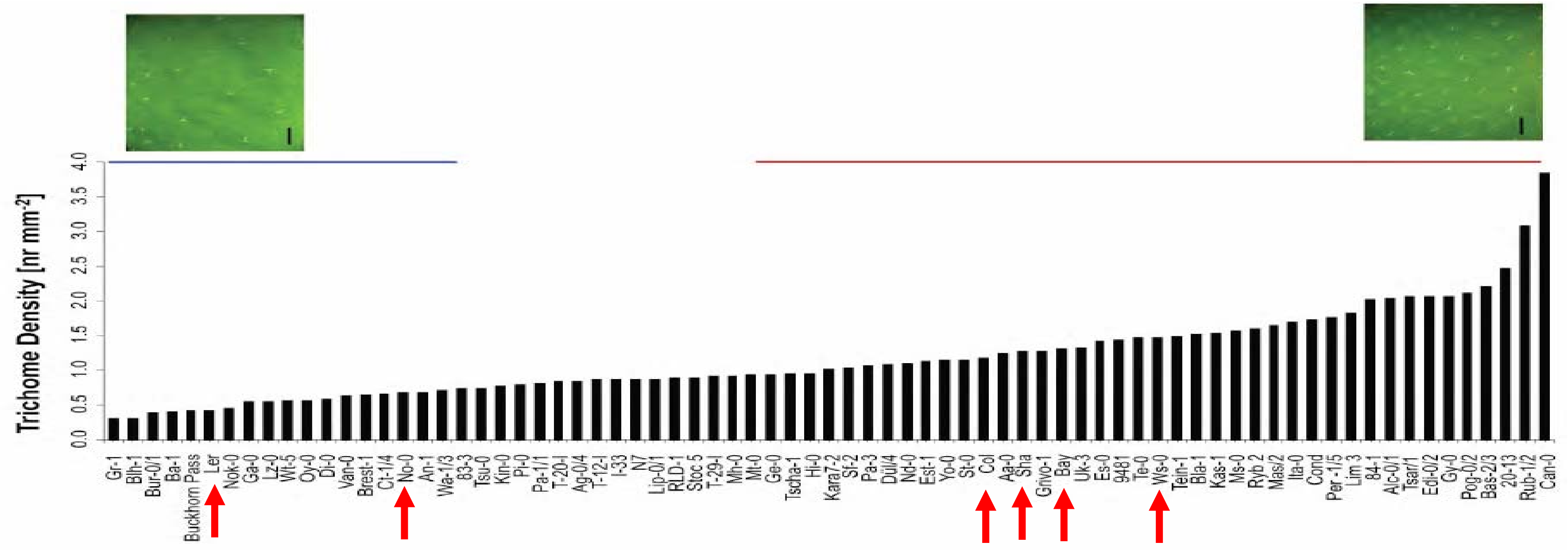
high trichome number accession

28.7.2009

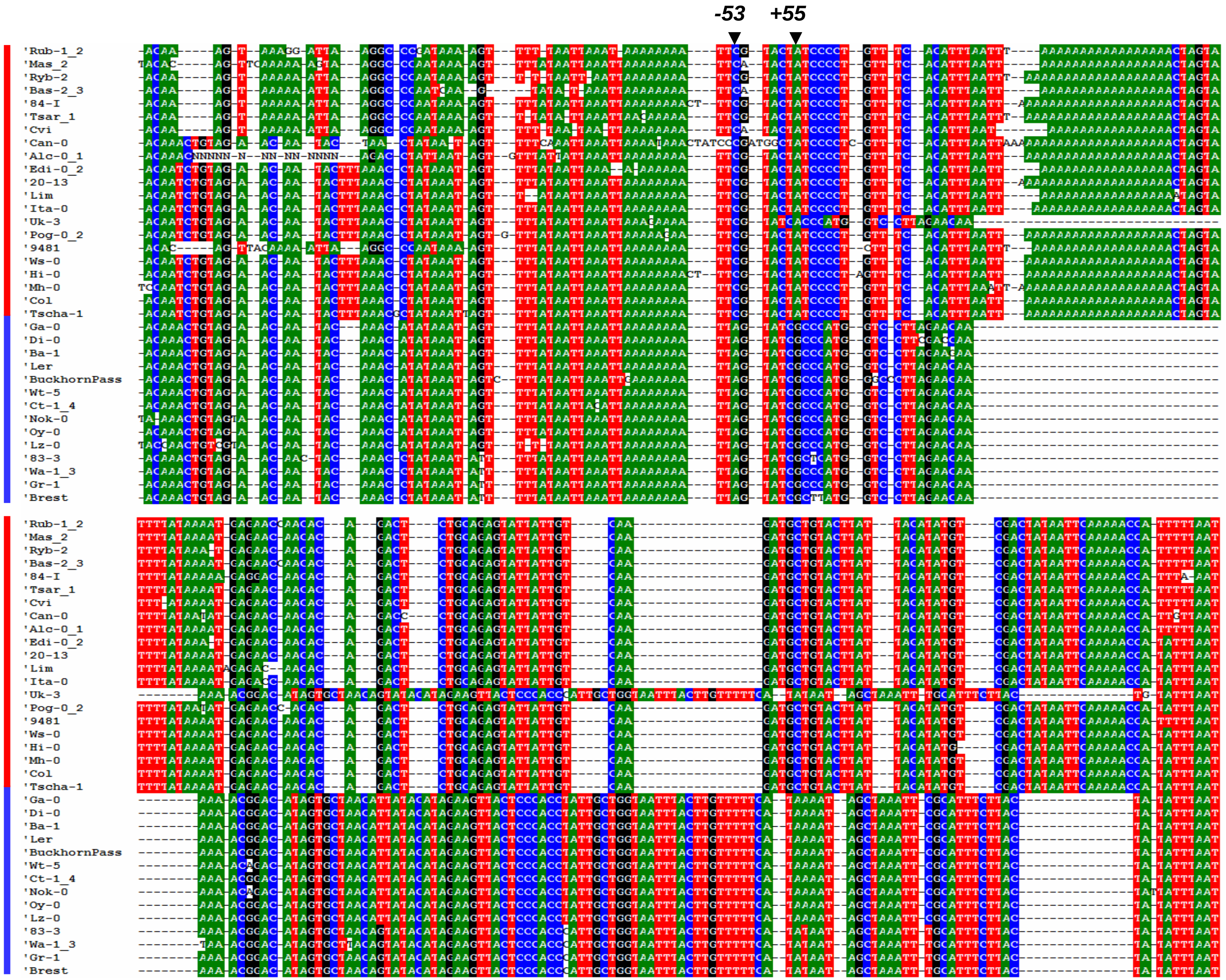
Identification of QTN

Many SNPs show association with trichome phenotype

Screen many high/low trichome accessions for recombinants

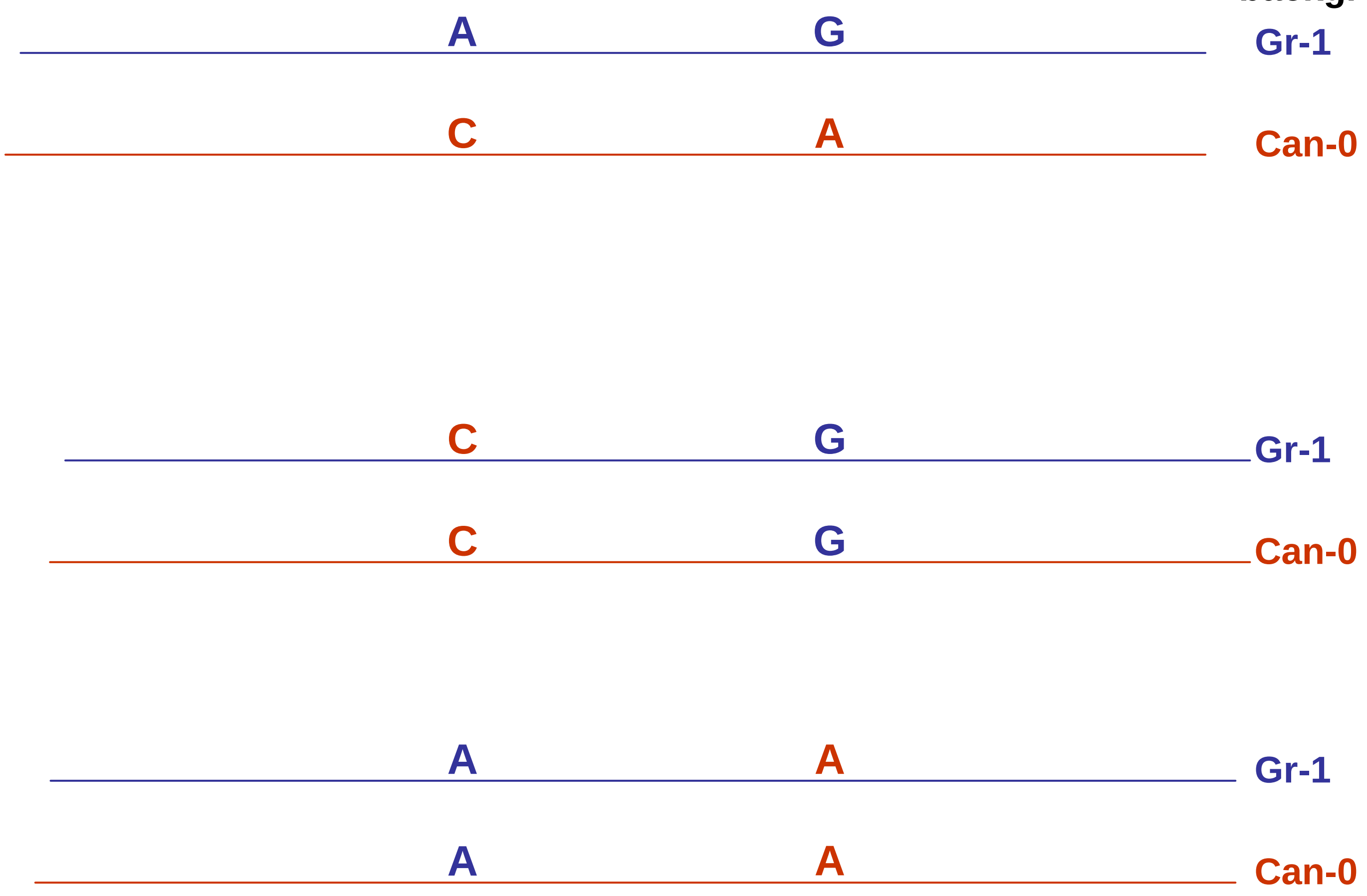


Identification of QTN: 2 candidates



28.7.2009

Identification of QTN

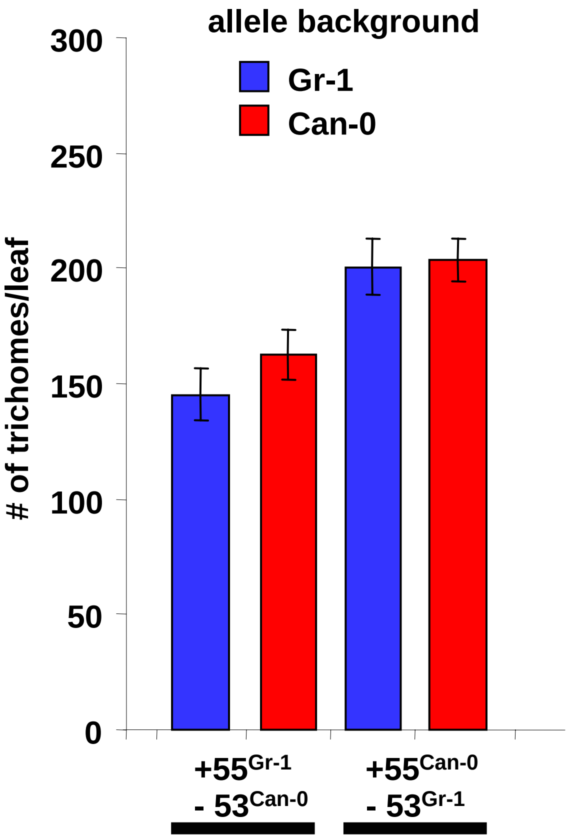


SNP status	position	-53 bp	+55 bp	allele background
wt		A	G	Gr-1
		C	A	Can-0
+55 Gr-1		C	G	Gr-1
-53 Can-0		C	G	Can-0
+55 Can-0		A	A	Gr-1
-53 Gr-1		A	A	Can-0

28.7.2009

Identification of QTN

Transgenic complementation



ANOVA

SNP state: p=0.001

allele background: p=0.35

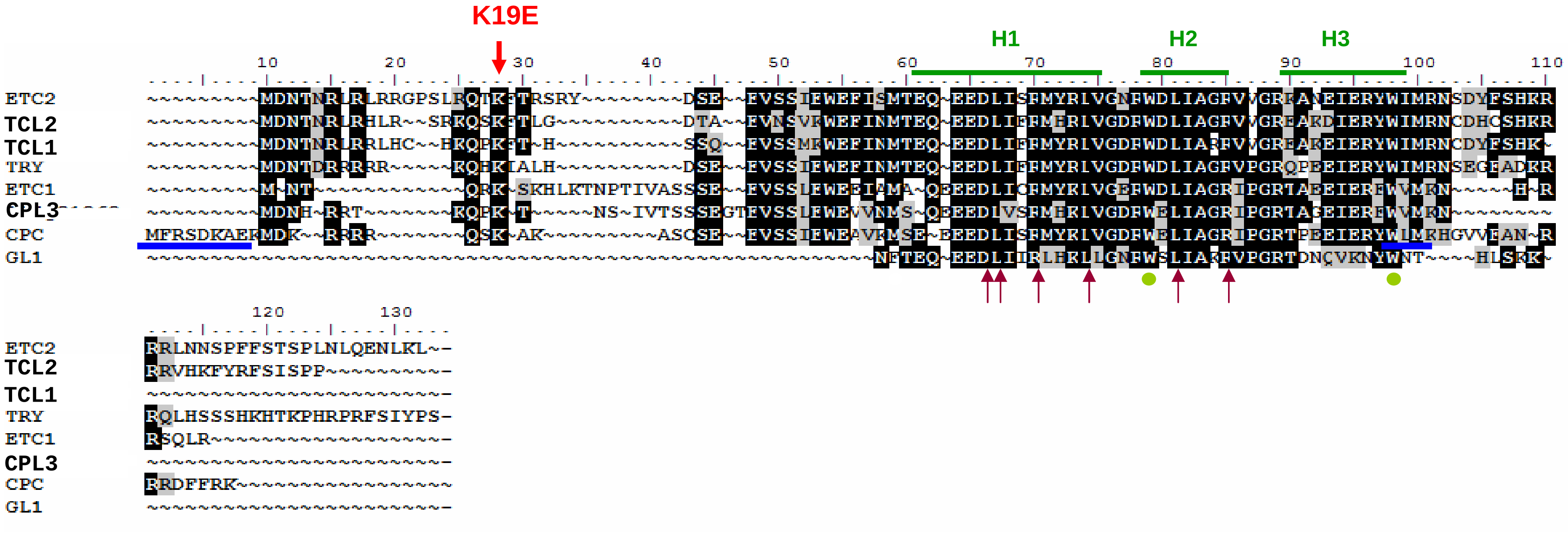
Mean ± s.e.m. of n>21 T1 lines each, 3 leaf positions

28.7.2009

QTN is highly conserved among family members

+55 mutation leads to an amino acid replacement: Lysine (K) to Glutamate (E)

K: ancestral, yet unknown importance



```
                    10        20        30        40        50        60        70        80        90       100       110
           ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
ETC2       ~~~~~~~~~MDNTNRLRLRRGPSLRQTKFTRSRY~~~~~~~~DSE~~EVSSIEWEFISMTEQ~EEDLISFMYRLVGNFWDLIAGFVVGRRANEIERYWIMRNSDYFSHKR
TCL2       ~~~~~~~~~MDNTNRLRHLR~~SRRQSKFTLG~~~~~~~~~~DTA~~EVNSVKWEFINMTEQ~EEDLIFFMHRLVGDFWDLIAGFVVGREAKDIERYWIMRNCDHCSHKR
TCL1       ~~~~~~~~~MDNTNRLRRLHC~~HKQPKFT~H~~~~~~~~~~~SSQ~~EVSSMKWEFINMTEQ~EEDLIFFMYRLVGDFWDLIARFVVGREAKEIERYWIMRNCDYFSHK~
TRY        ~~~~~~~~~MDNTDRRRRR~~~~~KQHKIALH~~~~~~~~~~~DSE~~EVSSIEWEFINMTEQ~EEDLIFFMYRLVGDFWDLIAGFVPGRQPEEIERYWIMRNSEGEADKR
ETC1       ~~~~~~~~~M~NT~~~~~~~~~~~~QRK~SKHLKTNPTIVASSSE~~EVSSLEWEEIAMA~QEEEDLICFMYRLVGEFWDLIAGRIPGRTAEEIERFWVMKN~~~~~H~R
CPL3       ~~~~~~~~~MDNH~RRT~~~~~~~KQPK~T~~~~~NS~IVTSSSEGTEVSSLEWEVVNMS~QEEEDLVSFMHRLVGDFWELIAGRIPGRTAGEIERFWVMKN~~~~~~~~
CPC        MFRSDKAEKMDK~~RRRR~~~~~~~QSK~AK~~~~~~~~~ASCSE~~EVSSIEWEAVKMSE~EEEDLISFMYRLVGDFWELIAGRIPGRTPEEIERYWLMRHGVVEAN~R
GL1        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~NFTEQ~EEDLIIRLHRLLGNFWSLIAKFVPGRTDNCVKNYWNT~~~~HLSKK~

                   120       130
           ....|....|....|....|....
ETC2       RRLNNSPFFSTSPLNLQENLKL~-
TCL2       RRVHKFYRFSISPP~~~~~~~~~-
TCL1       ~~~~~~~~~~~~~~~~~~~~~~~-
TRY        RQLHSSSHKHTKPHRPRFSIYPS-
ETC1       RSQLR~~~~~~~~~~~~~~~~~~-
CPL3       ~~~~~~~~~~~~~~~~~~~~~~~-
CPC        RRDFFRK~~~~~~~~~~~~~~~~~
GL1        ~~~~~~~~~~~~~~~~~~~~~~~-
```

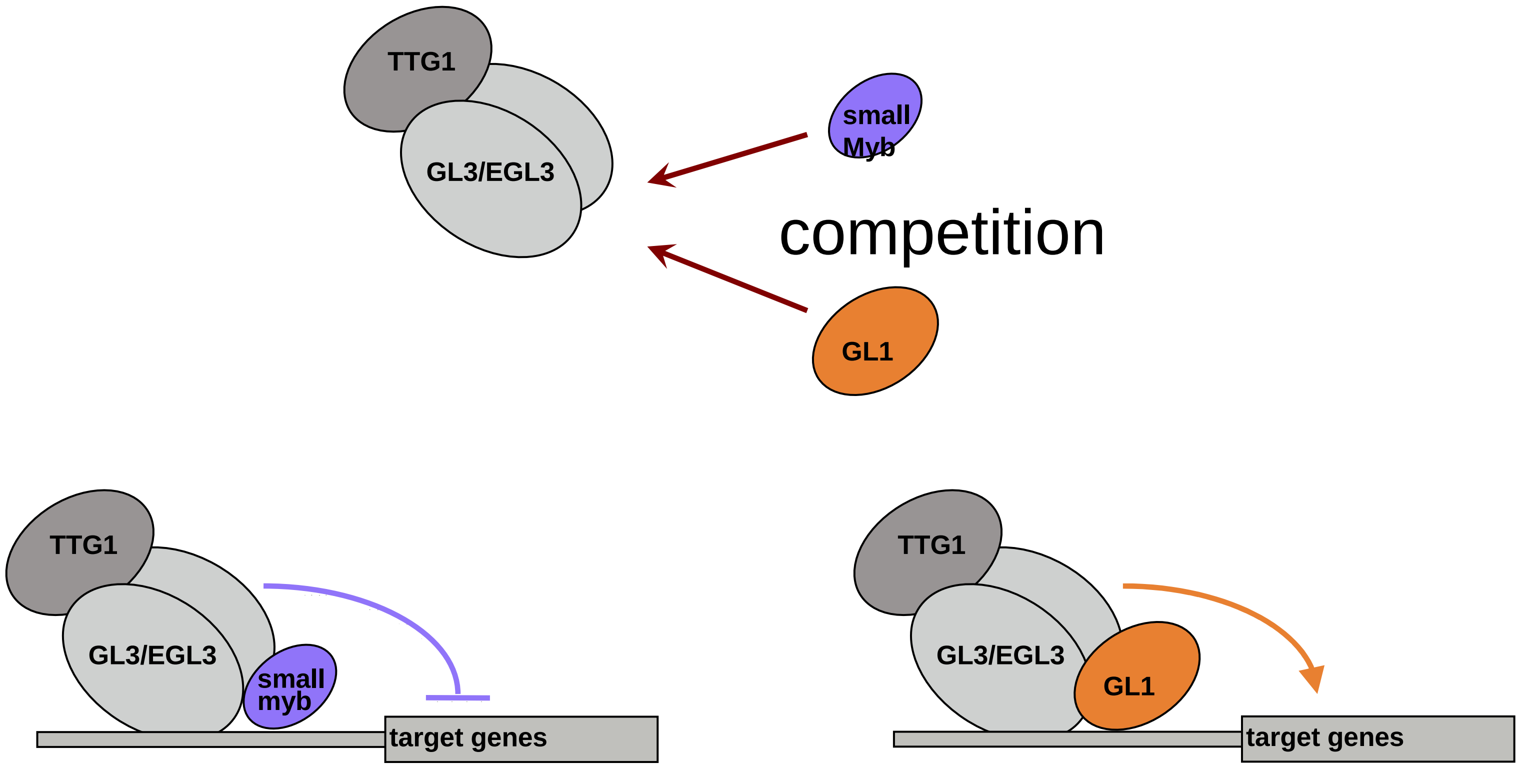
Alpha Helices 1-3 constituting R3 MYB domain with conserved W residues forming cluster

Predicted bHLH interaction motif: [DE]Lx2[RK]x3Lx6Lx3R (Zimmermann et al., Plant J 2004)

Required for CPC movement (Kurata et al., Development 2005)

28.7.2009

Competition between activators and repressors



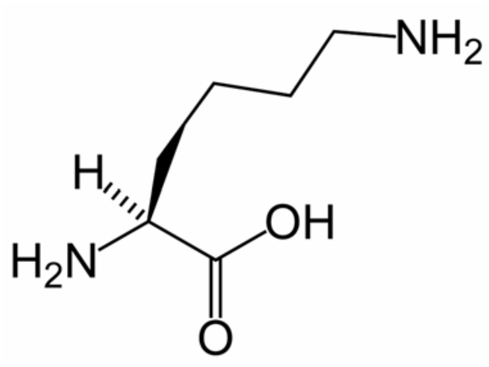
28.7.2009

3 possible factors for trichome patterning

Binding strength to GL3 or the regulatory region

Movement rate to neighboring cells

Stability



Lysine modification by: methylation
N-glycosylation
ubiquitylation
sumoylation
acetylation

Glutamate modification by: methylation

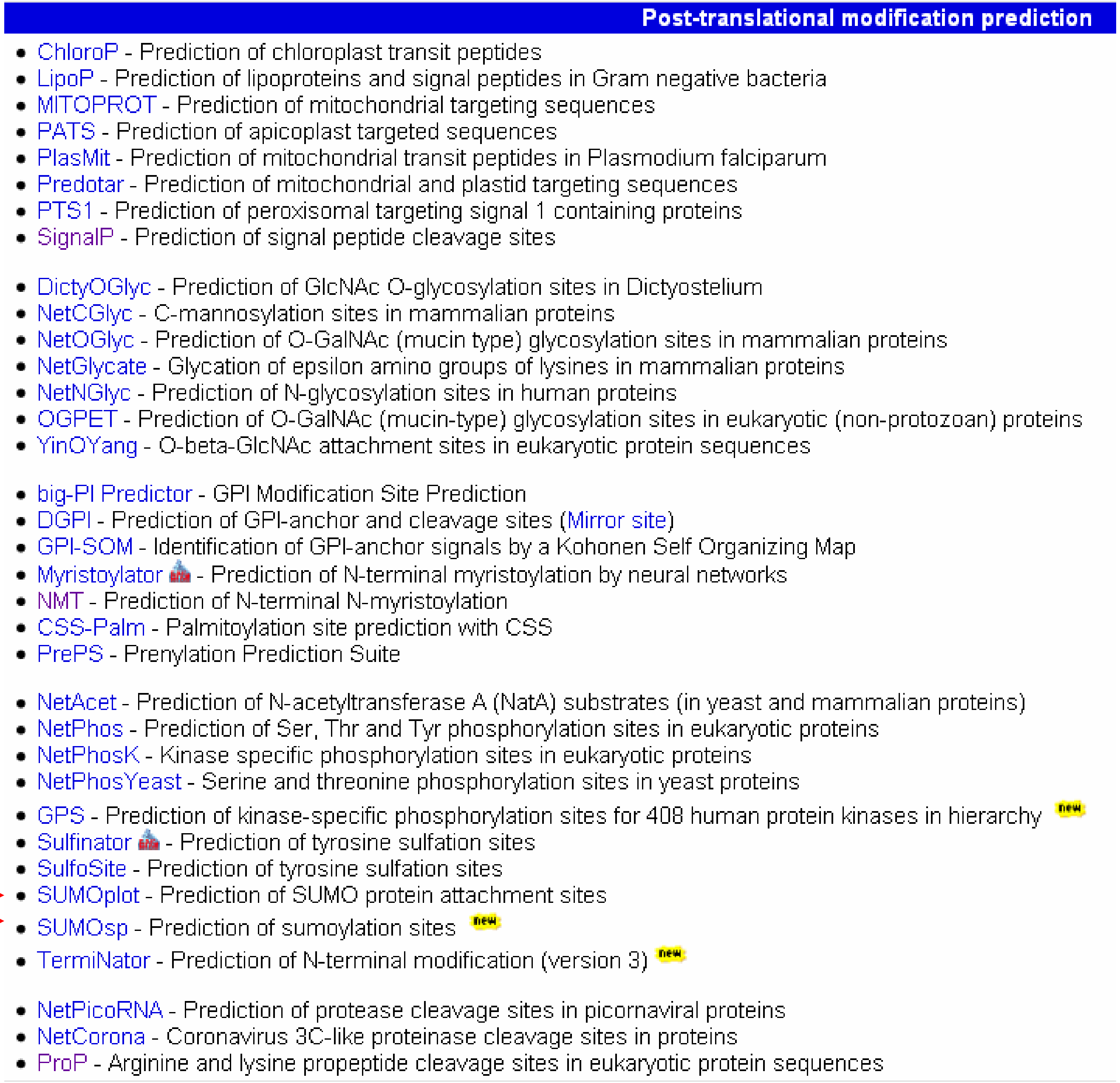
28.7.2009

Next steps

Biochemistry

Cell Biology

Sumoylation? Use ExPASy



Post-translational modification prediction

- ChloroP - Prediction of chloroplast transit peptides
- LipoP - Prediction of lipoproteins and signal peptides in Gram negative bacteria
- MITOPROT - Prediction of mitochondrial targeting sequences
- PATS - Prediction of apicoplast targeted sequences
- PlasMit - Prediction of mitochondrial transit peptides in Plasmodium falciparum
- Predotar - Prediction of mitochondrial and plastid targeting sequences
- PTS1 - Prediction of peroxisomal targeting signal 1 containing proteins
- SignalP - Prediction of signal peptide cleavage sites

- DictyOGlyc - Prediction of GlcNAc O-glycosylation sites in Dictyostelium
- NetCGlyc - C-mannosylation sites in mammalian proteins
- NetOGlyc - Prediction of O-GalNAc (mucin type) glycosylation sites in mammalian proteins
- NetGlycate - Glycation of epsilon amino groups of lysines in mammalian proteins
- NetNGlyc - Prediction of N-glycosylation sites in human proteins
- OGPET - Prediction of O-GalNAc (mucin-type) glycosylation sites in eukaryotic (non-protozoan) proteins
- YinOYang - O-beta-GlcNAc attachment sites in eukaryotic protein sequences

- big-PI Predictor - GPI Modification Site Prediction
- DGPI - Prediction of GPI-anchor and cleavage sites (Mirror site)
- GPI-SOM - Identification of GPI-anchor signals by a Kohonen Self Organizing Map
- Myristoylator - Prediction of N-terminal myristoylation by neural networks
- NMT - Prediction of N-terminal N-myristoylation
- CSS-Palm - Palmitoylation site prediction with CSS
- PrePS - Prenylation Prediction Suite

- NetAcet - Prediction of N-acetyltransferase A (NatA) substrates (in yeast and mammalian proteins)
- NetPhos - Prediction of Ser, Thr and Tyr phosphorylation sites in eukaryotic proteins
- NetPhosK - Kinase specific phosphorylation sites in eukaryotic proteins
- NetPhosYeast - Serine and threonine phosphorylation sites in yeast proteins
- GPS - Prediction of kinase-specific phosphorylation sites for 408 human protein kinases in hierarchy new
- Sulfinator - Prediction of tyrosine sulfation sites
- SulfoSite - Prediction of tyrosine sulfation sites
- SUMOplot - Prediction of SUMO protein attachment sites
- SUMOsp - Prediction of sumoylation sites new
- TermiNator - Prediction of N-terminal modification (version 3) new

- NetPicoRNA - Prediction of protease cleavage sites in picornaviral proteins
- NetCorona - Coronavirus 3C-like proteinase cleavage sites in proteins
- ProP - Arginine and lysine propeptide cleavage sites in eukaryotic protein sequences

28.7.2009

Take Home Message

Functional genetics & natural variation – powerful tool

Requirements for success

Strong QTL or Association

33% of natural variation in trichome number was explainable by a single aa replacement

Classical genetic studies failed to identify the major modifier of trichome number

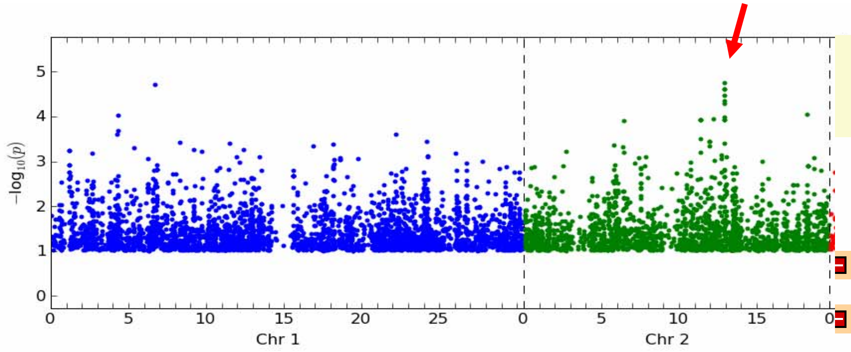
Typical accession used for functional tests (Ws, Col) have intermediate-high trichome number and the weak suppressor allele

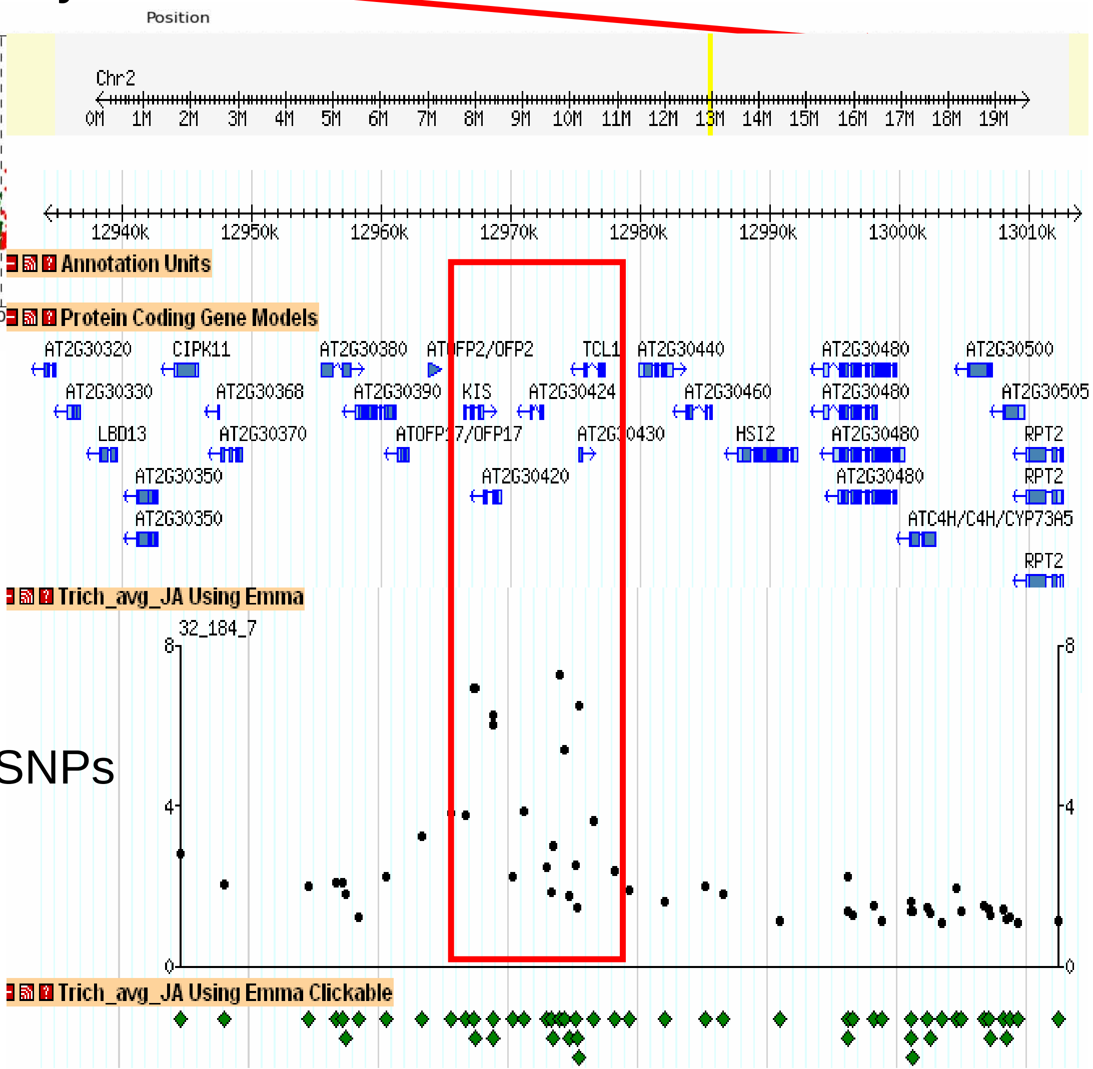
Association Mapping

Genomic polymorphism data in Arabidopsis thaliana

The Arabidopsis thaliana "HapMap" project

trichome density



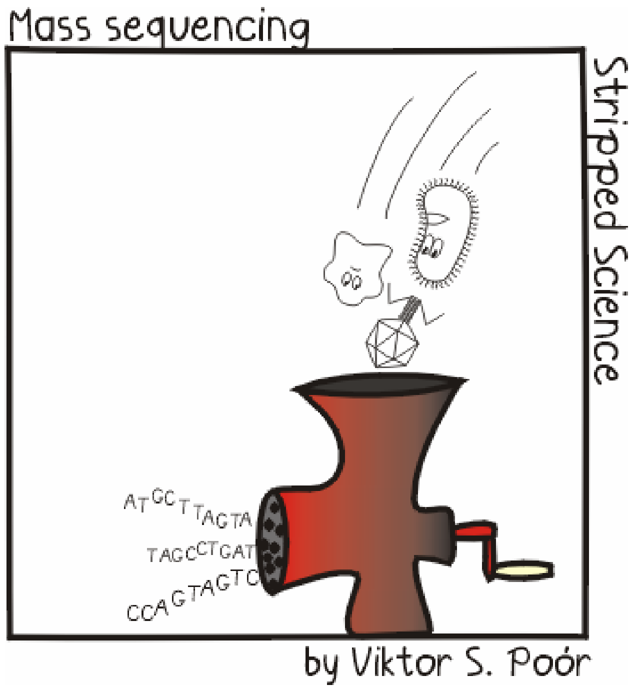


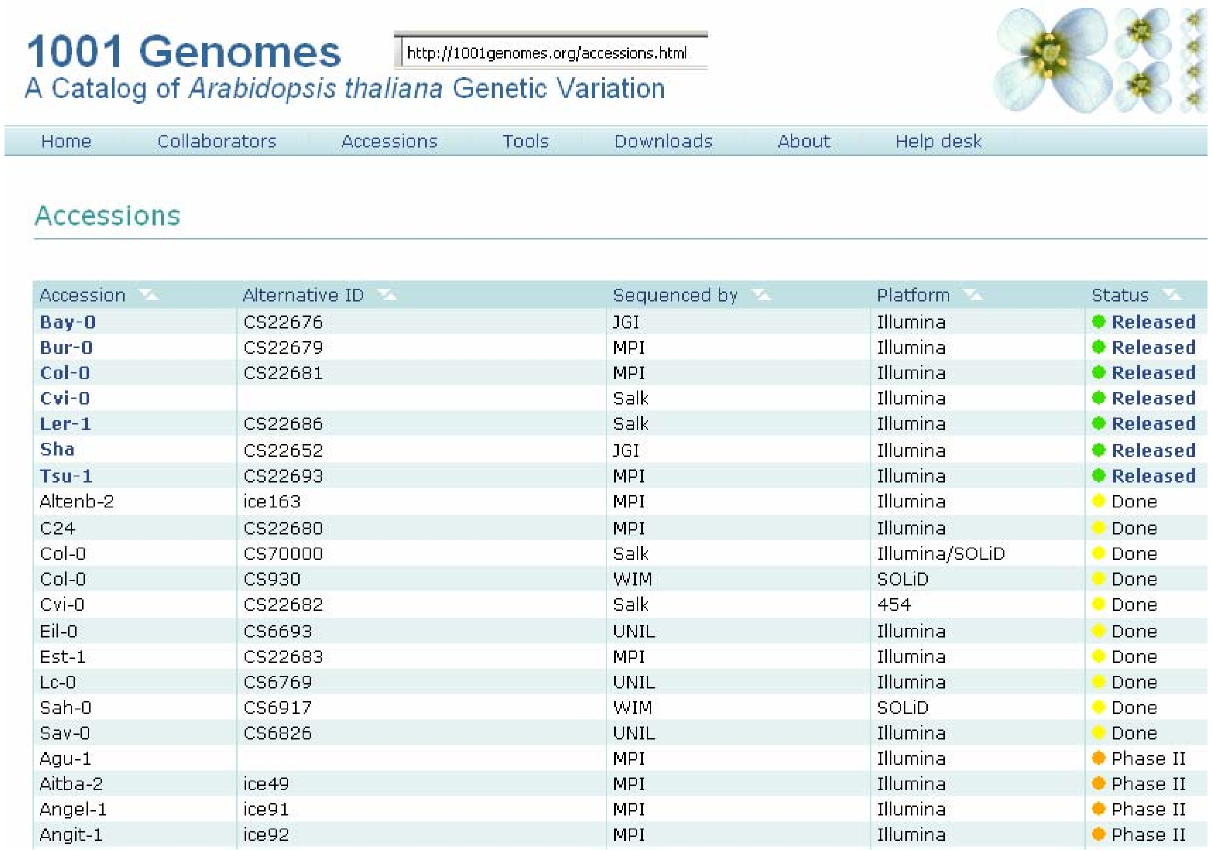
In future

– phenotyping ecotypes/accession

- look up into the association map
- identify the gene with the variable SNPs
- functional confirmation

More sequences * More functions * More to compare





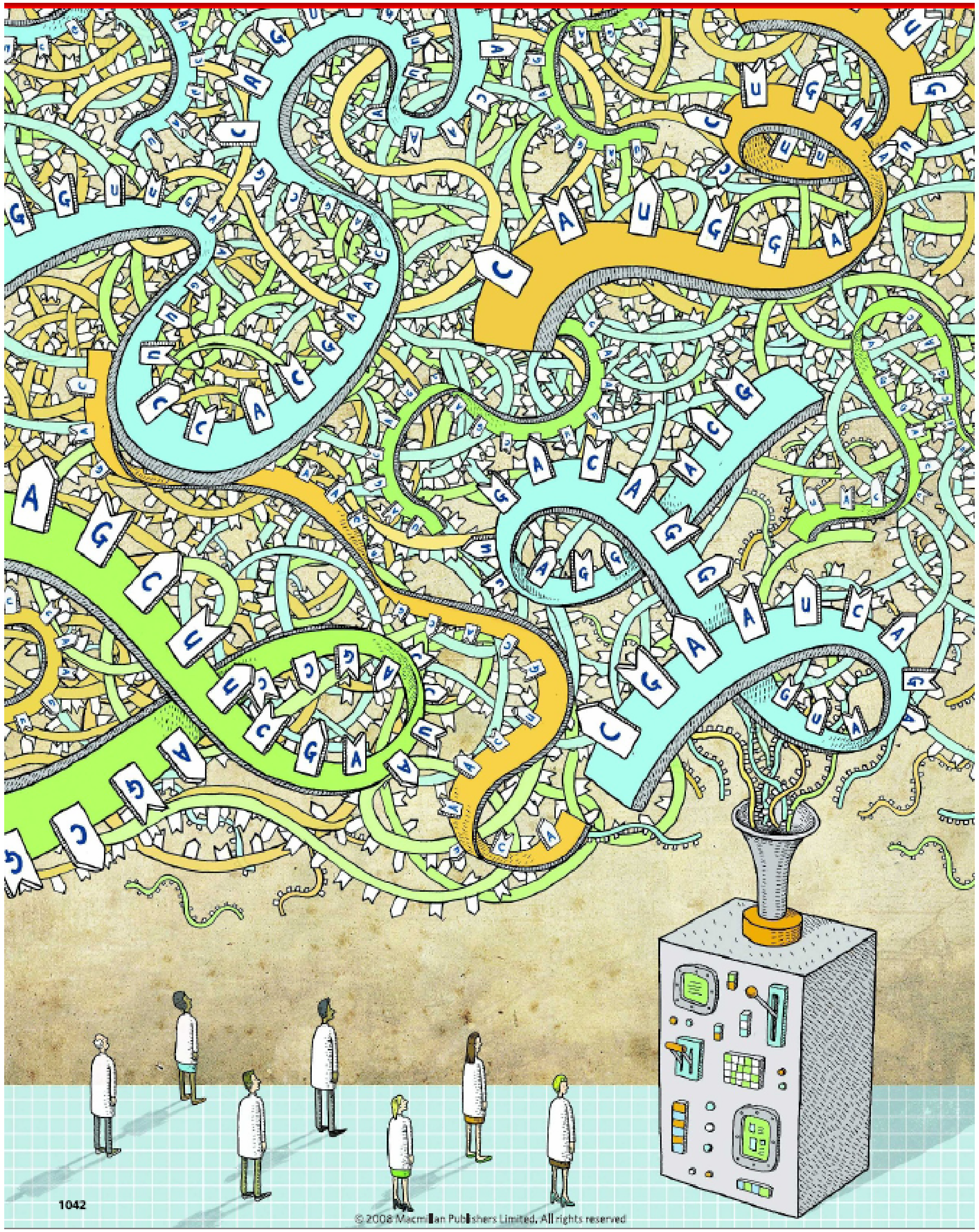
1001 Genomes

A Catalog of *Arabidopsis thaliana* Genetic Variation

Accessions

Accession	Alternative ID	Sequenced by	Platform	Status
Bay-0	CS22676	JGI	Illumina	● **Released**
Bur-0	CS22679	MPI	Illumina	● **Released**
Col-0	CS22681	MPI	Illumina	● **Released**
Cvi-0		Salk	Illumina	● **Released**
Ler-1	CS22686	Salk	Illumina	● **Released**
Sha	CS22652	JGI	Illumina	● **Released**
Tsu-1	CS22693	MPI	Illumina	● **Released**
Altenb-2	ice163	MPI	Illumina	● Done
C24	CS22680	MPI	Illumina	● Done
Col-0	CS70000	Salk	Illumina/SOLiD	● Done
Col-0	CS930	WIM	SOLiD	● Done
Cvi-0	CS22682	Salk	454	● Done
Eil-0	CS6693	UNIL	Illumina	● Done
Est-1	CS22683	MPI	Illumina	● Done
Lc-0	CS6769	UNIL	Illumina	● Done
Sah-0	CS6917	WIM	SOLiD	● Done
Sav-0	CS6826	UNIL	Illumina	● Done
Agu-1		MPI	Illumina	● Phase II
Aitba-2	ice49	MPI	Illumina	● Phase II
Angel-1	ice91	MPI	Illumina	● Phase II
Angit-1	ice92	MPI	Illumina	● Phase II

28.7.2009



THANKS

QUESTIONS?

28.7.2009