Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations

Michiel Vandenbussche[1], Günter Theissen[2], Yves Van de Peer[1] and Tom Gerats[1,3].

[1]Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium

[2]Lehrstuhl for Genetics, Friedrich Schiller University of Jena, Philosophenweg 12, D-07743 Jena, Germany

[3]present address: Department of Experimental Botany, University of Nijmegen, Toernooiveld 1, 6525E, Nijmegen, The Netherlands

Supporting Material

Contents:

Assembling the MADS-box sequence dataset

To extract the MIKC type MADS-box sequence dataset from the public databases, we have used the following approach: In a first step, one representative protein sequence of each known MADS-box subfamily (according to recently published phylogenetic analyses, see for example reference 1) was chosen. The full-length protein sequence of each selected subfamily member was used to search the public databases for homologous sequences, using the BLAST program (http://www.ncbi.nlm.nih.gov/BLAST/). The protein sequence dataset was built using the blastp option to search the available protein databases, while the nucleotide dataset was assembled using tblastn to search the available nucleotide databases. To ensure that even distantly related MADS-box genes would be retrieved in the homology searches, the standard settings of some search parameters of the BLAST program were changed. The expectancy value was raised to 100, and the number of homologous sequences to be displayed was increased to 500. Homologies shown in the output BLAST page were inspected visually, and MADS-box sequences were retrieved by using the sequence retrieval option and saved in batches as text files in Genbank format. These text files were subsequently imported in the GCG program using the FromGenbank function (Wisconsin Package Version 10.0, Genetics Computer Group (GCG), Madison, Wisc.). Because of the low stringency settings of the homology searches with the different subfamily representatives, a large proportion of the collected sequences was identified multiple times; such duplicates were removed automatically during import in GCG using the 'remove duplicates' option.

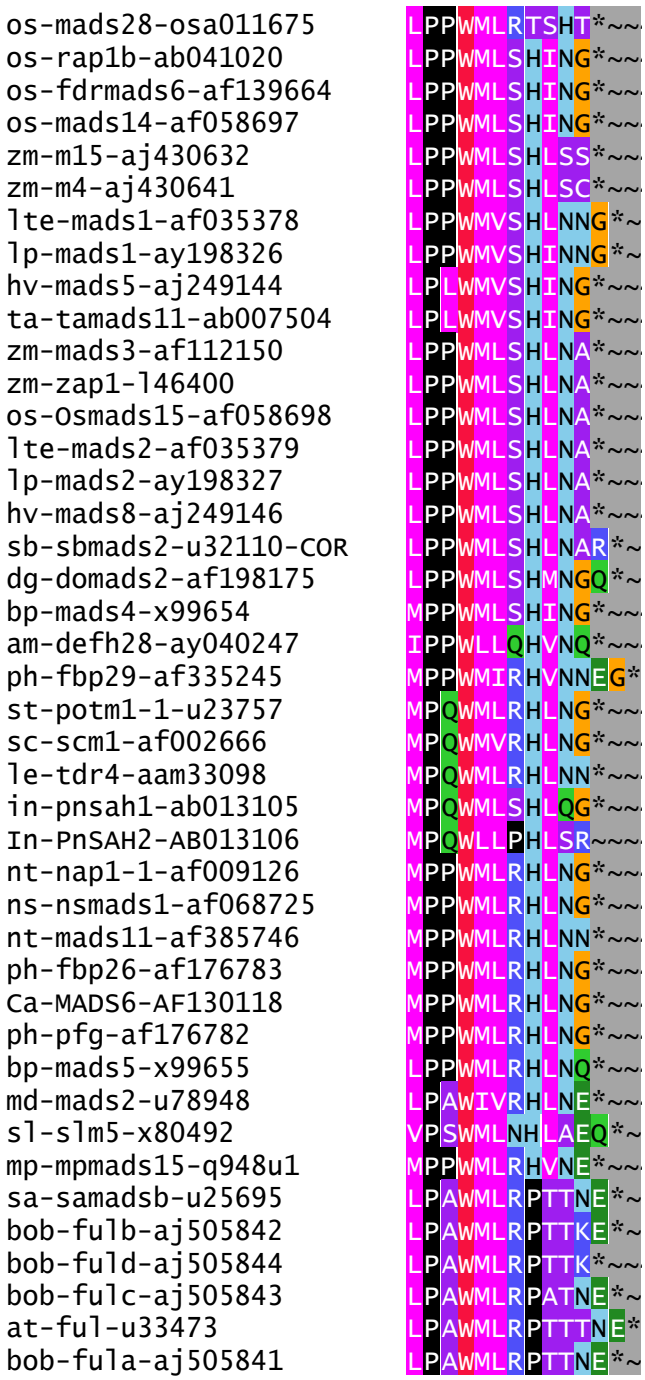
To yield a more comprehensive overview of the taxonomic distribution within subfamilies, the sequence identifiers were renamed manually as follows: they all start with a two- or three-letter code indicating the species name (see Species Table below for abbreviations), followed by the gene name as found in the database and terminating with the database Accession Number (according to either DNA, EST, Protein or Patent databases). Full-length sequences were aligned using the PILEUP function, followed by a manual alignment of the C-terminal regions using the Seqlab Editor of the GCG software package. To illustrate conservation of the C-terminal conserved motifs, we have displayed the alignments of the C-terminal motifs per subfamily. C-terminal motifs of sequences of which the sequence name terminates with 'COR' (corrected) have been identified in other reading frames than published in the database and/or beyond the proposed stopcodon. An asterisk behind a peptide motif indicates a stopcodon; motifs not terminating with an asterisk are followed by a variable number of less-conserved residues (not shown).

Species Table

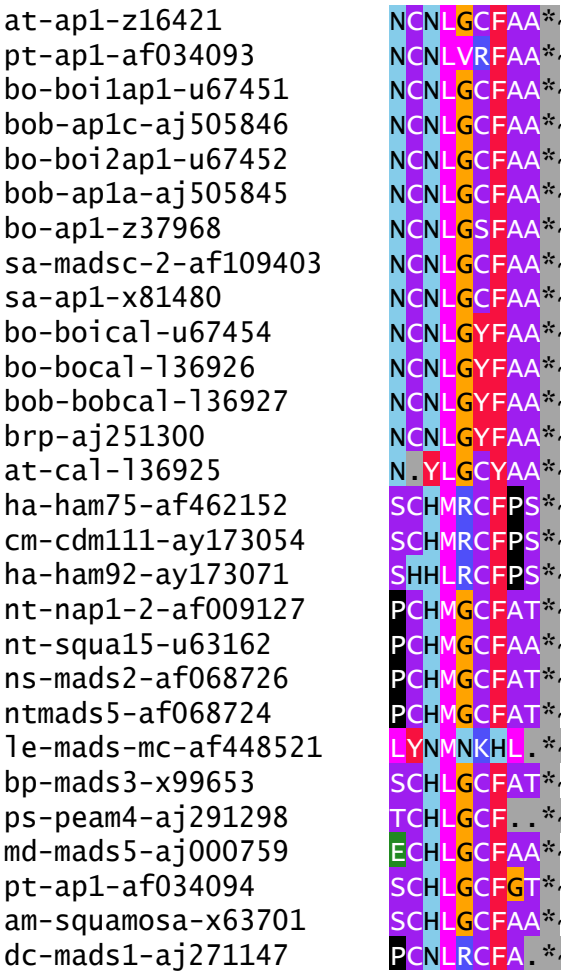
Species	code	Species	code
Agapanthus praecox	Ap	*Lolium temulentum*	Lte
Akebia quinata	Aq	*Lycopersicon esculentum*	Le
Anemone nemorosa	An	*Lycopodium annotinum*	La
Antirrhinum majus	Am	*Magnolia praecocissima*	Mp
Aquilegia alpina	Aa	*Malus domestica*	Md
Aquilegia caerulea	Ac	*Medicago sativa*	Ms
Arabidopsis lyrata	Al	*Michelia figo*	Mf
Arabidopsis thaliana	At	*Momordica charantia*	Mc
Aranda deborah	Ad	*Nicotinia sylvestris*	Ns
Asarum europaeum	Ae	*Nicotinia tabacum*	Nt
Berberis gilgiana	Bg	*Oncidium cv.*	Oc
Betula pendula	Bp	*Orchis italica*	Oi
Brassica napus	Bn	*Oryza sativa*	Os
Brassica olereacea	Bo	*Pachysandra terminalis*	Pt
Brassica olereacea var botrytis	Bob	*Panax ginseng*	Pg
Brassica rapa	Br	*Papaver californicum*	Pc
Calycanthus floridus	Cf	*Papaver nudicaule*	Pn
Canavalia lineata	Cl	*Paulownia kawakamii*	Pk
Capsicum annuum	Ca	*Peperomia hirta*	Phir
Ceratopteris richardii	Cr	*Petunia hybrida*	Ph
Chloranthus spicatus	Cs	*Petunia inflata*	Pi
Chrysanthemum x morifolium	Cm	*Petunia integrifolia*	Pin
Cimicifuga racemosa	Cra	*Phalaenopsis equestris*	Pe
Clematis chiisanensis	Cc	*Physcomitrella patens*	Pp
Clematis integrifolia	Ci	*Picea abies*	Pa
Corylus avellana	Cav	*Picea marinea*	Pm
Cryptomeria japonica	Cj	*Pimpinella brachycarpa*	Pb
Cucumis sativa	Cus	*Pinus radiata*	Pr
Cycas edentata	Ce	*Pinus resinosa*	Pres
Daucus carota	Dc	*Piper magnificum*	Pmag
Delphinium ajacis	Da	*Pisum sativum*	Ps
Dendrobium grex	Dg	*Platanus occidentalis*	Po
Dicentra eximia	De	*Poa annua*	Pan
Elaeis guineensis	Eg	*Populus balsamifera*	Pb
Eucalyptus globulus	Egl	*Populus tomentosa*	Pto
Eucalyptus grandis	Eug	*Populus tremuloides*	Pt
Fragaria x ananassa	Fa	*Ranunculus bulbosus*	Rb
Gerbera hybrida	Gh	*Ranunculus ficaria*	Rf
Glycine max	Gm	*Rosa rugosa*	Rr
Gnetum gnemon	Gg	*Rosa x hybrida*	Rh
Gnetum parvifolium	Gp	*Rumex acetosa*	Ra
Gossypium hirsutum	Ghi	*Sagittaria montevidensis*	Sm
Helianthus annuus	Ha	*Sanguinaria canadensis*	Sc
Helleborus orientalis	Hor	*Saururus chinensis*	Sch
Hemerocallis hybrid	Hh	*Silene latifolia*	Sl
Hieracium piloselloides	Hp	*Sinapis alba*	Sa
Hordeum vulgare	Hv	*Solanum tuberosum*	St
Hyacinthus orientalis	Ho	*Sorghum bicolor*	Sb
Hydrangea macrophylla	Hm	*Syringa vulgaris*	Sv
Ipomoea batatas	Ib	*Tacca chantieri*	Tc
Ipomoea nil	In	*Thalictrum thalictroides*	Tt
Juglans regia	Jr	*Trautvetteria carolinensis*	Tca
Lilium longiflorum	Ll	*Triticum aestivum*	Ta
Lilium regale	Lr	*Trollius laxus*	Tl
Liquidambar styraciflua	Ls	*Vitis vinifera*	Vv
Liriodendron tulipifera	Lt	*Zea mays*	Zm
Lolium perenne	Lp		

SQUAMOSA/AP1 SUBFAMILY

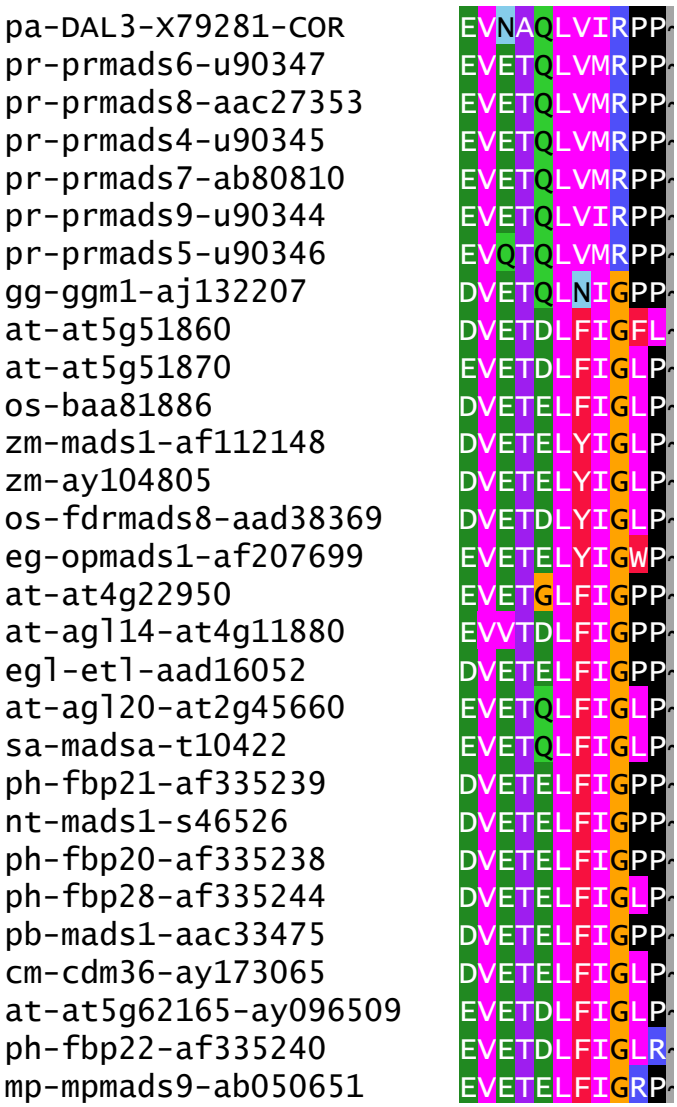
Paleo*AP1* Lineage



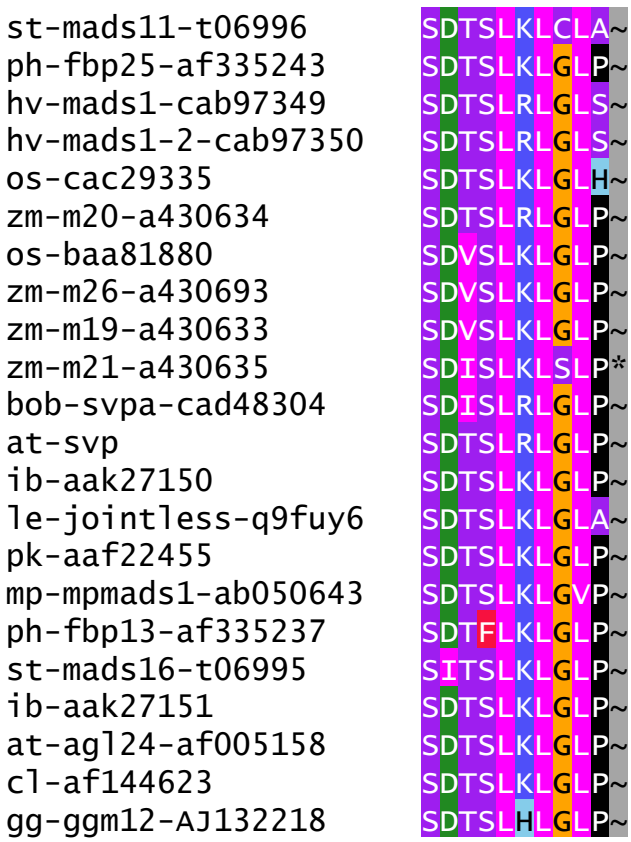
Eu*AP1* Lineage



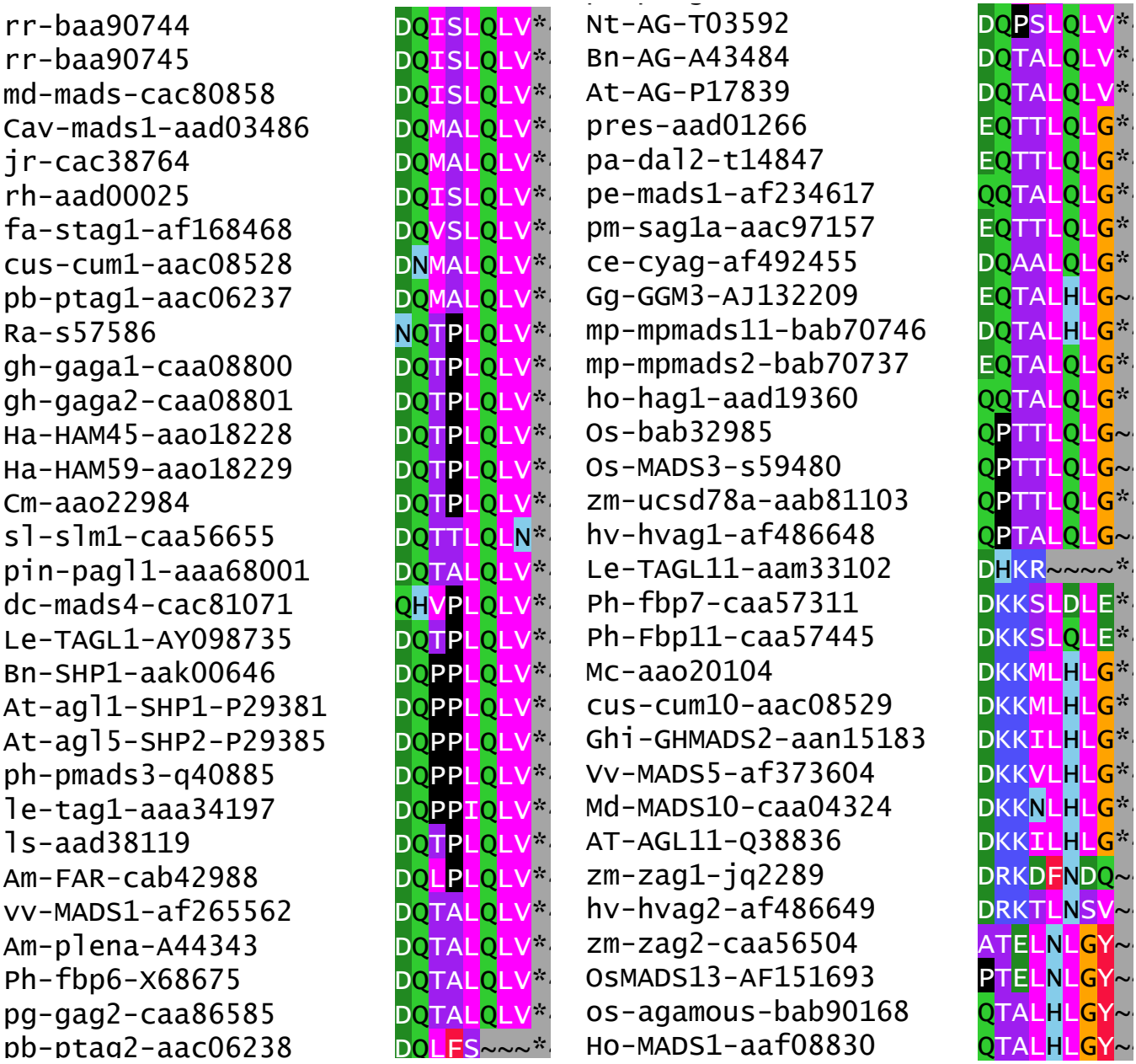
TM3 SUBFAMILY



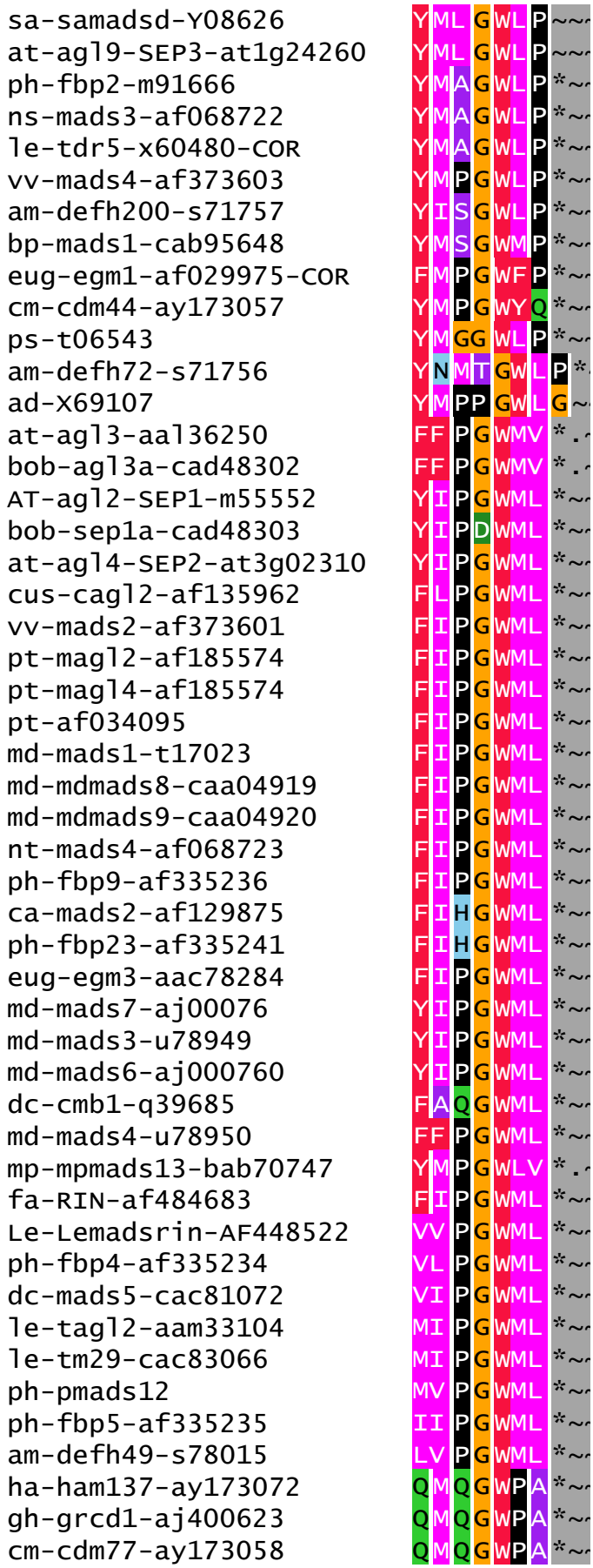
STMADS11 SUBFAMILY



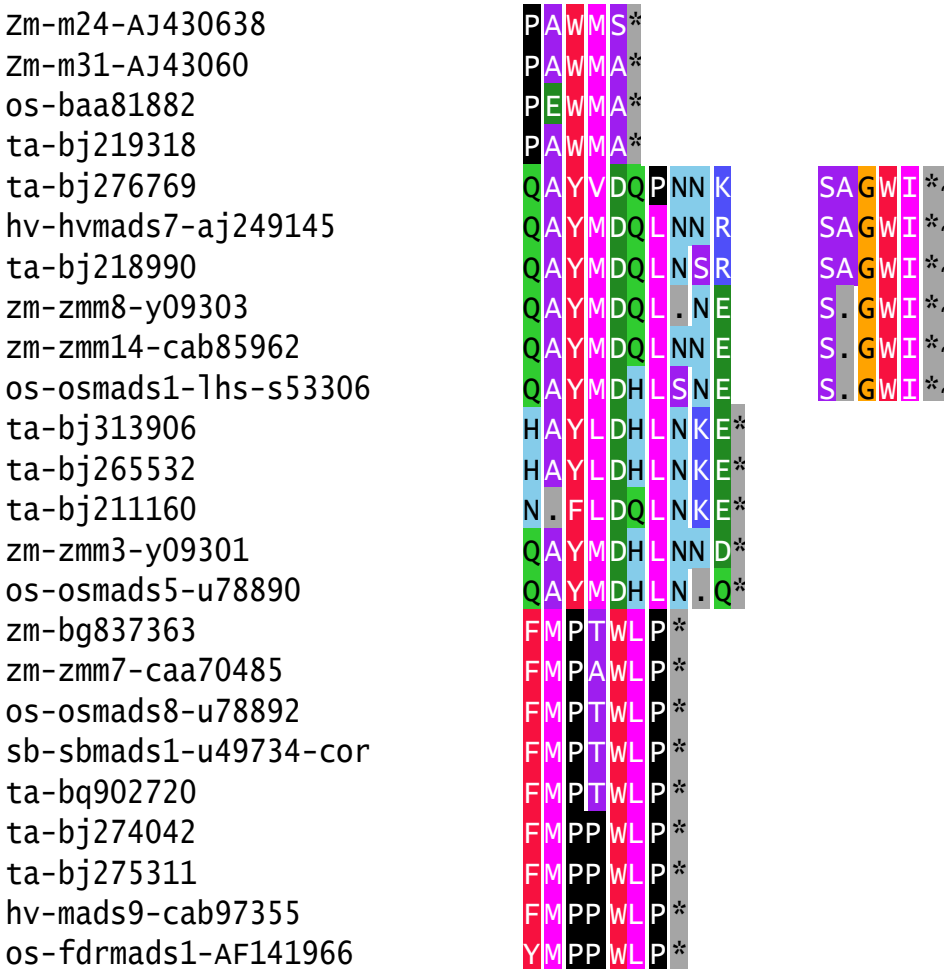
AGAMOUS SUBFAMILY



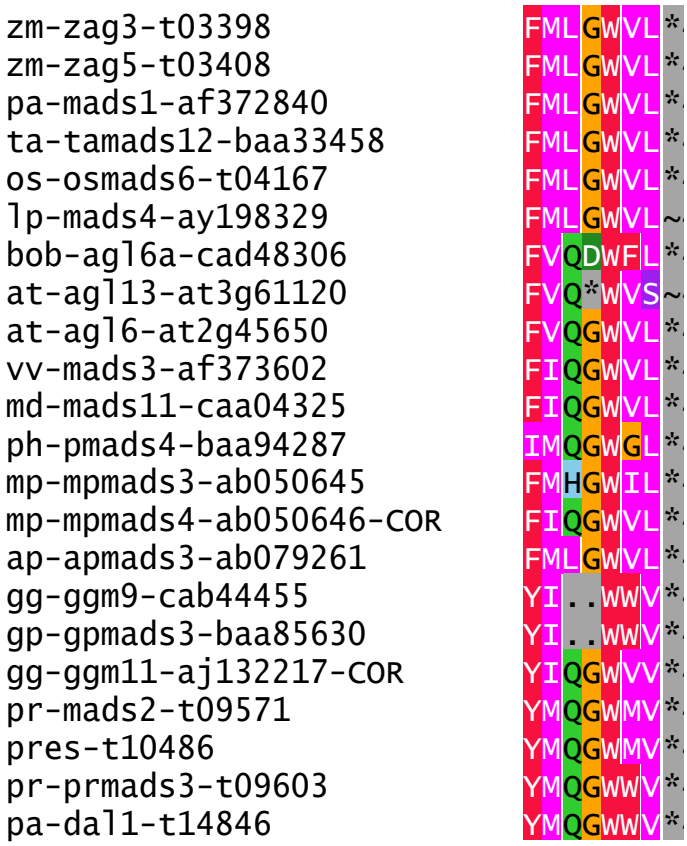
AGL2 SUBFAMILY PART I



AGL2 SUBFAMILY PART II (monocotyledons)

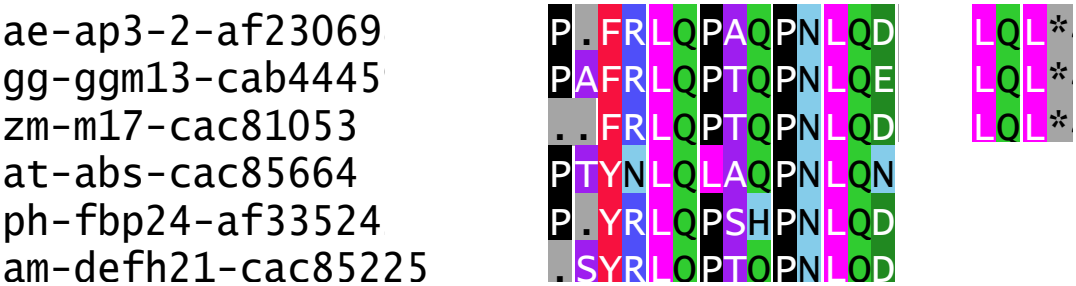


AGL6 SUBFAMILY



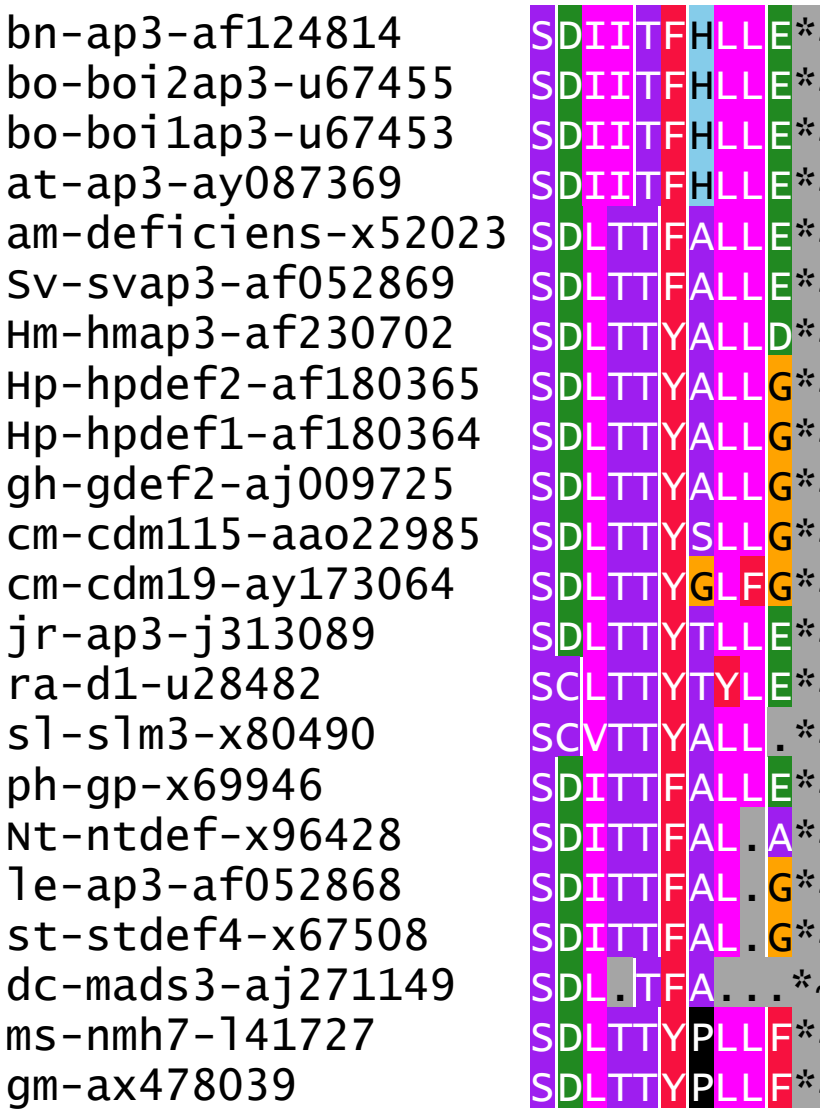
GLO/PI SUBFAMILY

Bs SUBFAMILY



DEF/AP3 SUBFAMILY

Eu*AP3* Lineage



DEF/AP3 SUBFAMILY

Paleo*AP3* Lineage