Evolution of sequence-diverse disordered regions in a protein family: order within the chaos

SUPPLEMENTARY INFORMATION | DOI 10.26181/5E33788AD3B54

Supplementary Figures

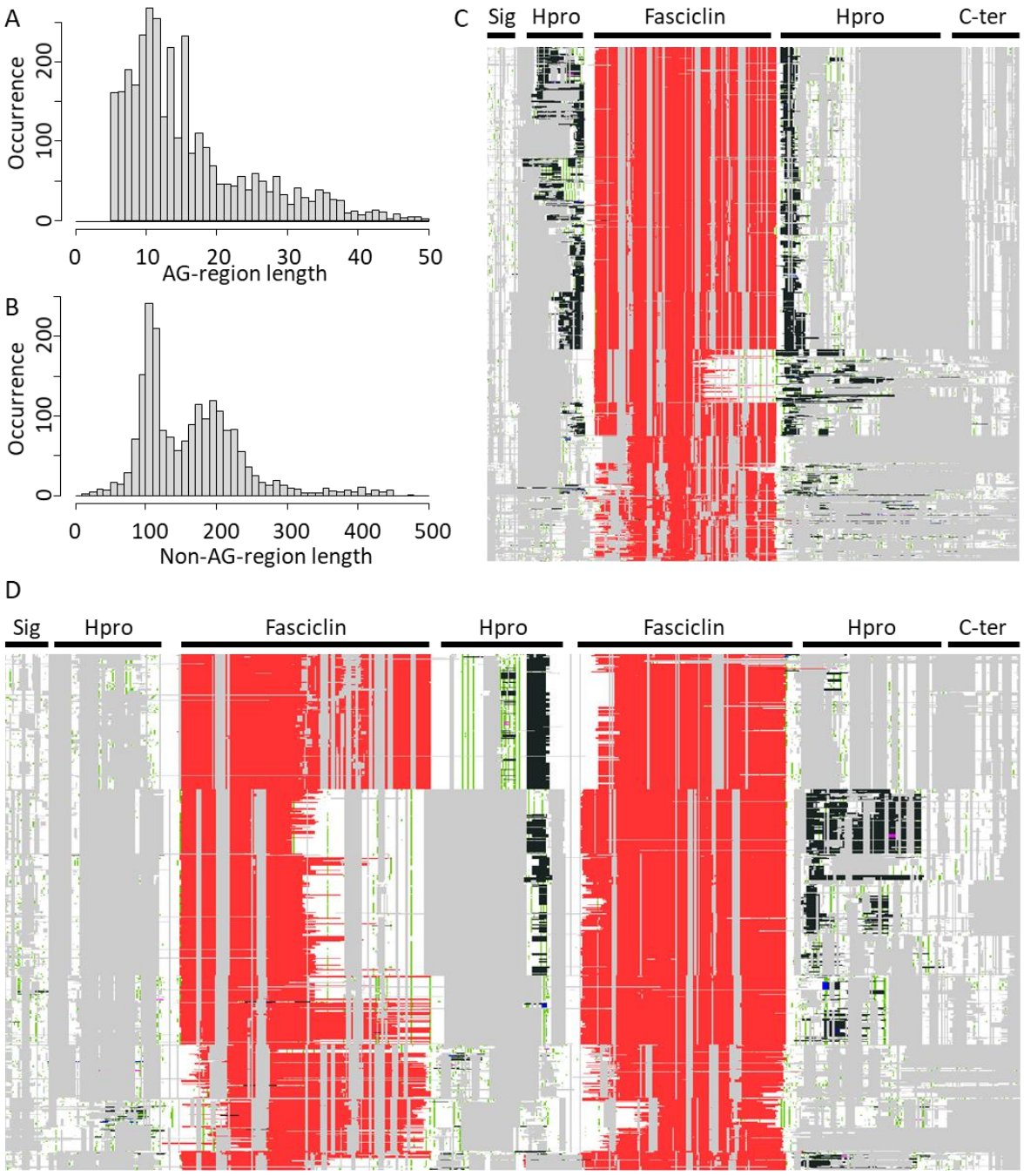
Supplementary Tables

Supplementary Data Files

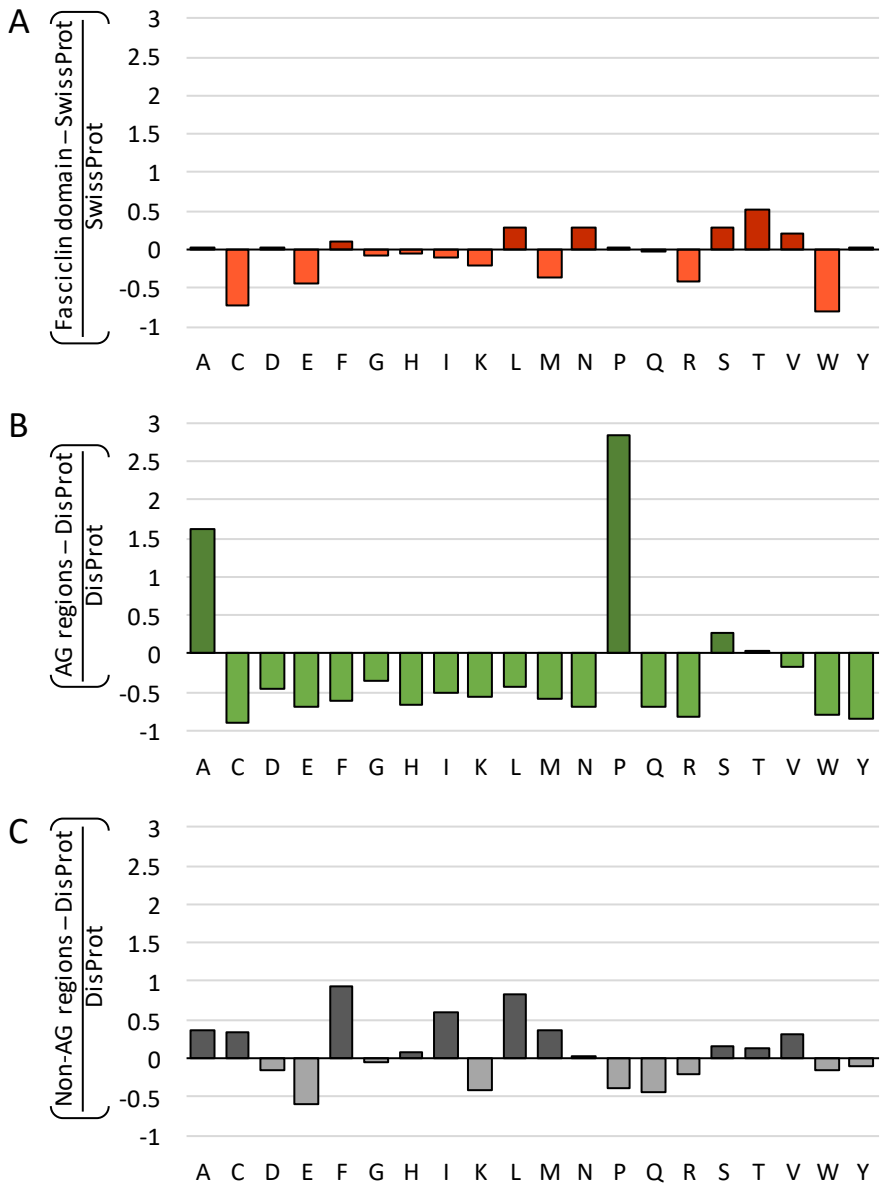
Supplementary Figures

FLA architecture	Taxa	Species in which each domain structure is found		
L-O-X	Alga	*Dunaliella salina x3*		
G-A-P	Monocot	*Setaria viridis*	*Panicum hallii x2*	
G-A-A	Rosid & monocot	*Populus trichocarpa*	*Salix purpurea*	*Zea mays*
Q-X-O	Alga	*Chlamydomonas reinhardtii*	*Volvox carteri*	
P-P-P-P	Alga	*Dunaliella salina x2*		
P-P-H	Rosid	*Anacardium occidentale*	*Linum usitatissimum*	
X-P-J	Superrosid	*Kalanchoe laxiflora*		
X-O-O-O-X-O-O	Alga	*Chlamydomonas reinhardtii*		
R-R-H	Asterid	*Mimulus guttatus*		
R-O-X	Alga	*Micromonas pusilla*		
R-O-P	Alga	*Micromonas sp. RCC299*		
R-H-P	Monocot	*Brachypodium stacei*		
Q-Q-Q-Q-Q	Alga	*Porphyra umbilicalis*		
Q-Q-Q-Q	Alga	*Chlamydomonas reinhardtii*		
Q-Q-Q	Alga	*Chlamydomonas reinhardtii*		
Q-Q-P	Alga	*Porphyra umbilicalis*		
Q-O-O	Alga	*Volvox carteri*		
Q-J-J	Basal eudicot	*Aquilegia coerulea*		
P-R-H	Asterid	*Olea europaea*		
P-P-P-X-X	Alga	*Chlamydomonas reinhardtii*		
P-P-O	Alga	*Coccomyxa subellipsoidea*		
P-O-O-P	Alga	*Ostreococcus lucimarinus*		
P-I-P	Rosid	*Medicago truncatula*		
O-X-Q-P	Alga	*Dunaliella salina*		
O-R-O	Alga	*Micromonas pusilla*		
O-O-P	Alga	*Chlamydomonas reinhardtii*		
O-O-O-P	Alga	*Chromochloris zofingiensis*		
O-K-P-P	Alga	*Chromochloris zofingiensis*		
L-P-P-O-O-P-P	Alga	*Volvox carteri*		
F-P-P	Rosid	*Trifolium pratense*		
F-F-F	Rosid	*Prunus persica*		
F-D-P	Rosid	*Manihot esculenta*		
E-C-X	Rosid	*Brassica rapa*		

Supp Figure S1 | Occurrence of FLAs with ≥3 fasciclin domains.
FLAs with ≥3 fasciclin domains occurring in algae in yellow, those in vascular plants in grey with species name indicated. FLA architecture indicates the fasciclin domain type (A-R, Figure 3) as defined by sequence space clustering, X indicates fasciclin domains not assignable to a cluster**.** The arrangement and number of the fasciclin domain types within the FLA are indicated.

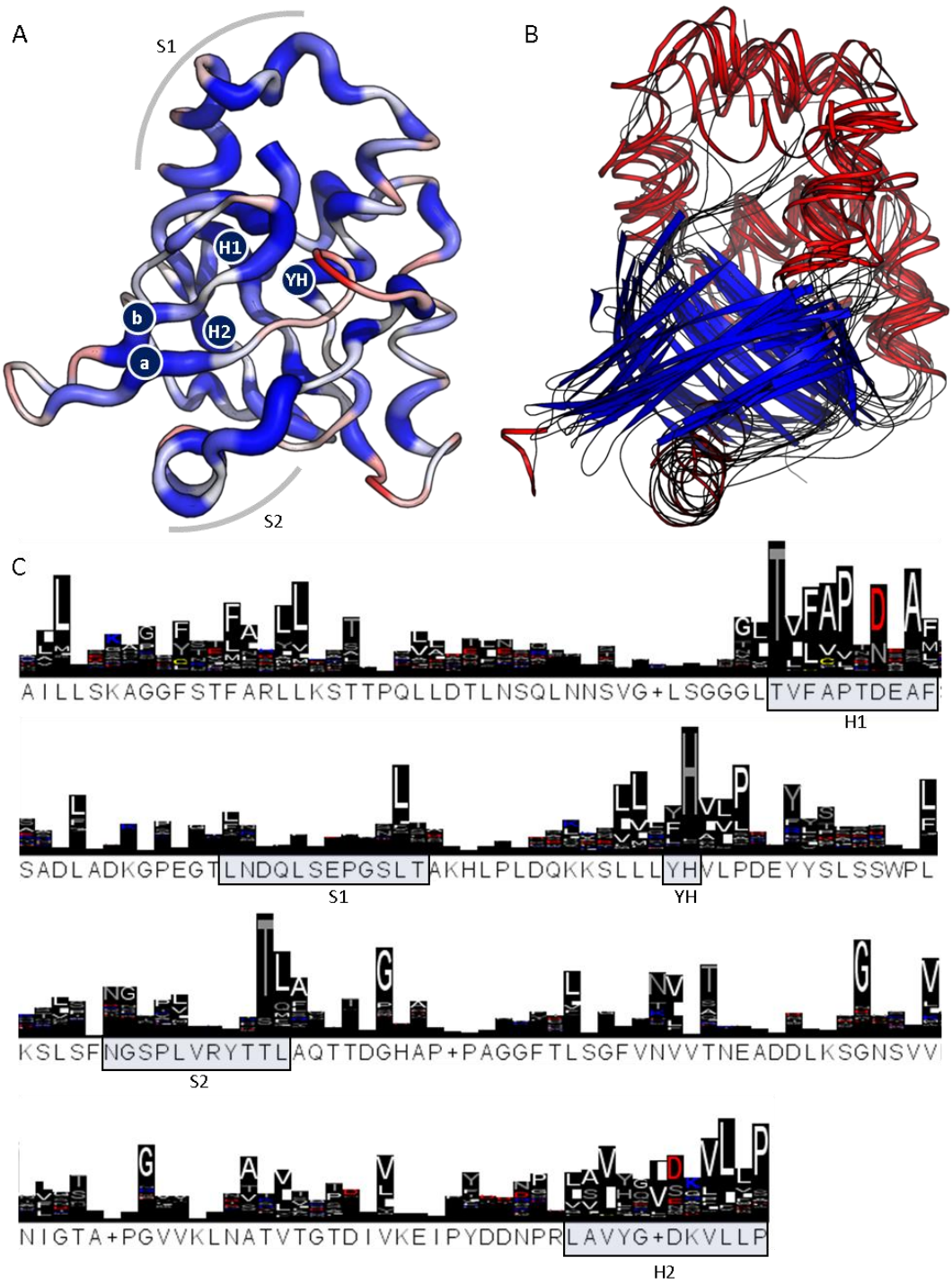


Supp Figure S2 | Length of AG and non-AG regions and multiple sequence alignment for FLAs.
A) Length distribution in amino acids for AG regions. **B)** Length distribution in amino acids for non-AG disordered regions. FLAs containing **C)** a single fasciclin domain (998 sequences) or **D)** two fasciclin domains (806 sequences). Regions that match the Pfam fasciclin (PF02469) HMM in red, AG-regions in black (mainly restricted to the columns labelled 'Hpro'), PRP regions in purple, extensin regions in blue, other prolines in green, all other residues in white, and gaps in grey. Some sequences with particularly long N- or C-terminal disordered regions have been trimmed to fit. Note: there are <100 PRP and extensin motifs in panels C and D. Note: despite the presentation of regions outside of the fasciclin domains in the alignment, the certainty of these regions is very low.

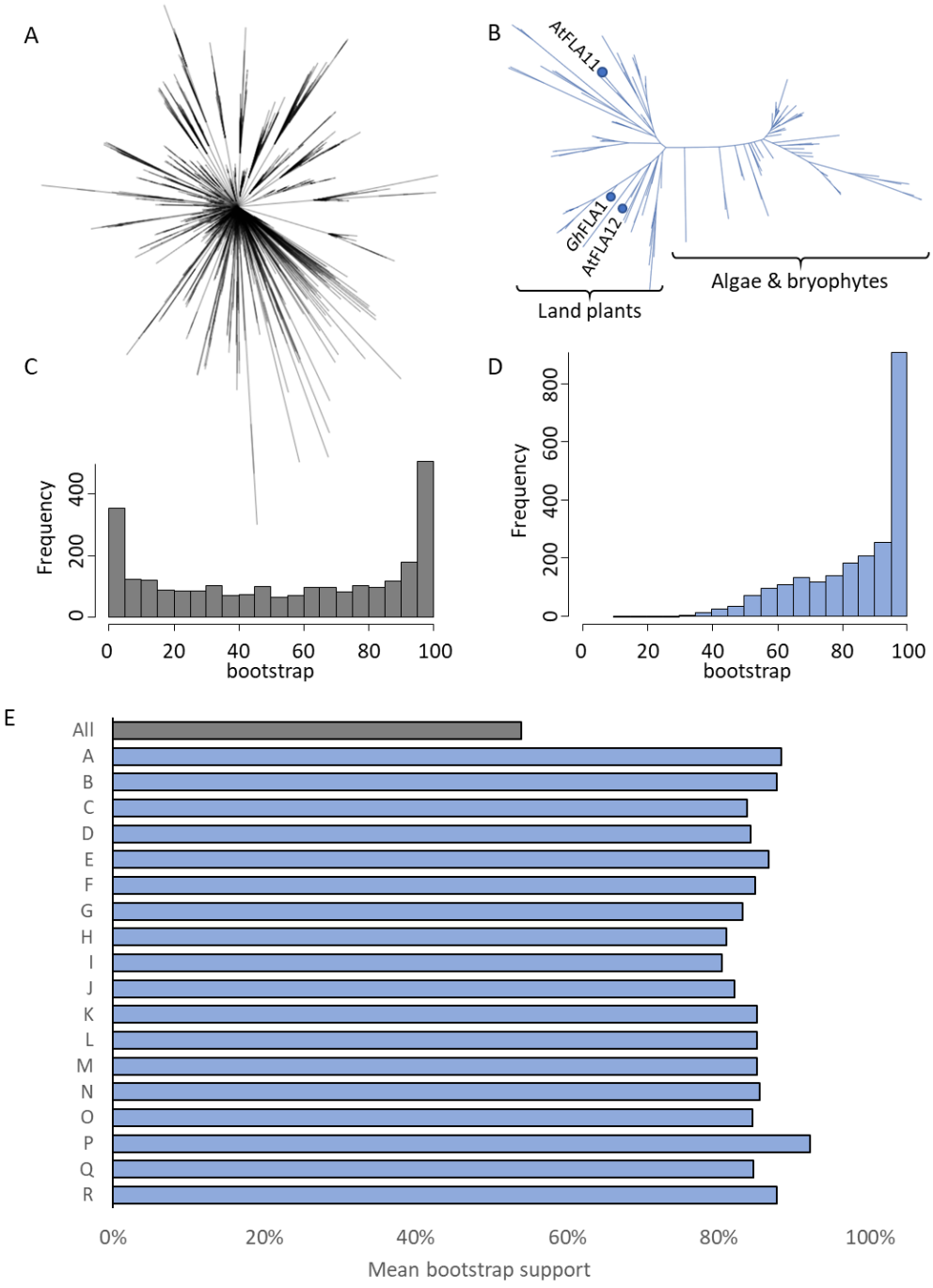


Supp Figure S3 | Amino acid composition benchmarked to relevant databases.

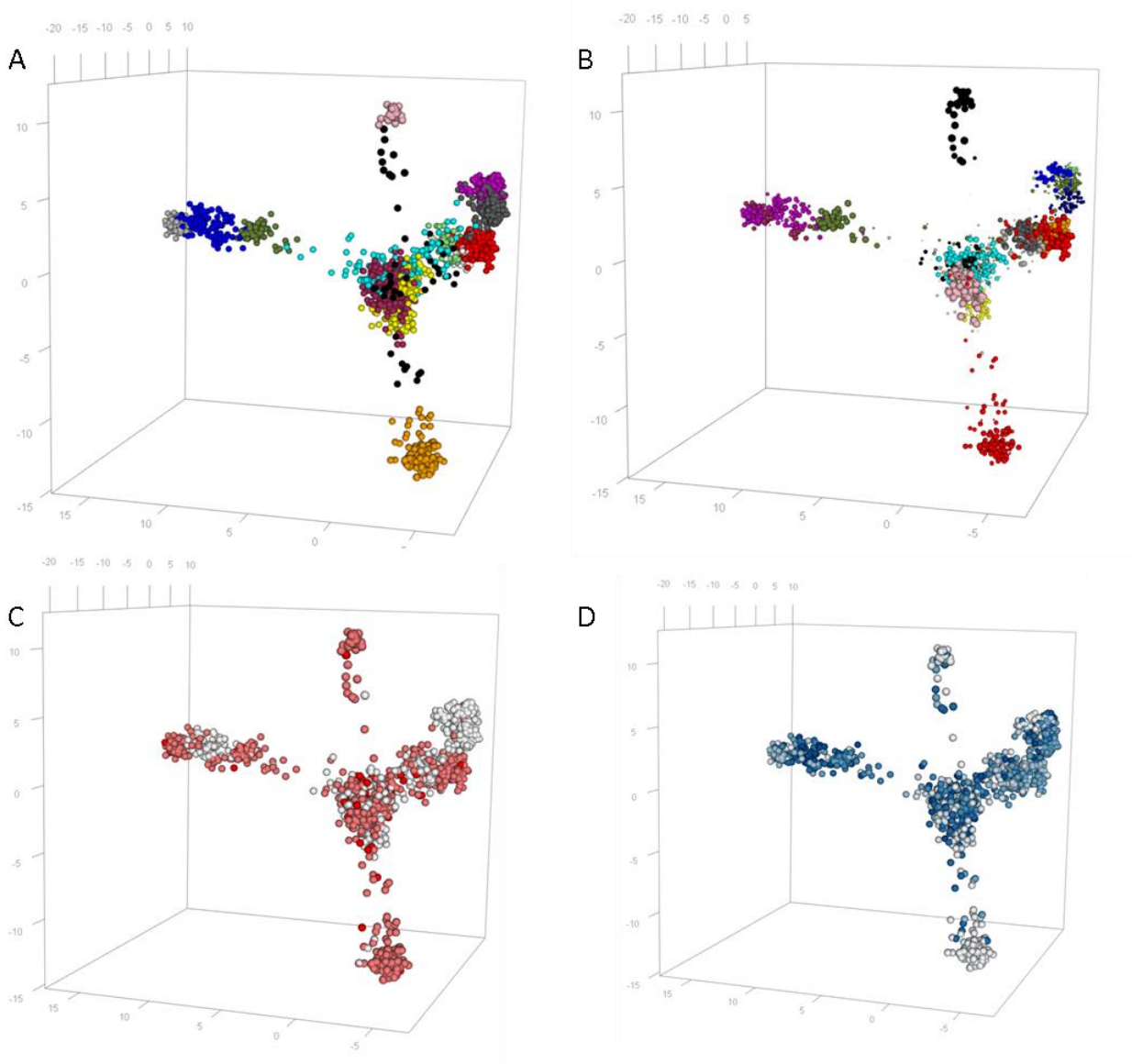
Relative amino acid composition ratios for **A)** the fasciclin domains as compared to SwissProt 51, **B)** the AG regions as compared to DisProt 3.4, and **C)** the non-AG regions as compared to DisProt 3.4.



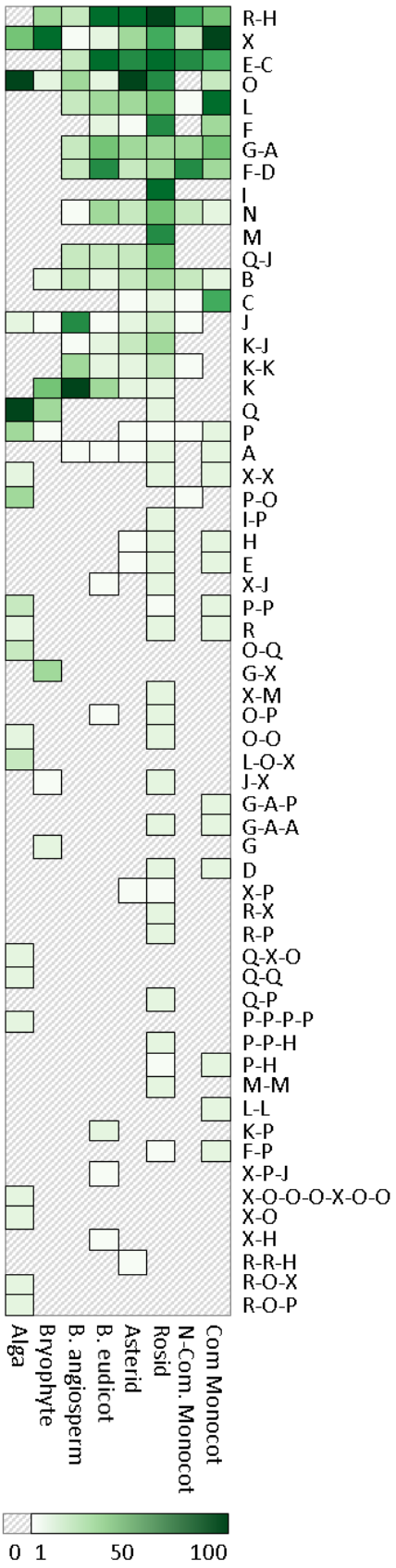
Supp Figure S4 | Conservation of fasciclin domain sequence and structure.
A) Sequence conservation across all 2644 FLA fasciclin domains in the Phytozome dataset is indicated by colour and width, mapped onto a representative homology model of the second fasciclin domain (based on PDB:1o70) from *At*FLA1 (as per Figure 2A). Conserved residues in blue, variable residues in red. The two most-conserved *N*-glycosylation sites are indicated as "a" and "b" as in Figure 2A. The highly conserved H1, YH and H2 motifs characteristic of fasciclin domains are buried in the core of the structure as indicated. **B)** Structure conservation as an overlay of all structurally characterised fasciclin domains (PDBs: 1o70; 5nv6; 1nyo; 5wt7; 1w7d; 2mxa). Beta strands in blue, alpha helices in red. **C)** Conservation sequence logo across 2644 FLA fasciclin domains.



Supp Figure S5 | Bootstrap support for fasciclin domain phylogenies.
A) Maximum likelihood phylogeny of all fasciclin domain sequences. Nodes with <50% bootstrap support collapsed, making the tree uninterpretable (45% of nodes have bootstrap support <50%, with particularly low support for deep nodes). **B)** Maximum likelihood phylogeny of fasciclin domain sequences of type O. Nodes with <50% bootstrap support collapsed (demonstrating higher overall bootstrap support). FLAs with characterised function annotated. A more detailed phylogeny of FLA sequences with fasciclin domain type O is shown in Supp Figure S14. **C)** Histogram of the bootstrap values for nodes in the phylogeny of all fasciclin domain sequences (panel A). **D)** Histogram of the bootstrap values for nodes in the phylogenies of all fasciclin domain sequence sets separated by type. **E**) Mean bootstrap values across all nodes in each of the phylogenies: all fasciclin domains (panel A), or the fasciclin domain sets separated by types from Figure 3A.

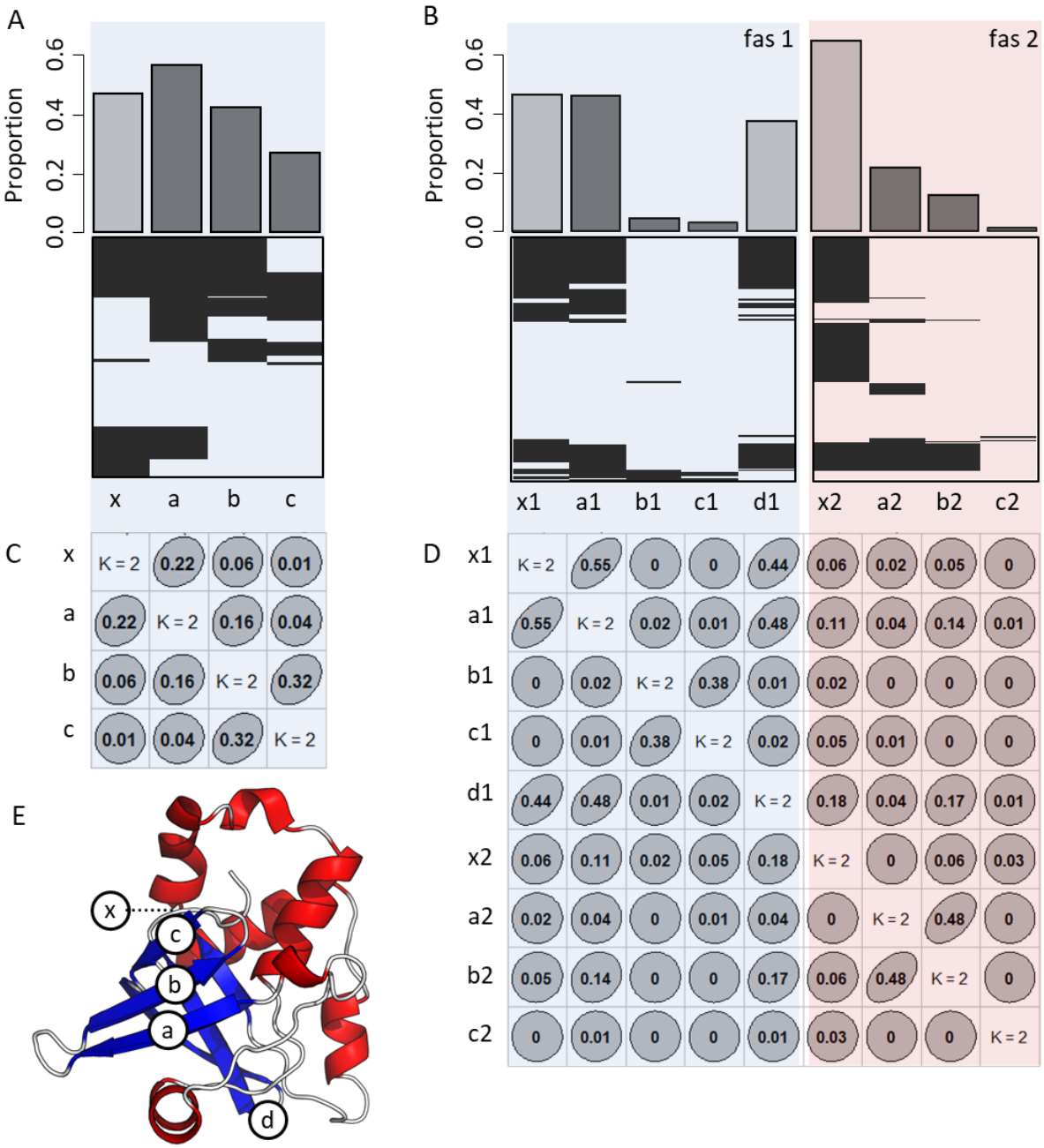


Supp Figure S6 | Sequence space of fasciclin domains, projected by PCA.
For each fasciclin domain, sequence biophysical properties projected by PCA. Coloured based on **A)** Bayesian clustering of PCA projection, **B)** HDBSCAN clustering of UMAP projection (cluster colours as per Figure 2A), **C)** number of fasciclin domains in the FLA, **D)** inter-proline distance.

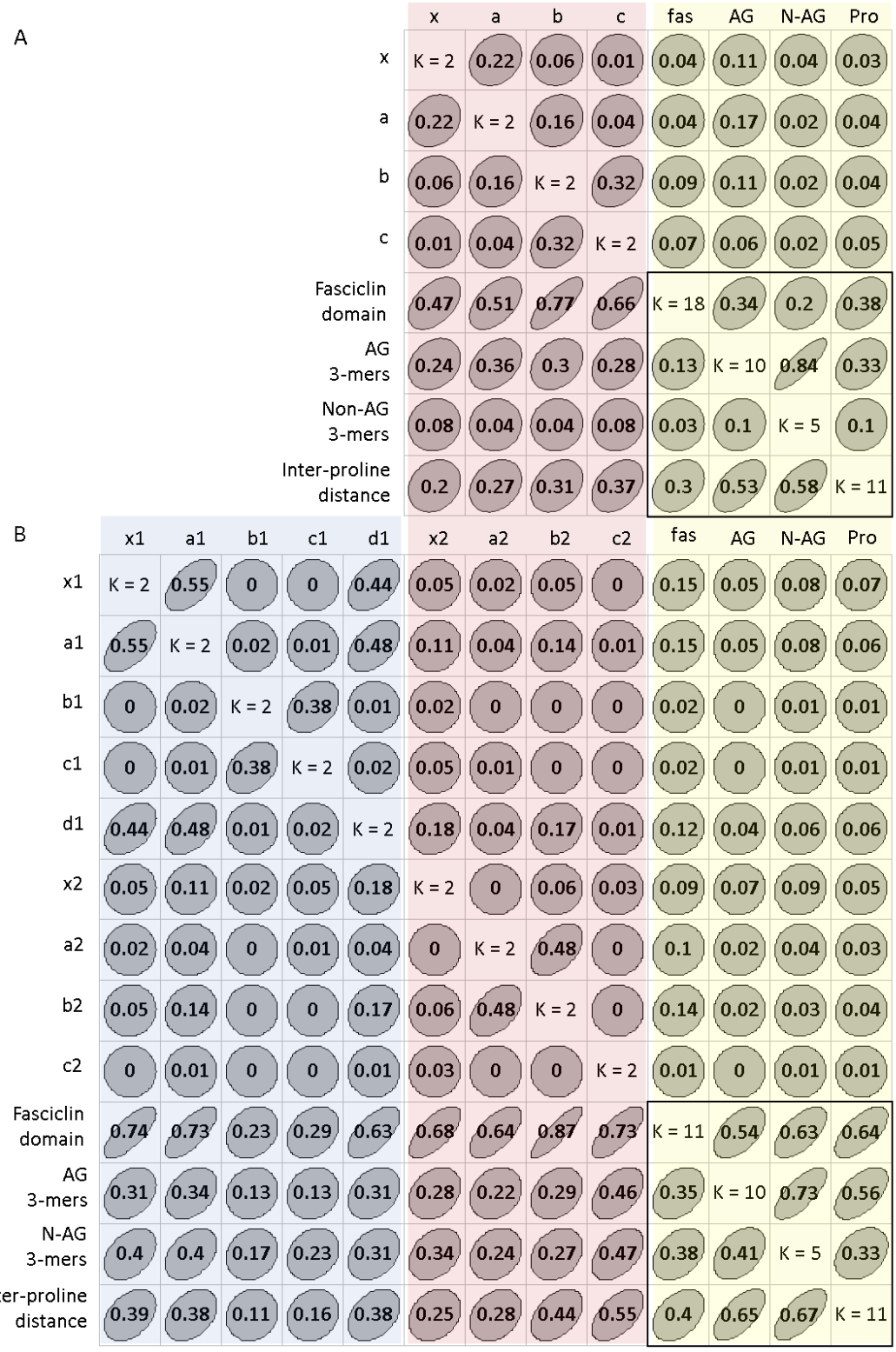


Supp Figure S7 | Fasciclin domain architectures across different taxonomic groups.

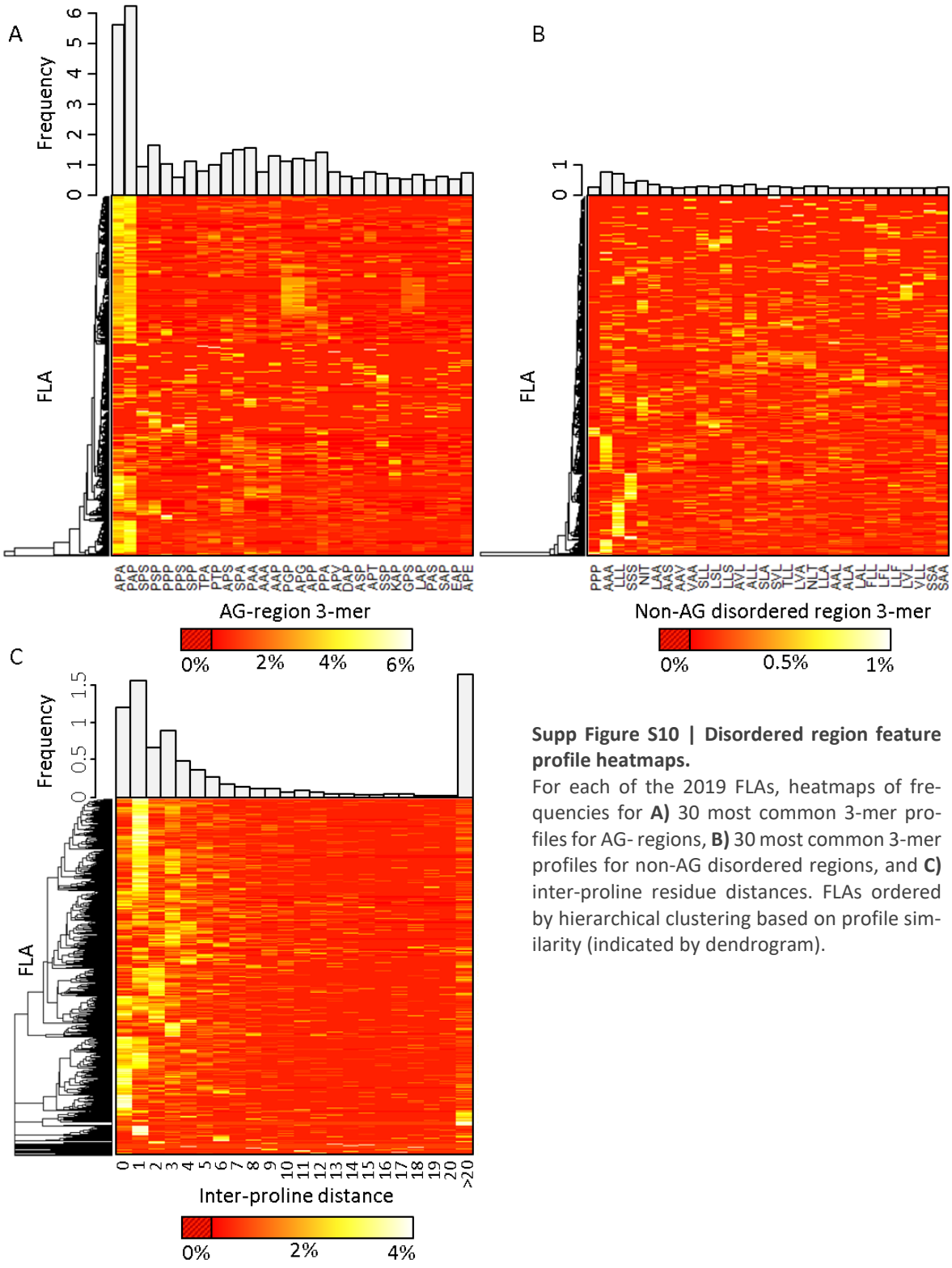
Relative proportion of domain architectures in different major taxonomic groups for the 60 most common domain combinations ordered by overall frequency of occurrence of each domain architecture.



Supp Figure S8 | *N*-glycosylation sites of 1-fas and 2-fas FLAs.
A) Presence of *N*-glycosylation sites for 1-fas FLAs. Overall proportion as bar chart above (darker bars indicate sites in the β-sheet characterised as *N*-glycosylated in other fasciclin domains) and presence/absence as heatmap below, where rows have been hierarchically clustered to indicate relative co-occurrence at each site. **B)** as above, but for 2-fas FLAs. **C)** All-vs-all matrix of Goodman and Kruskal's tau correlation measure for all 1-fas FLA sites. **D)** as above but for 2-fas FLAs. **E)** Locations of possible *N*-glycosylation sites in a fasciclin domain (model as in Figure 2A).

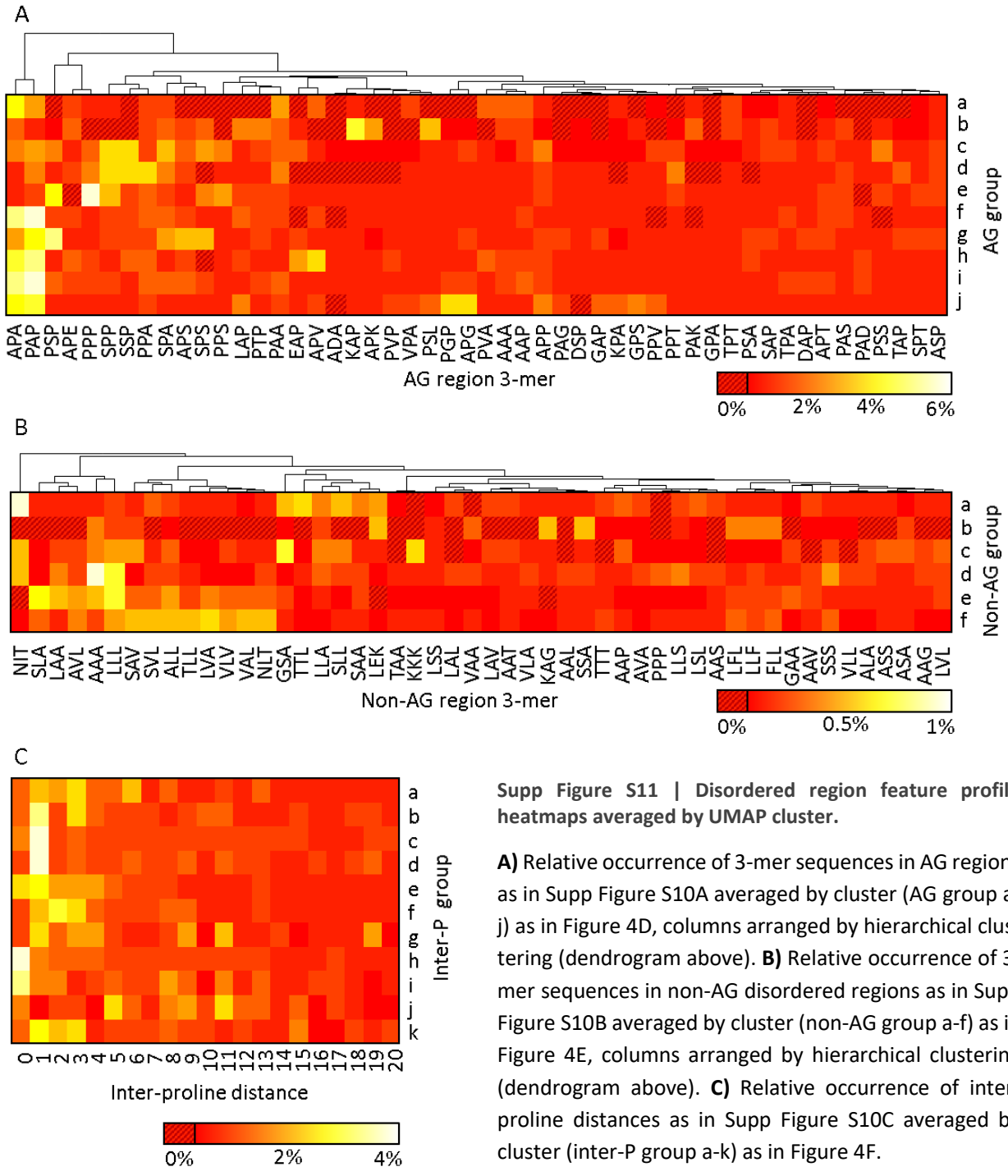


Supp Figure S9 | All-vs-all matrix of Goodman and Kruskal's tau feature correlation measure
Correlation of *N*-glycosylation sites with fasciclin domain type, non-AG region cluster, AG-region cluster and inter-proline distance cluster for all **A)** 1-fas FLAs, **B)** 2-fas FLAs. Coloured with *N*-glycosylation sites of the first domain in blue, *N*-glycosylation sites of the second domain in red, and UMAP clusters in yellow. Black boxes highlight the sections equivalent to Figure 4G.



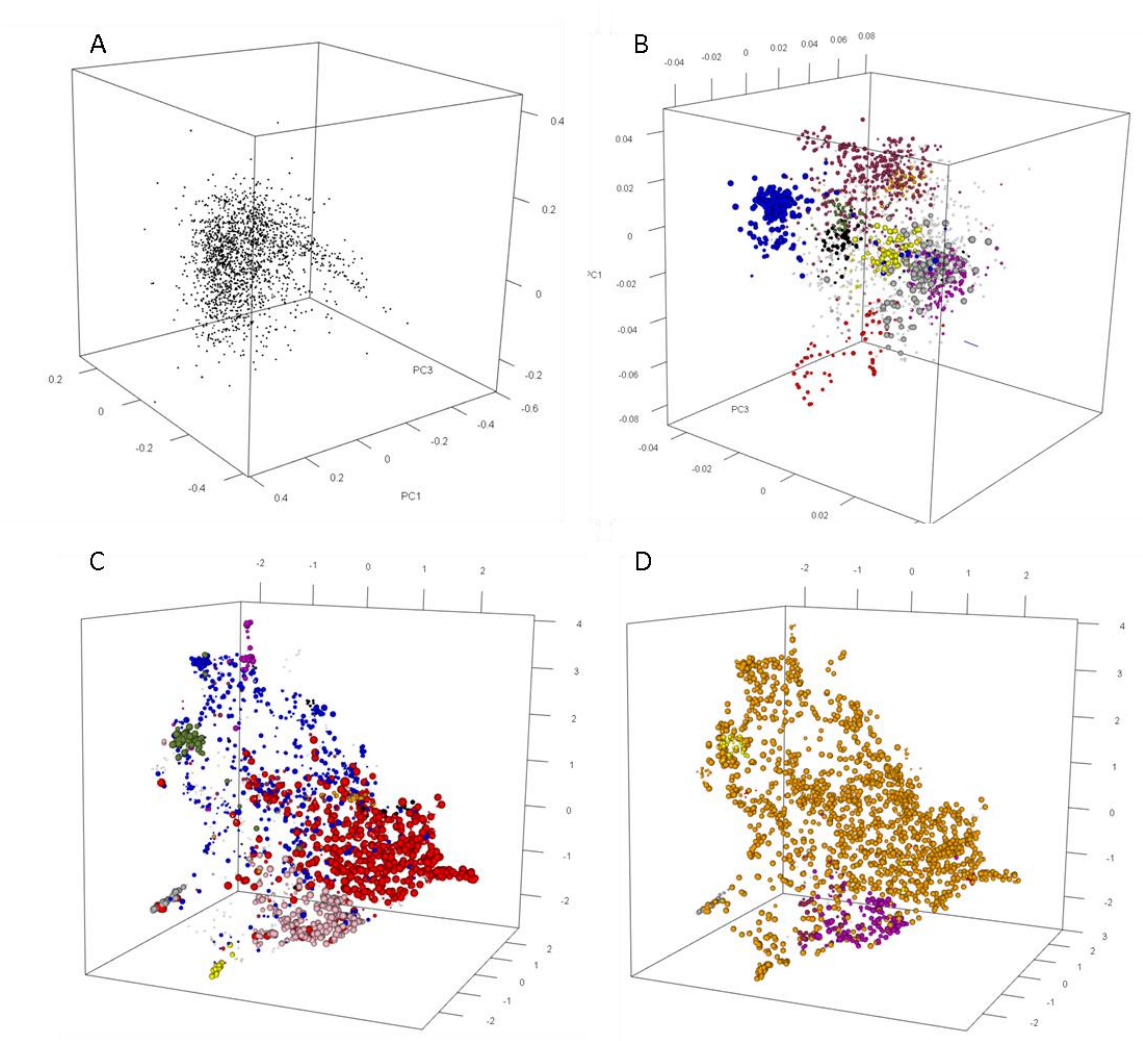
Supp Figure S10 | Disordered region feature profile heatmaps.

For each of the 2019 FLAs, heatmaps of frequencies for **A)** 30 most common 3-mer profiles for AG- regions, **B)** 30 most common 3-mer profiles for non-AG disordered regions, and **C)** inter-proline residue distances. FLAs ordered by hierarchical clustering based on profile similarity (indicated by dendrogram).

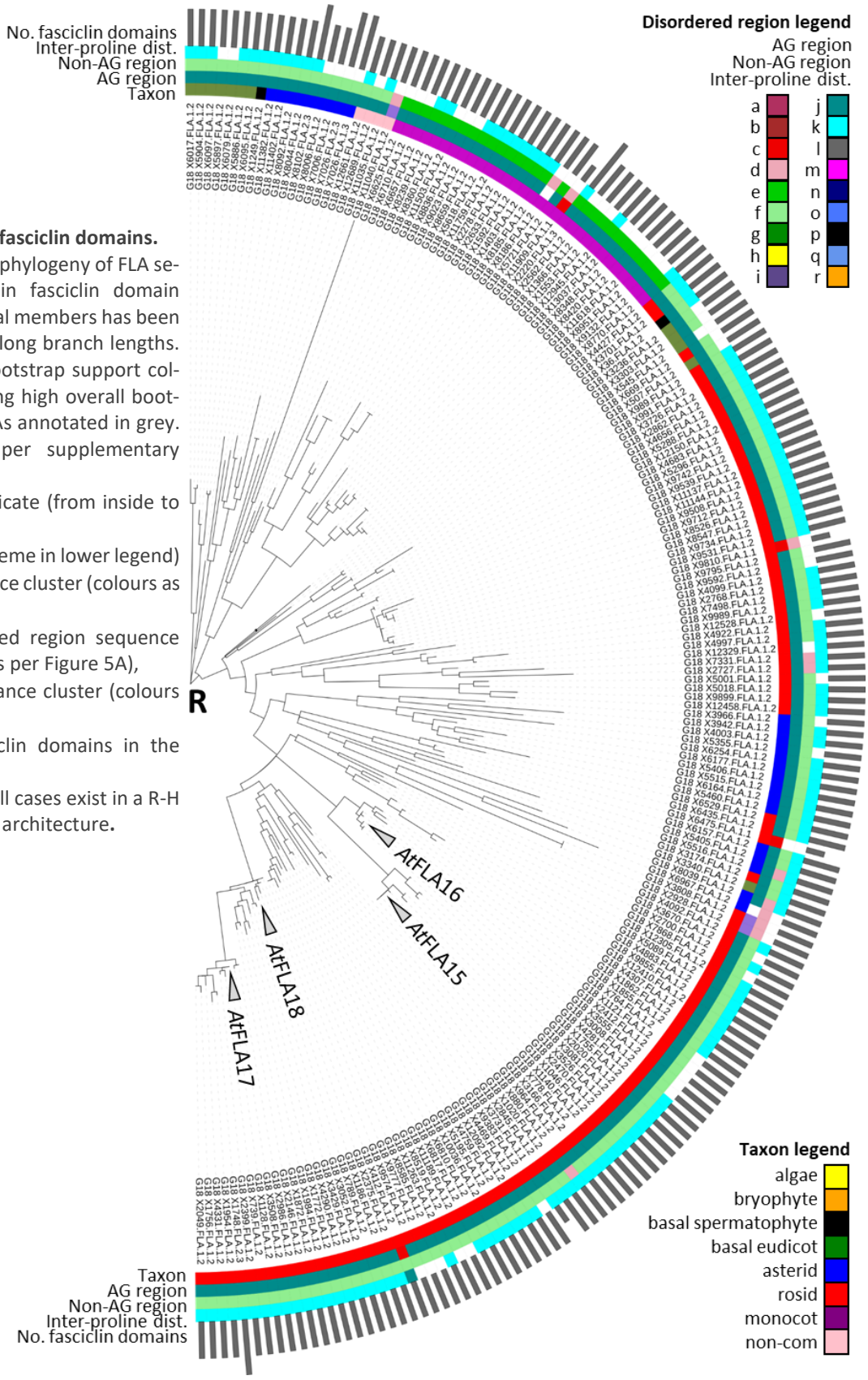


Supp Figure S11 | Disordered region feature profile heatmaps averaged by UMAP cluster.

A) Relative occurrence of 3-mer sequences in AG regions as in Supp Figure S10A averaged by cluster (AG group a-j) as in Figure 4D, columns arranged by hierarchical clustering (dendrogram above). **B)** Relative occurrence of 3-mer sequences in non-AG disordered regions as in Supp Figure S10B averaged by cluster (non-AG group a-f) as in Figure 4E, columns arranged by hierarchical clustering (dendrogram above). **C)** Relative occurrence of inter-proline distances as in Supp Figure S10C averaged by cluster (inter-P group a-k) as in Figure 4F.



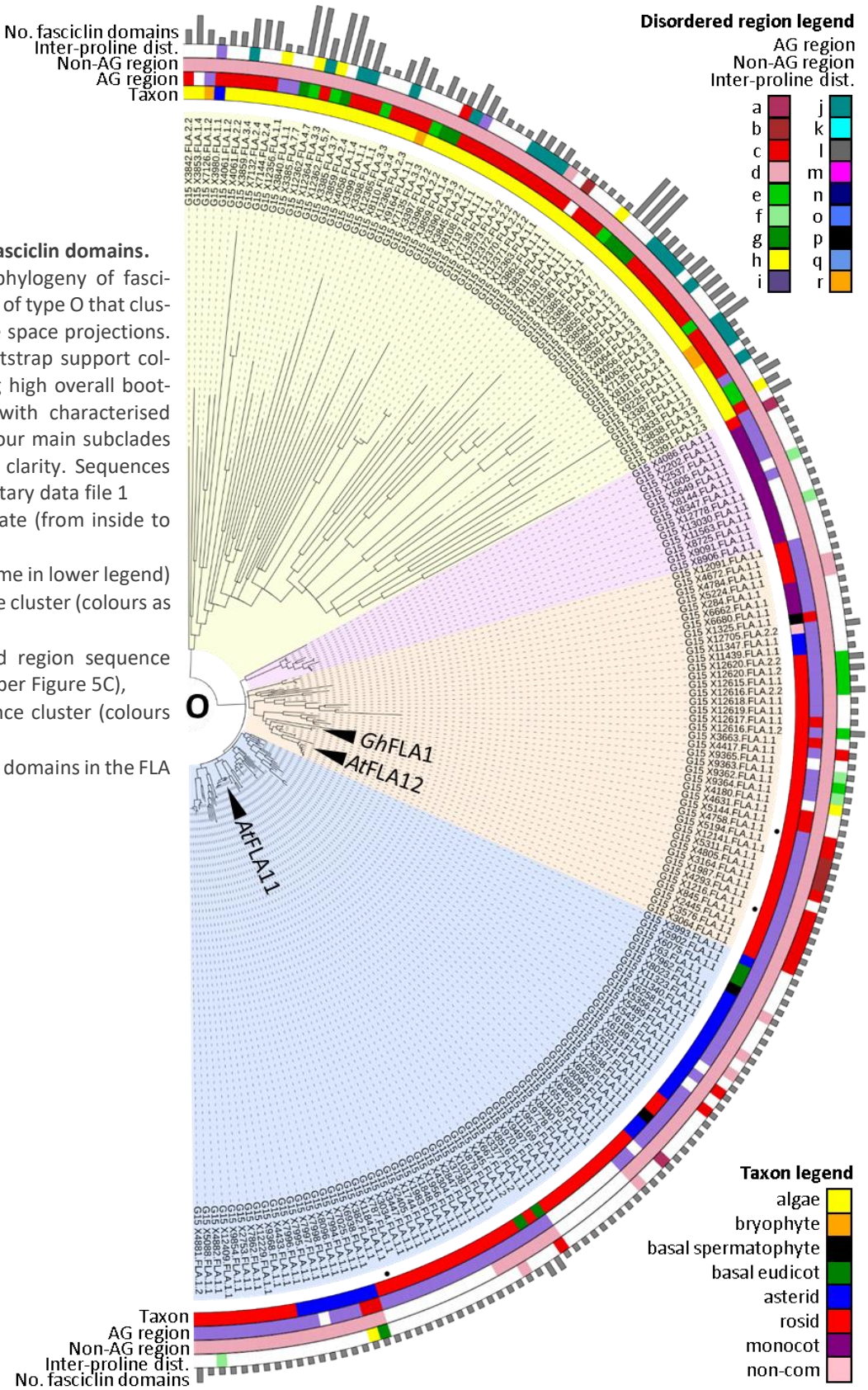
Supp Figure S12 | Inter-proline distance profile space projected by PCA.
A) For each FLA, distribution of inter-proline distances projected by PCA. No detectable clusters are identifiable by this method. **B)** PCA projection coloured by clusters identified by HDBSCAN clustering of UMAP projection of inter-proline residue distances (as in Figure 4F). **C)** UMAP projection coloured by HDBSCAN clustering of UMAP projection of AG region 3-mers (as in Figure 4C). **D)** UMAP projection coloured by HDBSCAN clustering of UMAP projection of non-AG region 3-mers (as in Figure 4D).



Supp Figure S13 |
Phylogeny of type R fasciclin domains.
Maximum likelihood phylogeny of FLA sequences that contain fasciclin domain type R. A clade of algal members has been omitted due to very long branch lengths. Nodes with <50% bootstrap support collapsed (demonstrating high overall bootstrap support). *At*FLAs annotated in grey. Sequences named per supplementary data file 1.
Annotation rings indicate (from inside to outside):

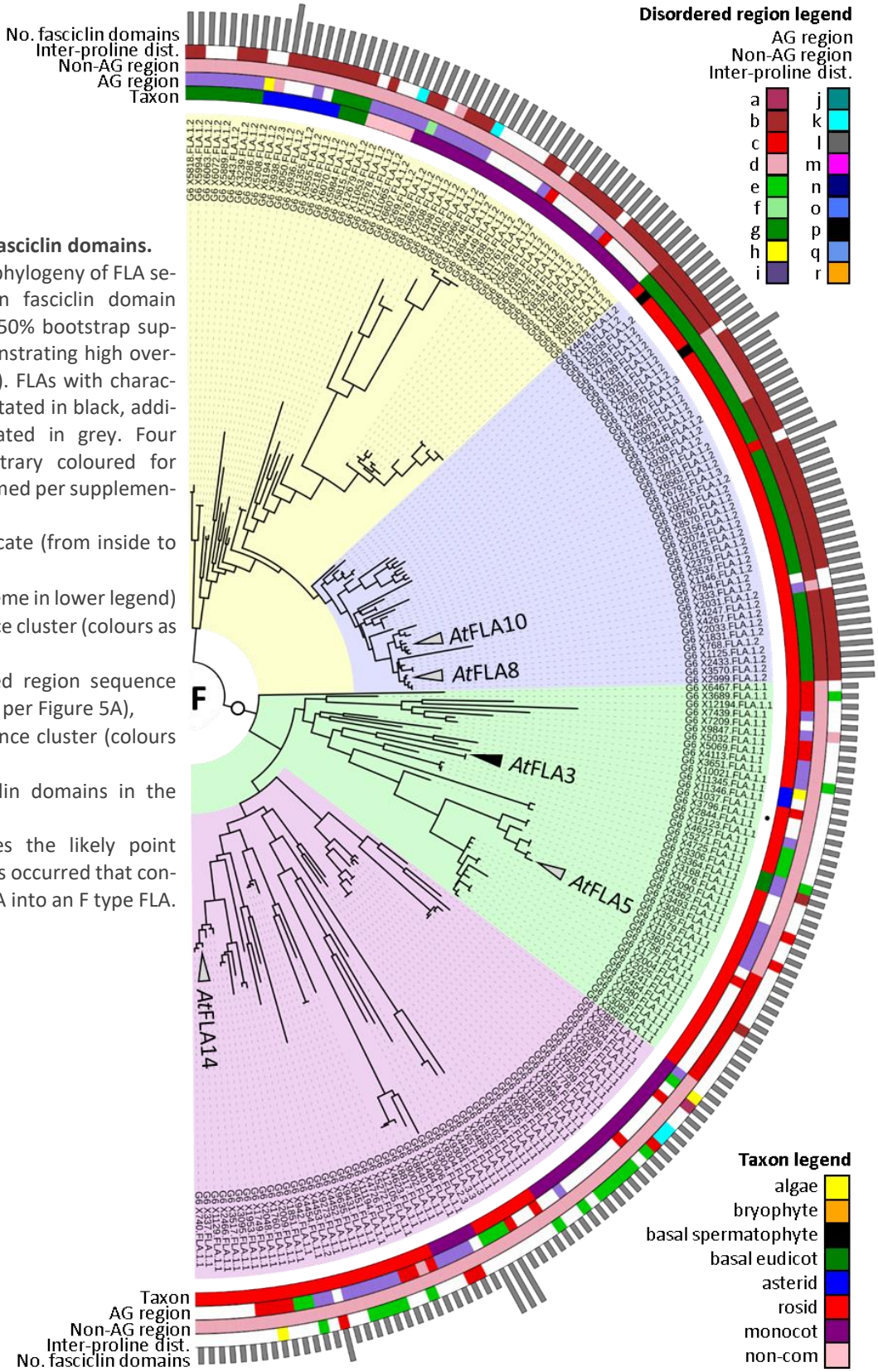
- Taxon (colour scheme in lower legend)
- AG region sequence cluster (colours as per Figure 4C)
- Non-AG disordered region sequence cluster (colours as per Figure 5A),
- Inter-proline distance cluster (colours as per Figure 5C)
- Number of fasciclin domains in the FLA (max = 3)

Note that in almost all cases exist in a R-H two-fasciclin domain architecture**.**



Supp Figure S14 |
Phylogeny of type O fasciclin domains.
Maximum likelihood phylogeny of fasciclin domain sequences of type O that cluster in UMAP sequence space projections. Nodes with <50% bootstrap support collapsed (demonstrating high overall bootstrap support). FLAs with characterised function annotated. Four main subclades arbitrary coloured for clarity. Sequences named per supplementary data file 1
Annotation rings indicate (from inside to outside):

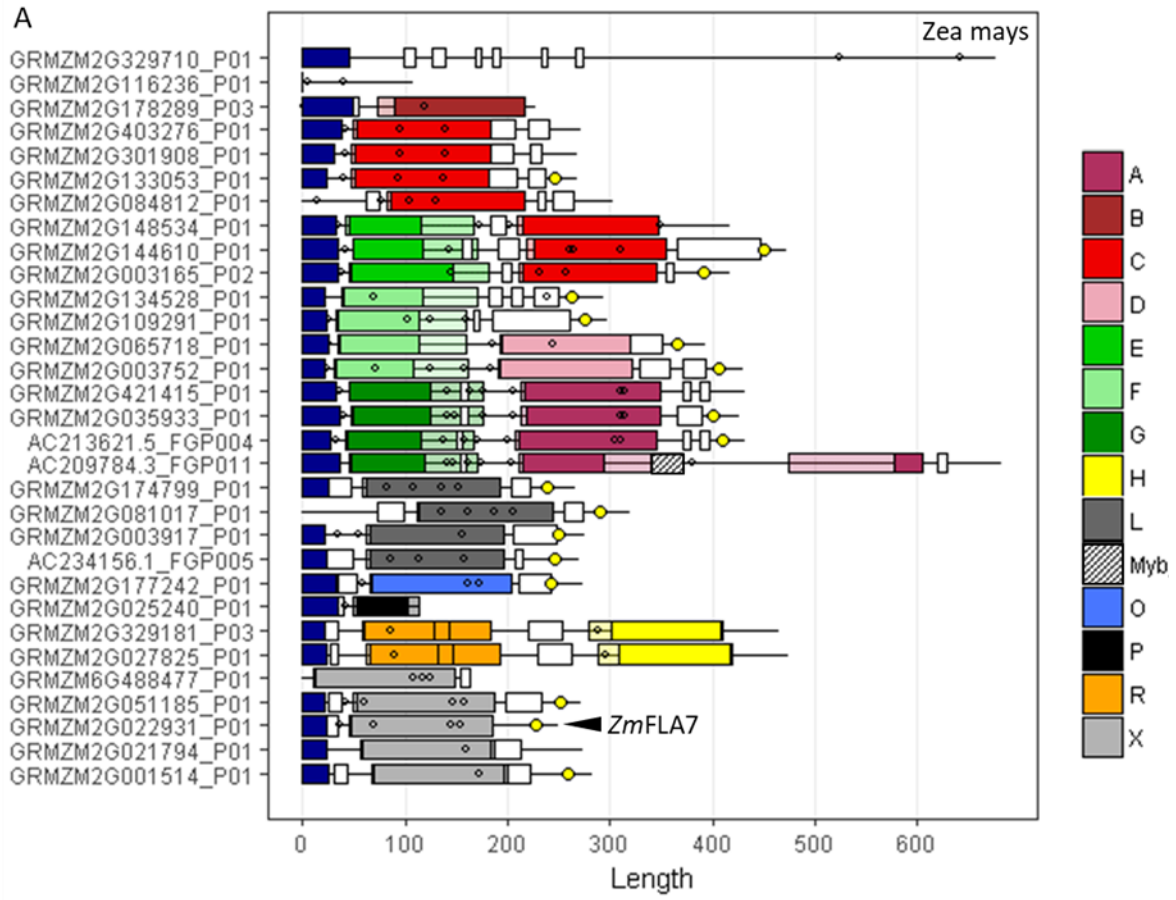
- Taxon (colour scheme in lower legend)
- AG region sequence cluster (colours as per Figure 4C)
- Non-AG disordered region sequence cluster (colours as per Figure 5C),
- Inter-proline distance cluster (colours as per Figure 5A)
- Number of fasciclin domains in the FLA (max = 7)



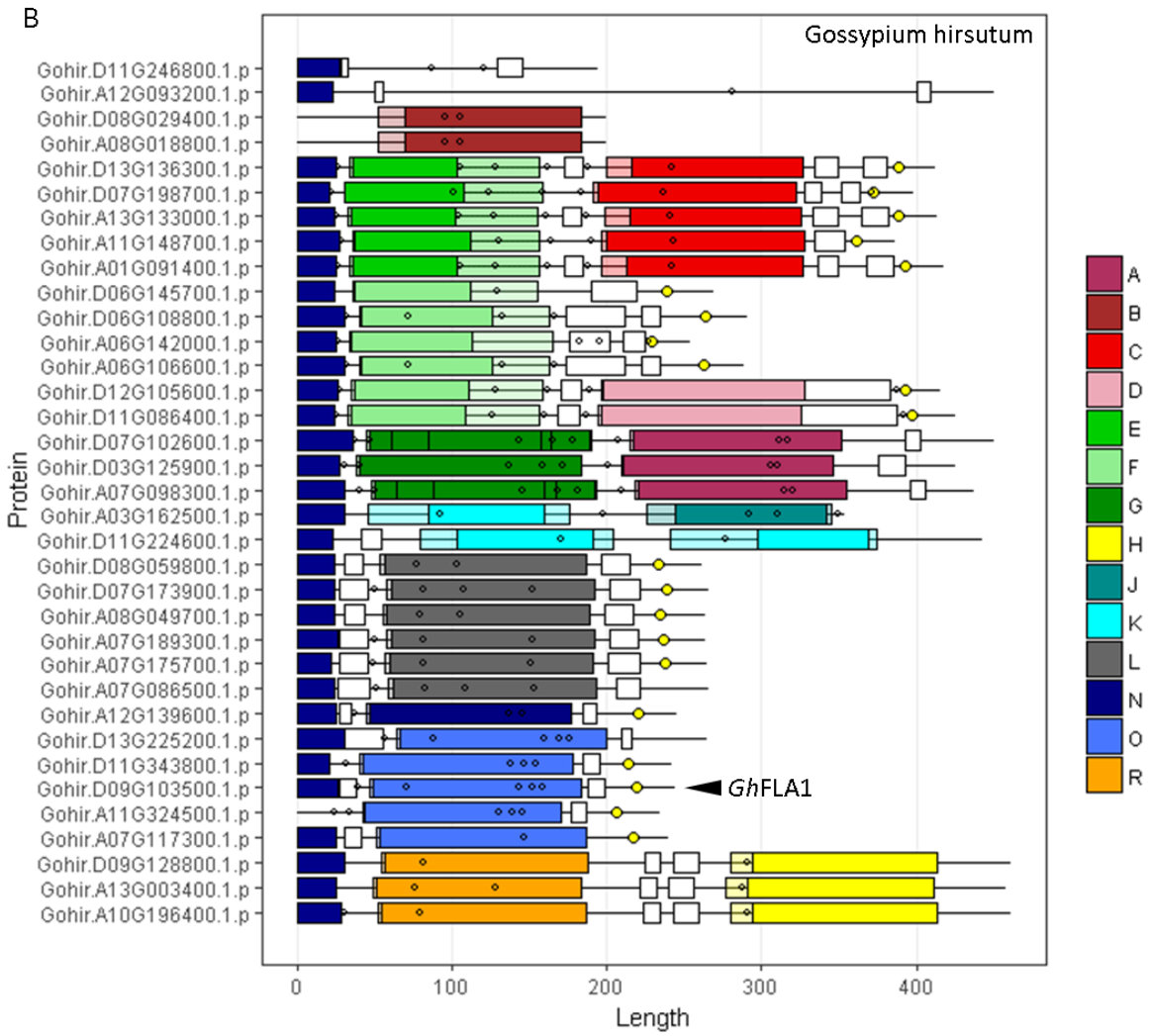
Supp Figure S15 |
Phylogeny of type F fasciclin domains.
Maximum likelihood phylogeny of FLA sequences that contain fasciclin domain type F. Nodes with <50% bootstrap support collapsed (demonstrating high overall bootstrap support). FLAs with characterised function annotated in black, additional *At*FLAs annotated in grey. Four main subclades arbitrary coloured for clarity. Sequences named per supplementary data file 1.
Annotation rings indicate (from inside to outside):

- Taxon (colour scheme in lower legend)
- AG region sequence cluster (colours as per Figure 4C)
- Non-AG disordered region sequence cluster (colours as per Figure 5A),
- Inter-proline distance cluster (colours as per Figure 5C)
- Number of fasciclin domains in the FLA (max = 3)

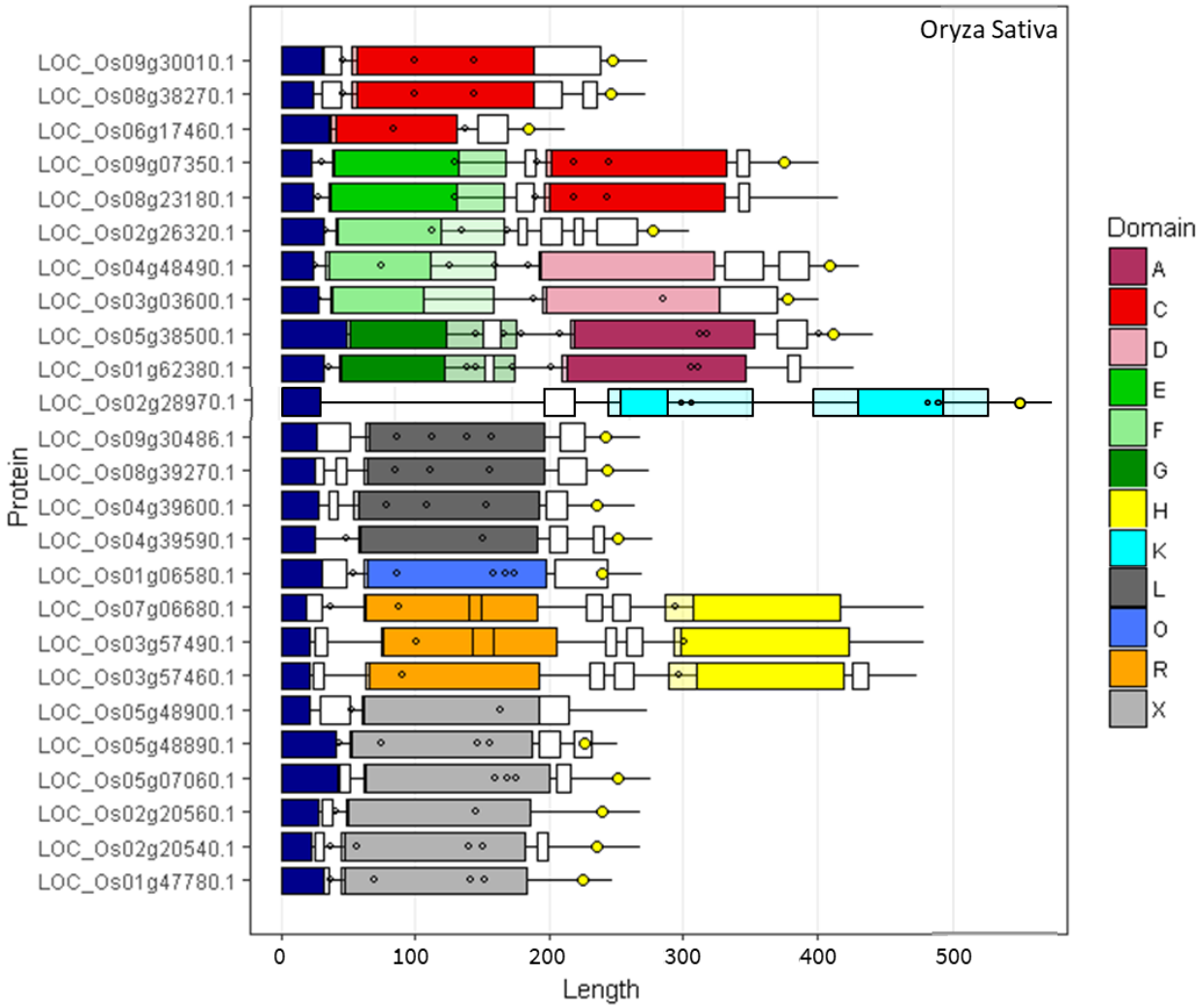
White circle indicates the likely point where the domain loss occurred that converted an F-D type FLA into an F type FLA.



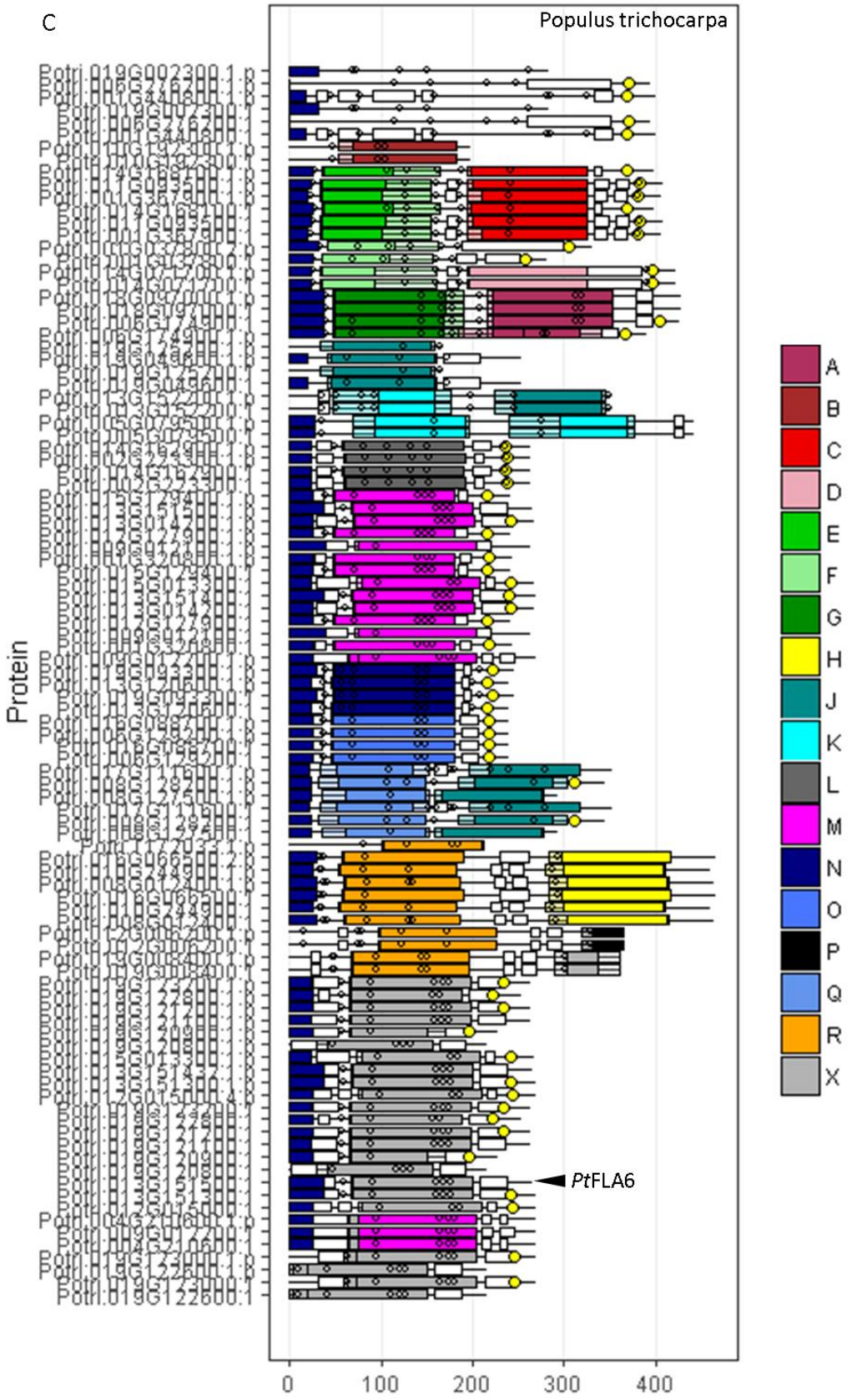
Supp Figure S16 | Schematic representation of FLA members in selected species.
FLA complement for **A)** *Zea mays*, **B)** *Populus trichocarpa*, **C)** *Gossypium hirsutum*, **D)** *Oryza sativum*, **E)** *Brassica rapa* (closest relative to *Brassica carinata* in Phytozome) and **F)** *Eucalyptus grandis*. Illustrated as in Figure 7.



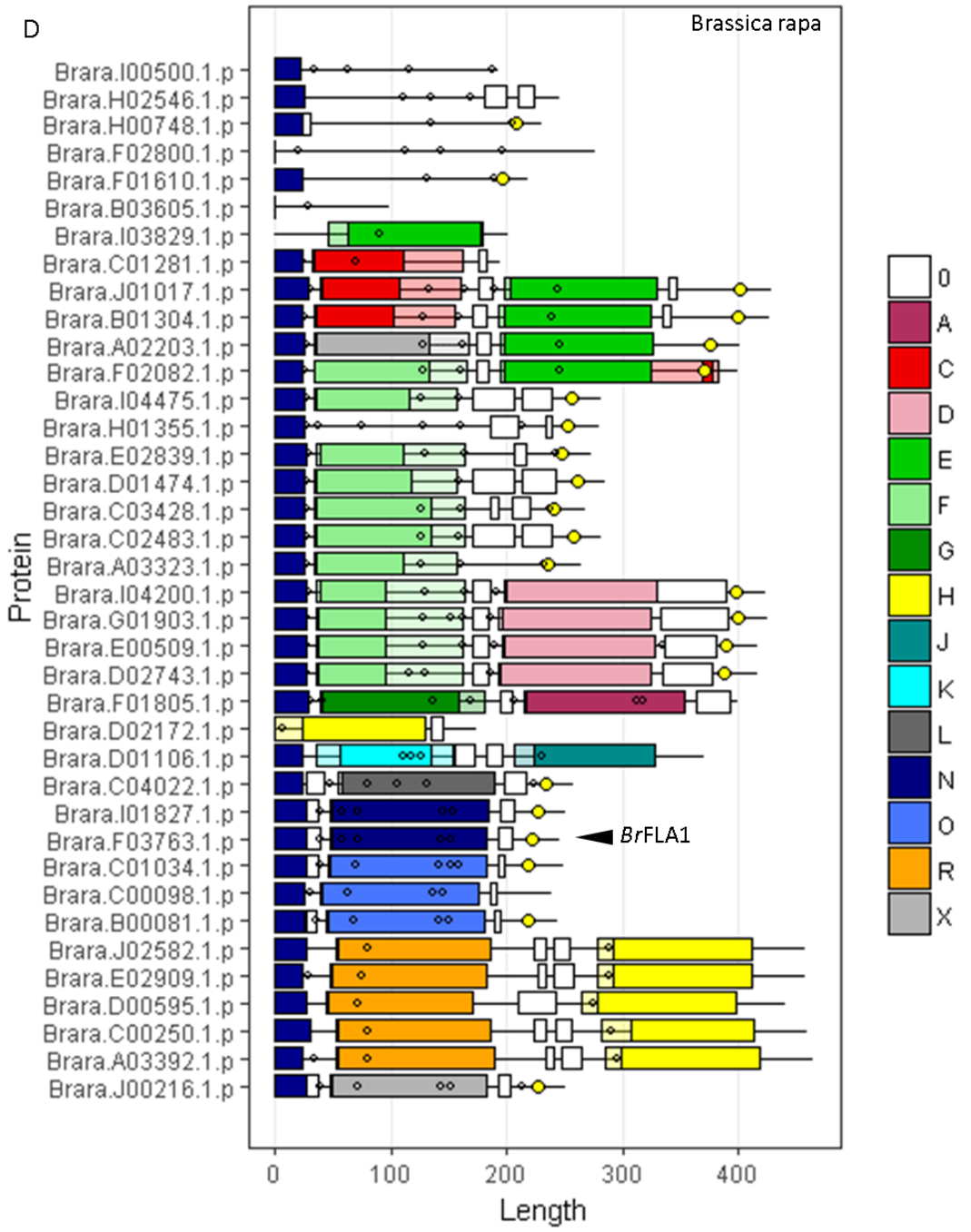
Supp Figure S16B (see legend above)



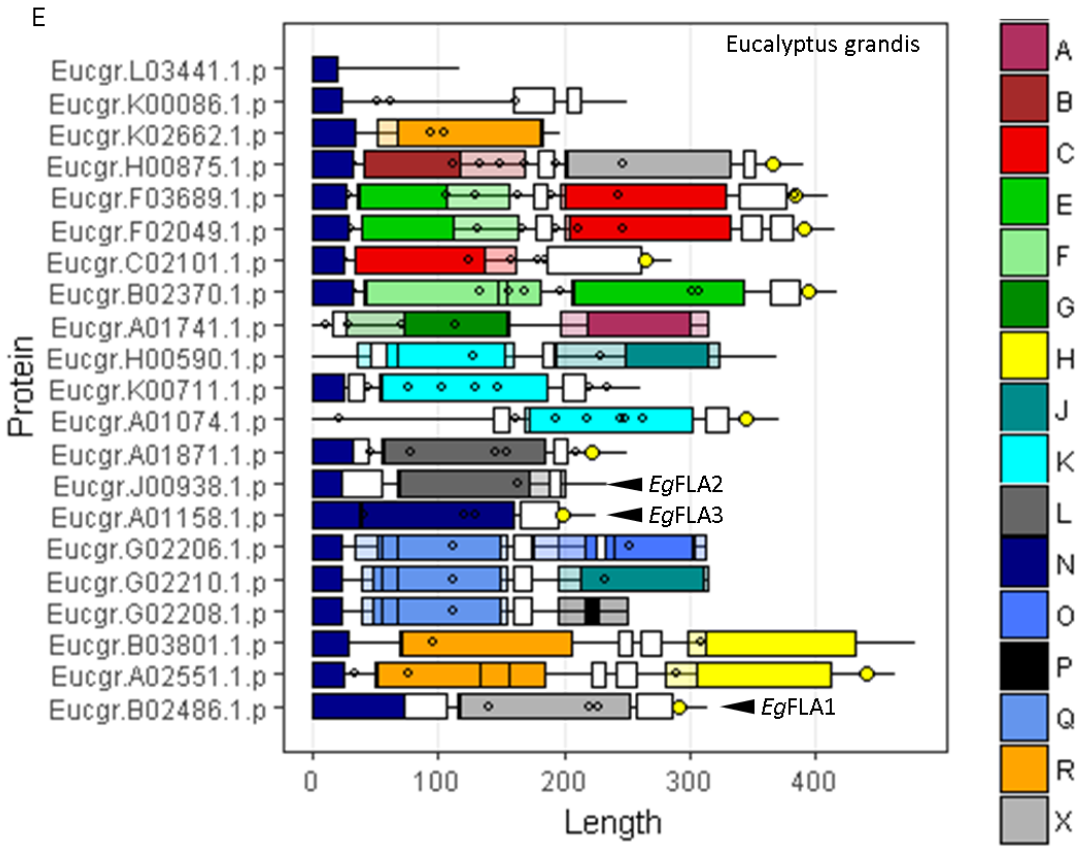
Supp Figure S16**C (see legend above)**



Supp Figure S16D (see legend above)



Supp Figure S16E (see legend above)



Supp Figure S16F (see legend above)

Supplementary Tables

Supp Table S1 | Fasciclin domain types and FLAs present in each species of the Phytozome dataset.

		Fasciclin domains																			FLAs			
	Broad group	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	X	1-fas	2-fas	≥3-fas	Total
Amaranthus hypochondriacus	basal eudicot	2	1	2	1	2	1	2	2	0	0	5	2	0	1	1	1	0	2	1	6	10	0	16
Amborella trichopoda	basal spermatophyte	2	1	1	1	1	1	1	1	0	1	4	1	0	0	2	0	0	1	0	6	6	0	12
Anacardium occidentale	rosid	1	1	3	1	2	3	1	6	0	4	1	2	0	1	2	3	3	5	11	18	14	2	34
Ananas comosus	monocot	2	0	2	1	1	3	1	1	0	0	0	1	0	1	1	1	0	0	4	7	6	0	13
Aquilegia coerulea	basal spermatophyte	1	1	1	1	1	1	1	1	0	9	9	1	0	1	1	0	3	1	1	15	8	1	24
Arabidopsis halleri	rosid	2	1	3	0	3	4	2	3	0	2	6	2	0	3	2	3	0	2	1	15	12	0	27
Arabidopsis lyrata	rosid	1	0	2	2	2	4	1	4	0	2	1	1	0	3	2	0	0	4	0	9	10	0	19
Arabidopsis thaliana columbia	rosid	0	0	2	2	2	5	0	5	0	2	2	2	0	3	2	0	0	5	0	10	11	0	21
Boechera stricta	rosid	1	1	2	2	2	4	1	4	0	2	1	1	0	2	2	0	0	4	1	10	10	0	20
Brachypodium distachyon	monocot	3	0	10	4	6	6	3	4	0	0	0	8	0	0	1	0	0	4	10	29	15	0	44
Brachypodium stacei	monocot	2	0	4	2	2	3	2	2	0	0	0	4	0	0	1	1	0	2	4	12	7	1	20
Brachypodium sylvaticum	monocot	2	1	4	1	2	2	2	2	0	0	0	4	0	0	1	0	0	2	5	14	7	0	21
Brassica oleracea capitata	rosid	1	1	3	3	5	8	0	7	0	1	0	1	0	1	2	0	0	7	1	17	11	1	29
Brassica rapa FPsc	rosid	1	1	4	4	5	11	1	6	0	1	1	1	0	2	3	0	0	5	2	17	14	1	32
Capsella grandiflora	rosid	1	1	2	2	2	4	1	4	0	2	0	1	0	2	2	0	0	4	1	13	8	0	21
Capsella rubella	rosid	1	1	2	2	2	5	1	4	0	2	1	1	0	2	2	0	0	4	1	11	10	0	21
Carica papaya	rosid	1	1	2	1	2	2	1	2	0	4	3	1	0	1	2	0	2	2	1	8	10	0	18
Chenopodium quinoa	basal eudicot	1	0	3	2	3	4	0	3	0	2	5	1	0	1	1	0	0	3	3	14	9	0	23
Chlamydomonas reinhardtii	algae	0	0	0	0	0	0	0	0	0	1	0	0	0	0	14	5	13	0	7	9	3	6	18
Chromochloris zofingiensis	algae	0	0	0	0	0	0	0	0	0	0	1	0	0	0	13	7	11	1	3	14	7	2	23
Cicer arietinum	rosid	1	1	2	1	2	1	1	3	10	5	3	1	0	1	1	1	1	3	2	16	12	0	28
Citrus clementina	rosid	1	1	2	1	2	2	1	3	0	4	1	1	0	1	1	0	2	3	3	9	10	0	19
Citrus sinensis	rosid	1	1	2	1	2	2	1	2	0	4	1	1	0	1	1	0	3	3	3	9	10	0	19
Coccomyxa subellipsoidea C-169	algae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5	0	0	1	3	1	1	5
Cucumis sativus	rosid	1	1	3	1	2	3	1	2	0	4	1	2	0	1	2	0	1	2	3	10	10	0	20
Daucus carota	asterid	2	1	3	1	3	1	2	3	0	4	4	2	0	1	2	0	2	3	1	9	12	1	22
Dunaliella salina	algae	0	0	0	0	0	0	0	0	0	0	0	4	0	0	6	11	1	0	5	2	1	7	10
Eucalyptus grandis	rosid	1	1	3	0	3	1	1	2	0	2	3	2	0	1	2	1	3	2	1	8	10	1	19
Eutrema salsugineum	rosid	1	1	2	2	2	4	1	4	0	1	2	1	0	1	2	0	0	4	2	10	10	0	20
Fragaria vesca	rosid	1	1	2	1	2	5	1	2	0	6	2	1	0	1	2	0	5	2	1	9	13	0	22
Glycine max	rosid	2	1	5	2	4	4	2	6	28	5	5	2	0	2	3	5	2	6	2	38	24	0	62
Gossypium hirsutum	rosid	3	2	5	2	5	6	3	3	0	1	3	6	0	1	5	0	0	3	0	18	15	0	33
Gossypium raimondii	rosid	2	1	4	2	4	4	2	3	0	1	3	4	0	1	4	0	0	3	0	12	13	0	25
Helianthus annuus	asterid	2	1	3	2	4	2	2	6	0	4	3	2	0	1	5	1	2	6	3	13	18	0	31
Hordeum vulgare	monocot	2	0	6	3	2	3	2	2	0	0	0	4	0	0	1	0	0	2	4	17	7	0	24
Kalanchoe fedtschenkoi	basal eudicot	2	1	4	3	4	3	2	4	0	2	2	1	0	2	1	1	1	3	1	5	16	0	21
Kalanchoe laxiflora	basal eudicot	4	1	4	5	4	5	4	4	0	5	4	1	0	2	1	2	3	4	1	5	23	1	29
Lactuca sativa	asterid	2	1	3	2	3	2	1	5	0	3	3	1	0	1	3	1	2	4	2	12	14	0	26
Linum usitatissimum	rosid	2	1	6	0	6	5	2	3	0	2	5	4	10	2	2	2	0	3	1	24	15	1	40
Manihot esculenta	rosid	2	0	3	1	3	1	2	2	0	5	3	2	16	1	1	1	3	2	0	21	12	1	34
Marchantia polymorpha	bryophyte	0	1	0	0	0	0	5	1	0	1	1	0	0	0	1	2	2	1	7	6	8	0	14
Medicago truncatula	rosid	2	1	2	1	2	3	1	3	13	3	6	1	0	1	1	3	2	3	2	21	13	1	35
Micromonas pusilla CCMP1545	algae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	2	2	2	1	5	1	2	8
Micromonas sp. RCC299	algae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	1	1	3	3	2	1	6
Mimulus guttatus	asterid	2	1	4	1	4	1	2	3	0	2	2	1	0	2	3	0	0	4	1	10	10	1	21
Musa acuminata	monocot	2	1	5	1	4	5	2	3	0	0	0	6	0	1	2	0	0	3	1	12	12	0	24
Olea europaea var. sylvestris	asterid	2	1	1	1	1	1	2	6	0	3	1	2	0	2	9	1	1	5	2	18	10	1	29
Oropetium thomaeum	monocot	2	0	4	2	1	2	2	3	0	0	0	1	0	0	1	0	0	3	4	9	8	0	17
Oryza sativa	monocot	2	0	5	2	2	3	2	3	0	0	0	4	0	0	1	0	0	3	6	15	9	0	24
Ostreococcus lucimarinus	algae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	2	1	1	1	6		1	7
Panicum hallii	monocot	6	3	18	6	12	12	6	6	0	0	0	10	0	0	3	2	0	6	15	43	28	2	73
Phaseolus vulgaris	rosid	1	1	2	1	2	2	0	3	19	3	3	1	0	1	1	3	1	3	3	24	13	0	37
Physcomitrella patens	bryophyte	0	1	0	0	0	0	2	2	0	1	4	0	0	0	3	0	2	2	3	16	2	0	18
Populus deltoides	rosid	2	1	3	1	3	2	2	4	0	4	4	2	9	2	2	0	2	4	9	26	15	0	41
Populus trichocarpa	rosid	5	2	6	2	6	4	4	6	0	12	6	4	18	4	4	2	6	11	32	53	39	1	93
Porphyra umbilicalis	algae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	13	0	0	2	3	2	7
Prunus persica	rosid	1	1	3	1	2	5	1	1	0	3	3	1	0	1	5	0	2	1	1	13	8	1	22
Ricinus communis	rosid	0	1	2	0	2	1	0	1	0	2	0	1	11	0	0	0	1	1	0	15	4	0	19
Salix purpurea	rosid	4	1	2	1	2	2	3	4	0	4	4	0	11	2	2	0	3	4	6	20	16	1	37
Selaginella moellendorffii	bryophyte	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1		1	0	1
Setaria viridis	monocot	2	1	6	1	4	3	2	2	0	0	0	5	0	0	1	1	0	2	5	16	8	1	25
Solanum tuberosum	asterid	2	1	5	1	4	3	2	2	0	3	3	2	0	1	4	0	1	2	2	14	12	0	26
Sorghum bicolor	monocot	1	0	2	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	4	4	3	0	7
Sphagnum fallax	bryophyte	0	1	0	0	0	0	1	2	0	0	1	0	0	0	0	1	1	2	6	11	2	0	13
Spirodela polyrhiza	non-com	1	1	1	2	1	2	1	2	0	0	2	0	0	2	0	0	0	2	0	3	7	0	10
Theobroma cacao	rosid	1	1	2	1	2	3	1	2	0	3	3	1	0	1	2	0	2	2	0	7	10	0	17
Trifolium pratense	rosid	5	0	2	0	2	2	1	2	12	4	3	1	0	1	1	4	1	2	2	22	10	1	33
Triticum aestivum	monocot	2	0	5	1	2	4	1	2	0	0	0	6	0	0	0	2	0	2	3	18	6	0	24
Vigna unguiculata	rosid	1	1	2	1	2	1	1	3	18	4	3	1	0	1	1	1	1	3	2	25	11	0	36
Vitis vinifera	rosid	0	0	2	0	2	0	0	1	0	0	7	0	0	1	8	2	0	0	0	7	8	0	15
Volvox carteri	algae	0	0	0	0	0	0	0	0	0	1	0	1	0	0	12	10	3	0	3	7	5	3	15
Zea mays	monocot	5	1	7	2	3	4	4	2	0	0	0	4	0	0	1	1	0	2	5	30	21	1	31
Zostera marina	non-com	2	1	5	3	4	3	2	2	0	1	0	1	0	0	1	2	0	2	2	7	12	0	19

Supp Table S2 | *N*-glycosylation and GPI anchor motif occurrence in different FLA types

	X	A	B	C	D	X2	A2	B2	C2	GPI
R-H	0%	0%	2%	1%	0%	99%	0%	0%	0%	1%
X	51%	60%	70%	37%						55%
E-C	91%	90%	0%	0%	71%	75%	10%	0%	1%	86%
O	72%	82%	82%	45%						74%
L	60%	89%	0%	0%						84%
F	81%	71%	0%	3%						90%
F-D	100%	82%	0%	0%	76%	88%	5%	0%	0%	90%
G-A	100%	82%	0%	0%	76%	88%	5%	0%	0%	43%
I	0%	54%	81%	96%						91%
M	42%	84%	84%	84%						60%
N	93%	91%	100%	0%						87%
Q-J	2%	0%	0%	0%	0%	0%	63%	0%	0%	11%
B	0%	2%	0%	2%						0%
C	71%	0%	0%	0%						60%
J	0%	15%	0%	0%						29%
K-J	0%	0%	3%	0%	5%	0%	3%	0%	0%	0%
K-K	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%
K	0%	0%	0%	0%						44%
Q	0%	4%	0%	17%						24%
P	8%	0%	8%	0%						7%

Supplementary Data Files

Supp data 1 | Domain names and labels for all FLA fasciclin domains

Excel file for the 2644 fasciclin domains, names and annotation information. In order to keep names short for phylogenies, FLAs given arbitrary identifier numbers, and fasciclin domains within them indicated by their (e.g. "`>X1234_FLA.2.3`" -> Fasciclin domain cluster 1, arbitrary FLA identifier number 1234, FLA fasciclin domain 2 out of 3). Numbers and colours given for fasciclin, AG, non-AG and inter-proline clusters.

Supp data 2 | Fasciclin domain alignments

Zip file of multiple sequence alignments as fasta files for all 2644 fasciclin domains, as well as separately for each cluster A-R.

Supp data 3 | Fasciclin domain phylogenies

Zip file of phylogenies as newick files for all 2644 fasciclin domains, as well as separately for each cluster A-R.

Supp data 4 | Fasciclin analysis script

Scripts of analyses in [R].