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

Thomas Shafee[1], Antony Bacic[1, 2], Kim Johnson[1, 2]*

[1] La Trobe Institute for Agriculture & Food, Department of Animal, Plant and Soil Sciences, La Trobe University, Melbourne, Australia
2 Sino-Australia Plant Cell Wall Research Centre, College of Forestry and Biotechnology, Zhejiang Agriculture and Forestry University, Lin'an, Hangzhou, China
*Correspondence via K.Johnson@latrobe.edu.au

Abstract

Approaches for studying the evolution of globular proteins are now well established yet are unsuitable for disordered sequences. Our understanding of the evolution of proteins containing disordered regions therefore lags that of globular proteins, limiting our capacity to estimate their evolutionary history, classify paralogs, and identify potential sequence-function relationships. Here, we overcome these limitations by using new analytical approaches that project representations of sequence space to dissect the evolution of proteins with both ordered and disordered regions, and the correlated changes between these. We use the Fasciclin-Like Arabinogalactan-proteins (FLAs) as a model family, since they contain a variable number of globular fasciclin domains as well as several distinct types of disordered regions: proline (Pro)-rich arabinogalactan (AG) regions and longer Pro-depleted regions.

Sequence space projections of fasciclin domains from 2019 FLAs from 78 species identified distinct clusters corresponding to different types of fasciclin domains. Clusters can be similarly identified in the seemingly random Pro-rich AG and Pro-depleted disordered regions. Sequence features of the globular and disordered regions clearly correlate with one another, implying co-evolution of these distinct regions, as well as with the *N*-linked and *O*-linked glycosylation motifs. We reconstruct the overall evolutionary history of the FLAs, annotated with the changing domain architectures, glycosylation motifs, number and length of AG regions, and disordered region sequence features. Mapping these features onto the functionally characterised FLAs therefore enables their sequence-function relationships to be interrogated. These findings will inform research on the abundant disordered regions in protein families from all kingdoms of life.

Abbreviations

AG	ArabinoGalactan
AGP	ArabinoGalactan-Protein
FLA	Fasciclin-Like Arabinogalactan-protein
GPI	GlycosylPhosphatidylInositol
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HMM	Hidden Markov Model
HRGP	Hydroxyproline-Rich GlycoProtein
MAAB	Motif and Amino Acid Bias
PCA	Principal Component Analysis
PRP	Proline Rich Protein
PTM	Post Translational Modification
UMAP	Uniform Manifold Approximation and Projection

Introduction

The evolution of globular proteins is well-studied [1, 2]. The requirement for a 3D folded structure leads to high conservation in sequence regions relevant to folding, function and interaction. Such sequences can be readily analysed using standard sequence alignment and phylogenetic approaches. However, evolutionary constraints are quite different for other protein architectures such as coiled proteins [3, 4], membrane proteins [5, 6], cysteine-rich proteins [7, 8] and disordered proteins [9, 10]. For example, intrinsically disordered proteins have a high proportion of disordering amino acids that disrupt the formation of stable structures. They frequently contain sequence repeats, and/or motifs for post-translational modification (PTM) [11]. Disordered proteins display rapid evolution with high mutation rates and often cannot be aligned in a meaningful way with the tools developed for folded proteins [12–14]. Tracking non-globular proteins/regions throughout evolution and relating changes to sequence-function relationships is therefore a significant challenge.

A third of eukaryotic proteins have chimeric structures that combine both globular domains and disordered regions of >30 residues (e.g. histones, p35, arabinogalactan-proteins (AGPs)). Indeed the majority of proteins in both *Arabidopsis thaliana* and in humans contain a disordered region >30 residues long [15, 16]. These classes of proteins play important functional roles, ranging from development and signalling to defence and disease, so are critical to our understanding of biological systems and in biotechnological applications [12]. Developing new methods to investigate these disordered and chimeric proteins will provide meaningful insight into their evolutionary pathways, functional associations between domains, and enable predictions of functionally equivalent members between species.

In plants, the hydroxyproline-rich glycoproteins (HRGPs) are a major class of cell wall glycoproteins. The protein backbones of HRGPs are intrinsically disordered, being rich in disordering amino acids such as proline (Pro) and containing repeat motifs that direct post-translational modification of Pro to hydroxyproline (Hyp) and glycan addition [17]. Within the HRGPs, the fasciclin-like arabinogalactan-proteins (FLAs) are an excellent example of a multigene family with a chimeric structure that have been implicated in a range of functions that impact plant growth and development. They combine both globular fasciclin-like domains, as well as long disordered sequences that also contain arabinogalactan (AG) regions where the Hyp residues are *O*-glycosylated with long AG side-chains [17]. In addition to the *O*-linked glycosylation of the AG region, FLAs also contain additional PTMs, including *N*-linked glycans in the fasciclin domains, a cleaved signal peptide at the N-terminus and many also contain a GPI-anchor signal sequence at the C-terminus. GPI-anchors consist of a glyco lipid structure that is attached to proteins and tethers them to the outer leaflet of the plasma membrane. *O*-linked glycosylation of the AG region is directed by the peptide sequence ‘glycomotifs’ and the resulting large glycans make a significant contribution to the molecular weight of the final glycoprotein [18, 19]. These features make FLAs an excellent model system for exploring the co-evolution between different protein regions.

Fasciclin domains have been present since the last universal common ancestor and are found in hundreds of different domain organisations [20]. They are part of the far wider β-grasp fold which includes functionally diverse proteins ranging from ubiquitin to some ribosomal subunits [21, 22]. Across eukaryotes and prokaryotes, fasciclin domains are involved in both carbohydrate and protein binding, so have evolved a variety of biological roles from cell-cell interactions to metabolite sensing [20]. Understanding the relationship between domains in FLAs can therefore be related to other fasciclin domain-containing proteins such as periostin, an extracellular matrix protein in humans associated with cancers [23].

Analysis of the FLA family has mainly focused on either the number of fasciclin domains or neighbour-joining phylogenies [24, 25], with limited investigation of their sequences [20, 26, 27]. This is due to the significant sequence variation leading to unresolved phylogenies for either the full sequence or even just the fasciclin domains [27]. The disordered regions outside of the fasciclin domains are even less conserved and have typically been omitted from analyses due to their frequent insertions, deletions and sequence repeats. However, it is these disordered regions that contain the glycomotifs that direct functionally relevant AG addition [28–30]. We developed a motif and amino-acid bias (MAAB) method that successfully categorises HRGPs into their various families [31]. This provided a basis for development of more sensitive computational methods that can distinguish features within these families.

Here, we present a set of complementary methods that better dissect sequence-diverse proteins that combine ordered and disordered regions, such as FLAs. By analysing 2019 FLAs from 78 species, we can detect order within the different kinds of seemingly random disordered regions, and co-evolution between these regions and the globular domains. This also allows a versatile classification system, that can act as a framework to further define sequence-function relationships in the FLAs (and other protein families containing disordered regions).

New Approaches

FLA sequences can be reliably identified via a combination of hidden Markov models (HMMs) of Pfam domains to identify fasciclin domains and MAAB to identify AG regions. Both HMMs and MAAB are probabilistic sequence models. HMMs make statistical predictions of amino acid (or insertion/deletion) likelihood at each position in a sequence, so are well suited to the high-complexity sequence found in globular sequence. MAAB is not position-specific, so is applicable to the repetitive, low-complexity sequence typical of AG disordered regions. Within the disordered sequence, AG regions are defined by strings of at least three "[S/T/A/G/V]P" dipeptide motifs within a ten residue window [31]. FLAs have been broadly categorised by the number of fasciclin domains. However, multiple sequence alignment of either full length FLAs or the fasciclin domains alone leads to phylogenies with very low bootstrap support, i.e. when re-running the phylogeny with random resampled subsets of the alignment, the nodes of the resulting phylogeny are not reproducible [27]. In order to infer evolutionary relationships sequences were therefore separated into their ordered fasciclin, disordered AG, and disordered non-AG regions.

Space-based methods, which organise sequences using their biophysical properties and place them within an explicit multidimensional sequence space, were adapted for each of these regions (Figure 1). For the fasciclin domains, a variant of a previously published method was used which has proven useful for other sequence-diverse protein families [32–36]. The one or more fasciclin sequences found in each FLA (2644 in total) were aligned and their position in the space was determined by the biophysical properties of each amino acid in that sequence. Such sequence spaces are highly multidimensional, so require a 'shadow' to be projected via dimension reduction techniques to be able to visualise the data in 3D plots, identify the major contributing biophysical properties, and calculate sequence clusters. For dimension reduction, Uniform Manifold Approximation and Projection (UMAP) [37] was used as it captures both local organisation of sequences within groups and the global structure between groups as well as being sensitive to non-linear relationships in the data. These clusters are likely to be monophyletic clades, however, this cannot be fully guaranteed for all sequences since a true phylogeny of the full sequence set is not possible, and so hereafter we will use the more general term 'types'.

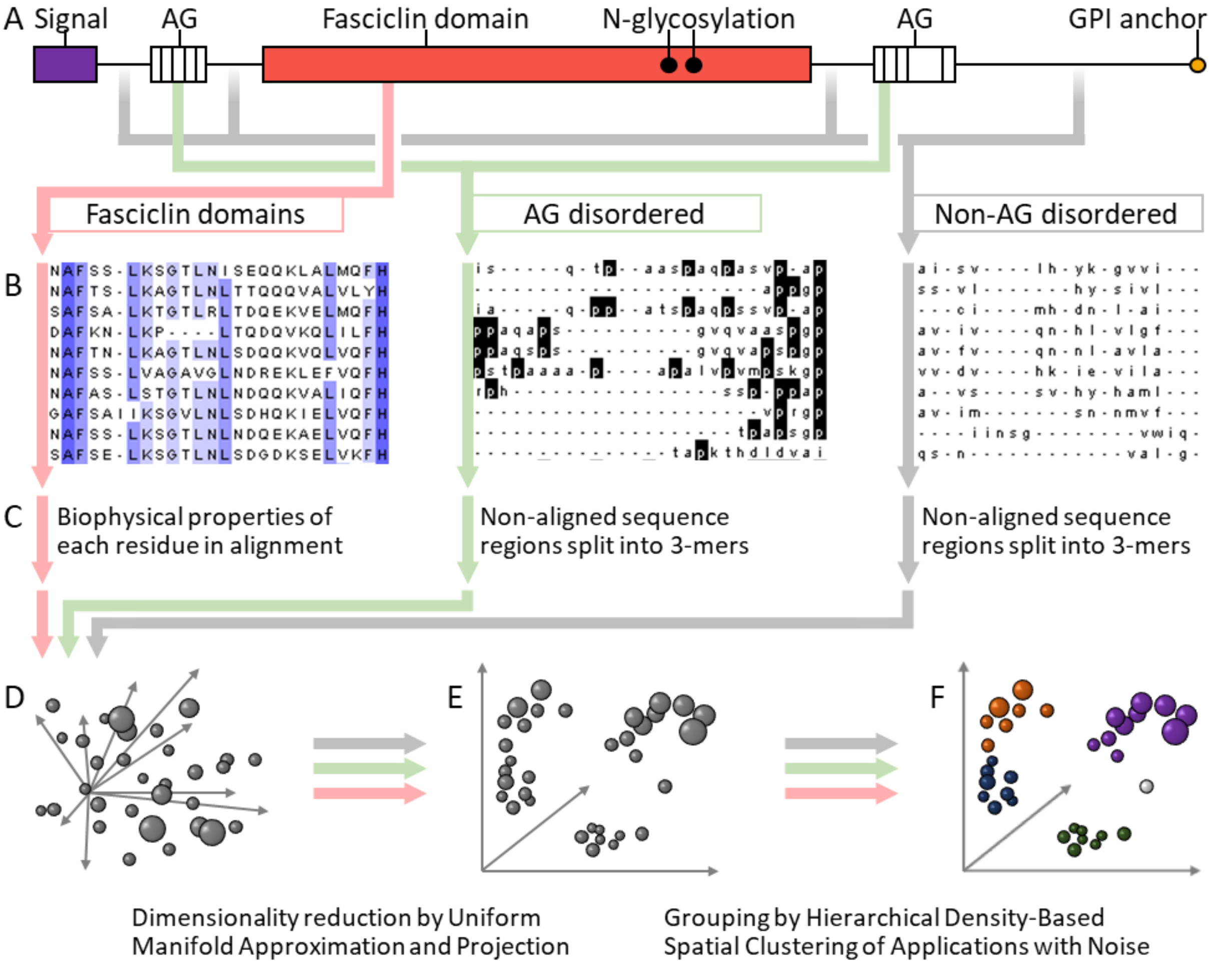

A
Signal
AG
Fasciclin domain
N-glycosylation
AG
GPI anchor
Fasciclin domains
AG disordered
Non-AG disordered
B
NAFSS-LKSGTLNISEQQKLALMQFH
NAFTS-LKAGTLNLTTQQQVALVLYH
SAFSA-LKTGTLRLTDQEKVELMQFH
DAFKN-LKP----LTQDQVKQLILFH
NAFTN-LKAGTLNLSDQQKVQLVQFH
NAFSS-LVAGAVGLNDREKLEFVQFH
NAFAS-LSTGTLNLNDQQKVALIQFH
GAFSAIIKSGVLNLSDHQKIELVQFH
NAFSS-LKSGTLNLNDQEKAELVQFH
SAFSE-LKSGTLNLSDGDKSELVKFH
C
Biophysical properties of each residue in alignment
Non-aligned sequence regions split into 3-mers
Non-aligned sequence regions split into 3-mers
D
E
F
Dimensionality reduction by Uniform Manifold Approximation and Projection
Grouping by Hierarchical Density-Based Spatial Clustering of Applications with Noise

Figure 1 | Overview of methods applied to different sequence regions of FLAs to investigate domain evolution and co-evolution.

A) FLAs are chimeric proteins, containing a range of differing numbers and types of domains, including signal peptide (purple), disordered regions containing glycomotifs that predict addition of large AG glycans (white boxed) and other, non-AG disordered regions, globular fasciclin domains (red) predicted to contain N-glycans and some contain a signal for GPI-anchor addition (orange circle). **B**) Different regions in FLAs have very different properties and levels of variation (conservation in blue, Pro shaded black, residues within detectable Pfam domain as capital letters). **C**) In the case of alignable regions (e.g. fasciclin domain), sequences can be described by the biophysical properties of their amino acids (hydrophobicity, molecular weight, charge, disorder propensity). For non-alignable regions (e.g. AG disordered) sequences can be described by the relative ratios of different 3 residue k-mers. **D**) Each sequence can therefore be represented as a point within a highly multidimensional 'sequence space'. **E**) The space can be projected into a smaller number of dimensions by UMAP. **F**) Naturally occurring clusters of sequences with similar properties can then be identified by Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). The groupings of sequences based on the different regions can then be compared via Goodman and Kruskal's tau method [38]. Within the resulting clustered sequences, it is possible to calculate phylogenies and identify correlated features, for example occurrence of *N*-glycosylation motifs, different domain architectures, inter-Pro distances, and taxonomic distributions.

The disordered regions were sub-divided into AG and non-AG regions. For each of these regions, the relative frequencies of all possible 3-residue k-mer sequences were similarly subjected to UMAP and clustering. In this study, k-mer refers to a subset of contiguous, overlapping amino acids of a defined length (k = 3) that occur within the disordered region sequence. Secondly, the profiles of inter-Pro distances present across a sequence were subjected to UMAP and HDBSCAN clustering. These are more indirect representations of the sequence space compared to the biophysical properties used for the fasciclin domain alignment, since they have many-to-one mapping from sequences to projected coordinates, however, this method is more compatible with such repetitive sequences.

These approaches have several advantages over previously used methods and are able to better dissect relevant sequence features. A given protein can therefore now be classified based on its combination of features from all these sequence regions (fasciclin biophysical properties, AG 3-mer profiles, non-AG 3-mer profiles, and inter-Pro distances). In addition, the predictive power of one feature for another (e.g. prediction of AG 3-mer profile based on fasciclin cluster) can be quantified for all possible associations. By cross-referencing this to the species taxonomy, the broad evolutionary history of the FLAs was able to be reconstructed and sequence-function predictions made.

Results

Sequence and domain diversity

In order to gather FLAs from diverse taxa, we obtained proteomes derived from 78 species of plant and algal genomes available at the Phytozome database [39]. Sequences predicted to contain any HRGP motifs were identified using a modified version of the MAAB pipeline [31] (implemented in [R] as ragp) [40]. Within this set, FLAs were then identified as proteins with a match to fasciclin domains HMM patterns (Pfam PF02469), yielding 2019 FLAs containing a total of 2644 fasciclin domains (Supp Data File 1).

In addition to the fasciclin domain, motifs were detected that direct the extensive PTMs: AG regions, *N*-glycosylation, signal peptides and GPI-anchors (Figure 2A). In particular, AG regions are defined by strings of at least three [S/T/A/G/V]P dipeptide motifs within a ten residue window in order to be recognised by the cell machinery for hydroxylation and subsequent *O*-glycosylation [31]. Overlaying a disorder prediction confirms that essentially all the regions outside of the fasciclin domains are likely unstructured (Figure 2A).

The most common domain structures were FLAs with either one or two fasciclin domains (termed 1-fas or 2-fas FLAs), with a similar number of AG regions (Supp Table S1). Even within FLAs with the same number of fasciclin domains, there was wide variation in the number and length of the disordered regions (Figure 2B). In addition, there were a smaller number of highly multi-domain FLAs with up to 7 fasciclin domain repeats (Figure 2C) which were mainly found in algae (Supp Figure S1).

Pro is typically enriched in the disordered peptides and protein regions from plants, animals and fungi (per the DisProt database). The amino acid composition of the AG regions was even more enriched in Pro than the DisProt average (Figure 2D), whereas the composition of disordered sequence outside of the AG regions was broadly consistent with the DisProt sequences (Supp Figure S3). The non-AG regions were, however, less enriched in charged residues and Pro than typically observed in DisProt (Figure 2D) [10, 41, 42]. Those few Pro residues that are present in the non-AG disordered regions do not sit within any known glycomotifs [31].

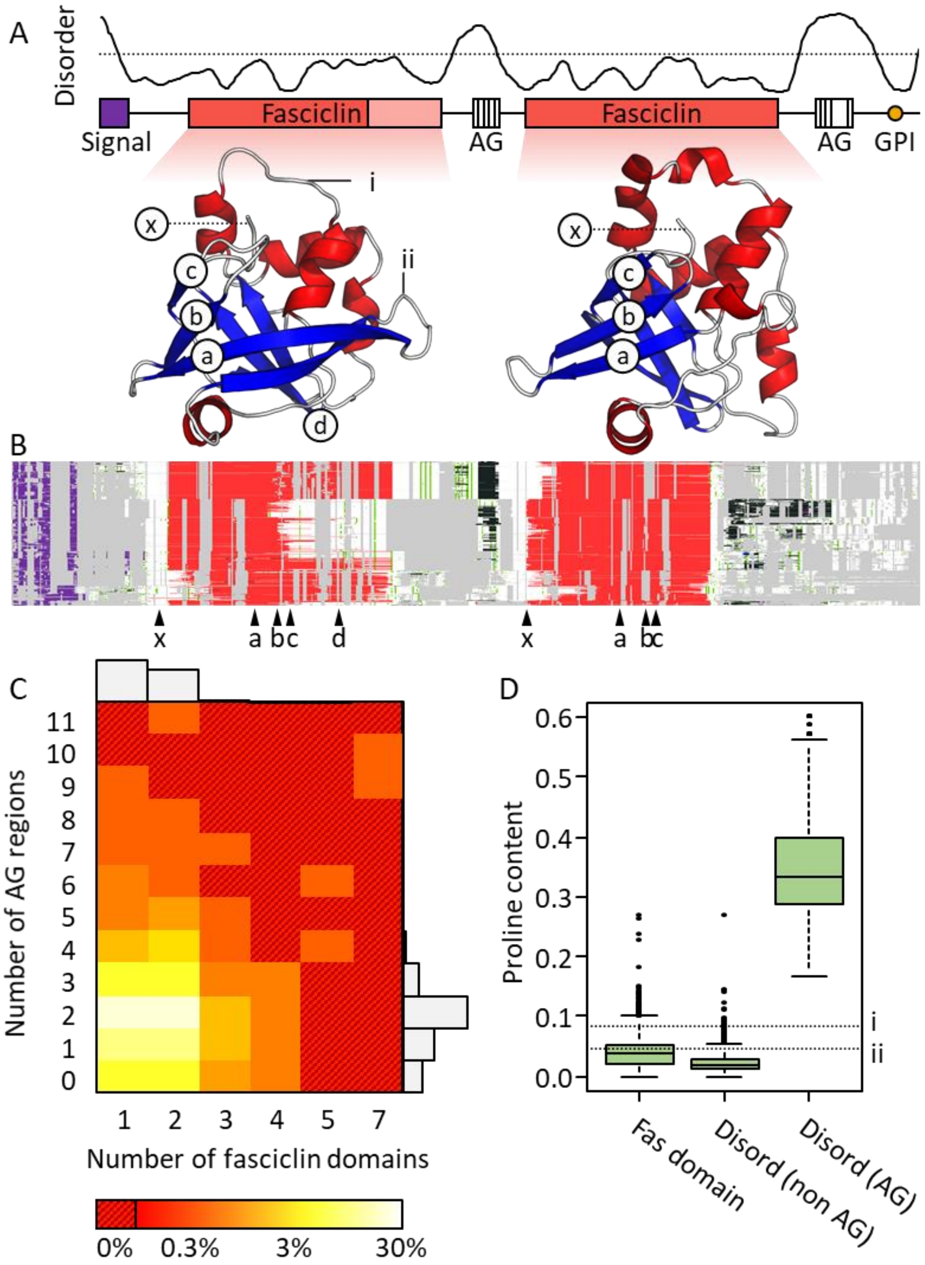


Figure 2 | FLA domain organisation and properties. A) Example FLA domain structure (*Arabidopsis thaliana* *At*FLA1; Q9FM65) with fasciclin domains in red, signal peptide in purple, AG regions in white with prolines as vertical black lines. The light red fasciclin domain region indicates the region that does not match the HMM pattern but is still likely part of the globular domain. Disorder prediction is shown above with dashed line indicating disorder 50% confidence prediction by IUPred2. Homology models of fasciclin domains indicated below are based on PDB:1o70 with beta strands in blue, alpha helices in red. Locations of *N*-glycosylation motifs observed in sequence datasets are shown as circles (locations a-d within solved crystal structure, x outside structurally solved region). The first of these domains is predicted to contain two regions of high divergence from the second domain, which has a classical fasciclin fold: i) reduced alpha helical region, ii) elongated beta strand region. **B**) Multiple sequence alignment of FLAs with two fasciclin domains. Regions that match the Pfam fasciclin (PF02469) HMM in red, AG-regions in black, prolines outside of AG regions shown in green, all other residues shown in white, gaps in grey. Locations of *N*-glycosylation motifs indicated by black arrowheads (labelled as in panel A). Some sequences with particularly long N- or C-terminal disordered regions have had their termini trimmed to fit diagram. Expanded version for 2-fas and 1-fas FLAs shown in Supp Figure S2. **C**) Frequency of occurrence of the number of fasciclin domains and AG-regions for each FLA identified in 78 species from Phytozome (log scale). Bars indicate frequency histogram. **D**) Proline content as fraction of amino acids in each FLA region. Dashed lines indicate average for i) disordered peptides in DISPROT (0.081), and ii) all proteins in UNIPROT (0.047). Expanded versions for all amino acids shown in Supp Figure S3.

Fasciclin domains

Structural regions necessary for fasciclin folding and glycosylation are relatively well conserved

The fasciclin domains present in FLAs are likely to fold into a very similar structure to fasciclin domains observed across the eukaryotes (best characterised in mammals) [20].The domain's buried hydrophobic core is highly conserved, and contains the motifs previously described as H1, H2 and YH [20] which are deeply buried (Supp Figure S4A, C). On the domain's surface, high sequence and structure conservation can be seen in the surface of the beta-sheet region (Supp Figure S4A, B), where fasciclin domains also typically have one or more *N*-glycosylation motifs N-x-[S/T] (where x≠P) within a surface-exposed beta sheet [43] (Supp Figure S4A). These include three sites in a β-sheet (labelled as 'a', 'b' and 'c' in Figure 2A), a potential site in an exposed loop regions (labelled as 'd' in Figure 2A) and a potential site immediately N-terminal of the structurally-characterised domain (labelled as 'x' in Figure 2A). Low variation is also observed in two surface patches of currently unknown function (Supp Figure S4A).

N-glycosylation motifs typically occur within a surface-exposed beta sheet [43]; (Supp Figure S8), with two other potential sites in exposed loop regions (Figure 1A). Of the 1-fas FLAs, 67% contain at least one *N*-glycosylation motif in the β-sheet, with site "a" being the most common (Supp Figure S8A). The fasciclin β-sheets of the 2-fas FLAs have fewer *N*-glycosylation motifs with 47% containing at least one motif in the first fasciclin domain, and 23% containing at least one in the second. In the 2-fas FLAs, it is more common to contain an *N*-glycosylation motif just N-terminal to each fasciclin domain. The presence of motifs at different sites through the sequence is highly correlated, such that some patterns are far more common. This correlation can be seen from how motif presence at one site often strongly predicts presence at another (Supp Figure S8C,D).

A notable point of sequence variation in the structure of fasciclin domains in FLAs from those found in animals is that a subset of FLA fasciclin domains do not match well to the common HMM pattern of the fasciclin domain superfamily and we term them 'non-classical' (Figure 2B, Supp Figure S2). This leads to a predicted tertiary structure with a contraction of the otherwise conserved alpha helical region and a large extension of part of the beta sheet (Figure 2A). This is particularly common in the first fasciclin domain of 2-fas FLAs which also tend to have much shorter inter-domain disordered regions (20-24 residues as opposed to the normal 93-100) and more extensive C-terminal AG regions.

Fasciclin domain types are identifiable from sequence space of amino acid biophysical properties

Structural differences in fasciclin domains of individual FLA members have not been studied in detail. It was therefore unclear how many distinct types exist and how that related to their functions. To analyse the structured regions in more detail, the fasciclin domains were extracted from each FLA (yielding 2644 fasciclin domains from 2019 FLAs) so that they could be analysed independently then matched back to their original FLA to see how the different domain architectures have evolved over time. Maximum likelihood phylogeny of the isolated fasciclin domain sequences yields very low bootstrap support, especially for deep nodes in the tree

(Supp Figure S5A, D). In contrast to purely phylogenetic methods, robust and interpretable clusters were detectable via the quantitative sequence space of the fasciclin domains based on the biophysical properties of their sequences (Figure 3A). It also extends the functionality of simpler PCA-based sequence space map methods (Supp Figure S6) [33].

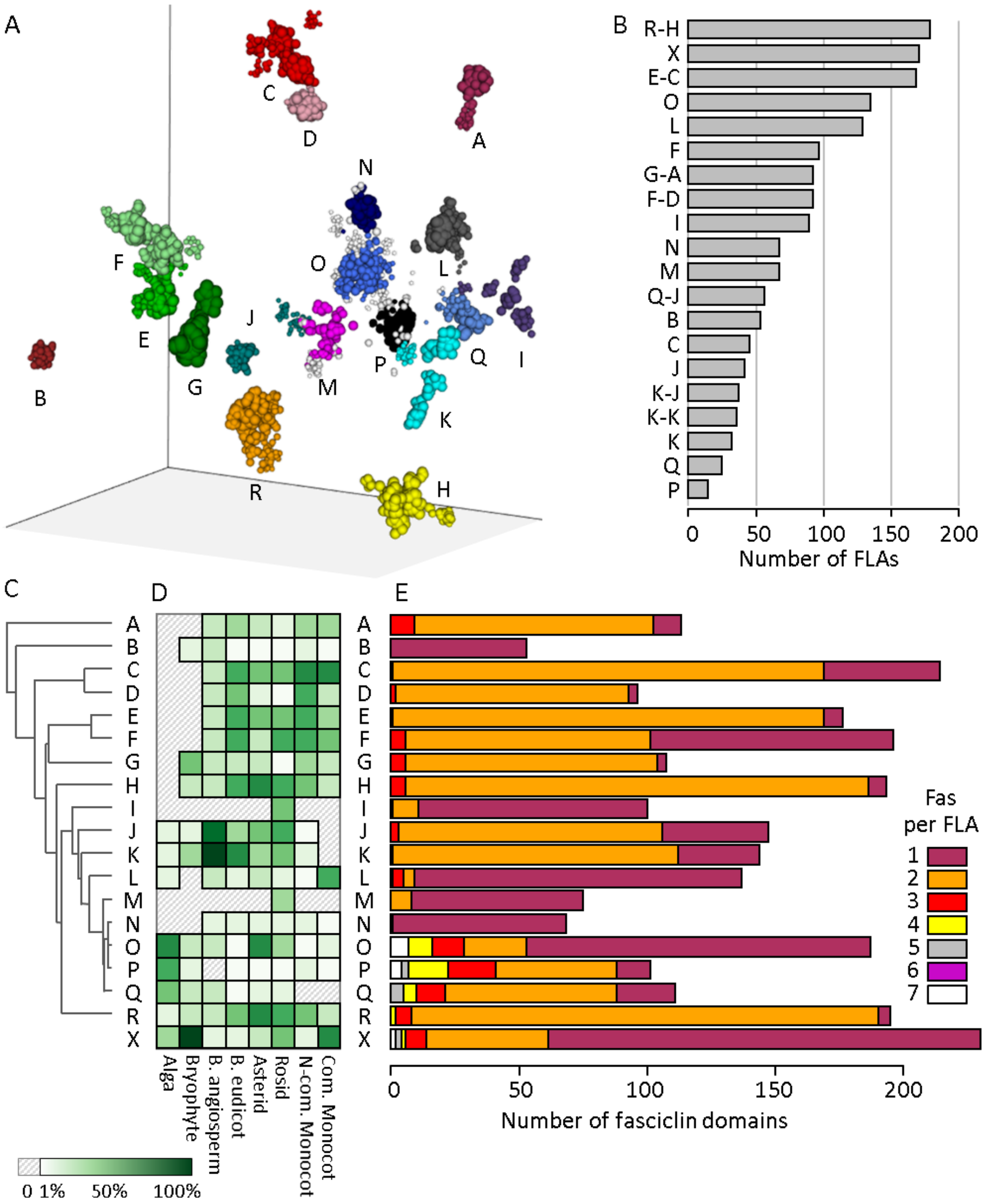


Figure 3 | Fasciclin domain types. A) Sequence space of fasciclin domains projected by UMAP. Families detected by HDBSCAN. Radius proportional to clustering certainty. Sequences not confidently assignable to a cluster in light grey. **B**) Co-occurrence of fasciclin types in the most common FLA architectures. X indicates fasciclin domains not confidently assignable to a cluster. **C**) Hierarchical clustering of fasciclin types by HDBSCAN distance. **D**) The relative frequency of occurrence of each fasciclin domain types in the major taxonomic groups. **E**) The number of facsiclin domains per FLA, separated for each fasciclin type.

The fasciclin domains were able to be assigned to 18 well-defined clusters, designated fasciclin types A-R. Their number and diversity explain why sequence alignments have proved challenging. Of the fasciclin domains, 91% were confidently assignable to a specific type, with the remainder designated as “X”, most of which sit around the clusters P, and O, indicative of either intermediate features or uniquely divergent features. Fasciclin domains from several types are found in a mixture of domain architectures (Figure 3B). Some fasciclin domain types are found only in a single architecture (e.g. type B is found only in 1-fas FLAs) or (e.g. type M is found only in the Malpighiales). However, most fasciclin domains types occur across broad taxa (Figure 3D) and in several architectures (Figure 3B, E) resulting from domain fusion and fission events. Sequences close in the projected space (Figure 3A) have more similar biophysical property sets.

Fasciclin domain types show evidence of expansion, rearrangement and divergence

Sequences from each of the fasciclin clusters (A-R, X) were aligned and phylogenies determined. The resulting phylogenies have far higher confidence than those that included all fasciclin domain sequences together (Supp Figure S5). In contrast to previously published phylogenies, the evolutionary history of each of these clusters can be accurately estimated. For example, the phylogeny of type O fasciclin domains separates out the algal and bryophyte sequences from a set of well-supported paralogous clades that originated in the early spermatophytes (Supp Figure S5B). The more basal evolutionary relationships between the clusters, can be more cautiously approximated from the similarities in their biophysical property sets, i.e. the distances in the projected sequence space (Figure 3C), and taxonomic distribution across different species (Figure 3D, Supp Figure S7).

The basal fasciclin domain type in FLAs appears to have been the O type. It is found across all plant taxonomic groups in 1-fas and multi-fas FLAs. Subsequent architectures arose from intra-gene domain duplications as well as inter-gene recombination events. As observed in other independently folding domains, repeat expansion in the FLAs occurs by both intra-gene duplications and recombination between paralogs [44], resulting in a range of domain architectures (especially in the algal lineages; Supp Figure S1). In the seed plants (spermatophytes), however, domain architecture is relatively stable, with most resulting from a very few intra-gene domain duplications (Supp Figure S7).

Types O, P, and Q are found in a large variety of organizations in the algae, including highly-multidomain FLAs with organizations such as “L-P-P-O-O-P-P”, “Q-Q-Q-Q-Q” and “X-O-O-O-X-O-O” suggesting that these resulted from relatively rapid internal domain duplications (Supp Figure S1). These domains are also found appended to the widest number of other architectures, including the infrequent occurrences of FLAs with more than two fasciclin domains in spermatophytes. This versatility in domain architecture suggests that these domains may be applicable to a variety of biological functions (Supp Figure S1). Conversely types B and N occur only in 1-fas FLAs, and types H, R, E, D have a strong preference for 2-fas FLAs, indicating that a more constrained architecture is necessary for their function.

Types G, H and B first appear in the land plants (embryophytes), and the majority of the remaining types are present in the spermatophytes (Supp Figure S7). Although most flowering plant (angiosperm) orders retain all these fasciclin domain types, individual species may lose one or more (Supp Table S1). In the spermatophyte lines, the domain organisation is relatively stable, with a small number of domain duplications, fusions and fissions resulting in most architectures (Supp Figure S7).

Disordered regions

Distinct subtypes of AG sequence regions are identifiable

Regions outside of the fasciclin domains could not be reliably aligned due to large length variation and seemingly random sequence for much of their length, similar to other intrinsically disordered peptides. For these regions alignment-free metrics are necessary [45, 46], as sequence alignments of these regions are highly misleading for any downstream analyses. The AG-regions tend to be short (median=14 amino acid residues) compared to the longer non-AG disordered regions (median=30) (Supp Figure S2).

In a new approach to distinguish features in disordered regions, the space projection and cluster detection techniques used for the fasciclin domains were combined with alignment-free methods [46]. For this, short k-mer compositions (of k=3) were calculated to summarise the frequency of all tripeptide motifs in disordered regions. Unsurprisingly, the most common 3-mers within the AG regions contain Pro residues (Figure 4A; Supp Figure S10), compared to the much broader spread of common 3-mers in the non-AG regions (Figure 4B).

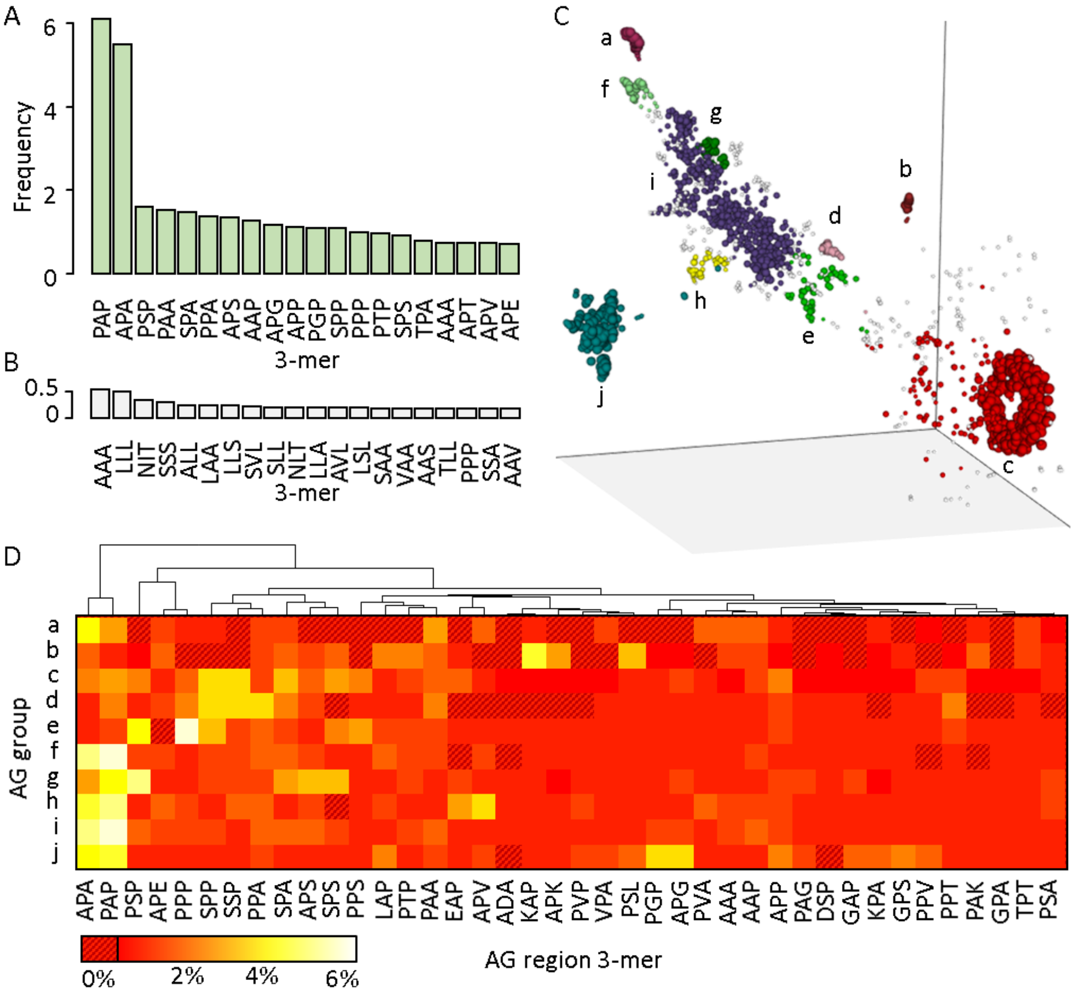


Figure 4 | AG region sequence features. Frequency of the 30 most common 3-mers within **A**) AG regions and **B**) non-AG regions of the FLAs. **C**) UMAP projection and clustering of 3-mer frequency profiles for AG regions. **D**) Relative frequency of the 30 most common 3-mers within each AG region 3-mer cluster in the FLA dataset. Version for non-AG region and inter-Pro distances shown in Supp Figure S11.

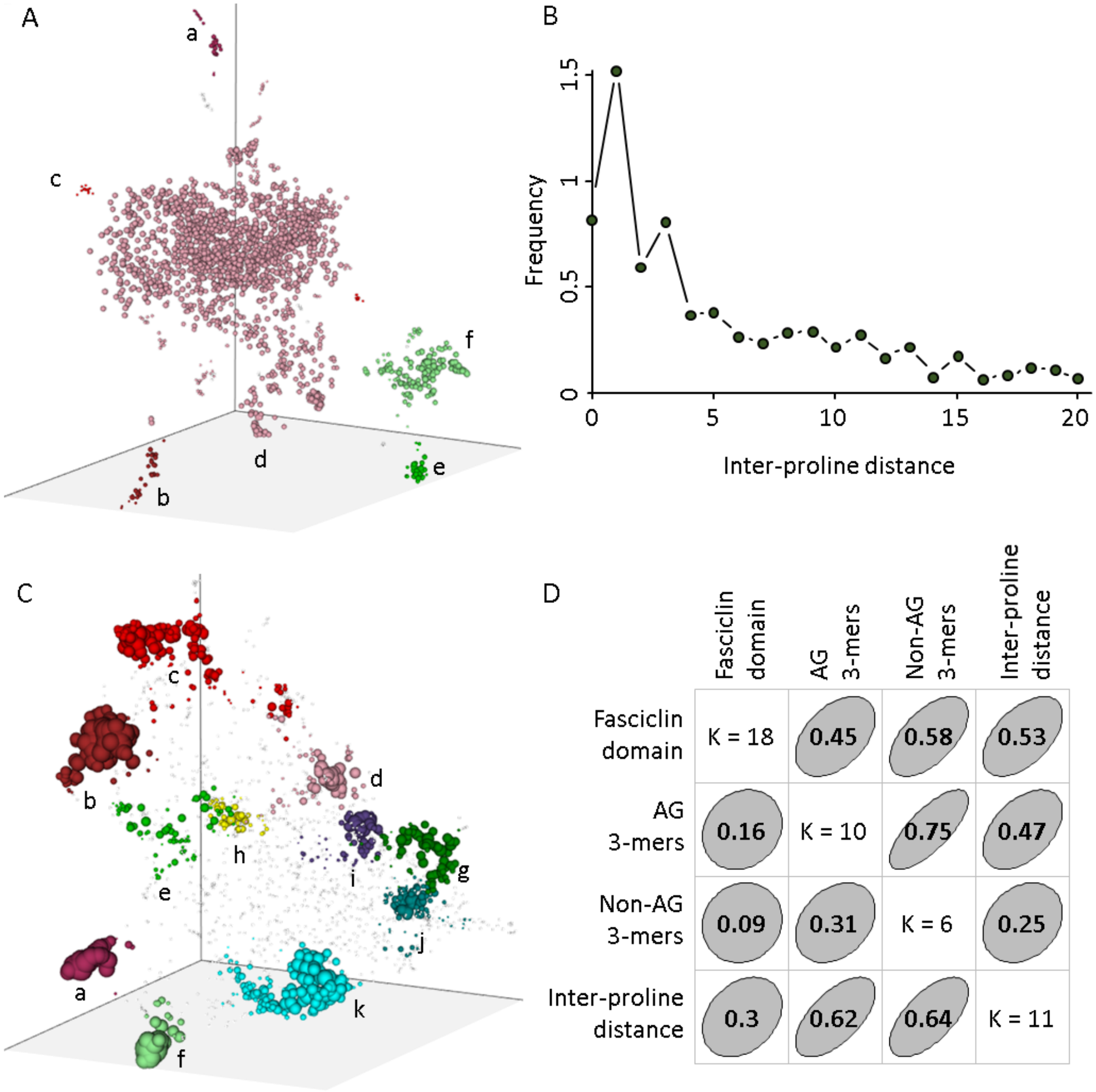
The structure of this 3-mer profile space of the AG regions is very different than that observed for the biophysical properties of the fasciclin domains (Figure 4C). Robust groups were identifiable, designated types a-j (82% of sequences assignable to a group); "x" indicates a sequence that could not be assigned to any cluster. The groups also have less well-defined boundaries than seen in the clustering of the fasciclin domain biophysical properties sequence space projection (Figure 3C) and a larger scatter of points around group c. The unusual ring structure of type c indicates a range of alternative AG region motifs but with certain combinations consistently missing. The non-random distribution of 3-mers into clusters clearly indicates that there is more order in these regions than merely enrichment of [S/T/A/G/V]P dimers.

AG region types a-j show a variety of 3-mer co-occurrence patterns (Figure 4D). Some of these correlations are trivial (e.g. "PAP" and "APA" inevitably co-occur in long $PA_{[n]}$ strings found in clusters f, h, i, j), whereas others are less intuitive, (e.g. "KAP", "APK" and "PSL" in cluster b). These 3-mer co-occurrence patterns can be used to cluster AG-region sequences in a way that captures more information than previously used methods that just calculate [P/A/S/T] percentage.

Patterns are detectable in even the seemingly random disordered sequence regions

The disordered sequence extends beyond just the AG regions (Figure 2A). As noted earlier, these regions are depleted in Pro residues (Figure 2D), but they show marked enrichment in other 3-mer sequences indicative of disorder propensity including repeating residue strings (Figure 4B; Supp Figure S11B) [42]. Those few Pro residues that are present are mainly in partially formed AG motifs that are either too short or too widely spaced to satisfy the Hyp contiguity pattern (three or more [S/T/A/G/V]P dipeptides over a 10-residue window) so are not designated as an AG region. When non-AG region 3-mer profiles are projected as before, clusters are similarly detectable (Figure 5A), designated a-f (98% of sequences assignable to a group). Again, the overall organisation of these sequences in the projection is different. The majority of non-AG disordered sequences sit in group d, with a set of smaller satellite groups. Again, those groups show biases in the 3-mer sequences, for example group d has a simple enrichment of poly-Ala and poly-Leu, whereas groups e and f have more complex patterns (Supp Figure S11B).

More global sequence features across the full-length FLA can be captured via the inter-Pro distances across the sequence (Figure 5B). The equivalent UMAP-projected space of inter-Pro distance profiles, once again, shows a different overall structure (Figure 5C). In this case, only 48% of sequences fall into discernible clusters. The remaining sequences fit in lower-density cloud between these due to intermediate properties. For example, clusters h and I are enriched in consecutive Pro residues, group f mainly has spacings of two or three residues between Pro residues and group b has spacings of one or three (Supp Figure S11C). Again, PCA is insufficient for detecting clusters (Supp Figure S12), but when the clusters identified by UMAP are painted back onto the PCA projection, they occupy discrete regions.



	Fasciclin domain	AG 3-mers	Non-AG 3-mers	Inter-proline distance
Fasciclin domain	K = 18	0.45	0.58	0.53
AG 3-mers	0.16	K = 10	0.75	0.47
Non-AG 3-mers	0.09	0.31	K = 6	0.25
Inter-proline distance	0.3	0.62	0.64	K = 11

Figure 5 | Disordered region sequence features. A) UMAP projection and clustering of 3-mer frequency profiles for non-AG regions. **B**) Frequency of inter-Pro distances. **C**) UMAP projection and clustering of inter-Pro distances ≤20 residues long. **D**) Goodman and Kruskal's tau score for correlation between fasciclin domain, AG-region 3-mer, non-AG 3-mer and inter-Pro distances. Expanded version including N-glycosylation shown in Supp Figure S9.

Coevolution of fasciclin domains and disordered regions

Applying our advanced sequence analysis methods, features distinguishing both globular and disordered regions in FLAs have been identified. These features can now be combined to identify patterns and correlations between these features and estimate a broader evolutionary history of the protein family. There is strong correlation between multiple sequence features (Figure 5D). For example, the fasciclin domain type strongly predicts *N*-glycosylation site position and GPI anchor presence (Supp Figure S9, Supp Table S2), even in cases where the rest of the disordered regions do not correlate. A caveat, however, is that prediction of GPI anchors in plants remains challenging and has a significant false negative rate [47].

Of major interest is the FLA's fasciclin domain type being highly predictive (45%) of the associated AG 3-mer profile cluster (Figure 5D). Even inter-Pro distance, despite the low separation between clusters in the UMAP-projected space, are highly predictive of both the associated fasciclin domain and k-mer-profile (Figure 5D). In *Arabidopsis*, this can be seen in the R-H type FLAs (*At*FLAs 15, 16, 17, 18) as strong co-occurrence with properties of their disordered regions (Figure 6, Figure 7, Supp Figure S13). Conversely, within the O type FLAs (*At*FLAs 11, 12), there is greater variation in the disordered regions that appears to be common in this type across other species (Supp Figure S14) and may correlate to the more subtle differences in their functions.

It may initially appear that all that is needed in order to achieve *O*-glycosylation of a FLA is at least one AG-region with some high [S/T/A/G/V]P dimer percentage. However, despite the seemingly random sequence of the AG-regions and their surrounding disordered peptides, there is order within the disorder. The different preferred sequence profiles, length and spacing of AG-region is strongly dependent on the type and number of associated fasciclin domains as well as features of the surrounding disordered sequence.

Sequence-function relationships

Given the characterised roles of FLAs in cell wall sensing and signalling, we looked at the evolution of the functionally characterised members, and whether either unique or enriched FLA types correlate to different taxonomic groups.

Combining the hierarchical clustering and phylogenies of fasciclin types with their first occurrence in different biological taxa allows a general reconstruction of the likely evolutionary history of the types and their domain architectures (Figure 6A). It is then possible to map the correlated changes in N-glycan occurrence, the length and number of AG regions, and disordered region sequence profiles (Figure 6B, C, D).

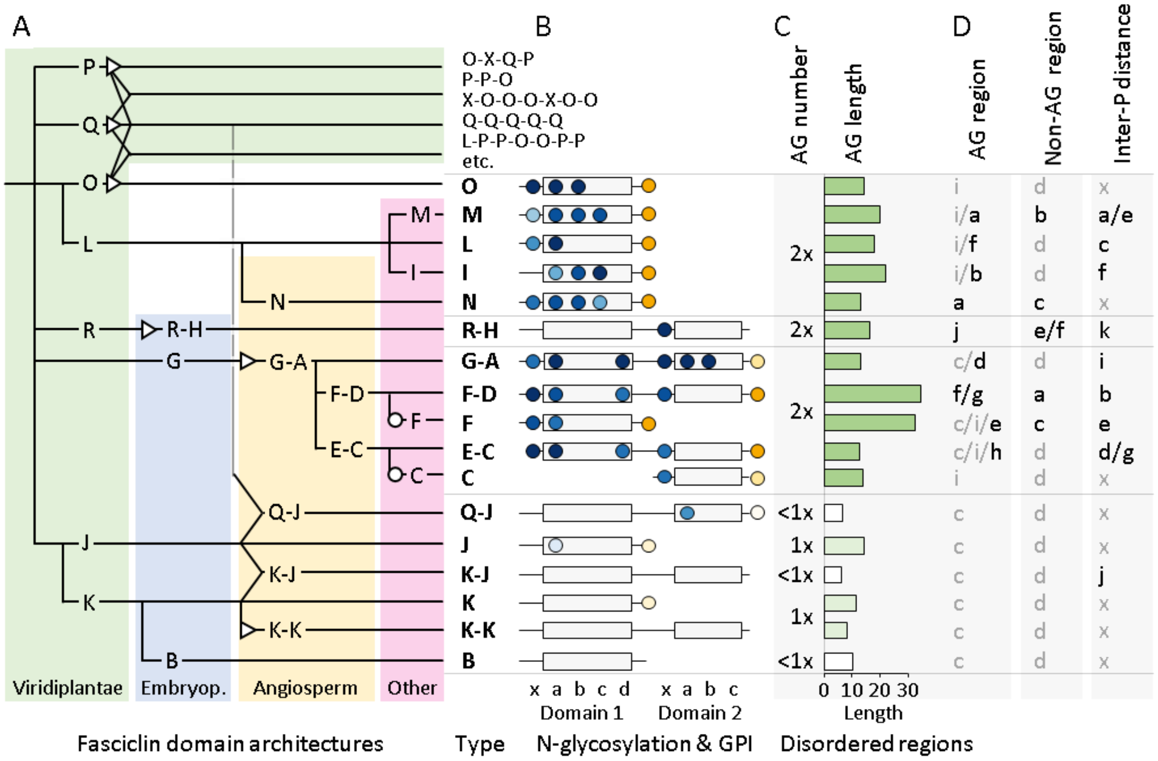


Figure 6 | Fasciclin domain types across different taxonomic groups. A) Estimated schematic evolution of most common domain architectures. Merging lines indicate domain fusion via recombination. Triangles indicate intra-gene domain duplication. Open circles indicate domain loss. Coloured panels indicate the lineage in which the main domain architectures first appeared: viridiplantae in green, embryophyte in blue, angiosperm in yellow and others in purple. **B**) Frequency of *N*-glycosylation and GPI anchor motifs at different sites across each domain. 10-100% occurrence as spectrum from white to blue for *N*-glycosylation and white to orange for GPI anchors. <10% occurrence omitted. **C**) Number of AG, average length of AG regions. **D**) most common cluster for AG region 3-mers (per Figure 4C), non-AG region 3-mers (per Figure 5A), and inter-Pro residue distances (per Figure 5C). Clusters commonly found across many FLA architectures in grey to highlight more architecture-specific clusters.

Several types of 1-fas FLAs are found in algae and have also arisen via several independent domain losses from 2-fas FLAs. Most notably, the C and F type 1-fas FLAs appear to have originated in the rosids and monocots from loss of the N-terminal domain of an E-C type FLA and C-terminal domain of an F-D type FLA, respectively (Supp Figure S7; Supp Figure S15). In both these cases, the *N*-glycosylation motifs and the length and number of AG regions is retained, however, the sequences of those AG regions shift between 3-mer clusters.

The non-classical fasciclin domains that we identified as having a shortened alpha helix and elongated beta sheet (Figure 2A) belong to types G, E and F and only match the Pfam HMM in their N-terminal portion. Of the 2-fas FLAs, 65% contain two ‘classical’ fasciclin domains (mainly types R-H, Q-J, K-J, K-K) and 35% contain a ‘non-classical’ variant first domain (types E-C, F-D, and G-A). The origin of the F type from the F-D type FLA also explains why it is the only 1-fas FLA type that uses a non-classical fasciclin domain.

In general, *N*-glycosylation motifs are relatively conserved within types. Appearance and disappearance of motifs has occurred several times independently at several sites. For example, an *N*-glycosylation motif at site b has been gained in types A, N, M, and I, with an additional motif at site c also appearing in the N, M, and I types. There appears to have been a similar convergence in the non-AG disordered regions cluster c (enriched in poly-Lys and “GSA” 3-mers) in both FLA types N and F. Whereas the length and sequence of disordered regions between the fasciclin domains varies, the position of such regions is highly conserved, in line with observations in other proteins [48].

The FLAs have been sparsely functionally characterised, with the majority of FLAs with known biological roles being from *Arabidopsis thaliana* (Figure 7). The effectiveness of the sequence space and alignment-free method to distinguish potential functional differences in FLAs is already apparent from analysis of the fasciclin domain types in FLAs from *Arabidopsis* (Figure 7). FLAs from *Arabidopsis* have previously been broadly divided into groups A-D [24]. These groups can now be seen to be quite heterogeneous. Group A included 1-fas domain *At*FLAs 6, 7, 9, 11, 12 and 13. The fasciclin domain types for these FLAs can now be distinguished and includes type L, N and O. *At*FLA9 (type N) has a role in ovule and seed development and *At*FLA11 and *At*FLA12 (type O) in secondary wall development whereas *At*FLA6, *At*FLA13 (type N), and *At*FLA7 (type L) remain uncharacterised. Fasciclin domain type O is one of the oldest, being common across all the green plants including those that do not build secondary cell walls, such as the algae. Type O in this group of FLAs therefore appears to have been recruited to roles in secondary cell wall development in the vascular plant lineage (Supp Figure S14). Type N, which occurs in *At*FLA9, only appears in the flowering plants and also includes *Zm*FLA7 from maize, which is involved in ovary drought tolerance. Type N also contains *Bc*FLA1, which appears to have a role in root hair elongation and suggests tissue-specific expression differences as well as other FLA features are likely to be important for function. The different roles for FLAs with different fasciclin types suggest functional specialisation. Disordered regions of these FLAs also display distinct patterns, for example *At*FLA11 and *At*FLA12 display different non-AG and inter-Pro types that may indicate distinct functions and/or interactions (Figure 7, Supp

Figure S14). This is further supported by phylogenetic analysis of type O fasciclin domain FLAs which show *At*FLA11 and *At*FLA12 fall into distinct clades (Supp Figure S14). Phylogenies also reinforce the coevolution of FLA ordered and disordered features (Supp Figure S14, Supp Figure S15).

FLAs with fasciclin domain type O also includes *Gh*FLA7 from cotton with a characterised role in fibre initiation and elongation and is part of the clade that includes *At*FLA12. As cotton fibres form secondary walls this suggests conservation of function in different species. *Pt*FLA6 has also been proposed to play a role in secondary wall formation, specifically in tension wood development and is the closest characterised example to the Malpighiales-specific type M domain, one of the only taxonomically restricted fasciclin domain types (the other being the Fabales-specific type I). The proximity of types M and O in the projected sequence space suggests that other members of this fasciclin type may have also been recruited to secondary walls.

Another group of 1-fas FLAs (*At*FLA3, 5, 14) in *Arabidopsis* was previously placed in Group C along with a number of 2-fas FLAs (*At*FLA1, 2, 8, 10) [24]. Overlaying the fasciclin type suggests these members were grouped largely on the presence of a fasciclin type F, with the exception of *At*FLA1 and 2 (E-C type). The relationship of this group is now much clearer given the evolutionary pathway of fasciclin type G (Figure 6). *At*FLA4 (involved in root salt tolerance and adherence) was previously placed in a different group is a G-A type, and also sits within this broad group. Type F first appeared in the flowering plants and only appears in 1-fas FLAs in the rosids (Supp Figure S15). The only characterised FLA member with fasciclin domain type F is *At*FLA3 and its role in microspore development is consistent with appearance in the flowering plants. Distinctions between FLA members with type F are now much clearer and even apparently similar members, for example *At*FLA3 and *At*FLA14, have distinct disordered regions and fall into separate clades (Supp Figure S15). FLA members with different fasciclin types are also involved in microspore development as shown by *Os*MTR1 from rice, involved in development of sporophytic and reproductive cells, that contains fasciclin type K-K (Figure 7) [49]. *Arabidopsis* does not contain corresponding K-K type FLAs, emphasising the importance of investigating FLAs from different species. Rice contains both *Os*MTR1 and a fasciclin type F member (*LOC_Os02g26320;* Supp Figure S15) expressed in the anthers (expression.ic4r.org) and the functional relationship between these is yet to be explored.

Some domain architectures have undergone significant expansions (e.g. K in the basal angiosperms and C in the commelinid monocots; Supp Figure S7). This may correlate to distinct cell wall compositions in cells/tissues and species. The commelinid monocots for example contain “type II” cell walls that are characterized by different matrix phase non-cellulosic polysaccharides (hemicelluloses) to those found in the “type I” cell walls of dicotyledonous plants, gymnosperms and non-commelinid monocots. Expansion of some fasciclin types (X, C, L), and domain combinations (G-A-P, L-L) are currently only observed in commelinid monocots and may indicate FLAs adapted specifically for function in walls of these species. Although most domain architectures are widely distributed, the I and M types are rosid-specific, found in the Fabales and Malpighiales, respectively. The FLA types defined throughout this process also highlight key FLA architectures that remain completely uncharacterised (e.g. the widespread B, K-J or R-H types).

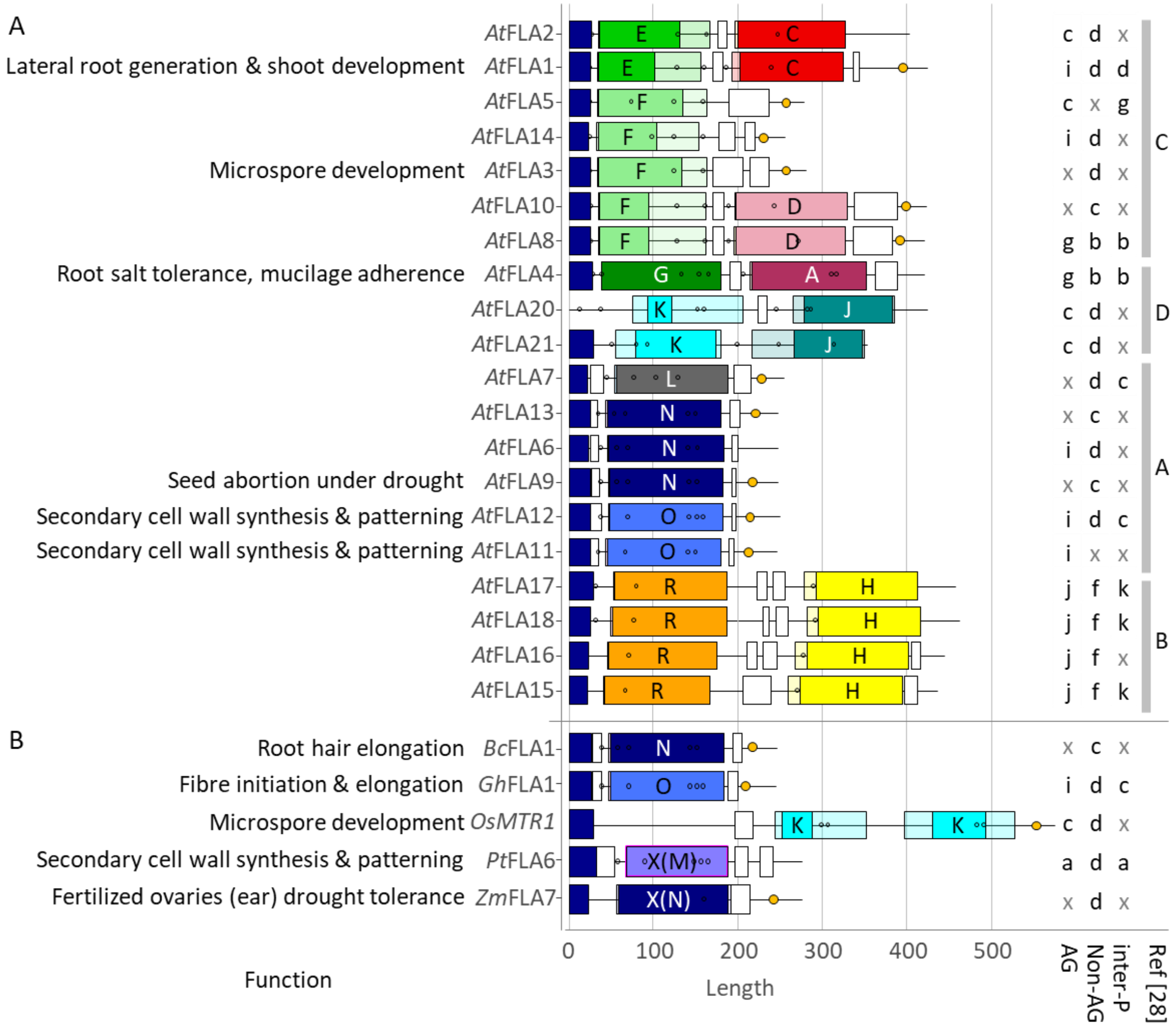


Figure 7 | Domain architecture of *Arabidopsis thaliana* and functionally characterised FLAs. Domain architectures for FLAs **A**) in *Arabidopsis* and, **B**) other plant species with genetic data on their function. Functions inferred from either knock-out mutants or RNAi knock-down lines. Fasciclin domains coloured as in Figure 3A. Regions not covered by the Pfam HMM but would fall inside the full-length domain shown as lighter coloured. For the type X FLAs, the fasciclin cluster that they are closest to is indicated in brackets. Clusters for disordered regions on the right. Sequences named with prefix: *At, Arabidopsis thaliana*; *Bc, Brassica carinata*; *Gh, Gossypium hirsutum*; *Pt, Poplar trichocarpa*; *Zm*, *Zea mays*. FLA architecture diagrams for species in panel B in Supp Figure S16. *Arabidopsis* FLA groups as defined in ref [24] indicated on the right.

Discussion

FLA evolutionary history and correlated sequence changes

The lack of robust phylogenetic clades has been a significant hinderance to understanding the evolution of the FLAs and similar chimeric ordered-disordered proteins. The methods presented here enable robust classification of fasciclin domains as well as the more challenging unstructured regions. Although hyper-variable sequence regions can appear random on first inspection, there are clear sequence preferences, both within the disordered sequence, as well as between it and the globular domains.

The sequences of glycomotifs that define different HRGP members (AGPs, PRPs, extensins) are well established and direct addition of different glycan types [19, 50]. What has remained unclear is if different features within these families exist and if this results in changes to the glycan structures, particularly in AG glycans which are the most complex. The potential micro-heterogeneity and complexity of the large AG glycans attached to the protein backbone of FLAs is astounding. This complexity has drawn parallels with the mammalian extracellular proteoglycans, such as the mucins, that have both structural and functional roles [51]. Research of animal proteoglycans show that altered glycosylation can influence function [52]. The use of antibodies that recognize different AG glycan structures suggest this is also the case in plants: AG glycan epitopes are developmentally regulated and change as tissues develop in a fashion that suggest functional significance [53]. Our approach using space projection and cluster detection techniques to distinguish features in disordered regions, combined with 3-mer co-occurrence patterns, identified 10 clustered AG-region sequences that captures significantly more information than previously used methods.

This is the first study to identify and investigate the non-AG disordered regions present in FLAs. The significance of the Pro residue depletion in the non-AG disordered regions is unknown, however it is possible that it is a specific adaptation to avoid mis-recognition by prolyl hydroxylases which might otherwise erroneously bind to those regions [54–56]. Pro residue depletion in these regions may therefore reflect a selection pressure to either minimise unproductive binding of the hydroxylation and *O*-glycosylation enzymes (sequestration) or minimise counter-productive mis-glycosylation. To date, experimental evidence defining the position and degree of hydroxylation and glycosylation on FLAs is extremely limited. Addition of AG-glycans on FLAs in *Arabidopsis* is supported by experiments showing they interact with β-glucosyl Yariv reagent that selectively precipitates AGPs [24]and AtFLA4 being labelled by AGP glycan-specific antibodies and observed at a higher molecular weight than the protein backbone alone[57]. Studies modifying the length and sequence motifs of the AG disordered and correlated non-AG disordered regions are now needed to reveal if they influence the position, degree and type of glycan structures attached to the protein backbone.

The robust clustering and correlation of a variety of sequence features across very different regions is a large advance for study of this family and more broadly for intrinsically disordered proteins. Distinct FLA types can be distinguished to an extent that it enables the establishment of a robust classification system and to interrogate their sequence-function relationships. In addition, evolutionary relationships of FLA types can now be determined in a manner that was impossible via previous phylogenetic methods [25, 27].

Fasciclin domains are found in a wide range of organisms and although some variation in the domains of FLAs was expected, the high number of fasciclin types was intriguing. Fasciclin domains have been identified in proteins across all kingdoms from a broad spectrum of taxonomic classifications indicating they are evolutionarily ancient [21]. Many of the plant fasciclin types identified in this study are present in algae but occur in a wide variety of architectures with up to 7 domains, with few paralogs. In the seed plants, the range of domain combinations becomes more conservative (mainly one or two domains per FLA) with most taxa then having a variable number of representative paralogs of those common FLA architectures (Figure 6). Since FLAs have been characterised as binding to a range of carbohydrate and protein partners [58], these features suggest that distinct functions and functional interactions are well established in the seed plants and paralogs are likely to be cell- and wall-type specific.

It is tempting to propose that the shift from the proteinaceous-rich walls of some algae to carbohydrate-rich walls of seed plants, along with specific wall types associated with different tissues, was a key factor driving specification of FLAs. Interestingly, the first fasciclin domain containing protein to be described in plants was from the algae *Volvox carteri* and called algal cell adhesion molecule (algal-CAM) due to its proposed role maintaining cell contacts during early embryogenesis [59]. Fasciclin domains are commonly associated with extracellular matrix proteins in a range of organisms and involved in cell adhesion and signalling functions [59, 60]. For example, periostin in humans is a multi-functional proteoglycan containing four FAS domains, which has been shown to regulate tissue mechanics and repair by binding fibronectins and collagens, and extracellular proteins through the fasciclin domains, to initiate signalling via the transforming growth factor-β superfamily [61]. Although these protein/proteoglycans do not exist in plants, it is possible that FLAs interact with other classes of membrane-bound protein(s) and carbohydrates to play analogous roles at the plasma membrane to cell wall interface. Wall proteins are key components in the maintenance of the biological and physical functions of the extracellular matrix in plants. Furthermore, they play important roles to perceive and transduce signals from the exterior to interior compartments [62]. FLAs have the potential to make protein-protein, protein-carbohydrate, carbohydrate-carbohydrate interactions and have been proposed to interact with pectins [63, 64], cellulose [25, 65] and receptor-like kinases [63]. Further work is required to establish how the features of these fasciclin types/FLAs influence function and how this corresponds to their specific expression patterns and plasma-membrane – wall interactions.

In many fasciclin-containing proteins the domains occur in tandem, for example *Dm*Fas1 in *Drosophila* and transforming growth factor β-induced protein (TGFBIp) from humans each have four tandem fasciclin domains [60, 66]. *Dm*Fas1 facilitates homophilic interactions [60] and the first and second, and third and fourth fasciclin domains in TGFBIp interact [67]. It remains unclear in plants if inter- or intra-molecular interactions exist. Identification of non-classical fasciclin domains (type E, F and G) with a predicted abnormal domain C-terminus was found to be common in the first fasciclin domain of 2-fas FLAs. FLAs containing these non-classical fasciclin domains also tend to have much shorter inter-domain disordered regions (20-24 residues as opposed to the normal 93-100) and more extensive C-terminal AG substituted regions. In both cases, any inter-domain interactions are likely to be different to those characterised in those animal and bacterial proteins with multiple fasciclin domains, which show buried interfaces between the domains to form a single structure (equivalent inter-domain distances of ~10 residues) [66–68]. The corresponding regions in all types of 2-fas FLAs show no evidence of conservation and their longer inter-fasciclin regions and intervening AG region(s) are also decorated with several tens of kilodaltons of *O*-glycosylation. This suggests that they likely do not interact via that structured interface but instead that the linkers are flexible and the globular domains separated by longer glycosylated linkers in the FLAs (in contrast to currently proposed models [26]). The influence of sequence features across long physical distances likely results from selective pressure and concerted function, whereby combinations of certain sequence properties yield higher fitness than others.

For the 2-fas FLAs, the only functionally characterised examples are, *At*FLA1 (E-C type) involved in root and shoot development, *At*FLA4 (G-A type), involved in root salt tolerance and adherence, and *Os*MTR1 (K-K type). A modified *At*FLA4 protein lacking the N-terminal fasciclin domain (type G) was able to complement the salt overly sensitive mutant phenotype [57]. Although naturally occurring FLAs containing only a type A fasciclin domain are rare, the equivalent N-terminal domain loss of type E-C FLAs has occurred in the commelinid monocots. Conversely FLAs with the R-H type domains have much more conserved domain architecture and disordered region sequence properties, so are expected to be less likely to function if truncated to a single domain. This highlights our limited knowledge of both the function and interactions of fasciclin domains in plants and the need for further experimental investigation. The FLA types defined throughout this process also highlight key FLA architectures that remain completely uncharacterised (e.g. the widespread B, K-J or R-H types). The extensive detail of FLA sequence features outlined in this study provide a platform for experimental studies to investigate their biological relevance.

FLAs have been implicated in a number of putative agronomic roles, including: cotton and hemp fibre development and quality [69–71], wood traits for agro-forestry and biofuel applications [72–74], wheat flour yield [75], drought and salt responses [76, 77]. The relationship of FLA members between species can now be more readily tracked so that functional conservation can be investigated. That the disordered regions fall into natural categories helps to explain how even FLAs with near-identical fasciclin domains can have different roles (such as *At*FLA11 and 12. Importantly, once function is established for FLA family members in model systems (such as Arabidopsis) we can now find the functional homologue(s) in important crop species thereby fast tracking discovery and translation in economically important species.

Generality of the method

The FLAs were used here as a model, but fasciclin domains are found in over 300 additional architectures across cellular life and include associations with carbohydrate-binding domains, epidermal growth factor domains and disordered regions of 'low complexity' in TGF-β induced protein IG-H3 proteins to name a few [20]. Exploring the co-evolution of fasciclin domains with their surrounding regions, the features that enable flexibility to form different associations and the functional significance will be an exciting area of upcoming research. These methods should also be useful for sequence analysis of other protein families which have been previously constrained by similar sequence analysis limitations as well as a diversity of proteins that contain functionally relevant non-globular regions. For each of the sequence regions, different methods of generating quantitative sequence spaces were found to be compatible with unsupervised machine learning via UMAP.

It is therefore possible to match homologs in commercially relevant species to sequences of known function with similar features, dissect sequence-function relationships, and detect correlations between different sequence regions, even in sequences and regions not compatible with phylogenetics. In plants, this includes other chimeric HRGP members such as phytocyanin-like AGPs (PAGs) also known as early-nodulin-like proteins (ENODLs), xylogen-like AGPs (XYLPs) that contain lipid-transfer-like domains and receptor-like kinases that include extensin- or AGP-like regions [78–81]. In addition to plant families, it is further relevant to animal proteins (e.g. cyclin-dependent kinase inhibitors [82], proline-rich proteins (PRPs) [83] and silk proteins [84]) and fungal proteins (e.g. AGP-like (AGL) proteins [85]). These sequence analysis techniques build on examples from globular proteins [86, 87] cysteine-rich proteins [35, 36, 88] coiled fibres [89, 90], or disordered peptides [91–93] highlighting the applicability to the majority of proteins that do not contain only ordered globular regions.

Methods

Structure comparison

Structures of fasciclin domains were determined using iterative searchers with the Dali server to search the PDB filtered by 25% redundancy [94].

Sequence gathering

FLA sequences were gathered from the Phytozome database using a modified version of the MAAB pipeline to extract 11048 HRGP sequences [31]. The MAAB pipeline as published excludes chimeric sequences (i.e. any that contain a recognised globular Pfam domain), so the standard pipeline was modified to include such sequences. Of these sequences, 2019 were identified as FLAs using the HMMER server's hmmscan function to search for Pfam domains [95] with E-value threshold at 0.01. Since some FLAs contain multiple fasciclin domains, the total number of identified fasciclin domains was 2644.

Sequence annotation

The AG-regions of the sequences were annotated using the MAAB pipeline implemented in [R] (ragp package [40]) and visualised using custom [R] scripts. Briefly, AG-regions were predicted by the presence of at least three [S/T/A/G/V]P dipeptide sequences over a 10 residue window. Fasciclin domains were annotated using the HMMER server's hmmscan function [95]. When only partial fasciclin domain sequences were identified, the region corresponding to what would be a full-length fasciclin domain sequence was further estimated based on the length of known full-length domains. For each FLA sequence, Signal peptides were estimated via SignalP5.0 (default 0.5 *Sec/SPI* threshold) [96], and GPI anchors were predicted by (p<0.001 threshold) [47]. Additionally, for each FLA sequence, statistics were gathered using custom R scripts for: the number of fasciclin domains, the number and lengths of AG regions, the number of proline residues, spacing between proline residues. Sequence profiles were calculated using the Composition Profiler webtool using SwissProt 51 as the comparator for the globular fasciclin domains, and DisProt 3.4 as the comparator for the disordered AG and non-AG regions [97].

The positions of *N*-glycosylation sites in 1-fas and 2-fas FLAs was predicted by the presence of the N[^P][T/S] motif. Since the motif is short, potential sites were checked against solvent exposure and secondary structure. The Goodman and Kruskal's tau correlation of categorical variables [38] was calculated for *N*-glycosylation sites using the GoodmanKruskal package in [R]. Goodman and Kruskal's tau correlation is a measure of association between different categorical variables in a dataset and is asymmetric: meaning the association between variables x and y is not assumed to be the same as that between y and x. It is expressed as proportional improvement in predictability of the dependent variable (scale from 0 to 1).

Sequence alignment and filtering

Since the number of fasciclin domains varies for different FLAs, multiple sequence alignments (MSAs) were made for 1-fas FLAs, and 2-fas FLAs using MAFFT (using BLOSUM30 matrix). Alignments were not generated for FLAs with >2 fasciclin domains as there were only a few sequences and highly variable. The fasciclin regions of these alignments were relatively robust and reproducible, however, alignments for all other regions of the sequences varied widely when comparing alignments by different algorithms (MAFFT[98], ClustalΩ[99] and the IDP-optimised KMAD[14]) and re-runs of the MAFFT [100, 101]

The 2644 regions annotated as fasciclin domains (plus overhang of 4 residues either side) were extracted and aligned using MAFFT using the BLOSUM30 matrix (Supp Data File 2). TrimAl's gappyout protocol was used to column-filter the alignments for phylogenetic analysis and columns that contained >0.1% occupancy were removed [102]. These final trimmed alignments were used for all for subsequent MSA-dependent analyses.

Phylogenetics

The best-fit substitution model for the fasciclin domain MSA was calculated with IQtree ModelFinder [103], indicating an optimal model of JTT+I+G4. Maximum-likelihood, 1000-bootstrap phylogenetic trees were generated for the full MSA of fasciclin domains under this substitution model using RaxML [104]. This was repeated using the full set of MSA columns, as well as column subsets resulting from processing with TrimAl under the 'gappyout' protocol, with a near-identical bootstrap distribution (the 'strict' protocol of TrimAl returned zero columns). Phylogenies were annotated in iTOL [105]. Subsequent phylogenies for sequence subsets were calculated in the same way, also using the JTT+I+G4 model and the same column set (Supp Data File 3).

Sequence space of fasciclin domains

The same multiple sequence alignment was also used to generate a quantitative sequence space for the fasciclin domains using their position-specific biophysical properties, based on reference [33]. Briefly, each position in the MSA was assigned a string of values based on its biophysical properties, and the resulting large matrix was subjected to dimensionality reduction and clustering to identify, examine, and visualise the resulting distribution in the reduced space.

The biophysical properties used were: R-group molecular mass (Daltons), net charge (Coulombs), hydrophobicity (Doolittle index), disorder propensity (TOP-IDP), and MSA column occupancy (binary descriptor). MSA gap positions (occupancy = 0) were given the column average values for each property (other than occupancy) such that those properties did not affect subsequent dimensionality reduction.

Dimensionality reduction and clustering were performed using two methods in [R]. Firstly, PCA was applied using the seqspace package, since the loadings of the resulting principal component axes summarizes the key covarying properties that most separate the sequences, with Bayesian clustering performed using the Mclust package (identifying optimal number, size and orientation of clusters based on goodness of fit). This was compared to Uniform Manifold Approximation and Projection (UMAP) using the UMAP [R] package [37]. UMAP was found to have several advantages in this case: being compatible with non-linear correlations in the data and preservation of local as well as global structure enabled more robust separation of clusters than PCA (though the mapping of clusters agreed between the methods) and was therefore used through the rest of this work. The resulting UMAP coordinates were clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) with minimum number of points in the ε-neighbourhood set to number_of_datapoints/50 using the dbscan package. Although coordinates resulting from UMAP are less easily interpretable than PCA-based coordinates, it can cluster sequences more robustly.

Similarity clustering of proline-rich and disordered regions

The disordered regions of the sequences are not reliably alignable and varied widely when aligned with different algorithms and even in repeat runs of the same algorithms [100, 101]. These regions were therefore analysed via alignment-free methods.

For each FLA, k-mer profiles (k=3) were also calculated for regions outside the Pfam domains using the k-mer [R] package for A) the AG regions and B) the non-AG non-fasciclin regions. Inter-proline distances were also calculated for the AG regions. UMAP and HDBSCAN were used for multidimensional scaling of the resulting matrices and to identify clusters of properties[37].

Homology model structure

For mapping sequence features onto their estimated 3D structural context, homology models were generated using Swiss-Model with the *Drosophila melanogaster* Fasciclin-1 (PDB:1O70) selected as the best template [106].

Data Availability

The datasets and [R] script used in this publication are available at https://doi.org/10.26181/5e33788ad3b54.

Additional scripts are available at https://github.com/TS404/FLAnnotator/.

Acknowledgements

TS, AB and KLJ were supported by the La Trobe Institute for Agriculture and Food and KJ by La Trobe Research Focus Area grant 2000004372. We would like to acknowledge Yingxuan Ma and Monika Doblin who provided constructive advice on the manuscript. AB and KJ would like to acknowledge a start-up grant for the Sino-Australia Plant Cell Wall Research Centre provided by Zhejiang A&F University.

References

1. Glasner ME, Gerlt JA, Babbitt PC. Evolution of enzyme superfamilies. Curr Opin Chem Biol. 2006;10:492–7. doi:10.1016/j.cbpa.2006.08.012.

2. Nepomnyachiy S, Ben-Tal N, Kolodny R. Global view of the protein universe. Proc Natl Acad Sci. 2014;:201403395. doi:10.1073/pnas.1403395111.

3. Jorda J, Kajava A V. Protein homorepeats: Sequences, structures, evolution, and functions. First. Elsevier Inc.; 2010. doi:10.1016/S1876-1623(10)79002-7.

4. Surkont J, Pereira-Leal JB. Evolutionary patterns in coiled-coils. Genome Biol Evol. 2015;7:545–56.

5. Pogozheva ID, Lomize AL. Evolution and adaptation of single-pass transmembrane proteins. Biochim Biophys Acta - Biomembr. 2018;1860:364–77. doi:10.1016/j.bbamem.2017.11.002.

6. Rapp M, Seppälä S, Granseth E, Heijne G Von. Emulating membrane protein evolution by rational design. Science (80-). 2007;315 March:1282–4.

7. Shafee T, Lay FT, Hulett MD, Anderson MA. The defensins consist of two independent, convergent protein superfamilies. Mol Biol Evol. 2016;33:1–23. doi:10.1093/molbev/msw106.

8. Craik D, editor. Plant Cyclotides. London: Advances in Botanical Research, Volume 76; 2015. https://www.elsevier.com/books/plant-cyclotides/craik/978-0-12-800030-4.

9. Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. Evolution and disorder. Curr Opin Struct Biol. 2011;21:441–6. doi:10.1016/j.sbi.2011.02.005.

10. Brown CJ, Johnson AK, Daughdrill GW. Comparing Models of Evolution for Ordered and Disordered Proteins. Mol Biol Evol. 2010.

11. Forman-Kay JD, Mittag T. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. Structure. 2013.

12. Van Der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of intrinsically disordered regions and proteins. Chem Rev. 2014;114:6589–631.

13. Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, et al. Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol. 2002.

14. Lange J, Wyrwicz LS, Vriend G. KMAD: Knowledge-based multiple sequence alignment for intrinsically disordered proteins. Bioinformatics. 2016;32:932–6.

15. Pietrosemoli N, García-Martín JA, Solano R, Pazos F. Genome-Wide Analysis of Protein Disorder in Arabidopsis thaliana: Implications for Plant Environmental Adaptation. PLoS One. 2013.

16. Mistry J, Coggill P, Eberhardt RY, Deiana A, Giansanti A, Finn RD, et al. The challenge of increasing Pfam coverage of the human proteome. Database. 2013.

17. Johnson KL, Cassin AM, Lonsdale A, Wong GK, Soltis DE, Miles NW, et al. Insights into the Evolution of Hydroxyproline-Rich Glycoproteins from 1000 Plant Transcriptomes. Plant Physiol. 2017;174:904–21. doi:10.1104/pp.17.00295.

18. Kieliszewski MJ. The latest hype on Hyp-O-glycosylation codes. Phytochemistry. 2001.

19. Kieliszewski MJ, Shpak E. Synthetic genes for the elucidation of glycosylation codes for arabinogalactan-proteins and other hydroxyproline-rich glycoproteins. Cellular and Molecular Life Sciences. 2001.

20. Seifert GJ. Fascinating fasciclins: A surprisingly widespread family of proteins that mediate interactions between the cell exterior and the cell surface. Int J Mol Sci. 2018;19.

21. Burroughs AM, Balaji S, Iyer LM, Aravind L. Small but versatile: The extraordinary functional and structural diversity of the β-grasp fold. Biol Direct. 2007;2:1–28.

22. Burroughs AM, Iyer LM, Aravind L. Structure and evolution of ubiquitin and ubiquitin-related domains. Methods Mol Biol. 2012.

23. Ruan K, Bao S, Ouyang G. The multifaceted role of periostin in tumorigenesis. Cellular and Molecular Life Sciences. 2009.

24. Johnson KL, Jones BJ, Bacic A, Schultz CJ. The Fasciclin-Like Arabinogalactan Proteins of Arabidopsis. A Multigene Family of Putative Cell Adhesion Molecules. Plant Physiol. 2003;133:1911–25. doi:10.1104/pp.103.031237.

25. MacMillan CP, Mansfield SD, Stachurski ZH, Evans R, Southerton SG. Fasciclin-like arabinogalactan proteins: Specialization for stem biomechanics and cell wall architecture in Arabidopsis and Eucalyptus. Plant J. 2010.

26. Turupcu A, Almohamed W, Oostenbrink C, Seifert GJ. A speculation on the tandem fasciclin 1 repeat of FLA4 proteins in angiosperms. Plant Signal Behav. 2018;13:1–5. doi:10.1080/15592324.2018.1507403.

27. He J, Zhao H, Cheng Z, Ke Y, Liu J, Ma H. Evolution Analysis of the Fasciclin-Like Arabinogalactan Proteins in Plants Shows Variable Fasciclin-AGP Domain Constitutions. Int J Mol Sci. 2019;20:1945. doi:10.3390/ijms20081945.

28. Basu D, Liang Y, Liu X, Himmeldirk K, Faik A, Kieliszewski M, et al. Functional identification of a hydroxyproline-O-galactosyltransferase specific for arabinogalactan protein biosynthesis in arabidopsis. J Biol Chem. 2013.

29. Basu D, Tian L, Wang W, Bobbs S, Herock H, Travers A, et al. A small multigene hydroxyproline-O-galactosyltransferase family functions in arabinogalactan-protein glycosylation, growth and development in Arabidopsis. BMC Plant Biol. 2015.

30. Basu D, Wang W, Ma S, DeBrosse T, Poirier E, Emch K, et al. Two hydroxyproline galactosyltransferases, GALT5 and GALT2, function in arabinogalactan-protein glycosylation, growth and development in Arabidopsis. PLoS One. 2015;10:1–36. doi:10.1371/journal.pone.0125624.

31. Johnson KL, Cassin AM, Lonsdale A, Bacic A, Doblin MS, Schultz CJ. Pipeline to Identify Hydroxyproline-Rich Glycoproteins. Plant Physiol. 2017;174:886–903. doi:10.1104/pp.17.00294.

32. Jackson MAM, Gilding EKEK, Shafee T, Harris KSK, Kaas Q, Poon S, et al. Molecular basis for the production of cyclic peptides by the plant asparaginyl endopeptidases. Nat Commun. 2018;2411:2411.

33. Shafee TMA, Anderson MAA. A quantitative map of protein sequence space for the cis-defensin superfamily. Bioinformatics. 2019;35:743–52. doi:10.1093/bioinformatics/bty697.

34. Mitchell ML, Shafee TMA, Papenfuss AT, Norton RS. Evolution of cnidarian *trans*-defensins: Sequence, structure and exploration of chemical space. Proteins Struct Funct Bioinforma. 2019;prot.25679:1–10.

35. Shafee T, Dash TS, Harvey PJ, Zhang C, Peigneur S, Deuis JR, et al. A Centipede Toxin Family Defines an Ancient Class of CSαβ Defensins. Structure. 2019;27:315-326.e7. doi:10.1016/j.str.2018.10.022.

36. Shafee T, Mitchell ML, Norton RS. Mapping the chemical and sequence space of the ShKT superfamily. Toxicon. 2019;165 December 2018:95–102. doi:10.1016/j.toxicon.2019.04.008.

37. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018. http://arxiv.org/abs/1802.03426.

38. Somers RH. A Similarity Between Goodman and Kruskal's Tau and Kendall's Tau, with a Partial Interpretation of the Latter. J Am Stat Assoc. 1962.

39. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res. 2012.

40. Dragićević MB, Paunović DM, Bogdanović MD, Todorović SI, Simonović AD. ragp: Pipeline for mining of plant hydroxyproline-rich glycoproteins with implementation in R. Glycobiology. 2019;30:19–35.

41. Obradovic Z, Vucetic S, Brown CJ, Dunker AK. Flavors of protein disorder. Proteins-Structure Funct Genet. 2003;52:573–84.

42. Uversky VN. The alphabet of intrinsic disorder: Act like a Pro: On the abundance and roles of proline residues in intrinsically disordered proteins. Intrinsically Disord Proteins. 2013;1:e24684. doi:10.4161/idp.24684.

43. Zielinska DF, Gnad F, Schropp K, Wiśniewski JR, Mann M. Mapping N-Glycosylation Sites across Seven Evolutionarily Distant Species Reveals a Divergent Substrate Proteome Despite a Common Core Machinery. Mol Cell. 2012;46:542–8.

44. Schüler A, Bornberg-Bauer E. Evolution of protein domain repeats in metazoa. Mol Biol Evol. 2016.

45. Rost B. Twilight zone of protein sequence alignments. Protein Eng Des Sel. 1999;12:85–94. doi:10.1093/protein/12.2.85.

46. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: Benefits, applications, and tools. Genome Biol. 2017;18:1–17.

47. Eisenhaber B, Wildpaner M, Schultz CJ, Borner GHH, Dupree P, Eisenhaber F. Glycosylphosphatidylinositol Lipid Anchoring of Plant Proteins. Sensitive Prediction from Sequence- and Genome-Wide Studies for Arabidopsis and Rice. Plant Physiol. 2003.

48. Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B. Protein disorder-a breakthrough invention of evolution? Current Opinion in Structural Biology. 2011.

49. Tan H, Liang W, Hu J, Zhang D. MTR1 Encodes a Secretory Fasciclin Glycoprotein Required for Male Reproductive Development in Rice. Dev Cell. 2012.

50. Shpak E, Barbar E, Leykam JF, Kieliszewski MJ. Contiguous Hydroxyproline Residues Direct Hydroxyproline Arabinosylation in Nicotiana tabacum. J Biol Chem. 2001.

51. Chaturvedi P, Singh AP, Batra SK. Structure, evolution, and biology of the MUC4 mucin. FASEB Journal. 2008.

52. Cascio S, Finn OJ. Intra-and extra-cellular events related to altered glycosylation of MUC1 promote chronic inflammation, tumor progression, invasion, and metastasis. Biomolecules. 2016.

53. Seifert GJ, Roberts K. The Biology of Arabinogalactan Proteins. Annu Rev Plant Biol. 2007.

54. Hieta R, Myllyharju J. Cloning and characterization of a low molecular weight prolyl 4-hydroxylase from Arabidopsis thaliana: Effective hydroxylation of proline-rich, collagen-like, and hypoxia-inducible transcription factor α-like peptides. J Biol Chem. 2002.

55. Tiainen P, Myllyharju J, Koivunen P. Characterization of a second Arabidopsis thaliana prolyl 4-hydroxylase with distinct substrate specificity. J Biol Chem. 2005.

56. Estévez JM, Kieliszewski MJ, Khitrov N, Somerville C. Characterization of synthetic hydroxyproline-rich proteoglycans with arabinogalactan protein and extensin motifs in arabidopsis. Plant Physiol. 2006.

57. Xue H, Veit C, Abas L, Tryfona T, Maresch D, Ricardi MM, et al. Arabidopsis thaliana FLA4 functions as a glycan-stabilized soluble factor via its carboxy-proximal Fasciclin 1 domain. Plant J. 2017.

58. Ma Y, Zeng W, Bacic A, Johnson K. AGPs Through Time and Space. In: Annual Plant Reviews online. Chichester, UK: John Wiley & Sons, Ltd; 2018. p. 1–38. doi:10.1002/9781119312994.apr0608.

59. Huber O, Sumper M. Algal-CAMs: isoforms of a cell adhesion molecule in embryos of the alga Volvox with homology to Drosophila fasciclin I. EMBO J. 1994.

60. Elkins T, Hortsch M, Bieber AJ, Snow PM, Goodman CS. Drosophila fasciclin I is a novel homophilic adhesion molecule that along with fasciclin III can mediate cell sorting. J Cell Biol. 1990.

61. Walker JT, McLeod K, Kim S, Conway SJ, Hamilton DW. Periostin as a multifunctional modulator of the wound healing response. Cell and Tissue Research. 2016.

62. Wolf S, Hématy K, Höfte H. Growth Control and Cell Wall Signaling in Plants. Annu Rev Plant Biol. 2012.

63. Griffiths JS, Crepeau M-J, Ralet M-C, Seifert GJ, North HM. Dissecting Seed Mucilage Adherence Mediated by FEI2 and SOS5. Front Plant Sci. 2016.

64. Griffiths JS, Tsai AYL, Xue H, Voiniciuc C, Šola K, Seifert GJ, et al. SALT-OVERLY SENSITIVE5 mediates arabidopsis seed coat mucilage adherence and organization through pectins. Plant Physiol. 2014.

65. Girault R, His I, Andeme-Onzighi C, Driouich A, Morvan C. Identification and partial characterization of proteins and proteoglycans encrusting the secondary cell walls of flax fibres. Planta. 2000.

66. Clout NJ, Tisi D, Hohenester E. Novel fold revealed by the structure of a FAS1 domain pair from the insect cell adhesion molecule fasciclin I. Structure. 2003.

67. García-Castellanos R, Nielsen NS, Runager K, Thøgersen IB, Lukassen M V., Poulsen ET, et al. Structural and Functional Implications of Human Transforming Growth Factor β-Induced Protein, TGFBIp, in Corneal Dystrophies. Structure. 2017.

68. Moody RG, Williamson MP. Structure and function of a bacterial Fasciclin I Domain Protein elucidates function of related cell adhesion proteins such as TGFBIp and periostin. FEBS Open Bio. 2013;3:71–7. doi:10.1016/j.fob.2013.01.001.

69. MacMillan CP, Birke H, Chuah A, Brill E, Tsuji Y, Ralph J, et al. Tissue and cell-specific transcriptomes in cotton reveal the subtleties of gene regulation underlying the diversity of plant secondary cell walls. BMC Genomics. 2017.

70. Qin LX, Chen Y, Zeng W, Li Y, Gao L, Li D Di, et al. The cotton β-galactosyltransferase 1 (GalT1) that galactosylates arabinogalactan proteins participates in controlling fiber development. Plant J. 2017.

71. Liu H, Shi R, Wang X, Pan Y, Li Z, Yang X, et al. Characterization and Expression Analysis of a Fiber Differentially Expressed Fasciclin-like Arabinogalactan Protein Gene in Sea Island Cotton Fibers. PLoS One. 2013.

72. Gritsch C, Wan Y, Mitchell RAC, Shewry PR, Hanley SJ, Karp A. G-fibre cell wall development in willow stems during tension wood induction. J Exp Bot. 2015.

73. Macmillan CP, Taylor L, Bi Y, Southerton SG, Evans R, Spokevicius A. The fasciclin-like arabinogalactan protein family of Eucalyptus grandis contains members that impact wood biology and biomechanics. New Phytol. 2015.

74. Wang H, Jin Y, Wang C, Li B, Jiang C, Sun Z, et al. Fasciclin-like arabinogalactan proteins, PtFLAs, play important roles in GA-mediated tension wood formation in Populus. Sci Rep. 2017.

75. Nirmal RC, Furtado A, Rangan P, Henry RJ. Fasciclin-like arabinogalactan protein gene expression is associated with yield of flour in the milling of wheat. Sci Rep. 2017.

76. Cagnola JI, Dumont de Chassart GJ, Ibarra SE, Chimenti C, Ricardi MM, Delzer B, et al. Reduced expression of selected FASCICLIN-LIKE ARABINOGALACTAN PROTEIN genes associates with the abortion of kernels in field crops of Zea mays (maize) and of Arabidopsis seeds. Plant Cell Environ. 2018.

77. Shi H, Kim YS, Guo Y, Stevenson B, Zhu JK. The Arabidopsis SOS5 locus encodes a putative cell surface adhesion protein and is required for normal cell expansion. Plant Cell. 2003.

78. Showalter AM, Keppler B, Lichtenberg J, Gu D, Welch LR. A bioinformatics approach to the identification, classification, and analysis of hydroxyproline-rich glycoproteins. Plant Physiol. 2010.

79. Schultz CJ, Rumsewicz MP, Johnson KL, Jones BJ, Gaspar YM, Bacic A. Using genomic resources to guide research directions. The arabinogalactan protein gene family as a test case. Plant Physiol. 2002.

80. Ma Y, Yan C, Li H, Wu W, Liu Y, Wang Y, et al. Bioinformatics prediction and evolution analysis of arabinogalactan proteins in the plant kingdom. Front Plant Sci. 2017.

81. Pazos F, Pietrosemoli N, García-Martín JA, Solano R. Protein intrinsic disorder in plants. Front Plant Sci. 2013.

82. Galea CA, Wang Y, Sivakolundu SG, Kriwacki RW. Regulation of cell division by intrinsically unstructured proteins: Intrinsic flexibility, modularity, and signaling conduits. Biochemistry. 2008.

83. Manconi B, Castagnola M, Cabras T, Olianas A, Vitali A, Desiderio C, et al. The intriguing heterogeneity of human salivary proline-rich proteins. J Proteomics. 2016.

84. Starrett J, Garb JE, Kuelbs A, Azubuike UO, Hayashi CY. Early events in the evolution of spider silk genes. PLoS One. 2012.

85. Schultz CJ, Harrison MJ. Novel plant and fungal AGP-like proteins in the Medicago truncatula-Glomus intraradices arbuscular mycorrhizal symbiosis. Mycorrhiza. 2008.

86. Cheng S, Karkar S, Bapteste E, Yee N, Falkowski P, Bhattacharya D. Sequence similarity network reveals the imprints of major diversification events in the evolution of microbial life. Front Ecol Evol. 2014;2 November:1–13. doi:10.3389/fevo.2014.00072.

87. Lopes de Carvalho L, Bligt-Lindén E, Ramaiah A, Johnson MS, Salminen TA. Evolution and functional classification of mammalian copper amine oxidases. Mol Phylogenet Evol. 2019.

88. Shafee T, Mitchell M, Papenfuss A, Norton R. Evolution of cnidarian trans-defensins: sequence, structure and exploration of chemical space. Proteins Struct Funct Bioinforma. 2019;(in press).

89. Exposito JY, Cluzel C, Garrone R, Lethias C. Evolution of collagens. Anat Rec. 2002.

90. Sutherland TD, Trueman HE, Walker AA, Weisman S, Campbell PM, Dong Z, et al. Convergently-evolved structural anomalies in the coiled coil domains of insect silk proteins. J Struct Biol. 2014.

91. Nguyen Ba AN, Yeh BJ, Van Dyk D, Davidson AR, Andrews BJ, Weiss EL, et al. Proteome-wide discovery of evolutionary conserved sequences in disordered regions. Sci Signal. 2012.

92. Bellay J, Han S, Michaut M, Kim TH, Costanzo M, Andrews BJ, et al. Bringing order to protein disorder through comparative genomics and genetic interactions. Genome Biol. 2011.

93. Mier P, Paladin L, Tamana S, Petrosian S, Hajdu-Soltész B, Urbanek A, et al. Disentangling the complexity of low complexity proteins. Brief Bioinform. 2019.

94. Holm L, Rosenström P. Dali server: conservation mapping in 3D. Nucleic Acids Res. 2010;38 Web Server issue:W545-9.

95. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res. 2018;46:W200–4. doi:10.1093/nar/gky448.

96. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019.

97. Vacic V, Uversky VN, Dunker AK, Lonardi S. Composition Profiler: A tool for discovery and visualization of amino acid composition differences. BMC Bioinformatics. 2007.

98. Ahola V, Aittokallio T, Vihinen M, Uusipaikka E. Model-based prediction of sequence alignment quality. Bioinformatics. 2008;24:2165–71.

99. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539. doi:10.1038/msb.2011.75.

100. Shafee T, Cooke I. AlignStat: a web-tool and R package for statistical comparison of alternative multiple sequence alignments. BMC Bioinformatics. 2016;17:434. doi:10.1186/s12859-016-1300-6.

101. Sela I, Ashkenazy H, Katoh K, Pupko T. GUIDANCE2: Accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res. 2015;43:W7–14.

102. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.

103. Kalyaanamoorthy S, Minh BQ, Wong TKFF, Von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–9. doi:10.1038/nmeth.4285.

104. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.

105. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44:W242–5. doi:10.1093/nar/gkw290.

106. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. Bioinformatics. 2006;22:195–201.

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

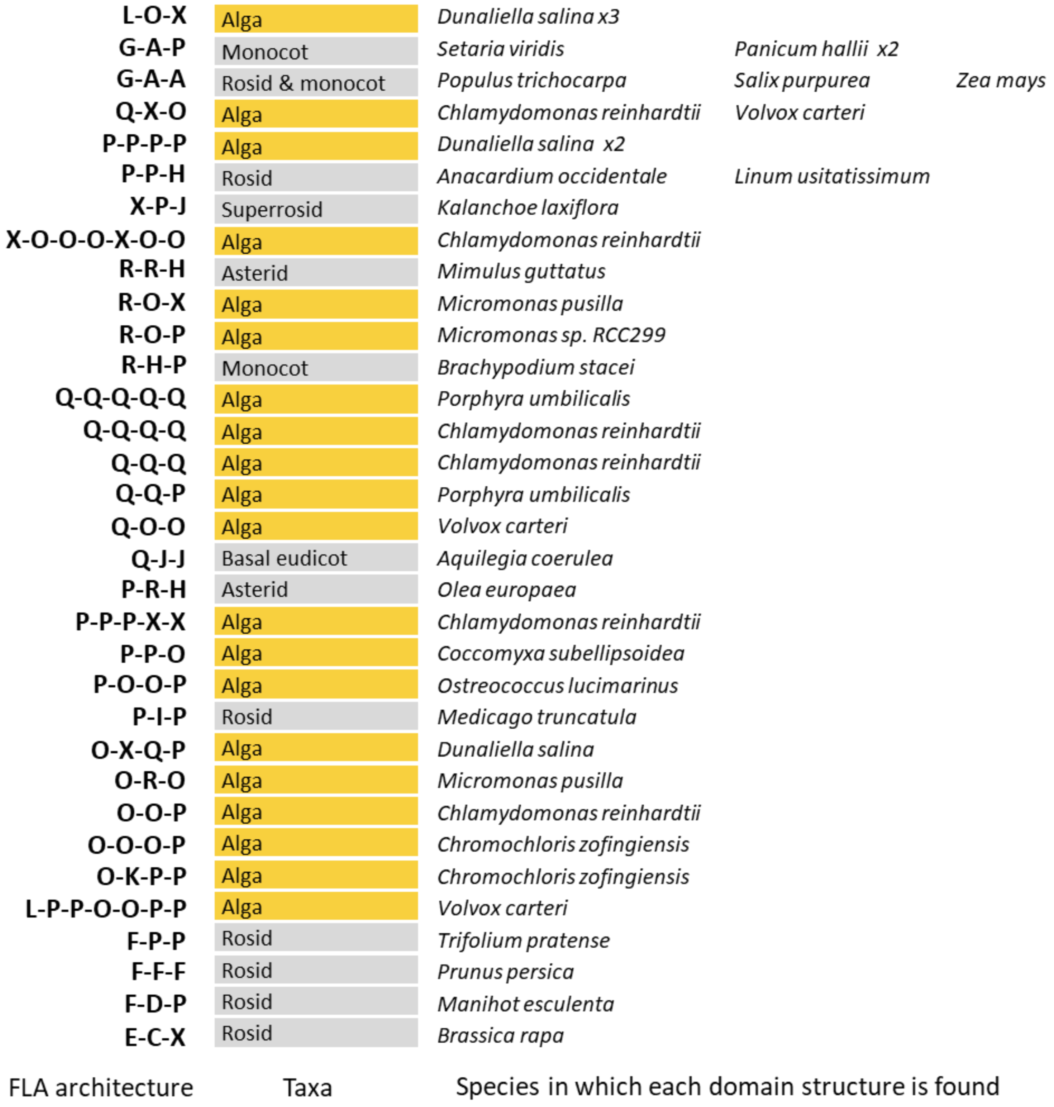
SUPPLEMENTARY INFORMATION | DOI 10.26181/5E33788AD3B54

Supplementary Figures

Supplementary Tables

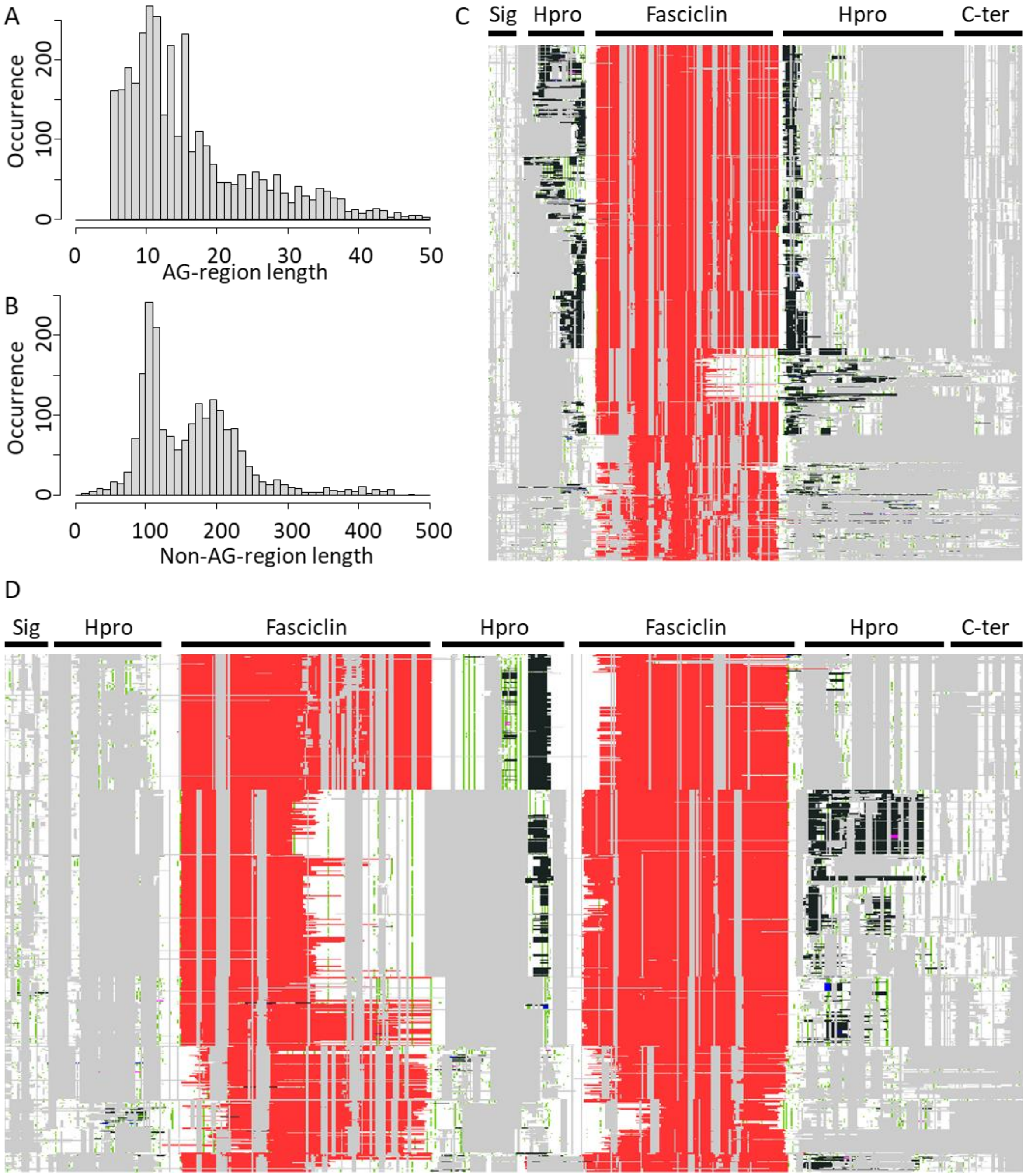
Supplementary Data Files

Supplementary Figures

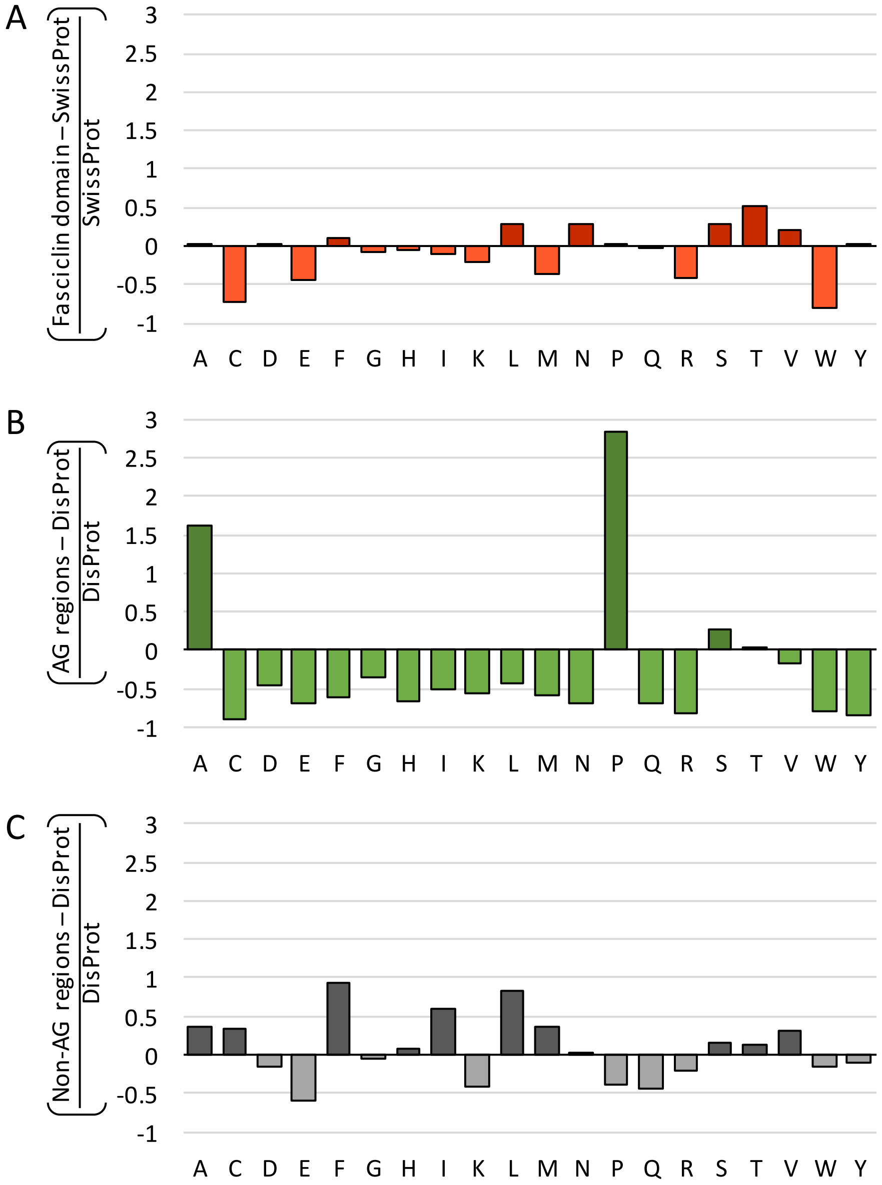


FLA architecture	Taxa	Species in which each domain structure is found		
L-O-X	Alga	*Dunaliella salina x3*		
G-A-P	Monocot	*Setaria viridis*	*Panicum hallii x2*	
G-A-A	Rosid & monocot	*Populus trichocarpa*	*Salix purpurea*	*Zea mays*
Q-X-O	Alga	*Chlamydomonas reinhardtii*	*Volvox carteri*	
P-P-P-P	Alga	*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.

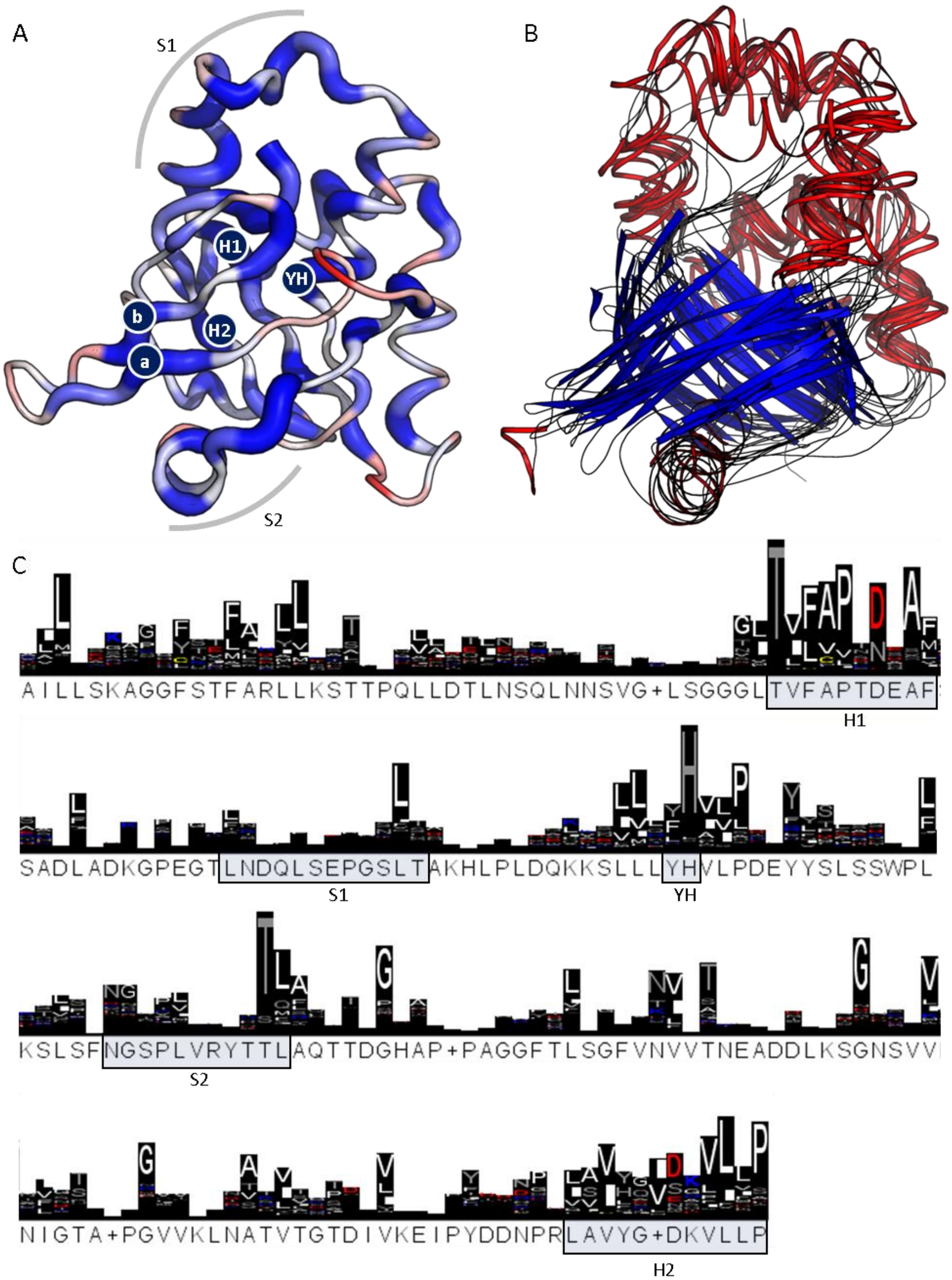


Supp Figure S2 | Length of AG and non-AG regions and multiple sequence alignment for FLAs.
A) Length distribution in amino acids for AG regions. **B)** Length distribution in amino acids for non-AG disordered regions. FLAs containing **C)** a single fasciclin domain (998 sequences) or **D)** two fasciclin 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.



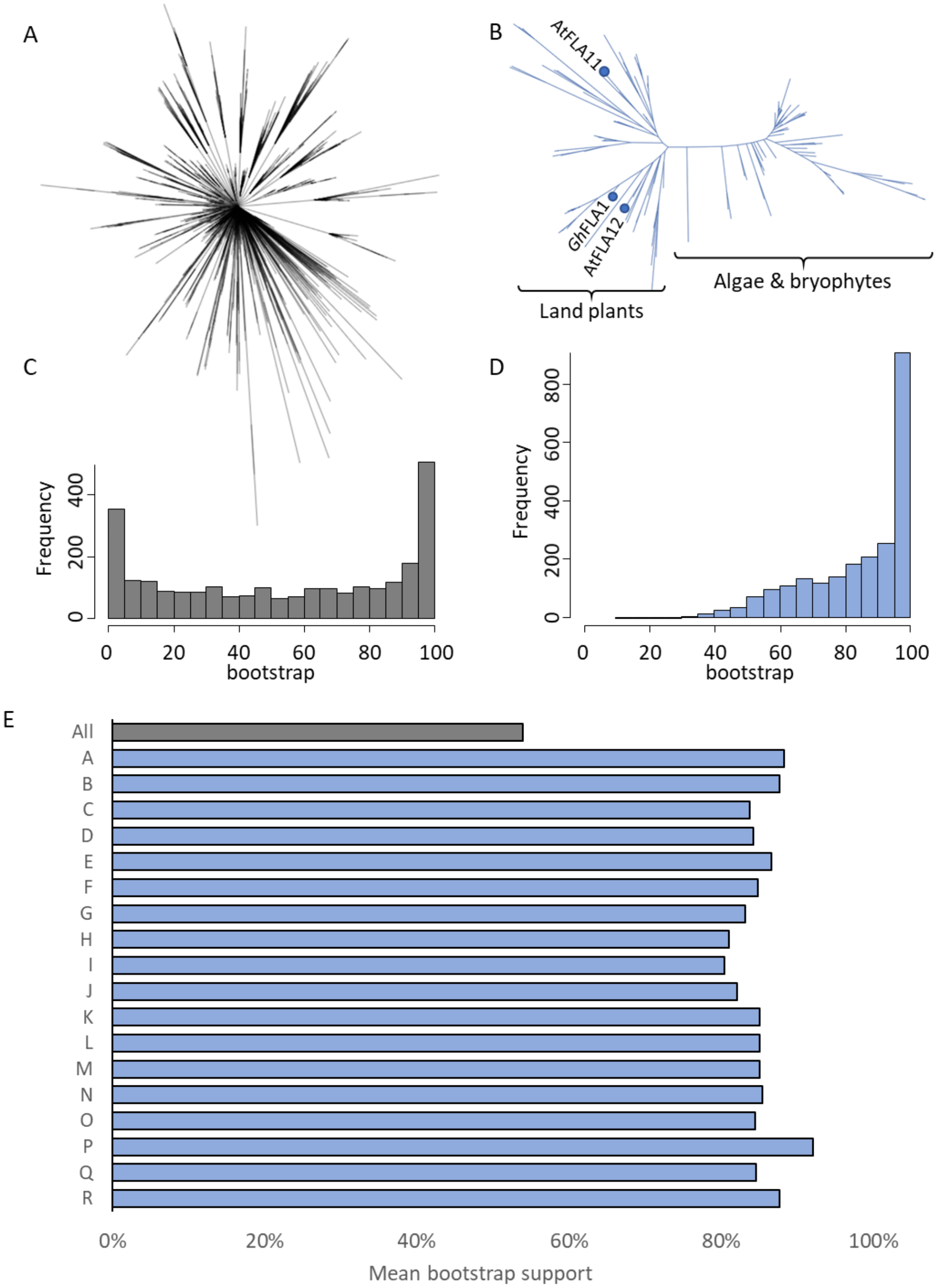
Supp Figure S3 | Amino acid composition benchmarked to relevant databases.

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.

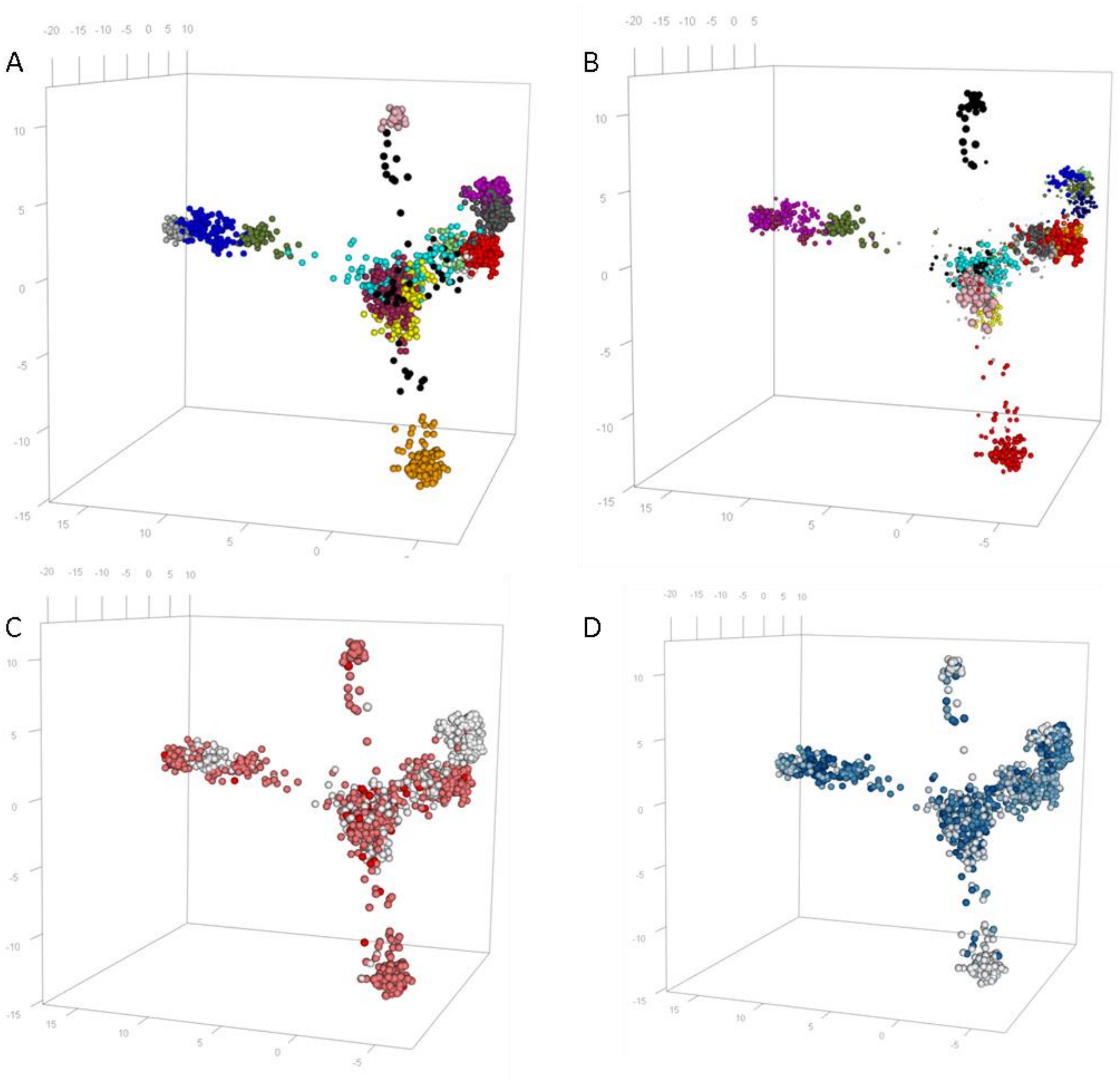

A
S1
H1
YH
b
H2
a
S2
B
C
AILLSKAGGFSTFARLLKSTTPQLLDTLNSQLNNSVG+LSGGGLTVFAPTDEAF
H1
SADLADKGPEGTLNDQLSEPGSLTAKHLPLDQKKSLLLYHVLPDEYYSLSSWPL
S1
YH
KSLSFNGSPLVRYTTLAQTTDGHAP+PAGGFTLSGFVNVVTNEADDLKSGNSVV
S2
NIGTA+PGVVKLNATVTGTDIVKEIPYDDNPRLAVYG+DKVLLP
H2

Supp Figure S4 | Conservation of fasciclin domain sequence and structure.

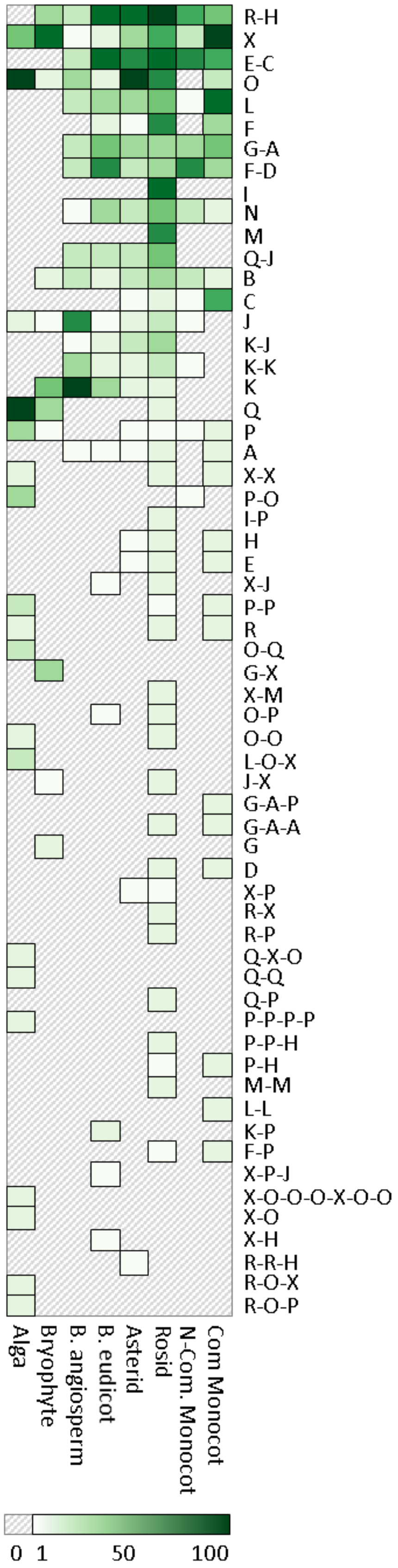
A) Sequence conservation across all 2644 FLA fasciclin domains in the Phytozome dataset is indicated by colour and width, mapped onto a representative homology model of the second fasciclin domain (based on PDB:1o70) from *At*FLA1 (as per Figure 2A). Conserved residues in blue, variable residues in red. 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.

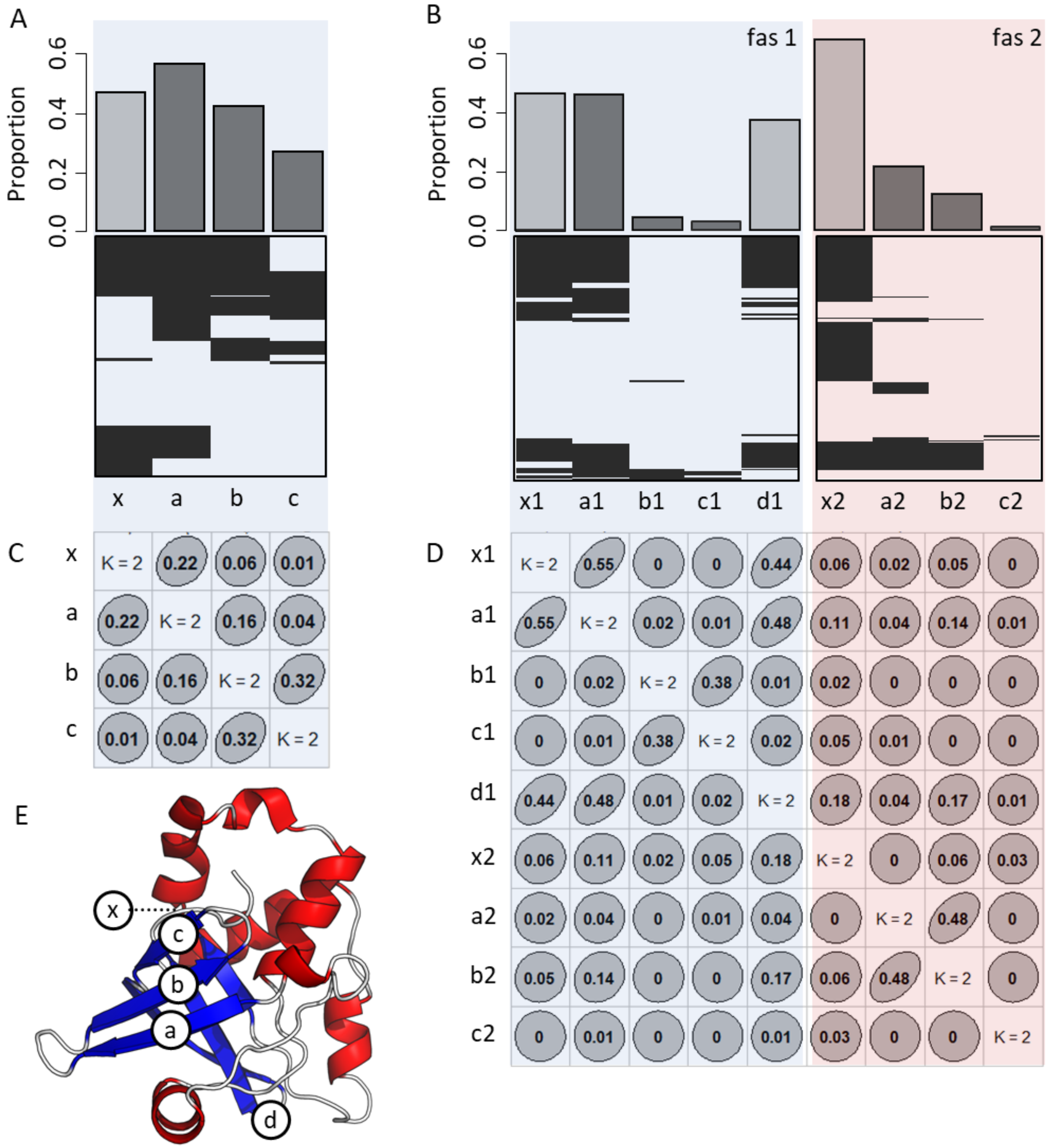


Supp Figure S6 | Sequence space of fasciclin domains, projected by PCA.
For each fasciclin domain, sequence biophysical properties projected by PCA. Coloured based on **A)** Bayesian clustering of PCA projection, **B)** HDBSCAN clustering of UMAP projection (cluster colours as per Figure 3A), **C)** number of fasciclin domains in the FLA, **D)** inter-proline distance.



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

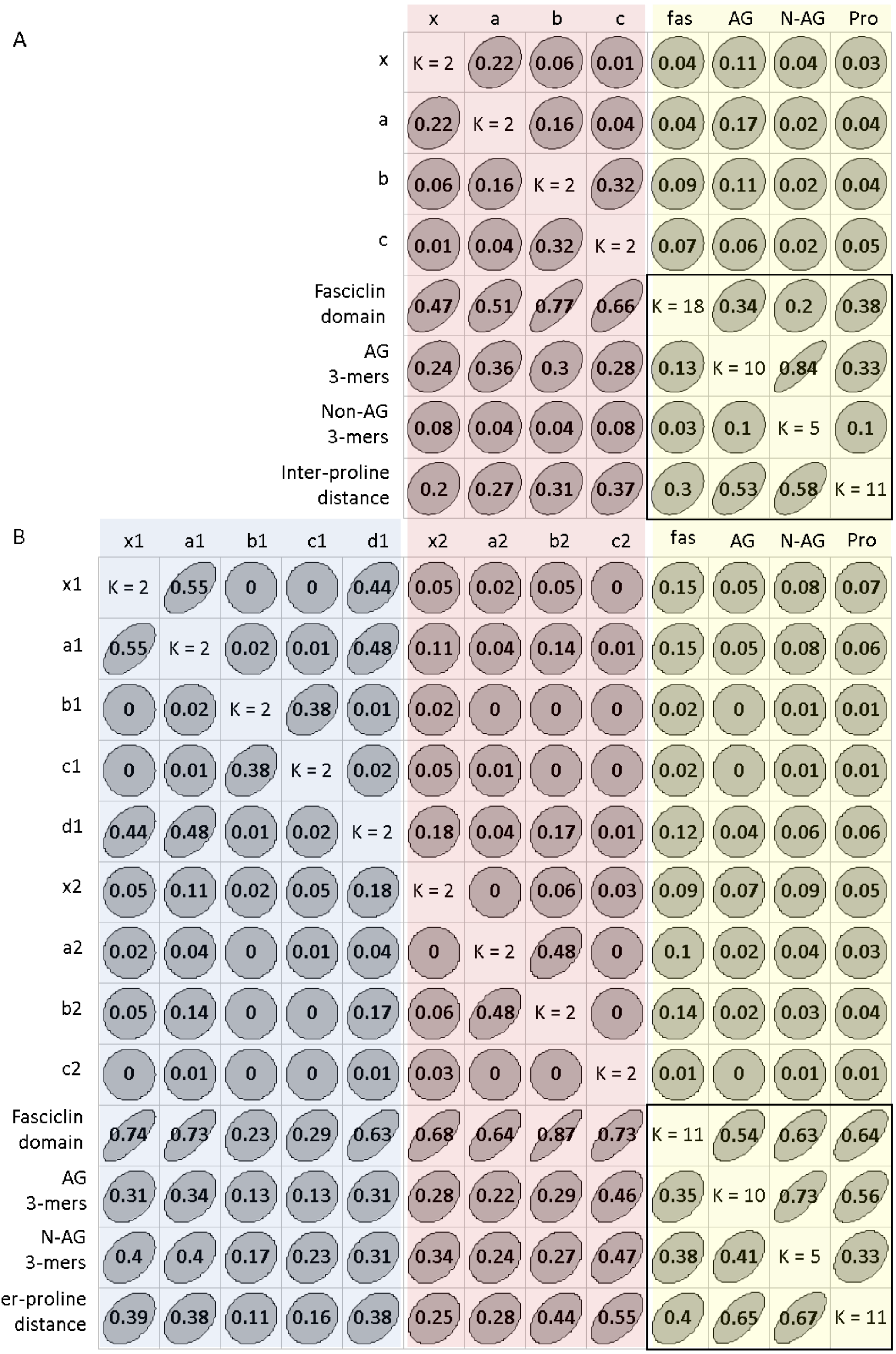
Relative proportion of domain architectures in different major taxonomic groups for the 60 most common domain combinations ordered by overall frequency of occurrence of each domain architecture.



	x	a	b	c
x	K = 2	0.22	0.06	0.01
a	0.22	K = 2	0.16	0.04
b	0.06	0.16	K = 2	0.32
c	0.01	0.04	0.32	K = 2

	x1	a1	b1	c1	d1	x2	a2	b2	c2
x1	K = 2	0.55	0	0	0.44	0.06	0.02	0.05	0
a1	0.55	K = 2	0.02	0.01	0.48	0.11	0.04	0.14	0.01
b1	0	0.02	K = 2	0.38	0.01	0.02	0	0	0
c1	0	0.01	0.38	K = 2	0.02	0.05	0.01	0	0
d1	0.44	0.48	0.01	0.02	K = 2	0.18	0.04	0.17	0.01
x2	0.06	0.11	0.02	0.05	0.18	K = 2	0	0.06	0.03
a2	0.02	0.04	0	0.01	0.04	0	K = 2	0.48	0
b2	0.05	0.14	0	0	0.17	0.06	0.48	K = 2	0
c2	0	0.01	0	0	0.01	0.03	0	0	K = 2

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).



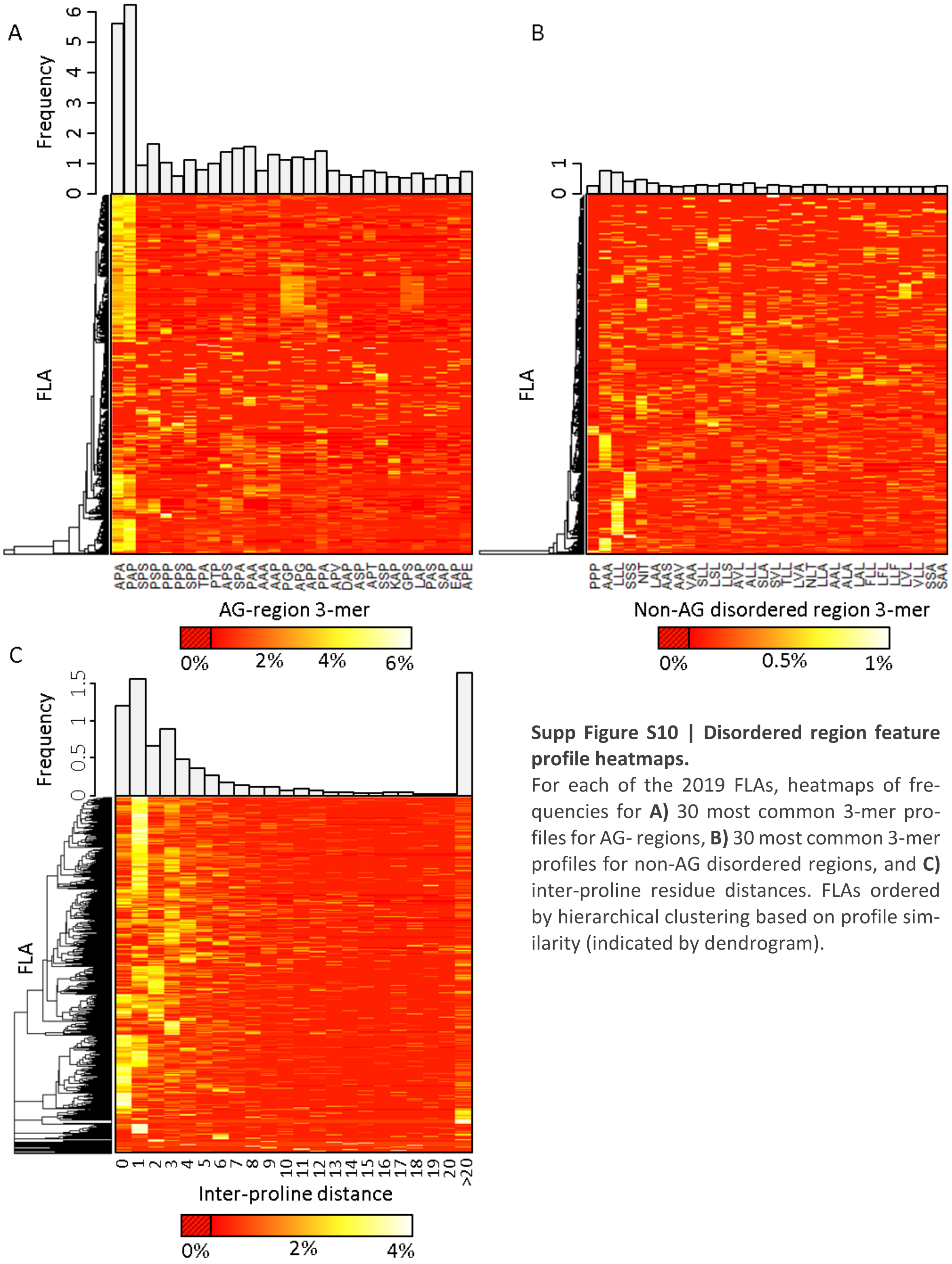
A

	x	a	b	c	fas	AG	N-AG	Pro
x	K = 2	0.22	0.06	0.01	0.04	0.11	0.04	0.03
a	0.22	K = 2	0.16	0.04	0.04	0.17	0.02	0.04
b	0.06	0.16	K = 2	0.32	0.09	0.11	0.02	0.04
c	0.01	0.04	0.32	K = 2	0.07	0.06	0.02	0.05
Fasciclin domain	0.47	0.51	0.77	0.66	K = 18	0.34	0.2	0.38
AG 3-mers	0.24	0.36	0.3	0.28	0.13	K = 10	0.84	0.33
Non-AG 3-mers	0.08	0.04	0.04	0.08	0.03	0.1	K = 5	0.1
Inter-proline distance	0.2	0.27	0.31	0.37	0.3	0.53	0.58	K = 11

B

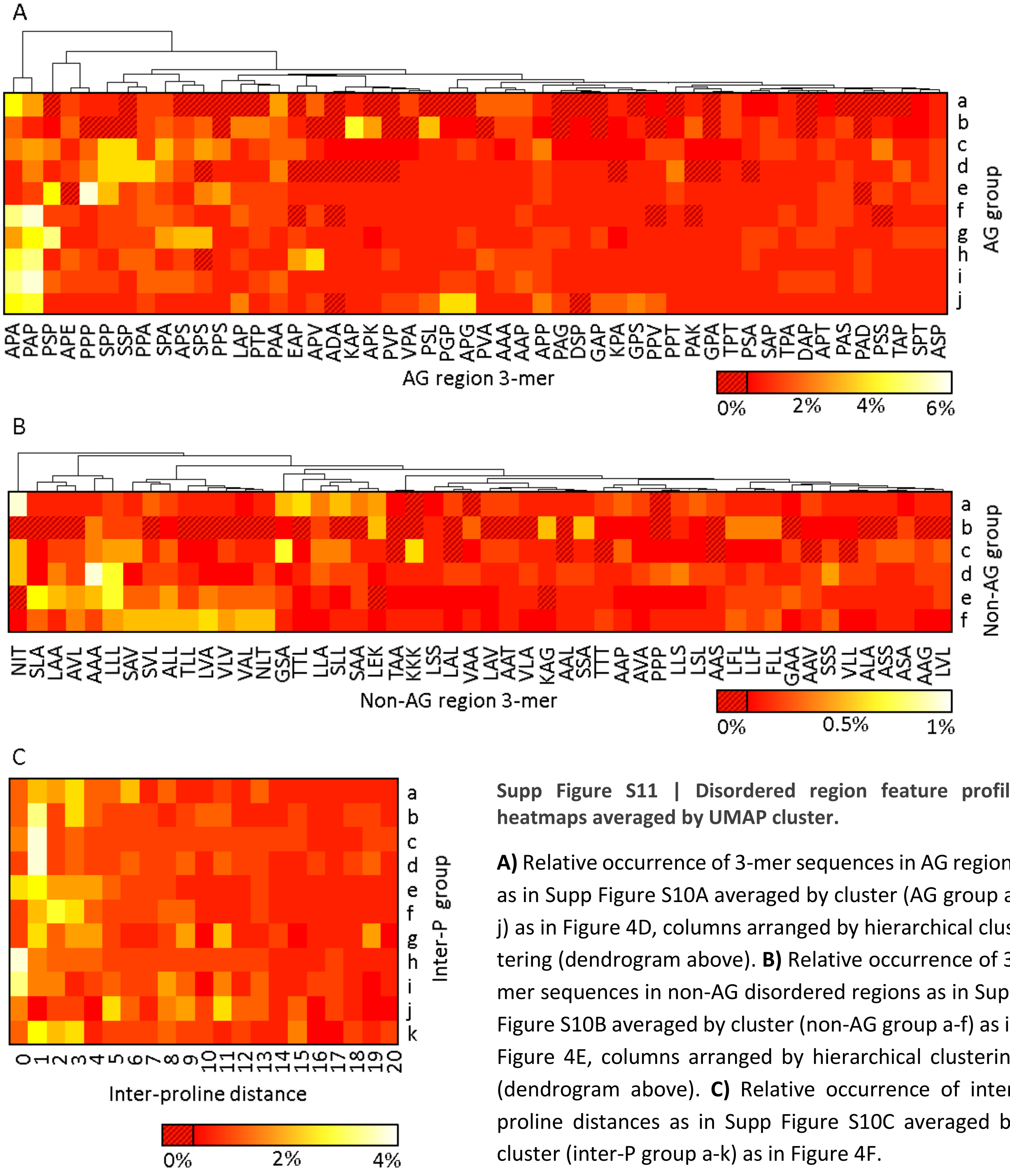
	x1	a1	b1	c1	d1	x2	a2	b2	c2	fas	AG	N-AG	Pro
x1	K = 2	0.55	0	0	0.44	0.05	0.02	0.05	0	0.15	0.05	0.08	0.07
a1	0.55	K = 2	0.02	0.01	0.48	0.11	0.04	0.14	0.01	0.15	0.05	0.08	0.06
b1	0	0.02	K = 2	0.38	0.01	0.02	0	0	0	0.02	0	0.01	0.01
c1	0	0.01	0.38	K = 2	0.02	0.05	0.01	0	0	0.02	0	0.01	0.01
d1	0.44	0.48	0.01	0.02	K = 2	0.18	0.04	0.17	0.01	0.12	0.04	0.06	0.06
x2	0.05	0.11	0.02	0.05	0.18	K = 2	0	0.06	0.03	0.09	0.07	0.09	0.05
a2	0.02	0.04	0	0.01	0.04	0	K = 2	0.48	0	0.1	0.02	0.04	0.03
b2	0.05	0.14	0	0	0.17	0.06	0.48	K = 2	0	0.14	0.02	0.03	0.04
c2	0	0.01	0	0	0.01	0.03	0	0	K = 2	0.01	0	0.01	0.01
Fasciclin domain	0.74	0.73	0.23	0.29	0.63	0.68	0.64	0.87	0.73	K = 11	0.54	0.63	0.64
AG 3-mers	0.31	0.34	0.13	0.13	0.31	0.28	0.22	0.29	0.46	0.35	K = 10	0.73	0.56
N-AG 3-mers	0.4	0.4	0.17	0.23	0.31	0.34	0.24	0.27	0.47	0.38	0.41	K = 5	0.33
Inter-proline distance	0.39	0.38	0.11	0.16	0.38	0.25	0.28	0.44	0.55	0.4	0.65	0.67	K = 11

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.



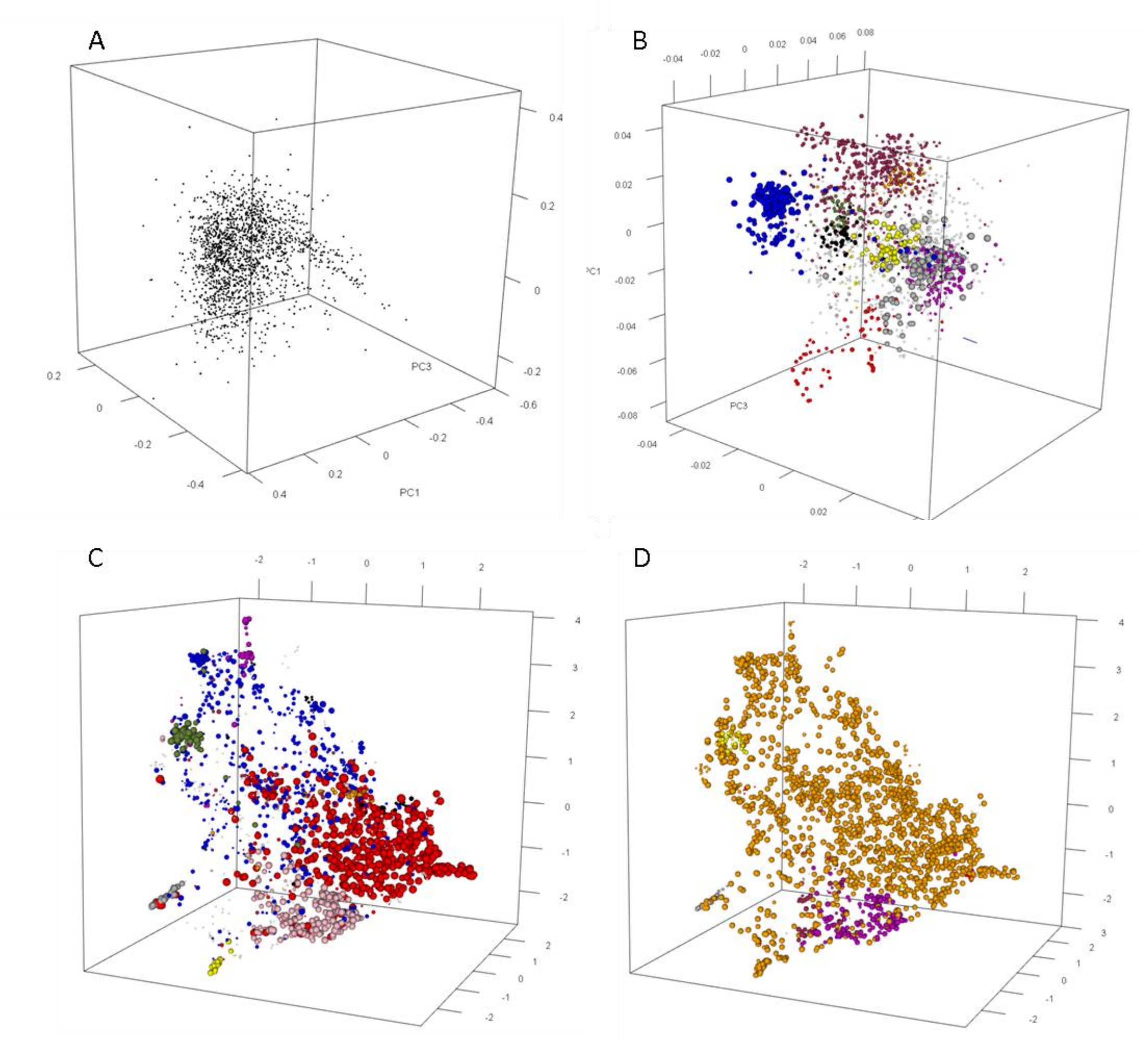
Supp Figure S10 | Disordered region feature profile heatmaps.

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.

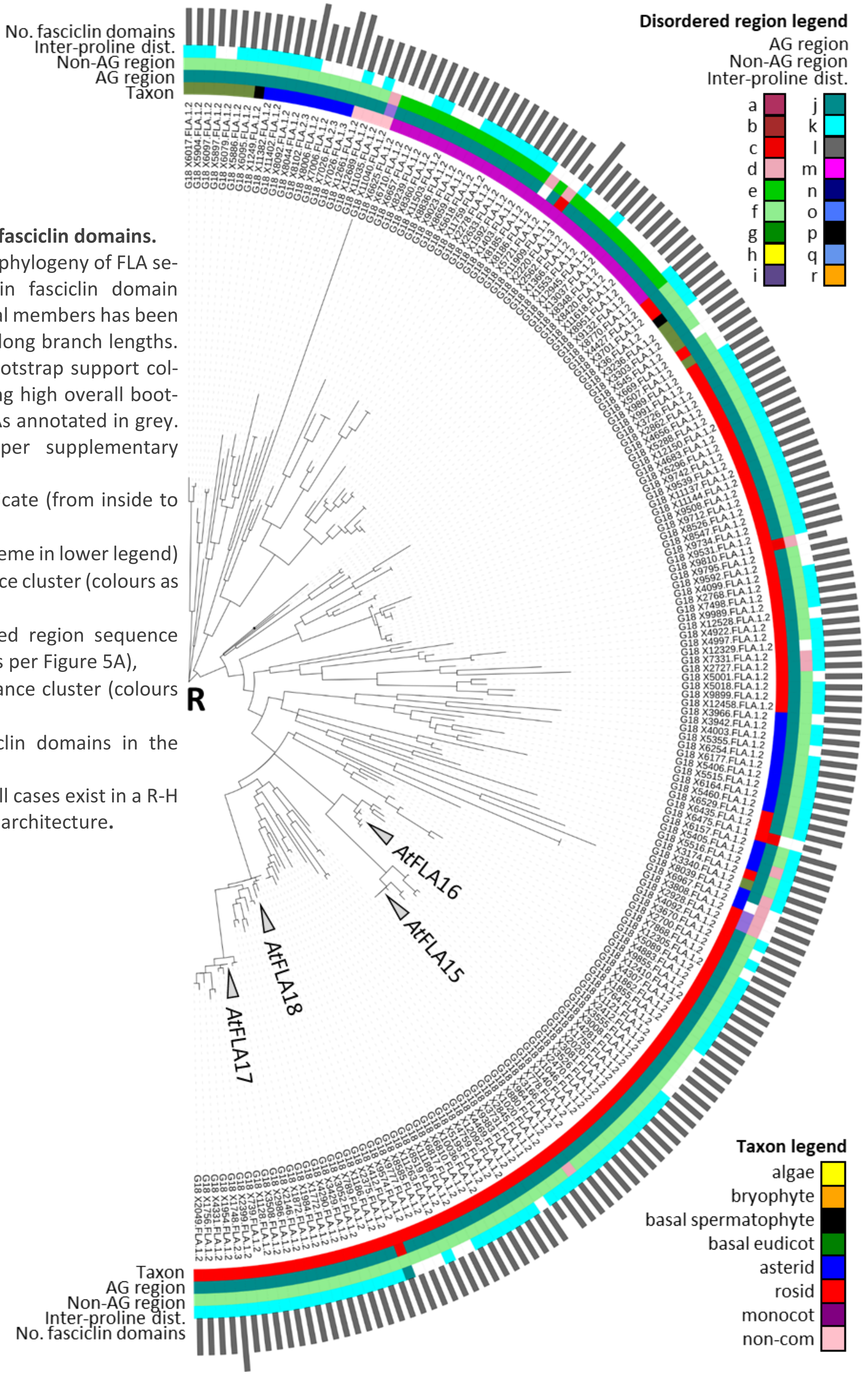
A) For each FLA, distribution of inter-proline distances projected by PCA. No detectable clusters are identifiable by this method. **B)** PCA projection coloured by clusters identified by HDBSCAN clustering of UMAP projection of inter-proline residue distances (as in Figure 4F). **C)** UMAP projection 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).

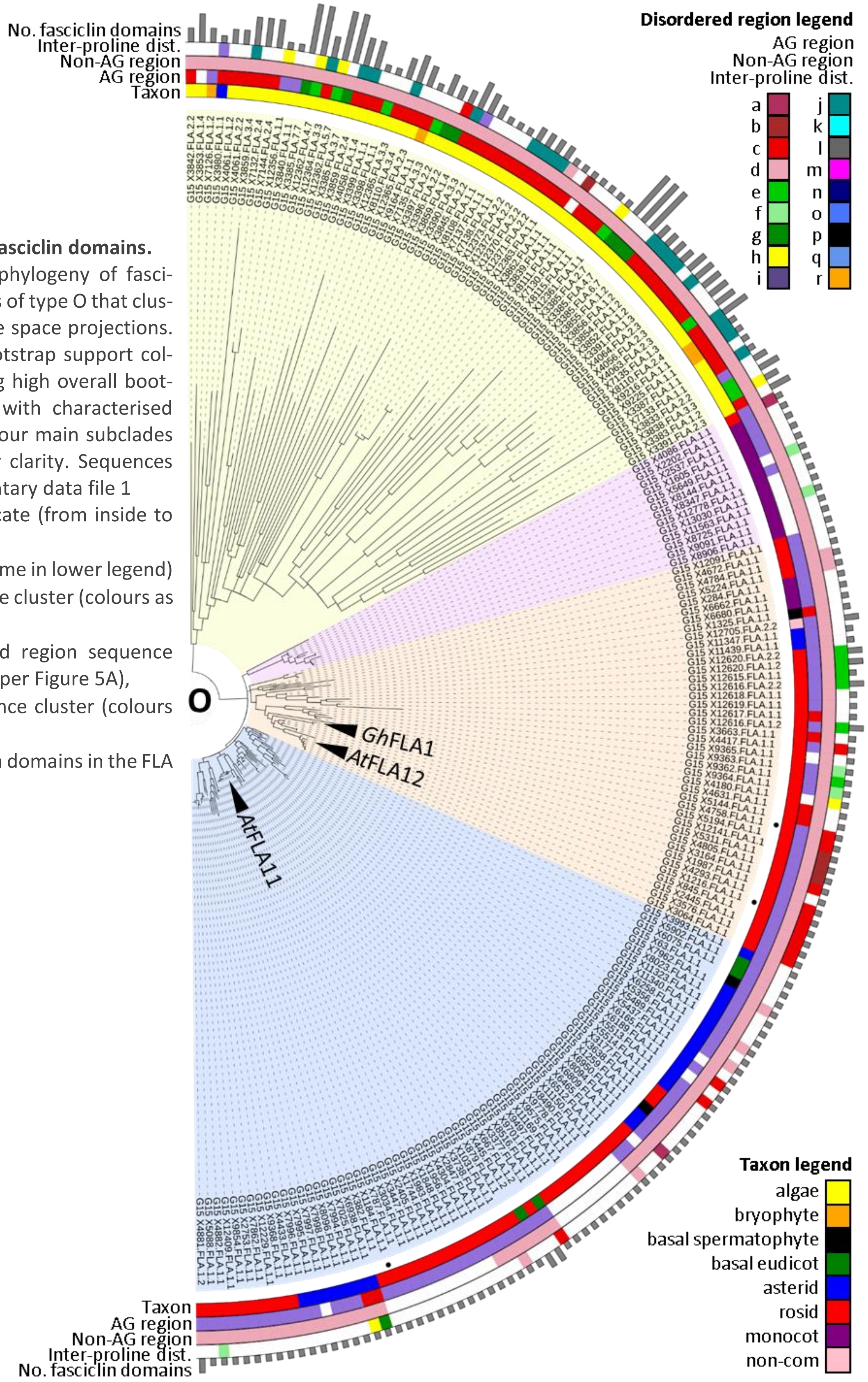
Supp Figure S13 |
Phylogeny of type R fasciclin domains.
Maximum likelihood phylogeny of FLA sequences that contain fasciclin domain type R. A clade of algal members has been omitted due to very long branch lengths. Nodes with <50% bootstrap support collapsed (demonstrating high overall bootstrap support). *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)

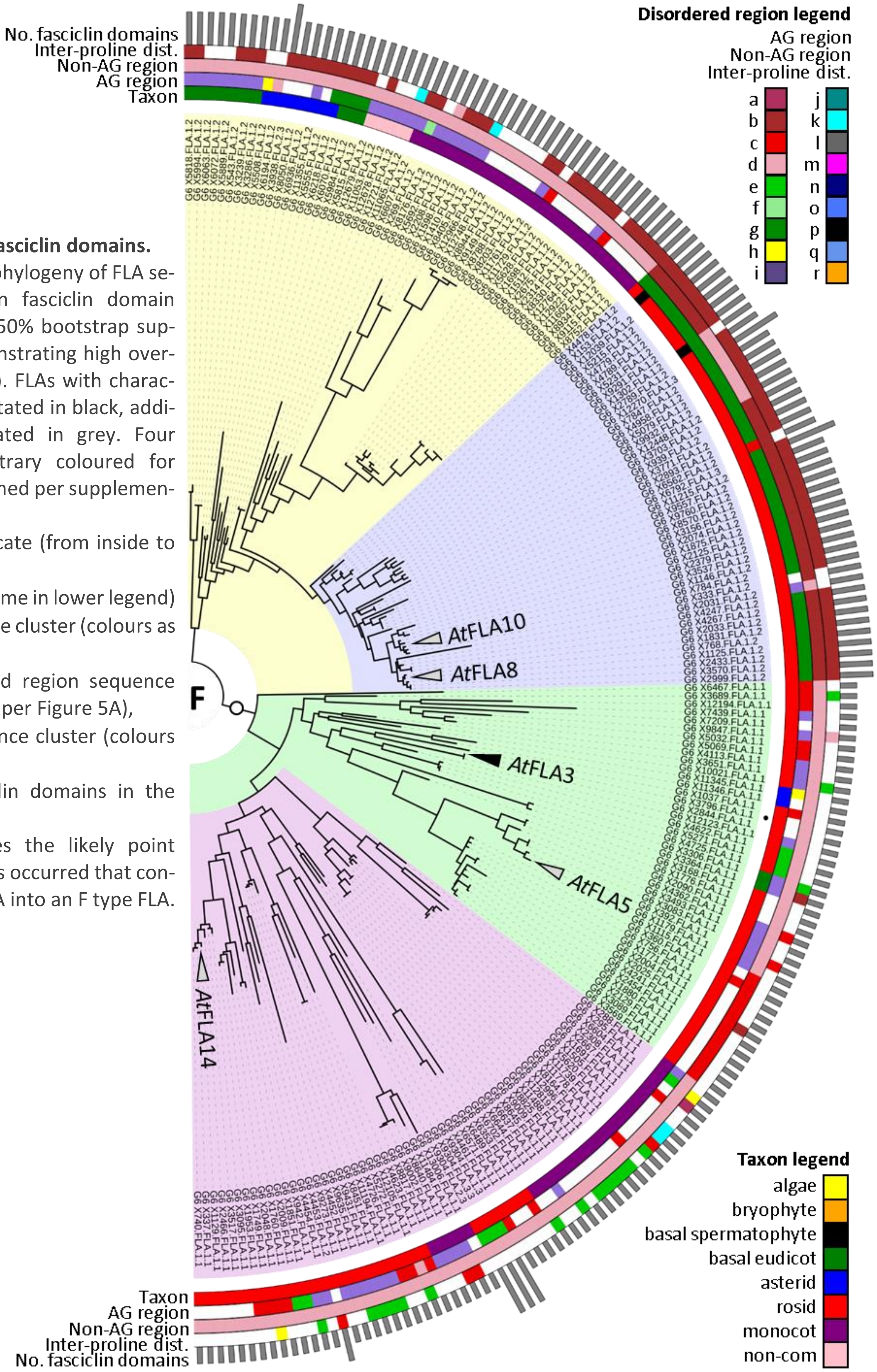
Note that in almost all cases exist in a R-H two-fasciclin domain architecture.





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 5A),
- Inter-proline distance cluster (colours as per Figure 5C)
- Number of fasciclin domains in the FLA (max = 7)

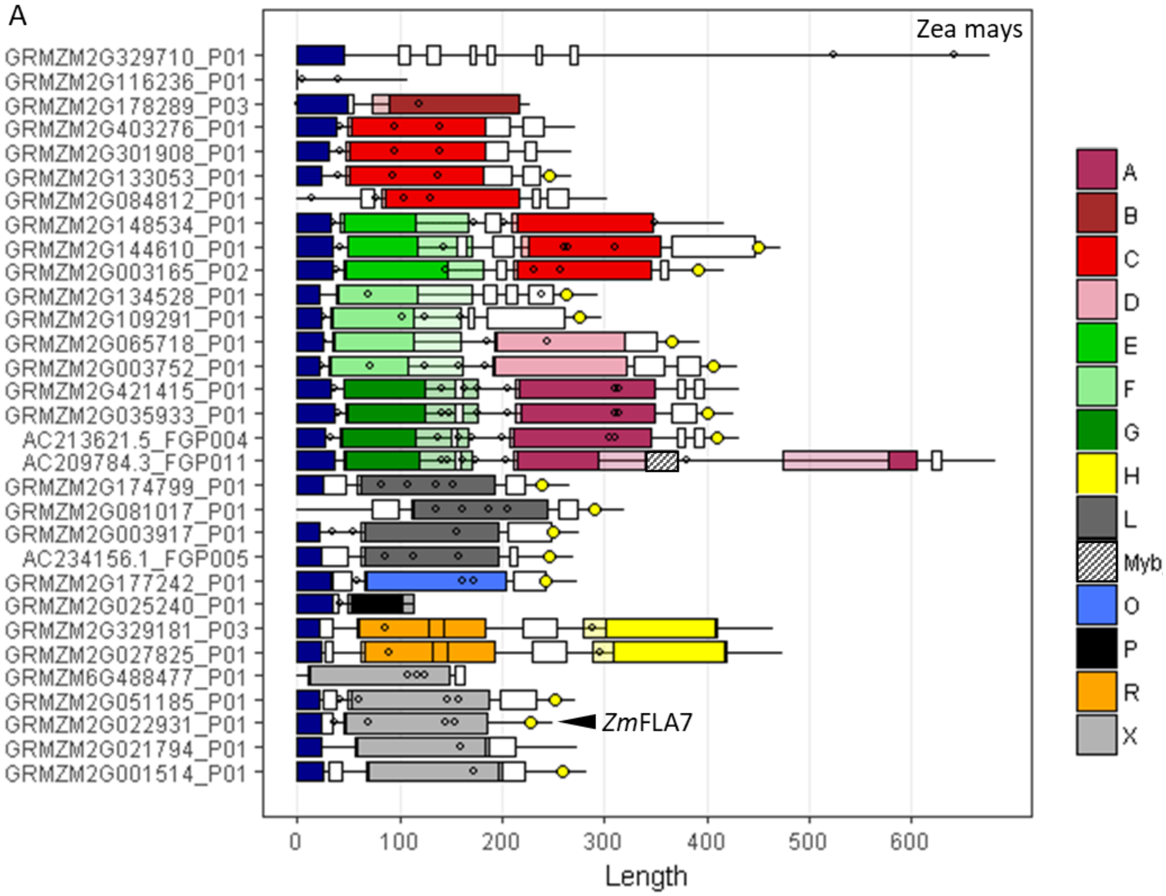


Supp Figure S15 |
Phylogeny of type F fasciclin domains.
Maximum likelihood phylogeny of FLA sequences that contain fasciclin domain type F. Nodes with <50% bootstrap support collapsed (demonstrating high overall bootstrap support). FLAs with characterised function annotated in black, additional *At*FLAs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

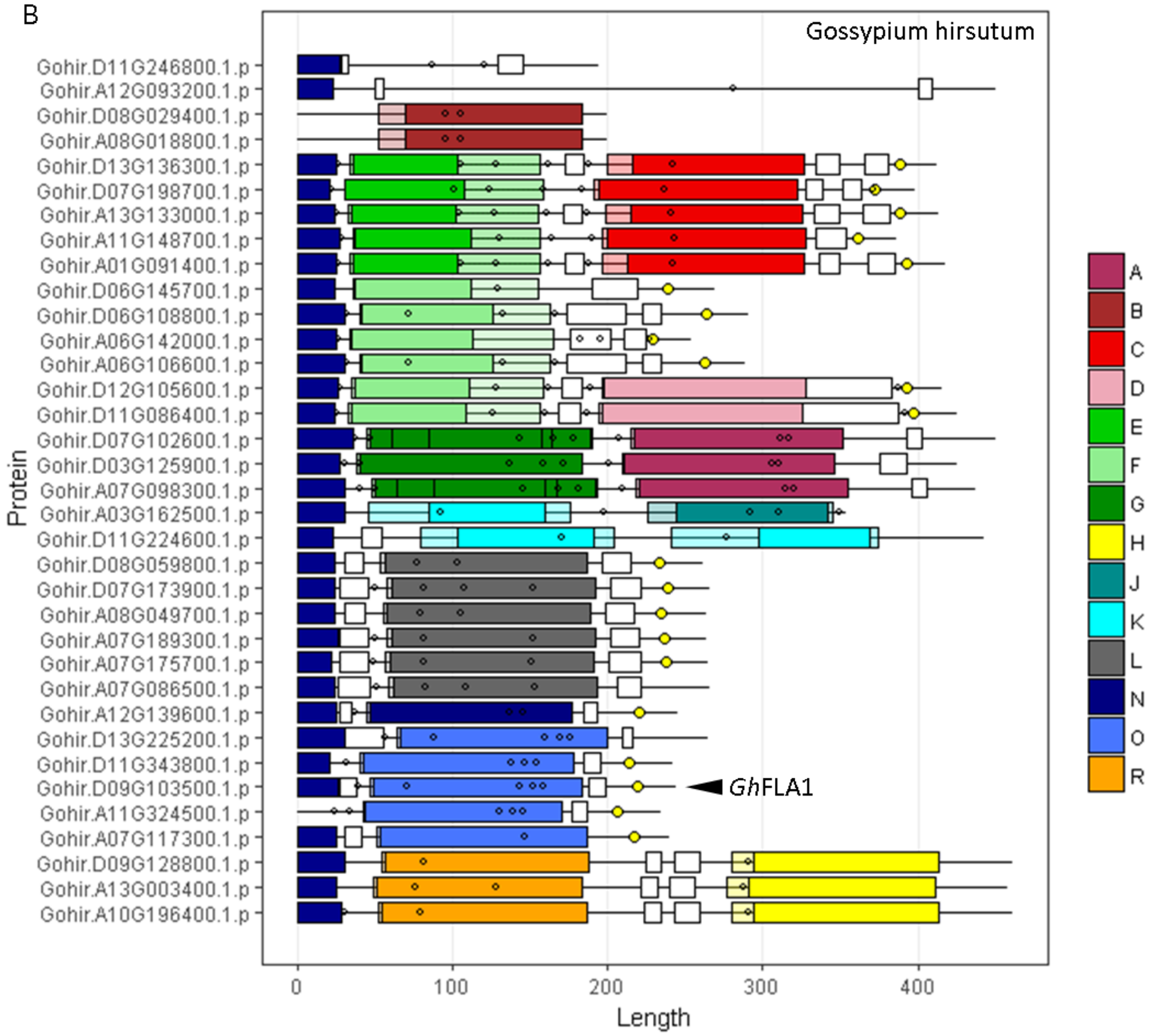
Annotation rings indicate (from inside to outside):

- 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)

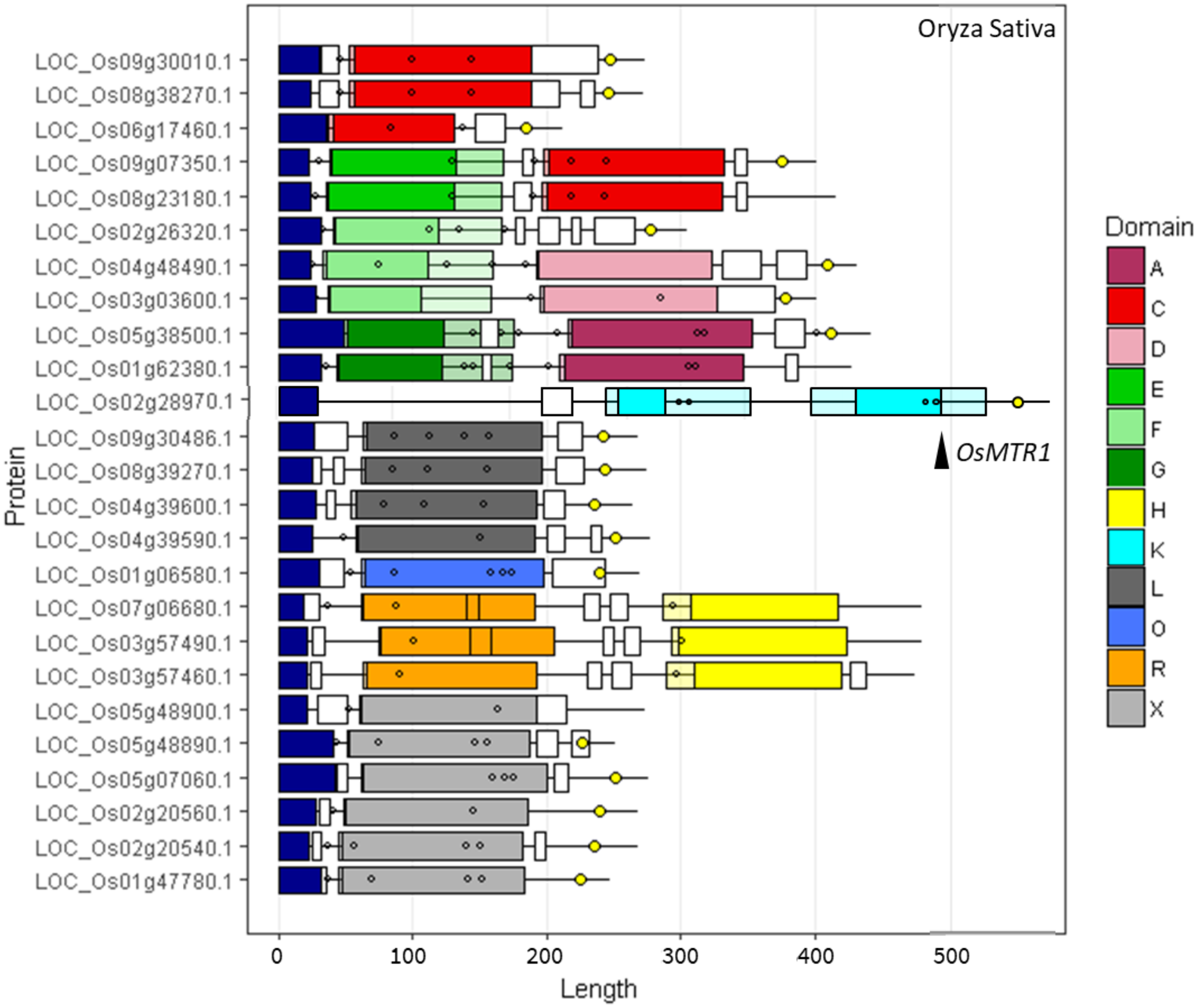
White circle indicates the likely point where the domain loss occurred that converted an F-D type FLA into an F type FLA.



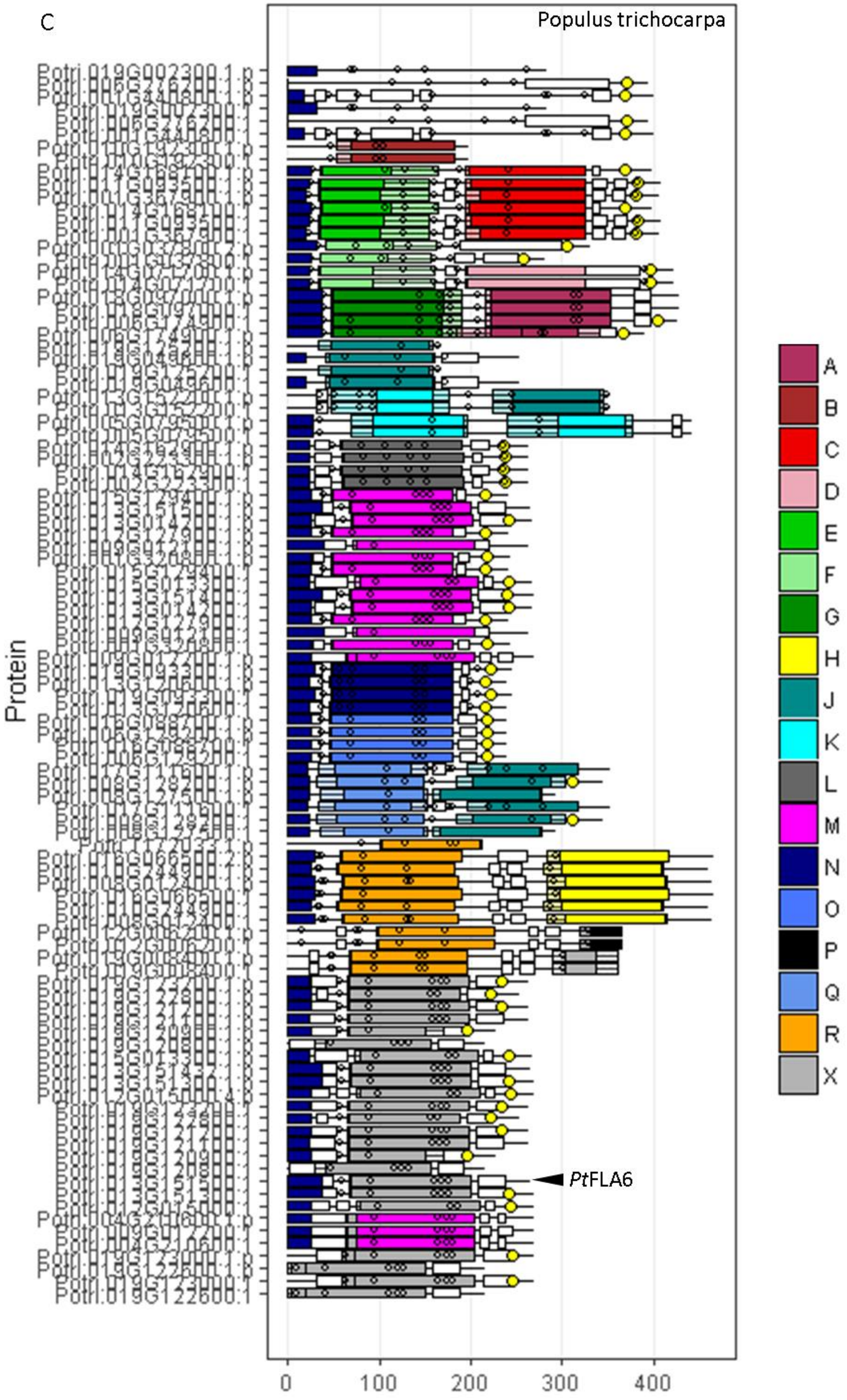
Supp Figure S16 | Schematic representation of FLA members in selected species.
FLA complement for **A)** *Zea mays*, **B)** *Populus trichocarpa*, **C)** *Gossypium hirsutum*, **D)** *Oryza sativum*, **E)** *Brassica rapa* (closest relative to *Brassica carinata* in Phytozome) and **F)** *Eucalyptus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



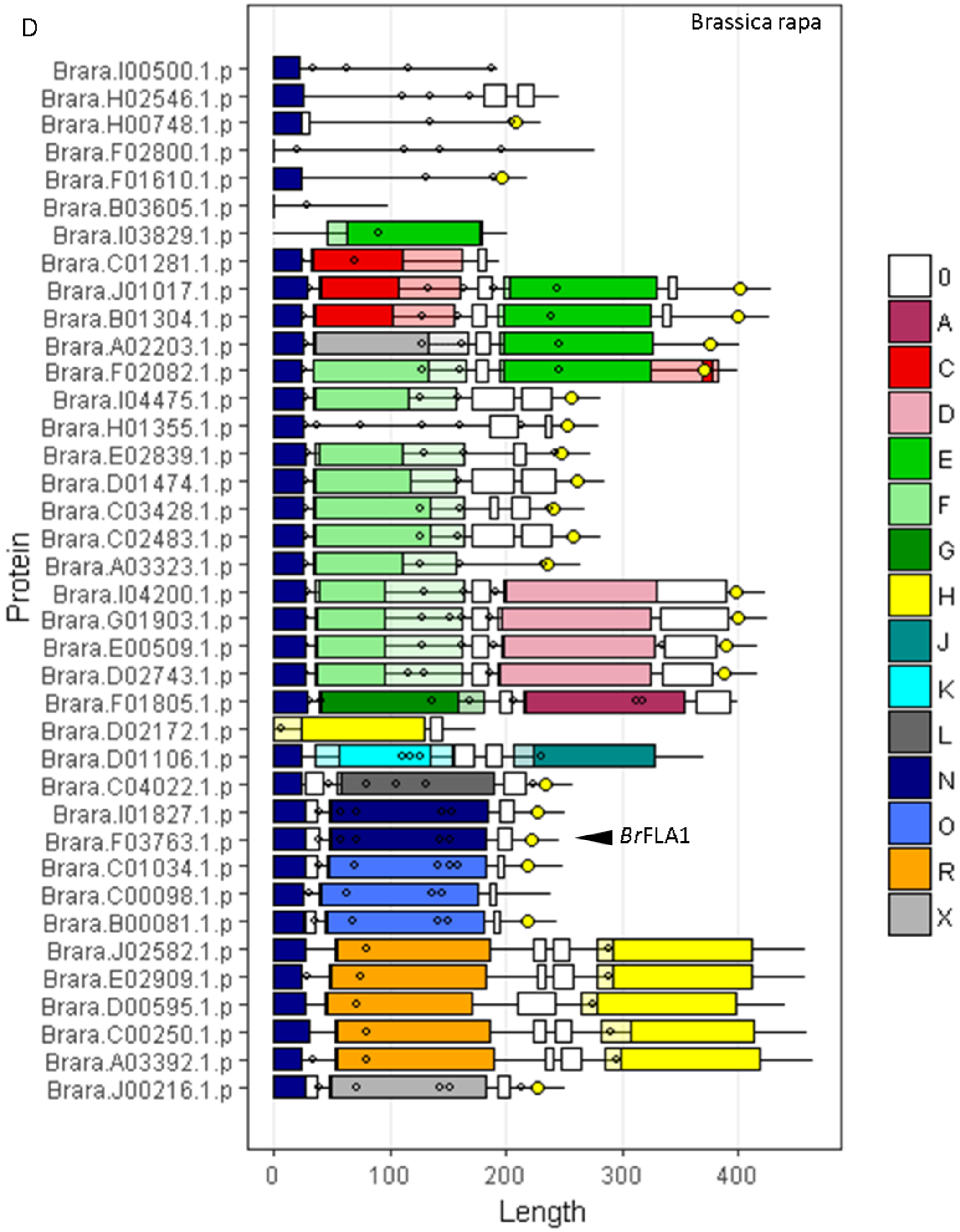
Supp Figure S16B (see legend above)



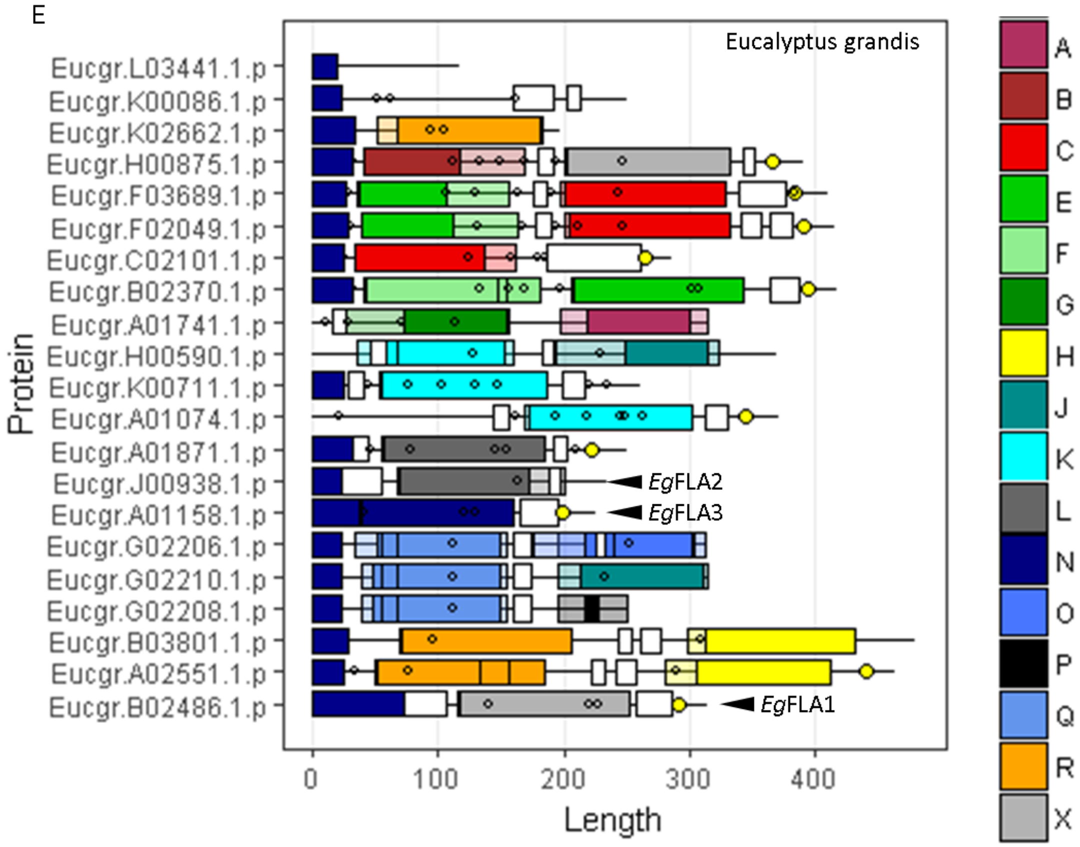
Supp Figure S16**C (see legend above)**



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].