PNAS RESEARCH ARTICLE | EVOLUTION

Proteomic analysis of the sponge Aggregation Factor implicates an ancient toolkit for allorecognition and adhesion in animals

Fabian Ruperti[a], Monika Dzieciatkowska[b], M. Sabrina Pankey[c], Cedric S. Asensio[d], Dario Anselmetti[e], Xavier Fernàndez-Busquets[f,g], and Scott A. Nichols[d,1]

Affiliations are included on p. 9.

Edited by Nicole King, University of California Berkeley, Berkeley, CA; received May 7, 2024; accepted November 12, 2024

The discovery that sponges (Porifera) can fully regenerate from aggregates of dissociated cells launched them as one of the earliest experimental models to study the evolution of cell adhesion and allorecognition in animals. This process depends on an extracellular glycoprotein complex called the Aggregation Factor (AF), which is composed of proteins thought to be unique to sponges. We used quantitative proteomics to identify additional AF components and interacting proteins in the classical model, *Clathria prolifera*, and compared them to proteins involved in cell interactions in Bilateria. Our results confirm MAFp3/p4 proteins as the primary components of the AF but implicate related proteins with calx-beta and wreath domains as additional components. Using AlphaFold, we unveiled close structural similarities of AF components to protein domains in other animals, previously masked by the mutational decay of sequence similarity. The wreath domain, believed to be unique to the AF, was predicted to contain a central beta-sandwich of the same organization as the vWFD domain (also found in extracellular, gel-forming glycoproteins in other animals). Additionally, many copurified proteins share a conserved C-terminus, containing divergent immunoglobulin (Ig) and Fn3 domains predicted to serve as an AF–interaction interface. One of these proteins, MAF-associated protein 1, resembles Ig superfamily cell adhesion molecules and we hypothesize that it may function to link the AF to the surface of cells. Our results highlight the existence of an ancient toolkit of conserved protein domains regulating cell–cell and cell–extracellular matrix protein interactions in all animals, and likely reflect a common origin of cell adhesion and allorecognition.

proteomics | evolution | Porifera | adhesion | allorecognition

Significance

Multicellularity in animals depends on the ability of cells to recognize and selectively attach to each other, to the exclusion of "non-self" cells. In sponges (Porifera), this is accomplished by a large, extracellular molecular complex called Aggregation Factor (AF). The AF is believed to be unique to sponges, contributing to a long-held view that they lack "true epithelia" as found in other animals. By leveraging proteomic and structural analyses of the AF, we provide evidence linking the AF to proteins involved in cell–cell/extracellular matrix protein (ECM) interactions in other animals. This work highlights the power of protein structure-based similarity searches to untangle ancient evolutionary questions and implicates an ancestral protein domain toolkit for cell attachment and self-recognition.

The theoretical requirements for the evolution of multicellularity include adhesion mechanisms for cell–cell attachment (1), signaling mechanisms to guide development and coordinate interactions between cells (2), and allorecognition mechanisms to distinguish self from nonself (3). In conventional animal models, much is known about how these conditions are met, but due to their phylogenetic placement (Fig. 1*A*) comparative studies of sponges, ctenophores, and cnidarians are needed to reconstruct the earliest events in the evolution of animal multicellularity.

In 1907, Henry Wilson's seminal experiments with *Clathria* (*Microciona*) *prolifera* showed that sponges can be dissociated into a heterogeneous suspension of cells which can reassemble into a functional, intact organism (4). For over a century since, researchers have used "aggregation assays" to study the mechanisms of cell adhesion and allorecognition in sponges. The prevailing model proposes that species-specific aggregation depends on a sulfated proteoglycan, termed the "Aggregation Factor" (AF), found in the extracellular matrix (ECM) and bound to the surface of cells (5–11). In aggregation assays performed with synthetic beads, only those carrying AFs from the same species associate, showing that AFs are sufficient for species-specific aggregation in vitro (12, 13).

The composition of the AF varies between species, consisting of 50 to 70% protein and 30 to 50% carbohydrate (14–16). In some species, individual AF molecules can be linear, similar to proteoglycans in other animals, but others have a unique "sunburst" architecture with a central ring and radiating arms (14, 16) (Fig. 1*B*). The sunburst-like AF of *C. prolifera* has a molecular weight of ~2×10^7 Da and is composed of two main proteins that derive from the precursor protein Microciona Aggregation Factor p3 (MAFp3)/p4 (Fig. 1*C*) (14, 17–19). Twenty copies of the ~50 kDa MAFp3 subunit comprise the "ring" of the sunburst, each decorated with a 200 kDa glycan (g-200) (20, 21). Twenty subunits of the ~400 kDa MAFp4 subunit comprise the radiating "arms" of the AF, attached to the MAFp3-ring and decorated with about 50 copies of a 6 kDa glycan (g-6) (14, 17–19, 22).

Author contributions: S.A.N. designed research; M.D., C.S.A., D.A., X.F.-B., and S.A.N. performed research; F.R., M.D., M.S.P., and S.A.N. analyzed data; and F.R. and S.A.N. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: scott.nichols@du.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2409125121/-/DCSupplemental.

Published December 18, 2024.

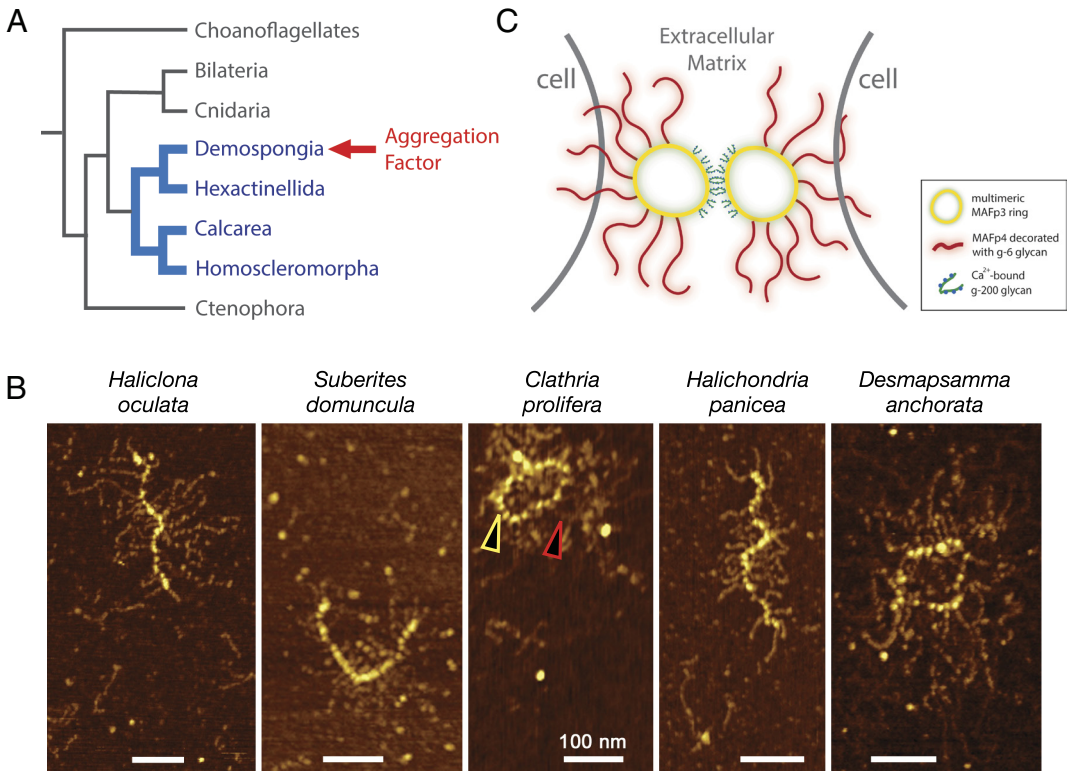

Fig. 1. The Aggregation factor (AF) is composed of glycoproteins that form linear or circular structures with radiating arms. (*A*) Phylogenetic placement of the four major lineages of sponges (blue). The AF is believed to be unique to demosponges. (*B*) Atomic force microscopy (AFM) images of linear vs. ring-like AF purified from different demosponge species. AF purification and AFM imaging were performed as described in the *Materials and Methods* Section (red arrowhead = MAFp4 arms, yellow arrowhead = MAFp3 ring). (*C*) Cartoon depiction of *Clathria prolifera* AF. Wreath domain-containing MAFp3 makes up the central ring and is decorated with g-200 glycans responsible for Ca^{2+}-dependent AF–AF interactions. Calx-beta domain-containing MAFp4 arms are decorated with g-6 glycan responsible for Ca^{2+}-independent AF–cell interactions, but it is unclear how the AF attaches to the surface of cells.

Ca^{2+}-dependent carbohydrate–carbohydrate interactions between g-200 glycans determine the specificity needed for AF–AF binding (20). In contrast, AF–cell binding is thought to be mediated by Ca^{2+}-independent interactions between the g-6 glycan with an unidentified 68 kDa lectin, hypothesized to then bridge to an integral membrane receptor (19, 23). In support of this, AF–AF and AF–cell binding affinities were recovered upon chemical crosslinking of protein-free glycans isolated from the AF complex into large multivalent structures (21, 22, 24).

The *C. prolifera* AF has been described to contain at least 11 different protein subunits that could be dissociated upon Ca^{2+} removal (25). However, MAFp3 and MAFp4 are the only unequivocally identified components, containing calx-beta domains and a novel domain termed the "wreath" domain due to its presence in the central ring of the AF (26). The predicted wreath domain in the MAFp3 subunit has previously only been found in demosponges (26), a finding interpreted as evidence that the AF is unrelated to adhesion and allorecognition mechanisms in other organisms. In contrast, calx-beta domains of MAFp4 were first identified in the cytoplasmic region of Na+/Ca2+ exchangers where they function to remove free cytosolic Ca^{2+} (27), and have subsequently been detected in the cytoplasmic tail of human integrin α6β4 subunit (28). In the extracellular space, calx-beta domains are known from Fras1-related ECM proteins (29, 30), the ectodomain of very large G protein-coupled receptor 1 (31), ECM3 in sea urchin (32), a beta-glucosidase from the acellular slime mold *Physarum polycephalum* (33), and from the cyanobacterium *Synechocystis* sp. PCC6803 (27). Calx-beta domains adopt an Ig-like beta-sandwich fold, similar to Ig-like and fibronectin 3 (Fn3) domains (SCOP family 48725), and typically bind Ca^{2+}.

Comparative genomics data support the possible existence of additional AF protein components. Specifically, demosponge genomes encode many additional genes homologous to MAFp3/p4. Some have a conserved wreath domain in combination with other elements including calx-beta, von Willebrand Factor A/D (vWFA/D), and Plexin-Semaphorin-Integrin domains. Lower confidence candidate AF components apparently lack a wreath domain but have conserved calx-beta domains and Basic Local Alignment Search Tool (BLAST) to MAFp3/p4 (26). These have not been experimentally validated to interact with the AF, but in the genome of *Amphimedon queenslandica* they have genomic features shared among allorecognition genes in diverse organisms (reviewed in ref. 3). For example, they are clustered together in tandem on a single chromosome, most have signal peptides indicating that they are secreted, and their structural features (such as calx-beta domains) vary in number between paralogs and exhibit high sequence variation within the population, suggesting rapid evolution. Repeated, highly polymorphic protein domains help ensure that cells with matching receptor/ligand combinations reliably distinguish between self/nonself and that compatible combinations do not occur by chance (3).

Other AF or AF-interacting proteins have been detected experimentally but remain incompletely identified at the sequence-level. Specifically, Varner (11, 34) used the AF as a probe to purify candidate AF-binding proteins in *C. prolifera*. Two proteins of ~68 kDa and ~210 kDa were found to bind with high affinity to the AF in the ECM, and to the cell surface (11, 23, 34). Also, using dissociative gel fractionation of the *C. prolifera* AF, Fernandez-Busquets and colleagues (18) identified two discrete bands at 210 kDa (p210) and 2,000 kDa (S2). They determined the sequence of short peptides from these bands, but protein database searches at the time did not yield significant similarities to known proteins. They speculated that the 210 kDa band was likely the same as that identified by Varner (11, 34), and demonstrated

that it was a glycosylated protein with interindividual polymorphism in the glycan moiety (17). This observation was consistent with the role of AFs as determinants of individuality within a species (17, 35) and suggests that differential glycosylation could be a way to distinguish between self and nonself.

To comprehend how the AF model of adhesion/allorecognition fits into the broader context of multicellular evolution, a deeper understanding of its protein backbone is required.

Here, we used a proteomics approach to identify AF and AF-associated proteins in *C. prolifera*. Like MAFp3/p4, other abundant proteins were found to contain wreath and/or calx-beta domains, or else a conserved C-terminus that potentially serves as an interaction interface with the AF. Structural analyses highlighted striking similarities to domains present in extracellular glycoproteins of bilaterians, such as vWFD, Ig-like, and Fn3 domains, despite low sequence similarity. Furthermore, identified AF components have low predicted isoelectric points, leading to negative charges at neutral pH—a common characteristic of secreted Ca^{2+}-binding proteins in Bilaterians. These results expand our understanding of the structure, composition, and endogenous interactions of the AF, and provide clues to the existence of an ancient protein domain toolkit shared between adhesion, ECM, and allorecognition proteins in sponges, cnidarians, and bilaterians.

Results

De Novo Assembly of a *C. prolifera* Proteome Database. To create a reference database for proteomic analyses, we first sequenced and assembled the transcriptome of *C. prolifera* using RNA isolated from whole, adult tissues. A Novaseq pe150 run (Novagene) produced 42,531,984 150 bp paired-end reads, which were assembled to yield 22,794 predicted transcripts. The final assembly had a BUSCO v3 completeness score of 89.6% and a BUSCO v4 completeness score of 85.1%. We deposited raw reads in the NCBI Sequence Read Archive (SRX18275041) and archived the TransPI assembly and corresponding peptide predictions on Figshare (36).

We prepared AF samples for proteomic analyses in three ways (*SI Appendix*, Fig. S1). First, we isolated a "Crude" AF extract based on the method of Humphreys (37). Briefly, we dissociated live tissue in calcium/magnesium-free seawater (CMFSW), removed cells and spicules by centrifugation, and then precipitated the AF from the supernatant by the addition of $CaCl_2$ to form a red, gel-like pellet. To remove pigments, membranes, and possible contaminants, we redissolved the Crude AF sample in CMFSW, centrifuged it, and passed it through a 0.22 µm filter (34). When we again added $CaCl_2$ to precipitate the AF it formed a fluffy, white pellet ("Filtered" sample). As a final purification step, we resolubilized the precipitate from the Filtered sample for fractionation by size exclusion chromatography. The upper limit of the fractionation range of this column was 2×10^7 Da, which is the reported molecular weight of the *C. prolifera* AF (14, 17, 18). We then combined and concentrated the fractions expected to contain the AF for analysis as the "Size-exclusion Chromatography (SEC)" sample.

Proteomic analyses of the Crude AF extract detected a total of 690 proteins. Analysis of the Filtered and SEC samples resulted in the detection of 437 and 696 proteins, respectively. A total of 319 unique proteins were common to all samples (Fig. 2*A*). Most detected proteins had a very low relative abundance (Fig. 2*B*). Proteins detected exclusively in any one of the three samples had a proportion <5% in their respective samples.

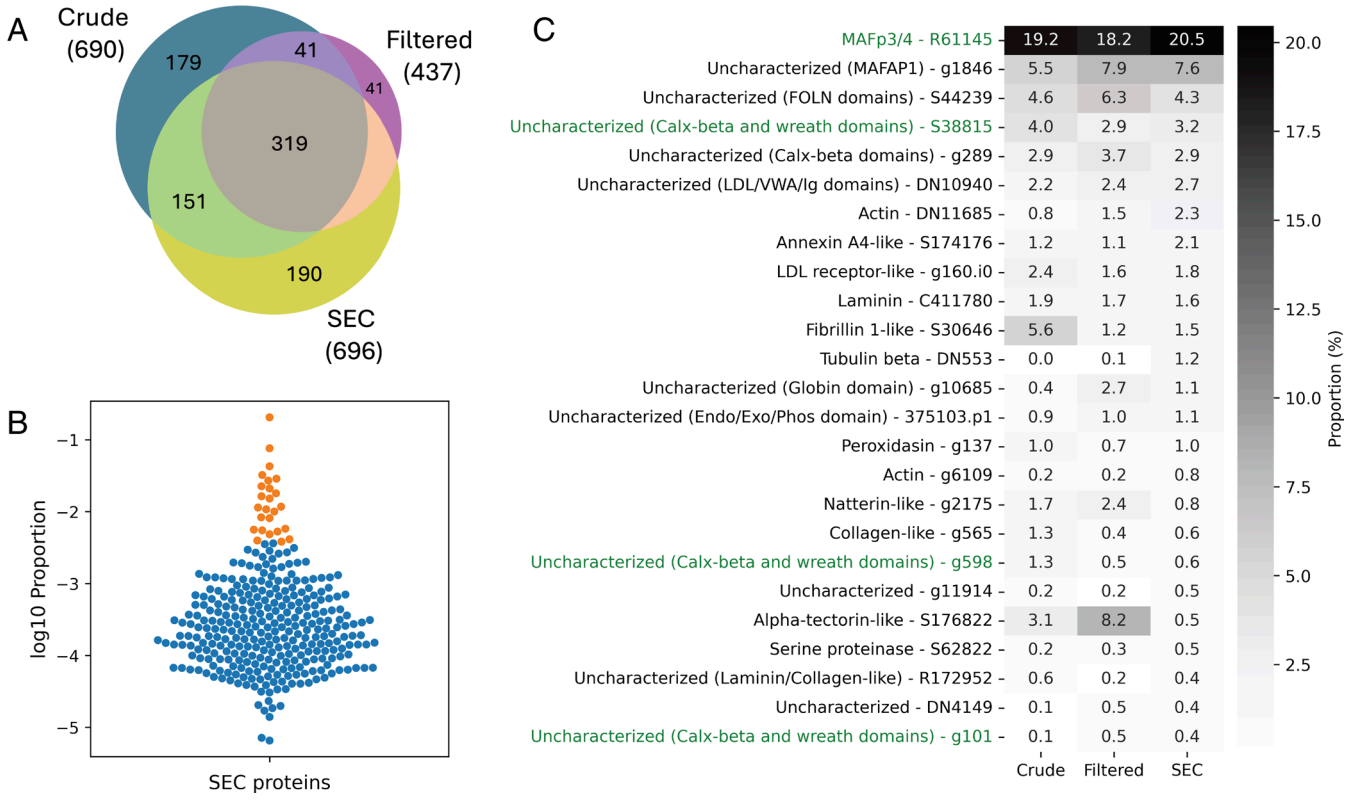

Fig. 2. Known and predicted AF components were abundant in all replicates. (*A*) Venn diagram of proteins detected in each of the Crude, Filtered, and SEC samples. Sample complexity was high, irrespective of the preparatory method, potentially indicating low purity. However, (*B*) a violin plot showing protein proportion in the SEC sample of the 319 proteins common to all samples (plotted on a log10 scale) illustrates that most proteins were actually very low in abundance. The 25 most abundant proteins are highlighted in orange. (*C*) These included the known AF components, MAFp3/p4, and predicted AF components that contained wreath and calx-beta domains (sorted by proportion in the SEC sample).

AF Samples Contain MAFp3/p4 and Related Wreath-Domain Proteins. The high diversity of proteins detected in all samples indicated that our methods did not produce highly purified preparations of the AF, or that high levels of glycosylation of the AF interfered with analysis by Liquid Chromatography–Tandem Mass Spectrometry leading to an apparent overrepresentation of low-level contaminants. Still, we reasoned that the best candidate AF components and/or binding partners were those that were highly represented in all samples (Fig. 2*B*). Indeed, the most abundant proteins detected in all samples were the known AF components MAFp3 and MAFp4, which together comprised 19.2% of the Crude sample on average, 18.2% of the Filtered sample, and 20.5% of the SEC sample (Fig. 2*C*).

The transcriptome assembly was found to encode a ~430 kDa MAFp3/p4 precursor protein (a prediction that does not take into account glycosylation). The MAFp4 region was predicted to have 29 calx-beta domains (Fig. 3*A*), and as previously described, the MAFp3 region contains the wreath domain. In addition to MAFp3/

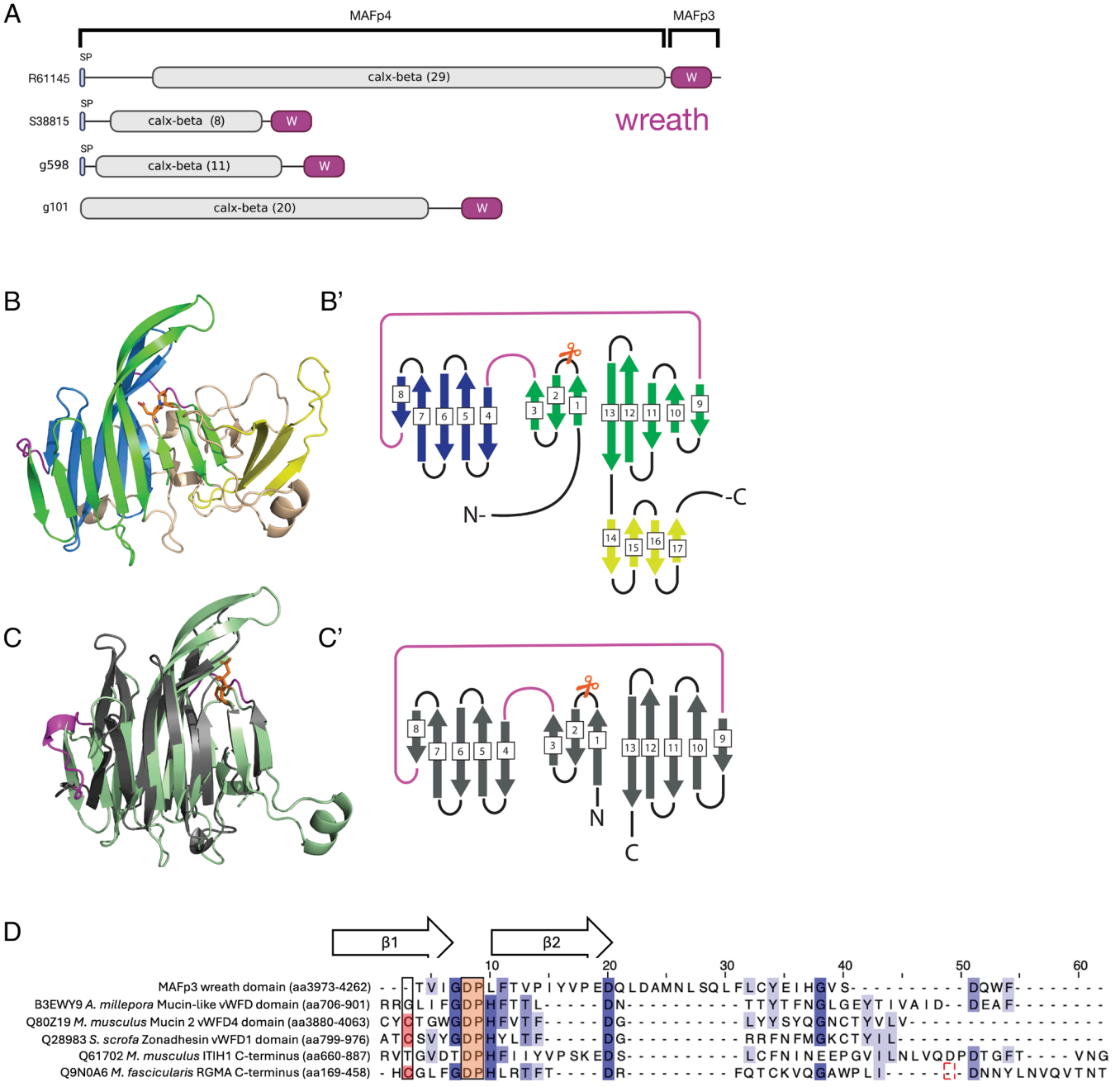

Fig. 3. Wreath domain-containing proteins detected in AF proteomic samples have structural similarity to the vWFD domain. (*A*) Wreath domain-containing proteins in order of their relative proportion in AF proteomic datasets. Wreath domain (W) is highlighted in magenta. Proteins are highlighted in green in Fig. 2*C*. (*B* and *B'*) AlphaFold3 prediction and secondary structure diagram of MAFp3 wreath domain, with beta-strands highlighted to reflect their position in the tertiary structure. Beta sheet switches between strands 3/4 and 8/9 are highlighted in magenta. Putative cleavage site residues Asp-Pro between strands 1 and 2 are highlighted in orange. Disordered N and C termini were manually trimmed for visibility. (*C*) Superposition of the central MAFp3 wreath domain beta-sandwich (green) with vWFD domain of *A. millepora* mucin-like (UniprotID: B3EWY9, aa 705 to 858; gray). rmsd = 4.15 over 200 atoms. Secondary structure diagram of the vWFD domain shown in *C'*. Conserved Asp-Pro autocleavage motif in the vWFD domain is highlighted in orange. Conserved beta sheet switches between strands 3/4 and 8/9 are highlighted in magenta (39). (*D*) Multiple sequence alignment of domains structurally similar to the MAFp3 wreath domain. Only the N terminus of the domains is shown, highlighting the conserved cleavage motif Asp-Pro between beta strands 1 and 2 (orange) (40). The cysteine residue needed for covalent attachment after cleavage is shown in red.

p4, we also detected an additional eight wreath domain-containing proteins in the *C. prolifera* transcriptome assembly (*SI Appendix*, Supplemental File 1). Of these, three (Fig. 3*A*) were also found within the top 25 most abundant proteins in our proteomics datasets (with S38815 being the fourth-most abundant protein in the SEC sample) (Fig. 2*C*) and each was predicted by Interpro (38) to have between 8 and 29 calx-beta domains.

The Wreath Domain is Structurally Similar to vWFD Domain. The wreath domain has previously only been proposed based on multiple sequence alignment of MAFp3 homologs (26). We used AlphaFold3 (41) to predict the MAFp3 wreath domain structure (average pLDDT for all atoms = 73.7, for all alpha-carbons = 76.2) [Supplement (42)]. The model indicated an essentially all-beta structure with isolated short alpha-helices in the periphery. A long, unfolded C-terminal domain is predicted to be disordered (prediction by IUPred3) (43). A central beta-sandwich is formed by two twisted antiparallel beta sheets, containing eight (green, strands 1-3, 9-13) and five (blue, strands 4-8) beta strands of varying lengths (Fig. 3*B*). The beta-sheets are connected by strands 3 and 4 as well as 8 and 9. Additionally, the structure has a curved, antiparallel beta sheet (four beta strands, 14-17) adjacent to the beta-sandwich (yellow).

Hidden-Markov model (HMM)-based searches for wreath domain–containing proteins in other species indicate that this domain is demosponge-specific (26), and we confirmed this result by searching the more inclusive eukaryotic protein dataset, Eukprot 3 (44). As a complement to this, we searched for structurally similar protein domains using the AlphaFold3 prediction for the MAFp3 wreath domain in Foldseek, as this method should find structurally similar proteins irrespective of their sequence conservation (45–47). However, a caveat is that in organisms other than humans, AlphaFold predictions are only available for proteins <2,700 aa in length. Top hits were to wreath domain-containing proteins in *C. prolifera* and other demosponges included in the AlphaFold DB, such as *Suberites domuncula* and *A. queenslandica*. Additionally, a structurally similar protein (UniProt A0A1X1QQN9) was detected from the bacterial species *Cycloclasticus* sp. M. (48). This sequence had 82% identity to the *C. prolifera* MAFp3/4 wreath domain (*SI Appendix*, Fig. S2), albeit only part of the beta-sandwich region is conserved. This bacterium was isolated as a symbiont of a poecilosclerid sponge, suggesting horizontal gene transfer from the sponge to the bacterium.

Notably, Foldseek also detected hits in Bilateria, including proteins such as alpha-tectorin, mucin, zonadhesin, or otogelin. A common feature of these proteins is the presence of a von Willebrand Factor D (vWFD) domain. Exemplary structural superposition of the wreath domain with a vWFD domain of *Acropora millepora* mucin-like protein highlighted a highly conserved beta-sandwich, showing identical position switch of beta strands between the two sheets of the sandwich (39) (Fig. 3*C*), despite an overall low sequence identity of 14.5%, which usually precludes detection by sequence-based similarity searches. Suggestively, the vWFD domain is found in secreted, gel-forming glycoproteins (49) and is pivotal for the platelet aggregation function of the name-giving von Willebrand Factor (50), in line with the function of wreath domain-containing AF proteins. Independently, a Foldseek search using the MAFp3 wreath domain as well as mucin vWFD as queries against the CATH protein structure classification database (51) revealed affiliation to the same large superfamily of proteins (2.60.120.200), sharing a beta-sandwich functionally related to cell adhesion. Additionally, among the top hits we found structural similarities to the C-terminal domain of the bilaterian inter-alpha-trypsin inhibitor heavy chain (ITIH) family members 1-3/5/6 and the C-terminus of repulsive guidance molecule A (RGMA) (39). A multiple sequence alignment of the domains showed a highly conserved pH-sensitive autocatalytic cleavage motif (Asp-Pro) forming a hairpin between beta strands 1 and 2 (Fig. 3*D*). Whereas the cleavage product of vWFD and RGMA stays connected to the protein via a disulfide bridge (40), the ITIH C-terminal domain is missing a required cysteine residue and is released in the acidic Golgi secretory pathway, allowing the terminal Asp to bind glycans to enhance ECM stability (52). Similarly, the wreath domain of MAFp3/p4 is missing the cysteine, offering a potential mechanism for MAFp3 cleavage and MAFp4 glycan binding.

AF-Associated Proteins Share a Conserved C-Terminus Composed of Divergent Immunoglobulin (Ig) and Fibronectin 3-Related Domains. After MAFp3/p4, the next most abundant protein in the SEC dataset (and in all datasets when combined) corresponded to transcript g1846 and is hereafter referred to as MAF-associated protein 1 (MAFAP1). In the Filtered sample, MAFAP1 was slightly less abundant (7.9%) than an uncharacterized NIDO and calx-beta domain-containing protein (8.2%), but this protein only showed a proportion of 0.5% in the SEC sample, indicating its effective separation from the AF by column chromatography. This may indicate that it is either not stably associated with the AF, or that this association was disrupted by short-term exposure to EDTA.

MAFAP1 is predicted as an ~100 kDa protein with a signal peptide, two N-terminal Ig-related domains, and an intrinsically disordered region composed of 69 pentapeptide repeats with the consensus amino acid sequence PETDA (Fig. 4*A*). Notably, it contains the peptide sequence ELIDYETFSDGRVL identified from the 210 kDa AF component purified by (18), but which could not be further resolved at the amino acid level at the time. BLAST searches using MAFAP1 as a query against the NCBI nonredundant protein database and against the more inclusive Eukprot3 database (44) resulted in low-confidence hits to proteins with repeat regions that are only superficially similar to the disordered PETDA repeat region. No candidate MAFAP1 homologs were detected by BLAST search in any other species, including in other sponges.

We next searched by BLAST for additional MAFAP1-related proteins in the *C. prolifera* transcriptome. Again, there were no candidate homologs, but 5 proteins had high identity-score matches to a 226 aa region of the C-terminus (Fig. 4 *A* and *B* and *SI Appendix*, Fig. S3). These hits were otherwise distinct from MAFAP1 and from each other, and included a fibrillin-like protein (S30646, containing ylk_9/nidogen-like, vWFD, and EGF domains), a secreted frizzled homolog (g2880, containing Fz and FN3 domains), a MAFp3/p4-related protein with calx-beta domains but lacking a wreath domain (g1044), a low-density lipoprotein receptor-related protein (g160), and an uncharacterized protein with calx-beta and Fn3 domains (S168704). All but S168704 were present in the proteomics datasets, and of these, the fibrillin-like protein (S30646) was among the top 25 most abundant proteins detected overall (Fig. 2*C*). This suggested that the shared C-terminal region may represent an important binding interface for interactions with the AF. Hereafter, we refer to this conserved region as a candidate "AF-interacting region."

Structural analysis of the candidate AF-interacting region by AlphaFold3 (average pLDDT for all atoms = 79.6, for all alpha-carbons = 83.1) indicated the presence of two discrete elements composed of beta-sheets (region 1 and 2) (Fig. 4*C*) [Supplement (42)]. The top five Foldseek search results against AFDB-Swissprot using MAFAP1 region 1 as a query were neural cell adhesion molecules (NCAM) from various Bilaterians, mapping to Ig-like domains, despite a sequence identity <15%. Structural superposition and sequence alignment supports their close structural similarity and highlights the conservation of two critical cysteine residues that form a stabilizing disulfide bridge in Ig-domains (53) (Fig. 5 *A* and *A*). Likewise, region 2 of the MAFAP1 C-terminal domain is structurally very similar to

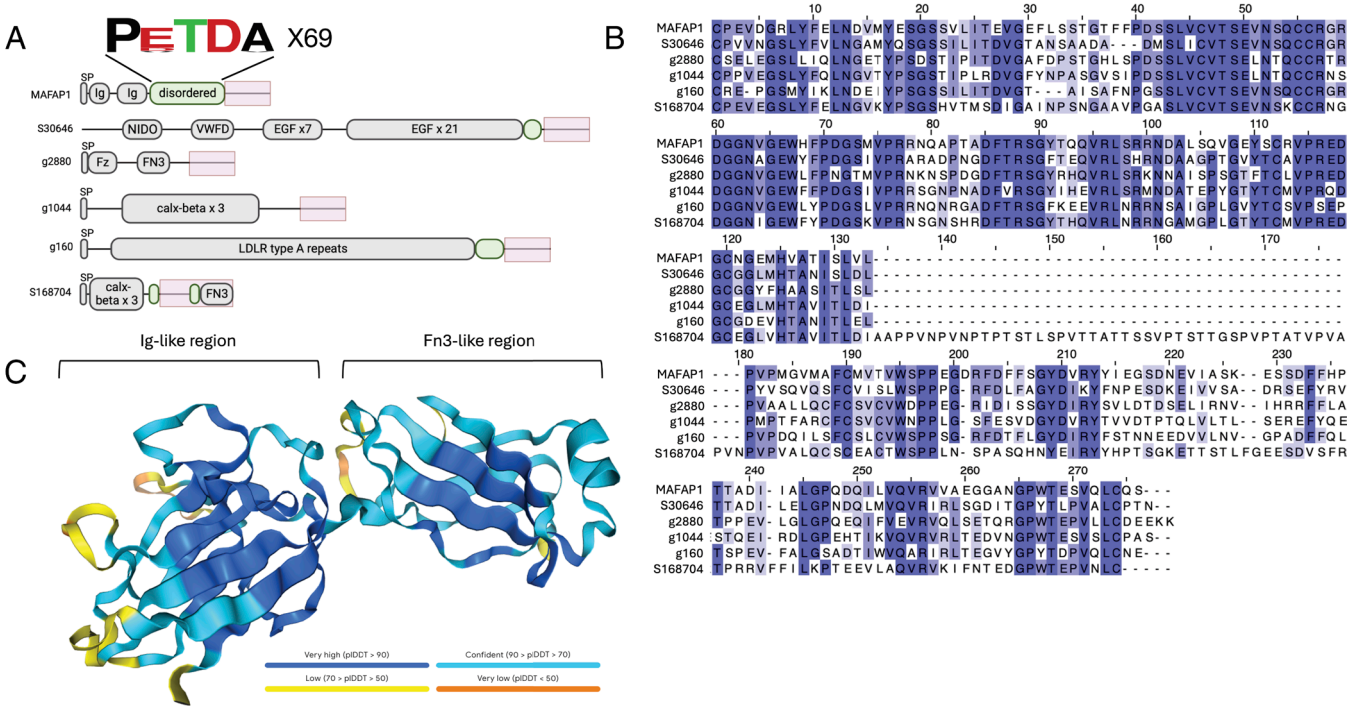

Fig. 4. Top MAFAP1 hits in the predicted proteome are unrelated but share a highly conserved, candidate AF-interaction domain. (*A*) Domain architecture of *C. prolifera* proteins that contain a conserved AF-interacting region (pink box) in common with MAFAP1. Also like MAFAP1, three of these also contain a disordered repeat region (green oval) adjacent to the AF-interacting region. All but S168704 were detected in proteomics results for the AF. (*B*) Multiple sequence alignment of the AF-interacting regions from proteins depicted in panel *A*. (*C*) AlphaFold3 prediction of the C-terminal AF-interacting region of MAFAP1, highlighting a divergent Ig-like domain (region 1) and a divergent Fn3-like domain (region 2).

the adjacent Fn3 domain of the respective NCAM proteins (Fig. 5 *B* and *B*′), and similar proteins are found in intermediately branching invertebrate lineages such as *Trichoplax adherens* and *Caenorhabditis elegans* (*SI Appendix*, Fig. S3). Foldseek searches against the CATH50 database sorted both domains into superfamily 2.60.40.10 (Ig fold).

NCAM belongs to the Ig superfamily of cell adhesion molecules (IgCAMs or IgSF CAMs) (54) which regulate cell–cell/ECM adhesion via extracellular Ig-like domains and are anchored via transmembrane (TM) regions or glycosylphosphatidylinositol (GPI). Although none of the proteins in *C. prolifera* with AF-interacting region are predicted to contain a TM region or GPI anchor [prediction via TMHMM and NetGPI; (55, 56), all are secreted extracellular proteins that presumably can interact with either the glycan or protein component of the AF.

Domain architecture analysis of the SMART database (57) indicated that Ig and Fn3 domains are found in combination in diverse animal protein families (and in some bacteria), but not in non-animal eukaryotes. To further test this we used HMMs to

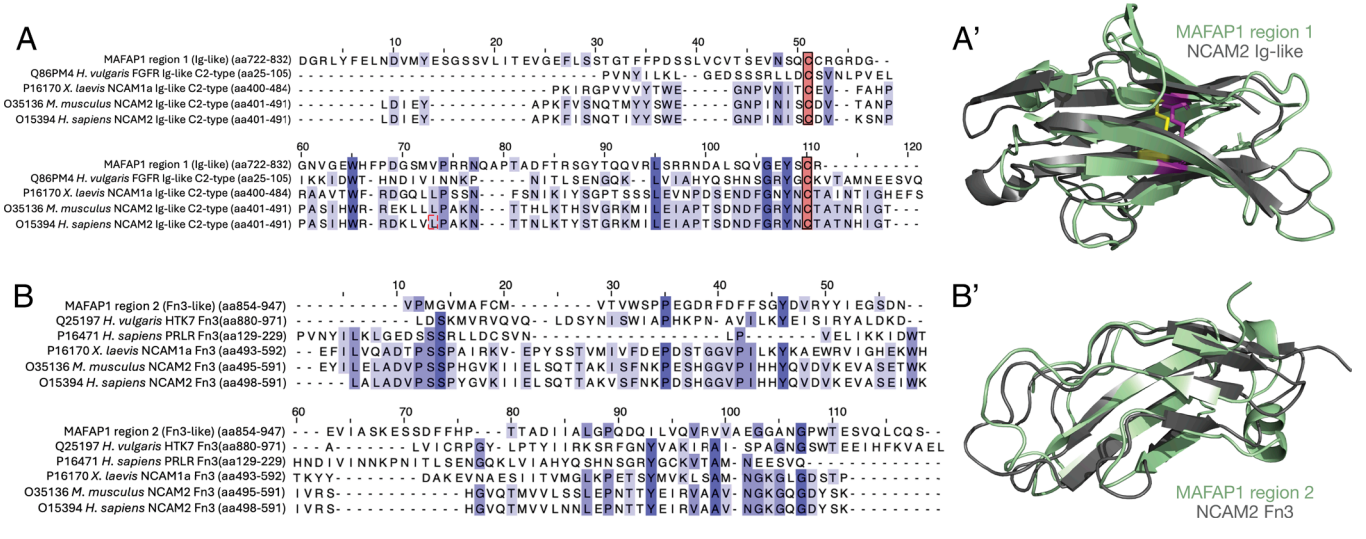

Fig. 5. MAFAP1 C-terminal domains resemble Ig and Fn3 domains. (*A*) Sequence alignment of region 1 of MAFAP1 C-terminal domain to Ig domains (C2-type) from Foldseek hits. Conserved cysteine residues, typical for Ig domains, are highlighted in red. (*B*) Sequence alignment of region 2 of MAFAP1 C-terminal domain to Fn3 domains from Foldseek hits and NCAM sequences shown in *A*. (*A*′) Structural superposition of AlphaFold3 prediction of region 1 of MAFAP1 C-terminal domain (green, aa722 to 853) to *M. musculus* NCAM2 Igdomain (UniprotID: O35136, aa21 to 108, gray) (rmsd: 3.14 over 350 atoms). Conserved cysteine residues that form a stabilizing, intramolecular disulfide bridge are highlighted in magenta (for MAFAP1) and yellow (for NCAM2). (*B*′) Structural superposition of AlphaFold3 prediction of region 2 of MAFAP1 C-terminal domain (green, aa854 to 947) to NCAM2 Fn3-domain (UniprotID: O35136, aa495 to 591, gray) (rmsd: 3.52 over 435 atoms).

search for Ig (PF00047) and Fn3 (PF00041) domains in the EukProt3 database. Using a cutoff of e < 0.00001 and a query-cover filter of 70%, we identified 5,524 animal proteins that contain both domains. None were detected outside of animals.

Proteins with Conserved AF-Interacting Regions in Other Sponge Species. MAFp3/p4 and other wreath domain-containing proteins are widely conserved in demosponges (26). To search for proteins that contain the putative AF-interacting region in other sponge species we developed an HMM from the alignment of the AF-interacting region of *C. prolifera* proteins [Supplement (42)]. We then searched the predicted proteomes of *Ephydatia muelleri* (58), *Tethya wilhelma* (59), *A. queenslandica* (60) (all heteroscleromorph demosponges), and *Cladhorizida* sp. [(61); a poecilosclerid demosponge, like *C. prolifera*]. None were found to have proteins with a conserved AF-interacting region.

When we expanded this search to include partial transcriptome data available from additional poecilosclerid sponges we found evidence for highly conserved AF-interacting regions in proteins from *Tedania anhelans* (62), *Crella elegans* (63), and *Phorbas areolata* (64). In all but two sequences from *T. anhelans*, the transcriptome assemblies were too fragmentary to characterize features of these proteins beyond the presence of the conserved AF-interacting region. But, like *C. prolifera* proteins that contain the AF-interacting region, two *T. anhelans* sequences were found to encode a CRD_Fz domain, and one was found to have calx-beta domains (*SI Appendix*, Supplement File 2). Using BLAST, the top hits of these proteins were secreted frizzled-related proteins and frizzled receptors in other sponges, and MAFp3/p4, respectively.

Other Abundant AF-Associated Proteins Contain Known Domains of Extracellular Proteins. Four additional proteins were relatively abundant in all proteomic datasets (Fig. 2*C*). These included a highly conserved annexin A4-like protein (found in the ECM of vertebrates) and three uncharacterized proteins. One of those (S44239) contains four follistatin-N-terminal domain-like domains which are also found in the ECM proteins Agrin and SPARC. The second uncharacterized protein (g289) again contains calx-beta domains, as detected in MAFp4 and other AF-associated proteins. The third uncharacterized protein (DN10940) has a signal peptide, LDL, Ig, and vWFA, and similarity by BLAST search to contactin, neurofascin, and NCAM—presumably due to the presence of Ig-like domains. The Ig/vWFA region is predicted by Interpro to relate to the Basigin family, which includes two protein subfamilies: neuroplastin and basigin. Both have extracellular Ig-like domains and are glycosylated. Neuroplastin functions in neuronal cell adhesion, whereas basigin has more diverse functions including stimulating the production of matrix metalloproteinases in fibroblasts (65). Among the top 25 proteins, we furthermore identified proteins such as laminin, which add to the pool of secreted glycoproteins that regulate cell–cell and cell–ECM interactions.

AF Proteins Share a Low Predicted Isoelectric Point (pI). The Glu and Asp residues of the pentapeptide repeats of MAFAP1 lead to an exceptionally low pI of 3.56 [prediction with the python implementation of "*Peptides*" (https://pypi.org/project/peptides/)] (66). Proteome-wide comparison of predicted pIs, Glu, and Asp proportions reveal that putative AF components share a significantly lower pI (4.4 ± 1.4 for top 25 AF proteins vs. 7.1 ± 2.1 for whole proteome) (*SI Appendix*, Fig. S5*A*) due to high proportions of Glu (7.5 ± 1.8%) and Asp (7.7 ± 2.1%) residues. Comparison to amino acid compositions from AF preparations of the related sponge *Clathria (Microciona) parthena* (14) supported our results and suggested that Henkart and colleagues likely worked with a pure MAFp3/p4 and MAFAP1 sample due to the high proportion of Glu and Asp residues (*SI Appendix*, Fig. S5*B*). In neutral extracellular environments, AF protein components are therefore predicted to be negatively charged, favoring Ca^{2+} binding.

Materials and Methods

Transcriptome Sequencing and Assembly. We acquired live samples of *C. prolifera* from the Marine Resources Center of the Marine Biological Laboratory at Woods Hole, cleaned tissues of debris and macroscopic contaminants, then ground them to a fine powder in liquid nitrogen. Using Trizol Reagent (Thermo Fisher Scientific), we isolated total RNA and shipped it to Novogene (Chula Vista, CA, USA) for library preparation and Novaseq pe150 bp sequencing. We assembled and annotated the transcriptome using the TransPi pipeline (67) as implemented on the University of Denver High Performance Computing Cluster.

AF Preparation. We purified crude AF using the (approximate) method of Humphreys (37). Briefly, we cut ~100 g of tissue into 1 cm pieces, washed them in cold CMFSW (2 mM $NaHCO_3$, 462 mM NaCl, 7 mM Na_2SO_4, 10.7 mM KCl, pH 7.2) for 10 min, and then rinsed them again briefly in CMFSW. We then squeezed tissue fragments through a fine synthetic mesh into a beaker containing 100 mL cold CMFSW, aliquoted dissociated cells into 50 mL conical tubes, and placed them on a rotator for 2 h at 4 °C. We separated cells from the AF-containing supernatant by centrifugation in a swinging bucket rotor at 1,500 g for 5 min at 4 °C. To remove any remaining insoluble debris, we then centrifuged the supernatant at 10,000 g for 15 min at 4 °C. To precipitate the AF from the supernatant, we added $CaCl_2$ to a final 20 mM concentration and placed the beaker on a stir plate at 4 °C for ~16 h. The AF precipitated as a red, gel-like substance which we collected by centrifugation at 10,000 g. We washed this pellet three times in Tris-buffered MBL-seawater (422 mM NaCl, 9.4 mM KCl, 9 mM $CaCl_2$, 49.4 mM $MgCl_2$, 28 mM $MgSO_4 \bullet 7H_20$, 0.85 mM $NaHCO_3$, 50 mM Tris, pH 7.2) for proteomic analysis as the Crude AF sample.

Following Varner et al. (11) we further purified the crude AF fraction by resuspending the pellet in CMFSW, passing it through a 0.22 μm filter, centrifuging at 20,000 g to remove red contaminating pigments and membranes, then precipitating it in 20 mM $CaCl_2$ overnight at 4 °C. The pellet that formed was white and fluffy in appearance. We again washed this precipitate in Tris-buffered MBL-seawater prior to proteomic analysis. This constituted the "Filtered" sample.

As the final purification step, we resolubilized the Filtered AF fraction in CMFSW + 1 mM EDTA (it remained insoluble in CMFSW alone). We then concentrated the sample as much as possible without causing the AF to precipitate in an Amicon-Ultra 10 column (EMD Millipore), and loaded it onto a Sephacryl S-500 h column (Cytiva). The AF eluted as a single broad peak which we concentrated in an Amicon-Ultra 10 column and analyzed by proteomics as the "SEC" sample.

Atomic Force Microscopy (AFM) Imaging and Instrumental Procedure. For AFM AF images, the demosponges *C. prolifera*, *Haliclona oculata*, *S. domuncula*, *Halichondria panicea*, and *Desmapsamma anchorata* were collected by the Marine Biological Laboratory Marine Resources Department. AFs were prepared as described for the Crude fraction in the *AF Preparation* section of this manuscript. AFM images were acquired under ambient conditions in the tapping mode of operation using standard monolithic Si cantilevers Tap300AI (NanoAndMore) on a commercial instrument Multimode Nanoscope IIIa instrument (Veeco). AF complexes were immobilized via physisorption from solution on muscovite mica surfaces (Plano), which were previously gas phase-silanized with aminopropyltriethoxysilane (Sigma) in a desiccator. The immobilization of the AFs from solution (50 μL of typically 0.5 mg/mL) for 15 min at room temperature was subsequently followed by a washing step with Milli-Q water as well as dried under N_2 flow.

Preparation of Samples for Proteomic analysis. We lyophilized the samples and treated the pellets with freshly prepared hydroxylamine buffer (1 M NH_2OH-HCl, 4.5 M guanidine-HCl, 0.2 M K_2CO_3, pH adjusted to 9.0 with NaOH). We briefly vortexed the samples and then incubated them at 45 °C for 6 h. Due to pressure build-up during incubation, we fastened the tubes shut during incubation. After incubation, we spun the samples for 15 min at 18,000 g, removed the supernatant, and stored it at −20 °C until further proteolytic digestion with trypsin.

We digested the samples following the filter-aided sample preparation protocol, using a 10 kDa molecular weight cutoff filter. Briefly, we mixed 50 μL of

samples in the filter unit with 8 M urea and 100 mM ammonium bicarbonate (AB), pH 8.0, and then centrifuged the mixture at 14,000 g for 15 min. We reduced the proteins with 10 mM dithiothreitol for 30 min at room temperature, centrifuged them, and alkylated them with 55 mM iodoacetamide for 30 min at room temperature in the dark. After centrifugation, we washed the samples three times with urea solution and three times with 50 mM AB, pH 8.0. We carried out protein digestion using sequencing grade modified trypsin (Promega) at a 1/50 protease/protein (w/w) ratio, incubating at 37 °C overnight. Finally, we recovered the peptides from the filter using 50 mM AB.

Mass Spectrometry. We loaded 20 μL of each sample onto individual Evotips (Evosep, Odense Denmark) for desalting, washed them with 20 μL of 0.1% formic acid (FA), and then added 100 μL of storage solvent (0.1% FA) to keep the Evotips wet until analysis. The Evosep One system (Evosep) was used to separate peptides on a Pepsep column (150 μm inner diameter, 15 cm) packed with ReproSil C18 1.9 μm, 120A resin using preset 15 samples per day gradient. We coupled the system to the timsTOF Pro mass spectrometer (Bruker Daltonics in Bremen, Germany) via its nano-electrospray ion source, Captive Spray.

We operated the mass spectrometer in Parallel Accumulation–Serial Fragmentation (PASEF) mode, setting the ramp time to 100 ms and acquiring 10 PASEF MS/MS scans per topN acquisition cycle. We recorded MS and MS/MS spectra from m/z 100 to 1,700 and scanned the ion mobility from 0.7 to 1.50 Vs/cm^2. We isolated precursors for data-dependent acquisition within ±1 Th and fragmented them with an ion mobility-dependent collision energy, linearly increasing it from 20 to 59 eV in positive mode. We repeatedly scheduled low-abundance precursor ions that had an intensity above a threshold of 500 counts but below a target value of 20,000 counts, and dynamically excluded them for 0.4 min.

The mass spectrometry proteomics Data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD058135.

Database Searching and Protein Identification. Mass spectrometry raw timsTOF ddaPASEF .d files were processed using the default LFQ-MBR workflow within Fragpipe v21.1 (68). Files were converted to mzBIN format and then searched using MSFragger v4.0 (69) against the predicted *C. prolifera* proteome including known contaminants and the reversed protein sequences. The default search parameters were as follows: strict tryptic digestion with a maximum of two missed cleavages, peptide length = 7 to 50, peptide tolerance = 20 ppm; MS/MS tolerance = 10 ppm; topN peaks = 150; fixed modifications = carbamidomethyl on cysteine; variable modifications = acetylation on protein N termini and oxidation of methionine. Peptide Spectrum Matches (PSMs) validation was performed by philosopher version 5.1.0 and the false discovery rate was fixed at 1% at the PSMs, peptides, and proteins level. Label-free quantification on the MS1 level using unique as well as razor peptides was conducted by Ionquant (70) using default settings.

The raw output files of FragPipe (protein.tsv files) were processed using the R and python programming languages. Contaminants and reverse proteins were filtered out and only proteins that were quantified with at least two razor peptides (Razor.Peptides ≥ 2) were considered for the analysis. To compare relative protein proportions within a sample (crude, filtered, or SEC), individual protein intensities (based on razor peptides) were divided by the sum of all protein intensities. The relative proportion of the proteins within each sample was calculated by dividing the protein intensity (based on MS1 label-free quantification of unique+razor peptides) by the sum of all peptide intensities from each sample.

Annotation and Analysis of Top Proteomic Hits. In addition to the annotation tools built-in to the TransPi pipeline, we annotated the *C. prolifera* proteome using EggNOG Mapper v2 (71). We then manually examined the most abundant hits in our proteomic analysis by BLAST search against the NCBI nr and Eukprot3 databases, and by using Interpro to predict the presence of signal peptides, conserved domains, and disordered regions.

We used the published Wreath domain HMM (26) to search for Wreath domain-containing proteins in *C. prolifera* using HMMer (71, 72). We used the identified wreath domain of *C. prolifera* MAFp3 as an input for the AlphaFold3 web server to predict its tertiary structure [cif file of best-performing model as (Supplement (42))]. Unfolded N- (aa 1 to 32) and C-termini (aa 318 to 360) were trimmed manually in PyMOL (73). The model was used as a query for the Foldseek web server in 3Di/AA search mode against AlphaFold/Uniprot50 v4, AlphaFold/UniProt v4, and CATH50 4.3.0 databases to identify structurally similar proteins and protein families [raw results as Supplement (42)].

We identified the candidate AF-interacting region by BLAST search of *C. prolifera* MAFAP1 against the *C. prolifera* predicted proteome. We then used an alignment of the conserved C-terminal region of top BLAST hits as an input into HMMer to create an HMM for the AF-interacting region. We used the AF-interacting region HMM to search additional sponge genomes and transcriptomes for homologous proteins. We used the AlphaFold3 web server to predict the tertiary structure of the MAFAP1 AF-interacting region [cif file of best-performing model as Supplement (42)]. We used region 1 (aa 722 to 851) and region 2 (aa 852 to 947) as independent queries for the Foldseek web server in 3Di/AA search mode against AlphaFold/Uniprot50 v4, AlphaFold/UniProt v4, and CATH50 4.3.0 databases to identify structurally similar proteins and protein families [Supplement (42)].

Visualizations and superpositions of protein structure models was done in Pymol (v2.3.5) using the "*super*" command. TM region prediction was performed using the DeepTMHMM web server (v.1.0.24) (55). GPI-anchor prediction was performed using the NetGPI 1.1 web server (56).

Analysis of MAFAP1-Like Proteins in Related Sponges. We examined the prevalence of AF-associated proteins using published sequence data from additional sponge species. First, we searched published datasets for *E. muelleri* (74), *Cladhoriza* sp. (61), and *Phorbas areolatus* (64). We then downloaded raw transcriptome reads from NCBI [Sequence Read Archive (SRA)] for available poecilosclerid sponges: *Asbestopluma hypogea* (ER216190), *C. elegans* (SRR648671), *Isodictya* (SRR6202908-12) *Kirkpatrickia variolosa* (SRR1916957) *Latrunculia apicalis* (SRR1915755), *Mycale grandis* (SRR3334580, SRR3339390, SRR3339394), *Mycale phyllophila* (SRR1711043, SRR2394941, SRR2402290), and *T. anhelans* (SRR3708911). Raw Illumina reads were error-corrected using Rcorrector (75) using a k-mer length of 31. The corrected reads were then quality-trimmed using Trimmomatic via Trinity v.2.4.0 with the default settings (76). The trimmed reads were combined within species and assembled with Trinity v.2.4.0 using the default settings. Raw 454 pyrosequencing reads were assembled using the default settings in MIRA v.4.0 (76, 77). We then translated the assemblies to peptide predictions using TransDecoder (https://github.com/TransDecoder/TransDecoder). The peptide predictions were then clustered to 99% similarity using Cluster Database at High Identity with Tolerance (78) to reduce redundancy from splice isoforms and inparalogs. To reduce potential protist and prokaryote contaminants, we filtered the peptide assemblies with Alien Indexing (https://github.com/josephryan/alien_index) using the provided metazoan and nonmetazoan representative datasets for the BLAST searches, with peptide models from the *T. wilhelma* genome (59) supplementing the metazoan dataset. We then searched the decontaminated and translated poecilosclerid assemblies, as well as genome models from *T. wilhelma* and *A. queenslandica*, for MAFAP1-related proteins using the *C. prolifera* candidates as queries in BLAST searches (e-values 1e−50 and 1e−150). We also searched the additional sponges using an HMM of the C-termini with hmmsearch in HMMER3 (v3.1b2).

Discussion

Studies of sponges have long emphasized that the AF provides the adhesive force and specificity needed for cell aggregation (79). However, the significance of the AF for understanding the early evolution of animal adhesion and allorecognition is less clear, as its known protein components—MAFp3/p4—lack obvious homology with proteins in other animals. The goal of this study was to combine structural analyses with proteomic methods to identify additional components of the AF, together with possible interacting proteins, that may clarify its evolutionary origins.

An Evolutionary Link between Wreath and vWFD Domains. The strongest clue linking the protein core of the AF to proteins known from other animals comes from the predicted structure of the MAFp3 wreath domain. We found that the central beta-sandwich of the wreath domain resembles the vWFD domain and the C-terminal domains of RGMs (39) and ITIH, despite that its sequence does not align well with any of these. However, the vWFD domain is ancient and found in diverse protein families throughout eukaryotes, whereas the wreath domain and

the C-terminal regions of RGMs and ITIH are found only in these specific protein families and are phylogenetically restricted within animals. Thus, we hypothesize that the wreath domain likely evolved through duplication and divergence from a vWFD domain-containing protein in the demosponge stem-lineage.

Named for its discovery in the von Willebrand Factor glycoprotein family (80, 81), the vWFD domain is found in various clotting factors that function at wound sites (82–84), in proteins that form mucus on the surface of epithelia (49), and as structural components of the basement membrane (85). In vitro-expressed vWFD domains from gel-forming mucins are known to multimerize into ring structures (86) of the same size as the wreath-containing ring-like core of the *C. prolifera* AF.

These structural parallels raise questions about the function and regulation of the AF in sponges. For example, it is typically accepted that the AF is a constitutively secreted component of the sponge ECM (79). In contrast, vWFD-containing proteins such as von Willebrand Factor, multimerin 1, thrombospondin-1, and gel-forming mucins are maintained within cellular granules and are subject to regulated secretion (82–84, 87). Considering that the AF has largely been studied in the context of dissociation/reaggregation assays rather than intact tissues, it is plausible that it may also be released as a response to stress. And there is some evidence for this, as sulfated glycosaminoglycans of the AF were found to be largely produced during initial aggregation and reduced after primmorph formation (88). As whole-tissue dissociation is unlikely to occur under natural conditions, we propose that the physiological functions of the AF may instead be related to wound healing and immune defense. Whereas the AF may be released to form nascent adhesions between cells (16) at wound sites, stable adhesion within tissues is likely to involve conserved cell junction proteins such as cadherins, catenins, integrins, and vinculin (89–91).

Ancient Roles for Ig and Fn3 Domains in Allorecognition and Adhesion. After MAFp3/p4, the next most abundant protein in our proteomics dataset, MAFAP1, is the same as the 210 kDa AF-binding protein independently detected by Varner (11) and Fernàndez-Busquets (18). Differences in its predicted molecular weight in our transcriptome assembly (~100 kDa) may reflect its glycosylation state when biochemically purified from endogenous lysates, or perhaps that it forms dimers. Although no clear MAFAP1 homologs were detected in other (even closely related) species, it was found to contain both Ig and Fn3 domains. Individually, these domains have a phylogenetically widespread distribution, but their combination in the same protein is a unique feature of animals, where they often function in ligand binding, particularly in the ECM.

An example with parallels to MAFAP1 is the IgCAM family of proteins (reviewed in ref. 92). Like MAFAP1, IgCAM proteins have Ig domains which often mediate homophilic or heterophilic adhesion between cells (93). Fn3 domains are also commonly present and contribute to the specificity of interactions in contexts such as the immune synapse (94, 95), but may also contribute to cis interactions that affect IgCAM clustering on the membrane (96, 97). Although their adhesion functions are usually not essential to epithelial integrity, they are often required for cell sorting that drives morphogenesis (98, 99). For example, *Drosophila* Echinoid serves recognition functions required for segregation of differentiated cell types (100, 101) a process with themes common to allorecognition. Although the exact interaction between the AF and MAFAP1 is still unresolved, it has been proposed that like IgCAMs the AF functions in inward self-recognition rather than outward recognition of foreign molecules (102).

Commonalities between allorecognition and IgCAM-mediated cell sorting notwithstanding, the very presence of Ig domains in MAFAP1 (and other candidate AF-interacting proteins) suggests a link between AF function and allorecognition systems in other animals. The vertebrate Major Histocompatibility Complex class I and class II molecules each contain Ig domains (103). Histocompatibility in the colonial tunicate *Botryllus schlosseri* is controlled by the FuHC gene which encodes a TM receptor with extracellular Ig domains (104). Allorecognition in the colonial cnidarian *Hydractinia* is controlled by two loci, *Alr1* and *Alr2*, which each encode TM proteins with extracellular Ig domains (105–107). Even outside of animals, self-recognition in the social amoeba *Dictyostelium discoideum* is controlled by two heavily glycosylated TM receptors, TgrB1 and TgrC1, which contain extracellular IPT/TIG domains—which contain Ig-like folds (108).

Conclusions

This study reveals a conserved toolkit of protein domains that are present in proteins regulating cell adhesion and recognition in cnidarians, bilaterians, and sponges. These include vWFD, Ig-like, Fn3, and calx-beta domains, with the latter three adopting similar Ig-like beta-sandwich folds. Although mutational decay of protein sequence similarity often prohibits the assumption of homology, the advance of tools such as AlphaFold allowed us to trace back the existence of those domains in the sponge AF and AF-associated proteins. On top of conserved protein domains, physico-chemical parameters such as low predicted pIs, which in turn lead to negative charges at neutral (extracellular) pH, appear to be conserved in secreted, Ca^{2+}-binding proteins such as osteopontin (pI = 3.5) (109, 110) or the cell adhesion protein fibronectin (predicted pI = 5.5). In the future, it will be interesting to apply the same unbiased proteomics approach that we used in this study to examine the AF composition and interactions in progressively distantly related demosponge species, which may provide additional clues to how the AF evolved and relates to adhesion and allorecognition mechanisms in other sponges and in nonsponge animals.

Data, Materials, and Software Availability. RNAseq and proteomics data have been deposited in NCBI SRA and PRIDE [SRX18275041 (111) and PXD058135 (112)].

ACKNOWLEDGMENTS. We thank Bernadette Doyle for help in assembling the *C. prolifera* transcriptome and in purification of the AF, the Savitski team at The European Molecular Biology Laboratory for their help with proteomic data analysis, Detlev Arendt and Mikhail Savitski for the support to F.R., and Ana Riesgo for sharing transcriptome assemblies of poecilosclerid sponges. This work was supported by a NSF grant (IOS:2015608) to S.A.N.

Author affiliations: [a]Developmental Biology Unit, European Molecular Biology Laboratory, Heidelberg 69117, Germany; [b]Department of Biochemistry and Molecular Genetics, Anschutz Medical Campus, University of Colorado, Aurora, CO 80045; [c]Department of Molecular, Cellular and Biomedical Science, University of New Hampshire, Durham, NH 03824; [d]Department of Biological Sciences, University of Denver, Denver, CO 80208; [e]Nanomalaria Group, Faculty of Physics, Experimental Biophysics, Bielefeld University, Bielefeld 33501, Germany; [f]Nanomalaria Group, Barcelona Institute for Global Health, Hospital Clínic Universitat de Barcelona, Barcelona 08036, Spain; and [g]Nanomalaria Group, Institute for Bioengineering of Catalonia, The Barcelona Institute of Science and Technology, Barcelona 08028, Spain

1. M. Abedin, N. King, Diverse evolutionary paths to cell adhesion. *Trends Cell Biol.* **20**, 734–742 (2010).
2. N. King, C. T. Hittinger, S. B. Carroll, Evolution of key cell signaling and adhesion protein families predates animal origins. *Science* **301**, 361–363 (2003).
3. L. F. Grice, B. M. Degnan, "How to build an allorecognition system: A guide for prospective multicellular organisms" in *Evolutionary Transitions to Multicellular Life: Principles and Mechanisms*, I. Ruiz-Trillo, A. M. Nedelcu, Eds. (Springer Netherlands, 2015), pp. 395–424.

4. H. V. Wilson, On some phenomena of coalescence and regeneration in sponges. *J. Exp. Zool. Part A* **5**, 245–258 (1907).
5. G. Weinbaum, M. M. Burger, Two component system for surface guided reassociation of animal cells. *Nature* **244**, 510–512 (1973).
6. D. R. McClay, Cell aggregation: Properties of cell surface factors from five species of sponge. *J. Exp. Zool.* **188**, 89–101 (1974).
7. J. E. Jumblatt, V. Schlup, M. M. Burger, Cell-cell recognition: Specific binding of Microciona sponge aggregation factor to homotypic cells and the role of calcium ions. *Biochemistry* **19**, 1038–1042 (1980).
8. G. N. Misevic, M. M. Burger, Reconstitution of high cell binding affinity of a marine sponge aggregation factor by cross-linking of small low affinity fragments into a large polyvalent polymer. *J. Biol. Chem.* **261**, 2853–2859 (1986).
9. C. Wagner-Hülsmann *et al.*, A galectin links the aggregation factor to cells in the sponge (Geodia cydonium) system. *Glycobiology* **6**, 785–793 (1996).
10. B. Blumbach *et al.*, The putative sponge aggregation receptor. Isolation and characterization of a molecule composed of scavenger receptor cysteine-rich domains and short consensus repeats. *J. Cell Sci.* **111**, 2635–2644 (1998).
11. J. A. Varner, M. M. Burger, J. F. Kaufman, Two cell surface proteins bind the sponge Microciona prolifera aggregation factor. *J. Biol. Chem.* **263**, 8498–8508 (1988).
12. O. Popescu, G. N. Misevic, Self-recognition by proteoglycans. *Nature* **386**, 231–232 (1997).
13. J. Jarchow, M. M. Burger, Species-specific association of the cell-aggregation molecule mediates recognition in marine sponges. *Cell Adhes. Commun.* **6**, 405–414 (1998).
14. P. Henkart, S. Humphreys, T. Humphreys, Characterization of sponge aggregation factor. Unique proteoglycan complex. *Biochemistry* **12**, 3045–3050 (1973).
15. W. E. Müller, R. K. Zahn, Purification and characterization of a species-specific aggregation factor in sponges. *Exp. Cell Res.* **80**, 95–104 (1973).
16. S. Humphreys, T. Humphreys, J. Sano, Organization and polysaccharides of sponge aggregation factor. *J. Supramol. Struct.* **7**, 339–351 (1977).
17. X. Fernàndez-Busquets, M. M. Burger, The main protein of the aggregation factor responsible for species-specific cell adhesion in the marine sponge Microciona prolifera is highly polymorphic. *J. Biol. Chem.* **272**, 27839–27847 (1997).
18. X. Fernàndez-Busquets, R. A. Kammerer, M. M. Burger, A 35-kDa protein is the basic unit of the core from the 2 × 10(4)-kDa aggregation factor responsible for species-specific cell adhesion in the marine sponge Microciona prolifera. *J. Biol. Chem.* **271**, 23558–23565 (1996).
19. J. Jarchow *et al.*, Supramolecular structure of a new family of circular proteoglycans mediating cell adhesion in sponges. *J. Struct. Biol.* **132**, 95–105 (2000).
20. G. N. Misevic *et al.*, Molecular recognition between glyconectins as an adhesion self-assembly pathway to multicellularity. *J. Biol. Chem.* **279**, 15579–15590 (2004).
21. G. N. Misevic, M. M. Burger, Carbohydrate-carbohydrate interactions of a novel acidic glycan can mediate sponge cell adhesion. *J. Biol. Chem.* **268**, 4922–4929 (1993).
22. G. N. Misevic, M. M. Burger, The species-specific cell-binding site of the aggregation factor from the sponge Microciona prolifera is a highly repetitive novel glycan containing glucuronic acid, fucose, and mannose. *J. Biol. Chem.* **265**, 20577–20584 (1990).
23. J. A. Varner, Cell adhesion in sponges: Potentiation by a cell surface 68 kDa proteoglycan-binding protein. *J. Cell Sci.* **108**, 3119–3126 (1995).
24. G. N. Misevic, J. Finne, M. M. Burger, Involvement of carbohydrates as multiple low affinity interaction sites in the self-association of the aggregation factor from the marine sponge Microciona prolifera. *J. Biol. Chem.* **262**, 5870–5877 (1987).
25. D. J. Rice, T. Humphreys, Two Ca2+ functions are demonstrated by the substitution of specific divalent and lanthanide cations for the Ca2+ required by the aggregation factor complex from the marine spongem, Microciona prolifera. *J. Biol. Chem.* **258**, 6394–6399 (1983).
26. L. F. Grice *et al.*, Origin and evolution of the sponge aggregation factor gene family. *Mol. Biol. Evol.* **34**, 1083–1099 (2017).
27. E. M. Schwarz, S. Benzer, Calx, a Na-Ca exchanger gene of Drosophila melanogaster. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 10249–10254 (1997).
28. N. Alonso-García, A. Inglés-Prieto, A. Sonnenberg, J. M. de Pereda, Structure of the Calx-beta domain of the integrin beta4 subunit: Insights into function and cation-independent stability. *Acta Crystallogr. D Biol. Crystallogr.* **65**, 858–871 (2009).
29. I. Smyth *et al.*, The extracellular matrix gene Frem1 is essential for the normal adhesion of the embryonic epidermis. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 13560–13565 (2004).
30. D. Kiyozumi, N. Sugimoto, I. Nakano, K. Sekiguchi, Frem3, a member of the 12 CSPG repeats-containing extracellular matrix protein family, is a basement membrane protein with tissue distribution patterns distinct from those of Fras1, Frem2, and QBRICK/Frem1. *Matrix Biol.* **26**, 456–462 (2007).
31. H. Nikkila *et al.*, Sequence similarities between a novel putative G protein-coupled receptor and Na+/Ca2+ exchangers define a cation binding domain. *Mol. Endocrinol.* **14**, 1351–1364 (2000).
32. P. G. Hodor, M. R. Illies, S. Broadley, C. A. Ettensohn, Cell-substrate interactions during sea urchin gastrulation: Migrating primary mesenchyme cells interact with and align extracellular matrix fibers that contain ECM3, a molecule with NG2-like and multiple calcium-binding domains. *Dev. Biol.* **222**, 181–194 (2000).
33. A. Maekawa, M. Hayase, T. Yubisui, Y. Minami, A cDNA cloned from Physarum polycephalum encodes new type of family 3 beta-glucosidase that is a fusion protein containing a calx-beta motif. *Int. J. Biochem. Cell Biol.* **38**, 2164–2172 (2006).
34. J. A. Varner, Isolation of a sponge-derived extracellular matrix adhesion protein. *J. Biol. Chem.* **271**, 16119–16125 (1996).
35. X. Fernàndez-Busquets, D. Gerosa, D. Hess, M. M. Burger, Accumulation in marine sponge grafts of the mRNA encoding the main proteins of the cell adhesion system. *J. Biol. Chem.* **273**, 29545–29553 (1998).
36. S. A. Nichols, Clathria prolifera transcriptome assembly. Figshare (2022). 10.6084/m9.figshare.21559569.v1. Accessed 15 November 2022.
37. T. Humphreys, Chemical dissolution and in vitro reconstruction of sponge cell adhesions. I. Isolation and functional demonstration of the components involved. *Dev. Biol.* **8**, 27–47 (1963).
38. P. Jones *et al.*, InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
39. X. Dong *et al.*, The von Willebrand factor D'D3 assembly and structural principles for factor VIII binding and concatemer biogenesis. *Blood* **133**, 1523–1533 (2019).
40. N. Yeshaya *et al.*, VWD domain stabilization by autocatalytic Asp-Pro cleavage. *Protein Sci.* **33**, e4929 (2024).
41. J. Abramson *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
42. F. Ruperti, S. A. Nichols, Supplementary Files for "Proteomic analysis of the sponge Aggregation Factor implicates an ancient toolkit for allorecognition and adhesion in animals." Zenodo. https://zenodo.org/records/13836934 (2024). Deposited 25 September 2024.
43. G. Erdős, M. Pajkos, Z. Dosztányi, IUPred3: Prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res.* **49**, W297–W303 (2021).
44. D. J. Richter *et al.*, EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. *Peer Commun. J.* **2**, e56 (2022).
45. M. Mirdita *et al.*, ColabFold: Making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
46. M. van Kempen *et al.*, Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2023), 10.1038/s41587-023-01773-0.
47. F. Ruperti *et al.*, Cross-phyla protein annotation by structural prediction and alignment. *Genome Biol.* **24**, 113 (2023).
48. M. Rubin-Blum *et al.*, Short-chain alkanes fuel mussel and sponge Cycloclasticus symbionts from deep-sea gas and oil seeps. *Nat. Microbiol.* **2**, 17093 (2017).
49. T. Lang, G. C. Hansson, T. Samuelsson, Gel-forming mucins appeared early in metazoan evolution. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 16209–16214 (2007).
50. R. N. Finn, Vertebrate yolk complexes and the functional implications of phosvitins and other subdomains in vitellogenins. *Biol. Reprod.* **76**, 926–935 (2007).
51. I. Sillitoe, N. Dawson, J. Thornton, C. Orengo, The history of the CATH structural classification of protein domains. *Biochimie* **119**, 209–217 (2015).
52. L. Zhuo, K. Kimata, Structure and function of inter-alpha-trypsin inhibitor heavy chains. *Connect. Tissue Res.* **49**, 311–320 (2008).
53. T. Kreis, R. Vale, *Guidebook to the Extracellular Matrix, Anchor, and Adhesion Proteins* (Oxford University Press, 1999).
54. Y. Yoshihara, "Immunoglobulin superfamily cell adhesion molecules" in *Encyclopedia of Neuroscience*, M. D. Binder, N. Hirokawa, U. Windhorst, Eds. (Springer, Berlin Heidelberg, 2009), pp. 1923–1926.
55. J. Hallgren *et al.*, DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.04.08.487609 (Accessed 30 March 2023).
56. M. H. Gíslason, H. Nielsen, J. J. Almagro Armenteros, A. R. Johansen, Prediction of GPI-anchored proteins with pointer neural networks. *Curr. Res. Biotechnol.* **3**, 6–13 (2021).
57. I. Letunic, T. Doerks, P. Bork, SMART 7: Recent updates to the protein domain annotation resource. *Nucleic Acids Res.* **40**, D302–D305 (2012).
58. N. J. Kenny *et al.*, Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge Ephydatia muelleri. *Nat. Commun.* **11**, 3676 (2020).
59. W. R. Francis, M. Eitel, S. Vargas, M. Adamski, The genome of the contractile demosponge Tethya wilhelma and the evolution of metazoan neural signalling pathways. bioRxiv [Preprint] (2017). https://www.biorxiv.org/content/10.1101/120998v3 (Accessed 30 March 2023).
60. M. Srivastava *et al.*, The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* **466**, 720–726 (2010).
61. D. T. Schultz *et al.*, Ancient gene linkages support ctenophores as sister to other animals. *Nature* **618**, 110–117 (2023).
62. C. Díez-Vives, L. Moitinho-Silva, S. Nielsen, D. Reynolds, T. Thomas, Expression of eukaryotic-like protein in the microbiome of sponges. *Mol. Ecol.* **26**, 1432–1451 (2017).
63. C. E. Laumer *et al.*, Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc. Biol. Sci.* **286**, 20190831 (2019).
64. A. Riesgo *et al.*, Recycling resources: Silica of diatom frustules as a source for spicule building in Antarctic siliceous demosponges. *Zool. J. Linn. Soc.* **192**, 259–276 (2020).
65. T. Kanekura, X. Chen, T. Kanzaki, Basigin (CD147) is expressed on melanoma cells and induces tumor cell invasion by stimulating production of matrix metalloproteinases by fibroblasts. *Int. J. Cancer* **99**, 520–528 (2002).
66. D. Osório, P. Rondón-Villarreal, R. Torres, Peptides: A package for data mining of antimicrobial peptides. *R J.* **7**, 4 (2015).
67. R. E. Rivera-Vicéns, C. A. Garcia-Escudero, N. Conci, M. Eitel, G. Wörheide, TransPi-a comprehensive TRanscriptome ANalysiS PIpeline for de novo transcriptome assembly. *Mol. Ecol. Resour.* **22**, 2070–2086 (2022).
68. F. Yu *et al.*, Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nat. Commun.* **14**, 4154 (2023).
69. A. T. Kong, F. V. Leprevost, D. M. Avtonomov, D. Mellacheruvu, A. I. Nesvizhskii, MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).
70. F. Yu, S. E. Haynes, A. I. Nesvizhskii, Ionquant enables accurate and sensitive label-free quantification with FDR-controlled match-between-runs. *Mol. Cell. Proteomics* **20**, 100077 (2021).
71. C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, J. Huerta-Cepas, eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
72. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
73. Z. Zhang, J. Cheng, X. Zhou, H. Wu, B. Zhang, Integrated network pharmacology and molecular docking to investigate the potential mechanism of Tufuling on Alzheimer's disease. *Heliyon* **10**, e36471 (2024).
74. J. F. Peña *et al.*, Conserved expression of vertebrate microvillar gene homologs in choanocytes of freshwater sponges. *Evodevo* **7**, 13 (2016).
75. L. Song, L. Florea, Rcorrector: Efficient and accurate error correction for Illumina RNA-seq reads. *Gigascience* **4**, 48 (2015).
76. M. G. Grabherr *et al.*, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
77. B. Chevreux *et al.*, Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* **14**, 1147–1159 (2004).

78. W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
79. E. Vilanova *et al.*, Carbohydrate–carbohydrate interactions mediated by sulfate esters and calcium provide the cell adhesion required for the emergence of early metazoans. *J. Biol. Chem.* **291**, 9425–9437 (2016).
80. Y.-F. Zhou *et al.*, Sequence and structure relationships within von Willebrand factor. *Blood* **120**, 449–458 (2012).
81. K. Titani *et al.*, Amino acid sequence of human von Willebrand factor. *Biochemistry* **25**, 3171–3184 (1986).
82. P. J. Lenting, O. D. Christophe, C. V. Denis, von Willebrand factor biosynthesis, secretion, and clearance: Connecting the far ends. *Blood* **125**, 2019–2028 (2015).
83. A. Bonnefoy, M. F. Hoylaerts, Thrombospondin-1 in von Willebrand factor function. *Curr. Drug Targets* **9**, 822–832 (2008).
84. C. P. Hayward *et al.*, Multimerin is found in the alpha-granules of resting platelets and is synthesized by a megakaryocytic cell line. *J. Clin. Invest.* **91**, 2630–2639 (1993).
85. M. S. P. Ho, K. Böse, S. Mokkapati, R. Nischt, N. Smyth, Nidogens-Extracellular matrix linker molecules. *Microsc. Res. Tech.* **71**, 387–395 (2008).
86. D. Ambort *et al.*, Calcium and pH-dependent packing and release of the gel-forming MUC2 mucin. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 5645–5650 (2012).
87. B. Demouveaux, V. Gouyer, F. Gottrand, T. Narita, J.-L. Desseyn, Gel-forming mucin interactome drives mucus viscoelasticity. *Adv. Colloid Interface Sci.* **252**, 69–82 (2018).
88. E. Vilanova, C. C. Coutinho, P. A. S. Mourão, Sulfated polysaccharides from marine sponges (Porifera): An ancestor cell-cell adhesion event based on the carbohydrate-carbohydrate interaction. *Glycobiology* **19**, 860–867 (2009).
89. K. J. Schippers, S. A. Nichols, Evidence of signaling and adhesion roles for β-catenin in the sponge Ephydatia muelleri. *Mol. Biol. Evol.* **35**, 1407–1421 (2018).
90. P. W. Miller *et al.*, Analysis of a vinculin homolog in a sponge (phylum Porifera) reveals that vertebrate-like cell adhesions emerged early in animal evolution. *J. Biol. Chem.* **293**, 11674–11686 (2018).
91. J. M. Mitchell, S. A. Nichols, Diverse cell junctions with unique molecular composition in tissues of a sponge (Porifera). *Evodevo* **10**, 26 (2019).
92. T. M. Finegan, D. T. Bergstralh, Neuronal immunoglobulin superfamily cell adhesion molecules in epithelial morphogenesis: Insights from Drosophila. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190553 (2020).
93. H. Volkmer, J. Schreiber, F. G. Rathjen, Regulation of adhesion by flexible ectodomains of IgCAMs. *Neurochem. Res.* **38**, 1092–1099 (2013).
94. T. A. Springer, Adhesion receptors of the immune system. *Nature* **346**, 425–434 (1990).
95. A. N. R. Cartwright, J. Griggs, D. M. Davis, The immune synapse clears and excludes molecules above a size threshold. *Nat. Commun.* **5**, 5479 (2014).
96. B. Kunz *et al.*, Axonin-1/TAG-1 mediates cell-cell adhesion by a cis-assisted trans-interaction. *J. Biol. Chem.* **277**, 4551–4557 (2002).
97. H. Tang *et al.*, Architecture of cell-cell adhesion mediated by sidekicks. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9246–9251 (2018).
98. C. Laplante, L. A. Nilson, Differential expression of the adhesion molecule Echinoid drives epithelial morphogenesis in Drosophila. *Development* **133**, 3255–3264 (2006).
99. H.-P. Lin *et al.*, Cell adhesion molecule Echinoid associates with unconventional myosin VI/Jaguar motor to regulate cell morphology during dorsal closure in Drosophila. *Dev. Biol.* **311**, 423–433 (2007).
100. R. Islam, S.-Y. Wei, W.-H. Chiu, M. Hortsch, J.-C. Hsu, Neuroglian activates Echinoid to antagonize the Drosophila EGF receptor signaling pathway. *Development* **130**, 2051–2059 (2003).
101. S. A. Spencer, R. L. Cagan, Echinoid is essential for regulation of Egfr signaling and R8 formation during Drosophila eye development. *Development* **130**, 3725–3733 (2003).
102. T. Humphreys, E. L. Reinherz, Invertebrate immune recognition, natural immunity and the evolution of positive selection. *Immunol. Today* **15**, 316–320 (1994).
103. M. Wieczorek *et al.*, Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation. *Front. Immunol.* **8**, 292 (2017).
104. A. W. De Tomaso *et al.*, Isolation and characterization of a protochordate histocompatibility locus. *Nature* **438**, 454–459 (2005).
105. H. Rodriguez-Valbuena, A. Gonzalez-Muñoz, L. F. Cadavid, Multiple Alr genes exhibit allorecognition-associated variation in the colonial cnidarian Hydractinia. *Immunogenetics* **74**, 559–581 (2022).
106. M. L. Nicotra *et al.*, A hypervariable invertebrate allodeterminant. *Curr. Biol.* **19**, 583–589 (2009).
107. S. F. P. Rosa *et al.*, Hydractinia allodeterminant alr1 resides in an immunoglobulin superfamily-like gene complex. *Curr. Biol.* **20**, 1122–1127 (2010).
108. S. Hirose, G. Chen, A. Kuspa, G. Shaulsky, The polymorphic proteins TgrB1 and TgrC1 function as a ligand-receptor pair in allorecognition. *J. Cell Sci.* **130**, 4002–4012 (2017).
109. E. Kläning, B. Christensen, E. S. Sørensen, T. Vorup-Jensen, J. K. Jensen, Osteopontin binds multiple calcium ions with high affinity and independently of phosphorylation status. *Bone* **66**, 90–95 (2014).
110. N. Azuma, A. Maeta, K. Fukuchi, C. Kanno, A rapid method for purifying osteopontin from bovine milk and interaction between osteopontin and other milk proteins. *Int. Dairy J.* **16**, 370–378 (2006).
111. University of Denver, Clathria prolifera transcriptome. NCBI Sequence Read Archive. https://www.ncbi.nlm.nih.gov/sra/?term=srx18275041. Deposited 16 November 2022.
112. F. Rupterti, S. Nichols, Proteomic analysis of the sponge (Porifera) Aggregation Factor implicates an ancient protein domain toolkit for allorecognition and adhesion in animals. PRoteomics IDEntifications Database. https://www.ebi.ac.uk/pride/. Deposited 21 November 2024.