

# **Developmental transcriptomics of the brittle star *Amphiura filiformis* reveals gene regulatory network rewiring in echinoderm larval skeleton evolution**

David Dylus<sup>1,2,#</sup>, Liisa M. Blowes<sup>3,##</sup>, Anna Czarkwiani<sup>1</sup>, Maurice R. Elphick<sup>3</sup> and Paola Oliveri<sup>1,\*</sup>

1. Research Department of Genetics, Evolution and Environment, UCL, Gower Street, London, WC1E 6BT, UK

2. CoMPLEX/SysBio, UCL, Gower Street, London, WC1E 6BT, UK

3. School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, UK

# Present address: Department of Ecology and Evolution, UNIL, Biophore, Lausanne, 1005, Switzerland

## Present address: Centre for Cell Biology & Cutaneous Research, Blizard Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, E1 2AT, UK

\* Corresponding Author:

Paola Oliveri, Research Department of Genetics, Evolution and Environment  
University College London, Darwin Building, Gower Street, London, WC1E 6BT, UK.  
Office: (+44) 020-767 93719  
Fax: (+44) 020 7679 7193  
Email: [p.oliveri@ucl.ac.uk](mailto:p.oliveri@ucl.ac.uk)

29 **Data Depositions**

30 *All sequence reads are available at Genbank SRR4436669 - SRR4436674. Any sequence*  
31 *alignments used are available by the corresponding author upon request.*

## ABSTRACT

Amongst the echinoderms the class Ophiuroidea is of particular interest for its phylogenetic position, ecological importance, developmental and regenerative biology. However, compared to other echinoderms, notably echinoids (sea urchins), relatively little is known about developmental changes in gene expression in ophiuroids. To address this issue we have generated and assembled a large RNAseq data set of four key stages of development in the brittle star *Amphiura filiformis* and a *de novo* reference transcriptome of comparable quality to that of a model echinoderm - the sea urchin *Strongylocentrotus purpuratus*. Furthermore, we provide access to the new data via a web interface: [http://www.echinonet.eu/shiny/Amphiura\\_filiformis/](http://www.echinonet.eu/shiny/Amphiura_filiformis/). With a focus on skeleton development, we have identified highly conserved genes associated with the development of a biomineralized skeleton. We also identify important class-specific characters, including the independent duplication of the *msp130* class of genes in different echinoderm classes and the unique occurrence of spicule matrix (sm) genes in echinoids. Using a new quantification pipeline for our *de novo* transcriptome, validated with other methodologies, we find major differences between brittle stars and sea urchins in the temporal expression of many transcription factor genes. This divergence in developmental regulatory states is more evident in early stages of development when cell specification begins, than when cells initiate differentiation. Our findings indicate that there has been a high degree of gene regulatory network rewiring in the evolution of echinoderm larval development.

## Keywords

RNA-seq, endoskeleton, biomineralization, transcription factors, regulatory states.

## 57 INTRODUCTION

58 A fundamental question in evolutionary biology is how complex characters originate.

59 Complex structures, such as the endoskeleton, sensory organs or central nervous system, are  
60 built during animal development and encoded by a precise program(s) that requires coordinated  
61 expression of many genes regulated by large networks. A comprehensive theory formulated a  
62 decade ago by Davidson and Erwin (2006) explains both macro and micro evolutionary transitions  
63 as changes in gene regulatory networks (GRN) or rewiring (Davidson & Erwin 2006). Therefore,  
64 comparative studies of gene expression during development have been used fruitfully in locating  
65 GRN rewiring that occurred during evolution (Israel et al. 2016).

66 The calcite endoskeleton of echinoderms provides an ideal system to study the evolution of  
67 complex characters at the level of GRNs. The phylum Echinodermata comprises five extant classes  
68 with well-supported phylogenetic relationships, which group echinoids (sea urchin) and  
69 holothuroids (sea cucumber) (Echinozoa) as a sistergroup to asteroids (sea star) and ophiuroids  
70 (brittlestar) (Asterozoa), with crinoids (sea lilies) an outgroup (Cannon et al. 2014; Telford et al.  
71 2014; O'Hara et al. 2014). While all echinoderms have calcitic skeleton as adults, only ophiuroids  
72 and echinoids develop an elaborate skeleton as larvae. In contrast, the larvae of the other three  
73 classes either develop only miniature ossicles, called spicules (holothuroids), or do not form a  
74 skeleton at all (McCauley et al. 2012; McIntyre et al. 2014). This provides an ideal evolutionary  
75 context to study appearance and/or reduction/loss of complex morphological characters. The most  
76 comprehensive GRN model so far studied for an animal describes the development of the larval  
77 skeleton in the sea urchin *Strongylocentrotus purpuratus* (Rafiq et al. 2012, 2014; Oliveri et al.  
78 2008). It explains how in the course of development dozens of regulatory genes act together to  
79 specify a mesodermal cell population, which later form two ventro-lateral clusters on each side of  
80 the primitive gut (archenteron) and finally secrete the calcitic endoskeleton typical of the sea urchin  
81 pluteus larva (reviewed in McIntyre et al. 2014). Interestingly, whereas around 30 transcription  
82 factors (TF) and a few signalling pathways are sufficient for the initiation, progression and  
83 maintenance of this process (Oliveri 2008), more than 800 genes participate in the final step of cell  
84 differentiation and biomineralization of organic matrix. These differentiation genes have been

identified using transcriptomic and proteomic experimental strategies (Rafiq et al. 2014; Mann et al. 2008, 2010; Barsi et al. 2014), although their role and GRN linkages are largely un-explored. The extensive level of detail of the sea urchin GRN underlying skeletogenesis provides a useful framework to address questions about the evolution of development mechanisms through comparison with other echinoderms. Indeed, expression data are already available for a few orthologs of sea urchin skeletogenic transcription factor genes that have been identified in representatives of all echinoderm classes except crinoids (Dylus et al. 2016; McCauley et al. 2012; McCauley et al. 2010; Seaver & Livingston 2015). However, there has been relatively little comparative analysis of genes involved in skeletal differentiation in echinoderms.

Recently, biological and evolutionary studies have been transformed by immense technological improvements in sequencing technology (Goodwin et al. 2016). Relevant to this study, RNA sequencing is now an established technique that provides a practical and cheap alternative to whole genome sequencing (Wang et al. 2009) because it allows rapid advancements in molecular genetic analysis of organisms for which limited or no genomic data are available but which are of great interest from an evolutionary and/or developmental perspective. Importantly, RNA sequencing enables a global quantitative analysis of gene expression at specific stages of life and/or in particular tissues/organs. In this way it is possible to reconstruct the timeline of expression of each individual gene and determine the progression of regulatory states, which is a key first step when analysing gene regulatory networks (Materna & Oliveri 2008).

The large amount of molecular genetic information in echinoids compared to other echinoderm classes can be attributed to the fact that sea urchins have been studied extensively for over 100 years. Furthermore, the genome of the sea urchin *Strongylocentrotus purpuratus* was sequenced 11 years ago (Sea Urchin Genome Consortium 2006) and together with several improvements and additional mRNA sequencing data provides a very high quality resource (Tu et al. 2012, 2014). So far within the echinoderms, only the *S. purpuratus* genomic resources are of a high standard, although many additional species have been sequenced at lower quality (Cameron et al. 2009). Very recently the genome sequence of the Indo-Pacific sea star *Acanthaster planci* was published (Hall et al. 2017). Furthermore, transcriptomic data are available for several

echinoderm species, but with significant variation in sequencing depth and quality and with most datasets limited to a single life-stage or tissue (Janies et al. 2016; Israel et al. 2016; Vaughn et al. 2012).

Within the echinoderms, the brittle star class has received growing attention in recent years (Purushothaman et al. 2014; Burns et al. 2011; Czarkwiani et al. 2013; Dupont et al. 2010) due to their phylogenetic position as a sister group of sea stars, their mode of development and regenerative capabilities. For instance, brittle stars develop a skeleton in the larvae similar to sea urchin (Dylus et al 2016; Primus 2005) and are thus a valuable model for addressing questions relating to differences and conservation of developmental genes involved in the formation of the larval skeleton. With this perspective, a single stage transcriptome identified many orthologs of sea urchin skeletogenic genes in a brittle star species (Vaughn et al. 2012), but no quantitative data on the dynamics of gene expression was provided. Furthermore, a comparison of skeletogenic regulatory states between an echinoid and an ophiuroid identified differences and similarities in the specification of the skeletogenic cell lineage (Dylus et al. 2016). Additionally, brittle stars regenerate their arms as part of their self-defence mechanism (Dupont and Thorndike 2006). The re-development of the skeleton has been characterized in detail with respect to morphology and gene expression during various phases of regeneration (Biressi et al. 2010; Czarkwiani et al. 2013, 2016; Burns et al. 2011; Purushothaman et al. 2014). Finally, brittle stars are used as important indicator species for ocean acidification studies (Dupont et al. 2010).

Here we present a *de novo* transcriptome for the brittle star *A. filiformis* obtained using four key stages of development, with the aim to provide a global quantitative assessment of developmental gene expression. We devised a computational strategy to generate a high-quality reference transcriptome, supported by several quality measures, and a reliable quantitative gene expression profile, validated on several candidates with other gene expression profile platforms, such as quantitative PCR and Nanostring. Focusing on the distinct feature of larval skeleton evolution within echinoderms, we assess the conservation of gene content by a large-scale comparison of our transcriptome with sequencing data from an asteroid, a crinoid and an echinoid. Our results reveal a high-degree of conservation of genes associated with skeleton formation in the

141 four species, consistent with the fact that all classes of echinoderms have a well-defined adult  
 142 skeleton that originated at the base of the *phylum*. Contrary to previous studies, we identify major  
 143 differences in the temporal expression of regulatory genes, which suggests a high-degree of re-  
 144 wiring for the developmental GRN. Furthermore, applying a fuzzy clustering approach, we find that  
 145 most skeletogenic differentiation genes exhibit an increasing trajectory of expression during  
 146 development consistent with their hierarchical position as the final tier of a GRN. We also present  
 147 an R-shiny application to allow access to all of the data presented here for future analysis.

148

## RESULT

### Assembly of a reference transcriptome for *A. filiformis*

Given the similarity of development between sea urchin and brittle star (Dylus et al. 2016; Primus 2005) we performed a global comparative analysis of the gene complement and gene expression profiles of representatives of these two classes of echinoderms. To enable this, we characterize for the first time the expression of genes in the brittle star *A. filiformis* using RNA-seq technology at four chosen key developmental stages that extend over the entire development of the larval skeleton, from early cell specification to final cell differentiation. The developmental stages are: end of cleavage stage (9 hours post fertilization; hpf), a hatched blastula stage (18 hpf), 3 samples for mesenchyme blastula stage (27 hpf) and a late gastrula stage (39 hpf) (Fig. 1A). For the sequencing, we multiplexed the 6 samples using 100bp paired-end reads on 2 lanes of Illumina HiSeq 2500, resulting in ~100Million reads per sample (STab. 1 and SFig. 1). Joining all the reads from each sequenced sample and following the khmer-protocols v0.84 (Brown et al. 2013), we assembled a reference transcriptome that would reflect all protein coding genes expressed in the analysed stages (Fig. 1B). In this three step assembly, we first trimmed all reads for Illumina adapters and low quality basepairs, then applied digital normalization to remove overrepresented reads and erroneous k-mers (Brown et al. 2012) and finally used the resulting reads as input for Trinity (Grabherr et al. 2011) (STab. 1). Our initial assembly resulted in 629,470 sequences (N50: 1094). To determine whether the digital normalization step introduced artefacts, we assembled each individual sample omitting this step and compared with the combined assembly. We recovered over 94% of sequences using a BLASTn search (e-value 1E-20) of each individual assembly against the combined assembly (SFig. 2). Thus, we concluded that the digital normalization step did not introduce any significant bias in the combined assembly.

Because the focus of this study was on protein-coding transcripts, we filtered our initial combined assembly for all open reading frames that have an uninterrupted coding region longer than 300bp (=100aa) using the TransDecoder package (Haas et al. 2013). This reduced our dataset to 92,750 protein coding sequences. We further removed any potential bacteria contaminates through application of a BLASTx search against 12,537,847 bacterial proteins



(Uniprot DB; Bacteria Release 2014\_06; 2563 species) and crosschecked the identified sequence for closer percentage of identity with hits obtained using a BLASTx (both e-value: 1E-20) search against the Uniprot SwissProt DB (Release 2014\_07). Finally, we were left with 91,311 contigs (N50: 1410) constituting our reference transcriptome (RefTr) (Tab. 1). The number of contigs produced by *de novo* transcriptome assemblers is typically large as assemblers cannot differentiate between isoforms or alternative transcripts of the same gene and thus report each separately (reviewed in Moreton et al. 2014). Moreover, artefacts such as repeats, sequencing errors, variation in coverage or genetic variation within a diploid individual, create contigs that are not truly representative of different isoforms. As a result, transcriptome assemblers often report repeated contigs that differ only by a single nucleotide polymorphism (SNP), indel or fragmented versions of a transcript (reviewed in Moreton et al. 2014). Moreover, simulation studies using error-free reads showed that *de novo* assemblers inevitably produce multiple contigs for the same gene (Vijay et al. 2013). To account for this type of variation in the absence of a reference genome, but without losing sequences, we partitioned similar contigs that differ due to SNPs or indels into transcript families that share a protein identity of at least 97%. On average this approach grouped 1.3 contigs to each transcript family, resulting in 67,945 total transcript families. Unfortunately, splice variants and other artefacts are not incorporated into this type of clustering, leading to a number still larger than expected when comparing with the gene set of the sea urchin *S. purpuratus* gene set (~21,000; Tu et al 2012), the only echinoderm for which high quality genome sequence data was available when this study was conducted.

To test the quality of our assembly, we compared our RefTr with 48 isolated clones containing coding (cumulative length of 32,769bp) and UTR regions (cumulative length of 7,091bp) sequenced using Sanger sequencing technology. Using BLASTn and collecting only the top hits, we obtained an average percentage of identity of 98.6%. On an average alignment length of 588bp we found ~7 mismatches in coding sequence, resulting in an average polymorphism in coding sequences of 1.2%, a value to be expected based on the fact that clones were obtained from various batches of cDNA that are different from the samples used for the RefTr. In conclusion, we

produced a high-quality reference transcriptome assembly that will provide a valuable resource for future studies in brittle star biology..

## Gene content of *A. filiformis* based on analysis of the developmental transcriptome

In order to have meaningful comparative analysis of gene expression between brittle star and sea urchin clades, which diverged roughly 480 mya (O'Hara 2014), we first classified and annotated the gene content of our RefTr and then assessed the evolutionary conservation of genes in the Echinodermata to better understand at a global level the conservation of genes and appearance of novel genes.

For this aim, and to be as comprehensive as possible, we applied independent search methods. First, we used the Blast2GO tool (Conesa & Götz 2008) that assigns gene ontology terms to each contig. Blast2GO first uses a BLASTx search (evalue: 1e-3) against the GenBank non-redundant database and this search resulted in hits for 62,388 Afi contigs corresponding to 26,010 unique genes from 1,334 different species. Consistent with ophiuroids being echinoderms, most hits were found for *Strongylocentrotus purpuratus* (25,882/62,388 contigs), followed by the hemichordate *Saccoglossus kowaleskii* (SFig. 3). The second step of the Blast2GO pipeline performs an InterProScan to find regions within contigs that have conserved protein-coding domains. This step found 66,071 contigs with at least one region that has a recognizable protein domain. The combination of the BLASTx and interpro searches was then used to assign gene ontology terms, which provided functional classifications for 27,923 of our contigs (SFig. 3).

To proceed with a general assesment of the evolution of gene content specifically in the Echinodermata, we collected in addition to the ophiuroid *A. filiformis* transcriptome (this study) representative datasets from the draft genome sequence of the asteroid *Patiria miniata* (Pmi) (Baylor College of Medicine: HP081117-HP139664), the genome sequence of the euechinoid *S. purpuratus* (Spu) (Tu et al. 2012; Sea Urchin Genome Consortium 2006), and the transcriptome of the skeleton-rich adult arm of the crinoid *Antedon mediterranea* (Ame) (Elphick et al. 2015) (Fig. 2A). Differences in samples, sequencing technology and assembly strategy make gene content comparisons from different species difficult. Therefore, we computed quantity and quality metrics

allowing us to make meaningful statements in relation to the properties of the individual datasets (see STab. 2 and 3 ; SFig. 4). Importantly, at the time of the study only the sea urchin dataset had a well-curated genome and was improved by additional deep coverage transcriptome data (Sea Urchin Genome Consortium 2006; Tu et al. 2012) and is thus used here as reference for comparative analysis. Our analysis indicated that all datasets are of comparable high quality and based on our tests can be ranked from lower to higher quality as follows: Ame, Afi, Pmi and Spu (see STab. 2 and 3 ; SFig. 4).

To gather information about the echinoderm-specific gene content we used a union of the Spu genesets predicted from genome and transcriptome databases (29,072) to identify genes in Afi and the other echinoderm species by applying a tBLASTx (evalue:  $1e-6$ ) search. For the identification we followed the khmer-protocols v0.84 (Brown et al. 2013). In this protocol a reciprocal blast is used on the sequences partitioned into transcript families. Reciprocally identified sequences are classified as orthologs and unidirectional identified sequences as homologs. Additionally, for contigs part of the same transcript family the BLAST result is propagated in order to ensure that the identification is consistent with the partition. Using this protocol, we found homologs of Spu proteins for ~58% of Afi RefTr sequences, for ~75% of Pmi genome and transcriptome derived contigs, but for only ~6% of Ame transcriptome-derived contigs. Detailed numbers are presented in Tab. 2. Importantly, the largest number of homologs of sea urchin proteins were identified in Pmi (16,211), followed by Afi (13,656) and Ame (12,982). This finding is consistent with the fact that the Pmi dataset is a combination of contigs derived from both genomic and a transcriptomic data, whereas the Afi and Ame datasets are derived solely from transcriptomes. As a positive control for our strategy, we searched the Spu dataset against itself and found 91% (Tab. 2) of hits with an evalue less than  $1e-6$ . The residual 9% of protein coding sequences are likely to be highly similar sequences such as recently duplicated genes, different alleles, or potential wrongly annotated genes, which in general fail to give a clear unequivocal result using a BLAST alone approach.

To determine the extent of sequence conservation in the echinoderm phylum we computed the overlap of contigs shared between species. Therefore, we searched reciprocally all versus all species (tBLASTx, evalue:  $1E-6$ ) using each time one of the four species as a reference (Fig. 2B).

Our analysis shows that around 6,000 sequences are common to all species analysed, corresponding to 25% of the protein coding sequences of the sea urchin reference species. Any other combination of 2-3 species identified at least 1,000-2,000 shared genes. This suggests that in each class a specific subset of ancestral genes has been retained and consequently that others have been lost or have diverged beyond recognition with the methods employed here. Notably, we observed a higher number of genes to be shared between Afi and Pmi compared to other pairs of species (Fig. 2B). This is consistent with the recently published phylogenetic analysis of echinoderm relationships, in which sea stars and brittle stars are sister groups (Telford et al. 2014; Cannon et al. 2014). To validate this result, we applied the orthology matrix algorithm (OMA) (Roth et al. 2008), which computes highly reliable groups of orthologous genes using the Smith-Waterman algorithm for sequence alignment. The set of orthologous genes obtained allowed us to clearly distinguish the differences in genes shared between species (Roth et al. 2008). Using OMA, we observe a much higher conservation between Pmi and Afi than in any other overlap of two species i.e. ~7000 orthologs compared to ~ 2000-4000 orthologs (SFig. 5). Moreover, the variation in the number of genes among species overlaps indicates a high level of evolutionary dynamicity in terms of gene conservation in the four classes of echinoderms analysed here. This is supported by the similar number of genes shared between two species and can be explained by the separation of the four classes early on in echinoderm evolutionary history (542 - 479 mya) followed by long periods of independent evolution (O'Hara et al. 2014; Pisani et al. 2012).

## **Functional characterisation of echinoderm genes reveals conservation of a regulatory toolkit in echinoderms**

A recent study explored in detail a developmental transcriptome of *S. purpuratus* in terms of gene content and established echinoderm specific ontology classifications (Tu et al. 2012). Our high quality RefTr and consistent data treatment allowed us to apply this ontology classification and to compare the abundance of specific functional classes with other echinoderms. We queried our three species for the identified genes that belong to sea urchin functional classes (SUFC; Fig. 3). From a total of 6,461 genes classified in 24 SUFCs we found 4,494 homologs in Afi, 4,407 in

288 Ame, and 4,976 in Pmi. We classified the SUFCs in three categories of conservation using  
 289 manually selected thresholds. In the first category of highly conserved SUFCs ( $\text{avg}(\text{Afi}, \text{Pmi},$   
 290  $\text{Ame}) > 80\%$  of identified Spu sequences), we find Cytoskeleton, Phosphatase, Signaling,  
 291 CalciumToolkit, CellCycle, TF, DNAReplication, GermLineDeterminant, TranslationFactorTF (Fig.  
 292 3). SUFCs that are conserved at a lower level (intermediate:  $\text{avg}(\text{Afi}, \text{Pmi}, \text{Ame})$  in between 70-  
 293 80% of identified Spu sequences) are Histone, Metabolism, Nervous, GTPase, Kinase and  
 294 EggActivation and whilst the lowest conservation of SUFCs ( $\text{avg}(\text{Afi}, \text{Pmi}, \text{Ame}) < 70\%$  of identified  
 295 Spu sequences) is observed for Biomineralization, Immunity, Oogenesis, Defensome, ZNF,  
 296 Apoptosis, Metalloprotease, Adhesion and GPCRRhodopsin (Fig. 3). Interestingly, the class of  
 297 Biomineralization, GPCRRhodopsin, Histones and ZNF shows the highest level of variation  
 298 between the three species ( $\text{sd} > 10\%$ ) and we find a high number of ZNFs only in brittle stars (Fig.  
 299 3).

300 To obtain a better picture of the conservation of the developmental program in general and the  
 301 evolution of the larval skeleton in particular, we focused our analysis on regulatory genes (TF and  
 302 Signalling) and on biomineralization differentiation genes. Out of 368 sea urchin TF, we identified  
 303 304 in the brittle star, 297 in the crinoid and 326 in the starfish. 304 TF in the brittle star correspond  
 304 to 82% of the sea urchin transcription factors (TF) and represent the cohort of TF used in this  
 305 species throughout development, a number comparable to estimates obtained for sea urchin  
 306 development (~80% of 283 transcription factors are expressed by late gastrula (Howard-Ashby et  
 307 al. 2006)). Consistent with the fact that the starfish dataset is a combination of genome and  
 308 transcriptome we find the largest number of homologs of sea urchin TFs (326) in this class of  
 309 echinoderms. On the contrary, the lowest degree of conservation was observed in the crinoid (297  
 310 out of 368), which might be attributable to the fact that the Ame transcriptome was obtained from a  
 311 single adult structure (the arm), although arms are formed from multiple tissue types. Generally, a  
 312 similar degree of conservation was observed for signalling molecules (~76-87%), but with more  
 313 variation between Pmi, Ame and Afi (Fig. 3). The high level of TF and signalling conservation  
 314 indicates that echinoderms share a similar regulome.

The biomineralization SUFC shows a higher degree of variation and we find generally less genes ~41-60%, or a lower percentage of conservation. Interestingly, when looking more thoroughly in the bio-mineralization class of genes, of the 14 spicule matrix (sm) genes only one gene in Afi and one gene in Pmi were identified, indicating that the sm class of genes is significantly reduced in the Asteroidea and Ophiuroides, by comparison with the Echinoidea. Homologs of more than 50% of Spu genes belonging to the collagen, cyclophilin and carbonic anhydrase categories (STab. 4) were found all species. Interestingly, in a first assessment we found few homologs of the nine Spu msp130 genes in the species analysed here (two sequences in Afi, three in Pmi, and four in Ame), although many contigs showed sequence matches. Therefore, we investigated if there are actually more msp130 genes in the other species that Blast algorithm alone is not able to discriminate. Using 18 candidate genes, we generated a multiple sequence alignment and built a hidden Markov model (<http://hmmer.org>, version 3.1b) in order to scan for other contigs with a msp130 signature. With this approach, we found several candidates in our dataset that had this signature but were different in terms of their amino acid sequence. In order to investigate their relation to the sea urchin msp130 genes we built phylogenetic trees using Bayesian and maximum likelihood methods using also genes found in outgroup species. Our trees support class specific duplications of msp130 genes, as displayed by their independent expansions in different branches of the tree (SFig. 6). This analysis suggests that while all echinoderms share a similar regulome, defined as the cohort of all TF and signalling genes encoded in a genome, some classes of sea urchin biomineralization genes are either absent or duplicated independently when compared to the other 3 species analysed here.

### **Skeletogenic genes are conserved within the echinoderms**

All echinoderms develop a calcite skeleton and hundreds of genes are involved in this process. However, the SUFCs of sea urchin contain only 56 genes that are classified as biomineralization genes. To obtain a more precise picture of genes involved in skeletogenesis and their evolution we gathered 1006 sea urchin skeletogenic candidates based on literature searches. This extended candidate list was compiled from proteomic studies based on skeletal elements

343 obtained from adults and larvae (Mann et al. 2010), a differential analysis of sea urchin  
344 mesenchyme blastula where skeletogenic mesenchymal cells were removed (Rafiq et al. 2014) or  
345 isolated (Barsi et al. 2014) and a large scale morpholino analysis (Oliveri et al. 2008); it is therefore  
346 representative of the skeleton developmental process from cell specification up to the deposition of  
347 the biomineralized skeleton. We updated this list with the latest annotation of the sea urchin  
348 genome and obtained 901 genes (SFile 1). Of these 901 candidates, 37 are TF and 32 are  
349 signalling molecules belonging to 5 different pathways (*i.e.* Fgf, Vegf, Delta/Notch, Wnt and BMP),  
350 whilst the rest of the genes belong to various classes of C-type lectin-type domain, carbonic  
351 anhydrases, matrix metalloproteases, known skeletogenic matrix genes (sm and msp130) and  
352 others. To maintain a very broad view, we searched the homologs of our annotated species for  
353 these candidates with the aim to find a core set of skeletogenic genes and possibly a set  
354 specifically used in the development of the larval skeleton in echinoids and ophiuroids. We find 601  
355 candidate skeletogenic genes in *Ame*, 622 in *Afi* and 672 in *Pmi* out of 901 *Spu* genes, which follow  
356 a trend similar to the whole gene set. To display the differences in skeletogenic gene conservation  
357 we computed the overlaps between the four species (Fig. 4). Due to the fact that skeletogenesis in  
358 the adult is a feature present in the common ancestor of extant echinoderms we wanted to check  
359 whether the 494 skeletogenic genes found in all four species are more highly conserved than a set  
360 of randomly selected genes. Therefore, we computed the overlap of 901 genes selected randomly  
361 1,000 times and compared it with the skeletogenic gene set (SFig. 8). Our analysis suggested that  
362 genes associated with the skeletogenic process are more conserved than a set of random genes  
363 (compare 494/757 to 278/613, chi-squared proportion test  $p < 0.001$ ; Fig. 4 and SFig. 7). This is in  
364 line with the evolution of the biomineralized ossicle in the form of stereoms at the base of the  
365 echinoderms and a high level of conservation of this structure throughout evolution. Although, this  
366 analysis gives us a good indication of the presence or absence of genes in the different classes of  
367 echinoderms, it does not provide evidence that these genes participate in skeleton formation.  
368 Recently, using a candidate approach we showed in a multi gene expression study that of 13 TFs  
369 involved in *Spu* skeletogenesis 10 are active in *Afi* development, whilst the other 3 although  
370 expressed during development are not localized in cells giving rise to skeleton (Dylus et al. 2016).



This highlights the importance to complementing transcriptomic data with spatial/temporal analysis of gene expression. Therefore, we selected from our list of 622 skeletogenic homologs 11 candidates of the differentiation cascade to investigate if they are expressed in the skeletogenic mesoderm (SM) lineage in brittle star (Fig. 4). We found that all of these genes are either expressed specifically or are enriched in skeleton-associated cells during the development of *A. filiformis*. Most of them seem to be specifically enriched in the SM lineage at late gastrula stages in cells where the skeleton is deposited. Together with our previous analysis of developmental regulatory states (Dylus et al 2016) a total of 24 genes show expression in cells associated with biomineralised skeleton conserved in two distant clades: sea urchin and brittle star. This indicates a largely similar molecular make up of calcitic endoskeleton (65%) in sea urchin and brittle star; and it is consistent with the ancient origin of the biomineralized skeleton in the form of stereom, which originated at the base of the Echinodermata phylum.

# **A quantitative developmental transcriptome for *A. filiformis* to assess the dynamics of gene expression.**

Our prior analysis indicates that skeleton forming genes are well conserved within the echinoderms, but what about the regulatory program? The developmental regulatory program is executed by a large GRN that tunes the expression of thousand of genes. To have an initial global assessment of the *A. filiformis* regulatory program we took advantage of the separate sequencing of four key developmental stages and the ability to obtain quantitative data from RNA-seq. While being relatively trivial to align reads when well curated gene models exists, this task is complicated for *de novo* assembled transcriptomes due to the high level of contig redundancy. In order to address this issue, we used the CORSET algorithm (Davidson & Oshlack 2014). CORSET removes sequences with less than 10 reads, which correspond to technical background level and groups contigs to expression clusters (EC) that share the same reads, thus resulting in expression values that are equivalent to potential gene counts. In a first step this algorithm removed 9,854 sequences that were expressed with less than 10 reads. The resulting 81,457 contigs were then clustered to 37,999 ECs (min: 1seq, max: 66seq, mean: ~2.1seq per cluster; SFig. 8). In order to



normalize the dataset relative to an internal standard, we computed the standard deviation for each EC between the 4 time points and selected 331 ECs with  $sd < 0.01$  (a list of all ECs can be found in SFile 2). We then divided the RPKM corresponding to each EC by the average of the 331 ECs and multiplied each by 1 million to normalize and to obtain EC counts in transcripts per million (tpm). Because of the grouping of contigs into ECs, the previous annotation could not be directly propagated. Therefore, we associated to each EC the most frequent annotation of its containing contigs, giving orthologs priority over homologs. This caused a reduction from 13,656 to 11,695 uniquely found sea urchin sequences in Afi. Of the reciprocally identified sequences only 350 were lost during this process, resulting in 9,429 reciprocally identified sea urchin sequences. Possible reasons for this reduction are the filtering of low level of expressed sequences (less than 10 reads; see above), and contigs mapping to different genes in sea urchin actually belonging to a single one. A summary for losses mapped to the SUFCs is presented in SFig. 9. To estimate the quality of our approach we compared 29 genes quantified using qPCR and 86 genes quantified using Nanostring in different RNA batches with the corresponding ECs. We obtained a high correlation between qPCR and ECs ( $r^2 = 0.84$ ) and between Nanostring (Geiss et al. 2008) and ECs ( $r^2 = 0.77$ ), supporting our quantification strategy (SFig. 10 and SFig. 11). These quantitative data are now available for evaluating dynamicity of gene expression and comparative analysis and will be used for comparative gene expression with sea urchin.

## **Temporal mode of TF expression in the brittle star shows many differences with the sea urchin.**

In order to obtain a global view of time-series expression during development and to group the genes by similar expression patterns we applied a fuzzy clustering approach (Futschik & Carlisle 2005). Based on the fact that between four time points there are 3 possible modes of expression (steady, increase or decrease) we decided to assign to each EC one of 27 fuzzy clusters (FC). This algorithm assigned 27 fuzzy clusters to the 37900 ECs. During this process 99 ECs were lost because they were not active throughout our 4 developmental time points but were

expressed in one of the other two 27hpf samples that were not used for this analysis. We re-iterated this algorithm 100 times and optimized the membership of each EC to a specific FC. A closer look into the 27 FC showed four distinct modes of dynamical behavior and we decided to use this grouping for future analysis. The groups were EARLY with 10,593 FCs, INTERMEDIATE with 8,531 FCs, LATE with 9,968 FCs, and BI-MODAL with 8,808 FCs (Fig. 5 A). EARLY FCs contained ECs that showed decreasing expression across the first 3 time-points and thus likely to have a role during very early development (9hpf, end of cleavage) . In these FCs we found genes that are responsible for early specification and are only transiently active. In total we found 59/287 TFs and 105/561 skeletogenic genes that showed a decreasing trajectory over the 4 time-points. In this group, only *Afi-pplx* was found as a gene involved in Afi skeleton specification. In the INTERMEDIATE group were genes whose expression trajectories peak either at 18hpf or 27hpf and then decrease steadily. Examples of genes found in this group are *Afi-alx1*, *Afi-tbr*, *Afi-gataC* and *Afi-erg*, TFs that have been shown to be expressed in mesodermal cells of the Afi embryo and known to play a role in the specification of mesoderm (Dylus et al. 2016). In total this group comprises 66/287 TFs and 68/561 skeletogenic genes. In order to form the extended larval skeleton, we expected most of the skeletogenic genes previously described to be expressed at the moment of the deposition of the calcite skeleton, and therefore to show an increasing pattern of gene expression. Indeed most of the skeletogenic genes were clustered in the LATE group 287/561. Among others, this group contained the biomineralisation genes *Afi-p19 (Cah10L)*, *Afi-p58a*, *Afi-p58b*, *Afi-ttrspn\_19*, *Afi-slc4a10/nbc* and *Afi-c-lectin*, all expressed in skeletogenic cells in brittle star (Fig. 3 and Dylus et al. 2016). Moreover, the LATE group contained most of the active TFs 132/287, consistent with the increasing complexity of cell types over developmental time. The final group, called BI-MODAL consists of two expression peaks throughout the 4 time-points and contains 30/287 TF and 101/561 skeletogenic genes. This group contains genes that might be expressed in different domains during development, potentially having two (or more) roles throughout development. Examples are *Afi-hesC* and *Afi-delta*, which are first expressed in the mesodermal cells at the vegetal side of the embryo at blastula stage (18 hpf) and then in scattered cells in the ectoderm at gastrula stage (39 hpf) and at the tip of the archenteron throughout

gastrulation (Dylus et al. 2016). Based on the fact that our 4 time-points correspond to 4 different stages of development, our grouping shows consistent activity of TF involved in multiple stages of cell specification.

The direct output of a GRN is the temporal expression profile of each gene throughout time and each expression profile is linked to its regulatory state. Therefore, comparing temporal expression profiles between two species provides a first glimpse on GRN rewiring and on heterochronic gene expression. In order to evaluate the differences and commonalities of TF usage between sea urchin and brittle star, we selected 4 time points that correspond to similar stages of development from the sea urchin transcriptome: they are 10, 18, 30, 40 hours post fertilization (hpf) in agreement with the comparative developmental stages previously described (Dylus et al 2016). On these *S. purpuratus* transcriptome stages we performed a fuzzy clustering as for *A. filiformis*, and we then grouped the clusters based on the above-mentioned criteria. We identified in the EARLY category 72, the LATE 110, the INTERMEDIATE 110 and in the BI-MODAL 70. out of the 368 TFs, and six genes are not classified due to the too low levels of expression. When comparing TF expression, and therefore the developmental regulatory states between *S. purpuratus* and *A. filiformis*, many differences emerge in the four categories. In all four categories we see more variation than overlap of TFs (Fig. 5 B). For example, only 15 transcription factors in the EARLY category are in common between the two species (e.g. *pmar1* and *soxB1*), whereas 44 Spu homologs in Afi categorized as EARLY differ from other 57 TF in Spu are expressed in this mode. Other examples of common transcription factors are for BI-MODAL *groucho*, for INTERMEDIATE *alx1*, *erg*, *foxM*, *mitf*, and for LATE *foxB*, *hnf4*, *tgif*. A summary of all TF can be found in SFile 3. This comparison highlights that TFs are used differently, or at least with a different timing of expression, during the development of the two species. Examples of such genes are *hesC* and *ets1/2*. Notably, there are more differences in the early phases of development when cell specification begins, than in the late stages when cells initiate the final differentiation. Given that the direct output of a GRN is reflected in the temporal gene expression, this suggests differences in the topology of gene regulatory networks between Afi and Spu.



## DISCUSSION

Here we present a *de novo* transcriptome of *A. filiformis* that samples four important stages of the embryonic development of this organism. We also present an overall strategy to effectively compare different data sets and to use RNA-seq quantitative data in the absence of a reference genome. Our data and assembly/annotation strategy is then used to obtain insights into two key evolutionary questions: how did the larval skeleton in echinoderms evolve and how conserved is the regulatory program of the pluteus larvae of sea urchins and brittle stars?

To assemble the *A. filiformis* RNA-seq data, we used a strategy with digital normalisation followed by application of the Trinity assembly. Our approach with digital normalisation allowed us to obtain a reference transcriptome that incorporated six independent samples within four weeks of computation on a server with only 64GB of memory RAM, with quality comparable to assembly obtained with non-normalised data. Our comparison is in agreement with what was observed by Lowe et al. 2014 for the assembly of sequence data from two closely related ascidians, for which a systematic comparison of assembly with and without digital normalisation showed no inclusion of computational artefacts, but a reduction of time and resources needed for the assembly. We showed that our RefTr is of high-quality by various computational and experimental methods and we also applied the computational quality control to the other datasets to strengthen the subsequent comparative analyses. In the developmental transcriptome the depth of sequencing (~100M reads per sample) and the combination of samples from multiple stages were important driving factors that made such a high quality assembly possible. Altogether our analysis shows that deep sequencing combined with a good pipeline can result in an assembly that is comparable to a genome in terms of gene capture. This is illustrated by the high number of genes that showed more than 90% identity to genes in the Swissprot database. Thus, our transcriptome performed best when compared to other genome and transcriptome datasets (SFig. 4). Interestingly, our extraction of protein-coding genes reduced the total number of contigs from ~600,000 to ~90,000 (15%) increasing the N50 value, however not affecting gene recovery, as shown in the CEGMA test (STab. 3). Based on our analysis only 15% of the Ref Tr sequences are protein-coding, giving rise to a particular question: what are the residual 85% of sequences? One possibility is that they

are part of non-coding sequences (e.g. non coding RNA, transcribed pseudogenes) or partially or wrongly assembled transcripts. Efforts to obtain genome sequence data for *A. filiformis* are underway to help obtain answers to these questions. Indeed, studies on human genomes show that more than 60% of the genome is reproducibly represented in long RNA sequences, while only 2.9% are represented by protein-coding sequences (The ENCODE Consortium 2012).

During the Cambrian period the rapid expansion of animal life was associated with acquisition of the capacity to form hard mineralized tissues, as testified by the first appearance of a fossil record for many phyla. Amongst others, echinoderms evolved their characteristic calcitic porous endoskeleton formed of magnesium rich calcium carbonate and occluded proteins (Bottjer et al. 2006; Gilbert & Wilt 2011). A first step towards understanding evolution and developmental genetics of a complex character such as mineralized skeleton is to perform a comparative and phylogenetic analysis of gene content (Fig. 2). For this reason we compared four echinoderm classes, 3 of the Eleutherozoa subphylum (Echinoidea, Ophiuroidea and Asteroidea) and a crinoid outgroup, with a focus on the genes involved in skeleton formation. Studies on sea urchins have shown that several genes used during adult skeleton formation are also used in larval skeleton (Gao & Davidson 2008; Mann et al. 2010), leading to the idea that an ancient regulative and differentiation module originated at the base of the phylum Echinodermata and then was secondarily co-opted to form larval skeleton. However, it is hotly debated whether this happened only once in the branch leading to the Eleutherozoa, or whether it occurred independently in both the sea urchin (Echinoidea) and brittlestar (Ophiuroidea) lineages. The two transcriptomes used in this analysis correspond to stages (late gastrula, for *A. filiformis*) or structure (adult arm for *A. mediterranea*) in which the biomineralized skeleton has been deposited. Therefore, expression of genes involved in this process must be highly represented. It is important to clarify that due to nature of this comparison, genome vs transcriptome, we can unequivocally evaluate only the gene (or protein coding transcripts) presence in at least two data set. On the other hand, in the context of skeleton development, the gene absence in *A. filiformis* and *A. mediterranea* can be interpreted as lack of expression in stages or structures with skeleton, sequenced in this study, and hence no-usage in building the skeleton of these two organisms.

Our analysis revealed a gene toolkit of 494 genes conserved in all four echinoderm classes (Fig. 4), which potentially corresponds to the echinoderm core of skeletogenic genes. Indeed our analysis of spatial expression shows that several of these genes are expressed in cells known to form the skeleton in the developing *A. filiformis* embryo (Fig. 4; Dylus et al. 2016) and for few of them also known to be expressed during *A. filiformis* adult arm regeneration (Czarkwiani et al. 2013, 2016). Of the initial 901 gene set only 37 are TFs and 32 signaling molecules. Of these regulatory genes, 84% (58/69 regulatory genes) are conserved in all the echinoderm classes analysed, while the other genes, which can generally be classified as differentiation genes, are conserved only for the 52% (436/832) in all the classes, indicating a higher conservation of the skeletogenic cells regulatory program and a rapid evolution of echinoderm skeleton-forming genes. A closer look into these 436 genes using the sea urchin functional classes revealed that metalloproteases and biomineralization genes are actually the most variable class of genes (SFig. 9). This observation indicates that solely looking into these two categories can produce a biased picture of evolution, because only these two categories of differentiation genes showed high level of variation and indicate low selective pressure. How can we explain the variation in the biomineralization genes? They are grouped in six categories of which collagens, cyclophilins, carbonic anhydrases and an unnamed category (Tu et al. 2014), which include P16 (Cheers & Ettensohn 2005) and other genes, are highly conserved in our selected representatives of the four classes of echinoderms. On the other hand, of these six categories, msp130 and spicule matrix (sm) genes show the highest level of variation. Indeed, of the nine sea urchin msp130 genes only two are found in all four species analysed (*Spu-Msp130r6* and *Spu-Msp130L*). An in depth look into the brittle star transcriptome, using a hidden Markov model, revealed also the presence of seven other msp130 contigs that show differences at the amino acid level higher than the 1.2 % of polymorphism identified in coding region, suggesting the presence of several genes. Indication of clade specific expansions took place is strongly supported by our phylogenetic analysis (SFig. 6), which shows a consistent group of sea urchin *Msp130* genes with various paralogues represented in both sea urchin species analysed (*S. purpuratus* and *L. variegatus*), a different group of Ophiuroid *Msp130s*, as well as other clade specific expansions consistent with what has already



shown for *Msp130* genes in molluscs and annelids (Szabó & Ferrier 2015). Concerning the spicule matrix (sm) genes, out of the 14 genes identified in sea urchin only the C-lectin that does not contain a proline rich region is conserved in all four species. Therefore, no spicule matrix genes, characterised by a C-lectin domain and a conserved proline-rich domain (Livingston et al. 2006), are found in any other class of echinoderm in stages when skeleton is built, making them likely to be a sea urchin specific set of skeletogenic matrix genes. Further support for this hypothesis is provided by the following observations: First, a proteomic study of skeletal elements in another species of brittle star *Ophiocoma wendtii* did not find orthologs of these genes (Seaver & Livingston 2015). However, other potential candidates of c-lectin type genes for brittle star skeletogenesis were obtained, which are also present in our transcriptome of *A. filiformis* and which are expressed during larval and adult skeletogenesis (Dylus et al. 2016; Czarkwiani et al. 2016). Second, in the *S. purpuratus* genome the sm genes are present in mini clusters of tandem repeated genes (STab. 6 and SFig. 12) suggesting a relatively recent duplication of these genes in the sea urchin lineage. Third, no such gene has been found in the hemichordate *Saccoglossus kowalevskii* (Cameron and Bishop, 2012), an outgroup of all echinoderms, confirmed by the absence of spicule matrix genes in the crinoid transcriptome analysed in this work (STab. 4). Both spicule matrix genes and *msp130* genes have been highly duplicated in sea urchin, as seen in the many tandem duplications, and the presence of both in the pencil urchin *Eucidaris tribuloidis* (Cameron et al., 2009) suggests their evolution already in the common ancestor of cidaroids and euechinoids. In this context, it would be interesting in future studies analyses holothuroids as sister class to the echinoids to pinpoint more exactly the evolutionary of this category of biomineralization genes. Interestingly, similar to these findings in echinoderms, the rapid parallel evolution in different lineages of genes associated with skeleton formation has also been reported for shell genes in molluscs and brachiopods (Jackson et al. 2010; Luo et al. 2015).

The fact that both *msp130* and sm genes are expressed in both adult and larval skeletal structures in sea urchin (Mann et al. 2010) suggests that the evolution of sm genes in echinoids and the independent expansion of *msp130* genes occurred before the evolution of the echinopluteus, the sea urchin larva with extended skeleton. Similarly, in brittle stars the specific *Afi-*



595 *Msp130L* is expressed in the larval skeletogenic cells, supporting the argument that larval  
596 skeletogenesis evolved independently in the two lineages, potentially in both cases as a co-option  
597 of the adult skeletogenic program after clade-specific gene expansion took place. Other evidence  
598 in support of evolutionary divergence of the echinoid and ophiuroid pluteus larvae is provided by  
599 our comparative analysis of regulatory states in developing embryos (Fig. 5), defined as the sum of  
600 transcription factors expressed in a given cell at a given developmental time. We compared the  
601 transcription factor usage in *S. purpuratus* (Tu et al. 2012) with the one of *A. filiformis*, taking  
602 advantage of the quantitative aspects of transcriptome data and the sequence data from four key  
603 developmental stages: cleavage stage (9hpf), when maternal mRNA are still present and the  
604 zygotic genome starts to be expressed; blastula stage (18hpf), when territories that will give rise to  
605 multiple cell types are specified and transcription factor genes are expressed in a territorially  
606 restricted manner (Dylus et al. 2016); mesenchyme blastula (27hpf), when territories are further  
607 subdivided, cells continue in their specification pathway, and morphogenetic movements  
608 commence; and finally gastrula stage (39 hpf), when cell types are specified, morphogenetic  
609 movements are almost completed and cell differentiation is underway. This comparison shows that  
610 the early regulatory states, which determine the developmental GRN, of these two species are  
611 quite different. On the contrary, when cell types are specified and terminal selector genes (LATE  
612 genes in this analysis) are expressed (Hobert 2008), they show a similar regulatory make up in  
613 these two classes of echinoderms, suggesting extensive GRN rewiring in the early stages of  
614 development. Taken together, our findings are in agreement with the hypothesis that the  
615 peripheries of the GRN (i.e. early regulatory input and differentiation gene batteries) are the least  
616 constrained and thus the most frequently changed (Davidson & Erwin 2006) part of a GRN, while  
617 the phylotypic stage (identified as the gastrula stage in echinoderms; (Raff 1996; Levin et al.  
618 2016)) is subject to strong evolutionary constraints. In this view our data support the idea that the  
619 regulatory states that define cell type identities, before differentiation, are the most evolutionary  
620 stable compared to early specification regulatory states. In the case of the developmental program  
621 for echinoderm skeleton, this likely corresponds to the transcription factors conserved in all four  
622 classes analysed here and known to be expressed in skeletal cells (Oliveri et al. 2008; Dylus et al.

2016; Czarkwiani et al. 2013). Indeed the high degree of conversation in all four classes is consistent with all echinoderms forming an adult skeleton by similar ossicle units: the stereom (Bottjer et al. 2006), and indicates that the GRN for adult skeletogenesis is a highly conserved feature. This is additionally supported by comparing expression patterns of several genes in juvenile or adult stages (Gao & Davidson 2008; Morino et al. 2012; Czarkwiani et al. 2013), which show a high degree of conservation in cells that participate in adult skeletogenesis. Additionally, in brittle star development most differentiation genes show an increasing trajectory over time consistent with their role in the final differentiation of the biomineral structure.

The modelling of developmental GRNs requires knowledge of spatial and temporal expression. For a GRN analysis comprising a few genes, the integration of such data is a relatively simple task. In a systems biology perspective, however, where hundreds or thousands of genes are considered simultaneously, it is easy to lose track of the important details of few or single genes, especially when working on novel systems with little to no access to the established data. Thus, we developed a website ([http://www.echinonet.eu/shiny/Amphiura\\_filiformis/](http://www.echinonet.eu/shiny/Amphiura_filiformis/)) using R-shiny that allows users to query different types of information, similar that implemented by Tu and collaborators in 2014 for *S. purpuratus* (Tu et al. 2014). Using the statistical programming language R as the backbone, our website provides a platform to easily query and find genes of interest. It gives access to annotations, expression levels, sequence information, differential screening and spatial expression patterns. Contigs can be queried by annotation, expression cluster id, contig id and additionally by the sea urchin functional classification. Thus, for example one can easily retrieve all transcription factors sequences and their expression temporarily and spatially (where available). Moreover, spatial expression data can be extended by simply adding a folder with the contig id and the individual pictures as JPEG files. The code is publicly available on github (<https://github.com/dvdyllus/Echinoderm-Web>). In future work, this website will be extended with data from regenerating arms produced in our laboratory and will thus create a unique resource to establish the brittle star *Amphiura filiformis* as developmental and regenerative model system.

651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678

## METHODS

### Experimental Techniques

#### *Embryological techniques*

*A. filiformis* cultures were set up as previously described (Dylus et al. 2016). At the desired stage, embryos were collected for RNA extraction and/or fixed for WMISH as described in (Dylus et al. 2016).

#### *Cloning and Probe synthesis*

All genes used for spatial expression analysis by WMISH were PCR amplified from *A. filiformis* cDNA and cloned in pGEM-T easy vector system (Promega) or Topo PCR cloning system (Invitrogen) accordingly to the manufacturer's instructions. Antisense probes labelled with DIG (Roche) were synthesized as previously described (Dylus et al. 2016). Primers are presented in Supplementary Table 5.

#### *Quantitative PCR (qPCR)*

qPCR was performed on different biological replicates to those used for the mRNA-seq, employing the procedures described previously (Dylus et al. 2016).

#### *Whole mount in situ hybridization (WMISH)*

Spatial expression of selected genes at mesenchyme blastula (24hpf) and (27 hpf) were characterised using WMISH as previously described (Dylus et al. 2016).

#### *RNA extraction*

For mRNA sequencing, embryo samples of a single male and single female culture were collected at 09hpf, 18hpf, 27hpf and 39hpf. At 27hpf three samples were collected. The RNA extraction was performed as previously described (Dylus et al. 2016). The quality of extraction and concentrations were checked using NanoDrop 2000 and Bioanalyser.

## 679 *mRNA sequencing*

680 Sequencing libraries were prepared using the TruSeq RNA library preparation protocol. The  
681 samples were sequenced with the Illumina v3 chemistry using the multiplex paired-end sequencing  
682 protocol. The sequencing was performed on an Illumina HiSEQ 2500 with 100bp paired-end reads.  
683 To reach optimal coverage we sequenced 2 lines multiplexing the 6 samples. Library preparation  
684 and sequencing were performed at the SickKids Hospital Toronto, Canada.

685

## 686 **Computational procedures**

687 If not otherwise stated, all computational work was performed on an Apple Mac OS X 10.6  
688 server with 24 cores and 64GB of memory.

689

## 690 *Assembly*

691 The assembly pipeline and annotation followed a set of unified protocols described in  
692 (Brown et al. 2013). The obtained reads were trimmed for adapters and for low quality sequences  
693 using Trimmomatic v0.27 (ILLUMINACLIP:Adapters.fasta:2:30:10; HEADCROP:12) (Bolger et al.  
694 2014). Quality filtering was performed using the FASTX-Toolkit (v0.0.13.2; fastq\_quality\_filter –  
695 Q33 –q 30 –p 50). The quality filtered and trimmed reads were then digitally normalized (Brown et  
696 al. 2012). Once all filtering was completed, reads from all stages were combined and the  
697 transcriptome was assembled using the Trinity package (v2013-02-25) (Grabherr et al. 2011).  
698 Partial and complete open reading frames (ORFs) with a minimum length of 100 amino acids were  
699 predicted using the TransDecoder (version rel16JAN2014) script. Bacterial contaminants were  
700 obtained using mpiBlast (v.1.6) (Darling et al. 2003) with e-value 1E-20 and crosschecked with hits  
701 obtained against UniProtKB -SwissProt with the same e-value. Searches with mpiBlast were run  
702 on the Legion HPC cluster at UCL on at least 40 cores. Sequences with higher similarity to the  
703 bacterial database were removed from the dataset. The cleaned ORF dataset represents the  
704 reference transcriptome (RefTr). All reads were deposited in the NCBI Short Read Archive (SRA)  
705 under accession numbers: SRR4436669 - SRR4436674.

706

## 707 *Preparation of other datasets*

708 Transcriptome sequence data from *Antedon mediterranea* was obtained by the Elphick lab  
709 at Queen Mary University of London, as reported previously (Elphick et al. 2015). To obtain a  
710 complete picture of coding sequences from *Patiria miniata*, we combined both genomic derived  
711 coding sequences and transcriptome sequences from <http://echinobase.org>.

712

## 713 *Quality Assessment*

714 Completeness of our transcriptome was estimated using CEGMA (v2.5) (Parra et al. 2007).  
715 Full-length distributions were estimated by considering all unique hits determined by BLASTx (1e-  
716 20) against the UniProtKB-SwissProt database and application of scripts included within the Trinity  
717 application.

718

## 719 *Annotation*

720 All BLAST (Altschul et al. 1990) searches were performed using a local NCBI-BLAST  
721 (v2.2.25) with e-value of 1e-6. The RefTr was annotated against the sea urchin *S. purpuratus*  
722 transcriptome sequences and against the UniProtKB-SwissProt database. One directional BLAST  
723 identified presumed homologs and reciprocal BLAST identified presumed orthologs. Gene  
724 Ontology classification was performed based on a previous sea urchin specific classification (Tu et  
725 al. 2012). For consistency purposes sequences obtained for the starfish *P. miniata*  
726 ([echinobase.org](http://echinobase.org)) and the crinoid *A. mediterranea* raw sequences (Elphick et al. 2015) were  
727 annotated using the same combination of one-directional and reciprocal BLAST (e-value 1e-6)  
728 against the sea urchin transcriptome database.

729

## 730 *Abundance estimation*

731 The quality filtered trimmed reads were re-aligned on the reference transcriptome using  
732 bowtie (v0.12.9) (Langmead 2010) with parameters set as in RSEM (Li & Dewey 2011). The bowtie  
733 output was loaded into CORSET in order to obtain counts for clusters of contigs that shared reads,

rather than individual contigs (Davidson & Oshlack 2014). This is equivalent to a potential “gene” count adding up all “isoform” counts. Normalization by internal standard was performed as follows: First, individual clusters were normalized by their peak of expression in the time-course data (9hpf, 18hpf, 27hpf and 39hpf); then, for each cluster the standard deviation was calculated and clusters with standard deviation below 0.01 were chosen as internal standard; and finally, an average of these clusters was used as normalization factor and each cluster was divided by this normalization factor and multiplied with 1,000,000. All downstream analysis was performed using customized R and bash scripts. In order to make statements about annotation content in the individual clusters, the most frequent annotations for each expression cluster were considered.

#### *Expression clustering of time-series data*

To sort expression clusters by their individual trajectories we applied the fuzzy clustering algorithm (Futschik & Carlisle 2005). We used 27 fuzzy clusters, based on the assumption that between 4 sampled time points the expression either increased, decreased or did not change giving  $3^3$  (27) possible paths for each trajectory. Note here the difference between a fuzzy cluster and an expression cluster: a fuzzy clusters describes a group of expression clusters that share similar trajectories over time. Since fuzzy clustering does not allocate each transcript always to the same cluster, we re-iterated this algorithm 100 times to find for each expression cluster the most probable fuzzy cluster membership.

#### *Estimation of phylogenetic trees*

Homologous sequences of Msp130 genes were selected from OMA output and used as input to build a HMM model using HMM 3.1 (<http://hmmer.org>, version 3.1b). Protein databases of 7 selected species were used to aggregate contigs with a conserved HMM domain. The determined contigs were filtered from redundant and small sequences with length below 100 amino acids. For the msp130 alignment specifically, additional sequences were obtained from *Ophiothrix spiculata* and *Lytechinus variegatus*. The sequences were aligned using PRANK (Löytynoja & Goldman 2008). The resulting alignment was then inspected using sea view and trees were

estimated using PhyML v3.1 (Guindon et al. 2009) and PhyloBayes MPI 1.6j (Lartillot et al. 2009).  
Topological differences are displayed using <http://phylo.io> (Robinson et al. 2016).

## AUTHORS' CONTRIBUTIONS

PO and DVD conceived the project, analyzed, and interpreted the data and wrote the manuscript.  
DVD acquired data, performed bioinformatics analyses and qPCR in *A. filiformis*. DVD, AC, and  
PO collected embryonic samples. DVD performed gene cloning and in situ hybridizations in *A.*  
*filiformis*. AC and PO performed Nanostring experiments. LMB and MRE contributed unpublished  
data from *A. mediterranea* and edited the manuscript. All the authors read and approved the  
manuscript.

## ACKNOWLEDGMENTS

The authors wish to Pok Wai (Prudence) Liu, Alun Jones, Luisana Carballo and Wendy Hart for  
experimental assistance. We wish to thank also the staff at the Sven Lovén Centre for Marine  
Sciences for their support.

## FUNDING

This work was supported by the EU FP7 Research Infrastructure Initiative ASSEMBLE (ref.  
227799), UCL System Biology, and KVA funds SL2015-0048 from the Royal Swedish Academy of  
Science. AC was supported by a Wellcome Trust PhD fellowship. LMB was supported by a  
studentship funded by Queen Mary University of London.

## REFERENCES

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–10. doi: 10.1016/S0022-2836(05)80360-2.
- Barsi JC, Tu Q, Davidson EH. 2014. General approach for in vivo recovery of cell type-specific effector gene sets. Genome Res. 24:860–868. doi: 10.1101/gr.167668.113.
- Biressi ACM et al. 2010. Wound healing and arm regeneration in *Ophioderma longicaudum* and *Amphiura filiformis* (Ophiuroidea, Echinodermata): Comparative morphogenesis and histogenesis. Zoomorphology. 129:1–19. doi: 10.1007/s00435-009-0095-7.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120. doi: 10.1093/bioinformatics/btu170.
- Bottjer DJ, Davidson EH, Peterson KJ, Cameron RA. 2006. Paleogenomics of echinoderms. Science (80-. ). 314:956–960. doi: 10.1126/science.1132310.
- Brown CT et al. 2013. khmer-protocols 0.8.4 documentation. doi: <http://dx.doi.org/10.6084/m9.figshare.878460>.
- Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH. 2012. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. arXiv. 1203.4802:1–18. doi: 10.1128/genomeA.00802-14.Copyright.
- Burns G et al. 2011. Dynamic gene expression profiles during arm regeneration in the brittle star *Amphiura filiformis*. J. Exp. Mar. Bio. Ecol. 407:315–322. doi: 10.1016/j.jembe.2011.06.032.
- Cameron RA, Samanta M, Yuan A, He D, Davidson E. 2009. SpBase: The sea urchin genome database and web site. Nucleic Acids Res. 37:750–754. doi: 10.1093/nar/gkn887.
- Cannon JT et al. 2014. Report Phylogenomic Resolution of the Hemichordate and Echinoderm Clade. Curr. Biol. 24:1–6. doi: 10.1016/j.cub.2014.10.016.
- Cheers MS, Etensohn CA. 2005. P16 is an essential regulator of skeletogenesis in the sea urchin embryo. Dev. Biol. 283:384–396. doi: 10.1016/j.ydbio.2005.02.037.
- Conesa A, Götz S. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int. J. Plant Genomics. 2008. doi: 10.1155/2008/619832.
- Czarkwiani A, Dylus D V, Oliveri P. 2013. Expression of skeletogenic genes during arm



811 regeneration in the brittle star *Amphiura filiformis*. *Gene Expr. Patterns*. 13:464–472.

812 Czarkwiani A, Ferrario C, Dylus DV, Sugni M, Oliveri P. 2016. Skeletal regeneration in the brittle

813 star *Amphiura filiformis*. *Front. Zool.* 13:18. doi: 10.1186/s12983-016-0149-x.

814 Darling AE, Carey L, Feng W. 2003. The Design, Implementation, and Evaluation of mpiBLAST.

815 *Proc. Clust.* 2003:13–15. doi: 10.1177/1740774509338234.

816 Davidson EH, Erwin DH. 2006. Gene regulatory networks and the evolution of animal body plans.

817 *Science* (80-. ). 311:796–800. doi: 10.1126/science.1113832.

818 Davidson NM, Oshlack A. 2014. Corset: enabling differential gene expression analysis for de novo

819 assembled transcriptomes. *Genome Biol.* 15:410. doi: 10.1186/s13059-014-0410-6.

820 Dupont S, Ortega-Martínez O, Thorndyke M. 2010. Impact of near-future ocean acidification on

821 echinoderms. *Ecotoxicology*. 19:449–462. doi: 10.1007/s10646-010-0463-6.

822 Dylus DV et al. 2016. Large-scale gene expression study in the ophiuroid *Amphiura filiformis*

823 provides insights into evolution of gene regulatory networks. *Evodevo*. 7:2. doi: 10.1186/s13227-

824 015-0039-x.

825 Elphick MR et al. 2015. Reconstructing SALMFamide Neuropeptide Precursor Evolution in the

826 Phylum Echinodermata: Ophiuroid and Crinoid Sequence Data Provide New Insights. *Front.*

827 *Endocrinol. (Lausanne)*. 6:1–10. doi: 10.3389/fendo.2015.00002.

828 Futschik ME, Carlisle B. 2005. Noise-robust soft clustering of gene expression time-course data. *J.*

829 *Bioinform. Comput. Biol.* 3:965–988. doi: 10.1142/S0219720005001375.

830 Gao F, Davidson EH. 2008. Transfer of a large gene regulatory apparatus to a new developmental

831 address in echinoid evolution. *Proc. Natl. Acad. Sci. U. S. A.* 105:6091–6096. doi:

832 10.1073/pnas.0801201105.

833 Geiss GK et al. 2008. Direct multiplexed measurement of gene expression with color-coded probe

834 pairs. *Nat. Biotechnol.* 26:317–25. doi: 10.1038/nbt1385.

835 Gilbert PUPA, Wilt FH. 2011. Molecular Aspects of Biomineralization of the Echinoderm

836 Endoskeleton. In: *Molecular Biomineralization: Aquatic Organisms Forming Extraordinary*

837 *Materials*. Müller, WEG, editor. Springer Berlin Heidelberg: Berlin, Heidelberg pp. 199–223. doi:

838 10.1007/978-3-642-21230-7\_7.

839 Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation  
840 sequencing technologies. *Nat Rev Genet.* 17:333–351. doi: 10.1038/nrg.2016.49.

841 Grabherr MG et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a  
842 reference genome. *Nat. Biotechnol.* 29:644–652. doi: 10.1038/nbt.1883.

843 Guindon S, Delsuc F, Dufayard JF, Gascuel O. 2009. Estimating maximum likelihood phylogenies  
844 with PhyML. *Methods Mol. Biol.* 537:113–137. doi: 10.1007/978-1-59745-251-9\_6.

845 Haas BJ et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity  
846 platform for reference generation and analysis. *Nat. Protoc.* 8:1494–512. doi:  
847 10.1038/nprot.2013.084.

848 Hall MR et al. 2017. The crown-of-thorns starfish genome as a guide for biocontrol of this coral reef  
849 pest. *Nature.* 544:231–234. doi: 10.1038/nature22033.

850 Hobert O. 2008. Regulatory logic of neuronal diversity: Terminal selector genes and selector  
851 motifs. *Proc. Natl. Acad. Sci.* 105:20067–20071. doi: 10.1073/pnas.0806070105.

852 Howard-Ashby M et al. 2006. High regulatory gene use in sea urchin embryogenesis: Implications  
853 for bilaterian development and evolution. *Dev. Biol.* 300:27–34. doi: 10.1016/j.ydbio.2006.10.016.

854 Israel JW et al. 2016. Comparative Developmental Transcriptomics Reveals Rewiring of a Highly  
855 Conserved Gene Regulatory Network during a Major Life History Switch in the Sea Urchin Genus  
856 *Heliocidaris*. *PLOS Biol.* 14:e1002391. doi: 10.1371/journal.pbio.1002391.

857 Jackson DJ et al. 2010. Parallel Evolution of Nacre Building Gene Sets in Molluscs. *Mol. Biol. Evol.*  
858 27:591–608. doi: 10.1093/molbev/msp278.

859 Janies DA et al. 2016. EchinoDB, an application for comparative transcriptomics of deeply-  
860 sampled clades of echinoderms. *BMC Bioinformatics.* 1–6. doi: 10.1186/s12859-016-0883-2.

861 Langmead B. 2010. Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinforma.* doi:  
862 10.1002/0471250953.bi1107s32.

863 Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: A Bayesian software package for  
864 phylogenetic reconstruction and molecular dating. *Bioinformatics.* 25:2286–2288. doi:  
865 10.1093/bioinformatics/btp368.

866 Levin M et al. 2016. The mid-developmental transition and the evolution of animal body plans.

867 Nature. doi: 10.1038/nature16994.

868 Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or  
869 without a reference genome. BMC Bioinformatics. 12:323. doi: 10.1186/1471-2105-12-323.

870 Livingston BT et al. 2006. A genome-wide analysis of biomineralization-related proteins in the sea  
871 urchin *Strongylocentrotus purpuratus*. Dev. Biol. 300:335–348. doi: 10.1016/j.ydbio.2006.07.047.

872 Lowe EK, Swalla BJ, Brown CT. 2014. Evaluating a lightweight transcriptome assembly pipeline on  
873 two closely related ascidian species. PeerJ Prepr. 2:e505v1. doi: 10.7287/peerj.preprints.505v1.

874 Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence  
875 alignment and evolutionary analysis. Science. 320:1632–1635. doi: 10.1126/science.1158395.

876 Luo Y-J et al. 2015. The Lingula genome provides insights into brachiopod evolution and the origin  
877 of phosphate biomineralization. Nat. Commun. 6:8301. doi: 10.1038/ncomms9301.

878 Mann K, Poustka AJ, Mann M. 2008. In-depth, high-accuracy proteomics of sea urchin tooth  
879 organic matrix. Proteome Sci. 6:33. doi: 10.1186/1477-5956-6-33.

880 Mann K, Wilt FH, Poustka AJ. 2010. Proteomic analysis of sea urchin (*Strongylocentrotus*  
881 *purpuratus*) spicule matrix. Proteome Sci. 8:33. doi: 10.1186/1477-5956-8-33.

882 Materna SC, Oliveri P. 2008. A protocol for unraveling gene regulatory networks. Nat. Protoc.  
883 3:1876–1887. doi: 10.1038/nprot.2008.187.

884 McCauley BS, Weideman EP, Hinman VF. 2010. A conserved gene regulatory network subcircuit  
885 drives different developmental fates in the vegetal pole of highly divergent echinoderm embryos.  
886 Dev. Biol. 340:200–208.

887 McCauley BS, Wright EP, Exner C, Kitazawa C, Hinman VF. 2012. Development of an embryonic  
888 skeletogenic mesenchyme lineage in a sea cucumber reveals the trajectory of change for the  
889 evolution of novel structures in echinoderms. Evodevo. 3:17. doi: 10.1186/2041-9139-3-17.

890 McIntyre DC, Lyons DC, Martik M, McClay DR. 2014. Branching out: origins of the sea urchin larval  
891 skeleton in development and evolution. Genesis. 52:173–85. doi: 10.1002/dvg.22756.

892 Moreton J, Dunham SP, Emes RD. 2014. A consensus approach to vertebrate de novo  
893 transcriptome assembly from RNA-seq data: assembly of the duck (*Anas platyrhynchos*)  
894 transcriptome. Front. Genet. 5:190. doi: 10.3389/fgene.2014.00190.

895 Morino Y et al. 2012. Heterochronic activation of VEGF signaling and the evolution of the skeleton  
896 in echinoderm pluteus larvae. *Evol. Dev.* 14:428–436. doi: 10.1111/j.1525-142X.2012.00563.x.

897 O'Hara TD, Hugall AF, Thuy B, Moussalli A. 2014. Phylogenomic resolution of the class  
898 ophiuroidea unlocks a global microfossil record. *Curr. Biol.* 24:1874–9. doi:  
899 10.1016/j.cub.2014.06.060.

900 Oliveri P, Tu Q, Davidson EH. 2008. Global regulatory logic for specification of an embryonic cell  
901 lineage. *Proc. Natl. Acad. Sci. U. S. A.* 105:5955–5962. doi: 10.1073/pnas.0711220105.

902 Parra G, Bradnam K, Korf I. 2007. CEGMA: A pipeline to accurately annotate core genes in  
903 eukaryotic genomes. *Bioinformatics.* 23:1061–1067. doi: 10.1093/bioinformatics/btm071.

904 Pisani D, Feuda R, Peterson KJ, Smith AB. 2012. Resolving phylogenetic signal from noise when  
905 divergence is rapid: A new look at the old problem of echinoderm class relationships. *Mol.*  
906 *Phylogenet. Evol.* 62:27–34. doi: 10.1016/j.ympev.2011.08.028.

907 Primus AE. 2005. Regional specification in the early embryo of the brittle star *Ophiopholis*  
908 *aculeata*. *Dev. Biol.* 283:294–309. doi: 10.1016/j.ydbio.2005.04.022.

909 Purushothaman S et al. 2014. Transcriptomic and proteomic analyses of *Amphiura filiformis* arm  
910 tissue-undergoing regeneration. *J. Proteomics.* 112:113–124. doi: 10.1016/j.jprot.2014.08.011.

911 Raff RA. 1996. *The shape of life: Genes, development, and the evolution of animal form.*

912 Rafiq K, Cheers MS, Ettensohn CA. 2012. The genomic regulatory control of skeletal  
913 morphogenesis in the sea urchin. *Development.* 139:579–590. doi: 10.1242/dev.073049.

914 Rafiq K, Shashikant T, McManus CJ, Ettensohn CA. 2014. Genome-wide analysis of the  
915 skeletogenic gene regulatory network of sea urchins. *Development.* 141:950–961. doi:  
916 10.1242/dev.105585.

917 Robinson O, Dylus D, Dessimoz C. 2016. Phylo.io: Interactive Viewing and Comparison of Large  
918 Phylogenetic Trees on the Web. *Mol. Biol. Evol.* .  
919 <http://mbe.oxfordjournals.org/content/early/2016/05/08/molbev.msw080.abstract>.

920 Roth ACJ, Gonnet GH, Dessimoz C. 2008. Algorithm of OMA for large-scale orthology inference.  
921 *BMC Bioinformatics.* 9:518. doi: 10.1186/1471-2105-10-220.

922 Sea Urchin Genome Consortium. 2006. The Genome of the Sea Urchin. *Science* (80-. ). 314:941–

923 952. doi: 10.1126/science.1133609.

924 Seaver RW, Livingston BT. 2015. Examination of the skeletal proteome of the brittle star

925 *Ophiocoma wendtii* reveals overall conservation of proteins but variation in spicule matrix proteins.

926 *Proteome Sci.* 13:1–12. doi: 10.1186/s12953-015-0064-7.

927 Szabó R, Ferrier DEK. 2015. Another biomineralising protostome with an msp130 gene and

928 conservation of msp130 gene structure across Bilateria. *Evol. Dev.* 17:195–197. doi:

929 10.1111/ede.12122.

930 Telford MJ et al. 2014. Phylogenomic analysis of echinoderm class relationships supports

931 Asterozoa. *Proc. R. Soc. B Biol. Sci.* 281:20140479–20140479. doi: 10.1098/rspb.2014.0479.

932 The ENCODE Consortium. 2012. An integrated encyclopedia of DNA elements in the human

933 genome. *Nature.* 489:57–74. doi: nature11247 [pii]n10.1038/nature11247.

934 Tu Q, Cameron RA, Davidson EH. 2014. Quantitative developmental transcriptomes of the sea

935 urchin *Strongylocentrotus purpuratus*. *Dev. Biol.* 385:160–167. doi: 10.1016/j.ydbio.2013.11.019.

936 Tu Q, Cameron RA, Worley KC, Gibbs R a., Davidson EH. 2012. Gene structure in the sea urchin

937 *Strongylocentrotus purpuratus* based on transcriptome analysis. *Genome Res.* 22:2079–2087. doi:

938 10.1101/gr.139170.112.

939 Vaughn R, Garnhardt N, Garey JR, Thomas WK, Livingston BT. 2012. Sequencing and analysis of

940 the gastrula transcriptome of the brittle star *Ophiocoma wendtii*. *Evodevo.* 3:19. doi: 10.1186/2041-

941 9139-3-19.

942 Vijay N, Poelstra JW, Künstner A, Wolf JBW. 2013. Challenges and strategies in transcriptome

943 assembly and differential gene expression quantification. A comprehensive in silico assessment of

944 RNA-seq experiments. *Mol. Ecol.* 22:620–634. doi: 10.1111/mec.12014.

945 Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev.*

946 *Genet.* 10:57–63. doi: 10.1038/nrg2484.

947

## TABLES

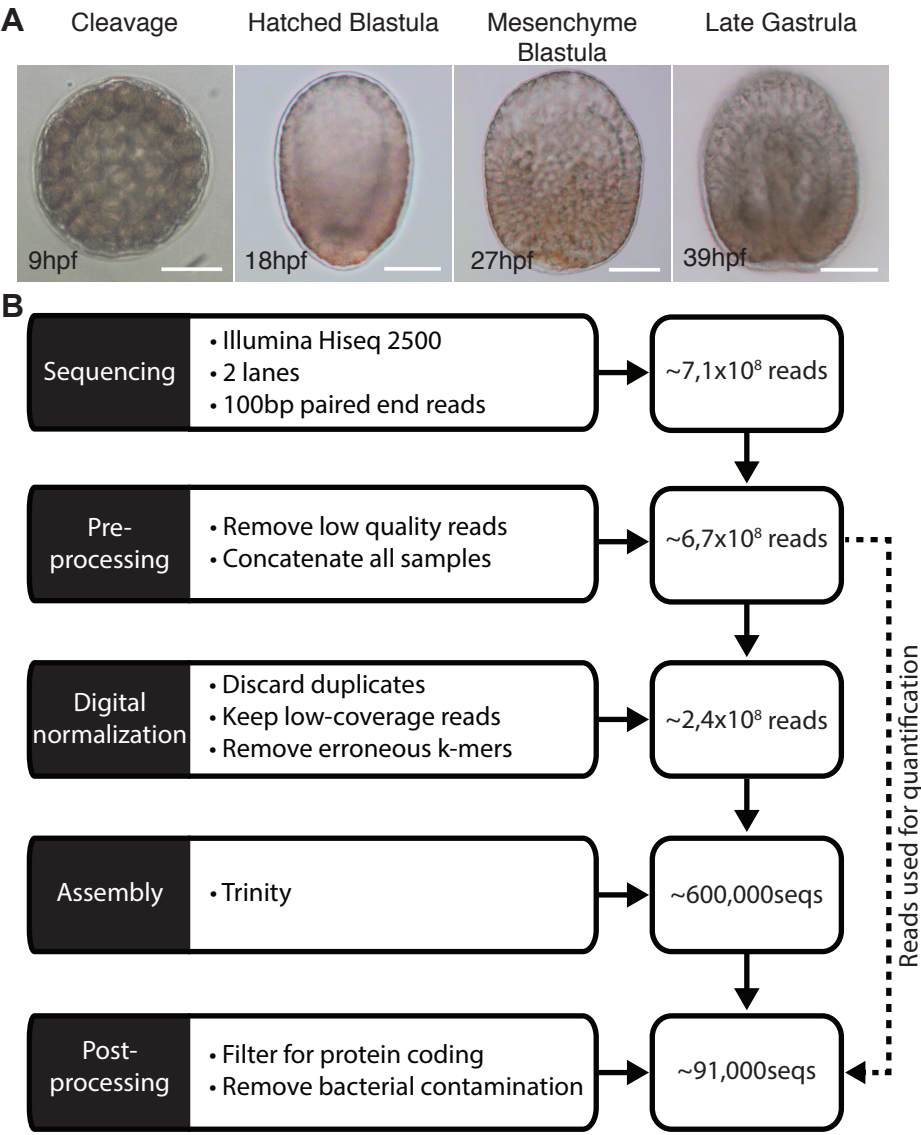
**Table 1.** Summary of quality statistics for the transcriptomic and genomic dataset used.

Species	N25	N50	N75	Longest	Mean	Median	Shortest	# Contigs
<i>Strongylocentrotus purpuratus</i>	6,297	4,108	2,438	22,850	2,821	2,139	400	29,072
<i>Amphiura filiformis</i>	2,601	1,410	639	18,993	925	525	297	91,311
<i>Patiria miniata</i>	1,131	474	351	23,898	524	369	63	263,867
<i>Antedon mediterranea</i>	1,508	443	186	36,836	302	156	100	607,455

**Table 2.** Homologs of sea urchin proteins in other echinoderms

Species	Source	E-value	Reciprocal BLAST	Single BLAST	# Unique Spu sequences	# Query Sequences
<i>A. filiformis</i>	Transcriptome	1E-06	9,779	41,492	13,656	91,311
<i>P. miniata</i>	Genome + Transcriptome	1E-06	10,208	77,576	16,211	263,867
<i>A. mediterranea</i>	Transcriptome	1E-06	9,164	26,997	12,982	607,454
<i>S. purpuratus</i>	Genome + Transcriptome	1E-06	26,395	2,675	26,475	29,072

957 **FIGURES**

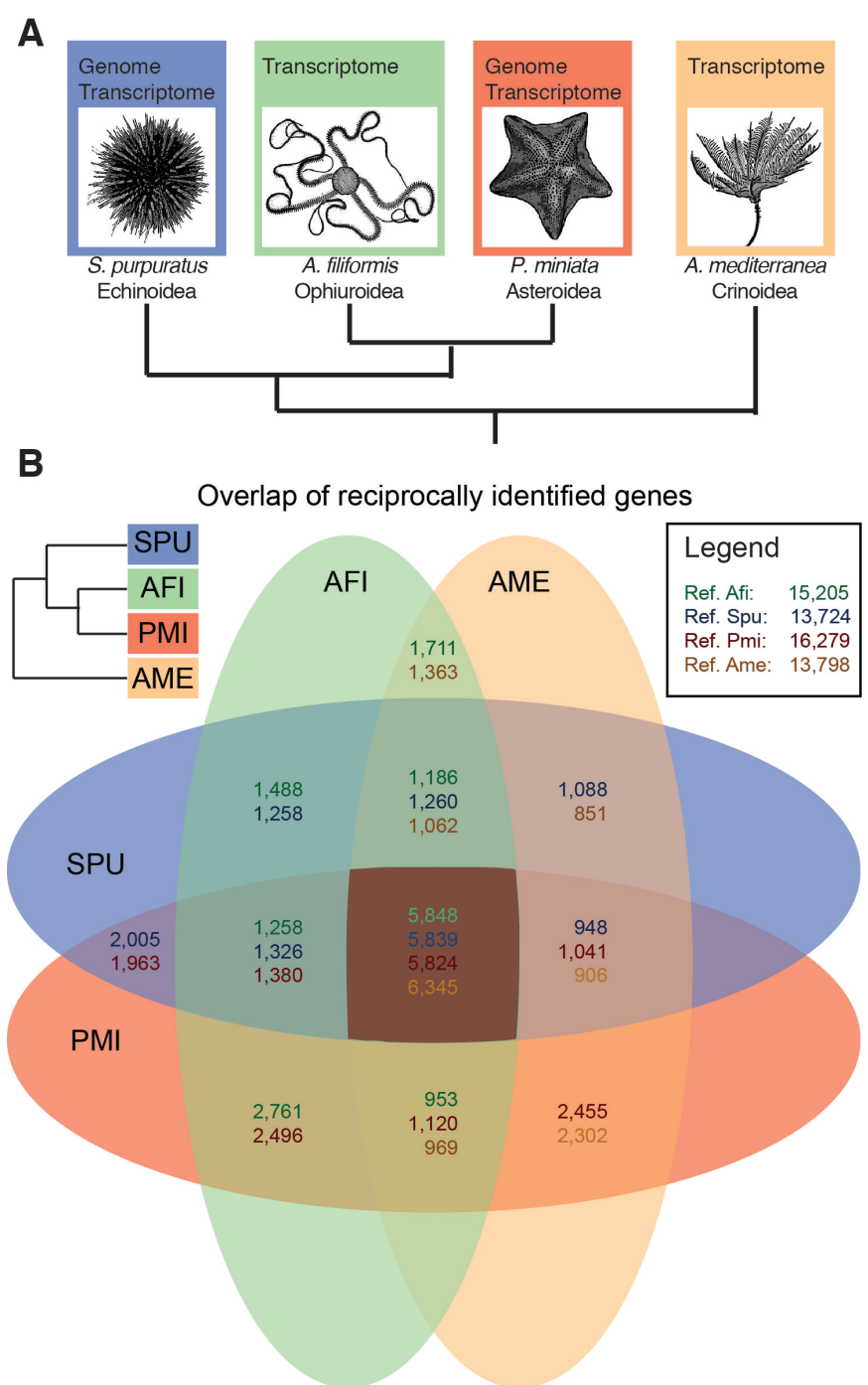


958  
959 **Figure 1. Pipeline used to obtain the *A. filiformis* developmental transcriptome. (A)**

960 Developmental timepoints used for RNA-seq: 9hpf corresponds to a late cleavage stage, 18hpf to  
961 a blastula stage, 27hpf to a mesenchyme blastula stage and 39hpf to a late gastrula stage. (B)  
962 Assembly pipeline showing the individual steps and the reduction in sequences.

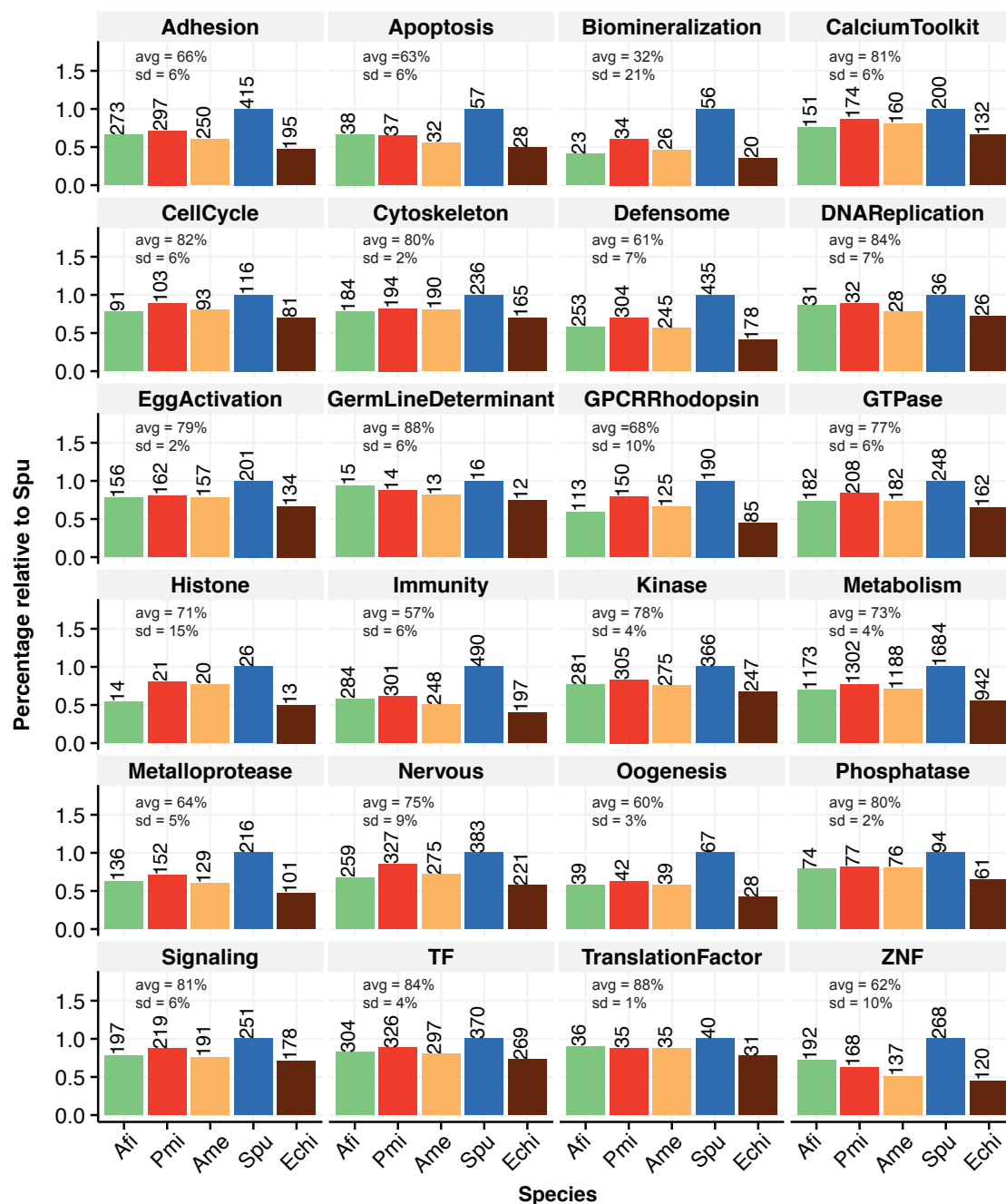
963



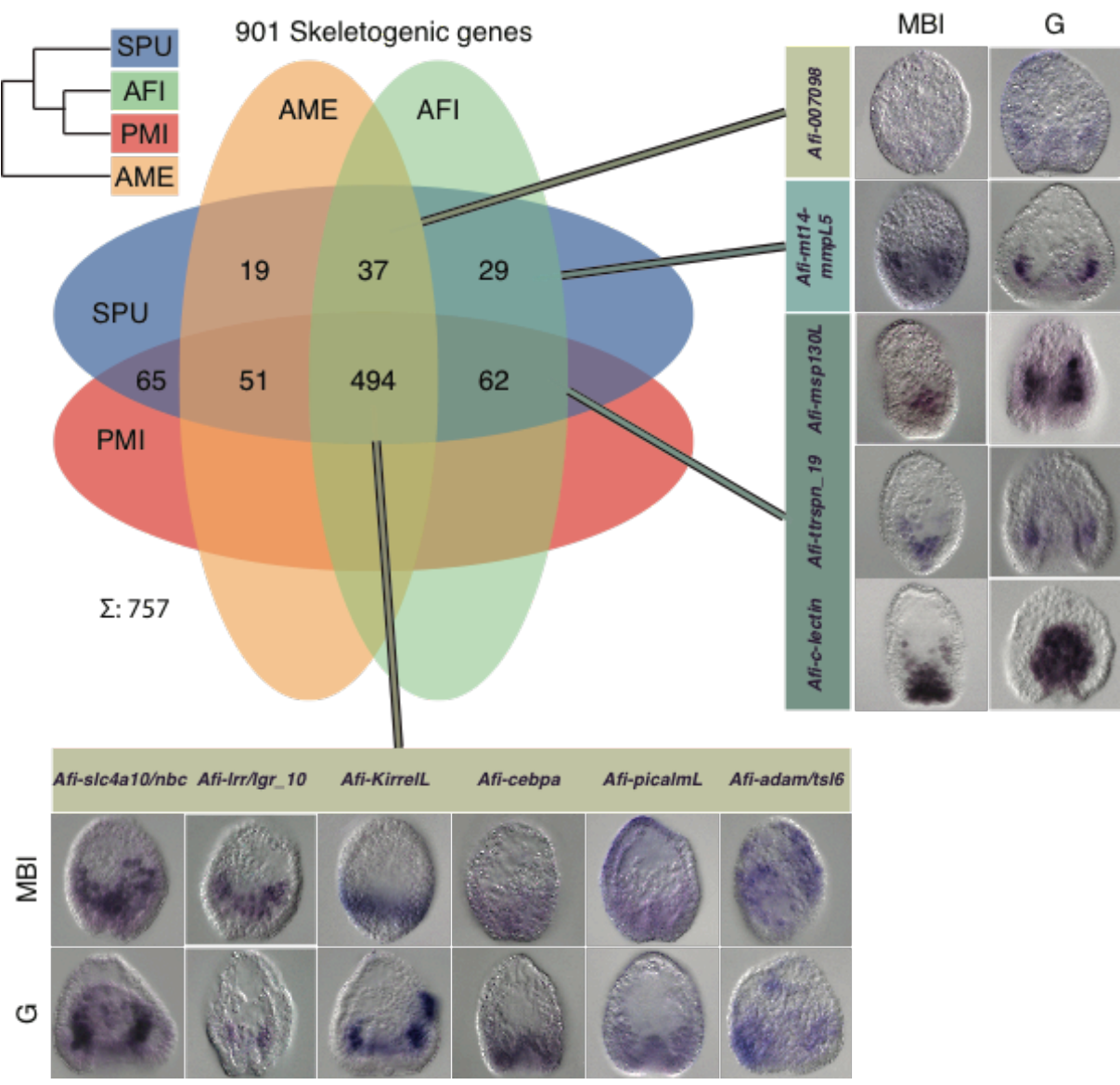


**Figure 2. Gene content in representatives of four echinoderm classes.** (A) Phylogenetic relationships of the four species compared in this study according to the currently most supported phylogeny for the classes these species belong to. (B) VENN-diagram showing the overlaps of genes that were identified using a reciprocal tBLASTx (e-value 1e-6) strategy. The different numbers in each overlap field indicate the species that was used as reference for the Blast search. Afi – *Amphiura filiformis*, Pmi – *Patiria miniata*, Ame – *Antedon mediterranea*, Spu – *Strongylocentrotus purpuratus*, Echi – Echinoderm core (overlap of all four classes).

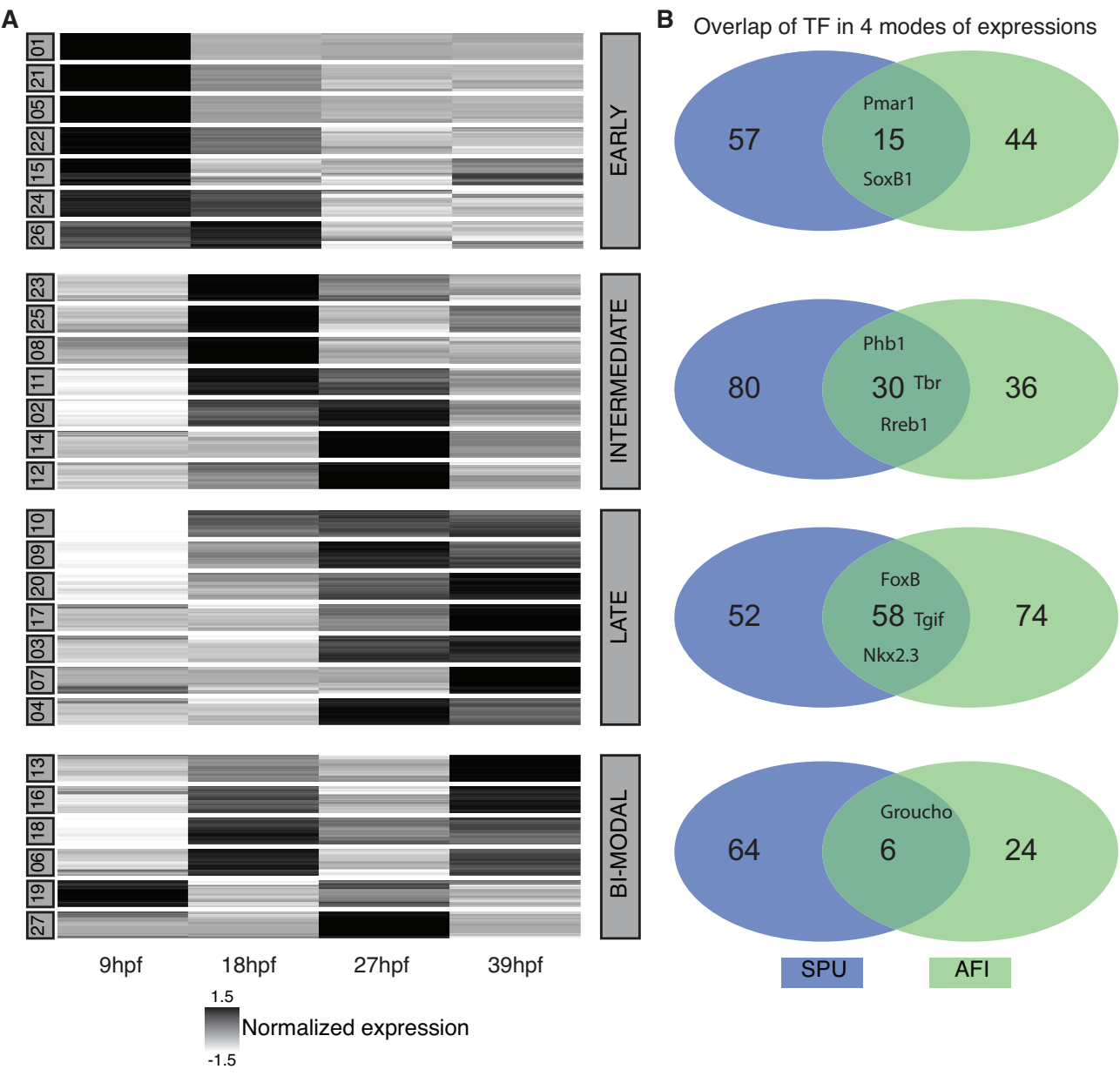




**Figure 3. Conservation of gene functional classes in echinoderms.** Sea urchin functional classes are based on *S. purpuratus* (Tu et al. 2012) and show proportions identified in the other 3 echinoderms. In the top right corner is depicted the currently most supported phylogeny for these 4 species. Average and standard deviation are calculated between Afi, Pmi and Ame and are normalized based on the sea urchin. Afi – *Amphiura filiformis*, Pmi – *Patiria miniata*, Ame – *Antedon mediterranea*, Spu – *Strongylocentrotus purpuratus*, Echi – Echinoderm core (overlap of all four classes).



**Figure 4. Homologs of sea urchin skeletogenic genes identified in other echinoderms and expression patterns for selected candidates.** Venn diagram showing the overlap of genes involved in sea urchin skeletogenesis with homologs found in other echinoderms. 494/901 are shared between 4 classes of echinoderms, which is a higher proportion than a set of random genes (SFig. 7). Whole mount in situ expression patterns in two important brittle star developmental stages for several selected candidates from different regions of overlap reveals an association with cells associated with skeleton formation. In the top right corner is depicted the currently most supported phylogeny for these 4 species. Afi – *Amphiura filiformis*, Pmi – *Patiria miniata*, Ame – *Antedon mediterranea*, Spu – *Strongylocentrotus purpuratus*, Echi – Echinoderm core (overlap of all four classes). MBI - mesenchyme blastula, G – gastrula.



**Figure 5. Global *A. filiformis* gene expression and comparison of larval regulatory states.** (A) Fuzzy clustering of 39,000 ECs in 27 clusters of four developmental time-points sorted in four distinct modes of expression (EARLY, LATE, INTERMEDIATE, BI-MODAL). Each line represents the expression of a single gene, and the grey intensity indicates the normalized expression (B) Comparison of TF in the 4 modes of expression between sea urchin (SPU) and brittle star (AFI). The majority of TFs show differences in expression.