Analysis of the *P. lividus* sea urchin genome highlights contrasting trends of genomic and regulatory evolution in deuterostomes

**Graphical abstract**

- Ancestral linkages are retained in sea urchins, but local gene order is evolving rapidly.
- Genes duplicated in sea urchins are expressed in novel structures.
- A new family of Pmar1-related genes is involved in specifying the micromere lineage.
- Gene-regulatory modules are conserved between sea urchin and chordates.

**Highlights**

**Authors**
Ferdinand Marlé taz, Arnaud Couloux, Julie Poulain, ..., Maria Ina Arnone, Christian Gache, Thierry Lepage

**Correspondence**
f.marletaz@ucl.ac.uk (F.M.), thierry.lepage@unice.fr (T.L.)

**In brief**
The chromosome-scale genome and characterization of the genomic elements that control gene expression in development reveal distinct trends of genome evolution in sea urchins and vertebrates as well as conserved gene-regulatory modules in deuterostomes.

Marlé taz et al., 2023, Cell Genomics 3, 100295
April 12, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.xgen.2023.100295
Analysis of the *P. lividus* sea urchin genome highlights contrasting trends of genomic and regulatory evolution in deuterostomes

Ferdinand Marléťat,1,23,* Arnaud Couloix,2 Julie Poulain,2 Karine Labadie,2 Corinne Da Silva,2 Sophie Mangenot,2 Benjamin Noel,2 Albert J. Poustka,4,5 Philippe Dru,6 Cinta Pegueroles,7 Marco Borra,8 Elijah K. Lowe,6 Guy Lhomond,6 Lydia Besnardeau,6 Stéphanie Le Gras,7 Tao Ye,9 Daria Gavriouchkina,10 Roberta Russo,11 Caterina Costa,11 Francesca Zito,11 Letizia Anello,11 Aldo Nicosa,11 Maria Antonietta Ragusa,11 Marta Pascual,7 M. Dolores Molina,13,14 Aline Chessel,11 Marta Di Carlo,15 Xavier Turon,16 Richard R. Copley,6 Jean-Yves Exposito,17 Pedro Martinez,13,18 Vincenzo Cavallini,12 Smadar Ben Tabou de Leon,19 Jennifer Croce,6 Paola Oliveri,1 Valeria Matranga,11 Maria Di Bernardo,20 Julia Morales,21 Patrick Cormier,21 Anne-Marie Genovèве,22 Jean Marc Aury,2 Valérie Barbe,2 Patrick Wincker,2 Maria Ina Arnone,8 Christian Gache,6 and Thierry Lepage14,*

1Center for Life’s Origin & Evolution, Department of Genetics, Evolution, & Environment, University College London, WC1 BBT London, UK
2Génomique Métabolique, Genoscope, Institut de Biologie François Jacob, Commissariat à l’Énergie Atomique, CNRS, Université Évry, Université Paris-Saclay, 91057 Évry, France
3Genoscope, Institut de Biologie François-Jacob, Commissariat à l’Énergie Atomique (CEA), Université Paris-Saclay, Évry, France
4Evolution and Development Group, Max-Planck-Institut für Molekulare Genetik, 14195 Berlin, Germany
5Dahlem Center for Genome Research and Medical Systems Biology (Environmental and Phylogenomics Group), 12489 Berlin, Germany
6Laboratoire de Biologie du Développement de Villefranche-sur-Mer (LBDV), Sorbonne Université, CNRS, 06230 Villefranche-sur-Mer, France
7Institute for Research on Biodiversity (IRBio), Department of Genetics, Microbiology, and Statistics, University of Barcelona, 08028 Barcelona, Spain
8Biologie et Evolution of Marine Organisms, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Napoli, Italy
9Plateforme GenomEast, IGBMC, CNRS UMR7104, INSERM U1258, Université de Strasbourg, 67404 Illkirch Cedex, France
10Molecular Genetics Unit, Okinawa Institute of Science and Technology, 904-0495 Onna-son, Japan
11Consiglio Nazionale delle Ricerche, Istituto per la Ricerca e l’Innovazione Biomedica (IRIB), 90146 Palermo, Italy
12Department of Biological, Chemical and Pharmaceutical Sciences and Technologies, University of Palermo, 90128 Palermo, Italy
13Department of Genética, Microbiologia i Estadística, Universitat de Barcelona, 08028 Barcelona, Spain
14Institut Biologie Valrose, Université Côte d’Azur, 06108 Nice Cedex 2, France
15Institute for Biomedical Research and Innovation (CNR), 90146 Palermo, Italy
16Department of Marine Ecology, Centre d’Estudis Avançats de Blanes (CEAB, CSIC), 17300 Blanes, Spain
17Laboratoire de Biologie Tissulaire et d’Ingénierie Thérapeutique (LBTI), UMR CNRS 5305, Institut de Biologie et Chimie des Protéines, Université Lyon 1, 69367 Lyon, France
18Institut Català de Recerca i Estudis Avançats (ICREA), 08028 Barcelona, Spain
19Department of Marine Biology, Charney School of Marine Sciences, University of Haifa, 31095 Haifa, Israel
20Consiglio Nazionale delle Ricerche, Istituto di Farmacologia Traslazionale, 90146 Palermo, Italy
21Integrative Biology of Marine Models (LBI2M), StationBiologique de Roscoff, CNRS, Sorbonne Université, 29680 Roscoff, France
22Sorbonne Université, CNRS, Biologie Intégrative des Organismes Marins, BIOM, 66650 Banyuls/Mer, France
23Lead contact
*Correspondence: f.marletaz@ucl.ac.uk (F.M.), thierry.lepage@unice.fr (T.L.)
https://doi.org/10.1016/j.xgen.2023.100295

**SUMMARY**

Sea urchins are emblematic models in developmental biology and display several characteristics that set them apart from other deuterostomes. To uncover the genomic cues that may underlie these specificities, we generated a chromosome-scale genome assembly for the sea urchin *Paracentrotus lividus* and an extensive gene expression and epigenetic profiles of its embryonic development. We found that, unlike vertebrates, sea urchins retained ancestral chromosomal linkages but underwent very fast intrachromosomal gene order mixing. We identified a burst of gene duplication in the echinoid lineage and showed that some of these expanded genes have been recruited in novel structures (water vascular system, Aristotle’s lantern, and skeletogenic micromere lineage). Finally, we identified gene-regulatory modules conserved between sea urchins and chordates. Our results suggest that gene-regulatory networks controlling development can be conserved despite extensive gene order rearrangement.
INTRODUCTION

Sea urchins are benthic marine animals that have attracted the curiosity of scientists since antiquity. They are one of the most amenable model systems in developmental biology because they produce abundant gametes and transparent embryos, and they have enabled major discoveries, such as the chromosomal nature of heredity, regulators of the cell cycle, and gene-regulatory networks controlling development. Beyond their original symmetry, echinoderms possess intriguing novelties, such as their calcium carbonate endoskeleton and their water vascular system running through five canals and associated tube feet. Among echinoderms, sea urchins also stand out with some remarkable embryological and adult morphological characteristics. During embryogenesis, they proceed with early specification of a micromere cell lineage that gives rise to the embryonic skeleton, a mechanism absent in other echinoderms. As adults, they also possess an intricate calcified masticatory organ: Aristotle’s lantern.

The first sequenced sea urchin genome, that of the Pacific sea urchin Strongylocentrotus purpuratus, revealed the extensive conservation of its gene repertoire with that of the human genome and served as a support for the elucidation of gene-regulatory networks at play during development. However, while comparative genomics has helped us understand functional differences across different vertebrate model species, the genomic differences between distinct echinoderms have not been thoroughly examined. A possible clue could come from the Hox genes, a classic locus linking genomes and body plan establishment. Because Hox genes are expressed during metamorphosis of echinoderm larvae, the pentaradial body plan of echinoderms has been classically related to the original observation of a rearranged Hox cluster in S. purpuratus, where this unique locus has been kept intact in most bilaterian animals. The subsequent discovery of an intact Hox cluster in the sea star and sea cucumber suggested, however, that this rearrangement is not the primary reason for the novel body plan of echinoderms. The sea urchin Lytechinus variegatus has revealed the same rearranged Hox cluster in this species as in S. purpuratus, which diverged ~60 mya. We profiled genome-wide chromatin accessibility throughout embryonic development. We investigated the expression and regulation of genes that originated at the origin of the echinoids clade. We identified a novel family of Pmar1-related transcriptional regulators specific to euechinoids, expressed in the micromere lineage and capable of converting any cell of the embryo into skeletogenic precursors. By integrating genomic and regulatory datasets, we demonstrated how regulatory changes could be associated with the origin of the novel body plan of urchins and other echinoderms.

RESULTS

The genome and developmental regulatory landscape of P. lividus

We sequenced the genome of a single male individual of P. lividus using 230× Illumina sequencing (Table S1). The genome was initially assembled using stringent parameters, and haploid copies were filtered out because the 2.9% polymorphism caused both haplotypes to be present in the initial contigs (Figure S1A). The assembly was then improved using long-read PacBio sequencing for gap closing, followed by scaffolding using proximity ligation (Chicago and Hi-C). The draft assembly of P. lividus amounts to 927.4 Mb with 5% of gaps and only 1.7% missing and 0.7% duplicated BUSCO genes (STAR Methods). Half of the assembly is comprised of 8 scaffolds larger than 41 Mb (N50), and there are 18 main scaffolds (>2 Mb) that represent distances in lineages such as cnidarians, molluscs, or other spiralian species with very limited fusions or rearrangements. Little is known, however, about the genomic rearrangements that took place in the sea urchin lineage and might underlie the evolution of sea urchin-specific novelties in embryogenesis and body plan.
88.3% of the assembly (Figure 1A; Table S2). These 18 scaffolds correspond to the chromosome number observed in cytogenetic studies (Figure 1B). Other sea urchin model species possess 21 chromosomes per haploid genome (S. purpuratus) and 19 (L. variegatus). While most of the repetitive fraction of vertebrate genomes is usually made of retrotransposons, DNA transposons...
are dominant in the sea urchin, similar to other deuterostomes, such as amphioxus24,25 (Figure S1B).

In addition, transcriptomes for 17 embryonic stages, from the fertilized egg to the pluteus stage, and 5 adult tissues were generated (Table S3). We annotated 32,041 genes, of which 80.1% have similarity to other species, and 92.1% are expressed (TPM > 1). We also annotated 5,087 high-confidence long non-coding RNAs (lncRNAs). By using gene expression clustering, we determined that 4,965 lncRNAs show stage-specific expression and, hence, play a potential role in the regulation of development (Figures S2C and S2D).

To gain further insight into their developmental regulatory program, we applied assay for transposase-accessible chromatin sequencing (ATAC-seq) to 6 stages from the 16-cell stage to the pluteus larva stage (Figure S3A; Table S4). Genomic regions with higher-than-background chromatin accessibility were labeled as open chromatin regions (OCRs) and classified depending on their location (Figure 1D) and their activity profile during development (Figure 1C). Among 64,701 OCRs distinct from transcription start sites, we found 40% “dynamic” regions (peaks) that were specific to one or two successive developmental stages and 57% constitutive regions present in three or more stages (Figures 1C and 1G). To further elucidate the regulatory activity in these regions, we applied Cut&Tag targeting the H3K27ac histone modification that is associated with active enhancers41,42 (Figures S3D, S3E, and S3G). Many OCRs also exhibit a level of transcription consistent with regulatory activity (Figure S3D).43 Our data efficiently recovered the activity of previously well-characterized regulatory elements (Table S5), such as, for example, the proximal and intronic enhancers of the nodal gene44 (Figure 1F). We compared the distribution of OCR locations with that of amphioxus and zebrafish, for which developmental ATAC-seq is available45 (Figures 1D and 1E). It has been proposed that OCRs that play an important regulatory role usually lie either in the vicinity of the promoter (proximal, <5 kb) or within introns (gene body).46 In vertebrates, a large fraction of these elements is located more distantly than in amphioxus, which has been associated with more complex gene-regulatory processes (Figure 1D).45 P. lividus shows a higher proportion of distal OCRs compared with the cephalochordate amphioxus and a lower proportion compared with zebrafish (Figures 1C and 1D).

We noticed that 16-cell-stage embryos (27,501 peaks) as well as pluteus-stage larvae (39,623 peaks) exhibit fewer OCRs than the other sampled stages. The blastula (137,938 peaks) and gastrula (147,188 peaks) stages show many stage-specific active elements, which are likely involved in the extensive gene-regulatory events that control embryo patterning and cell fate specification at these stages (Figure 1C; Table S5). In early 16-cell-stage and pluteus larva, the lower number of peaks could be interpreted either as a generally more relaxed chromatin state, potentially related to the absence of transcription before zygotic genome activation, or a more condensed state (Figure S3C). We hypothesize that the 16-cell stage possesses large open chromatin domains, as seen in other species,48 while the larval stage (pluteus) likely undergoes a reduction of open chromatin domains, as seen in other lineages, such as annelids47 (Figure 1C). Similar patterns are also recovered with an H3K27ac enrichment signal (Figure S3G). The sea urchin regulatory landscape therefore generally resembles the architecture observed in vertebrates or amphioxus in terms of enhancer location and developmental dynamics.48

### Opposite trends of genomic architectural changes in sea urchins and vertebrates

Many animal genomes retain chromosomal linkage over time as sets of genes stay localized on homologous chromosomes, even in distantly related species.49 Particularly, comparison of high-quality genomes of vertebrates, cephalochordates, and molluscs has revealed the existence of 17 ALGs ancestral to chordates.50 However, the occurrence of chromosomal fusions in some lineages suggests that more genomes and lineages should be taken into account to infer the ancestral bilaterian complement of ALGs. We examined the distribution of these ALGs in *P. lividus* as well as two other available sea urchin genomes (Figures 1A and 1B).51,52 Such representations pinpoint pairs of chromosomes or scaffolds that exhibit significant mutual ortholog enrichment (Fisher’s exact test, p < 0.05). Our comparisons indicate that some of the ALGs actually fused in the vertebrate and amphioxus lineages and suggest the existence of 23 ALGs for bilaterians, in agreement with recent studies53 (Figures S4A and S4B). Our comparisons of the sea urchin chromosomes further support the validity of these 23 ALGs as independent genomic units and reveal a missing linkage group that was not detected previously (dubbed “ALG R”), likely dispersed across multiple chromosomes in chordates, which indicates 24 ALGs (Figure 2). This ALG R is merged with ALG Q on *P. lividus* chromosome 4 (chr4) but remained intact in *S. purpuratus* and *L. variegatus* (Figures S4C and S4D). We found that 13 of 18 chromosomes of *P. lividus* descend from a single ALG, four are derived from the fusion of two ALGs, and one (chr2) receives contributions from three more (Figure 2).

Despite the conservation of ancestral linkages in *P. lividus*, the distribution of ALGs in the chromosomes of *L. variegatus* and *S. purpuratus* indicates some species-specific chromosomal fusions10,21 (Figure 2B). The fusions observed in *P. lividus* are absent in the two other species, but, conversely, some independent fusions took place, particularly in *L. variegatus*, where the two largest chromosomes (chr1 and chr2) derived from a mixture of distinct ALGs. Uniquely, ALG A1 and A2 are fused in *P. lividus* but not in the other sea urchin species, a fusion event that also took place in amphioxus.26,50 The higher number of chromosomes of *S. purpuratus* reflects the occurrence of only two fusions, the rest of the ALGs being represented as a single chromosome. In contrast, some chromosomal fusions are more ancient; two fusions are shared by all three urchin species: fusion of the three largest chromosomes (chr1 and chr2) derived from a mixture of distinct ALGs. Unlike vertebrates, sea urchins did not undergo a sizable rearrangement of units derived from ancestral linkages in most of their chromosomes, even when secondary fusions are observed (Figures 2C, S4C, and S4D). Incidentally, the distinct sea urchin lineages under scrutiny here have diverged since 48–68 mya, less than for human and mouse.50 In contrast to this echnoid chromosomal stability, we observed an extensive reshuffling of the microsynteny intrachromosomal gene order, which results in the absence of an observable “collinear” gene order visible as linear segments across pairs of homologous chromosomes.
as seen when comparing human and mouse genomes (Figures 3B and 3C). To quantify the rate at which gene collinearity is eroded, we compared the retention of microsynteny with the divergence time for selected sea urchin and vertebrate species (Figure 3A) and showed in this way that intrachromosomal gene order appears to evolve at a much slower pace in vertebrates than in sea urchins (Figure 3A).

Consequently, the gene order in the genomes of sea urchin appears to evolve following a trend distinct from that of vertebrates: the rate of interchromosomal rearrangement appears to be very low, while the rate of intrachromosomal gene order change appears to occur much faster. This distinct trend could be due to a relaxation of functional regulatory constraints on gene order compared with vertebrates.

The expression of recently evolved sea urchin genes
Exploration of gene content and gene expression can inform on how gene gains and duplications can play a role in organismal...
To this end, we performed gene family reconstruction (Table S6) and applied phylogenetic reconciliation to detect duplication events (Figure 4A). Sea urchins, like other echinoderms, do not show a particularly increased occurrence of gene losses in their genomes, such as that observed in the tunicate lineage (Figure 4A). Some losses, however, are shared by all three sea urchin species (Table S7); for instance, some members of the transforming growth factor β (TGF-β) signaling pathway, like BAMBI or BMP9, present in other echinoderms, like the sea cucumber.

Interestingly, we noticed a burst of gene duplication events in the echinoid ancestor with the second-highest number of duplicated gene families in deuterostomes after the origin of vertebrates and its whole-genome duplications. However, in the sea urchin, these duplicates are located in close genomic proximity, present on the same chromosomes (49%) and at close distances (36% closer to 100 kb), indicating a common origin by tandem duplication rather than by large-scale genomic duplications, as reported recently for cephalochordates. The gene duplicates show enrichment of Gene Ontology (GO) terms associated with membrane transport and the circulatory system, with, for instance, multiple ABC transporters and solute carriers encoding genes (Figure 4C), which possibly suggests that these genes could play a role in the functioning of the water vascular system (Figure 4C). In the most extreme cases, some gene families underwent large expansions in the sea urchin lineages, as detected by hypergeometric tests (p < 0.01; Table S8): the glycoprotein Kirrel, which plays a role in sea urchin skeletogenesis; the SLC16 transporter family, which is enriched during early embryonic stages or in adult structures, such as tube feet or the body wall (comprising the water vascular system and skeleton). We also found such increased expression of novel genes in Aristotle’s lantern, a calcified buccal apparatus specific to echinoids. We similarly observed preferred expression of genes duplicated in the echinoid lineage during the earliest embryonic stages and among the maternally expressed genes, suggesting that some of them could be involved in cell lineage specification (see below the example of pmar1) (Figure 4B). We then examined the putative regulatory elements (OCRs) that are associated with genes duplicated in echinoids; we identified 3.45 elements per gene (on average) for echinoid duplicates in contrast with 4.13 elements for genes arising in older duplications (Figure 4E). These elements are closer to promoters and less conserved in sequence than the elements associated with older duplicates or even the single-copy genes (Figure 4F). At the gene expression level, very recent duplicates show a higher organ specificity than other genes according to the tau estimate (Figure 4D), indicating that novel, fast-evolving regulatory elements are driving their expression in new organs and stages.

We assessed whether some transcription factor binding sites (TFBSs) are enriched in the OCRs associated with sea urchin gene duplicates and recovered significant enrichments for 22 TFBSs (hypergeometric test, p < 0.01; Table S9). This list comprises five nuclear receptors, including homologs of HNF4, thyroid receptor, and COUP-TF, as well as other factors known for controlling organogenesis (AP2, ELK, and ATF1), findings compatible with the organ-specific expression pattern recovered for a number of these duplicates (Figure 4D).

By studying the evolution, expression, and regulation of the sea urchin genes, we showed that newly evolved and newly duplicated genes are preferentially expressed in specific embryonic stages and anatomical structures. Some of these structures, such as Aristotle’s lantern or tube feet, are novelties of
Echinoids and echinoderms, respectively. This observation parallels similar findings regarding the younger transcriptome age of early developmental stages 62 or gene expression in the mollusc shell and radula, 63 but further studies will be necessary to explore their functional significance.

The origin of pmr/hbox12 genes

To understand how tandem gene duplications underlie the emergence of novelties, we investigated the origin of the pmr1 genes that play a critical role in specifying micromeres and their skeletogenic primary mesenchyme cell (PMC) derivatives, a specific character of euechinoids. 64 We identified 12 copies of genes related to the original pmr1/hbox12 sequence clustered within a 350-kb region in P. lividus (Figure 5A). 65 In L. variegatus, a similar number of pmr-related genes has recently been described in two distinct genome locations of chr2. 21 Moreover, eight copies of a pmr1-related gene have also been identified in the cidaroid Eucidaris tribuloides, the sister group of euechinoids. 66 pmr1 genes do not have clear orthologs in other echinoderms, but the identification of phb as a pmr1-related gene in brittle stars (Ophiuroidea) suggested that pmr1 genes are divergent copies of an ancient class of paired-type homeobox genes in this clade. 8,66 To clarify their respective relationships, we cataloged the multiple pmr copies found in sea urchins and copies of other pmr-related echinoderm genes, including the Pplx gene from brittle stars and the Phb gene from sea stars. 8 We constructed a phylogenetic tree that confirms the relationship of pmr and phb genes but, surprisingly, also shows that the aprd genes found in amphioxus and spiralians are related to these genes. 8,67 Our phylogeny suggests an independent expansion of pmr genes in multiple sea urchin lineages because the different pmr genes group into species-specific clades. Intriguingly, we identified a novel family of pmr-related genes that group in a distinct clade, showing a particularly fast rate of evolution, and dubbed them parent of pmr (pop) genes to reflect that they are sister genes of pmr (Figure S6E). The architecture of pmr/pop clusters appears distinct in all four sea urchins: in L. variegatus, pops and pmars are in distant loci, while in S. purpuratus, one locus contains 3 pmars and another 2 pmars and 2 pops. In E. tribuloides, eight pmr1-related genes are present in two different scaffolds, while all of these genes are closely located in P. lividus (Figure 5A). In the cidaroid E. tribuloides, only one set of pmr/pop-related genes is found, which could indicate that the duplication that gave rise to the ancestors of pop and pmr genes took place within the Euechinoidea lineage, followed by an independent expansion of pmr genes in the lineages leading to each sea urchin species.

Because pop genes seem to constitute a distinct clade of pmr1-related genes, we sought to determine whether their expression and function are the same as other pmr1 genes. In P. lividus, the pop1 gene is expressed maternally and ubiquitously, while the two other pop genes show an expression spatially
restricted to the micromere lineage in the late cleavage and early blastula stage, consistent with these genes being, like pmar1, involved in specification of the micromeres and PMC lineage (Figures S5D and S5E). Indeed, overexpression of these pop genes caused massive delamination of PMC-like mesenchymal cells that popped up from the vegetal pole at the onset of gastrulation and that was accompanied by ectopic expression of the PMC-specific marker genes delta and alx1 (Figure S5F). This phenotype is identical to the phenotype caused by overexpression of pmar1, reinforcing the idea that pop genes also act to specify the PMC lineage. These observations illustrate how pervasive pmar1 is identical to the phenotype caused by overexpression of pmar -related paired homeobox genes using the homeobox residues (IQTREE LG4X+R model).

Conservation and divergence of gene expression modules across deuterostomes

The dynamic gene order across sea urchin species contrasts with their highly conserved embryonic development and corresponding gene expression. To understand how the gene-regulatory program is affected by these gene order changes, we assessed the extent of the evolutionary conservation of sea urchin regulatory programs by comparing clusters of genes with similar temporal expression profiles in P. lividus, S. purpuratus, and the cephalochordate Branchiostoma lanceolatum. We found significant pairwise enrichments of genes belonging to identical gene families in clusters of co-expressed genes, either between sea urchins or when comparing sea urchins and chordate species (Figure 6A) and chordates (Figure 6B). This enrichment indicates conservation of genes involved in temporal gene expression modules, as observed previously within chordates and insects. Interestingly, the patterns of conservation at short and long divergence times (Figures 6A and 6B) show remarkable similarities, with most highly conserved modules being the stage-specific ones and the ones active during mid- to late development. We noticed that evolutionarily conserved modules can involve gene expression at the same developmental stage (homochrony) or shifted in their expression timing (heterochrony) at short and long evolutionary distances (Figures 6A and 6B). Despite the difference between the pluteus larva and the amphioxus, we observed several clusters that share a significant homologous gene set. For instance, the cluster (Pliv18) that shares the highest number of homologous genes appears to be enriched in GO terms associated with sensory perception and neuronal function (ion and amino acid transport) (Figure 6C). An examination of TFBSs enriched in OCRs related to genes belonging to these clusters revealed multiple factors associated with circadian regulation of gene expression, such as ARNT (BMAL) and CLOCK, which constitute the “core clock” associated regulators such as USF; downstream circadian effectors such as ATF and CREB; plus genes such as SREB, which is related to nutrient-related circadian adjustments (Figure 6D). These results suggest an evolutionarily conserved mechanism of circadian-based activation of neural activity at the end of embryonic development and the onset of larval life.

Beyond gene expression, we focused on the dynamics of the OCRs during development in sea urchins to determine key regulatory steps and how evolutionarily conserved they are. We identified conserved non-coding regions in P. lividus by performing an alignment of all three available sea urchin genomes and applying a statistical model to infer non-coding conservation (Phastcons). We found that 39% of OCRs with a putative regulatory role (non-repetitive, non-transcription start site [TSS]) overlapped with the 81,142 evolutionarily conserved regions that otherwise represent 1.94% of the genome (17.9 Mb) (Fisher’s exact test, p < 1e−9). We further assessed the sequence conservation of putative regulatory elements across development by assessing the evolutionary conservation of OCRs active at specific stages (Figure 1C). We identified the OCRs specific to the blastula, gastrula, and prism stages as the most conserved in sequence (Figure 7A).

We also compared the transcriptomic distance (Jensen-Shannon) between embryonic stages in sea urchin and deuterostome species to determine which embryonic stages exhibit the closest transcriptomic proximity (Figures 7B and S6A). The early blastula, gastrula, and prism stages that show the highest sequence conservation are also the ones that have the lowest transcriptomic divergence in the comparison of the two sea urchins (P. lividus and S. purpuratus) and the comparison with the cephalochordate amphioxus (Figure 7A). However, sea urchins appear to show a higher level of conservation at earlier stages (Figures 7B and S7A). This conservation can be explained by an earlier onset of major cell lineage specification at the early blastula stage.

Regulatory landmarks during sea urchin development

To determine which TFs control successive phases of sea urchin development, we employed a footprinting approach to detect the most likely occurrences of TF binding events in the ATAC-seq signal at successive stages (Figure 7). Then, by using reconstructed gene families, we identified 815 TFs in P. lividus and assigned 568 of them to TFBSs derived from the JASPAR database, which were later employed to analyze the footprint of the ATAC-seq signal (Figure 7D; Table S10) that we compared with their gene expression across developmental stages (Figure S7A). To ascertain the validity of the regulatory interactions deduced from our chromatin profiling data, we also

Figure 5. The evolution of pmar/hbox12 genes in echinoids

(A) Genomic organization of the pmar and pop loci in all three echinoid genomes.
(B) Regulatory landscape with RNA-seq (red) and ATAC-seq signal (blue) and OCRs in P. lividus.
(C) Phylogeny of pmar-related paired homeobox genes using the homeobox residues (IQTREE LG4X+R model).
(D) Expression of pmar and pop genes.
(E) In situ expression of pmar and pop genes. pop1 is expressed maternally and ubiquitously, while pop2 and pop3 are expressed in the PMC precursors. Scale bar. 30 μm.
(F) Phenotypes caused by overexpression of pmar or pop genes. Overexpression of pmar1, pop1, pop2, pop3, or pop2 fused to the repressor domain of Engrailed causes massive production of PMC-like mesenchymal cells and ectopic expression of PMC marker genes such as Delta and alx1. Inset: ventral view. VV, vegetal pole view; DIC, differential interference contrast.
reconstructed, for each stage, the global gene-regulatory network (GRN) by leveraging chromatin accessibility in OCRs, the assignment of OCRs to genes, and gene expression. The resulting networks represent genes as vertices and regulatory interactions as edges and can be interrogated to highlight a specific subnetwork. This analysis recovered, for instance, the endomesoderm specification GRN (Figure 7C) and the skeletogenic GRN (Figure S7E) described previously for S. purpuratus, validating the idea that the OCRs we identified are indeed cis-regulatory modules that bind key TFs of these GRNs.

While a small number of loci appear to be transcribed during cleavage stages (Figure S7A), we found that the onset of large-scale zygotic genome activation is taking place at the end of cleavage, in the early blastula stage, as pointed out previously (Figures 7D and S7B). Gene expression clustering (WGCNA) identifies small sets of genes specifically expressed in each of the cleavage stages, such as G-protein receptors (Oprx1, 16-cell), ionic channels (Scsα2, 60-cell), Toll-like receptors (2-cell), or homeodomain genes (Figure S6D). However, zygotic expression of TFs appears to only start at the 60-cell to late cleavage stages and gradually increases to mid-blastula stage (Figure S6E). Transcripts encoding TFs are present as maternal messages and remain detectable in later stages. At the 16-cell stage, most OCRs correspond to constitutively active and not stage-specific elements (Figure 1C), with a limited ATAC signal at TSS, confirming that zygotic expression is not yet taking place (Figure S3C). Accordingly, the TFBSs enriched at the 16-cell stage (Figures 7D and 7E) are also enriched in the early blastula stage, with few of corresponding TFs showing detectable expression at these stages (Figure S7A), which could indicate that some chromatin regions are present in a relatively open configuration before zygotic genome activation.

Conversely, the early blastula stage is characterized by the enrichment in TFBSs for maternally expressed TFs such as Ets4 and SoxB1 (Figures 7D and S7B). Interestingly, SoxB1 and Ets4 have been implicated as major determinants in regulation of the expression of early-expressed genes, such as those encoding the Hatching enzyme (He) and the SpAN/BP10 protease, whose transcripts start to accumulate in the ectoderm of the embryo at the 8- to 16-cell stages. Consistent with this, we detected Ets4 binding motifs in an OCR located in the vicinity of the He2 gene (Figure 7E). These factors are therefore excellent candidates for regulators involved in triggering onset of zygotic transcription. Additional factors of the homeobox family (OTX2,
Figure 7. *cis*-Regulatory landscape conservation during sea urchin development
(A) Sequence conservation scores in OCRs showing stage-specific activation (distinct from the TSS and consecutively expressed). Arrowheads indicate the stages’ highest non-coding conservation. Const., constitutive. Dev1 and Dev2 correspond to populations of OCRs that show broad activation domains.

(legend continued on next page)
PITX2, and MIX-1), nuclear receptor family (HNF4), and bZIP family (CREB1 and Myc) also have their binding sites enriched at the 16-cell and early blastula stages and may likely participate in activation of the zygotic genome (Figure 1D).

Later developmental stages showed increased TF motifs for developmental regulators expected to be active by mid-development, such as Sox, Otx, or Gsc factors. For instance, ATAC-seq footprints of SOXE, expressed in the left coelomic pouch that gives rise to the rudiment, and PITX2, expressed in the right coelomic pouch, are also enriched at the prism stage, consistent with establishment of the left-right asymmetry of the larva at this stage. Accessibility to FOX binding sites is restricted to the gastrula stage and subsequent stages. At the prism and then pluteus larva stage, we detected binding motifs for a set of circadian-controlled factors, such as CLOCK and ARNT (see above), HNF factor and bZIP factor, as well as CEBP-related factors. The pluteus larva stage similarly shows fewer OCRs (Figure 1) and limited enrichment of TFBSs. The most characteristic larva-specific factors are Ari5b, which plays a role in organ growth in vertebrates, and Fuk-1, which bears similarity to NFATC2, which is a putative role in immune system activation in vertebrates. The role of TFs at the successive developmental stages is confirmed by analysis of their influence in the GRNs reconstructed at each stage, as estimated by their prominence in the network architecture or centrality (Figure 7D).

Our examination of chromatin occupancy and TF binding informs on the mechanism of zygotic genome activation as well as on the subsequent sequence of regulatory events coordinating development, such as establishment of cell lineages, segregation of germ layers, and activation of circadian gene expression, and ultimately should help to reconstruct GRNs controlling developmental processes.

**DISCUSSION**

Sea urchins, as echinoderms, exhibit a unique and derived body plan with pentaradial symmetry, which has originally been related to a reshuffling of GRNs caused by an extensive gene order change. This reshuffling was exemplified by the original description of a rearranged Hox cluster in *S. purpuratus*. Here, we performed gene order comparison between the available sea urchin models and between sea urchin and amphioxus and demonstrated that the chromosomal architecture of sea urchins retained the ALGs inherited from a bilaterian ancestor.

These linkage groups were extensively duplicated after whole-genome duplication in vertebrates or completely lost in tunicates and nematodes. Interestingly, while chromosomal architecture is very conserved, the local intrachromosomal gene order appears to evolve much faster in sea urchins than in vertebrates (Figure 3).

While it could be tempting to link this observation with a faster evolution of gene regulation in sea urchins, most developmental and transcriptomics studies suggest conservation of GRNs and developmental mechanisms between sea urchin models, such as *S. purpuratus*, *L. variegatus*, and *P. lividus*. The relatively limited number of high-quality echinoderm and invertebrate genomes available at distinct evolutionary distances makes it difficult to determine whether the evolution gene order in sea urchins or in vertebrates constitutes the exception or the norm among bilaterians. This observation also suggests, more practically, that many loci will have a different organization and gene order between models such as *P. lividus* and *S. purpuratus*; for instance, the *pmar1* gene locus (Figure 5A). We also identified extensive echinoid-specific gene duplications and expansions and pinpointed that some novel echinoderm and echinoid structures, such as tube feet and Aristotle’s lantern, captured the expression of many of these recently evolved genes. A similar enrichment of new genes has, for instance, been witnessed with the mollusc radula or mollusc shell. Specialization of gene expression after duplication is similar to that observed for WGD duplicates in vertebrates or other lineage-specific duplicates in chordates.

We also explored the origin of the *pmar1* class of homeobox genes involved in cell lineage determination in an echinoid-specific fashion and identified a novel family of *pmar1*-like homeobox genes, the pop genes, that arose in euechinoids and that share with *pmar1* the ability to promote formation of the skeletogenic mesoderm. Remarkably, homedomain genes acting at early cleavage stages, such as Paired or TALE genes, have shown the propensity to undergo lineage-specific duplication in other animal lineages, such as mammals or spiralians, suggesting a recurrent phenomenon during animal evolution.

Interestingly, by using RNA sequencing (RNA-seq) to monitor gene expression and ATAC-seq to explore the open chromatin domains, we were able to perform evolutionary comparisons of gene expression profiles and gene-regulatory modules across species. Our data suggest possible long-range conservation of regulatory logic between sea urchins and chordates (Figure 6), which is consistent with the idea of deeply conserved regulatory mechanisms controlling animal development, although experimental testing of these regulatory mechanisms will be required to validate this idea. We showed that early development (cleavage) was the most divergent stage at the gene expression level, even between two sea urchin species that split 75 mya, possibly suggesting that the dynamics of zygotic genome activation might vary across species (Figure 7). Finally, we confirmed that tightly regulated and large-scale transcription of the zygotic genome did not occur before the early blastula stage in *P. lividus* (Figure 7).

In sum, we identified conserved principles of gene regulation across deuterostomes in terms of genomic location and of
activity of regulatory elements during development. We showed that these gene-regulatory modules are maintained despite the fast intrachromosomal gene order change in sea urchin genomes. The detected conservation of cis-regulatory modules complements previous observations comparing vertebrates and classic model systems like Drosophila or C. elegans and suggests that, despite their derived body plan, the presence of a larval stage and of mechanisms of early determination of embryonic cell lineages, echinoderms preserved such principles.

Limitations of the study
Our analysis of gene order in sea urchins is limited by the extent of available genomes. With more echinoderm and invertebrate genomes, we could generalize observed trends of intrachromosomal and interchromosomal gene order changes and determine which, of sea urchins or vertebrates, is representative of the ancestral state. We also think that functional work targeting some of the candidate genes proposed in this study to be expressed preferentially in urchin novelties would be important to appreciate their role; for instance, in the water vascular system or Aristotle’s lantern. Moreover, our characterization of the accessible and active genomic region is limited to the “bulk” of the genome and could be expanded across germ layers and cell lineages using approaches such as single-cell RNA-seq and ATAC-seq.

STAR METHODS
Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - DNA extraction and sequencing
  - Genome assembly
  - Annotation
  - Long non-coding RNAs
  - Synteny and gene family reconstruction
  - Gene expression analyses
  - ATAC-seq
  - Cut&Tag
  - In situ hybridization
  - Overexpression of mRNA
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION
Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100295.

ACKNOWLEDGMENTS
F.M. is supported by a Royal Society University research fellowship (URF/R1/191161) and a BBSRC research grant (BB/V01109X/). The authors acknowledge support from the Marine Genomics Network of Excellence (MGEN), the European Marine Biological Resource Centre (EMBRC), and the Coordinated Research Infrastructures Building Enduring Life-science (CORBEL). Proximity ligation Chicago and HIC were supported by Dovetail Genomics (Cantata Bio). Pacbio sequencing was funded by a grant from the Agence Nationale de la Recherche (ANR) to T.L. (project ANR-14-CE11-0006-01) and by a grant from CORBEL to M.I.A. E.K.L. was supported by a fellowship from EMBRC. We thank Hiroshi Wada and Atsuko Yamazaki for providing the pmar sequences of Eucidaris tribuloides. C.P. and X.T. are funded by the grant PID2020-118550RB (MarGeCh), funded by MCIN/AEI/10.13039/501100011033 (Spanish Government). We thank Alexandre de Mendoa for insightful comments. We also thank Carla Falugi, Sonia Manzo, and Maeva S. Kelly for support in the early phases of the project.

AUTHOR CONTRIBUTIONS

DECLARATION OF INTERESTS
The authors declare no competing interests.

Received: May 8, 2022
Revised: December 24, 2022
Accepted: March 6, 2023
Published: April 5, 2023

REFERENCES

Cell Genomics, 3, 100295, April 12, 2023 13


### KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical commercial assays</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TruSeq DNA library</td>
<td>Illumina</td>
<td>FC-121-2001</td>
</tr>
<tr>
<td>Nextera Mate-pair Kit</td>
<td>Illumina</td>
<td>FC-132-1001</td>
</tr>
<tr>
<td>TruSeq RNA Library</td>
<td>Illumina</td>
<td>RS-122-2001</td>
</tr>
<tr>
<td>TRIzol reagent</td>
<td>Invitrogen</td>
<td>15596026</td>
</tr>
<tr>
<td>pGEM-T Easy</td>
<td>Promega</td>
<td>A1360</td>
</tr>
<tr>
<td>DIG RNA Labeling Kit (SP6/T7)</td>
<td>Roche</td>
<td>11175025910</td>
</tr>
<tr>
<td>NBT/BCIP</td>
<td>Roche</td>
<td>11681451001</td>
</tr>
<tr>
<td>Anti-Digoxigenin-AP, Fab fragments</td>
<td>Roche</td>
<td>11093274910</td>
</tr>
<tr>
<td>mMessage mMachine kit</td>
<td>Invitrogen</td>
<td>AM1344</td>
</tr>
<tr>
<td>Nextera DNA Library prep</td>
<td>Illumina</td>
<td>FC-121-1030</td>
</tr>
<tr>
<td>Digitonin</td>
<td>Promega</td>
<td>G9441</td>
</tr>
<tr>
<td>NEBNex Ultra II Q5 Master Mix</td>
<td>NEB</td>
<td>M0544S</td>
</tr>
<tr>
<td>DNA Clean &amp; Concentrator Kit-5</td>
<td>Zymo</td>
<td>D4004</td>
</tr>
<tr>
<td>CUT&amp;Tag-IT Assay Kit</td>
<td>Activemotif</td>
<td>S3160</td>
</tr>
<tr>
<td>Histone H3K27ac antibody (pAb)</td>
<td>Activemotif</td>
<td>39134</td>
</tr>
<tr>
<td>SPRSelect</td>
<td>Beckman Coulter</td>
<td>B23318</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Software and algorithms</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Meraculous (v2.2.2.2)</td>
<td>Chapman et al.</td>
<td><a href="https://jgi.doe.gov/data-and-tools/software-tools/meraculous/">https://jgi.doe.gov/data-and-tools/software-tools/meraculous/</a></td>
</tr>
<tr>
<td>Haplomerger (v2, 20151124 build)</td>
<td>Huang et al.</td>
<td><a href="https://github.com/mapleforest/HaploMerger2">https://github.com/mapleforest/HaploMerger2</a></td>
</tr>
<tr>
<td>Busco (v5.1.3)</td>
<td>Simão et al.</td>
<td><a href="https://busco.ezlab.org/">https://busco.ezlab.org/</a></td>
</tr>
<tr>
<td>Trinity (2.11.0)</td>
<td>Grabherr et al.</td>
<td><a href="https://github.com/trinityrnaseq/trinityrnaseq">https://github.com/trinityrnaseq/trinityrnaseq</a></td>
</tr>
<tr>
<td>Portcullis</td>
<td>Mapleson et al.</td>
<td><a href="https://github.com/El-CoreBioinformatics/portcullis">https://github.com/El-CoreBioinformatics/portcullis</a></td>
</tr>
<tr>
<td>Mikado (v1.2.1)</td>
<td>Venturini et al.</td>
<td><a href="https://github.com/El-CoreBioinformatics/mikado">https://github.com/El-CoreBioinformatics/mikado</a></td>
</tr>
<tr>
<td>Augustus (v3.3.3)</td>
<td>Stanke et al.</td>
<td><a href="https://github.com/Gaius-Augustus/Augustus">https://github.com/Gaius-Augustus/Augustus</a></td>
</tr>
<tr>
<td>FEELnc</td>
<td>Wucher et al.</td>
<td><a href="https://github.com/tderrien/FEELnc">https://github.com/tderrien/FEELnc</a></td>
</tr>
<tr>
<td>MMSeqs2 (12-113e3)</td>
<td>Steinegger et al.</td>
<td><a href="https://github.com/soedinglab/MMseqs">https://github.com/soedinglab/MMseqs</a></td>
</tr>
<tr>
<td>Broccoli (v1.2)</td>
<td>Dereille et al.</td>
<td><a href="https://github.com/broccolijs/broccoli">https://github.com/broccolijs/broccoli</a></td>
</tr>
<tr>
<td>etc3</td>
<td>Huerta-Cepas et al.</td>
<td><a href="http://etetoolkit.org/">http://etetoolkit.org/</a></td>
</tr>
<tr>
<td>Msaprobgs (v0.9.7)</td>
<td>Liu et al.</td>
<td><a href="https://msaprobs.sourceforge.net/homepage.htm">https://msaprobs.sourceforge.net/homepage.htm</a></td>
</tr>
<tr>
<td>IQTREE (v2.1.1)</td>
<td>Nguyen et al.</td>
<td><a href="http://www.iqtree.org/">http://www.iqtree.org/</a></td>
</tr>
<tr>
<td>Generax (v1.2.2)</td>
<td>Morel et al.</td>
<td><a href="https://github.com/BenoitMorel/GeneRax">https://github.com/BenoitMorel/GeneRax</a></td>
</tr>
</tbody>
</table>

(Continued on next page)
RESOURCE AVAILABILITY

Lead contact
Further information and requests for resources and reagents should be directed to the lead contact, Ferdinand Marlé taz (f.marletaz@ucl.ac.uk).

Materials availability
Requests for clones and constructs of pmar and pop genes should be addressed to Thierry Lepage (Thierry.Lepage@unice.fr).

Data and code availability
The pop1, pop2 and pop3 transcripts have been deposited to NCBI under the accessions Genbank: ON325581, Genbank: ON325582, ON325583.

The genome (Genbank: GCA_940671915.1) and sequencing reads have been deposited to NCBI under the accession PRJEB25800.

The RNA-seq and ATAC-seq have also been deposited to NCBI Gene Expression Omnibus (GEO) under the accession GEO: GSE202034.

Other data files including genome assembly and annotation are available on zenodo (https://doi.org/10.5281/zenodo.7459274). Code underlying the analyses is available on github: https://github.com/paracentrotus/genome.

METHOD DETAILS

DNA extraction and sequencing
DNA was extracted from sperm from a single male individual collected in Naples (Italy). Sperm cells were concentrated by centrifugation, washed repeatedly, and subsequently embedded in 2% low melting agarose. Sperm cells were lysed in a solution of 1% SDS, 10mM Tris (pH 8) and 100mM EDTA and then resuspended in a solution of 0.2% N-laurylsarcosine, 2mM Tris (pH 9) and 0.13 mM EDTA. DNA was released from the agarose blocks using agarase.
Transcripts assembled with Trinity (see below) were aligned to the genome using GMAP (version of 2018-03-25). These alignments and scaffolder. The final assembly shows the following BUSCO statistics (v5.1.3) when using the Metazoa gene set: C:94.7% reads in 2 library preparation kit (New England Biolabs, Ipswitch). Chicago and HiC libraries were sequenced for 476M and 210M paired-end from the beads and the DNA fragments were purified again on SPRI beads. Sequencing library was constructed using the NEB Ultra. After end-labelling, proximity ligation was carried out using T4 DNA ligase and cross-linking reversed using Proteinase K, removed bler relying on a OLC (overlap-layout-consens) design. Resulting error-prone assembled sequences are used to fill gaps and extend scaffolds. We used the parameters '-minMatch 8 -sdpTupleSize 8 -minPctIdentity 75 -bestn 1 -nCandidates 10 -maxScore/C0.500'. The resulting patched assembly has a reduced fraction of gaps (4.9%) and incorporates 1,475Mb of assembled DNA for a 1,551Mb total size (Table S2). To fill the gaps, and improve the Illumina assembly, we performed gap-filling and local reassembly using the PBJelly tool (v15.8.24). PBJelly aligns the Pacbio reads to the assembly using the Blasr aligner designed to account for underlying read error profile, then collects reads surrounding and spanning gaps and scaffold extremity regions and assemble them with the ALLORA assembler relying on a OLC (overlap-layout-consensus) design. Resulting error-prone assembled sequences are used to fill gaps and extend scaffolds. We used the parameters '-minMatch 8 -sdpTupleSize 8 -minPctIdentity 75 -bestn 1 -nCandidates 10 -maxScore/C0.500' for blasr alignment. The resulting patched assembly has a reduced fraction of gaps (4.9%) and incorporates 1,475Mb of assembled DNA for a 1,551Mb total size (Table S2).

To generate a haploid reference genome from our diploid assembly, we employed the Haplomerger (v2, 20151124 build), which relies on a graph of reciprocal LASTZ alignments to extract the best path across haplotype scaffolds. Before reciprocal alignments, we masked repetitive regions in the diploid assembly with RepeatMasker using a custom repeat library (see below) and a custom scoring matrix obtained with the script lastz_D_Wrapper.pl. We then applied module A aimed at detecting and splitting mis-assembled scaffolds, and module B aimed at performing haplotype reconciliation. Finally, residual haplotype sequences smaller to be processed by module A (<5000bp) were screened using module G relying on residual similarity to resolved haplotigs.

To further extend the contiguity of our haploid reference, we used long-range contact information from Chicago and HiC library prepared by Dovetail Genomics (Santa Cruz, USA). The Chicago method relies on the reconstitution of synthetic chromatin in controlled condition followed by chromatin conformation capture and evenly distributed contact information in the 150 kb range. Library preparation is described in detail in. Alternatively, HiC chromatin conformation capture provides contact information at a broader range, and particularly helps reconstruct chromosomes which usually do not present much interactions in trans. Tissue was crosslinked in 1% PFA, and chromatin subsequently extracted, immobilised on SPRI beads, washed and digested with DpnII. After end-labelling, proximity ligation was carried out using T4 DNA ligase and cross-linking reversed using Proteinase K, removed from the beads and the DNA fragments were purified again on SPRI beads. Sequencing library was constructed using the NEB Ultra library preparation kit (New England Biolabs, Ipswitch). Chicago and HiC libraries were sequenced for 476M and 210M paired-end reads in 2 x 150bp mode on a HiSeq4000 instrument. Chicago and HiC data were processed through two distinct runs of the HiRise scaffold. The final assembly shows the following BUSCO statistics (v5.1.3) when using the Metazoa gene set: C:94.7% [S:94.0%,D:0.7%];F:3.6%,M:1.7%,m:954.

Annotation
Transcripts assembled with Trinity (see below) were aligned to the genome using GMAP (version of 2018-03-25). These alignments and the merged stringtie assemblies were leveraged using Mikado (v1.2.1) to generate a high-quality reference transcriptome. A set of curated splice-junctions generated from RNA-seq alignments using Portcullis (v1.0.2) was also provided to Mikado. Putative fusion transcripts were detected by Blast comparison against Swissprot and ORFs were annotated using Trans-decoder (Haas et al. 2008). Transcript scripts derived from the reference transcriptome were selected to train the Augustus de novo gene prediction tool (Stanke et al. 2006). Exon and intron positions derived from the mikado consensus transcriptome were converted into hints for Augustus gene prediction. We annotated repetitive regions in the genome by constructing a repeat library using RepeatModeler (v1.0.11) that was subsequently used for masking with RepeatMasker (v4.0.7). Repeat landscape was subsequently inferred by computing Kimura 2-parameters distances with the consensus for each repeat category. Gene models with half or more of their exons overlapping at 50% with repeats were discarded, yielding 41717 filtered gene models. Alternative transcripts and UTRs were subsequently incorporated using the PASA pipeline (Haas et al. 2008). These gene models contain a total number of 4915 distinct PFAM domains.

Long non-coding RNAs
For lncRNA annotation, we used all RNA-seq data from P. lividus available in SRA archive (accessions: PRJEB10269, PRJNA392084, PRJNA376650, PRJNA288758, PRJNA264358, 4787.4 M reads in total, 14.7 M reads per sample). After quality control with FastQC v0.11.6 and trimming with Trimmomatic v0.39 (LEADING:20 TRAILING:20 MINLEN:25 ILLUMINACLIP:adapters.fa:2:30:10
SLIDINGWINDOW:10:20,\textsuperscript{120} we mapped the reads to the reference genome using HISAT2 v2.2.1\textsuperscript{121} and obtained a transcriptome assembly for each sample using StringTie v2.1.4\textsuperscript{122} by providing our gene models as a reference annotation file. We obtained 175 GTF files corresponding to the individual assemblies which were merged to obtain a reference GTF using the merge option from StringTie. The merged GTF was used to identify the candidate IncRNAs using FEELnc software\textsuperscript{99} in a three steps pipeline: i) filter out transcripts shorter than 200nt, monoexonic and overlapping protein-coding exons, ii) discard transcripts with coding-potential and iii) classify the candidate IncRNAs according to their relative position in the genome. Given that there is no previous knowledge of IncRNAs in \textit{P. lividus}, we used two different strategies to compute the coding-potential: i) we took a set of mRNAs and shuffled them while preserving 7-mer frequencies (shuffle approach), and ii) we provided as training set IncRNAs from a closely related species (\textit{S. purpuratus}, reference approach). The number of candidate IncRNA transcripts obtained was 32,147 and 32,107 genes for the shuffle and reference approaches, respectively, with more than 99% overlap. After filtering the transcripts uniquely annotated by a single approach, we obtained a list of 32,001 candidate IncRNAs genes (56,259 transcripts). Subsequently, we estimated the abundance of the candidate IncRNAs in each SRA sample using the feature Counts function of the Rsubread package from Bioconductor\textsuperscript{123} and we classified genes according to their expression range using the filter ByExpr function from EdgeR package\textsuperscript{123} for R by setting the min.count parameter to 10. This allowed us to identify a high confidence IncRNAs set that includes genes that are expressed in most of the samples, consisting of 5,087 IncRNA genes.

**Synteny and gene family reconstruction**

To compute pairwise synteny comparisons, we used mutual-best-hits based on MMSeqs2 (MMSeqs2/12-113e3) comparisons between proteomes after selecting for the longest protein for each locus. After reindexing of gene coordinates, we used Fisher’s exact test to determine mutual enrichment of orthologues between chromosomes or scaffolds. To estimate the relationship between gene order and divergence time, we estimated as the fraction of orthologues located in blocks of 2 or more consecutive genes in the same order with no more than one interspersed gene (Figure 3). Corresponding python and R codes upload to https://github.com/paracentrotus/urchinpaper.

We used Broccoli (v1.2) for gene family comparisons using the species specified in Table S5\textsuperscript{101} and inferred clade specific gains and losses by comparing the content of each family with that of a reference species tree using the ete3 library.\textsuperscript{102} To assess duplication within gene families, families with less than 500 genes and at least 5 genes in 3 species were subjected to phylogenetic reconstruction: after alignment with Masprops (v0.9.7)\textsuperscript{103} and alignment trimming using clipkit (v0.1, option -m gappy),\textsuperscript{124} a tree was re-constructed using IQTREE (v2.1.1) assuming a LG4X + R model. Then, Generax (v1.2.2) was used to perform genes and species tree reconciliation and to detect duplication events in a maximum likelihood framework.\textsuperscript{105}

**Gene expression analyses**

We extracted RNA for successive embryonic stages and several organs (Table S3) using Trizol reagent (Invitrogen). Strand-specific RNA-seq libraries were prepared using the TruSeq RNA Library (Illumina) and sequenced in a 2x150bp layout with an average of 64.6M reads per sample. Reads were aligned to the genome using STAR (v2.5.2b) at an average rate of 65.64% of unique mapping.\textsuperscript{125} A transcriptome was assembled for each sample using Stringtie (v1.3.3b)\textsuperscript{108} and sample-specific transcriptome assemblies were merged using Taco.\textsuperscript{109} The reads from all the samples were also assembled de novo using Trinity.\textsuperscript{94} Gene expression was quantified from reads aligned using STAR (v2.5.2b) and using featureCounts from the Subread package (v1.6.3)\textsuperscript{119} and counts converted in FPKM. We also evaluate the coverage of OCRs for both strands using featureCounts to evaluation their level of transcription on both strands.

We used WGCNA (v1.7.0)\textsuperscript{126} clustering for the full set of embryonic stages and organs. After filtering out genes with limited variance and counts, then, the ‘softpower’ parameters were estimated and set at 13, and clustering was run with a ‘signed’ network type.\textsuperscript{95} For all clustering analyses, FPKM were calculated and replicates if available were merged to obtain a single gene expression value. Then, we used mfuzz (v2.54.0)\textsuperscript{127} clustering to compare subsets of 8 embryonic stages in sea urchin, amphioxus and zebrafish using datasets for the two later as described in.\textsuperscript{25} After filtering genes with low expression or limited variability, expression values were normalised, the fuzzifier parameters (m) was estimated and the optimal number of clusters was determined by computing with minimal distance between cluster centroids (Dmin) for various numbers of clusters. To compare the evolutionary conservation of gene expression modules, a hypergeometric test was performed on the number of genes belonging to distinct gene families shared between each pair of alternate gene expression modules (accounting for possible many-to-many paralogy relationships between individual genes). Gene ontology enrichments were computed using the topGO package using terms transferring from PFAM annotation and Swissprot best hits (evalue 1e-10).\textsuperscript{128} Comparisons of transcriptomic distances across embryonic stages and were computing as Jensen-Shannon divergence and single-copy orthologues inferred using Broccoli as performed in.\textsuperscript{25} \textit{S. purpuratus} data analyzed corresponds to PRJNA81157 and \textit{Branchiostoma floridae} to PRJNA416866.

**ATAC-seq**

ATAC-seq was performed following the Omni-ATAC protocol using Digitonin (Promega) in addition to NP40 and Tween in the cell lysis buffer (Corces et al., 2017). Sea urchin eggs were fertilised in filtered sea water with 2 mM of Paraminobenzoic acid (Sigma) to prevent hardening of the fertilisation envelope. The fertilization envelope was then removed by repeated filtration on a nylon net (70 μM). Embryos at the appropriate stage were collected and washed in NaCl 0.55M twice then dissociated in Calcium-Magnesium artificial sea water by
energetic pipetting. For each time point, different numbers of nuclei were tested, typically 50,000, 150,000 and 250,000 and at least two biological replicates generated Table S3). The integrity of the nuclei was checked on a microscope after staining the nuclei with Hoechst. After tagmentation the libraries were purified with a Zymo DNA clean up kit then a qPCR was performed on an aliquot of the eluted libraries using the Ad1 and Ad2.x primers and the corresponding Ct was determined for each sample.126 The libraries were then amplified at Ct+ 2 cycles. The quality of the libraries was checked on a 2% agarose gel to verify the size of the fragments and the nucleosome phasing.

Reads obtained for each library were mapped using Bowtie2 (v2.4.1) with the parameters ‘-v very-sensitive’ and ‘-k 10’127 and peak calling for open chromatin region was performed using Generich (v0.6) available at https://github.com/jsh58/Generich using ATAC model (-j), keeping unpaired alignments (-y), removing PCR duplicates (-r) and excluding reads mapped to mitochondria (-e MT). At this step, we pooled biological and technical replicates together as Generich analyses each replicate separately and then combined them by summarising p values using Fisher’s method.

We then used pybed tools to generate a set of unified peaks from the peaks called at each stage, classify peaks according to the activity during development and perform intersection with masked repetitive regions and conserved regions, as well as assigning peaks to genes (see jupyter notebook). Unified OCRs were then classified as open at a given stage based on the intersection with peaks called at each stage (Figure 1C). For comparative coverage analyses (ATAC-seq, Cut&Tag, eRNA, Figure S3D), a randomised set of genome intervals of the same size distribution as the predicted OCRs was generated using bedtools shuffle.128

Conserved regions were calculated using Phastcons with the parameters ‘-target-coverage 0.25 –expected-length 12 –rho 0.4’ from a multi-alignment of 3 echinoid genomes performed with Roast (ref) that started with Lastz alignment to P. lividus species using parameters ‘--inner = 2000 –ydrop = 3400 –gapped thresh = 6000 –hspthresh = 2200’ and the ‘HoxD55’ substitution matrix.

Footprinting analysis was performed using TOBIAS (v0.12.10) and the JASPAR binding motifs using the ‘-time-series’ parameter on the scored footprint bigwig files at successive stages, after correction yielding pairwise (Figure S8D) and overall (Figure 5D) enrichment scores. We associated JASPAR motifs with sea urchin transcription factors by identifying the sea urchin genes present in the same gene families as the human genes for which the TFBS were characterised. When multiple P. lividus genes were present in a given gene family, we determine orthology relationships by selecting the urchin gene that had the most recent last common ancestor with the motif-associated gene in the phylogenetic trees reconstructed for each gene family (Table S10). For enrichment analysis, the number of TFBS in the population of peaks associated with the gene of interest (<150kb of the TSS) was contrasted with the population of peaks associated with other genes in a hypergeometric test for all TFBS and subjected to BH correction for multiple testing (e.g. Table S8).

Network reconstruction was performed using ANANSE18 from ATAC-seq aligned BAM and the ‘gimme.vertebrate.v5.0’ binding motif database assigned to P. lividus TFs using gimme motif2factor script. Resulting network was reconstructed using the igraph package in R retaining edges with probably above 90% quantile. were for plotting and computation of centrality (degree index).

Cut&Tag
The assay was performed using the CUT&Tag-IT™ Kit and H3K27ac antibody (Active Motif Ref 39135)). Dissociated cells were lysed in a hypotonic buffer in the presence of NP40, Tween20 and Digitonin each at 0.1% and nuclei were washed in Resuspension Buffer (RSB) (RSB:10mM Tris pH: 7.4, 10 mM NaCl, 3 mM MgCl2 and stored at -80°C in RSB containing 20% glycerol. For stages swimming blastula, gastrula, prism and pluteus, 10^5 nuclei were used while for stage early blastula 15000 nuclei were used. Nuclei were thawed and washed once in wash buffer (20 mM Hepes pH: 7.5, 150 mM NaCl, 0.5 mM spermidine, 1x EDTA free Protease inhibitor cocktail mixed with activated Concanavalin A beads and magnetized to remove liquid. Nuclei were then incubated with the primary antibody in Wash buffer containing digitonin at 0.05% for 2h at room temperature, washed in Wash buffer and incubated for 1h with secondary antibody (guinea pig anti-rabbit) diluted in Wash buffer supplemented with digitonin at 0.05%. The beads were then washed and resuspended in assembled proteinA-Tn5 transposomes diluted in wash buffer+ 300 mM NaCl for 1h. At the end of the incubation, beads were washed with wash buffer supplemented with 300 mM NaCl. After the wash, tagmentation was performed by adding 125 μl of tagmentation buffer supplemented with 10 mM MgCl2 and continued for 1h at 37°C. Following tagmentation, beads were magnetized, washed, and incubated for 1h at 37°C with proteinase K and SDS to digest the chromatin and release DNA fragments. At the end of the incubation, beads were magnetized and DNA was purified using 625 μl DNA binding buffer. Following purification, libraries were amplified by using 25 μl of DNA sample and 2.5 μl of of a uniquely barcoded i5 primer and 2.5 μl of a uniquely barcoded i7 primer in a 50 μl PCR reaction with Q5 high-fidelity DNA polymerase. The program included gap-filling at 72°C then 14 cycles at 98°C for 10 seconds and 63°C for 10 seconds. After PCR, clean-up was performed by adding 55 μl of SPRI beads and eluting in 21 μl of DNA purification buffer. Libraries were then analyzed using the Agilent 4200 Tapestation instrument and sequenced. Reads were aligned using bowtie2 with the parameters ‘–local –very-sensitive –no-mixed –no-discordant –phred33 -10 -X 700 -3 75’ and coverage calculated for previously defined ATAC peaks using BAMScale.129 Coverage density heatmaps for ATAC-seq and Cut&Tag datasets were plotted using Deeptools ‘plotHeatmap’ function.130

**In situ hybridization**

The sequence of pop2 was retrieved from an EST library while the sequences of pop1 and pop3 were obtained from available transcriptomes. For **in situ** hybridisation the full-length sequence of all three genes were cloned into pGemT. pop1 and pop2 plasmids were linearised with Ncol and transcribed with SP6 polymerase. pop3 plasmid was linearised with SpeI and transcribed with T7
polymerase. *In situ* hybridisation was performed using standard methods (Harland 1991) with Digoxigenin labelled RNA probes and developed with chromogenic substrates NBT and BCIP. The Delta and ali1 probes have been described previously (Röttinger et al. 2004). Control and experimental embryos were developed for the same time in the same experiments. Embryos were imaged with an Axio Imager M2 microscope.

**Overexpression of mRNA**

For overexpression studies, the open reading frame of each pop gene was amplified by PCR and cloned into the pCS2 vector at the BamHI and XhoI sites. Capped mRNAs were synthesized from NolI-linearized templates using mMessage mMachine kit (Ambion) and SP6 polymerase. After synthesis, capped RNAs were purified on Sephadex G50 columns and quantitated by spectrophotometry. RNAs were mixed with Tetramethylrhodamine Dextran (10000 MW). pmar1 mRNA was injected at 30 μg/ml, pop1, pop2 and pop3 mRNAs were injected at concentrations in the range 30–100 μg/mL with similar effects.

**QUANTIFICATION AND STATISTICAL ANALYSIS**

Various statistical tests were used to calculate p values as indicated in the methods section, figure legend, or text, where appropriate. Results were considered statistically significant when p < 0.05 or FDR<0.05 when multiple hypothesis correction was applied, unless stated otherwise. Statistical analyses were conducted using R (v4.1.0).