Finding cell-specific expression patterns in the early Ciona embryo with single-cell RNA-seq

Garth R. Ilsley[1,6], Ritsuko Suyama[1,2,6], Takeshi Noda[1,3], Nori Satoh[1], Nicholas M. Luscombe[1,4,5,*]

[1] Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan

[2] Current address: Osaka University, 1-3 Yamadaoka, Suita, Osaka 565-0871, Japan

[3] Current address: Shinshu University, Matsumoto, Nagano Prefecture 390-8621, Japan

[4] The Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK

[5] UCL Genetics Institute, University College London, Gower Street, London WC1E 6BT, UK

[6] These authors contributed equally to this work.

* Correspondence should be addressed to nicholas.luscombe@oist.jp.

Abstract

Single-cell RNA-seq has been established as a reliable and accessible technique enabling new types of analyses, such as identifying cell types and studying spatial and temporal gene expression variation and change at single-cell resolution. Recently, single-cell RNA-seq has been applied to developing embryos, which offers great potential for finding and characterising genes controlling the course of development along with their expression patterns. In this study, we applied single-cell RNA-seq to the 16-cell stage of the *Ciona* embryo, a marine chordate and performed a computational search for cell-specific gene expression patterns. We recovered many known expression patterns from our single-cell RNA-seq data and despite extensive previous screens, we succeeded in finding new cell-specific patterns, which we validated by in situ and single-cell qPCR.

Introduction

Cell types can be characterised at high resolution with single-cell RNA-seq. Apparently homogenous groups of cells can be clustered to identify novel and rare subtypes[1–5] and cells undergoing differentiation at different rates can be ordered and analysed[6–9]. This offers great potential for studying developing embryos with their high diversity of cell types[10–20]. Making sense of this diversity is a goal of developmental biology and an important first step is identifying and characterising the relatively few genes controlling the course of development and the subsets of cells they are expressed in, or in other words, their gene expression patterns.

We applied single-cell RNA-seq (scRNA-seq) to the 16-cell stage of *Ciona*, a simple chordate, and looked for gene expression patterns using a novel computational approach. Zygotic expression begins around the 8-cell stage[21] and many gene expression

patterns, both maternal and zygotic, are known from comprehensive in situ screens in *Ciona*[22] (Fig. 1b, Supplementary Table 1). We identified many of these from our scRNA-seq data. In addition, we found unknown gene expression patterns, which we validated by in situ and single-cell qPCR.

Results

Ciona development follows an invariant lineage allowing cells to be unambiguously identified at the 16-cell stage[23–25], whether before sequencing as in our work (Fig. 1a) or afterwards based on expression, as in recent work of the same stage[26]. We collected cells from four 16-cell stage embryos, each from different individuals fertilised on different days. The left and right side of the embryo at this stage are thought to be symmetrical, but to avoid any potential bilateral variation in our small sample, we collected all eight cell types from the right half of each embryo. We multiplexed the cells of each embryo (or batch) and sequenced them on Illumina MiSeq, as well as on three lanes across two Illumina HiSeq 2500 runs, leading to scRNA-seq measurements for 32 cells (Supplementary Table 2).

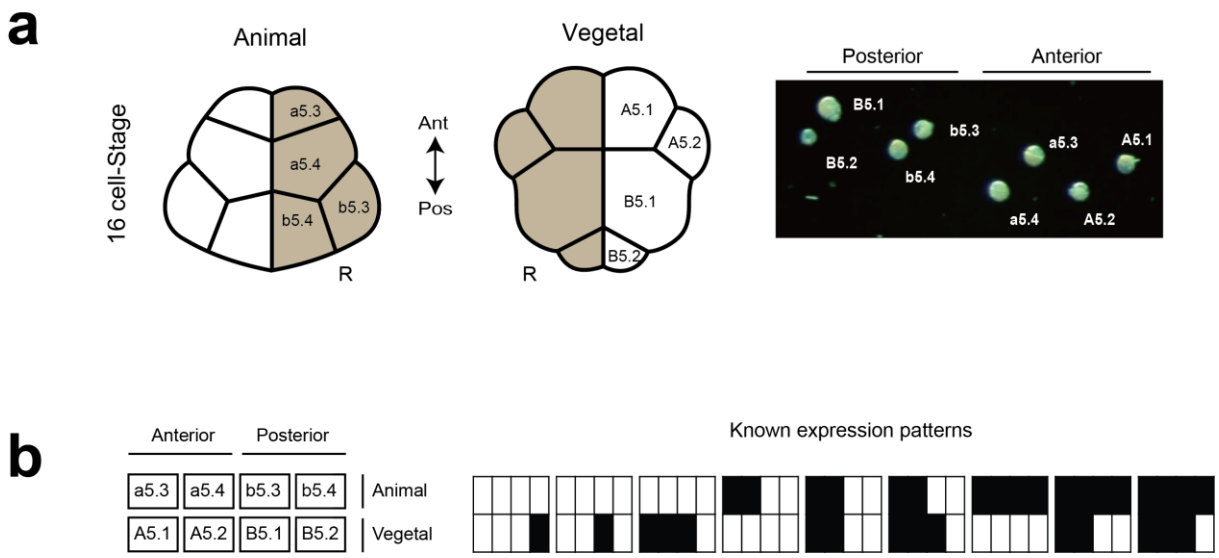


Figure 1: The cell types and gene expression patterns of the 16-cell *Ciona* embryo are well characterised. **(a)** Drawing of a 16 cell–stage embryo showing *Ciona* cell type nomenclature (left). All eight cells were dissected and isolated by glass needle from the right side of four 16-cell embryos, and identified visually under microscope (right). **(b)** Nine known expression patterns of *Ciona* (right) are shown according to a schematic layout (left), with solid rectangles indicating expression in the corresponding cells.

Our cells were sequenced with relatively high coverage (around 20 million reads per cell on Illumina HiSeq and 4.5 million per cell on MiSeq). Although we had only four replicates per cell type, we produced a reliable dataset.

Assessment of technical variability and reproducibility

Our results show limited technical variation within each batch: the expression levels in different cell types from the same embryo are well correlated for embryos 2, 3 and 4. They are, in fact, more similar to each other than the same cell types are across different individuals (Fig. 2a-b). Although we cannot separate out the sources of cross-embryo variation, this result is consistent with a previous report showing that maternal mRNA levels vary significantly between unfertilized eggs from different individuals[27]. It is also worth noting that very little of the variation between embryos is from the sequencing run. This can be seen by comparing our sequence results from MiSeq with HiSeq—the correlation between unnormalised counts from the two platforms is over 99% for every cell type, whether genes with zero counts are included or excluded. This is consistent with previous results showing high correlation between expression measurements from tens of millions of reads per cell and those from lower coverage of a million or fewer reads[28,29].

This embryo batch effect is further demonstrated by a Principal Components Analysis (Fig. 2b), which shows a similar result with the cell types of embryos 2, 3 and 4 being close to each on the first two components (which explain 56% of the variance) and the cell types of embryo 1 being more spread out.

The close clustering of cells from the same embryo, as well as their high correlation, suggests that our technical variability is low, leading to reproducibility

78 within each batch (or embryo). A confirmation of the reproducibility of our results is the
79 tight distribution of genes detected across samples within embryos (Fig. 2c-d). Genes
80 were considered detected when the measured count was greater than zero. These
81 results show that slightly more genes were detected on HiSeq than MiSeq, but that the
82 median difference for each embryo is less than 10%. This can be compared with a
83 previous result showing a reduction of genes detected of around 39% when lowering
84 sequence coverage to less than a million reads per cell[28]. As before, embryo 1 showed
85 more variability across samples than the other embryos.

86 Further, our data can be compared to previously published data for the 16-cell
87 stage that was generated using gene expression microarrays from many pooled single
88 cells[27]. We found good agreement with our scRNA-seq expression patterns for the key
89 genes shown in their paper (Supplementary Fig. 1).

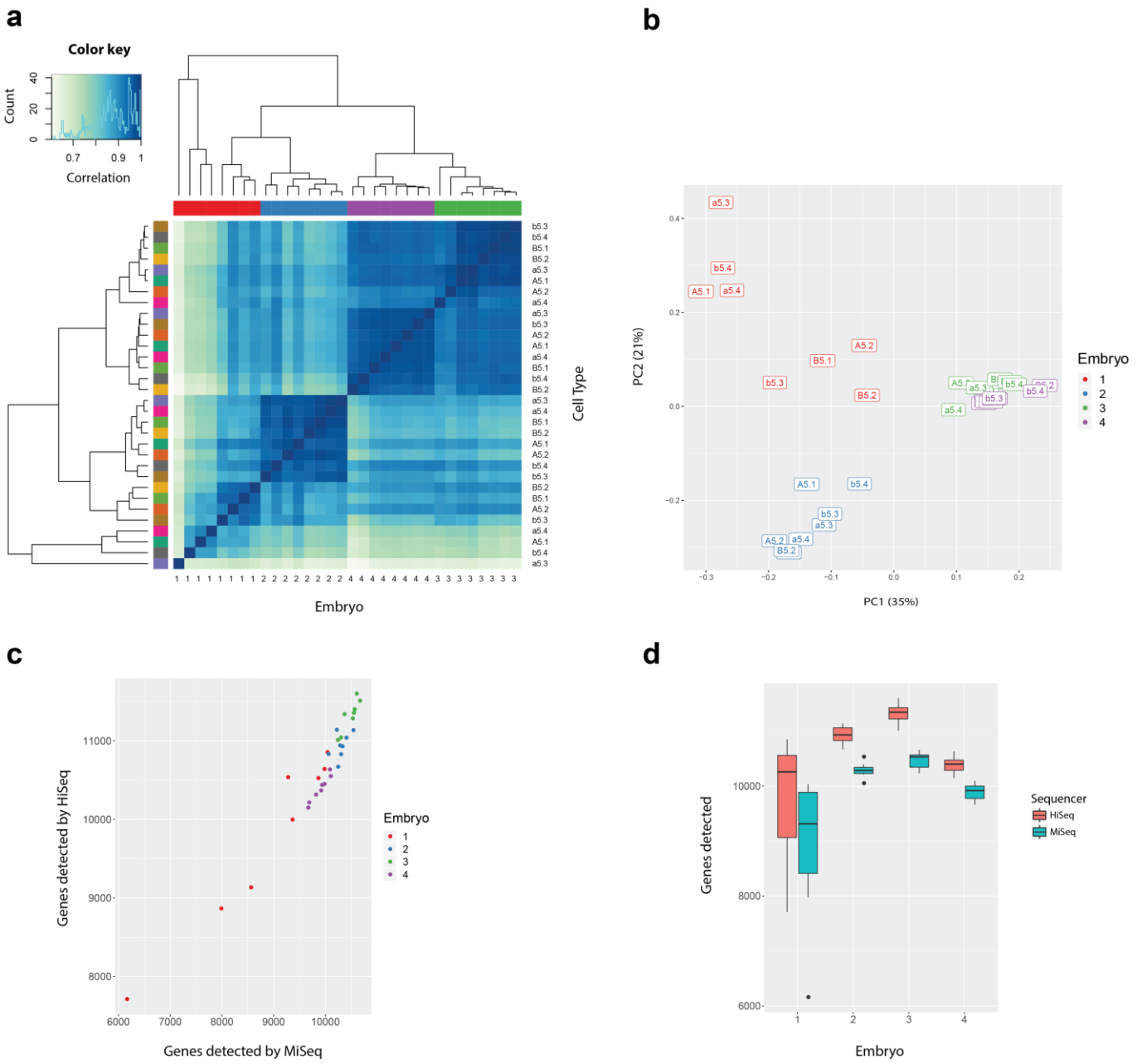


Figure 2: Our dataset exhibits low technical variability within batches and between Illumina MiSeq and HiSeq. (**a**) Clustered heat map of the symmetrical correlation matrix of normalised expression data from HiSeq samples (excluding ERCC counts and genes with zero counts), with the histogram in the top left providing the colour key for the correlation matrix. The embryo and cell type annotations for each sample are shown on different axes, but the samples are arranged in the same order. The samples are colour-coded by embryo along the top and by cell type on the left with the same hierarchical clusters shown in both cases. (**b**) PCA plot of the same data showing the first two components, which explain 56% of the total variance. The samples are colour-coded by embryo and labelled by cell type. (**c and d**) The number of genes detected is consistent across the four embryo replicates whether the libraries were sequenced on MiSeq or HiSeq. (**c**) Scatter plot showing the consistent relationship between MiSeq and HiSeq, with more zeros or undetected genes for some cells of embryo 1 compared to the others. (**d**) Boxplots showing the narrower distribution of genes detected for embryos 2 to 4 compared to embryo 1 and the consistent increase from MiSeq to HiSeq.

Normalisation of counts

After producing a dataset of counts, we normalised our data. Since we were interested in gene expression differences between different cell types, we did not normalize across genes (such as by GC content or transcript length), but only for sequencing depth by dividing by the total number of reads per sample. More complex normalisation steps could be considered[30], but this simple approach suffices for our data. It gives a natural measure of expression for each gene, namely the proportion it contributes to the total, which we assume is independent of the total number of reads.

We then made use of a suitable transformation of proportions, the arcsine of the square root (referred to here as the φ transformation), namely

$$\varphi_{ij} = 2 \arcsin\left(\sqrt{\frac{k_{ij}}{N_j}}\right)$$

where k_{ij} is the count for the ith gene and j^{th} sample and N_j is the total number of counts for the jth sample. The difference between φ_i of two samples can be interpreted as an effect size index for proportions, namely Cohen's h[31].

In practice, the arcsine function in the φ transformation is largely redundant because most genes are expressed at low proportions. At these low values, the arcsine transformation is close to identity, meaning that a square root transformation of proportions performs equivalently in many cases.

110 **Finding putative gene expression patterns**

111 We then took a simple and effective approach to find gene expression patterns.
112 Instead of grouping cells into a single set of clusters, we clustered the gene expression
113 pattern of each gene separately. For each gene, we grouped the different cell types into
114 two classes of ON and OFF expression. We took advantage of our known cell types and
115 calculated the Euclidean distance between vectors of replicate measurements for each
116 cell and performed single-linkage hierarchical clustering of these. The resulting top-
117 level clusters determined the ON/OFF pattern of each cell type for each gene.

118 We then ranked these results, by taking an approach that does not require
119 parametric estimation of variation or dispersion. We calculated our cluster reliability
120 score as the difference between the first quartile of the ON cluster and the third quartile
121 of the OFF cluster, which we call the Transquartile Range (TQR). The TQR is larger when
122 the difference in cluster means is larger, but it penalizes higher variation for a given
123 difference in means.

124

125 **Ranking cell-specific gene expression patterns**

126 We applied this approach and produced a list of ranked genes and their expression
127 patterns. We examined our results for 77 genes (Supplementary Table 1) with known
128 in situ patterns (Imai et al. 2004; Matsuoka et al. 2013; Miwata et al. 2006; Prodon et al.
129 2007). Our top 35 results matched the known in situ patterns (Supplementary Fig. 2a),
130 except for KH.L152.12, which is validated below as a new pattern by in situ and qPCR
131 (Fig. 4). However, the lower ranked results did not correspond well (Supplementary
132 Fig. 2b).

There were a few reasons for this. For example, clustering did not produce a correct pattern for *Lefty* (Supplementary Fig. 2b). The cell types, A5.2 and B5.1, are normally considered part of the *Lefty* pattern, but they have intermediate expression in our scRNA-seq results and, in this instance, the clustering algorithm places them in the OFF cluster. For a few other genes, e.g. *DPOZ* (KH.C12.589) and *Dlx.b*, the clustering is correct, but the TQR is low. For a few other genes, no reads were mapped, e.g. *Sox7/17/18*, *Fringe 2* and KH.C13.22, but in most cases where our scRNA-seq data does not agree with published in situ patterns, our expression measurements were low or relatively uniform across the eight cells and thus the algorithm functions correctly in attributing lower score to these results. Many of these genes are expressed maternally in the *Ciona* embryo and hence are not easily detectable by our RNA-seq protocol, which amplifies from the poly(A) tail of mRNA. In situ hybridisation can detect localised maternal signal in the cytoplasm, which could explain the discrepancy. It is also possible that some in situ results are false positives.

Validating cell-specific gene expression patterns

Using these known in situ results, we assessed that the top 40 ranked results from all genes were likely to be reliable and focused on these for further validation. We found 12 distinct patterns in the top 40, which included all known patterns as well as three potentially new patterns (highlighted in orange in Fig. 3). Ten patterns are currently known at this stage in *Ciona*[27,32–37], and although we matched only nine of these directly (Fig. 1b), the tenth pattern has a single known case, *Tfap2-r.b* (*AP-2-like2*), which we did pick up, but without expression in A5.2. This agrees with previous observations that expression is not consistent in this cell across embryos[27,35]. Further, it is in agreement

157 with the average over many embryos as measured by microarray[27] (Supplementary Fig.
158 1).
159 We validated one of the potentially new patterns by in situ and single-cell qPCR,
160 namely the pattern for KH.L152.12 (Fig. 4e). We could not validate the second pattern
161 for KH.C1.933 that has relatively ubiquitous expression, although with apparently
162 reduced expression in b5.4. We also did not manage to validate the third potentially new
163 pattern (expression in A5.1 and A5.2 only), although the scRNA-seq expression data in
164 this case seems to show clear expression in these cells.
165 In addition to recovering all known patterns and validating one novel pattern, we
166 found at least 28 genes with known in situ patterns. Of the 12 patterns we found, the
167 pattern with expression in the B5.2 cell only was the most represented. This is also the
168 most frequent pattern in known in situ patterns, i.e. postplasmic/PEM RNAs[38]. The
169 majority of our results for B5.2 are confirmed by previous in situ datasets, but we
170 identified new B5.2-specific genes, such as KH.C13.98 and KH.C12.212, confirming their
171 expression by in situ and single-cell qPCR (Fig. 4a).
172 We also validated other classes of uncharacterized genes, namely KH.S1497.1,
173 which expresses specifically in the animal hemisphere, and KH.C11.529 on the anterior
174 side (Fig. 4c-d). These results are particularly striking, since it was expected that no
175 further zygotic genes expression patterns would be found at this stage in *Ciona*[22].
176

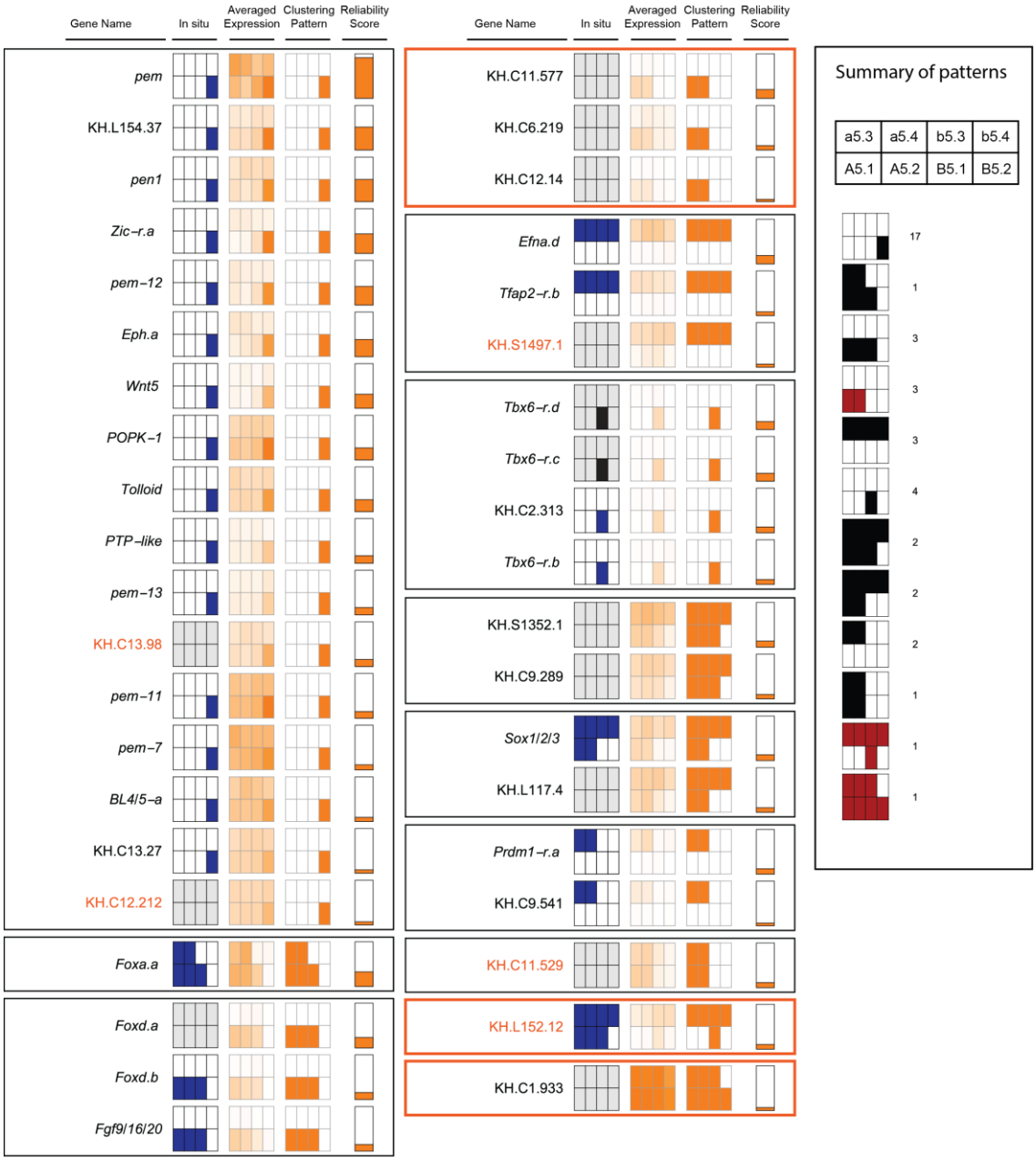


Figure 3: All previously known patterns occur in the top 40 genes when ranked according to their Transquartile Range (TQR). Expression summary plots of the top 40 genes, grouped by pattern, with an orange border indicating an unknown pattern. The arrangement of cells in each pattern is shown in the summary box on the right, together with the number of genes per pattern, whether known (black) or unknown (red). Genes whose patterns we validated are highlighted in orange. For each gene, the columns indicate the gene name, any previously known in situ pattern in blue (grey if not known), the average of the transformed expression values (clipped above 0.05), the pattern resulting from clustering, and finally, the TQR as the reliability score, scaled for visualization. The in situ probe for Tbx6-r.b also detects Tbx6-r.c and Tbx6-r.d (indicated by a black in situ pattern).

We also looked more widely in the top 60 (Supplementary Fig. 3) and validated additional genes, KH.C8.450, KH.C9.289 and KH.C4.260, by single-cell qPCR and in situ hybridisation (Fig. 4b). The first is another example of B5.2 expression, whereas the last two genes are expressed in all cells except B5.2, a pattern known previously from *Hes.a*[39]. While developing our approach, we also applied the TQR to the microarray data of this stage[27] and looked for reliable B5.2 cell-specific expression using both datasets. As a result, we also found and validated KH.L60.2 and KH.C14.501 (ranked 88 and 2403 in our data; see Supplementary Table 7).

In conclusion, we have recovered many known patterns, as well as patterns and genes that had not been detected previously despite extensive in situ screens. These results open up opportunities for further research into developmental patterning in *Ciona*. In addition, we have demonstrated that single-cell RNA-seq is a viable alternative to extensive in situ screens, offering a promising approach for finding genes with cell-specific expression in less well-studied organisms.

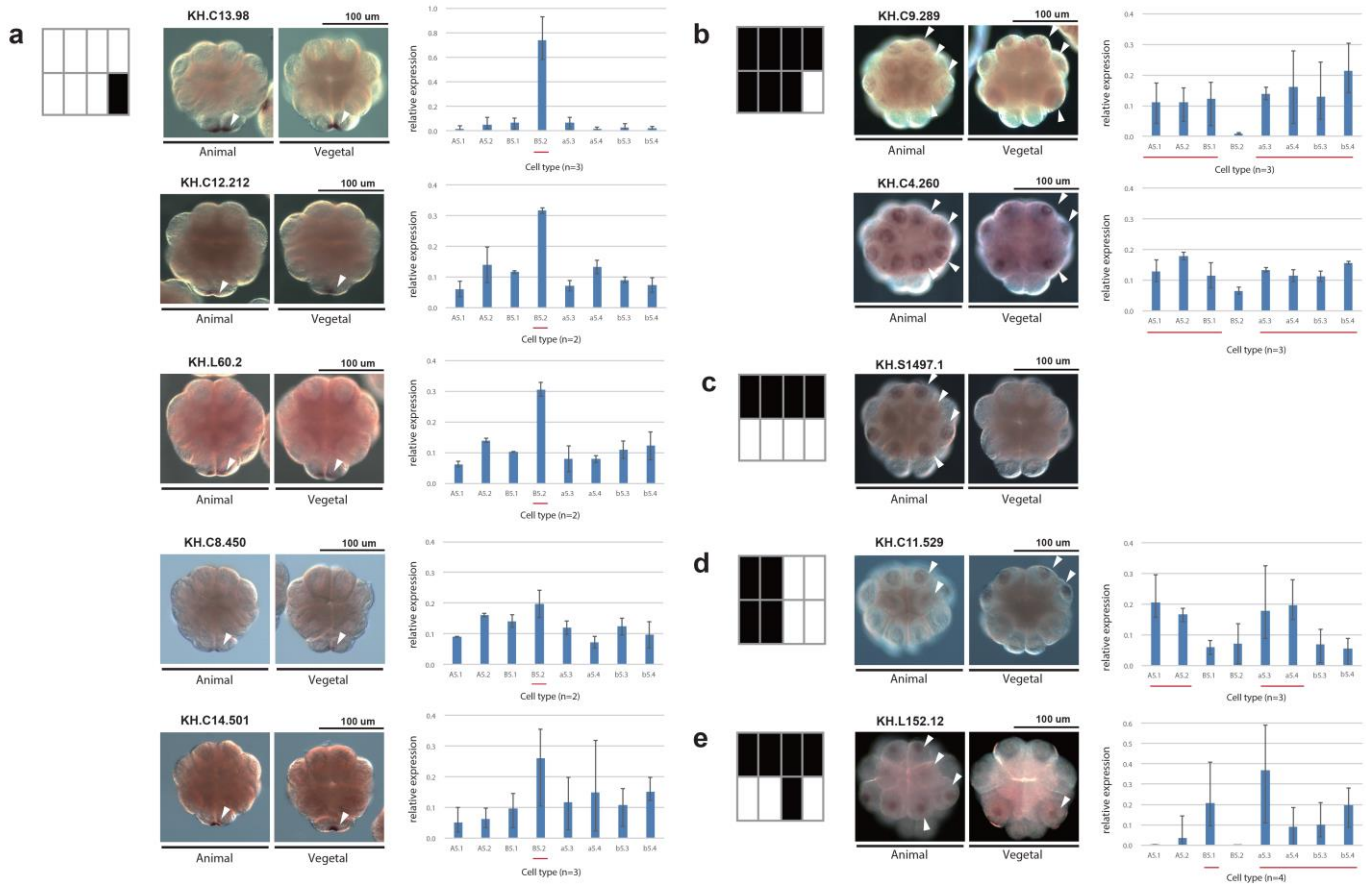


Figure 4: Pattern discovery results are validated by in situ hybridisation and single-cell qPCR.

(**a-e**) For each pattern being tested (left), the in situ hybridisation results are shown from the animal and vegetal side (middle). The arrowheads pointing to the expressing cells are shown on only one side of the embryo. The scale bar indicates 100 µm. Gene expression levels for each cell type were measured by single-cell qPCR. The qPCR measures for the cell types of each embryo were scaled to add to 1. The bar charts (right) show the mean of the resulting values of biological replicates, with whiskers indicating the range. The lines under the bars indicate the cells of the expression pattern being tested. The number of replicates of each gene is shown under each bar chart.

194

Methods

Study design

We isolated cells from five 16-cell stage *Ciona* embryos, each on a different day (Supplementary Table 2). Early ascidian embryos are thought to be bilaterally symmetrical, but to avoid any potential bilateral variation in our small sample, we collected eight cells from the right side of each embryo. The cells were collected individually in batches of eight cells from the same embryo on the same day, with sequencing libraries prepared in parallel, barcoded and then sequenced together. This means that biological variation between embryos and technical variation between batches cannot be distinguished. The advantage of this design is that it minimizes technical variation between cell types of the same embryo and controls for confounding technical and biological variation between embryos. Averaging across the cell types of different batches reduces this unwanted variation, maintaining cell-specific variation. Our results show that cells from the same embryo are more similar to each other than the same cell types are across individuals, with a similar number of genes detected per cell type (Fig. 2a-d).

Preparation of *Ciona* embryos

Ciona intestinalis type A, recently designated *Ciona robusta*[40,41], adults were obtained from Maizuru Fisheries Research Station (Kyoto University) and Misaki Marine Biological station (The University of Tokyo) under the National Bio-Resource Project for *Ciona*. They were maintained in aquarium in our laboratory at Okinawa Institute of Science and Technology Graduate University under constant light (Calcitrans, Nisshin Marinetech Co., Ltd.) for three days apart from a few hours of darkness a day with

219 feeding to induce spawning of the old eggs. After this, the *Ciona* were maintained under
220 constant light to induce oocyte maturation. Eggs and sperm were obtained surgically
221 from the gonoducts. Embryos were dechorionated after insemination using a solution of
222 0.07% actinase and 1.3% sodium thioglycolate. Eggs were reared to reach the 16-cell
223 stage in Millipore-filtered seawater (MFSW) at about 18 °C. Embryos from each
224 insemination batch were kept to check the ratio that developed into morphologically
225 normal tailbud. We only used embryos from batches where more than 70% developed
226 normally to tailbud (10 hours post fertilization at 18 degrees) (see Supplementary
227 Table 2 for embryo batch information).
228
229 **Naming of cells**
230 In *Ciona*, cells are named using the nomenclature of Conklin[23]: the animal side is
231 prefixed with a lowercase letter (a or b) and the vegetal with an uppercase letter; the
232 anterior with A or a and the posterior with B or b. The initial letter is followed by a
233 number that indicates the embryo stage since fertilization, with individual cells
234 numbered according to their lineage. At the 16-cell stage, the animal domain
235 corresponds to a5.3, a5.4, b5.3 and b5.4, the vegetal domain to A5.1, A5.2, B5.1 and B5.2,
236 and postplasmic RNAs are localized to B5.2.
237
238 **Isolation of single cells at the 16-cell stage**
239 At a defined point in development of the 16-cell embryo i.e., at the stage
240 immediately after compaction of the embryo (2.5 ~ 2.6 hours post fertilization), the
241 embryo was transferred to 4°C to slow its development. Each blastomere was isolated
242 with a fine glass needle in a mannitol solution (0.77 M mannitol : MFSW, 9:1) under a

243 stereo microscope at 4 °C regulated by a thermo plate (Tokai Hit Co., Ltd.) and its
244 identity noted. Isolated blastomeres were picked up and transferred immediately with a
245 mouth pipet into a lysis buffer[42] for reverse transcription.
246
247 **Library preparation**
248 We followed the single-cell library preparation method of Tang et al[42,43] with some
249 modification. We added ERCC spike-in RNA (Thermo Fisher scientific, 4456740,
250 1:80000) to each lysis buffer and applied 14 and then 9 cycles of PCR amplification after
251 second strand synthesis. Amplified cDNA was purified with MinElute PCR Purification
252 kit (28006, QIAGEN) and QIAquick PCR Purification Kit (28106, QIAGEN) after each PCR
253 reaction respectively and its concentration measured with Qubit® 2.0 Fluorometer
254 (Q32866, Life Technologies) to have more than 150 ng total yield of cDNA. The quality of
255 the amplified cDNA and distribution of DNA fragment size were confirmed by Agilent
256 2100 Bioanalyzer (Agilent Technologies) with High Sensitivity DNA Kit (5067-4626,
257 Agilent) to consist mainly of 1.0-1.5 kb fragments.
258 Amplified cDNAs were sheared using sonication Covaris S2 System to produce
259 DNA of 300 bp on average. The settings were as follows: Duty cycle: 20%, Intensity: 5,
260 Cycles per burst: 200, Power mode - Frequency sweeping, Treatment time: 90 seconds,
261 Temperature: 12°C.
262 NEB Next® ChIP-Seq Library Prep Master Mix Set for Illumina® (E6240, NEB) was
263 applied to sheared cDNA for preparation of the library for the Illumina platform.
264 NEBNext® Multiplex Oligos for Illumina (E7335, E7500, NEB Next Multiplex Oligos for
265 Illumina, NEB) were combined to introduce an index and adaptor to the double-
266 stranded DNA. After extraction of the 300 bp fraction of adaptor-ligated DNA by E-Gel

267 Size Select 2% Agarose (G661002, Invitrogen), DNA was amplified with individual index
268 primers using PCR with 19 cycles.
269 The amplified DNA fragment composition was purified with Agencourt AMPure XP
270 twice (A63881, Beckman) and again checked by Qubit (> 60 ng of cDNA in total yield)
271 and by Bioanalyzer to ensure that the fragment size was sharply distributed around 300
272 bp (on average, about 320 bp with a standard deviation of 40). The concentration of
273 fragments with appropriate index adapters was quantified by KAPA Library
274 Quantification Kits (KAPA Library Quantification Kits, Illumina GA/Universal, KK4825,
275 Genetics) to ensure that the final libraries had adapters for both ends and their
276 concentration was at least 20 pM.
277
278 **Data generation and quality checking**
279 Libraries were sequenced on Illumina's (San Diego, CA) MiSeq benchtop sequencer
280 and Illumina HiSeq 2500. Libraries were prepared with different index primers and
281 sequenced on MiSeq using paired 150 nt reads (No. MS-102-2002, MiSeq Reagent Kit
282 v2) with eight multiplexed samples per run with the standard Illumina protocols. The
283 same libraries were sequenced on an Illumina HiSeq 2500 with 150 bp paired end reads
284 (No. PE-402-4001 and FC-402-4001, TruSeq Rapid Cluster - Paired-End and SBS Kits)
285 with 16 multiplexed samples per lane following standard Illumina protocols. Our results
286 from using HiSeq and MiSeq were similar (Fig. 2c-d, cf. Supplementary Figs 3 and 4).
287 The resulting reads were aligned using Bowtie[44] version 2.2.6 to the *Ciona* KH
288 genome assembly[45,46], downloaded from Ghost
289 (http://ghost.zool.kyoto-u.ac.jp/download_kh.html). Reads were mapped using local
290 alignment (--local), with other settings at their default. We did not trim or filter reads,

291 but instead made use of local alignment to find the optimal match. This had the
292 additional benefit that we did not need to split up reads to handle transcripts spanning
293 more than one intron, as is done, for example, in TopHat[47]. Gene counts were calculated
294 from the resulting alignment files using htseq-count[48] with the non-stranded option and
295 mode "intersection-nonempty" against the KH gene models (version 2013) downloaded
296 from Ghost.

297 We assessed our samples for mapping quality. We excluded one embryo from
298 subsequent analysis since it had oligo-dT primer sequence in more than 50% of its read
299 pairs; the remaining four embryos had less than 1% of read pairs affected. All remaining
300 samples mapped well to the genome (Supplementary Table 3) and a uniform number of
301 genes were detected (about 60%), although embryo 1 had noticeably fewer detected
302 genes for some of its cells.

303

304 **Pattern discovery**

305 Hierarchical clustering to determine candidate patterns was performed with
306 ClusteringComponents in Mathematica 10.4 with the Agglomerate method and
307 Euclidean distance function. This is equivalent to *hclust* in R with the single linkage
308 method. The Transquartile Range (TQR) was calculated as the difference between the
309 first quartile of the ON cluster and the third quartile of the OFF cluster. The quantile
310 method for the TQR used linear interpolation equivalent to type 5 in the R *quantile*
311 function (the hydrologist method). The resulting patterns and TQR score are listed in
312 Supplementary Table 7.

313

314 **Single-cell qPCR analysis**

315 cDNA was reverse transcribed from all cells of one embryo per gene replicate
316 using the same protocol we used for single-cell RNA-seq[42,43]. Quantitative PCR was
317 performed using a StepOnePlus PCR machine (Applied Biosystems) with the SYBR green
318 method (No. RR820B, Takara). Each gene was measured with either two or three
319 replicates, except KH.L152.12, which had four. We did not get reliable measurements for
320 KH.S1497.1. The qPCR measures for the cell types of each embryo were scaled between
321 0 and 1 and then averaged for each cell type across replicates. If there was insufficient
322 target mRNA, it was first amplified using primers covering a wider region of the target
323 gene than those used for single-cell qPCR. Amplification of a specific product in each
324 reaction was confirmed by determining a dissociation curve and comparing with the
325 relevant standard plasmid to estimate the copy number in each cell type. The primers
326 for single-cell qPCR analysis, the IDs for the cDNA clones and the resulting data are
327 listed in Supplementary Tables 4-6.

328

329 **In situ hybridisation**

330 Whole-mount in situ hybridisation was carried out as previously described with
331 minor modification[49]. Dig-labeled antisense RNA probes were synthesized in vitro from
332 cDNAs from the *Ciona* cDNA project[50]. mRNA expression was visualized using the
333 NBT/BCIP system (Roche, No.11681451001) and detected on a Zeiss Axio Imager Z1
334 microscope using DIC (Differential Interference Contrast). The images were acquired
335 with Axiovision SE64 release 4.9.1. Contrast and brightness were adjusted for some
336 images using Adobe Photoshop. The IDs for the cDNA clones are shown in
337 Supplementary Table 5.

338

339 **Microarray processing**

340 Previously published microarray data[27] was processed with the limma R
341 package[51]. Background was corrected using *normexp* and arrays were normalised with
342 the *quantile* method.

343

344 **Gene models and names**

345 Gene names for the KH 2013 gene models were downloaded from Ghost
346 (http://ghost.zool.kyoto-u.ac.jp/TF_KH.html and http://ghost.zool.kyoto-
347 u.ac.jp/ST_KH.html) and supplemented with names from Prodon et al[38].

348

349 Code and Data Availability

350 RNA-seq data have been deposited in the ArrayExpress database at EMBL-EBI
351 (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-6117. Software is
352 available at https://github.com/ilsley/Ciona16.

353

354 Acknowledgements

355 We thank the staff of the Maizuru Fisheries Research Station of Kyoto University
356 and Misaki Marine Biological Station of the University of Tokyo for collecting and
357 cultivating *Ciona* under the National BioResource Project (NBRP) of MEXT, Japan, and
358 RIKEN BRC for providing *Ciona* EST clones through the NBRP. We thank Vladimir Benes
359 and Dinko Pavlinic in the Genomics Core Facility at the European Molecular Biology
360 Laboratory (EMBL) for initial advice on the library preparation protocol and the
361 members of the OIST DNA Sequencing Section for their support in running our samples

on their Illumina MiSeq and HiSeq machines. We also thank Sylvain Guillot for his technical support, Filipe Tavares-Cadete for early feedback on the method and Kenji Kobayashi for a helpful comment on the validation experiment. This work was supported by core funding from OIST to the Genomics & Regulatory Systems and Marine Genomics Units.

Author contributions

NML and NS established and supervised the project. All authors contributed to the study design. RS and TN optimized the experimental protocols, collected and prepared the samples and sequencing libraries. RS designed and performed the in situ and qPCR analysis. GRI conceived and designed the normalization, gene testing and pattern discovery method and performed the bioinformatics analysis. GRI, RS and NML wrote the paper and all authors edited and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

1. Björklund, Å. K. *et al.* The heterogeneity of human CD127+ innate lymphoid cells revealed by single-cell RNA sequencing. *Nat Immunol* **17,** 451–460 (2016).
2. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525,** 251–255 (2015).

3. Jaitin, D. A. *et al.* Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science* **343,** 776–779 (2014).
4. Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat Meth* **14,** 483–486 (2017).
5. Usoskin, D. *et al.* Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nature Neuroscience* **18,** 145–153 (2015).
6. Mojtahedi, M. *et al.* Cell Fate Decision as High-Dimensional Critical State Transition. *PLOS Biology* **14,** e2000640 (2016).
7. Olsson, A. *et al.* Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537,** 698–702 (2016).
8. Richard, A. *et al.* Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process. *PLOS Biology* **14,** e1002585 (2016).
9. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotech* **32,** 381–386 (2014).
10. Biase, F. H., Cao, X. & Zhong, S. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Res.* **24,** 1787–1796 (2014).
11. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357,** 661–667 (2017).
12. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science* **343,** 193–196 (2014).

13. Goolam, M. *et al.* Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell* **165,** 61–74 (2016).
14. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* **2,** 666–673 (2012).
15. Ibarra-Soria, X. *et al.* Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nat. Cell Biol.* **20,** 127–134 (2018).
16. Karaiskos, N. *et al.* The Drosophila embryo at single-cell transcriptome resolution. *Science* **358,** 194–199 (2017).
17. Scialdone, A. *et al.* Resolving early mesoderm diversification through single-cell expression profiling. *Nature* **535,** 289–293 (2016).
18. Tintori, S. C., Osborne Nishimura, E., Golden, P., Lieb, J. D. & Goldstein, B. A Transcriptional Lineage of the Early C. elegans Embryo. *Developmental Cell* **38,** 430–444 (2016).
19. Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500,** 593–597 (2013).
20. Yan, L. *et al.* Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol* **20,** 1131–1139 (2013).
21. Rothbächer, U., Bertrand, V., Lamy, C. & Lemaire, P. A combinatorial code of maternal GATA, Ets and β-catenin-TCF transcription factors specifies and patterns the early ascidian ectoderm. *Development* **134,** 4023–4032 (2007).
22. Satou, Y. & Imai, K. S. Gene regulatory systems that control gene expression in the *Ciona* embryo. *Proceedings of the Japan Academy, Series B* **91,** 33–51 (2015).

23. Conklin, E. G. The organization and cell-lineage of the ascidian egg. *Journal of the Academy of Natural Sciences of Philadelphia* **13,** 1–119 (1905).
24. Lemaire, P. Unfolding a chordate developmental program, one cell at a time: Invariant cell lineages, short-range inductions and evolutionary plasticity in ascidians. *Developmental Biology* **332,** 48–60 (2009).
25. Nishida, H. Specification of embryonic axis and mosaic development in ascidians. *Dev. Dyn.* **233,** 1177–1193 (2005).
26. Treen, N., Heist, T., Wang, W. & Levine, M. Depletion of Maternal Cyclin B3 Contributes to Zygotic Genome Activation in the Ciona Embryo. *Curr. Biol.* **28,** 1150-1156.e4 (2018).
27. Matsuoka, T., Ikeda, T., Fujimaki, K. & Satou, Y. Transcriptome dynamics in early embryos of the ascidian, Ciona intestinalis. *Developmental Biology* **384,** 375–385 (2013).
28. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat Biotech* **32,** 1053–1058 (2014).
29. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510,** 363–369 (2014).
30. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* **14,** 565–571 (2017).
31. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. (Routledge, 1988).

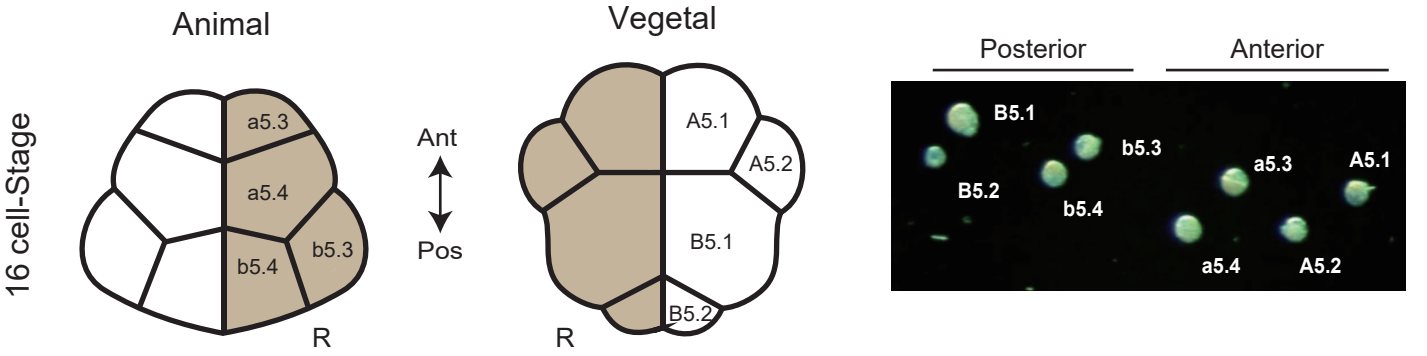
452 32. Bertrand, V., Hudson, C., Caillol, D., Popovici, C. & Lemaire, P. Neural Tissue in
453 Ascidian Embryos Is Induced by FGF9/16/20, Acting via a Combination of Maternal
454 GATA and Ets Transcription Factors. *Cell* **115,** 615–627 (2003).
455 33. Hamaguchi, M., Fujie, M., Noda, T. & Satoh, N. Microarray analysis of zygotic
456 expression of transcription factor genes and cell signaling molecule genes in early
457 Ciona intestinalis embryos. *Development, Growth & Differentiation* **49,** 27–37 (2007).
458 34. Hudson, C. & Yasuo, H. Patterning across the ascidian neural plate by lateral Nodal
459 signalling sources. *Development* **132,** 1199–1210 (2005).
460 35. Imai, K. S., Hino, K., Yagi, K., Satoh, N. & Satou, Y. Gene expression profiles of
461 transcription factors and signaling molecules in the ascidian embryo: towards a
462 comprehensive understanding of gene networks. *Development* **131,** 4047–4058
463 (2004).
464 36. Imai, K. S., Levine, M., Satoh, N. & Satou, Y. Regulatory Blueprint for a Chordate
465 Embryo. *Science* **312,** 1183–1187 (2006).
466 37. Shi, W. & Levine, M. Ephrin signaling establishes asymmetric cell fates in an
467 endomesoderm lineage of the Ciona embryo. *Development* **135,** 931–940 (2008).
468 38. Prodon, F., Yamada, L., Shirae-Kurabayashi, M., Nakamura, Y. & Sasakura, Y.
469 Postplasmic/PEM RNAs: A class of localized maternal mRNAs with multiple roles in
470 cell polarity and development in ascidian embryos. *Dev. Dyn.* **236,** 1698–1715
471 (2007).
472 39. Satou, Y., Kawashima, T., Shoguchi, E., Nakayama, A. & Satoh, N. An Integrated
473 Database of the Ascidian, Ciona intestinalis: Towards Functional Genomics.
474 *Zoological Science* **22,** 837–843 (2005).

40. Hoshino, Z. & Tokioka, T. An unusually robust Ciona from the northeastern coast of Honsyu Island, Japan. *Publ. Seto Mar. Biol. Lab.* **15,** 275–290 (1967).
41. Pennati, R. *et al.* Morphological Differences between Larvae of the Ciona intestinalis Species Complex: Hints for a Valid Taxonomic Definition of Distinct Species. *PLOS ONE* **10,** e0122879 (2015).
42. Tang, F. *et al.* RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat. Protocols* **5,** 516–535 (2010).
43. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Meth* **6,** 377–382 (2009).
44. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9,** 357–359 (2012).
45. Dehal, P. *et al.* The Draft Genome of Ciona intestinalis: Insights into Chordate and Vertebrate Origins. *Science* **298,** 2157–2167 (2002).
46. Satou, Y. *et al.* Improved genome assembly and evidence-based global gene model set for the chordate Ciona intestinalis: new insight into intron and operon populations. *Genome Biology* **9,** R152 (2008).
47. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14,** R36 (2013).
48. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).
49. Wada, S., Katsuyama, Y., Yasugi, S. & Saiga, H. Spatially and temporally regulated expression of the LIM class homeobox gene Hrlim suggests multiple distinct functions in development of the ascidian, Halocynthia roretzi. *Mechanisms of Development* **51,** 115–126 (1995).

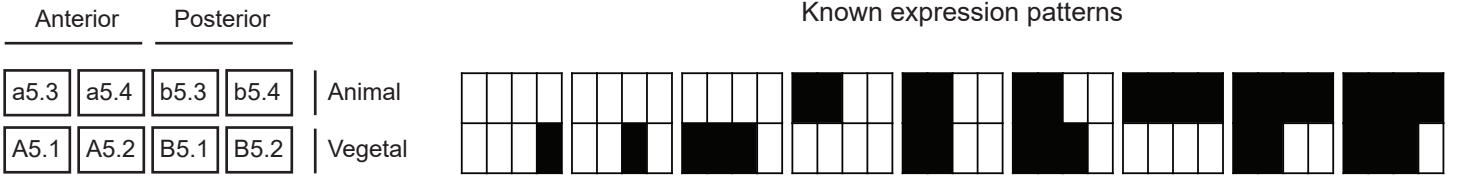
50. Satou, Y. *et al.* A cDNA resource from the basal chordate Ciona intestinalis. *Genesis* **33,** 153–154 (2002).

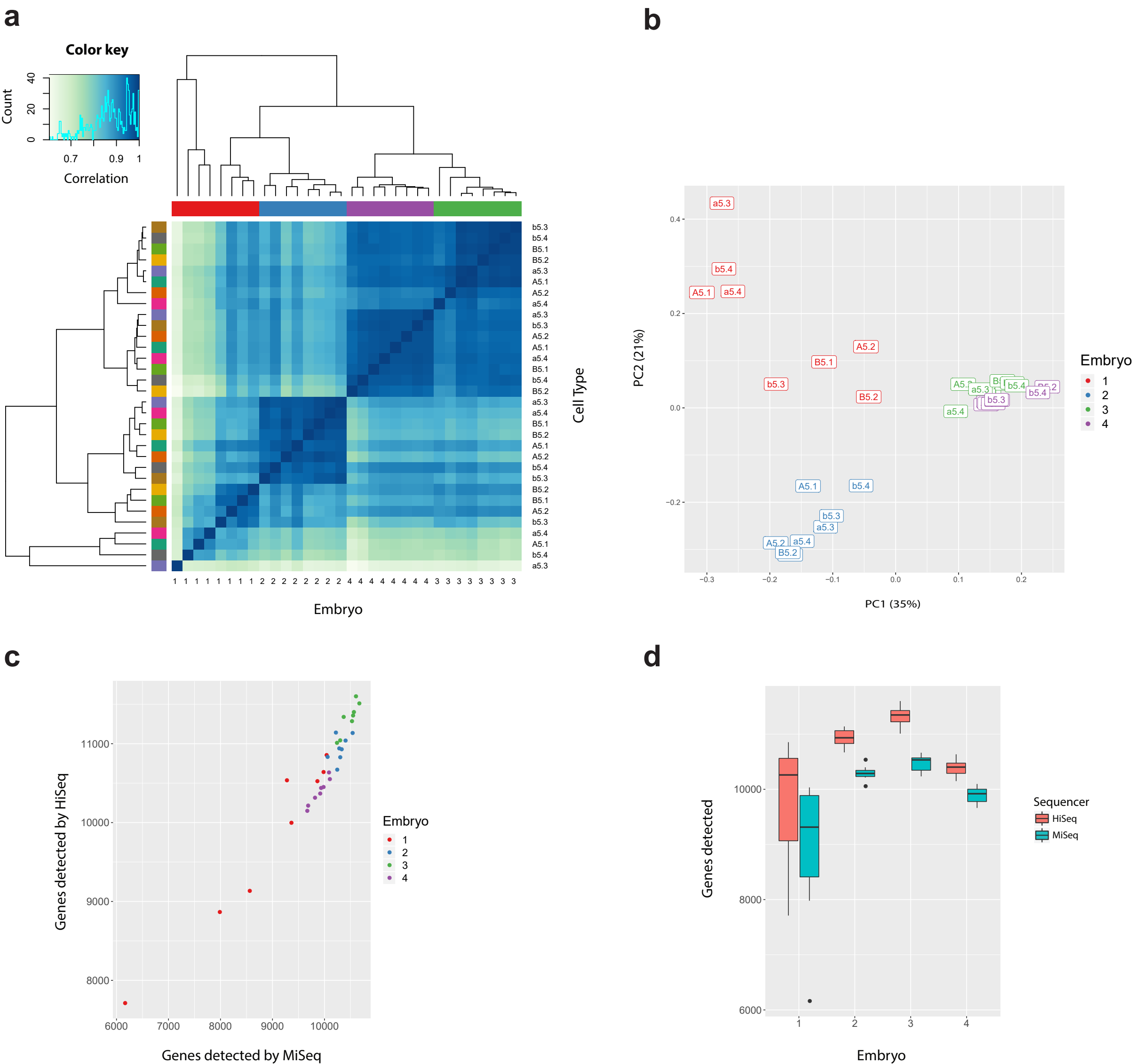
51. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43,** e47–e47 (2015).

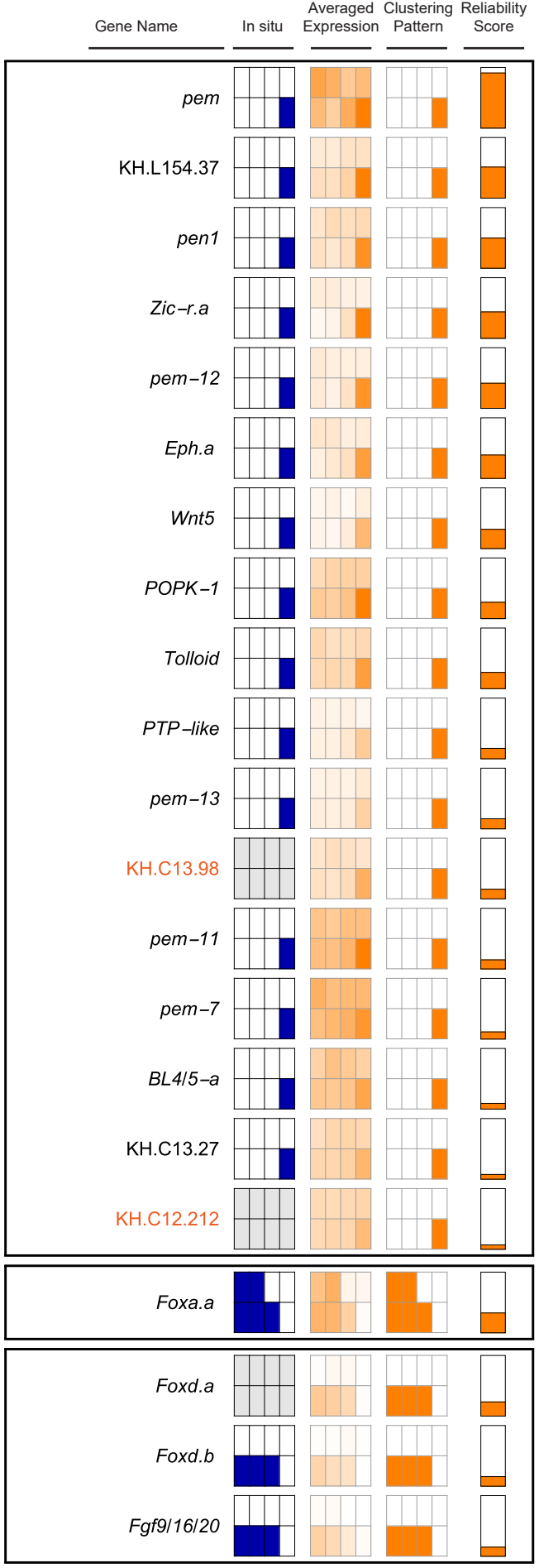
a

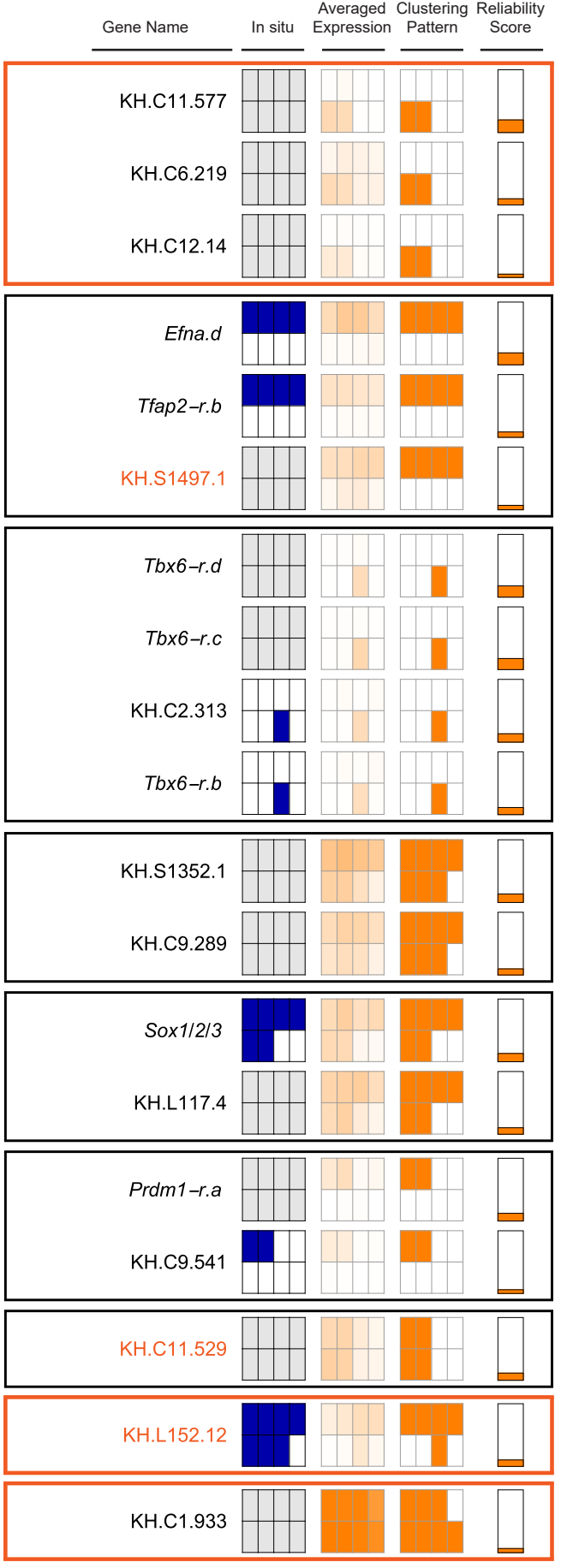

Animal
Vegetal
16 cell-Stage
a5.3
a5.4
b5.3
b5.4
R
Ant
Pos
A5.1
A5.2
B5.1
B5.2
R
Posterior
Anterior
B5.1
b5.3
B5.2
b5.4
a5.3
A5.1
a5.4
A5.2

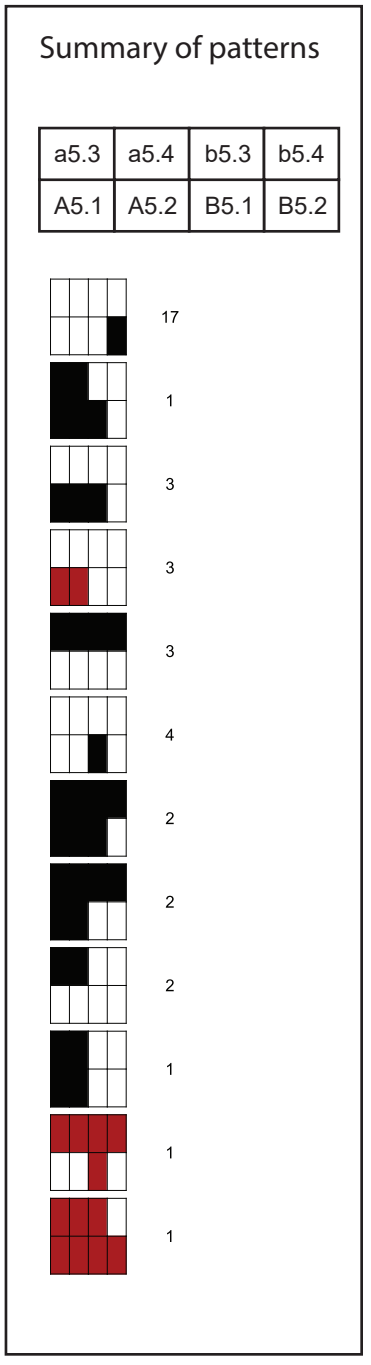
b

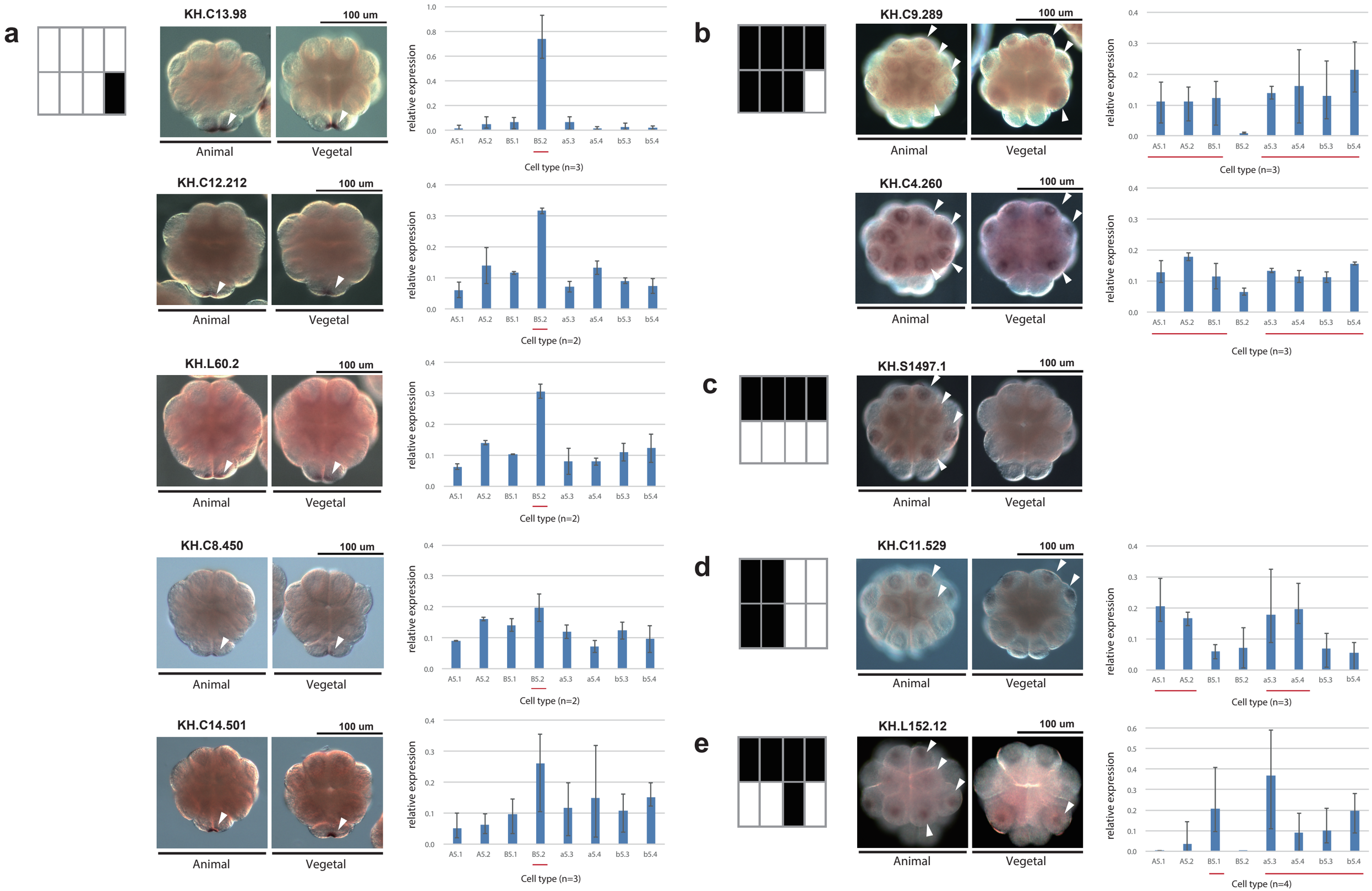

Anterior
Posterior
a5.3
a5.4
b5.3
b5.4
Animal
A5.1
A5.2
B5.1
B5.2
Vegetal
Known expression patterns


a
Color key
Count
Correlation
Cell Type
Embryo
b
PC2 (21%)
PC1 (35%)
Embryo
c
Genes detected by HiSeq
Genes detected by MiSeq
Embryo
d
Genes detected
Embryo
Sequencer
HiSeq
MiSeq


Gene Name
In situ
Averaged Expression
Clustering Pattern
Reliability Score
pem
KH.L154.37
pen1
Zic−r.a
pem−12
Eph.a
Wnt5
POPK−1
Tolloid
PTP−like
pem−13
KH.C13.98
pem−11
pem−7
BL4/5−a
KH.C13.27
KH.C12.212
Foxa.a
Foxd.a
Foxd.b
Fgf9/16/20


Gene Name
In situ
Averaged Expression
Clustering Pattern
Reliability Score
KH.C11.577
KH.C6.219
KH.C12.14
Efna.d
Tfap2−r.b
KH.S1497.1
Tbx6−r.d
Tbx6−r.c
KH.C2.313
Tbx6−r.b
KH.S1352.1
KH.C9.289
Sox1/2/3
KH.L117.4
Prdm1−r.a
KH.C9.541
KH.C11.529
KH.L152.12
KH.C1.933


Summary of patterns
a5.3
a5.4
b5.3
b5.4
A5.1
A5.2
B5.1
B5.2
17
1
3
3
3
4
2
2
2
1
1
1


a
KH.C13.98
100 um
Animal
Vegetal
relative expression
A5.1 A5.2 B5.1 B5.2 a5.3 a5.4 b5.3 b5.4
Cell type (n=3)
KH.C12.212
100 um
Animal
Vegetal
relative expression
A5.1 A5.2 B5.1 B5.2 a5.3 a5.4 b5.3 b5.4
Cell type (n=2)
KH.L60.2
100 um
Animal
Vegetal
relative expression
A5.1 A5.2 B5.1 B5.2 a5.3 a5.4 b5.3 b5.4
Cell type (n=2)
KH.C8.450
100 um
Animal
Vegetal
relative expression
A5.1 A5.2 B5.1 B5.2 a5.3 a5.4 b5.3 b5.4
Cell type (n=2)
KH.C14.501
100 um
Animal
Vegetal
relative expression
A5.1 A5.2 B5.1 B5.2 a5.3 a5.4 b5.3 b5.4
Cell type (n=3)
b
KH.C9.289
100 um
Animal
Vegetal
relative expression
A5.1 A5.2 B5.1 B5.2 a5.3 a5.4 b5.3 b5.4
Cell type (n=3)
KH.C4.260
100 um
Animal
Vegetal
relative expression
A5.1 A5.2 B5.1 B5.2 a5.3 a5.4 b5.3 b5.4
Cell type (n=3)
c
KH.S1497.1
100 um
Animal
Vegetal
d
KH.C11.529
100 um
Animal
Vegetal
relative expression
A5.1 A5.2 B5.1 B5.2 a5.3 a5.4 b5.3 b5.4
Cell type (n=3)
e
KH.L152.12
100 um
Animal
Vegetal
relative expression
A5.1 A5.2 B5.1 B5.2 a5.3 a5.4 b5.3 b5.4
Cell type (n=4)