Identifying developmentally important genes with single-cell RNA-seq from an embryo

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

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

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

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

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

[5] These authors contributed equally to this work.

* Correspondence should be address to garth.ilsley@oist.jp or nicholas.luscombe@oist.jp.

Gene expression studies have typically focused on finding differentially expressed genes or pathways between two or more conditions. More recently, single-cell RNA-seq has been established as a reliable and accessible technique enabling new types of analyses, such as the study of gene expression variation within cell types from cell lines or from relatively similar cells in tissues, organs or tumors. However, although single-cell RNA-seq provides quantitative and comprehensive expression data in a developing embryo, it is not yet clear whether this can replace conventional in situ screens for finding developmentally important genes; moreover, current single-cell data analysis approaches typically cluster cells into types based on a common set of genes or identify more variable or differentially expressed genes using predefined groups of cells, limiting their use for finding genes with novel expression patterns. Here we present a method that comprehensively finds cell-specific patterns of developmentally important regulators directly from single-cell gene expression data of the *Ciona* embryo, a marine chordate. We recover many of the known expression patterns directly from our single-cell RNA-seq data and despite extensive previous screens, we succeed in finding new cell-specific patterns and genes, which we validate by in situ and single-cell qPCR.

One early application of single-cell sequencing has been the study of gene expression variation within cell types such as from cell lines or from relatively similar cells in tissues, organs or tumors[1–7], an analysis not possible with bulk RNA-seq where expression is averaged over thousands of cells. Single-cell data has enabled finer resolution approaches: apparently homogenous groups of cells can be clustered to identify novel and rare subtypes[8–11]. Cells undergoing differentiation at different rates

can be ordered and grouped and cell-to-cell variation underlying differentiation decisions can be studied[12-15].

A further distinct application for single-cell sequencing is to probe the very different and changing (nonterminal) cell types of developing embryos[3,16–22]. An important goal in developmental biology is to identify the relatively few genes controlling the course of development. They are expressed in various, overlapping subsets of cell types and it is the combination of these that gives rise to the multiplicity of cell types. Ideally, we would like to find these key genes and the subsets of cells (the patterns) they are expressed in. However, since these subsets are not known a priori, finding cell-specific patterns from differential expression analysis requires many pairwise comparisons between different groupings of cell types, leading to many false positives from multiple testing. In the eight cells we are considering (the right half of the 16-cell embryo), there are 127 pairwise comparisons required: specifically, there are eight possible comparisons for one cell type against seven; 28 comparisons for two cell types against six and so on. At the 32-cell stage, more than 2 billion comparisons would be required. Moreover, the increased number of false positives does not result in many true positives since most of the pairwise comparisons do not correspond to future lineage or cell fate decisions, and when they do, only a few key genes will be specifically expressed. Hence, the methods to date have focused on comparisons between a few embryo stages or lineages, or have looked for genes that express more heterogeneously within the early (2- or 4-cell) stages of the embryo[16,17].

Cell clustering approaches don't address this problem either. If cells are clustered based on the expression of either all genes or the most variable genes, cells from the same embryo, development stage or batch tend to cluster together[3,11,17,21,23] (Figure 1a-b

for our data). Clustering can be more informative regarding cell types if a subset of genes is used, but again the relevant gene subsets vary depending on the cell types being compared. In other words, the choice of genes will predetermine what cell type differences can be resolved.

Given these limitations, the standard approach to finding developmentally important genes still requires extensive use of in situ hybridization assays, applied to a subset of genes selected by genomic techniques, for example, genes with sequence similarity to known developmental regulators in other animals or candidate genes from a whole-embryo differential expression analysis.

A pattern discovery method

To find developmentally important patterns directly from our single-cell sequencing data of *Ciona* embryos, we developed a method that can scale to many cells. *Ciona* develops according to a stereotyped or invariant lineage[24–26], with zygotic expression beginning around the 8-cell stage[27]. This allowed us to collect precisely defined replicates of all eight cell types of the right half of the embryo at the 16-cell stage of *Ciona*, which has comprehensive in situ data and many known gene expression patterns[28] as well as microarray data at cellular resolution from the pooling of large numbers of single cells[23].

First, we checked the quality of our data and produced counts for each gene with four replicates per cell type (Online methods, Figure 1c-e and Supplementary Table 3). Although other transformations are possible, our method begins with a simple transform to the counts, which has the advantage of being easy to interpret. Since our primary interest was to study how genes change 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. This 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. The arcsine of the square root is a suitable transformation of proportions, so we use

$$\varphi_i = 2\ arcsin\left(\sqrt{\frac{k_i}{N}}\right)$$

where k_i is the count for the ith gene and N the total number of counts ($\sum k_i$). The difference between these transformed values measured in two different conditions can be interpreted as an effect size for proportions[29], namely Cohen's h, but it is worth noting that a square root transformation of the proportion (or normalized count) performs equivalently.

Unlike general cell clustering approaches, which seek to find a classification of cell types, the first step of our method is to cluster the cells separately for every gene, which leads to putative gene expression patterns. We take the simplest approach, which is to assume that expression can be classified into two classes i.e. high and low expression. In our implementation, we use single-linkage hierarchical clustering of Euclidean distance between vectors of replicates. The resulting top-level clusters of ON and OFF then determine the relevant pairwise comparison on a per gene basis.

The next step in differential expression analysis is to rank genes, with the p-value being a common choice for parametric tests. Parametric tests typically require an estimate of variance for each gene that incorporates information (shrinkage) from many genes[30,31]. However, as discussed below, there are problems with this approach when the appropriate groups for comparison vary by gene and are not known a priori. Therefore,

we take a more direct approach, which does not require parameter estimation, but nevertheless ranks genes by how well the two classes (ON and OFF) separate. One approach to calculating cluster separation is to rank genes by the difference between the lowest expressing cell in the ON cluster and the highest expressing cell in the OFF cluster, but this approach is sensitive to outliers. Since we have transformed values, the φs, we could rank according to the difference in the mean φ between each cluster, but an important objective in differential expression analysis is not only to downrank ubiquitous or low expression, but also differences associated with higher variability. For these reasons, we calculate our cluster reliability score as the difference between the first quartile of the ON cluster and the third quartile of the OFF cluster, which we call the Transquartile Range (TQR). The TQR is larger when the difference in cluster means is larger, but it penalizes higher variation for a given difference in means. Further, the p-value in general is strongly affected by sample size (in this case, the number of cells being compared), whereas the TQR is relatively robust to outliers and different sample sizes, making it a suitable choice for comparisons across different patterns without requiring parameter estimation.

Discovery of cell-specific gene expression patterns

The dataset we generated consists of single-cell RNA-seq measurements from all eight cells of the right half of four embryos from different individuals fertilized on different days (Supplementary Table 2). We applied our method to this dataset to search comprehensively for cell-specific gene expression patterns. We selected the 40 most reliable cell-specific genes, and found these generated 12 distinct patterns (Figure 2 and Figure 4a), which include nine of the ten currently known patterns[23,32–37], and 25 genes

with known in situ patterns. The missing pattern is for a single exemplar gene, AP-2-like2, which does not show consistent expression across embryos in one of the cells, A5.2[23,35]. Our result is in agreement with the average over many embryos as measured by microarray[23]. Thus, our approach demonstrates the power of single-cell RNA-seq surveys for finding developmentally relevant genes without extensive in situ screens, an approach which offers great potential for studying organisms that do not have the same experimental heritage as *Ciona*.

Out of the 12 patterns we found, the pattern with the most genes was for specific expression in the B5.2 cell type, which is also the most frequent pattern in known in situ patterns (i.e. postplasmic/PEM RNAs[38]). The majority of our results for B5.2 are confirmed by previous in situ datasets. Despite extensive previous screens, we identified new B5.2-specific genes, such as KH.C13.98 and KH.C12.212, confirming their expression by in situ and single-cell qPCR (Figure 3a). We also looked at further B5.2-specific genes outside of the top 40 and found and validated additional genes, such as KH.C8.450 and KH.L60.2, thus demonstrating the value of single-cell RNA-seq for finding developmentally important genes.

Knowing the full range of patterns is important for understanding the progressive specification of cell fate in the early embryo. From our study, we found three potentially new patterns, highlighted in red in Figure 4a, one of which was validated by in situ and single-cell qPCR, namely KH.L152.12 (Figure 3e). We also validated further uncharacterized genes, namely KH.S1497.1, which expresses specifically in the animal hemisphere, and KH.C11.529 on the anterior side (Figure 3c-d).

In addition, we looked more widely in the top 60 results (Supplementary Figure 1), validating new genes, KH.C9.289 and KH.C4.260, by single-cell qPCR and in situ

hybridization (Figure 3b). These are expressed in all cells—except B5.2, a pattern known previously from Hes-a[39]. These results open up avenues for further research into developmental patterning in *Ciona*.

Comparison with known in situ expression patterns

Looking at it in the other direction and comparing our results to 76 genes with known in situ patterns[23,35,38,40] (Supplementary Table 1), we find that clustering agrees with the known in situ pattern for 38 of the 76 genes (Supplementary Figure 3a). Further, when the results are ranked according to how well the ON or OFF expression clusters separate (see Methods), all the top 34 results match known in situ patterns, with the exception of KH.L152.12. However, as described above, we tested this gene by in situ hybridization and single-cell qPCR, validating our RNA-seq measurement (Figure 3e), which is in agreement with results from gene expression microarrays[23].

In the lower ranked results (Supplementary Figure 3b) it could be argued that the algorithm fails in a few cases because it does not cluster correctly, such as for KH.C3.411 (lefty/antivin) where the assumption of only two levels of expression does not seem to apply. For a few other genes, e.g. KH.C12.589 (DPOZ) and KH.C7.243 (Dll-B), the clustering is correct, but the effect size is small. For a few other genes, no reads were mapped, e.g. KH.C7.407 (SoxF), KH.C9.576 (Fringe 2) and KH.C13.22. However, in most cases where our single-cell RNA-seq does not agree with published in situ patterns, our expression measurements are low or relatively uniform across the eight cells—hence the algorithm functions correctly in attributing lower score to these results.

Thus, the method, in combination with single-cell RNA-seq, is effective in recovering many known patterns. In most cases, the discrepancy between the method

179 results and known in situ patterns occurs because the differences between cells as
180 measured by single-cell RNA-seq is not as dramatic as the equivalent measurements
181 from in situ hybridization. This suggests that in situ hybridization could be more
182 sensitive at detecting differences in expression between cell types because the protocol
183 can be optimized for each gene separately—although in some cases these might be false
184 positives.
185
186 **Comparison with a standard differential expression method**
187 Dispersion or variance estimation is an important aspect of parametric methods of
188 differential expression analysis[30,31]. Many studies include only a few replicates, and
189 hence it is not clear if the observed variance for any specific gene is from an underlying
190 difference in gene regulation, a result of limitations in measurement precision, an
191 aberrant outlier, or biological variation (e.g. across embryos). Information from all genes
192 is therefore used to estimate the within-group variance, where this is often assumed to
193 be related to mean expression level. Using this estimate, it is possible to identify genes
194 that vary more than expected or where the level of expression is significantly different
195 than expected under the null hypothesis of no change in mean expression between
196 groups.
197 However, in the case of a developing embryo, there are many different groupings of
198 cells (the samples) that are relevant depending on the specific genes being considered—
199 and for pattern discovery these are not known in advance. This means that dispersion
200 can't be estimated assuming the same groups for all genes and nor are the dispersion
201 estimation algorithms designed to operate with small numbers of genes, as is the case,
202 for example, if a subset of genes is chosen in advance based on the clustering pattern.

Suitable extensions or strategies could deal with this, but at present it is reasonable to assume that parametric methods based on dispersion estimation might not work across different pattern subsets from the same dataset; that is, p-values will not be meaningfully comparable across patterns, thus limiting their applicability to pattern discovery. By contrast, our method takes a more direct approach that does not require parametric estimation nor assumptions regarding the source of variation. A further advantage is that the results are not affected by the proportions of the different patterns, including when only a few genes belong to a pattern of interest. This latter point is particularly important for early development where a few genes can have a critical impact on cell fate determination. However, our method does not produce p-values nor adjust its score based on the proportion of ON and OFF cells being compared. Thus, it is instructive to see how it performs in practice by comparing our top results with the top results of an approach using a standard differential expression method like DESeq2.

Therefore, we performed an exhaustive differential expression analysis applying DESeq2 to all 127 possible comparisons for eight cells (Figure 4) by estimating dispersion and fitting the DESeq2 negative binomial model for each comparison, with embryo and the cell pattern (ON or OFF) as factors. Unlike our method, which can scale to many more cell types, an exhaustive approach will not be possible in general because of combinatorial explosion—32 cells would require more than two billion DESeq2 comparisons. Nevertheless, it provides a useful baseline for how DESeq2 performs by default— without having a smaller set of patterns as a guide. After applying DESeq2 to all comparisons, we combined the results by selecting, for each gene, the pattern with the lowest adjusted p-value (not adjusting further for the comparison across all patterns) and summarized the resulting patterns of the top 40 genes in Figure 4b, row i.

By comparing this with the result from our method (Figure 2 and Figure 4a), it is clear this approach finds more spurious patterns and fewer known patterns (i.e. seven known patterns compared to nine from our method).

As described above, the first step of our method is to find clusters for each gene. However, using these clusters as predefined comparisons does not significantly reduce the number to be tested by a standard differential expression method. In our data, there are 242 clusters out of a maximum possible 254 when considering both directions (increase and decrease in expression). However, the reduced list from our method (e.g. the 12 patterns in Figure 2 and Figure 4a) can be used to guide further analysis using established methods of differential expression analysis: by limiting the initial DESeq2 comparisons to only these patterns, the number of spurious patterns are reduced. In Figure 4b, row ii, each gene is assigned the pattern with the lowest p-value from all comparisons. If the adjusted p-value is instead chosen from the pattern (or comparison) given by our method, the number of unknown patterns is further reduced and an extra known pattern is found (Figure 4b, row iii), thus demonstrating the value of using the patterns from our method as a guide.

In summary, there are more known patterns in the top results of our method than from an exhaustive application of DESeq2 (nine patterns compared to seven), showing that our method performs well in terms of sensitivity. Also, considering the well-known class of B5.2-specific genes, we find more in our top 40 results than DESeq2: 17 compared to 7 in DESeq2's top 40 results. In both cases, these results largely comprise known B5.2-specific genes, with others validated as above (Figure 2 and Figure 3a). Nevertheless, there could be value in using additional methods, particularly when guided by given patterns, for example DESeq2 identifies further known B5.2 genes using

an adjusted p-value cutoff of 0.01, specifically pem2, Dll-B and midnolin. However, other patterns apparently produce more false positives when using the same threshold (Figure 4c), indicating that further work is required to adapt parametric methods to this type of data. In the meantime, our method offers a scalable and comprehensive approach for finding developmentally important expression patterns in single-cell RNA-seq data.

Conclusion

In conclusion, we have demonstrated that single-cell RNA-seq is a suitable replacement for extensive in situ screens during early embryo development. We recovered many known patterns, as well as new patterns and genes that had not been detected previously despite extensive efforts. This significantly broadens the possibilities for finding the key developmental regulators of less well studied organisms.

Author contributions

NML and NS conceived 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 implemented the pattern discovery method and performed the bioinformatics analysis. GRI, RS and NML wrote the paper and all authors edited and approved the final manuscript.

Code and data access

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

Acknowledgements

We thank the staff of the Maizuru Fisheries Research Station of Kyoto University and Misaki Marine Biological Station of the university of Tokyo for collecting and cultivating *Ciona* under the National BioResource Project (NBRP) of MEXT, Japan, and RIKEN BRC for providing *Ciona* EST clones through the NBRP. We thank Vladimir Benes and Dinko Pavlinic in the Genomics Core Facility at the European Molecular Biology Laboratory (EMBL) for initial advice on the library preparation protocol and the members of the Sequencing Core facility of OIST for their support in running our samples on their Illumina MiSeq and HiSeq machines. We also thank Sylvain Guillot for his technical support and Filipe Tavares-Cadete for early feedback on the method. This work was supported by core funding from OIST to the Genomics & Regulatory Systems and Marine Genomics Units.

290 **References:**

291

292 1. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq
293 experiments. *Nat. Methods* **10,** 1093–1095 (2013).
294 2. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell
295 RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33,**
296 155–160 (2015).
297 3. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-Cell RNA-Seq Reveals
298 Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science* **343,**
299 193–196 (2014).
300 4. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell
301 transcriptomics. *Nat. Methods* **11,** 637–640 (2014).
302 5. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular
303 heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat.*
304 *Biotechnol.* **32,** 1053–1058 (2014).
305 6. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and
306 splicing in immune cells. *Nature* **498,** 236–240 (2013).
307 7. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular
308 variation. *Nature* **510,** 363–369 (2014).
309 8. Björklund, Å. K. *et al.* The heterogeneity of human CD127+ innate lymphoid cells
310 revealed by single-cell RNA sequencing. *Nat. Immunol.* **17,** 451–460 (2016).
311 9. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell
312 types. *Nature* **525,** 251–255 (2015).

10. 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).
11. Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14,** 483–486 (2017).
12. Mojtahedi, M. *et al.* Cell Fate Decision as High-Dimensional Critical State Transition. *PLOS Biol.* **14,** e2000640 (2016).
13. Olsson, A. *et al.* Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537,** 698–702 (2016).
14. 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 Biol.* **14,** e1002585 (2016).
15. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32,** 381–386 (2014).
16. 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).
17. Goolam, M. *et al.* Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell* **165,** 61–74 (2016).
18. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2,** 666–673 (2012).
19. Scialdone, A. *et al.* Resolving early mesoderm diversification through single-cell expression profiling. *Nature* **535,** 289–293 (2016).

20. Tintori, S. C., Osborne Nishimura, E., Golden, P., Lieb, J. D. & Goldstein, B. A Transcriptional Lineage of the Early C. elegans Embryo. *Dev. Cell* **38,** 430–444 (2016).
21. Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500,** 593–597 (2013).
22. 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).
23. Matsuoka, T., Ikeda, T., Fujimaki, K. & Satou, Y. Transcriptome dynamics in early embryos of the ascidian, Ciona intestinalis. *Dev. Biol.* **384,** 375–385 (2013).
24. Conklin, E. G. The organization and cell-lineage of the ascidian egg. *J. Acad. Nat. Sci. Phila.* **13,** 1–119 (1905).
25. Lemaire, P. Unfolding a chordate developmental program, one cell at a time: Invariant cell lineages, short-range inductions and evolutionary plasticity in ascidians. *Dev. Biol.* **332,** 48–60 (2009).
26. Nishida, H. Specification of embryonic axis and mosaic development in ascidians. *Dev. Dyn.* **233,** 1177–1193 (2005).
27. 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).
28. Satou, Y. & Imai, K. S. Gene regulatory systems that control gene expression in the *Ciona* embryo. *Proc. Jpn. Acad. Ser. B* **91,** 33–51 (2015).
29. Cohen, J. Statistical Power Analysis for the Behavioral Sciences. (Routledge, 1988).
30. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).

31. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40,** 4288–4297 (2012).
32. Bertrand, V., Hudson, C., Caillol, D., Popovici, C. & Lemaire, P. Neural Tissue in Ascidian Embryos Is Induced by FGF9/16/20, Acting via a Combination of Maternal GATA and Ets Transcription Factors. *Cell* **115,** 615–627 (2003).
33. Hamaguchi, M., Fujie, M., Noda, T. & Satoh, N. Microarray analysis of zygotic expression of transcription factor genes and cell signaling molecule genes in early Ciona intestinalis embryos. *Dev. Growth Differ.* **49,** 27–37 (2007).
34. Hudson, C. & Yasuo, H. Patterning across the ascidian neural plate by lateral Nodal signalling sources. *Development* **132,** 1199–1210 (2005).
35. Imai, K. S., Hino, K., Yagi, K., Satoh, N. & Satou, Y. Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: towards a comprehensive understanding of gene networks. *Development* **131,** 4047–4058 (2004).
36. Imai, K. S., Levine, M., Satoh, N. & Satou, Y. Regulatory Blueprint for a Chordate Embryo. *Science* **312,** 1183–1187 (2006).
37. Shi, W. & Levine, M. Ephrin signaling establishes asymmetric cell fates in an endomesoderm lineage of the Ciona embryo. *Development* **135,** 931–940 (2008).
38. Prodon, F., Yamada, L., Shirae-Kurabayashi, M., Nakamura, Y. & Sasakura, Y. Postplasmic/PEM RNAs: A class of localized maternal mRNAs with multiple roles in cell polarity and development in ascidian embryos. *Dev. Dyn.* **236,** 1698–1715 (2007).

39. Satou, Y., Kawashima, T., Shoguchi, E., Nakayama, A. & Satoh, N. An Integrated Database of the Ascidian, Ciona intestinalis: Towards Functional Genomics. *Zoolog. Sci.* **22,** 837–843 (2005).

40. Miwata, K. *et al.* Systematic analysis of embryonic expression profiles of zinc finger genes in Ciona intestinalis. *Dev. Biol.* **292,** 546–554 (2006).

41. 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).

42. 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).

43. Tang, F. *et al.* RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat. Protoc.* **5,** 516–535 (2010).

44. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6,** 377–382 (2009).

45. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

46. Dehal, P. *et al.* The Draft Genome of Ciona intestinalis: Insights into Chordate and Vertebrate Origins. *Science* **298,** 2157–2167 (2002).

47. 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 Biol.* **9,** R152 (2008).

48. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

405 49. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-
406 throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).
407 50. External RNA Controls Consortium. Proposed methods for testing and selecting the
408 ERCC external RNA controls. *BMC Genomics* **6,** 150 (2005).
409 51. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21,**
410 1543–1551 (2011).
411 52. Wada, S., Katsuyama, Y., Yasugi, S. & Saiga, H. Spatially and temporally regulated
412 expression of the LIM class homeobox gene Hrlim suggests multiple distinct
413 functions in development of the ascidian, Halocynthia roretzi. *Mech. Dev.* **51,** 115–
414 126 (1995).
415 53. Satou, Y. *et al.* A cDNA resource from the basal chordate Ciona intestinalis. *genesis*
416 **33,** 153–154 (2002).
417 54. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-
418 sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47–e47 (2015).
419

Online 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 so 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 (Figure 1a-d).

Preparation of *Ciona* embryos

Ciona intestinalis type A, recently designated *Ciona robusta* [41,42], 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 feeding to induce spawning of the old eggs. After this, the *Ciona* were maintained under

444 constant light to induce oocyte maturation. Eggs and sperm were obtained surgically
445 from the gonoducts. Embryos were dechorionated after insemination using a solution of
446 0.07% actinase and 1.3% sodium thioglycolate. Eggs were reared to reach the 16-cell
447 stage in Millipore-filtered seawater (MFSW) at about 18 °C. Embryos from each
448 insemination batch were kept to check the ratio that developed into morphologically
449 normal tailbud. We only used embryos from batches where more than 70% developed
450 normally to tailbud (10 hours post fertilization at 18 degrees) (see Supplementary Table
451 2 for embryo batch information).
452
453 **Naming of cells**
454 In *Ciona*, cells are named using the nomenclature of Conklin[24]: the animal side is
455 prefixed with a lowercase letter (a or b) and the vegetal with an uppercase letter; the
456 anterior with A or a and the posterior with B or b. The initial letter is followed by a
457 number that indicates the embryo stage since fertilization, with individual cells
458 numbered according to their lineage. At the 16-cell stage, the animal domain
459 corresponds to a5.3, a5.4, b5.3 and b5.4, the vegetal domain to A5.1, A5.2, B5.1 and B5.2,
460 and postplasmic RNAs are localized to B5.2.
461
462 **Isolation of single cells at the 16-cell stage**
463 At a defined point in development of the 16-cell embryo i.e., at the stage
464 immediately after compaction of the embryo (2.5 ~ 2.6 hours post fertilization), the
465 embryo was transferred to 4°C to slow its development. Each blastomere was isolated
466 with a fine glass needle in a mannitol solution (0.77 M mannitol : MFSW, 9:1) under a
467 stereo microscope at 4 °C regulated by a thermo plate (Tokai Hit Co., Ltd.) and its

identity noted. Isolated blastomeres were picked up and transferred immediately with a mouth pipet into a lysis buffer[43] for reverse transcription.

Library preparation

We followed the single-cell library preparation method of Tang et al[43,44] with some modification. We added ERCC spike-in RNA (Thermo Fisher scientific, 4456740, 1:80000) to each lysis buffer and applied 14 and then 9 cycles of PCR amplification after second strand synthesis. Amplified cDNA was purified with MinElute PCR Purification kit (28006, QIAGEN) and QIAquick PCR Purification Kit (28106, QIAGEN) after each PCR reaction respectively and its concentration measured with Qubit® 2.0 Fluorometer (Q32866, Life Technologies) to have more than 150 ng total yield of cDNA. The quality of the amplified cDNA and distribution of DNA fragment size were confirmed by Agilent 2100 Bioanalyzer (Agilent Technologies) with High Sensitivity DNA Kit (5067-4626, Agilent) to consist mainly of 1.0-1.5 kb fragments.

Amplified cDNAs were sheared using sonication Covaris S2 System to produce DNA of 300 bp on average. The settings were as follows: Duty cycle: 20%, Intensity: 5, Cycles per burst: 200, Power mode - Frequency sweeping, Treatment time: 90 seconds, Temperature: 12°C.

NEB Next® ChIP-Seq Library Prep Master Mix Set for Illumina® (E6240, NEB) was applied to sheared cDNA for preparation of the library for the Illumina platform**.** NEBNext® Multiplex Oligos for Illumina (E7335, E7500, NEB Next Multiplex Oligos for Illumina, NEB) were combined to introduce an index and adaptor to the double-stranded DNA. After extraction of the 300 bp fraction of adaptor-ligated DNA by E-Gel

Size Select 2% Agarose (G661002, Invitrogen), DNA was amplified with individual index primers using PCR with 19 cycles.

The amplified DNA fragment composition was purified with Agencourt AMPure XP twice (A63881, Beckman) and again checked by Qubit (> 60 ng of cDNA in total yield) and by Bioanalyzer to ensure that the fragment size was sharply distributed around 300 bp (on average, about 320 bp with a standard deviation of 40). The concentration of fragments with appropriate index adapters was quantified by KAPA Library Quantification Kits (KAPA Library Quantification Kits, Illumina GA/Universal, KK4825, Genetics) to ensure that the final libraries had adapters for both ends and their concentration was at least 20 pM.

Data generation and quality checking

Libraries were sequenced on Illumina's (San Diego, CA) MiSeq benchtop sequencer and Illumina HiSeq 2500. Libraries were prepared with different index primers and sequenced on MiSeq using paired 150 nt reads (No. MS-102-2002, MiSeq Reagent Kit v2) with eight multiplexed samples per run with the standard Illumina protocols. The same libraries were sequenced on an Illumina HiSeq 2500 with 150 bp paired end reads (No. PE-402-4001 and FC-402-4001, TruSeq Rapid Cluster - Paired-End and SBS Kits) with 16 multiplexed samples per lane following standard Illumina protocols. Our results from using HiSeq and MiSeq were similar (Figure 1c-d, cf. Supplementary Figures 1 and 2).

The resulting reads were aligned using Bowtie[45] version 2.2.6 to the *Ciona* KH genome assembly[46,47], downloaded from Ghost (http://ghost.zool.kyoto-u.ac.jp/download_kh.html). Reads were mapped using local

alignment (--local), with other settings at their default. We did not trim or filter reads, but instead made use of local alignment to find the optimal match. This had the additional benefit that we did not need to split up reads to handle transcripts spanning more than one intron, as is done, for example, in TopHat[48]. Gene counts were calculated from the resulting alignment files using htseq-count[49] with the non-stranded option and mode "intersection-nonempty" against the KH gene models (version 2013) downloaded from Ghost.

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

Assessment of expression data 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 (mostly above 0.8 for embryos 2, 3 and 4). They are, in fact, more similar to each other than the same cell types are across different individuals (Figure 1a-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[23]. 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 unnormalized counts from the

two platforms is over 99% for every cell type, whether 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[5,7].

This embryo batch effect is further demonstrated by a Principal Components Analysis (Figure 1b), 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 experimental measurements are reliable and reproducible within each batch (or embryo). A confirmation of the reproducibility of our results is the tight distribution of genes detected across samples within embryos (Figure 1c-d). Genes were considered detected when the measured count was greater than zero. These results show that slightly more genes were detected on HiSeq than MiSeq, but that the median difference for each embryo is less than 10%. This is comparable with a previous result showing a reduction of genes detected of around 39% when lowering sequence coverage to less than a million reads per cell[5]. As before, embryo 1 showed more variability across samples than the other embryos.

We also made use of ERCC spike-in controls[50,51] to assess the quality of our library preparation, including the steps of reverse transcription and PCR amplification by comparing the measured counts with known spiked-in mRNA concentrations. We added the spike-in at a low concentration (1:80,000 dilution), and yet found good agreement between the known spike-in concentrations and expression measurements. To assess this, we regressed, with no intercept, the square root of the unnormalized counts against

the square root of the known spike-in concentrations. The resulting R^2 value was greater than 85% for every cell in embryos 2, 3 and 4 (Figure 1e). The poorer fit for the spike-ins of embryo 1 also reveals that the somewhat anomalous expression measurements of embryo 1 likely result from the library preparation step, particularly since PCR amplification produced less RNA from most cells of this embryo compared to other embryos.

A further validation of our data is a comparison of our results with previously published data for the 16-cell stage that was generated using gene expression microarrays[23]. We found good agreement with the key genes analyzed in the associated paper (Supplementary Figure 4).

Pattern discovery

Hierarchical clustering to determine candidate patterns was performed with ClusteringComponents in Mathematica 10.4 with the Agglomerate method and Euclidean distance function. This is equivalent to *hclust* in R with the single linkage method. The quantile method used linear interpolation equivalent to type 5 in the R *quantile* function (the hydrologist method).

Single-cell qPCR analysis

cDNA was reverse transcribed from all cells of one embryo per gene replicate using the same protocol we used for single-cell RNA-seq[43,44]. Quantitative PCR was performed using a StepOnePlus PCR machine (Applied Biosystems) with the SYBR green method (No. RR820B, Takara). Each gene was measured with three replicates, except for KH.L152.12, which had four. The qPCR measures for the cell types of each embryo were

scaled between 0 and 1 and then averaged for each cell type across replicates. If there was insufficient target mRNA, it was first amplified using primers covering a wider region of the target gene than those used for single-cell qPCR. Amplification of a specific product in each reaction was confirmed by determining a dissociation curve. The primers for single-cell qPCR analysis are listed in Supplementary Table 4.

In situ hybridization

Whole-mount in situ hybridization was carried out as previously described with minor modification[52]. Dig-labeled antisense RNA probes were synthesized in vitro from cDNAs from the *Ciona* cDNA project[53]. The IDs for the cDNA clones are shown in Supplementary Table 5.

Microarray processing

Previously published microarray data[23] was processed with the limma R package[54]. Background was corrected using *normexp* and arrays were normalized with the *quantile* method.

Gene models and names

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

609 **Differential expression analysis**

610 The DESeq2 package from R was used for differential expression analysis. A
611 DESeqDataSet was created from the matrix of counts. The DESeq function was used with
612 default values. The design formula included the embryo and the cells' grouping (ON or
613 OFF) for the relevant pattern.

614 **Figure Legends:**

615

616 **Figure 1.** (**a and b**) Gene expression is more similar between cells of the same embryo
617 or batch than between cell types across batches. (**a**) Clustered heatmap of the
618 correlation matrix of transformed expression data (φ) from HiSeq samples (excluding
619 ERCC counts and genes with zero counts), with the histogram in the top left providing
620 the color key. (**b**) PCA plot showing the first two components, which explain 56% of the
621 total variance. (**c and d**) The number of genes detected is consistent across the four
622 embryo replicates whether the libraries were sequenced on MiSeq or HiSeq. (**c**) Scatter
623 plot showing the consistent relationship between MiSeq and HiSeq, with more zeros or
624 undetected genes for some cells of embryo 1 compared to the others. (**d**) Boxplots
625 showing the narrower distribution of genes detected for embryos 2 to 4 compared to
626 embryo 1 and the consistent increase from MiSeq to HiSeq. (**e**) R-squared values
627 resulting from linear regression of the square root of unnormalized counts from the
628 HiSeq data against the square root of known concentrations as the independent
629 variable, with no intercept term.

630

631 **Figure 2.** All previously known patterns occur in the top 40 genes when ranked
632 according to their Transquartile Range (TQR). (**a**) Schematic of the eight cell types
633 showing their arrangement in the expression summary plots in (b). (**b**) Expression
634 summary plots of the top 40 genes, grouped by pattern, with a red border indicating an
635 unknown pattern. A summary of the patterns is shown in Figure 4a. Genes with
636 previously uncharacterized, but now validated patterns are highlighted in red. For each
637 gene, the columns indicate the gene name, any previously known in situ pattern in blue

(gray 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.

Figure 3. Pattern discovery results are validated by in situ hybridization and single-cell qPCR. (**a-e**) For each pattern being tested, a schematic (left) indicates the expected pattern of expression using the layout of Figure 2a. The photomicrographs show the results of situ hybridizations (middle) viewed from the animal and vegetal side. 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 was measured by single-cell qPCR. The qPCR measures for the cell types of each embryo were scaled between 0 and 1 and then averaged for each cell type across replicates. The means for each cell type are shown in the bar charts (right).

Figure 4. Standard differential expression analysis finds many false positives with fewer true positives in the top results than when ranking by Transquartile Range. (**a**) The patterns of the top 40 genes are shown with the number of associated genes below. Black indicates a known pattern, and red a novel or spurious pattern. The layout for each pattern is shown in the key on the right. (**b**) The top 40 genes and their corresponding patterns from different approaches. (**i**) Exhaustively running DESeq2 against all 127 possible comparisons and selecting, for each gene, the pattern with the lowest adjusted p-value. The maximum adjusted p-value for this set is 2.6×10^{-8}. (**ii**) Running DESeq2 separately for each of the patterns from (a) and selecting, for each gene, the pattern with the lowest adjusted p-value. DESeq2 finds genes up- and down-

regulated and hence the pattern for B5.2 and its complement are part of the same run, leading to 11 runs in total. Similarly, for some genes, DESeq2 selects a pattern complementary to the given 12 patterns. (**iii**) DESeq2 against the 12 patterns shown in (a), but using the adjusted p-value for the given pattern. (**c**) DESeq2 results for a selection of patterns. Each DESeq2 analysis is run for all genes and the top results ranked by adjusted p-value with a cut-off of $p<0.01$. For illustration, a line is drawn showing a stricter threshold, which is equivalent to the highest p-value of the top 40 results found using the approach of row i in (b). The results for each pattern show previously known in situ results as well as averaged expression. For comparison, the genes are only counted once in row i of (b), i.e. for the pattern that gives the lowest p-value.

673 **Supplementary Material**

674

Supplementary Table 1	List of known in situ patterns	SuppTable1.xlsx
Supplementary Table 2	Sample, library preparation and sequencing dates	SuppTable2.xlsx
Supplementary Table 3	Sequencing statistics	SuppTable3.xlsx
Supplementary Table 4	Primers used for qPCR	SuppTable4.xlsx
Supplementary Table 5	Single-cell qPCR measurements	SuppTable5.xlsx
Supplementary Table 6	IDs for the cDNA clones	SuppTable6.xlsx

675

676 **Supplementary Figure 1.** The top 60 results from sequencing our libraries on Illumina
677 HiSeq 2500, with the genes grouped by pattern. For each gene, the columns indicate the
678 gene name, any previously known in situ pattern in blue (gray if not known), the
679 average of the transformed expression values (clipped above 0.05), the pattern resulting
680 from clustering, and finally, the TQR as the reliability score.

681

682 **Supplementary Figure 2.** The top 60 results from sequencing our libraries on Illumina
683 MiSeq, with the genes grouped by pattern. For each gene, the columns are the same as in
684 Supplementary Figure 1.

685

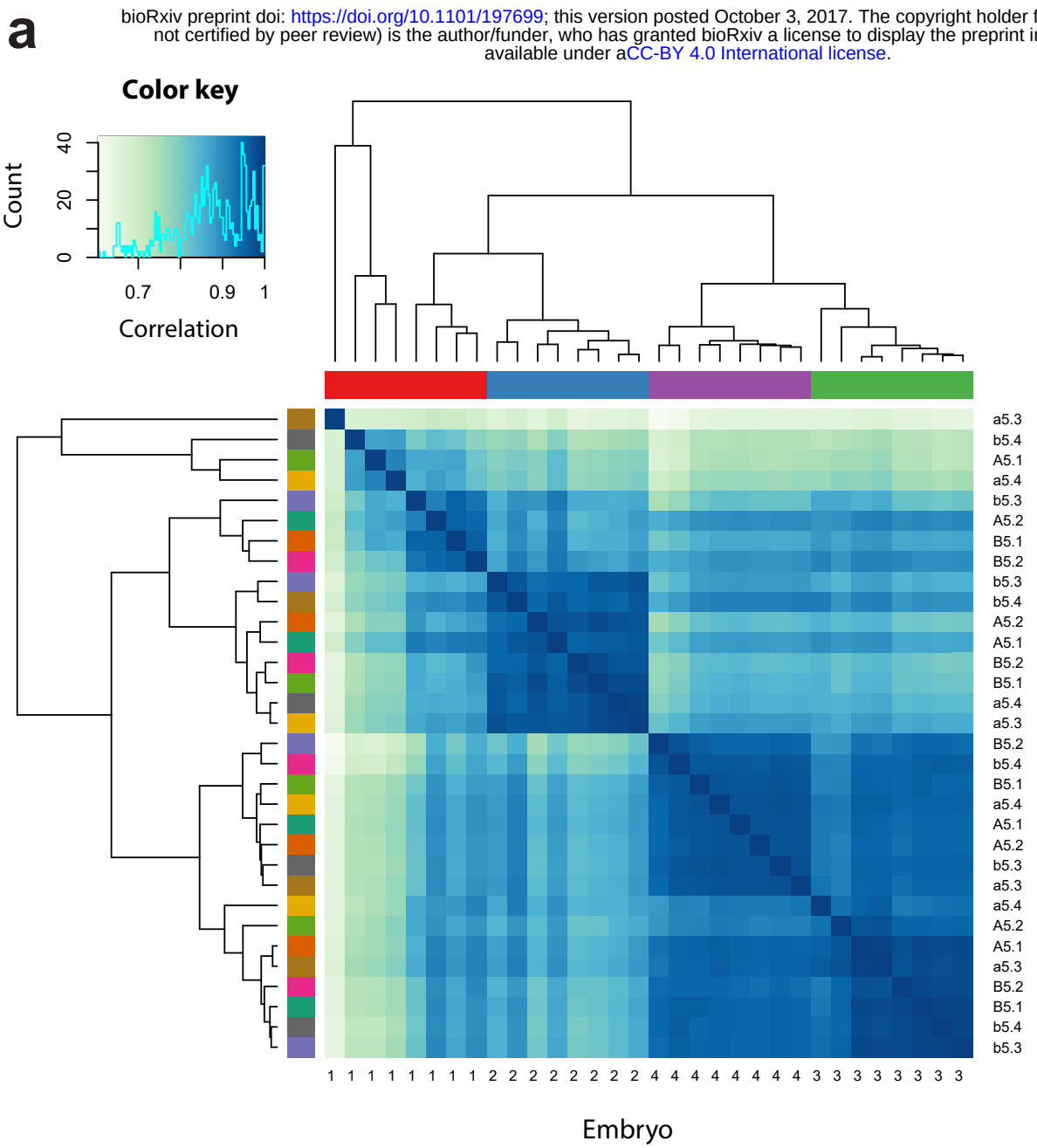
686 **Supplementary Figure 3.** Comparison with known in situ patterns. For each gene, the
687 columns are the same as in Supplementary Figure 1. The results are divided into (**a**) the
688 top 34 results and (**b**) the next 42 results from pattern discovery for genes with known
689 in situ patterns.

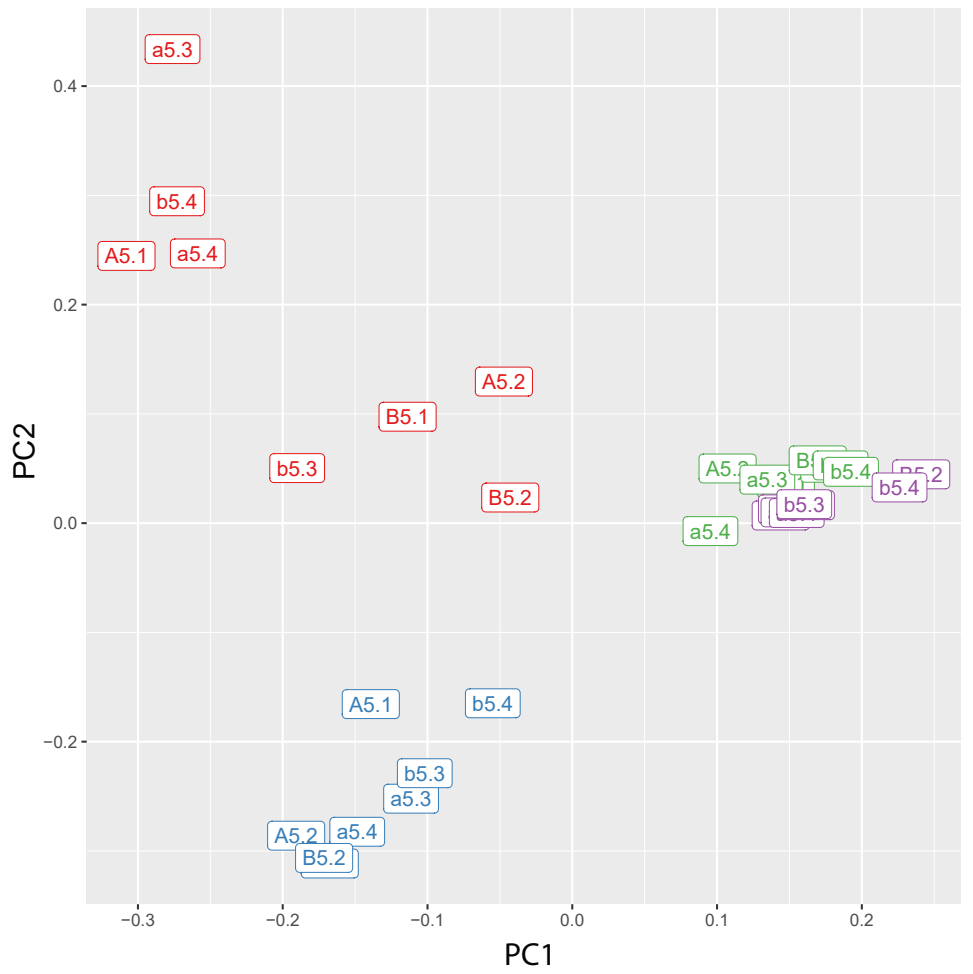
690

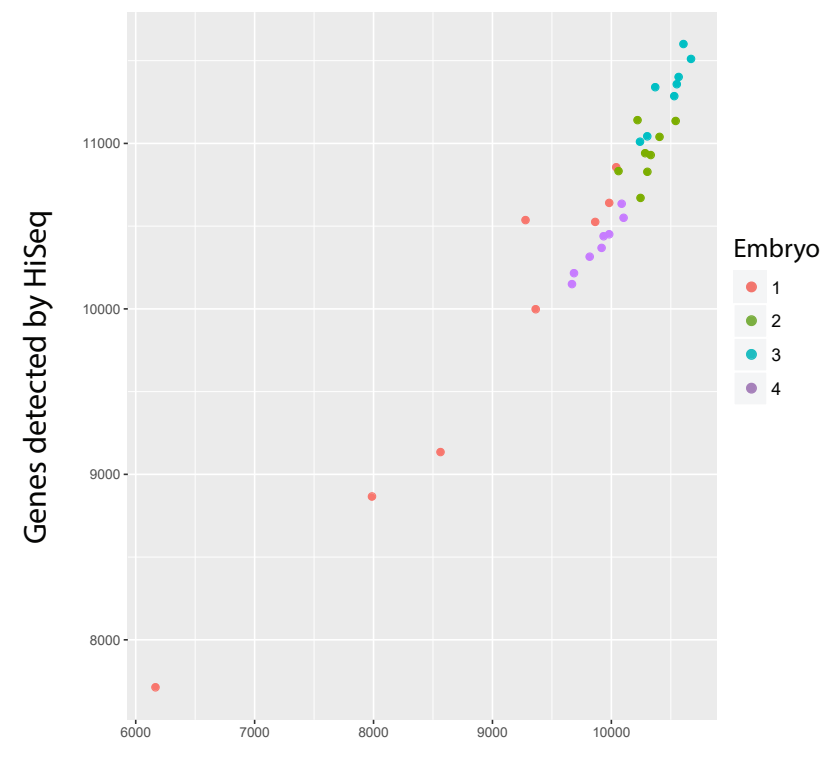
691 **Supplementary Figure 4**

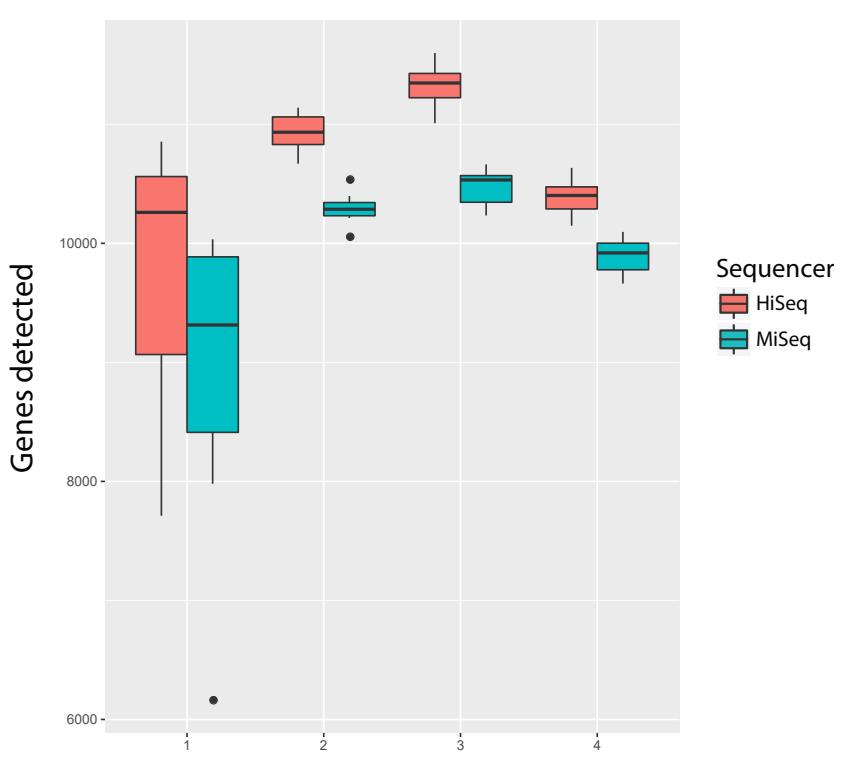
692 Comparison of heatmaps from cellular resolution microarray data [23] with φ-

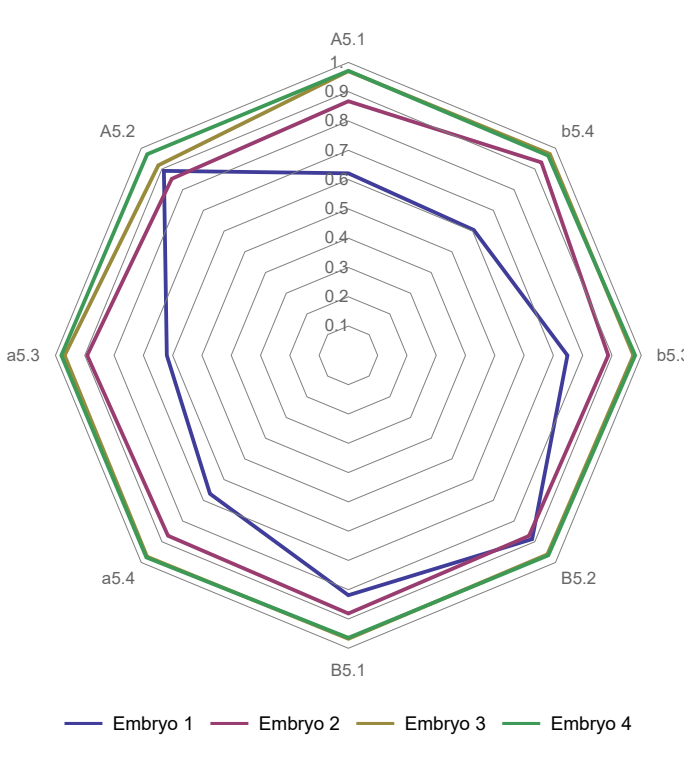
693 transformed single-cell RNA-seq.







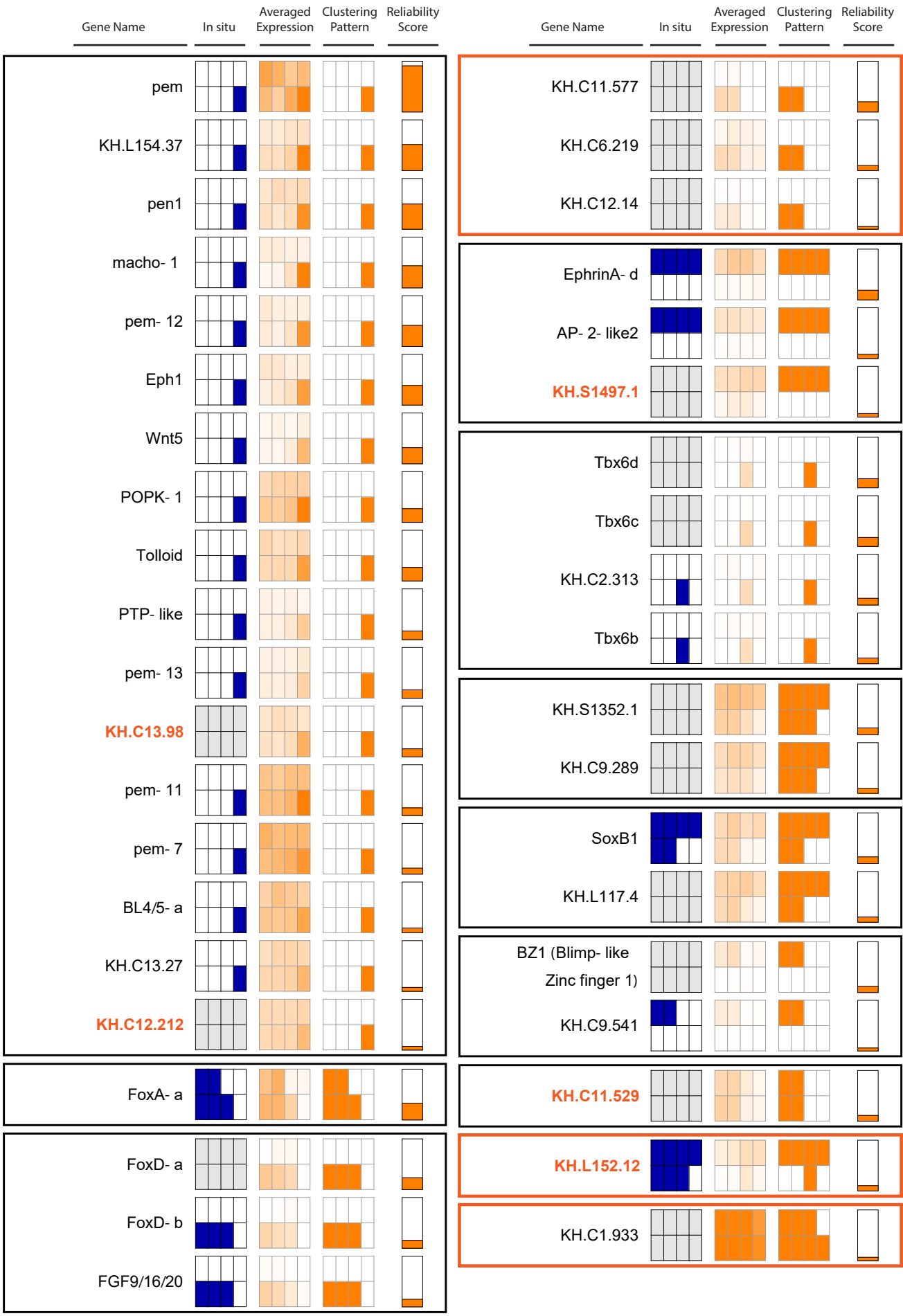




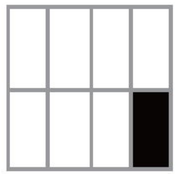
a

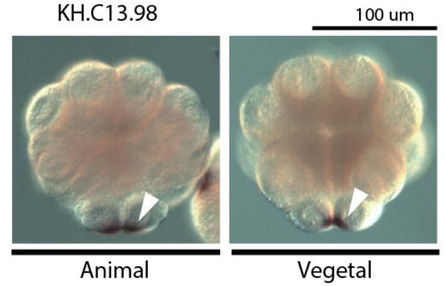
a5.3	a5.4	b5.3	b5.4
A5.1	A5.2	B5.1	B5.2

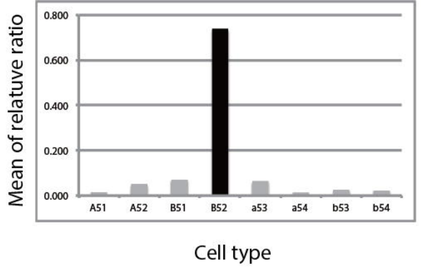
b

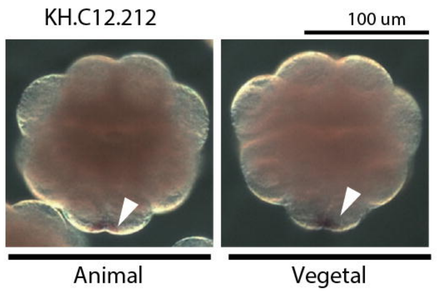


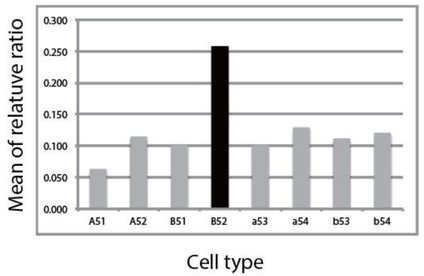
a

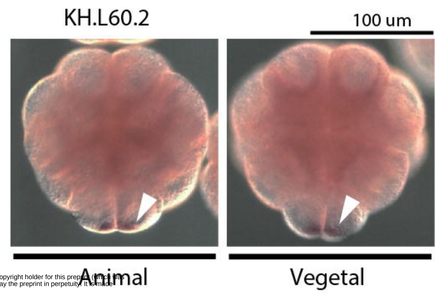


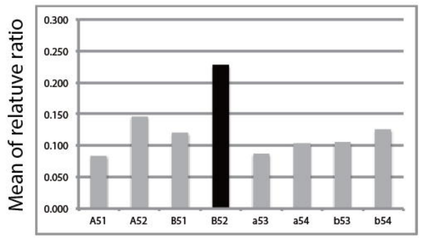




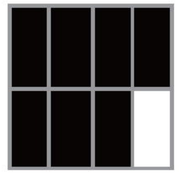


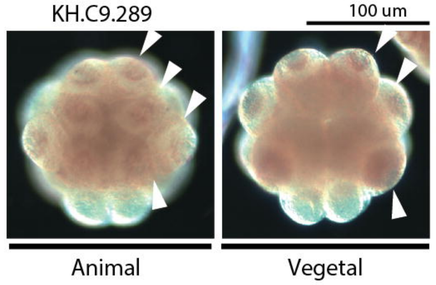


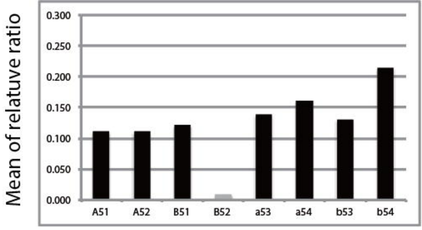


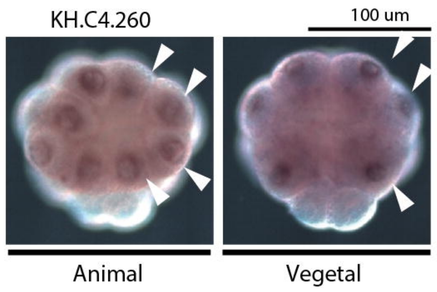


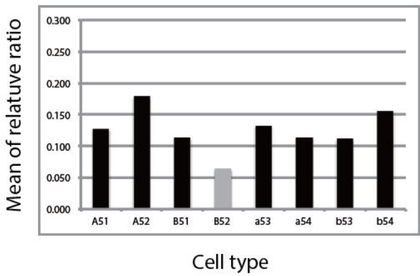
b



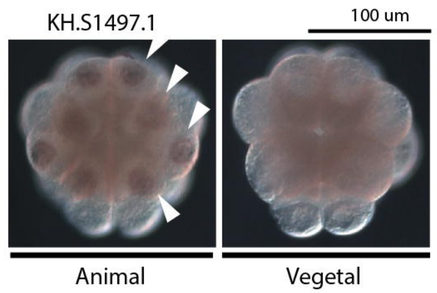


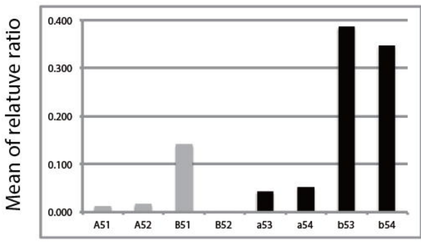


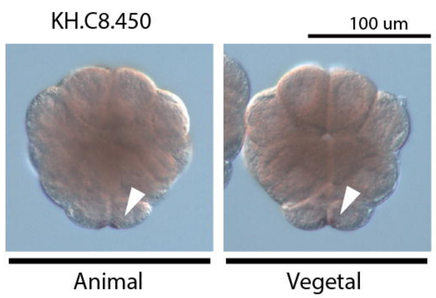


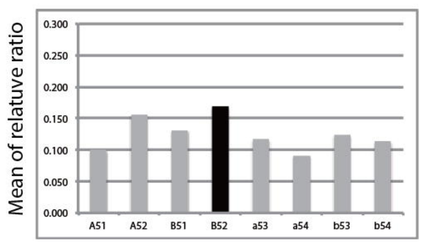


c

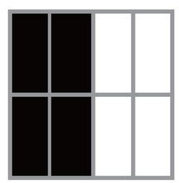


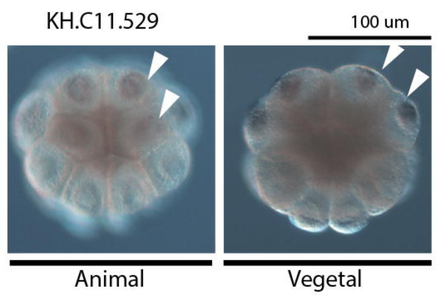


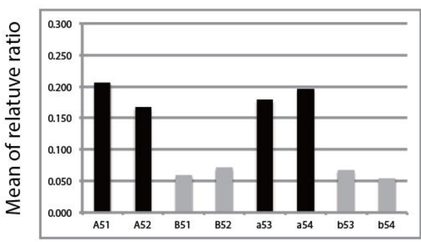


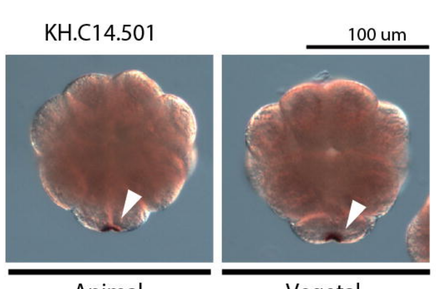


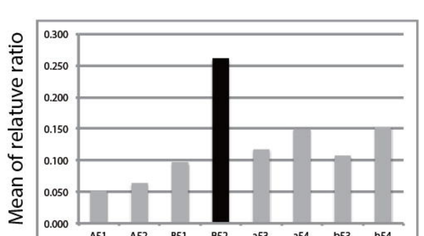
d



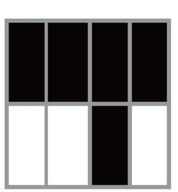


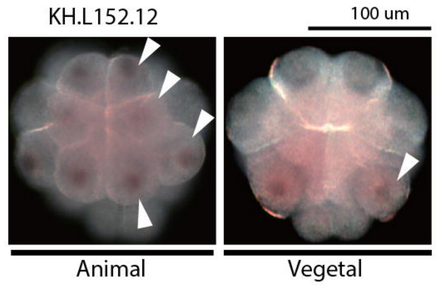


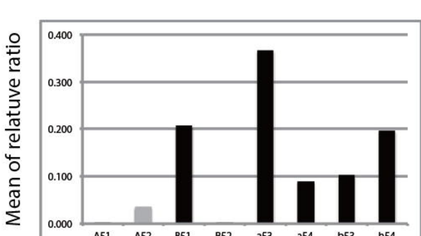




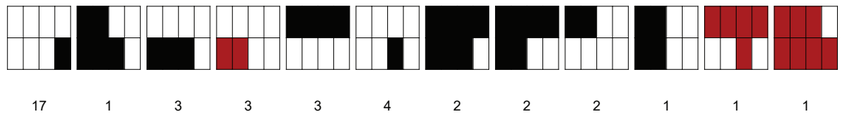
e





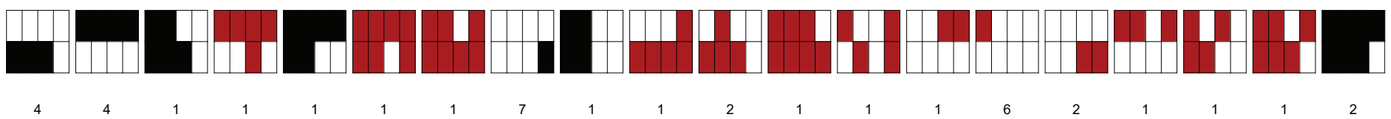


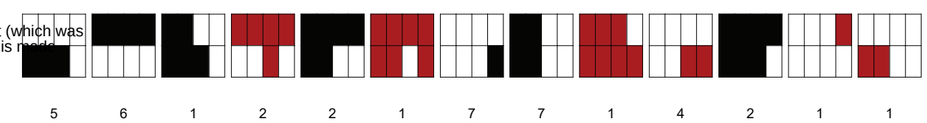
a

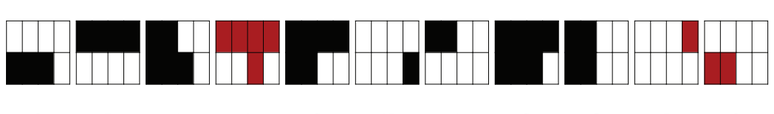


b

	Number of runs	Pattern selection criterion per gene	Highest adjusted p-value	Patterns from the top 40 results
i	127	Lowest p-value	2.6×10^{-8}	
ii	11	Lowest p-value	1.3×10^{-6}	
iii	11	Prescribed by (A)	3.3×10^{-4}	







c