

Physical Modeling of Embryonic Transcriptomes Identifies Collective Modes of Gene Expression

Dominic J. Skinner^{1,2,*}, Patrick Lemaire³, and Madhav Mani^{2,4,†}

¹Center for Computational Biology, *Flatiron Institute*, 162 5th Ave, New York, New York 10010, USA

²NSF-Simons Center for Quantitative Biology, *Northwestern University*, 2205 Tech Drive, Evanston, Illinois 60208, USA

³CRBM, *Université de Montpellier*, CNRS, 34293 Montpellier, France

⁴Department of Engineering Sciences and Applied Mathematics, *Northwestern University*, Evanston, Illinois 60208, USA



(Received 2 August 2024; accepted 29 August 2025; published 29 September 2025)

Starting from one totipotent cell, complex multicellular organisms form through a series of differentiation and morphogenetic events, culminating in a multitude of cell types arranged in a functional and intricate spatial pattern. To do so, cells coordinate with each other, resulting in dynamics which follow a precise developmental trajectory, constraining the space of possible embryo-to-embryo variation. Using recent single-cell sequencing data of early ascidian embryos, we leverage natural variation together with modeling and inference techniques to investigate development at the level of a complete interconnected embryo—an embryonic transcriptome. After developing a robust and biophysically motivated approach to identifying distinct transcriptomic states or cell types, a statistical analysis reveals correlations within embryos and across cell types demonstrating the presence of collective variation. From these intraembryo correlations, we infer minimal networks of cell-cell interactions, which reveal the collective modes of gene expression. Our work demonstrates how the existence and nature of spatial interactions along with the collective modes of expression that they give rise to can be inferred from single-cell gene expression measurements, opening up a wider range of biological questions that can be addressed using sequencing-based modalities.

DOI: [10.1103/7gs5-w48q](https://doi.org/10.1103/7gs5-w48q)

I. INTRODUCTION

Development reliably produces complex and highly structured organisms in the presence of external perturbations and intrinsic stochasticity [1]. Strikingly, the dynamics which give rise to these complex structures are also constrained or canalized in the language of Waddington [2]. Yet phenotypic variation across individuals must be present to facilitate evolution [3]; how the nature of phenotypic variation is shaped by the nature of the dynamical constraints is largely unknown. Single cell gene expression measurements, though limited in quality and an incomplete measurement of a cell's state, provide a high-dimensional cellular phenotype [4] that has led to an avalanche of cell atlases focused on phenotypic identification of cell types across a multitude of species [5–7]. Phenotypic variation, however, exists not solely at the scale of cells but also at the scale of embryos, producing variation in organismal function. Cells within an embryo interact with one another through intricate mechanochemical mechanisms, which correlate the states of cellular gene expression. Much like how the couplings between infinitesimal material

portions of a violin string sustain large-scale collective oscillations, we expect the complex coupling of cells in embryos to give rise to collective modes of gene expression. Far from producing a pure note, we anticipate that the intricate intercellular couplings creates a rich timbre of complex collective modes of gene expression in embryos. These modes represent emergent phenotypes at the multicellular scale beyond individual cellular gene expression. To our knowledge, the present study is the first attempt to identify and characterize these collective modes of embryonic gene expression using a combination of simple physical modeling and careful statistical analyses of embryonic transcriptomic data. In doing so, we aim to contribute to the growing body of work demonstrating how single-cell sequencing data can be leveraged to address a much broader range of biological questions than simply cell-type identification [8–12]. In addition we wish to propose that, much like in condensed-matter systems, the lower-dimensional phenotypic space spanned by the collective modes of gene expression in an embryo are functionally relevant, encoding a space in which developmental dynamics ensue.

We follow the tightly controlled program of the developing ascidian embryo, observed using single-cell RNA sequencing data of individuals at distinct stages of development, starting from the initiation of zygotic transcription. Increasing numbers of genes are transcriptionally activated in precise spatial locations as development ensues, and cells are increasingly coupled through inheritance of transcripts and intercellular signaling. Through statistical means, we find that variation in gene expression is coupled across cells within the same

*Contact author: dskinner@flatironinstitute.org

†Contact author: madhav.mani@gmail.com

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

embryo, even after accounting for differences in staging time and the finiteness of our datasets. In particular, we find this to be the case for zygotic transcripts, in contrast to variation in maternal transcripts, whose variation can be explained by a simple shift in the overall level of gene expression from embryo to embryo. To ascertain the nature of zygotic cell-cell couplings, whether through inheritance of transcripts via cell divisions or intercellular signaling, we infer minimal, or sparse, networks of cell-cell interactions that give rise to the observed complex, or dense, set of cell-cell correlations. To perform this inference, we turn to statistical physics which studies the collective self-organization and fluctuations of systems composed of many individual components, strongly interacting with each other. Specifically, we use a spin-glass-like model [13], following approaches that have proved insightful for the study of complex systems from bird flocks [14] to gene regulatory networks [15,16] and biological neural networks [17]. The inferred network of interactions allows us to quantify and visualize embryowide variation by decomposing this variation into collective modes of gene expression, which are not directly accessible from the single-cell sequencing data. This work represents an attempt to combine data-driven and physical modeling to reveal new perspectives on the developing organism and to analyze single-cell sequencing data at the level of the embryo.

II. BACKGROUND

A. Biological context

In order to ground our goals and approaches in a concrete setting we focus on the early development of the ascidian, or sea squirt, *Phallusia mammillata*. This organism displays several attractive features, which we now discuss, that make it an ideal setting to develop our approach. The ascidian cell lineage is invariant, with the same series of synchronous cell divisions occurring in each embryo [18–20], Fig. 1(a), which also provides a consistent way to name cells across embryos [18]. The physical arrangement, or adjacency matrix, of these cells is conserved and even the geometry, such as cell-cell contact areas, is broadly conserved up to an overall size scaling [23]. The paradigm of cell-cell communication in these embryos is one of immediate neighboring cells interacting through mechanochemical mechanisms. The long-range morphogen signals, which sculpt the dynamics of fly and frog embryos are believed to be largely absent or, at least, a higher-order effect that are subdominant to nearest-neighbor interactions [23]. This local mode of communication, which occurs after zygotic genome activation, follows from the initial localization of a handful of maternal mRNAs and proteins to the posterior and vegetal poles of the zygote, defining the anterior-posterior (head-to-tail) axis. From this relatively unstructured start, and transcriptional inheritance from the individual’s mother, cells self-organize into distinct transcriptional states, precisely arranged in space; there are five such states, referred to as cell types, as early the 16-cell stage [Fig. 1(b)]. The fate map of each cell, the set of terminal cell types its progeny will become, is restricted [18], Fig. 1(a), and thus many of these cells remain diverged in gene expression permanently. As development progresses, and the number of

cells increases through cell divisions, additional cell types appear and increasing numbers of genes become zygotically expressed [22], again, all in a precise spatial arrangement. The period from the onset of embryonic transcription to the point when zygotic transcripts predominate is known as the maternal-to-zygotic transition (MZT). In ascidian, zygotic transcription begins at the 8-cell stage and zygotic transcripts dominate by the 112-cell stage [24]. A multitude of molecular mechanisms are known or suspected to contribute to the embryo’s feat of self-organization including localized maternal mRNA [25], FGF signaling [26], β -catenin gradients [27], and asymmetric mRNA partitioning during cell-division [28]. Major questions about the program of self-organization remain and many would-be insightful experiments, such as a complete space-time recording of live signaling, remain intractable. Yet measuring the gene expression, which is a proxy for the internal “state” of the cell [4], of every cell in an embryo is possible through single-cell sequencing. A central question addressed in this study is whether such measurements can be leveraged to study a broader spectrum of biological phenomena than identification of cell types. As such, the ascidian with its stereotyped cellular topology, geometry, and conserved lineages represents a simpler, but authentic, developmental system where general quantitative and physical approaches to identifying the collective modes of gene expression can be developed.

B. Data

In addition to the developmental considerations that make ascidian an attractive model system, there exists a publicly available and rich sequencing dataset [22]. Single-cell RNA sequencing was performed on individual *P. mammillata* embryos up to the 64-cell stage, with multiple biological replicates at each stage. Here we will only consider the 8- to 64-cell stages. Notably, the sequencing platform used in Ref. [22] was Smart-Seq2 [29], which, together with relatively large cells, results in a sequencing depth higher than typical scRNA experiments [30], with on average 300 reads per gene at the 64-cell stage (SM Table S1 [36]), although we note that these are not unique molecular identifier (UMI) counts. This depth of sequencing presents a unique opportunity to develop novel theoretical and computational frameworks.

The procedure of single-cell RNA sequencing creates some difficulty in asking what are essentially spatially minded questions that motivate our work. Before sequencing, embryos are dissociated into individual cells [Fig. 1(c)], and the location and identity of each cell is lost. Nevertheless, the cell type, and thus approximate spatial location, of each cell can be reconstructed from gene expression, as we will detail later (Fig. 3). During the sequencing procedure, individual cells from the dissociated embryos were transferred to a 96-well plate. Which embryo each cell came from was recorded [Fig. 1(c)], an essential requirement for identifying embryo-to-embryo variation. Embryos were sourced from a number of mothers, and several embryos were typically used from each mother; this information was also recorded. While a careful dissection and sorting of cells presequencing is technically possible [31], it is tedious and impractical for more than a handful of embryos, although a crude sorting of cells

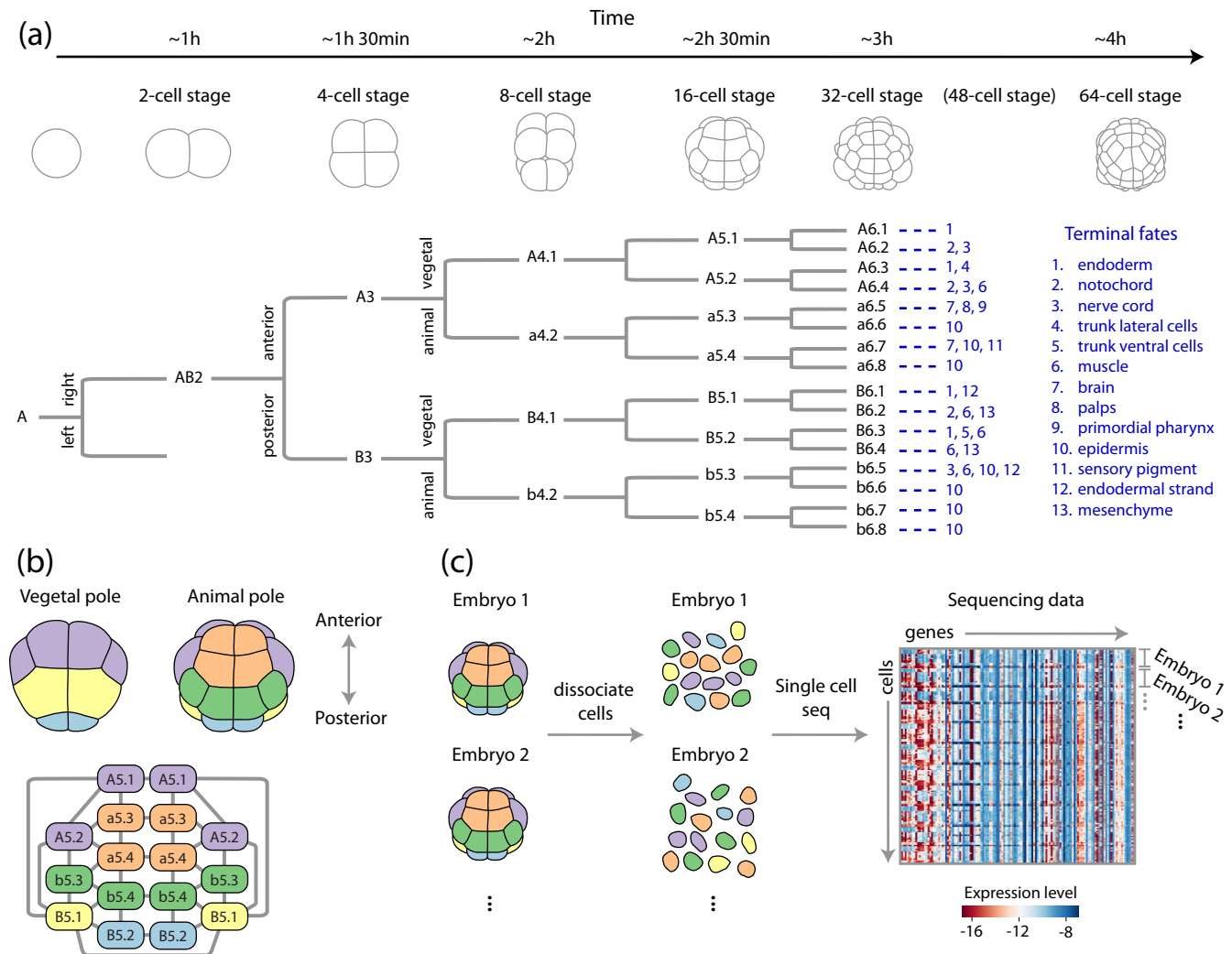


FIG. 1. Developmental dynamics of the ascidian embryo: lineage, geometry, and single-cell gene expression data. (a) Early development of the ascidian *P. mammillata*. Representative morphologies (top) shown along with right side of lineage tree (bottom) with the left being mirror symmetric. Cell fates at the larval stage indicated in blue for each 64-cell stage cell (adapted from Ref. [18]). Time is in hours postfertilization at 18° [21]. (b) Representative cell geometry for the 16-cell stage (top), together with the cell-cell contact graph (bottom). Colors indicate transcriptomically distinct cell states, as identified in Fig. 3, which are also conserved across embryos. The animal-vegetal, anterior-posterior, and left-right axes form the three morphological axes of the embryo. (c) Experimental setup of Ref. [22], where large numbers of embryos at each cell stage are first dissociated before single-cell RNA sequencing at high depth was performed. The precise identity of each cell is lost after dissociation, but the embryonic identity is retained through barcoding [22]. The resultant cell-by-gene expression matrix is shown for the 16-cell stage with 168 cells across 11 embryos and 265 genes which are differentially expressed across the embryo (see Sec. V).

into the left and right side of the embryo would allow for an almost complete spatial reconstruction after cell-typing. Altogether, the combination of high-quality sequencing, a significant number of biological replicates (see Sec. V), and a relatively constrained space of variation due to invariant cell number and geometry make the ascidian an ideal model system for exploring collective variation in development.

C. Roadmap

Having introduced the system, we can now give an overview of our analysis (Fig. 2). Our first major task is to identify cell types and assign each cell to one of them [Fig. 2(a)]. While every cell has a unique position in the embryo, the mRNA composition, or transcriptomic state, is

not unique. For instance a B5.1 cell on the left side of the embryo has the same mRNA composition, on average, as a B5.1 cell on the right side. We refer to this smaller set of distinct states as the cell types. To identify these cell types we cluster the sequencing data, ensuring that clusters respect the structure of the ascidian embryo. We then use prior knowledge of the spatial expression pattern for a handful of marker genes to map these clusters onto the physical embryo, allowing for a partial spatial reconstruction.

Our next task is to determine whether there is any statistical evidence of collective variation or whether noisy single-cell expression along with measurement error dominates. Said another way, could the expression data be explained by each cell's expression being drawn independently from some cell-type-specific probability distribution or must the expression

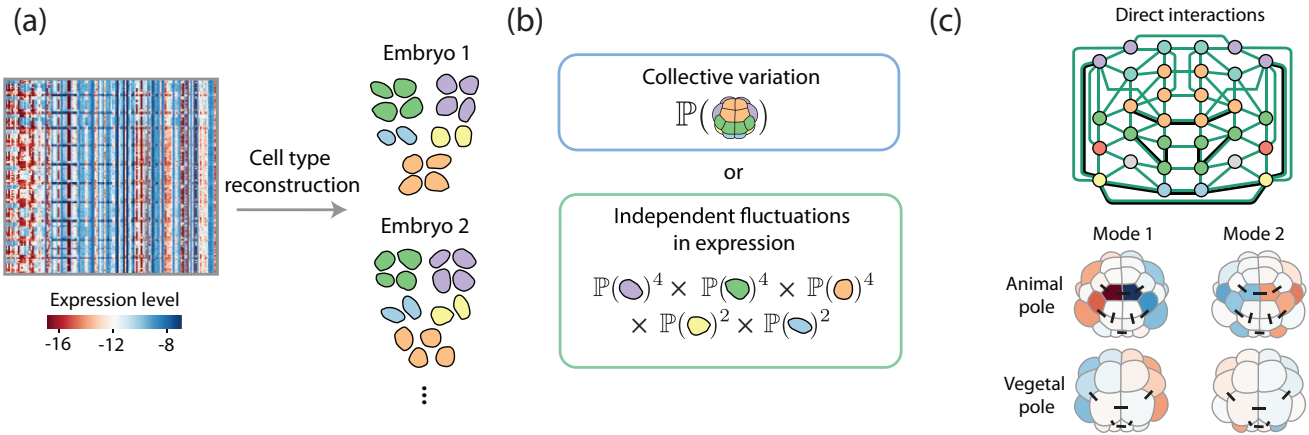


FIG. 2. Overview of identifying collective variation from scRNA-seq data. (a) First we cluster the scRNA-seq expression matrix, identifying each cluster with a specific set of cells through expression of marker genes (see Fig. 3). This partially recovers a cell’s spatial position, but it remains unknown whether, say, a cell came from the left or right of the embryo. Expression matrix same as in Fig. 1(c). (b) Next, we statistically test whether the data could have been generated by independent cells or whether it must have come from a joint distribution. (c) Finally, we fit maximum entropy models to the data, identifying cell-cell interactions and revealing the modes of variation (see Fig. 8).

data of cells in the same embryo come from a joint distribution [Fig. 2(b)]? We do this by carefully considering and accounting for potential sources of variation in the data, such as differences in staging times between embryos, and testing whether shuffling cells of the same cell type between different embryos alters the covariances between cell types. To ensure these conclusions are robust to batch effects, we perform the analysis for both maternal and zygotic mRNAs, finding that maternal mRNAs only vary in a nonspatial way primarily due to mother-specific differences, unlike the zygotic mRNAs. As part of this analysis, we estimate the small number of principal components that are statistically significant and project the zygotic mRNA expression onto these components.

Finally, we derive and fit a model for the probability distribution of zygotic gene expression across an embryo. We take our model as the distribution which maximizes entropy while holding a few key cell-cell covariance statistics fixed, where regularization determines which cell-cell covariance statistics are fixed. We then ask what typical variation looks like in this model. Since our statistical model is ultimately a Gaussian distribution, we can decompose the observed variation into variation along the principal axes of the covariance matrix. We refer to these directions as the collective modes of variation, since variation along these axes couples cells across the embryo [Fig. 2(c)].

III. RESULTS

A. A robust statistical approach to cell-type identification

Before any quantitative analysis, we need to preprocess the data. Starting from the raw transcripts, we align them to a reference genome (see Sec. V), resulting in a cell-by-gene integer matrix of transcript counts, from which we infer the level of gene expression. Specifically, if a_{cg} is the expected number of mRNA molecules of gene g in cell c , then $\alpha_{cg} = a_{cg} / \sum_{g'} a_{cg'}$ is the fraction of mRNA molecules within that cell given to gene g , and we will refer to $\log \alpha_{cg}$ as the “gene expression,” where the log accounts for variation in fold change rather

than absolute value and serves to stabilize variance. There is uncertainty in this inference due to fluctuations in the number of mRNA molecules arising from inherent biological stochasticity and from the measurement process, which can be significant. A host of data processing pipelines are used by the community, each one taking a different approach to normalizing the data in some manner to account for the uncertainty in the measurement [32,33]. We normalize our gene expression data with Sanity [34], a biophysically motivated Bayesian approach to estimating a posterior distribution for gene expression. The resulting posterior is approximated as an independent Gaussian for each cell and gene, so the estimate for expression is $\mathcal{N}(X_{cg}, \epsilon_{cg})$, effectively giving error-bars to the inferred expression matrix X_{cg} . We adopt the notational convention that X_{cg} refers to the inferred expression matrix containing all embryos at a given stage. Later we will introduce models for the expression of a single embryo, which we will denote as x_{cg} . During PCR, a single mRNA molecule can be unevenly amplified [35]. Without UMIs it is impossible to know whether one original molecule has been counted multiple times during sequencing, and Smart-Seq2 does not use UMIs. While Sanity assumes UMI counts as input and does not explicitly model noise introduced by the amplification process, we found that adjusting the error bars resulted in a consistent estimate of gene expression (Supplemental Material (SM) Sec. II [36]). Throughout, we make use of this uncertainty estimate, finding error bars for statistics, such as covariances, and for robustly identifying cell types. We have found that propagating the uncertainty to downstream analyses greatly strengthens the robustness of our statistical inferences and scientific conclusions.

A major technical challenge is to find the distinct transcriptomic states [37], or cell types, at each stage and identify which state each cell belongs to. Identification of cell types allows us to partially reconstruct the embryo’s spatial structure; an accurate classification of each cell is crucial to compute quantities like mean expression levels and cell-cell covariances around that mean. In order to robustly recover clusters,

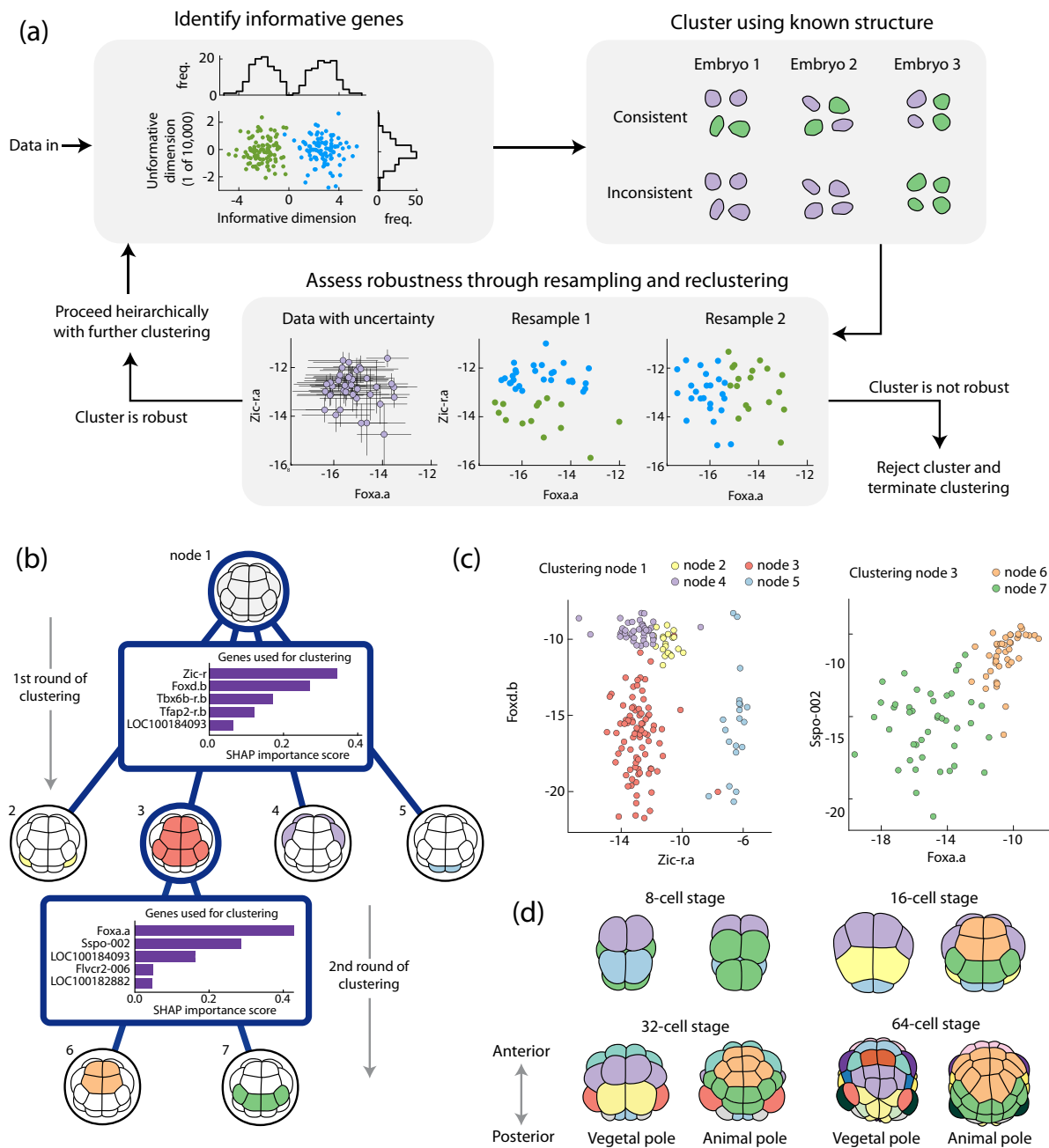


FIG. 3. Principled and physically motivated clustering robustly recovers cell types. (a) Overview of clustering algorithm highlighting the three key underlying principles. (1) Identify informative genes. Informative dimensions split the data into separable clusters (green vs blue points, top left); uninformative dimensions do not and make identifying clusters harder. (2) Cluster using known structural priors. A cell type should have equal numbers of cells in each embryo, a clustering without that property is inconsistent (purple and green cells represent clusters, top right). (3) Assess robustness through resampling and reclustering. We have a posterior for our expression levels [34] (purple points with error bars, bottom), and by repeatedly resampling and clustering (blue vs green points, bottom), we assess how consistent the clusters are and decide whether to keep or reject a cluster. For full details see SM Sec. III [36]. (b) Worked example of the hierarchical clustering for the 16-cell stage. Starting from all cells at node 1, the algorithm splits into four clusters using the five genes shown. The SHAP importance score is a measure of how useful each gene was for performing this round of clustering [38] (SM Sec. III [36]). The algorithm then splits cluster 3 into two further clusters using a different set of five genes. The algorithm terminates here as further clusters are assessed to be not robust (SM Sec. III [36]). (c) The two most important genes with expression values drawn from the posterior shown for both clustering stages, showing separation between identified clusters. (d) Eight-cell through 64-cell stages colored by cell type showing increasing transcriptomic specialization in time.

we developed an algorithm [Fig. 3(a)] which accounts for known properties of the developmental system in focus, listed as follows. First, most genes are not expressed in an informative way [39], either not expressed at all or expressed equally in every cell in the embryo. A few genes will be expressed in some cell types but not others, and finding these informative genes is crucial to distinguishing clusters; this information gets lost when using all genes [Fig. 3(a)]. At each stage of clustering, we adapt the algorithm of Ref. [39] to find a set of informative genes by generating trial clusters and searching for a sparse hyperplane defined by a handful of genes which discriminates between these trial clusters (SM Sec. III [36]). We find this to be an effective way of identifying informative genes, although many alternative approaches are available, such as biclustering [40,41]. Second, clusters must be consistent across embryos, a cluster with two cells in every embryo is conceivably a cell type, a cluster with four cells in one embryo and zero cells in the others is not a cell type [Fig. 3(a)]. Throughout, we enforce all clusters to be consistent with the known embryo structure (SM Sec. III [36]). Third, we use uncertainty about expression to assess whether a cluster is robust [42–44] and is to be kept or not. If a cluster is robust, then that cluster should be consistently identified after perturbing the data and reclustering. Hence, upon repeatedly resampling from the posterior, subsampling the data, and clustering, we should repeatedly find the same clusters [Fig. 3(a)]. Having performed this resampling, we compute a cluster consistency score, which we then compare against cluster consistency scores for null datasets generated by randomly assigning cells to embryos. These null datasets preserve the variance in the expression but destroy the known structure of cell types, a true cluster should have a significantly better consistency score than these null datasets. This provides a criteria to assess whether a cluster is robust and should be rejected or not (SM Sec. III [36]). Overall, we combine these three principles into a hierarchical algorithm, motivated by the hierarchical nature of cell differentiation [45]. We search for a small number of well-separated clusters before searching these for further subclusters, potentially using a different set of genes to do so. The 16-cell stage is shown as an example in Figs. 3(b) and 3(c); for full details of the algorithm and results for other stages see SM Sec. III [36]. Having found the clusters, it is straightforward to identify them with cell identities and their position in the embryo [Fig. 3(d)] using existing *in situ* hybridization data for select genes [46] (SM Sec. III [36]). We recover cell types previously identified [22,47], but note that the challenge is not only to identify cell types but to accurately assign a cell type for every single cell including those with seemingly ambiguous expression levels. Working through each embryo and manually assigning cells to a cell type by inspecting the expression of known marker genes is possible in theory but would rapidly become infeasible at later stages where many genes are needed to assign cells or for higher-throughput data. Our approach for identifying clusters is fully automated and could in principle identify novel genes or cell types at later stages. Altogether, the approach outlined here and described in detail in the SM can be used directly for systems with invariant cell lineage, such as ascidian, and many of the ideas can be adapted broadly for a robust and principled approach

to cell-type identification [36]. The Bayesian approach to normalization as well as the resampling and shuffling inherent to our method could make it computationally expensive for large datasets, adapting the algorithm to be feasible for large datasets is an outstanding challenge worth overcoming.

B. Identifying variation in the embryonic transcriptome

With cell types recovered, we can combine embryos to get an average transcriptomic state for the embryo at each cell stage. This typical state, or wild type (WT), contains information about the system. For instance, measuring activation of the *Otx* gene in a specific cell at the late 32-cell stage tells us that cell has received an FGF signal emitted by a neighboring cell. Probing this usually involves making a significant perturbation, e.g., genetically knocking out or pharmacologically perturbing a signaling pathway, and studying the result which may differ significantly from the WT [Fig. 4(a)]. Here complementary to this approach, we study the system using only the average WT trajectory and the natural phenotypic variation around it [Fig. 4(b)]. Though cells occupy a vast transcriptomic space spanned by the number of genes in the organism, n_{genes} , embryos occupy an even larger transcriptomic space, $n_{\text{cells}} \times n_{\text{genes}}$. At a particular developmental time along the one-dimensional (parameterized by time) WT trajectory, embryos have a $n_{\text{cells}} \times n_{\text{genes}} - 1$ -dimensional space to explore [Fig. 4(b)]. We refer to this form of variation in gene expression as natural since each individual can deviate from the consensus trajectory due to genetic and environmental causes. In this formulation, the dimensionality of the embryo changes after cells undergo cell division. While this phenomenon is worthy of attention, here we consider each cell stage separately and hence the dimensionality does not dynamically change. We will study the subspace that is accessible to typical biological variation and, from quantitative analyses and physical modeling, infer interactions between cells.

It is not obvious that the scRNA-seq measurement reveals any form of natural variation at all, perhaps all that can be garnered from such measurements are cell types. Said another way, Is there any signal above the noise after cell types are identified? Can we indeed study natural variation and the emergent patterns manifest in space?

Turning again to the experimental data, prior to any modeling, we first ask whether we can statistically detect collective variation within an embryo. To ensure that any variation we find is biologically meaningful, at each cell stage we consider two sets of genes: (1) a set of purely zygotic genes transcribed in the embryo and not present initially in the zygote and (2) a corresponding set of maternal genes, which were present in the zygote (we exclude the handful of maternal transcripts that are localized to the posterior pole which determine the germ cell lineage). During this time, zygotic transcripts are serving to pattern the embryo, create distinct cell states, and transfer information across cells. We expect our set of zygotic genes to collectively vary due to differences in staging time, as well as other sources of collective difference induced through cell-cell interactions or lineage relationships. In contrast, while nonlocalized maternal transcripts are essential for the self-organizational process, we expect them to lack the complex

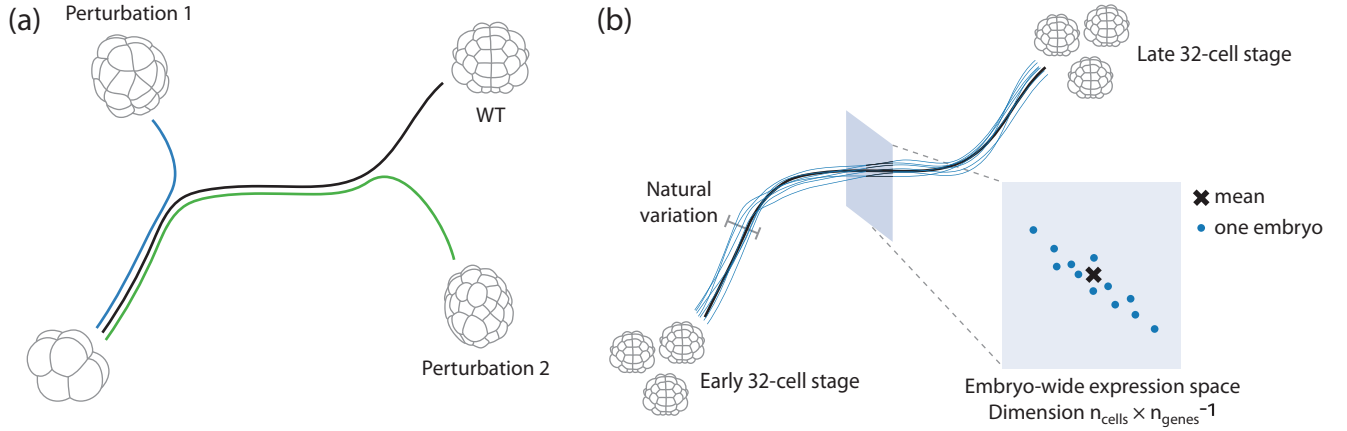


FIG. 4. Studying the mechanisms underlying collective modes of gene expression through quantifying natural phenotypic variation. (a) Development is thought to be canalized, implying that embryos robustly follow a typical wild-type trajectory (black). Typically, the tools of molecular biology used to study development produce perturbations that are strong enough to kick the system far from its canalized trajectory. (b) Even though development is canalized, each embryo will follow a unique trajectory, slightly deviating from the average. Taking a slice at a specific time point along the developmental trajectory, there exists a high-dimensional space of variation that embryos can explore. Trajectory from early to late 32-cell stage is shown as an example.

correlations of the zygotic genes even if their overall level may vary across embryos due to genetic or mother-specific differences in mRNA composition. In this way, maternal genes can provide a check that spurious correlations are not induced by technical batch effects. To ensure similarity to the zygotic genes, we created a set of matching maternal genes by selecting the maternal gene of closest average expression to each zygotic gene, and working with this set. We also restrict to maternal genes which are not differentially expressed at any stage, excluding those whose transcripts may be selectively degraded in specific cell types (see Sec. V).

C. Variation in maternal transcripts

Let us first consider the set of maternal genes, with corresponding cell-by-gene matrix of expression, X_{cg} . We will account for known biological variation before identifying if any further sources of variation exist. To begin we consider the singular values of the matrix after subtracting the mean expression levels, $X_{cg} \mapsto Y_{cg} = X_{cg} - \sum_d X_{dg} / \sum_d 1$ [Fig. 5(b)]. Beyond the mean, for the maternal genes we expect there to be variation arising from mother-to-mother differences in mRNA deposition, which we can subtract from the expression matrix, $Y_{cg} \mapsto Z_{cg} = Y_{cg} - v_{mg}$, where $v_{mg} = \sum_{d \in m} Y_{dg} / \sum_{d \in m} 1$ is the mother-specific mean expression, m is the mother of cell c , and $d \in m$ denotes the set of all cells d with mother m . Further, we anticipate additional embryo-to-embryo variation, which we subtract similarly, $Z_{cg} \mapsto A_{cg} = Z_{cg} - \eta_{eg}$, where $\eta_{eg} = \sum_{d \in e} Y_{dg} / \sum_{d \in e} 1$ is the embryo-specific mean expression and $d \in e$ is set of all cells d in embryo e . The singular values of the expression matrix, after this known variation has been removed, are shown in Fig. 5(b).

Having accounted for these known sources of variation, we can now investigate the structure of the remaining matrix A_{cg} . For matrices arising from single-cell sequencing experiments, retaining singular values which exceed what one would expect from a random matrix, and setting the remainder to zero is an effective method of removing technical and stochastic

expression noise while retaining biological signal [48,49]. Here we obtain the singular values of the random matrix by noting that if each entry of A_{cg} were drawn from independent gene-specific distributions, then the singular values would be invariant under permuting the values in each column of this matrix. Applying such a permutation preserves the marginal distribution of each gene but destroys the correlation structure. Computing the singular values of the permuted matrix and comparing against those of the unpermuted matrix finds a handful of nontrivial singular values at each stage. At the 32-cell stage we estimate there are at most nine significant components [Fig. 5(b)], although future statistical tests are not sensitive to the exact number. We project the expression levels for each cell onto these principal components (PCs), effectively setting to zero the singular values that were plausibly noise, and work in this low-dimensional space from here onward.

Even though several principal components were kept, we have not shown that genuine spatially ordered collective embryo-to-embryo variation is present, beyond that accounted for by the overall embryo-specific level of expression. The reason for this is that correlations between genes in the same cell could give a matrix with significant singular values, even if there are no correlations between cells. To test this, we compute covariances for cells within the same embryo and examine how this changes after shuffling cells between embryos [Fig. 5(d)]. Note that we cannot compute the covariance between, say, neighboring a5.3 and a5.4 cells since we only know that a cell is in the a5.3-a5.4 cell type, not its exact location [Fig. 1(b)]. Yet, we can still compute the average covariance between a cell in the a5.3-a5.4 cell type and one in the B5.1 cell type within the same embryo. Specifically, we can measure the covariance

$$\gamma_{ij}^{t,t'} = \frac{1}{n_t n_{t'}} \sum_{c \in t, c' \in t'} \langle x_{ci} x_{c'j} \rangle \quad (1)$$

for cell types t, t' , principal components i, j , where n_t is the number of cells in cell type t and x_{ci} is the expression for an

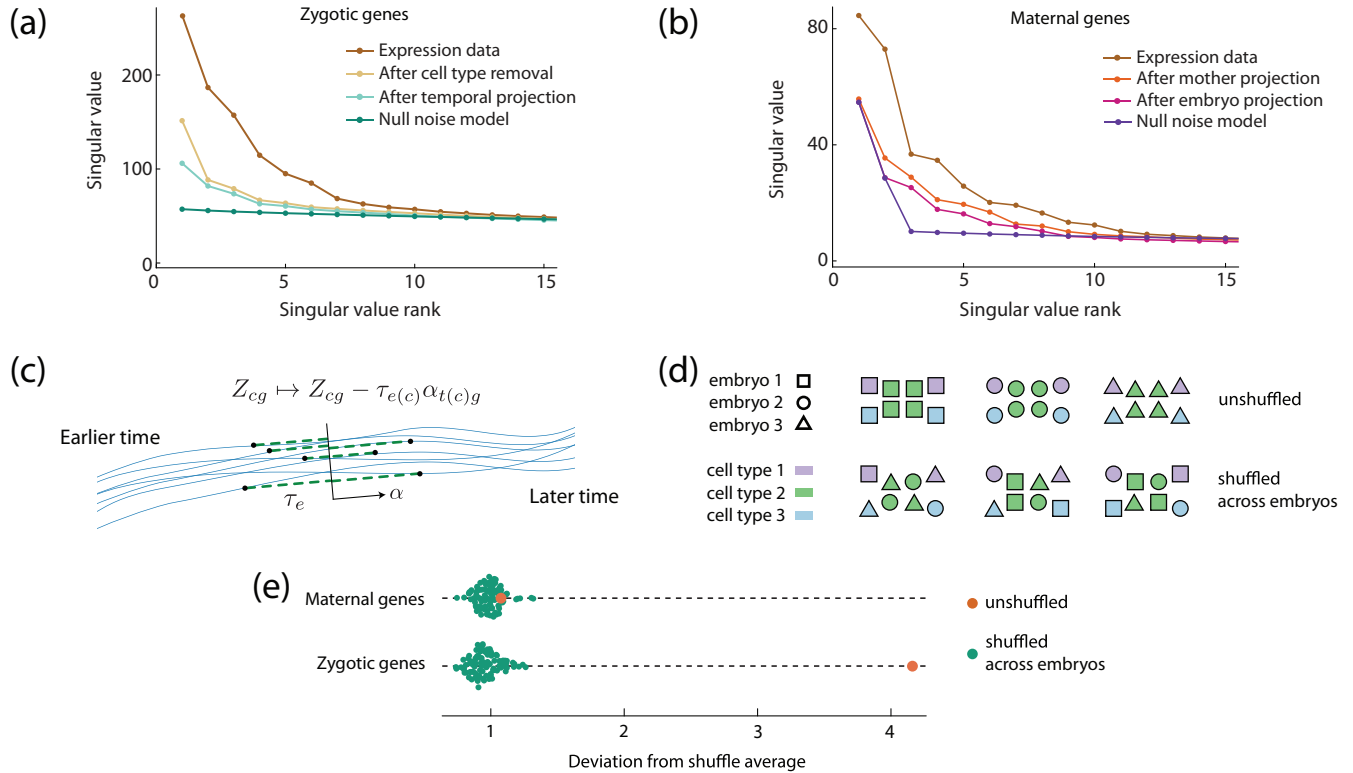


FIG. 5. A quantitative assessment of single-cell gene expression data demonstrates correlated expression between cells within an individual embryo. [(a) and (b)] Singular values of the gene expression matrix, for zygotic (a) and maternal (b) genes at the 32-cell stage before and after known sources of variation, such as cell types, variations from mother to mother or embryo to embryo, and variations in developmental staging [see (c)], are removed. Singular values of null noise model are shown for comparison. Other stages are shown in SM Figs. S17 and S18 [36]. (c) Local developmental time, τ_e for each embryo (black points, full trajectories are blue lines) will vary owing to slight discrepancies in staging. We remove this variation by projecting each embryo onto a common time (black line) by identifying the axis tangent to the average trajectory, α , and then subtracting the temporal variation, $\tau_e\alpha$ from the gene expression matrix Z . (d) To test for cell-cell correlations we shuffle cells of the same cell type between embryos. Color is cell type, shape is embryo identity. (e) Results of shuffling as in (d) for both maternal and zygotic genes at the 32 cell stage (other stages in Figs. S17 and S18 [36]). Each parameter configuration has 100 random shuffle trials. In every case, all covariances were measured and the deviation Δ from the average covariance values is plotted.

embryo after known variance has been removed. We can also measure covariance within a cell for each cell type,

$$C_{ij}^t = \frac{1}{n_t} \sum_{c \in t} \langle x_{ci} x_{cj} \rangle. \quad (2)$$

Together, we say these make up the set of empirical covariances $\{C_\alpha\}$. For a robust empirical estimate of $\gamma_{ij}^{tt'}$, and C_{ij}^t , we repeatedly resample expression from the posterior to account for uncertainty from the expression measurement, as well as bootstrapping to account for uncertainty from having finite data; see SM Sec. IV for details [36].

To test for collective variation, we compute the intercellular covariance statistics, $\gamma_{ij}^{tt'}$, before and after shuffling the data [Fig. 5(d)]. Taking 100 shuffles as well as the unshuffled data, we compute the component-wise average covariance, $\bar{\gamma}_{ij}^{tt'}$, as well as the component-wise standard deviation, $\sigma_{ij}^{tt'}$, from which we calculate the deviation Δ ,

$$\Delta = \sum_{i,j,t,t'} \left(\frac{\gamma_{ij}^{tt'} - \bar{\gamma}_{ij}^{tt'}}{\sigma_{ij}^{tt'}} \right)^2. \quad (3)$$

If the value of Δ for the unshuffled data is significant compared with values for the shuffled data, then there is significant collective variation; see SM Sec. IV for full details [36].

For these maternal genes, shuffling within an embryo does not affect the covariance statistics at any stage, Δ is not significant, suggesting a lack of spatial structure [Fig. 5(e)]. Hence, the variation can be explained by embryo-to-embryo variation in initial mRNA deposition, without needing to invoke spatial distributions or coupling through cell divisions. Biologically, this is what we expect for the nonlocalized maternal mRNA transcripts; there is an embryo-wide variation in the total expression levels, and no other detectable sources of variation were introduced through batch effects or processing.

D. Variation in zygotic transcripts

For the zygotic genes, we know that each cell type, whose identification we have described in the previous section, has specific expression patterns which we remove, $Y_{cg} \mapsto Z_{cg} = Y_{cg} - \mu_{tg}$, where t is the cell type of cell c , $\mu_{tg} = \sum_{c \in t} Y_{cg} / \sum_{c \in t} 1$ is the cell-type-specific mean, and $d \in t$ is the set of all cells d in cell type t . For this remaining

expression Z_{cg} , we ask whether there is any signature of collective variation.

We expect that each embryo differs in developmental time or how far the embryo is along the WT trajectory [3], even though care was taken to stage them at a consistent time [22]. We want to identify whether collective variation exists around the WT trajectory, beyond simply variation along the WT trajectory. To do so, we remove the temporal variation by projecting onto one fixed developmental time, fitting an embryo-specific time τ_e , and a local cell-type-specific tangent vector to the developmental trajectory α_{tg} by minimizing the variance left in the matrix $Z_{cg} \mapsto Z_{cg} - \tau_e \alpha_{tg}$ [Fig. 5(c)] (see Sec. V). Without orthogonal measurements it is difficult to guarantee that $\tau\alpha$ represents a staging time correction, and not some other mode of variation. That said, should development proceed in a purely stereotypical way, the only source of collective variation would be staging time, and this correction would remove that variation.

Indeed, for the 16-cell stage, having subtracted the temporal variation, there are no significant singular values remaining (SM Sec. IV [36]). However, for the 32- and 64-cell stages this is not the case; the data have nontrivial singular values and hence signal [Fig. 5(b)] (SM Sec. IV [36]). The cell-type-specific expression accounts for much of the variance with the WT trajectory projection typically accounting for the largest remaining singular value [Fig. 5(b)].

To test whether collective variation is present, having first accounted for cell state, staging time, and retaining only a small number of singular values, we perform shuffle tests across embryos. Unlike the maternal genes, the shuffle tests reveal genuine collective variation in gene expression within and across embryos for the zygotic genes for the 32- and 64-cell stages [Fig. 5(e)] (SM Fig. S18 for 64-cell stage [36]). This agrees with our biological intuition, that the space of zygotic transcripts is tightly controlled and thus the fluctuations that do occur are coupled across cells within an embryo, in contrast to the maternal transcripts which can vary from cell to cell within an embryo and are not as tightly restricted in their variation. So far we have treated the zygotic genes differently to their maternal counterparts, accounting for staging time rather than subtracting an overall embryo-wide effect. Instead, we can repeat this analysis treating the zygotic genes identically to the maternal ones, by removing an overall embryo effect, identifying singular values to keep, and testing whether there is significant variation through shuffle tests. As before, this finds significant detectable variation at the 32- through 64-cell stages, but now it also finds variation at the 16-cell stage, suggesting our explicit accounting for staging time explains the variance better than an overall embryo-wide effect (SM Sec. IV [36]). Altogether, our detailed statistical approach set out to ask whether there was measurable collective fluctuations having accounted for known sources of variation. We find that for the set of zygotic genes there exists collective embryo-to-embryo variation.

E. Modeling collective variation

So far, we have established the existence of collective fluctuations within an embryo for zygotic genes, in contrast to the maternal transcripts. We now attempt to understand the nature

of this collective behavior. Constructing and then studying the empirical joint probability distribution of an embryo's state of gene expression, the embryonic transcriptome, is impractical, since each embryo counts as only one data point in a high-dimensional space and we only partially know a cell's identity. Despite these limitations, the principle of maximum entropy gives us a way to rationalize statistics of this system, making minimal assumptions about the nature of interactions and using the statistics we can estimate, in particular the covariances between different cell types. In short, we suppose there are certain important statistics which are prescribed by the system. For instance, the variance of a gene within a cell is constrained by the gene regulatory network, or the covariance between two sister cells is fixed by the mechanism of division. There are infinitely many distributions that are consistent with these statistics, but out of all such distributions there is one that makes minimal further assumptions: the maximum entropy distribution [15,50,51]. Here the first nontrivial statistics we can consider to be constrained by the system are covariances (including variances), as we have already subtracted the mean.

For instance, we could take the average covariance between PC i and PC j for cells in cell type t ,

$$C_{ij}^t = \frac{1}{n_t} \sum_{c \in t} \langle x_{ci} x_{cj} \rangle \quad (4)$$

to be fixed by the system, where n_t is the number of cells in cell type t . Similarly, we could consider the covariance between cells p and q , $\Lambda_{(pi)(qj)} = \langle x_{pi} x_{qj} \rangle$ to be fixed. Since there is a left-right symmetry in the embryo at this early stage, there is no biological reason to vary the left and right covariances independently. Instead if p and q are on the left side of the embryo, with $\tilde{p}$ and $\tilde{q}$ their equivalent cells on the right side, then we may fix $\Lambda_{ij} = \frac{1}{2} (\langle x_{pi} x_{qj} \rangle + \langle x_{\tilde{p}i} x_{\tilde{q}j} \rangle)$. In general, we can introduce an asymmetric directed adjacency matrix, Ω_{rs} , to fix covariances of the form

$$\Lambda_{ij} = \frac{1}{|\Omega|} \sum_{p,q} \Omega_{pq} \langle x_{pi} x_{qj} \rangle, \quad (5)$$

with $\Omega_{sr} \in \{0, 1\}$ keeping track of which cells are being included.

Since $\langle x_{pi} x_{qj} \rangle \neq \langle x_{pj} x_{qi} \rangle$, covariances between cells can be directed, for example, the covariance between PC 1 in A5.1 and PC 2 in A5.2 can be different than the covariance between PC 2 in A5.1 and PC 1 in A5.2. Indeed, we might directly fix the covariance between PC 1 in A5.1 and PC 2 in A5.2, but not fix the covariance between PC 2 in A5.1 and PC 1 in A5.2, implemented by having an asymmetric adjacency matrix Ω . However, in certain cases we might explicitly fix symmetric covariances, for instance the covariance between PC i in one cell and PC j in a neighboring cell, averaged over all pairs of neighbors without choosing a directional arrow between these pairs. In this case we denote the adjacency matrix as Ψ to explicitly indicate that it is symmetric. In general there will be a set of such symmetric constraints and a corresponding set $\{\Psi^k\}$ of symmetric adjacency matrices, along with a set of asymmetric constraints and corresponding set $\{\Omega^k\}$ of directed adjacency matrices that we consider to be fixed.

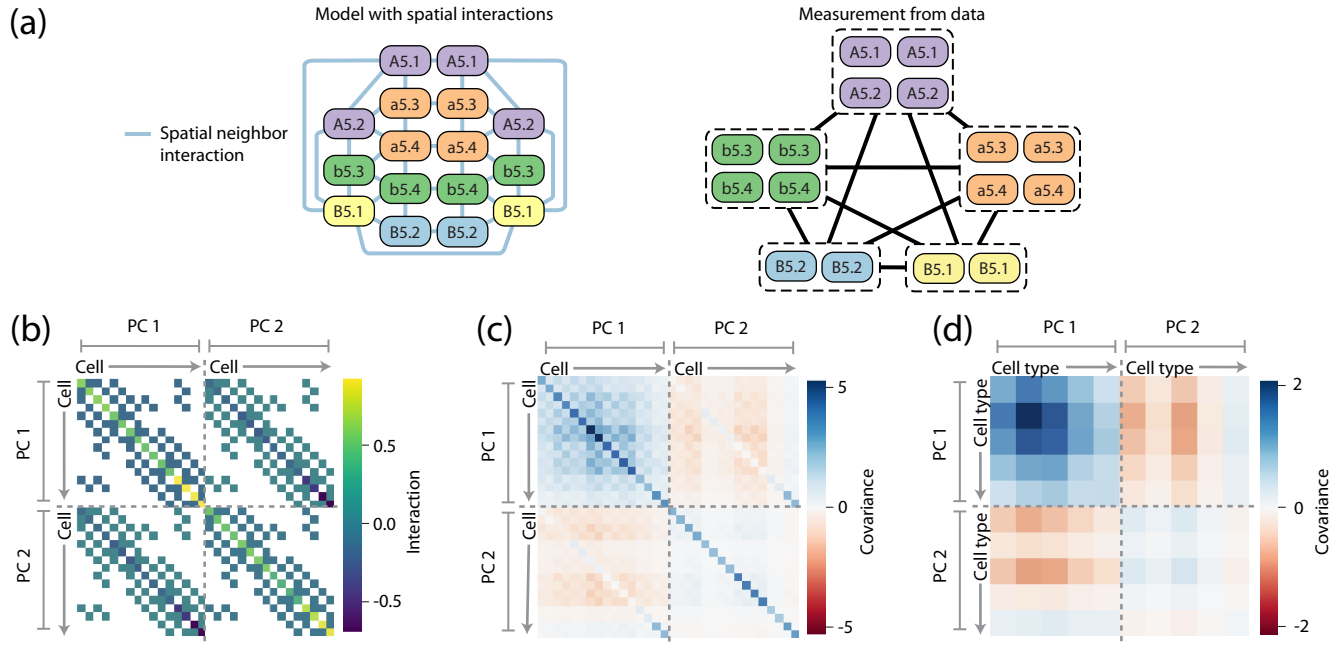


FIG. 6. Structure of maximum entropy model with sparse interactions. (a) Model of the 16-cell stage, where spatial neighbors are directly interacting. In the model we know every cell's identity and the covariance between any two cells (left). However, in the data we only know the cell type of each cell and not its exact identity (right), hence we can only compute the covariances between cell types. (b) The sparse interaction matrix J , for model $p(\mathbf{x}) = \mathcal{N}(0, J^{-1})$ where only spatial neighbors have direct interactions, shown for the 16-cell stage with two principle components as an example. (c) Dense covariance matrix J^{-1} . Note that while J^{-1} is always symmetric, $(J^{-1})_{(pi)(qj)} \neq (J^{-1})_{(pj)(qi)}$ in general, even if $J_{(pi)(qj)} = J_{(pj)(qi)}$. (d) Dense matrix of cell type-to-cell type covariances, corresponding to the model J , which, unlike (c), can be directly compared to measurements from data.

Whatever set of covariances we elect to fix, the distribution with maximum entropy that is consistent with those covariances is a Heisenberg-like model, $p(\mathbf{x}) = \mathcal{N}(0, J^{-1})$ with $\mathbf{x} = (x_{pi})$ with i indexing over principal components, p indexing over cells within an embryo, $\mathbf{x}$ is flattened into a vector, and

$$J_{(pi)(qj)} = \sum_t M_{ij}^t I_{pq}^t + \sum_{k=1}^{n_\Psi} S_{ij}^k \Psi_{pq}^k + \frac{1}{2} \sum_{k'=1}^{n_\Omega} (U_{ij}^{k'} \Omega_{pq}^{k'} + U_{ji}^{k'} \Omega_{qp}^{k'}), \quad (6)$$

where $I_{pq}^t = \delta_{pq} \delta_{q \in t}$ is an indicator function with t a cell type; Ψ^k and Ω^k are adjacency matrices of which there are n_Ψ and n_Ω , respectively, and M , S , U are interaction parameters; M couples genes within the same cell, S symmetrically couples different cells, U is a general coupling between different cells, all arise as Lagrange multipliers enforcing the constraints (SM Sec. V [36]). This distribution arises as it is the simplest one consistent with biologically and physically motivated constraints on the system, namely that principal components within the same cell are coupled with a cell-type-specific coupling and that principal components between cells can be coupled through an adjacency matrix.

As an example, suppose the developmental dynamics directly sets the intracellular covariance matrix for each cell type, as well as the covariance matrix for cells in spatial contact. For illustrative purposes we plot this for the 16-cell stage with two principal components [Fig. 6(a)], although

these components are not statistically significant (Fig. S18 [36]). The interaction matrix, $J_{(pi)(qj)} = M_{ij}^{(p)} \delta_{pq} + S_{ij} \Psi_{pq}$, is sparse, where Ψ_{pq} is the spatial adjacency matrix of an embryo at a given stage [Fig. 6(b)]. The nonzero terms of this matrix specify direct interactions, since if $J_{(pi)(qj)} = 0$ for all i, j , then cells p and q are conditionally independent given the expression of all other cells. While J is sparse, the covariance matrix, J^{-1} is dense [Fig. 6(c)] as is the matrix of observed covariances between cell types [Fig. 6(d)]. The only assumption is that cells in spatial contact have correlated gene expression, the form of the distribution follows from the maximum entropy principle and the parameters are fit to experimental measurements of covariances (SM Sec. V [36]). We can ask how well this simple model explains the data and compare against a model which assumes sisters are interacting, so $\Psi_{pq} = 1$ when p and q are sister cells, or a model where all cells are equally interacting, so $\Psi_{pq} = 1$ for any pair $p \neq q$ [Fig. 7(a)]. We quantify the quality of the fit by comparing the empirical covariances, $\{C_\alpha\}$, the set of covariances measured from the data, to the model predicted covariances $\{\tilde{C}_\alpha\}$, which we can directly compute from the model [Fig. 6(d)]. A perfectly fitting model would have $C_\alpha = \tilde{C}_\alpha$, $\forall \alpha$, and we quantify the fit by measuring the variance in $\{C_\alpha\}$ explained by $\{\tilde{C}_\alpha\}$, namely the quantity $\eta = 1 - \sum_\alpha (C_\alpha - \tilde{C}_\alpha)^2 / \sum_\alpha C_\alpha^2$. At the 32-cell stages, the embryo-wide model and neighbor model can reasonably explain the experimental data ($\eta = 0.87$ and 0.87 , respectively) [Fig. 7(b)] with the spatial model fitting the data better at the 64-cell stage ($\eta = 0.83$ embryo-wide vs 0.92 for spatial). In contrast, the sister model cannot explain the experimental observations [Fig. 7(b)] as this model can never

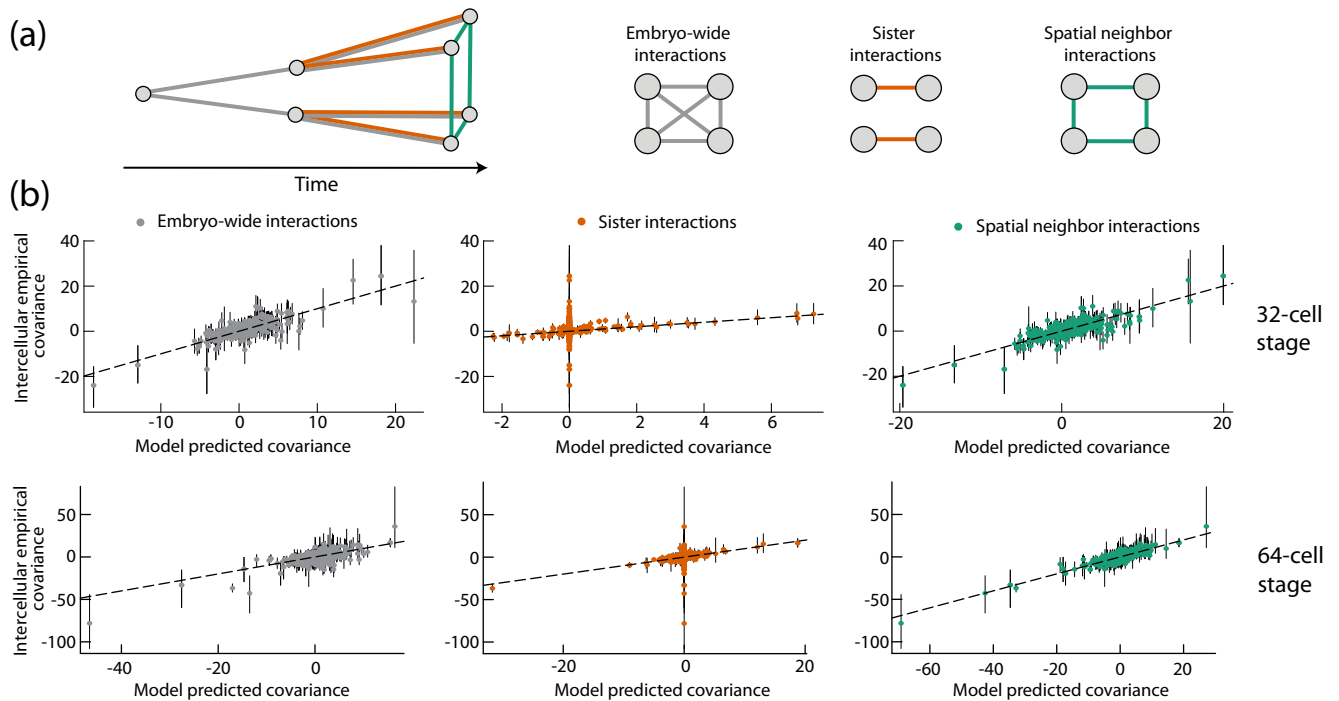


FIG. 7. Comparing maximum entropy models with data identifies plausible cell-cell interaction rules. (a) Correlations between different cells could arise from signaling resulting in interactions between spatial neighbors, cell division resulting in interactions between sister cells, or variations in the zygote resulting in all-to-all coupling across the embryo. We consider how well these different hypotheses explain the experimental data. (b) The intercellular empirical covariances (IEC) are covariances between different cell types measured from data. The model predicted covariances (MPC), are covariances between different cell types in the maximum entropy model [Fig. 6(d)]. A model which explains the data should have MPC values equal to the IEC values. Here we test models which assume that all cells within an embryo directly interact (gray) only sisters interact (orange) or only spatial neighbors interact (green). Error bars are from the 25th to 75th percentile of covariance estimates.

cause embryo-wide correlations, only correlations between sister cells ($\eta < 0.64$ at 32-cell stage, $\eta < 0.78$ at 64-cell stage).

While comparing these hypotheses for interactions is revealing, there are embryo-wide, neighborhood, and sister cell interactions at play, not to mention interactions between specific cells representing cell-cell signaling. Consider a model which admits embryo-wide, spatial neighbor, and sister-cell interactions, together with specific cell-cell interaction terms which respect the left-right symmetry of the embryo for all spatial neighbors. Such a model has too many parameters (1080 U and S values for the 32-cell stage with five principal components) and will overfit the data. Therefore, and motivated by the biology of the system, we assume the nature of interactions are sparse and use regularization [52,53], together with uncertainty estimates for the covariances, to identify which of the many possible interactions are important and can robustly explain the experimental measurements. Specifically, we apply a L_1 regularization penalty for each interaction parameter (U and S terms) with strength λ . We draw $N = 100$ empirical covariance matrices from the posterior distribution and fit a regularized model to each of them. We then assess which parameters are a consistent and nonzero size across different fits and use this to rank terms by importance. We sample across a number of λ values, making a ranking for each, and then aggregate the rankings with the Schulze method [54], resulting in an overall ranking of terms by importance. Full

details in SM Sec. VI [36]. Given this ranking, starting from the most important terms, we include sufficient terms in the model to reach $\eta > 0.9$, finding a 19-term model at the 32-cell stage [Fig. 8(b)] and a 47-term model at the 64-cell stage (SM Fig. S21 [36]).

The interactions we identify must be cross-referenced with what is known and suspected about the signaling interactions at this stage of development. These interactions have been studied through the tools of molecular embryology, which are wholly distinct from the approach we have pursued here. Neural induction, where FGF signaling causes a6.5 and b6.5 cells to adopt a neural fate, occurs at the late 32-cell stage and is the first known cell-cell signaling that causes a change of state. However, levels of the FGF-target gene *Otx* were not found to be elevated in potential a6.5 or b6.5 cells relative to the remaining animal cells in the sequencing data, suggesting that cells were sequenced before induction occurred. In general, cell-cell communication induces correlations in time, with the transcriptomic machinery taking a finite time to respond to new information. A static snapshot may capture some of those temporal couplings, but only after the signal is received. Neighbor, embryo-wide, and a handful of specific cell-cell interactions, two of which are between sister cells reflecting interaction through inheritance, appear in the top inferred terms (SM Table S3 [36]). All of the inferred specific cell-cell interactions are between animal cells or between vegetal cells with no animal-vegetal interactions which indeed

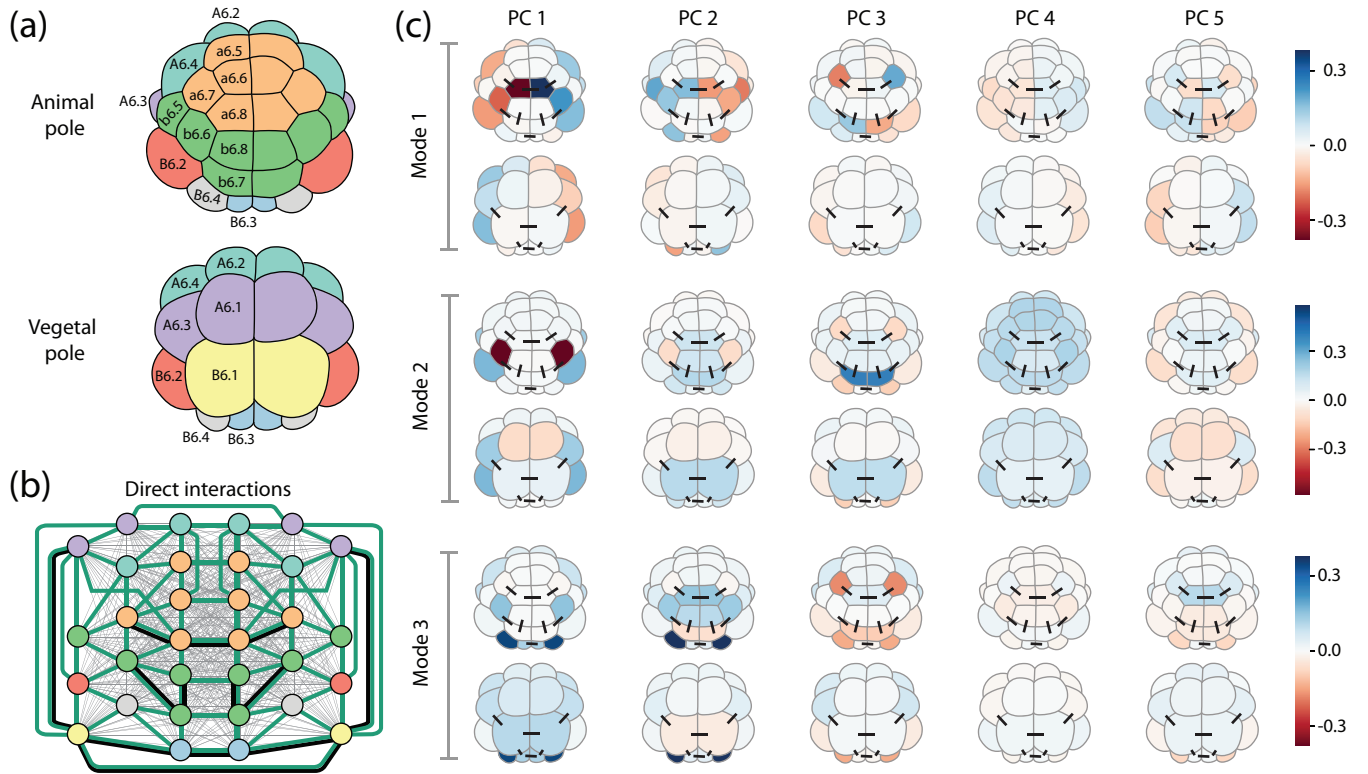


FIG. 8. Statistical modeling identifies spatially ordered collective modes in expression data and reveals underlying interactions. (a) Schematic of the 32-cell stage embryo colored by cell type. (b) Graph representation of the 32-cell stage embryo, colored by cell type with edges representing inferred direct interactions in regularized statistical model. A direct interaction simply means there is a nonzero term coupling two cells in the J matrix of our model. Embryo-wide (gray lines) and neighbor (green) cell-cell interactions are present in model, as well as specific cell-cell interactions (black lines). The ranking of interaction terms is shown in SM Table S3 [36]. (c) Principal modes of variation from inferred regularized model. The inferred model from (b) is a Gaussian which can be interpreted as a multidimensional ellipsoid defined by the covariance matrix J^{-1} . Here we visualize the first three principal directions of that covariance matrix (called modes 1–3), which are the principal directions along which variation occurs. Each such direction, or mode ϕ_{cg} , exists in a $n_{\text{cell}} \times n_{\text{PC}}$ -dimensional space, which we can visually represent by $n_{\text{PC}} = 5$ columns (PC's = 1, $\dots$, 5) where column g shows an embryo (both animal and vegetal view shown) with cell c colored by the value ϕ_{cg} . Black lines show direct interactions between specific cells.

should not have occurred yet. Communication between animal cells is expected to occur through EphrinA signaling [55] at this stage, which would support model's inference of multiple direct interactions between animal cells. For the remainder of nonsister inferred interactions, there is no known molecular mechanism for them, say, left and right B6.1 cells, to communicate at this stage in a way analogous to FGF signaling. Instead, perhaps they could reflect embryo-to-embryo genetic differences in levels of expression for expressed genes, such as *nodal* or *snail* in the case of B6.1, which would correlate expression between these cells across embryos. While cell-cell inductions have occurred by the time of the 64-cell sequencing, the data have fewer and less complete embryos, as well as noisier sequencing which limits our ability to robustly infer interactions. Interactions which could reflect neural induction (A7.4–a7.9, and b7.11–B7.4) were identified (Fig. S21), although many other terms including embryo-wide, neighbor, and, for the first time, sister-cell interactions were also identified as significant. Altogether, at the 32-cell stage, we infer specific animal cell-cell interactions which could arise from EphrinA signaling, along with embryo-wide and neighbor interactions. At the 64-cell stage, we identify

some interactions plausibly arising from neural induction, but the lack of data at this stage prevents truly robust inference.

While we do not know the ground truth of cell-cell interactions, in addition to the results being biologically plausible, we can check to see whether our inference procedure is self-consistent. Specifically, given a particular sparse model, we can sample from it to create a dataset of the same size and quality as the experimental data (SM Sec. VII [36]). After putting this simulated data through the inference procedure, we do not recover exactly the same ranking, but true sparse interactions are identified as among the most important (SM Fig. S23 [36]), showing that there is robustness to the inference pipeline despite the limitations of the data.

F. Identifying the collective modes of gene expression

Having found a model with a sparse set of interactions, which explains the observed covariances, we can examine this model and ask what the spatially ordered modes of collective variation look like in this system. Given any Gaussian model, $\rho(\mathbf{x}) \sim \mathcal{N}(0, \Sigma)$, one can picture the distribution as an ellipsoid defined by the principal axes of the covariance

matrix $\Sigma = J^{-1}$. Specifically, since Σ is symmetric, we have the standard mode decomposition

$$\Sigma_{(ip)(jq)} = \sum_{\beta} \lambda^{\beta} \phi_{ip}^{\beta} \phi_{jq}^{\beta}, \quad (7)$$

where the ϕ^{β} are eigenvectors of Σ with eigenvalues λ^{β} . The ϕ^{β} corresponding to the largest eigenvalues give the directions of the principal modes of variation. Analogously to principal component analysis, projecting $\mathbf{x}$ to the k largest eigenvectors, retains $\sum_{\beta \leq k} \lambda^{\beta} / \sum_{\beta} \lambda^{\beta}$ of the variance of $\mathbf{x}$. For our sparse model at the 32-cell stage, around 30% of the variation is captured in the first three principal axis or modes of variation, shown in Fig. 8(c), and around 90% is captured in the first 15. These modes represent coordinated embryo-wide variations in gene expression, in stark contrast to the modes found for a model without cell-cell interactions (SM Fig S20 [36]). We emphasize that each mode [rows of Fig. 8(c)] corresponds to a set of zygotic genes, which collectively express in particular spatial patterns across the cells of an embryo. The pattern of a specific gene is some linear combination of the PCs [columns of Fig. 8(c)]. Importantly, since we do not know the exact identity of each cell, only which cell type it belongs to, these modes reflect an inference of the model.

These modes are not computed directly from the data, they are computed indirectly through first inferring a model. It is relevant, therefore, to ask how robust they are to the choice of model. We find that the 15 modes that account for around 90% of the variance in the sparse 19-term model account for around 50% of the variance of models ranging from 16–24 terms. Moreover, these modes account for 40% of the variance in the models of 16–24 terms inferred from the synthetic data which has a slightly different set of interaction terms (SM Fig. S23 [36]). These modes exist in a high-dimensional space ($n_{\text{cells}} \times n_{\text{PC}} = 160$ for the 32-cell stage), in which two random vectors will typically be almost orthogonal. Applying a random rotation to these 15 modes results in a subspace that explains less than 10% of the variance: If the mode inference was sensitive to the model terms, then this would be the typical level of variance explained.

The first collective mode in Fig. 8(c) represent a left-right mode of variation, which agrees with biological intuition that left and right sides are relatively uncoupled and hence could vary independently or be at slightly different time points along their developmental progression. Strong fluctuations in the a6.8 cells and b6.6 cells in modes 1 and 2, respectively, could represent further cell types beginning to form but not strongly differentiated enough to be separated by the clustering. These modes represent a testable prediction of the theory. Were expression levels measured without losing cell identity, say, with *in situ* hybridization or scRNA-seq with careful dissection of embryos, then one could directly measure the subspace of variation around the mean expression levels. Our prediction is that this subspace is characterized by the leading modes of our inferred model.

IV. DISCUSSION

Our goal was to study variation in ascidian development at the level of the embryo, both to understand what typical

variation looks like as well as to construct generative models of the variation. Starting from high-quality transcriptomic data [22], across multiple stages and multiple embryos, we developed a new statistically robust approach to characterize the WT transcriptional states at successive stages and to accurately classify every cell into one of those states. For zygotic transcripts, we found that embryo-to-embryo variation along the WT time axis, as well as expression noise at the level of single cells, was sufficient to explain variation at the 8- and 16-cell stages. At the 32- and 64-cell stages, we show that there are additional collective degrees of freedom in the data that vary from embryo to embryo. To understand the source of the collective variation, we introduced statistical physics models, allowing us to infer both interactions as well as the collective modes of variation from limited experimental observations. Our findings shows the maternal to zygotic transition in action, which begins at the 8-cell stage as zygotic transcription turns on in the embryo, with a handful of genes transcribed at the 16-cell stage. As increasing numbers of zygotic genes are expressed at later stages, and different cells become coupled through inheritance and cell-cell communication, nontrivial collective variation in the embryo appears.

A central goal of this work was to identify these spatially organized collective modes of variation at the scale of the whole embryo. By searching for such modes, we reveal natural variation present in an otherwise tightly controlled developmental program, in contrast to strong perturbative approaches typically taken to study development. As such, our work represents the desire to go beyond cell typing and leverage single-cell expression measurements to address a broader scope of biological questions, including analyzing the collective behavior of multiple cells rather than one cell at a time. Further, we propose that, since selection acts on the organism rather than its individual cells, the collective modes we identify are also phenotypic traits that naturally vary and are under selection.

To perform much of our quantitative analysis, we required a high-quality identification of cell type from single-cell data, and as such we formulated a statistically robust approach to cell typing which can be leveraged by the community broadly. Central to this was the emphasis on utilizing a posterior distribution for gene expression that permits resampling for downstream statistical tasks. Gene expression data are widely recognized as noisy, yet often the only statistical challenge is seen as estimating the most likely value of expression. Instead, by accepting that these measurements have potentially large uncertainty and including this uncertainty in our analysis, we can strengthen our ability to perform not only cell typing but also improve our estimation of covariances and inference of interactions along with related statistical problems. The sequencing data used here were of high depth, but we hope that future studies can investigate a diverse array of developmental programs with shallower sequencing, using some of the techniques developed here.

To build statistical models and remove noise from the data, instead of using hundreds of zygotic genes, we work with a handful of principal components. The principal components span across many genes, with genes often contributing to multiple components. While the building blocks of the developmental process are undoubtedly genes, by working

with principal components we are searching for a level of description beyond the single gene and at the collective level of multiple genes, much as in statistical physics we work with collective variables like density rather than individual molecules. Here the choice was data-driven principal components, although one could coarse grain over biologically motivated collections of genes, say, in different regulatory pathways. In principle, one could even take different principal components or regulatory pathways for distinct cell types, reflecting that each cell type may have its own coordinates of typical variation. Recent deep learning and large language model approaches [56], trained on a vast corpus of single-cell sequencing among other biological data, find a nonlinear transcriptome latent space. It would be intriguing to take such a latent space as the input degrees of freedom to then be modeled within a statistical physical approach, which could provide an interpretable model for such data.

Statistical physics models have proved powerful at explaining related biological phenomena such as pattern formation in cell layers [57], with maximum entropy models in particular used to study gene regulatory networks [15,16], protein dynamics [58], and antibody diversity [59]. While recent work has begun directly applying maximum entropy modeling to scRNA-seq data [16,60], to the best of our knowledge, these techniques have not yet been used to study the collective expression of multiple cells. Here our knowledge of the stereotyped spatial connectivity and mean expression level of cells within an embryo allows us to statistically model the joint probability distribution of gene expression for all the cells in an embryo at a given stage—an embryonic transcriptome—with one sample from the generative model representing one embryo. As statistical physics suggests, interacting systems, which cells in an embryo are, have rich collective behavior and should be modeled as such.

Finally, we remind the reader that our approach does not model temporal dynamics. Each stage of the embryo is treated separately despite the fact that they are connected in time. This is a consequence of the destructive nature of scRNA-seq together with the significant length of time between measurements. Expression profiles between stages are seemingly time discontinuous and so the data are intrinsically nondynamical; indeed, our inferred models are even specific to the time of sequencing within each cell stage. Including deeper lineage statistics rather than just sister-cell interactions or creating a joint distribution of all stages at once could begin to incorporate more dynamical information into a model. Perhaps more dynamical measurements are required to fit a more dynamical model, but a dynamical perspective is worth considering, as these collective modes are ultimately a spatiotemporal feature of development.

V. METHODS

A. Raw scRNA data

All experimental data come from the single-cell RNA sequencing experiments of Ref. [22], available with accession numbers, ArrayExpress: E-MTAB-6506, E-MTAB-6508, E-MTAB-6528, and E-MTAB-6530. We only consider cell stages 8 to 64 as there is no zygotic expression before stage 8

[22]. In total there are 8 embryos at the 8-cell stage, 11 embryos at the 16-cell stage, 14 embryos at the 32-cell stage, and 8 embryos at the 64-cell stage. Not all embryos are complete.

B. Aligning transcripts

We use the reference transcriptome of Ref. [22], ENA accession number PRJEB3758, which was assembled from scRNA transcripts and validated against a *Ciona robusta* reference. We also compared against the Aniseed genome [46] which has a more complete set of genes but does not have untranslated regions, limiting the reads and hence quality of expression estimate of important genes (SM Sec. II [36]). Aligning was performed using the STAR method [61].

C. Differentially expressed genes

Given a particular gene, we test whether it is differentially expressed in two cell types by collecting expression values for all cells in belonging to the two cell types and performing the standard Wilcoxon rank-sum test [62]. To find all genes that are differentially expressed across any two cell types, we perform the above test but adjust p -values to account for testing across multiple genes and cell-type pairs. We consider genes at significance $p < 0.01$ for adjusted p -value as differentially expressed. Figure 1(c) shows the differentially expressed genes at the 16-cell stage.

D. Zygotic genes

At the k -cell stage, we identify a set of purely zygotic genes. First we take all genes which have fewer than 10 reads in every cell at the four-cell stage, which identifies nonmaternal genes which are not expressed at early stages. Of those genes, if any have more than 100 reads in any cell at the k -cell stage, we include them in our set of zygotic genes at that stage. This criterion excludes genes that have maternal transcripts present which are then additionally are transcribed zygotically, as well as lowly expressed zygotic genes. It correctly identifies the major zygotic genes and excludes any maternal factors. Next, we exclude any zygotic genes that are not differentially expressed across cell types. We do this as certain genes are expressed in every cell in some embryos and in no cells in others. While interesting, we choose to focus on the differentially expressed zygotic genes which are actively patterning each embryo.

E. Maternal genes

To select a set of purely maternal genes, we take all genes with more than 500 reads in any cell at the four-cell stage that are also not differentially expressed in any subsequent stage. This gives us 2211 genes that are purely maternal factors and not zygotically transcribed. To create a set of maternal genes similar to the zygotic set at some stage, for each zygotic gene we take the maternal gene closest in average expression to that zygotic gene, working through the zygotic genes without replacement until we have assembled a set of maternal genes the same length as the zygotic set and with similar average expression levels.

F. Temporal projection

To remove temporal variation from the expression matrix Z_{cg} , we want to find a set of embryo-specific times τ_e and tangent vector α_{tg} that minimize the variation

$$M = \sum_{g,c} (Z_{cg} - \tau_{e(c)} \alpha_{t(c)g})^2, \quad (8)$$

where we have adopted the notation $t(c)$ to mean the cell type corresponding to cell c and $e(c)$ the embryo corresponding to cell c . We solve this by alternately factorizing this expression,

$$M = \sum_e \left[\tau_e^2 \sum_{g,c \in e} \alpha_{t(c)g}^2 - 2\tau_e \sum_{g,c \in e} Z_{cg} \alpha_{t(c)g} \right] + \text{const.} \quad (9)$$

$$= \sum_{t,g} \left[\alpha_{tg}^2 \sum_{c \in t} \tau_{e(c)}^2 - 2\alpha_{tg} \sum_{g,c \in e} Z_{cg} \tau_{e(c)} \right] + \text{const.} \quad (10)$$

and taking an iterative strategy of setting

$$\tau_e = \sum_{g,c \in e} Z_{cg} \alpha_{t(c)g} / \sum_{g,c \in e} \alpha_{t(c)g}^2 \quad (11)$$

and then

$$\alpha_{tg} = \sum_{g,c \in t} Z_{cg} \tau_{e(c)} / \sum_{g,c \in t} \tau_{e(c)}^2, \quad (12)$$

together with normalization step of $\sum_{g,t} \alpha_{tg}^2 = 1$ to resolve the undetermined scaling degree of freedom. This iterative minimization converges well in practice.

G. Maximum entropy model

We find the maximum entropy distribution $\rho(\mathbf{x})$, subject to the constraints that intracellular covariances are fixed

$$C_{ij}^t = \frac{1}{n_t} \sum_{c \in t} \langle x_{ci} x_{cj} \rangle, \quad (13)$$

and n_ψ symmetric covariances of the form

$$\Gamma_{ij} = \frac{1}{2|\Psi|} \sum_{p,q} \Psi_{pq} \langle x_{pi} x_{qj} \rangle, \quad (14)$$

are fixed and n_Ω asymmetric correlations of the form

$$\Lambda_{ij} = \frac{1}{|\Omega|} \sum_{p,q} \Omega_{pq} \langle x_{pi} x_{qj} \rangle, \quad (15)$$

are fixed. The total Lagrangian is

$$\begin{aligned} L = & - \int d\mathbf{x} \rho(\mathbf{x}) \log \rho(\mathbf{x}) + \mu \int d\mathbf{x} \rho(\mathbf{x}) \\ & + \frac{1}{2} \sum_{i,j,p,q} \left(\sum_t M_{ij}^t I_{pq}^t + \sum_{k=1}^{n_\psi} S_{ij}^k \Psi_{pq}^k + \sum_{k'=1}^{n_\Omega} U_{ij}^{k'} \Omega_{pq}^{k'} \right) \\ & \times \int d\mathbf{x} \rho(\mathbf{x}) x_{pi} x_{qj}, \end{aligned} \quad (16)$$

where μ , M , S , and U are Lagrange multipliers. Maximizing this over ρ leads to $\rho(\mathbf{x}) = \mathcal{N}(0, J^{-1})$, where

$$\begin{aligned} J_{(pi)(qj)} = & \sum_t M_{ij}^t I_{pq}^t + \sum_{k=1}^{n_\psi} S_{ij}^k \Psi_{pq}^k \\ & + \frac{1}{2} \sum_{k'=1}^{n_\Omega} (U_{ij}^{k'} \Omega_{pq}^{k'} + U_{ji}^{k'} \Omega_{qp}^{k'}), \end{aligned} \quad (17)$$

and we have symmetrized the final term to ensure J is symmetric. There is no known analytic formula for the inverse of J in terms of the M , S , and U matrices. Typically, one would numerically search for M , S , and U that satisfy the constraints. However, with the exception of C^t , we do not directly measure the constraints, we measure correlations like $\Gamma_{ij}^{tt'} = (\sum_{r \in t, q \in t'} \langle x_{ri} x_{qj} \rangle) / n_t n_{t'}$, where we have defined $\Gamma_{ij}^{tt'}$ and t and t' are cell types. We therefore choose our parameters so that the model matches these observables. Specifically we minimize

$$\begin{aligned} \tilde{L}(\{M^t, S^k, U^k\}) = & \sum_{i,j,t} \left(C_{ij}^t - \frac{1}{n_t} \sum_{r \in t} (J^{-1})_{(rj)(ri)} \right)^2 \\ & + \sum_{i,j,t < t'} \left(\Gamma_{ij}^{tt'} - \frac{1}{n_t n_{t'}} \sum_{r \in t, q \in t'} (J^{-1})_{(ri)(qj)} \right)^2, \end{aligned} \quad (18)$$

over $\{M^t, S^k, U^k\}$. Assuming, for now, that a solution exists with minimum $\tilde{L} = 0$, then we have found that a maximum entropy model, fit to correlation matrix C and some *a priori* unknown Λ , Γ , which we can calculate from M , S , and U . We minimize the objective $\tilde{L}$ starting from the initial guess $M^t = (C^t)^{-1}$, $S^k = U^k = 0$, using gradient descent with a hand computed gradient (SM Appendix A [36]).

H. Determining cell-cell contact network

From light-sheet microscopy images of embryos segmented with the ASTEC framework [23], two cells were determined to be in contact if, on average, the area of contact represented at least 5% of one of the cells surface area. This average was taken over the left and right side of the embryos, assuming a symmetric contact map, as well as across two embryos.

ACKNOWLEDGMENTS

We thank Kilian Biasuz for providing data and assistance in determining cell-cell contacts, and Christelle Dantec for discussions about aligning transcripts. D.J.S. thanks Eric Siggia and Dillon Cislo for helpful discussions and valuable comments on the manuscript. M.M. and D.J.S. were supported by Simons Foundation Grant No. 597491 and NSF Grant No. 1764421. M.M. is a Simons Investigator. P.L. is a senior CNRS employee. This project has been made possible in part by Grant No. DAF2023-329587 from the Chan Zuckerberg

Initiative DAF, an advised fund of the Silicon Valley Community Foundation. This work was supported by a binational “NSF-ANR Collaborative Research: A mechanical atlas for embryogenesis at single-cell resolution.” (National Science Foundation [2204237] and ANR ANR-21-CE13-0046) to M.M. and P.L. This research was supported in part by Grants No. NSF PHY-1748958 and No. NSF PHY-2309135 and the

Gordon and Betty Moore Foundation Grant No. 2919.02 to the Kavli Institute for Theoretical Physics (M.M.).

DATA AVAILABILITY

The data that support the findings of this article are openly available [63].

-
- [1] J. M. W. Slack, *From Egg to Embryo: Regional Specification in Early Development*, 2nd ed., Developmental and Cell Biology Series (Cambridge University Press, Cambridge, UK, 1991).
- [2] C. H. Waddington, *The Strategy of the Genes* (Allen & Unwin, Sydney, 1957).
- [3] V. Alba, J. E. Carthew, R. W. Carthew, and M. Mani, Global constraints within the developmental program of the *Drosophila* wing, *eLife* **10**, e66750 (2021).
- [4] S. R. Quake, The cell as a bag of RNA, *Trends Genet.* **37**, 1064 (2021).
- [5] J. Cao, J. S. Packer, V. Ramani, D. A. Cusanovich, C. Huynh, R. Daza, X. Qiu, C. Lee, S. N. Furlan, F. J. Steemers, A. Adey, R. H. Waterston, C. Trapnell, and J. Shendure, Comprehensive single-cell transcriptional profiling of a multicellular organism, *Science* **357**, 661 (2017).
- [6] Tabula Sapiens Consortium, The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans, *Science* **376**, eabl4896 (2022).
- [7] Tabula Muris Consortium, Single-cell transcriptomics of 20 mouse organs creates a tabula muris, *Nature (London)* **562**, 367 (2018).
- [8] G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastrioti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko, RNA velocity of single cells, *Nature (London)* **560**, 494 (2018).
- [9] C. Weinreb, A. Rodriguez-Fraticelli, F. D. Camargo, and A. M. Klein, Lineage tracing on transcriptional landscapes links state to fate during differentiation, *Science* **367**, eaaw3381 (2020).
- [10] S. L. Freedman, B. Xu, S. Goyal, and M. Mani, A dynamical systems treatment of transcriptomic trajectories in hematopoiesis, *Development* **150**, dev201280 (2023).
- [11] A. Alemany, M. Florescu, C. Baron, J. Peterson-Maduro, and A. van Oudenaarden, Whole-organism clone tracing using single-cell sequencing, *Nature (London)* **556**, 108 (2018).
- [12] S. Jin, C. F. Guerrero-Juarez, L. Zhang, I. Chang, R. Ramos, C.-H. Kuan, P. Myung, M. V. Plikus, and Q. Nie, Inference and analysis of cell-cell communication using cellchat, *Nat. Commun.* **12**, 1088 (2021).
- [13] S. Kirkpatrick and D. Sherrington, Infinite-ranged models of spin-glasses, *Phys. Rev. B* **17**, 4384 (1978).
- [14] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, Statistical mechanics for natural flocks of birds, *Proc. Natl. Acad. Sci. USA* **109**, 4786 (2012).
- [15] T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Fedoroff, Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns, *Proc. Natl. Acad. Sci. USA* **103**, 19033 (2006).
- [16] J. Jiang, S. Chen, T. Tsou, C. S. McGinnis, T. Khazaei, Q. Zhu, J. H. Park, I.-M. Strazhnik, J. Hanna, E. D. Chow, D. A. Sivak, Z. J. Gartner, and M. Thomson, D-SPIN constructs gene regulatory network models from multiplexed scRNA-seq data revealing organizing principles of cellular perturbation response, *bioRxiv*: [10.1101/2023.04.19.537364](https://doi.org/10.1101/2023.04.19.537364) (2023).
- [17] Y. Roudi, E. Aurell, and J. Hertz, Statistical physics of pairwise probability models, *Front. Comput. Neuro.* **3**, 22 (2009).
- [18] E. G. Conklin, Mosaic development in ascidian eggs, *J. Exp. Zool.* **2**, 145 (1905).
- [19] H. M. Funakoshi, T. T. Shito, K. Oka, and K. Hotta, Developmental table and three-dimensional embryological image resource of the ascidian *ascidiella aspersa*, *Front. Cell Dev. Biol.* **9**, 789046 (2021).
- [20] G. Kumano and H. Nishida, Ascidian embryonic development: An emerging model system for the study of cell fate specification in chordates, *Dev. Dynamics* **236**, 1732 (2007).
- [21] K. Hotta, K. Mitsuhashi, H. Takahashi, K. Inaba, K. Oka, T. Gojobori, and K. Ikeo, A web-based interactive developmental table for the ascidian *ciona intestinalis*, including 3d real-image embryo reconstructions: I. From fertilized egg to hatching larva, *Dev. Dynam.* **236**, 1790 (2007).
- [22] H. L. Sladitschek, U.-M. Fiuza, D. Pavlinic, V. Benes, L. Hufnagel, and P. A. Neveu, Morphoseq: Full single-cell transcriptome dynamics up to gastrulation in a chordate, *Cell* **181**, 922.e21 (2020).
- [23] L. Guignard, U.-M. Fiuza, B. Leggio, J. Laussu, E. Faure, G. Michelin, K. Biasuz, L. Hufnagel, G. Malandain, C. Godin, and P. Lemaire, Contact area-dependent cell communication and the morphological invariance of ascidian embryogenesis, *Science* **369**, eaar5663 (2020).
- [24] J. Wei, W. Zhang, A. Jiang, H. Peng, Q. Zhang, Y. Li, J. Bi, L. Wang, P. Liu, J. Wang, Y. Ge, L. Zhang, H. Yu, L. Li, S. Wang, L. Leng, K. Chen, and B. Dong, Temporospatial hierarchy and allele-specific expression of zygotic genome activation revealed by distant interspecific urochordate hybrids, *Nat. Commun.* **15**, 2395 (2024).
- [25] H. Nishida and K. Sawada, *macho-1* encodes a localized mRNA in ascidian eggs that specifies muscle fate during embryogenesis, *Nature (London)* **409**, 724 (2001).
- [26] V. Bertrand, C. Hudson, D. Caillol, C. Popovici, and P. Lemaire, 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 (2003).
- [27] K. Imai, N. Takada, N. Satoh, and Y. Satou, β -catenin mediates the specification of endoderm cells in ascidian embryos, *Development* **127**, 3009 (2000).

- [28] T. Negishi and H. Nishida, Asymmetric and unequal cell divisions in ascidian embryos, in *Asymmetric Cell Division in Development, Differentiation and Cancer*, edited by J.-P. Tassan and J. Z. Kubiak (Springer International, Cham, 2017), pp. 261–284.
- [29] S. Picelli, O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg, Full-length RNA-seq from single cells using Smart-seq2, *Nat. Prot.* **9**, 171 (2014).
- [30] C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, and W. Enard, Comparative analysis of single-cell rna sequencing methods, *Mol. Cell* **65**, 631 (2017).
- [31] N. Treen, T. Heist, W. Wang, and M. Levine, Depletion of maternal cyclin b3 contributes to zygotic genome activation in the ciona embryo, *Cur. Biol.* **28**, 1150 (2018).
- [32] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, Spatial reconstruction of single-cell gene expression data, *Nat. Biotech.* **33**, 495 (2015).
- [33] M. D. Robinson and A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data, *Genome Biol.* **11**, R25 (2010).
- [34] J. Breda, M. Zavolan, and E. van Nimwegen, Bayesian inference of gene expression states from single-cell RNA-seq data, *Nat. Biotech.* **39**, 1008 (2021).
- [35] J. Lause, C. Ziegenhain, L. Hartmanis, P. Berens, and D. Kobak, Compound models and pearson residuals for normalization of single-cell RNA-seq data without UMIs, *bioRxiv* 10.1101/2023.08.02.551637 (2023).
- [36] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/7gs5-w48q> for additional algorithmic and numerical details, Figs. S1–S25, and Tables S1–S3.
- [37] P. Grobecker, T. Sakoparnig, and E. van Nimwegen, Identifying cell states in single-cell RNA-seq data at statistically maximal resolution, *PLoS Comput. Biol.* **20**, e1012224 (2024).
- [38] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, From local explanations to global understanding with explainable ai for trees, *Nat. Mach. Intel.* **2**, 56 (2020).
- [39] S. Melton and S. Ramanathan, Discovering a sparse set of pairwise discriminating features in high-dimensional data, *Bioinformatics* **37**, 202 (2021).
- [40] Y. Cheng and G. M. Church, Biclustering of expression data, in *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, Vol. 8 (AAAI Press, 2000), pp. 93–103.
- [41] J. Xie, A. Ma, A. Fennell, Q. Ma, and J. Zhao, It is time to apply biclustering: A comprehensive review of biclustering applications in biological and biomedical data, *Brief. Bioinform.* **20**, 1450 (2019).
- [42] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data, *Mach. Learn.* **52**, 91 (2003).
- [43] Y. Senbabaoğlu, G. Michailidis, and J. Z. Li, Critical limitations of consensus clustering in class discovery, *Sci. Rep.* **4**, 6207 (2014).
- [44] E. Levine and E. Domany, Resampling method for unsupervised estimation of cluster validity, *Neural Comput.* **13**, 2573 (2001).
- [45] S. Jang, S. Choubey, L. Furchtgott, L.-N. Zou, A. Doyle, V. Menon, E. B. Loew, A.-R. Krostag, R. A. Martinez, L. Madisen, B. P. Levi, and S. Ramanathan, Dynamics of embryonic stem cell differentiation inferred from single-cell transcriptomics show a series of transitions through discrete cell states, *eLife* **6**, e20487 (2017).
- [46] J. Dardailon, D. Dauga, P. Simion, E. Faure, T. A. Onuma, M. B. DeBiasse, A. Louis, K. R. Nitta, M. Naville, L. Besnardeau, W. Reeves, K. Wang, M. Fagotto, M. Guérault-Bellone, S. Fujiwara, R. Dumollard, M. Veeman, J.-N. Volff, H. Roest-Crolius, E. Douzery, J. F. Ryan, B. Davidson, H. Nishida, C. Dantec, and P. Lemaire, ANISEED 2019: 4D exploration of genetic data for an extended range of tunicates, *Nucl. Ac. Res.* **48**, D668 (2019).
- [47] A. Madgwick, M. S. Magri, C. Dantec, D. Gailly, U.-M. Fiuza, L. Guignard, S. Hettinger, J. L. Gomez-Skarmeta, and P. Lemaire, Evolution of embryonic cis-regulatory landscapes between divergent phallusia and ciona ascidians, *Dev. Biol.* **448**, 71 (2019).
- [48] L. Aparicio, M. Bordyuh, A. J. Blumberg, and R. Rabadan, A random matrix theory approach to denoise single-cell data, *Patterns* **1**, 100035 (2020).
- [49] M. Nitzan and M. P. Brenner, Revealing lineage-related signals in single-cell gene expression using random matrix theory, *Proc. Natl. Acad. Sci. USA* **118**, e1913931118 (2021).
- [50] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill, Principles of maximum entropy and maximum caliber in statistical physics, *Rev. Mod. Phys.* **85**, 1115 (2013).
- [51] J. Shore and R. Johnson, Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy, *IEEE Trans. Inform. Theory* **26**, 26 (1980).
- [52] Y. Kleeorin, W. P. Russ, O. Rivoire, and R. Ranganathan, Undersampling and the inference of coevolution in proteins, *Cell Syst.* **14**, 210 (2023).
- [53] G. Stepaniants, A. D. Hastewell, D. J. Skinner, J. F. Totz, and J. Dunkel, Discovering dynamics and parameters of nonlinear oscillatory and chaotic systems from partial observations, *Phys. Rev. Res.* **6**, 043062 (2024).
- [54] M. Schulze, A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method, *Soc. Choice Welfare* **36**, 267 (2011).
- [55] N. Ohta and Y. Satou, Multiple signaling pathways coordinate to induce a threshold response in a chordate embryo, *PLoS Genet.* **9**, e1003818 (2013).
- [56] Y. Rosen, Y. Roohani, A. Agarwal, L. Samotorčan, T. S. Consortium, S. R. Quake, and J. Leskovec, Universal cell embeddings: A foundation model for cell biology, *bioRxiv*: 10.1101/2023.11.28.568918 (2023).
- [57] M. Smart and A. Zilman, Emergent properties of collective gene-expression patterns in multicellular systems, *Cell Rep. Phys. Sci.* **4**, 101247 (2023).
- [58] A. J. Dear, G. A. Garcia, G. Meisl, G. A. Collins, T. P. J. Knowles, and A. L. Goldberg, Maximum entropy determination of mammalian proteome dynamics, *Proc. Natl. Acad. Sci. USA* **121**, e2313107121 (2024).
- [59] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Maximum entropy models for antibody diversity, *Proc. Natl. Acad. Sci. USA* **107**, 5405 (2010).

- [60] C. Sarra, L. Sarra, L. Di Carlo, T. GrandPre, Y. Zhang, C. G. Callan, and W. Bialek, Maximum entropy models for patterns of gene expression, [Phys. Rev. E **112**, 014408 \(2025\)](#).
- [61] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner, [Bioinformatics **29**, 15 \(2013\)](#).
- [62] G. C. Tseng, D. Ghosh, and E. Feingold, Comprehensive literature review and statistical considerations for microarray meta-analysis, [Nucl. Ac. Res. **40**, 3785 \(2012\)](#).
- [63] D. Skinner, P. Lemaire, and M. Mani, Physical modeling of embryonic transcriptomes identifies collective modes of gene expression [Data set], Zenodo (2025), doi:[10.5281/zenodo.16782577](#).