# Approximated Gene Expression Trajectories (AGETs) for Gene Regulatory Network Inference on Cell Tracks

**Kay Spiess**[1,2], **Timothy Fulton**[1], **Seogwon Hwang**[1], **Kane Toh**[1], **Dillan Saunders**[1], **Brooks Paige**[2,3*], **Benjamin Steventon**[1*], **Berta Verd**[1,4*]

***For correspondence:**
b.paige@ucl.ac.uk (BP);
bjs57@cam.ac.uk (BS);
berta.verdfernandez@zoo.ox.ac.uk (BV)

[1]Department of Genetics, University of Cambridge, Cambridge, UK;
[2]The Alan Turing Institute, London, UK;
[3]Centre for Artificial Intelligence, University College London, UK;
[4]Department of Zoology, University of Oxford, UK

**Abstract** The study of pattern formation has benefited from reverse-engineering gene regulatory network (GRN) structure from spatio-temporal quantitative gene expression data. Traditional approaches omit tissue morphogenesis, hence focusing on systems where the timescales of pattern formation and morphogenesis can be separated. In such systems, pattern forms as an emergent property of the underlying GRN. This is not the case in many animal patterning systems, where patterning and morphogenesis are simultaneous. To address pattern formation in these systems we need to adapt our methodologies to explicitly accommodate cell movements and tissue shape changes. In this work we present a novel framework to reverse-engineer GRNs underlying pattern formation in tissues experiencing morphogenetic changes and cell rearrangements. By combination of quantitative data from live and fixed embryos we approximate gene expression trajectories (AGETs) in single cells and use a subset to reverse-engineer candidate GRNs using a Markov Chain Monte Carlo approach. GRN fit is assessed by simulating on cell tracks (live-modelling) and comparing the output to quantitative data-sets. This framework outputs candidate GRNs that recapitulate pattern formation at the level of the tissue and the single cell. To our knowledge, this inference methodology is the first to integrate cell movements and gene expression data, making it possible to reverse-engineer GRNs patterning tissues undergoing morphogenetic changes.

## Introduction

Embryonic pattern formation underlies much of the diversity of form observed in nature. As such, one of the main goals in developmental biology is to understand how spatio-temporal molecular patterns emerge in developing embryos, are maintained and change over the course of evolution. Over the past three decades, the interest of the field has focused on elucidating the function and dynamics of the gene regulatory networks (GRNs) underlying these processes. GRNs can be formulated mathematically as non-linear systems of coupled differential equations whose parameters can be inferred from quantitative gene expression data: a methodology known as reverse-engineering ***Reinitz and Sharp*** (***1996***); ***Liang et al.*** (***1998***); ***D'haeseleer et al.*** (***2000***); ***Gardner and Faith*** (***2005***); ***Rockman*** (***2008***); ***He et al.*** (***2009***); ***Jaeger and Monk*** (***2010***); ***Crombach et al.*** (***2012***).

Reverse-engineering has been successfully applied to a myriad of systems, from the *Drosophila* blastoderm to the vertebrate neural tube **Verd et al.** (**2017**, 2018); **Manu et al.** (**2009**); **Balaskas et al.** (**2012**), generating a wealth of knowledge on the mechanisms by which GRNs read out morphogen gradients **Verd et al.** (**2019**); **Jaeger et al.** (**2004**); **Balaskas et al.** (**2012**); **Cohen et al.** (**2015**); **Kicheva et al.** (**2014**); **El-Sherif et al.** (**2014**), scale patterns **Wu et al.** (**2015**), control the timing of differentiation **Averbukh et al.** (**2018**); **Schröter et al.** (**2012**); **Rayon et al.** (**2020**), synchronise cellular fates **Uriu et al.** (**2010**) and evolve pattern formation **Crombach et al.** (**2016**).

Much of what we know about pattern formation has been learnt from reverse-engineering GRN structure from spatio-temporal quantitative data in systems where the timescales of pattern formation and morphogenesis are different and can therefore be separated. In such systems, spatio-temporal gene expression profiles are typically obtained by measuring gene expression levels across the tissue of interest in fixed stained samples, and interpolating between measurements at different time points **Crombach et al.** (**2012**). The underlying and seldom stated assumption, is that the patterning dynamics are much faster than the cell movements in the developing tissue, and that therefore cell movements can be ignored during the timescales at which the pattern forms. This is true in many systems and processes, such segmentation patterning in early *Drosophila* embryogenesis. When this is indeed the case, pattern formation can be considered an emergent property of GRN dynamics alone **Kicheva et al.** (**2012**) and much insight can be drawn from analysing reverse-engineered GRNs. However, it should not be regarded a property of developmental patterning systems in general.

In systems where tissue patterning and tissue morphogenesis are coupled and occurring simultaneously, GRNs alone do not generate the resulting patterns and can therefore not fully explain them. This has been recently highlighted by work in organoids, where shape, size and cell type distribution are difficult to control as a result of altered patterning due to abnormal morphogeneses in unconstrained tissue geometries **Huch et al.** (**2017**). To understand developmental pattern formation we have to address how morphogenesis and GRNs together control fate specification and embryonic organisation. Importantly, to be able to do this, we have to adapt our current reverse-engineering methodologies to explicitly accommodate cell movements and tissue shape changes.

In this work we present an inference methodology that we have developed to reverse-engineer GRNs underlying pattern formation in tissues that are experiencing morphogenetic changes and cell rearrangements. As a case study we focus on T-box gene patterning in the developing zebrafish presomitic mesoderm (PSM) (Fig.1A). T-box genes coordinate fate specification along the PSM as cells move out of the tailbud and make their way towards the somites **Fulton et al.** (**2022**). Cell movements in the PSM can be live-imaged and followed in 3D **Thomson et al.** (**2021**). By the time they reach a somite, cells in the PSM will have undergone a stereotypical progression of T-box gene expression: Tbxta and Tbx16 in the tailbud, followed by Tbx16 in the posterior PSM and Tbx6 in the anterior PSM (Fig.1A&D). The Tbx16/Tbx6 boundary roughly marks the cells' transition out of the tailbud and in zebrafish it is thought to correlate with marked changes in cell behaviours where extensive cell mixing in the tailbud gives way to reduced, almost nonexistent, cell mixing and neighbourhood cohesion in the PSM **Mongera et al.** (**2018**). Therefore, while all cells will eventually have undergone the same gene expression progression, their expression dynamics will differ as cells spend variable amounts of time in the tailbud **Fulton et al.** (**2022**). Despite this, a tissue-level pattern forms which scales with PSM length during the course of posterior development and somitogenesis **Fulton et al.** (**2022**). T-box pattern formation in the developing zebrafish PSM is therefore a good example of a developmental process where the molecular pattern across the tissue is an emergent property of the GRN and the cell movements and tissue shape changes involved in the tissue's morphogenesis.

The reverse-engineering methodology presented in this paper accommodates cell movements within the developing tissue, hence taking tissue morphogenesis explicitly into account while reverse-engineering GRNs. To be able to do so, this methodology requires a combination of quantitative data: cell tracking data obtained from live-imaging the developing tissue and three-dimensional

quantitative gene expression of the genes and signalling pathways of interest over developmental time. We project the 3D gene expression data onto the cell tracks to approximate gene expression trajectories (AGETs) in single cells, hence approximating the gene expression dynamics in single cells. Using a subset of AGETs from ten cells randomly spaced within the tissue we were able to reverse-engineer candidate GRNs using a Markov Chain Monte Carlo (MCMC) approach. The fit of the resulting reverse-engineered GRNs is assessed by simulating them in each cell in the tracks using initial and boundary conditions extracted directly from the gene expression data, a methodology that we refer to as "live-modelling". The resulting well-fitting GRNs clustered into 22 clusters, generating candidate GRNs that can be further investigated and challenged using experimental work ***Fulton et al.*** (***2022***).

To our knowledge, this inference methodology is the first to integrate cell movements and gene expression data, making it possible to reverse-engineer GRNs patterning tissues undergoing morphogenetic changes. We hope that the reverse-engineering toolbox provided by our work will contribute to broaden the types of patterning systems that can be studied quantitatively and mechanistically, increasing our understanding of how pattern formation in development and evolution.

## Results and Discussion

### Approximating gene expression dynamics on single cell tracks: AGETs

The ideal data to reverse-engineer gene regulatory networks are temporally accurate quantifications of the gene expression dynamics at the single cell level as the tissue develops. Unfortunately, current state of the art in live gene expression reporter technology, while very advanced, cannot follow three genes and two signalling pathways simultaneously in space and time, while also ensuring that the dynamics of all reporters faithfully recapitulate the expression dynamics of the genes of interest. For this reason, it has been necessary to develop an alternative approach based on approximating the single cell gene expression trajectories in the developing PSM, which we will from now on refer to as AGETs (approximated gene expression trajectories).

In brief, AGETs are obtained by projecting spatial quantifications of gene expression in the PSM obtained using HCRs and antibody stains, onto each time frame of a time lapse of the developing PSM at approximately the same stage, and then reading out the projected expression level for each gene and signal, in every cell in the time lapse at every time point. The output result is an approximated gene expression trajectory for every cell in the time lapse, which can now be used to reverse-engineer gene regulatory networks which, when simulated on the tracks recapitulate T-box pattern formation on the developing PSM.

#### Data requirements and preparation

Two kinds of data are required to produce AGETs: cell tracks obtained from live-imaging the developing tissue of interest and quantitative spatial gene expression data at each developmental stage covered by the tracks.

In this case study, cell tracks were obtained by live-imaging a fluorescently labelled developing zebrafish tailbud between the 22nd and 25th somite stages using a two-photon microscope (see ***Thomson et al.*** (***2021***) and Materials and Methods). The raw data obtained consists of a series of point clouds representing the position of single cells in 3D space in 61 consecutive frames, taken at two minute intervals. The raw data were processed using a tracking algorithm in the image analysis software Imaris to obtain the position of single cells over time, and selected tracks were validated manually. The resulting data are a collection of cell tracks that describe the how individual cells move as the zebrafish tailbud and PSM develops. A cell track provides spatial information over time but is devoid of any information regarding gene expression levels in each cell.

Gene expression levels were approximated from fixed tailbud samples stained for the genes and signals of interest using HCR ***Choi et al.*** (***2018***) for the T-box gene products and antibody stains for the signals (see Materials and Methods). If gene expression patterns don't scale with the development of the tissue, stage-specific stains should be used separately. Otherwise, if the pattern of interest scales with tissue growth over developmental time - as is the case in the developing PSM - images at different stages can be quantified and pooled together. T-box genes - Tbxta, Tbx16 and Tbx6 - were simultaneously stained for on zebrafish tailbuds that had been fixed between the 23rd and 25th somite stages (SS) (Fig.1A). Of a total of 13 images, ten were processed and used for fitting (2x 23SS, 3x 24SS and 5x 25SS). Three separate antibody stained samples were used to quantify signals Wnt and FGF. In addition to the gene expression, tailbuds were stained with DAPI to be able to locate the cells. Only one side of the zebrafish PSM was used.

A processing pipeline (Fig.1A) was developed to quantify the imaging data, again using the image anaysis software Imaris (Fig.1). The first step in the pipeline consists of isolating the PSM from the surrounding tissues in the tailbud, including the spinal cord and the notochord. This was achieved by drawing a surface around the PSM using morphological and gene expression landmarks as a guide to identify different tissue boundaries (Fig.1B). Next, in order to consider only gene expression levels inside of the isolated PSM, all gene expression outside of the defined surface was set to zero (Fig.1C). Background noise in the data was reduced by setting lower-bound thresholds for every gene. These thresholds were chosen such that Tbxta and Tbx16 would appear restricted to the posterior end of the PSM (Fig.1Di, Dii, Ei and Eii) with their expression in the anterior PSM reduced to zero. Similarly, thresholds were set for Tbx6 expression such as to eliminate any background expression in the posterior PSM (Fig.1Diii and Eiii). Each gene is then normalized; normalization had to be robust to be able to deal with expression levels that could be very noisy in places. A Savitzky-Golay filter was applied to each gene to smoothen the signal (Fig.1D) and the smoothened maximum for each gene was set to one. Finally, spots were created in each detected nucleus from which a point cloud consisting of the 3D spatial coordinates and associated Tbxta, Tbx16 and Tbx6 levels were extracted (Fig.1E). The same pipeline was used to obtain the levels of signals Wnt and FGF in single cells.

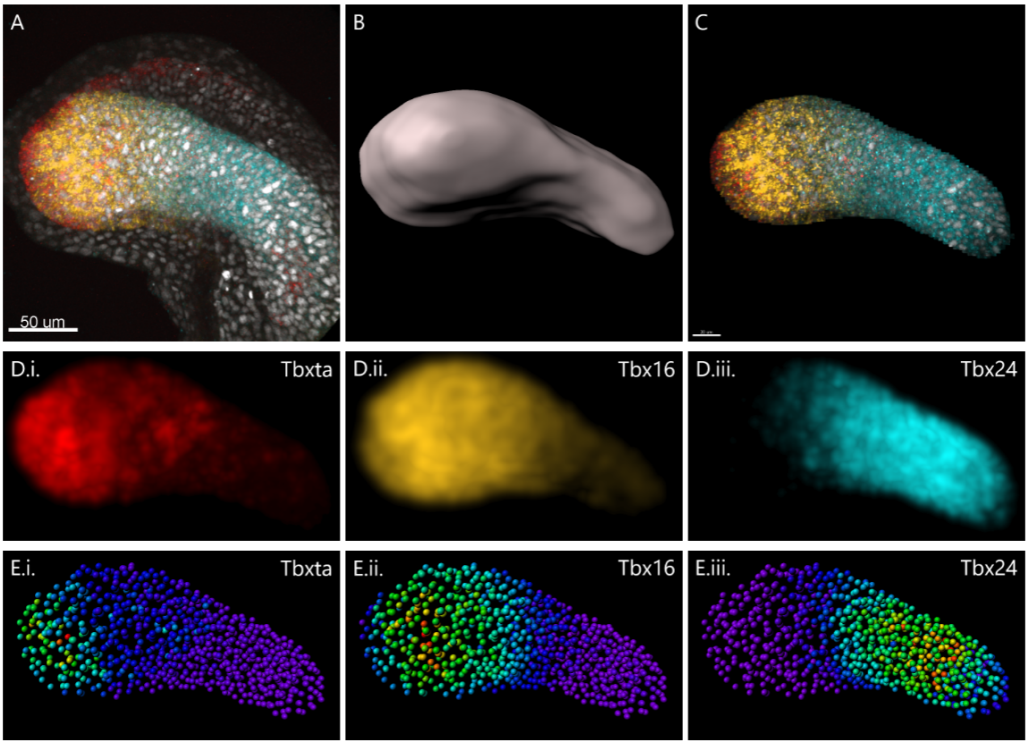


**Figure 1. Gene expression data preparation pipeline (A)** Typical HCR image of a 22 somite stage tailbud stained for Tbxta (red), Tbx16 (yellow), Tbx6 (blue) and DAPI (gray). Anterior to the right, posterior to the left, dorsal up and ventral down from here on. **(B)** Surface masking the PSM based on T-box expression and morphological landmarks. **(C)** Gene expression and nuclear marker in the isolated PSM (as before Tbxta in red, Tbx16 in yellow, Tbx6 in blue and DAPI in gray). **(D)** Normalising gene expression levels: Tbxta and Tbx16 levels in the anterior PSM are normalised to zero while posterior PSM levels of Tbx6 are normalised to zero, to eliminate background expression. A Gaussian filter has been then applied to each T-box gene to smoothen gene expression across the PSM. **(E)** Nuclei are segmented using the DAPI channel creating spots in 3D space. Spots are coloured according to the median intensity of each gene (i) Tbxta, ii) Tbx16 and iii) Tbx6 ), where purple denotes zero expression and red, highest expression. The spatial coordinates of the spots together with the median intensities were exported and used to generate the AGETs.

## AGET construction

AGETs are constructed to approximate the gene expression dynamics of single cells as they move and undergo complex re-arrangements during tissue morphogenesis. This requires live-imaging data, which provides information of the cell's spatial trajectories over time, to be combined with quantitative gene expression data at the single cell level. To achieve this, we begin by projecting the pre-processed HCR data onto the tracks, to in this way obtain a read-out of the gene expression and signalling levels that each cell experiences as it moves.

In order to project the extracted quantitative gene expression data onto the cell tracks, a first step is to align the point clouds representing the positions of the cells in 3D space processed from the HCRs (Fig.1E) with the point clouds for each of the 61 time frames in the time lapse (Fig.2A). We use point-to-plane ICP (iterative closest point) to perform this alignment (**??**), which in brief, is an iterative algorithm that seeks to map two point clouds onto each other by recursively minimising the distance between them (see Materials and Methods, and Fig.2). Once the point clouds have been aligned, equivalent regions of the different PSMs will overlap in space (Fig.2A) making it possible to map the quantitative gene expression from the processed HCRs onto the cells (represented by points) at each given time frame in the time lapse (Fig.2B and Algorithm1).

To approximate the gene expression and signalling values in a cell from the time lapse, we first

find its five closest neighbouring cells from the processed HCR data. Since all PSMs have been aligned as point clouds, we now have a point cloud representing cells from both the PSM in the time lapse and those from the HCRs. The median gene expression and signalling values are calculated from the expression and signalling values of the five nearest neighbouring cells and assigned to the cell from the time lapse (Fig.2B and see Algorithm 1 for a more detailed description of the process). Fig.2Bi shows the result of mapping T-box gene expression data from ten pre-processed HCR images onto the first frame of the tracking data while Fig.2Bii shows the quantified gene expression levels for all cells along the posterior to anterior axis. We repeat this procedure for each of the 61 frames in the time lapse resulting in an approximated gene expression trajectory (AGET) for every cell in the timelapse (Fig.2C).

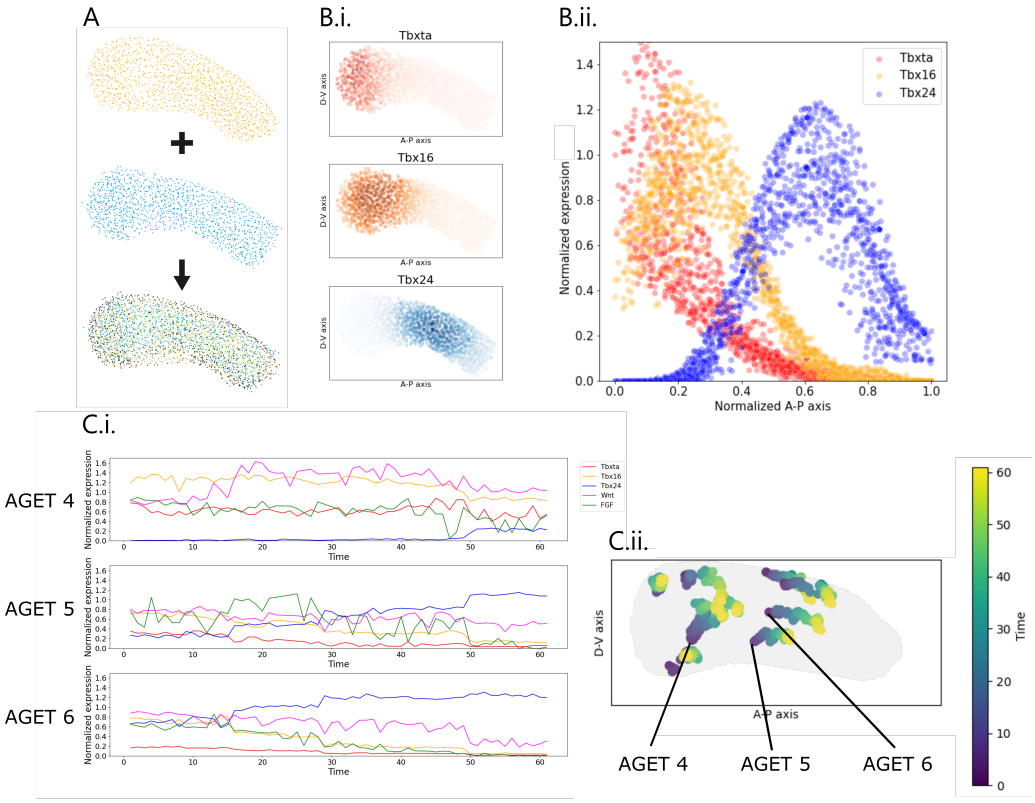


**Figure 2. Calculating AGETs. (A)** In orange is the processed HCR image showing the positions of the cells in the PSM (source point cloud) and in blue are the positions of the cells taken from the first frame of the tracking data (target point cloud). Using ICP, all the source point clouds obtained from the HCR images are aligned with the target point cloud obtained from each frame (61 in total) of the tracking data. This is illustrated by the overlapping orange, blue and black point clouds. **(B) i.** T-box gene expression from ten pre-processed HCR images, projected onto the first frame of the tracking data. Tbxta in red, Tbx16 in yellow and Tbx6 in blue. **ii.** Maximum projection of the data in i. quantified along the posterior to anterior axis. **(C) i.** For illustration purposes, three AGETs representing approximated T-box gene and signaling dynamics in three single cells at different position in the developing PSM (shown in C.ii). X-axis represents relative gene expression levels and y-axis reflects the time frame in the time lapse (from 1 to 61). Tbxta in red, Tbx16 in yellow and Tbx6 in blue, Wnt in pink, FGF in green. **(C) ii.** Ten cell tracks spanning the length of the PSM, whose AGETs were subsequently used for the GRN inference process. The color gradedness indicates position over time, where the initial position of the cell is show in purple and the final position is shown in yellow. The ten cells have been chosen semi-randomly. The outline illustrates the shape of the PSM. AGETs associated with cells 4, 5 and 6 are shown in panel C.i.

## Using AGETs to reverse-engineer gene regulatory networks that recapitulate pattern formation on a developing tissue

The motivation behind developing a methodology to construct AGETs is to be able to use them to reverse-engineer candidate gene regulatory networks that might be underlying a given patterning process when dynamic measurements or reliable approximations of gene expression at the single cell level are unavailable. In the previous section we have described how AGETs are constructed by combining tracking data with quantitative gene expression data. In this section, we present how AGETs can be used to reverse-engineer GRNs.

GRN models are often formulated as systems of coupled differential equations where state variables describe the concentrations of the genes of interest and parameters, the interactions between genes, as well as other factors such as production and degradation rates. In the case of the T-box genes, there are three state variables representing Tbxta, Tbx16 and Tbx6 levels and a total of 24 parameters to be fit (see Materials and Methods). Dynamic data are required to constrain and fit such models, and in this case these will be provided by the AGETs calculated previously. AGETs will therefore be used as the target expression dynamics for the fitting procedure instead of directly measured gene expression dynamics. As with other fitting procedures, an optimal parameter set will be one that minimises the difference between the target and the simulated data. We chose to adapt a Markov Chain Monte Carlo (MCMC) algorithm to use as our parameter sampling method since MCMC has been extensively used and repeatedly validated for GRN inference ***Ram and Chetty*** (***2009***). In addition, MCMC and has the advantage of providing an array of candidate networks by approximating the entire posterior distribution of all GRN parameters.

Using all 1903 available AGETs to fit our models would be ideal, as together they represent the tissue scale patterning dynamics that we seek to recapitulate. However, this is currently computationally expensive and unfortunately, unfeasible. Instead, we chose to fit to an ensemble of only ten AGETs that span the length of the PSM. These AGETs were selected semi-randomly, where a randomly chosen set of ten AGETs would be visually inspected to ensure that they together represented cells across the length of the PSM, and would otherwise be discarded. In addition, we only selected AGETs of maximal duration, namely those that corresponded to cells that had been consecutively tracked for the entire duration of the time lapse (61 frames). The ten AGETs used for reverse engineering and their approximate position in an idealised PSM is shown in Fig.2C.ii. For this case study, we found that reverse-engineering using ten AGETs generated well-fitting candidate networks while avoiding over-fitting and optimising the computational time required (see Materials and Methods).

MCMC inference outputs a collection of parameter sets or combinations (samples) that together approximate the posterior distribution of the GRN parameters: for every parameter, we obtain a probability distribution across its values, which provides information about the values that are most likely to produce good fits. We first chose to explore the network corresponding to the parameter set with the overall highest posterior probability score, that is the maximum a posteriori or MAP sample (Fig.3A). We simulated each of the ten AGETs used during the fitting procedure and then proceeded to simulate all 1903 available AGETs, that is, we simulated on the tracks (which we refer to as live-modelling). We qualitatively validate the quality of the inference by both comparing single AGETs with their simulations (Fig.3B), and by comparing the whole tissue-level gene expression profiles over time (Fig.3C). We are especially interested in how well the simulations recapitulate the whole tissue patterns, as these result from simulating AGETs that had not been used for model training.

We discard parameter sets that simulate clear pattern aberrations, and consider a good fit when the position of gene expression domain intersections does not differ by more than the inter-embryonic biological boundary range (<10% A-P position) in the simulated versus the approximated patterns (Fig.3C). While quantitative measures of the goodness of fit can be easily defined,

such as comparing the log-likelihood between parameter sets or calculating least-squares measures, these don't necessarily reflect whether aspects of the pattern that are of notable biological importance are being captured, and were therefore not favoured in this part of the analysis.

Fig.3B.i compares four of the ten AGETs (relative positions shown in Fig.3B.ii) (solid lines) used for model fitting with the resulting simulations (dotted lines). The simulated expression recapitulates well the target expression for the AGETs. The model was formulated as a deterministic system without added stochasticity which explains the smoothness of the simulated curves, which nonetheless can be seen to recapitulate AGET gene expression levels and trends. In other systems, fits might be improved by setting smaller standard deviations. Given that the AGETs had been obtained from a small data set and are noisy, we avoided over-fitting and were satisfied with these fits. Fig.3C shows simulated T-box expression for each cell along the normalized posterior to anterior axis of the PSM (dots). The simulated data have been fit at each separate time point by curves which are then normalised (dotted curves) and compared to the curves previously fit in the same way to all AGETs (shown as solid curves). A comparison between AGETs and simulations is shown at three different time points in Fig.3C (simulation outputs at 33%, 66% and 100% total time respectively). Importantly, the overall position of the domains is recapitulated and the the position of domain intersections is within the preset biological range of 10% A-P position. The full simulation is shown in Movie 1.

Notably, there is a discrepancy between the AGETs and the simulated anterior Tbx6 expression. The formulated GRN is unrealistic in this region, where additional factors secreted from the somites are known to be down-regulating this transcription factor ***Kawamura et al.*** (***2005***). For this reason, it is reassuring and expected that the model doesn't recapitulate the pattern well in this region. In addition, the model predicts that over time, a small percentage of posterior cells will express low levels of Tbx6. Although unexpected, there is evidence suggesting that this is indeed the case ***Fulton et al.*** (***2022***). Such low and sparse anterior expression of Tbx6 would have been lost during the smoothing step in our data preparation pipeline, which is unable of capturing patterns of such fine resolution as it stands. It is encouraging that candidate GRNs consistently recapitulate this unexpected feature of the biology and might suggest that the three genes considered are indeed causally responsible for most of the biological pattern.

In summary, we have been able to infer the parameters of candidate GRNs which recapitulate the global pattern of T-box expression in the zebrafish PSM by fitting to ten spaced and pseudo-randomly chosen AGETs.

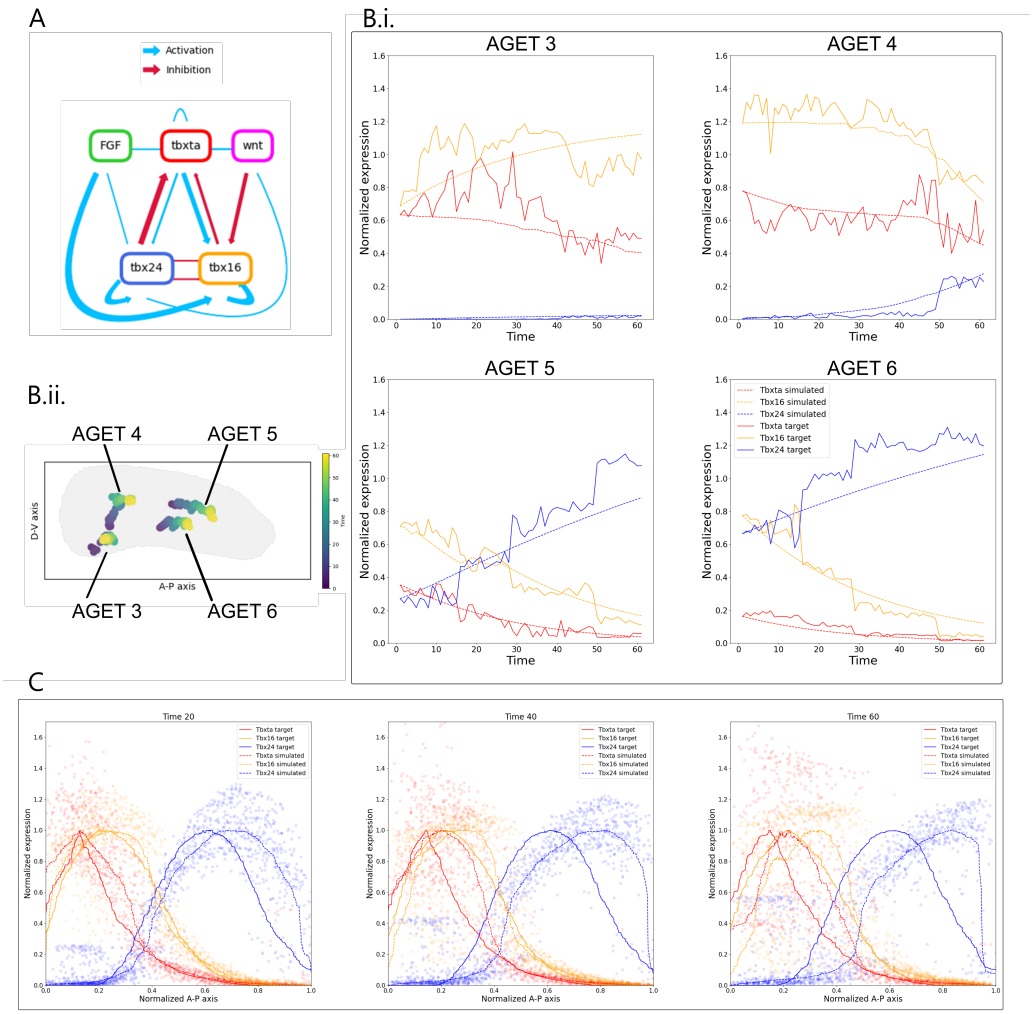


**Figure 3. Performance and fit of the GRN corresponding to the maximum a posteriori (MAP) parameters. (A)** GRN topology with MAP parameters obtained from the MCMC inference. **(B)**i. Simulated data (dotted curves) for four of the ten AGETs (solid curves) used for model fitting **(B)**ii. Illustrative spatial location in the PSM of the four AGETs shown in B.i. **(C)** Snapshots showing simulated T-box gene expression along the normalized posterior (0) to anterior (1) axis of the PSM (dots) at 33%, 66% and 100% of total simulation time respectively. Dotted curves represent the simulated data once they have been fit by smooth curves at each separate time point and normalised. Solid curves represent all AGETs fit and normalised in the same way

## Parameter determinability and model clustering

MCMC is a parameter sampling algorithm, and as such it will return an approximated posterior distribution for the GRN parameters instead of a single estimate. This provides a range of candidate networks that can be subsequently analysed and challenged in combination with experimental approaches. Such parameter distributions also provide valuable information regarding which model parameters — and therefore genetic interactions — are tightly constrained by the data, and which aren't and therefore appearing to take on a broad range of values across the inferred networks. Such information can lead to interesting hypotheses regarding which aspects of the pattern selection might be most strongly working on.

While in the previous section we analysed the parameter set with the maximal posterior probability (MAP) to asses the goodness of fit of one of the candidate GRNs, in this section we assess how well the posterior distribution has been approximated across candidate GRNs (Fig.4)). To do this,

we selected 1000 parameter sets at random from the posterior distribution, representing 1000 distinct candidate networks. We then proceeded to cluster them according to the similarity of their parameter values using agglomerative hierarchical clustering (see Materials and Methods). In order to be able to choose a representative to explore further for each cluster, we set the condition that the parameter distributions within clusters should be uni-modal. After imposing this condition, the algorithm returned 30 clusters and the network with mean parameter values was picked as the representative for each cluster. We simulated the resulting 30 networks and compared them with AGETs 1-10 used for fitting. The simulations were visually inspected and networks returning aberrant patterns were discarded along with all the networks in the cluster that they belonged to. This process left a total of 22 clusters of well-fitting GRNs (Fig.4).

Fig.4 shows the topology of the respresentative GRNs in each of the resulting 22 clusters. By topology we mean whether parameters are positive (blue) or negative (negative). This provides only a superficial illustration of the clusters which is useful for visualisation purposes only, leaving out much of the complexity within these classes since the clustering was done on the quantitative value of the parameters. For this reason too, it might appear that representative networks of different clusters are the same, however although that might be the case qualitatively (taking only into account parameter signs), it isn't the case quantitatively (networks 26, 22, 13, 12, 10, 2 and 6). 10 out of 24 parameters were set as positive in the priors (Fig.4, round blue circles; see Materials and Methods for justification), the remaining 14 could adopt either positive or negative values. These correspond to parameters that represent the interaction strengths between T-box genes and from Wnt and FGF to the T-box genes. The global probability of an activation (positive parameter) is shown above each corresponding column in Fig.4. Generally, for each parameter there is a clear preference across all clusters, suggesting a degree of constraint in the determinability of parameters. We also recorded the in-sample log-likelihood of each network as a measure of how well these networks fit the data (Fig.4, right). Given how close these values are, we want to emphasise at this point that they should all be treated as likely candidates and that further biological knowledge and experiments will be required to discriminate between them ***Fulton et al.*** (***2022***).


**Figure 4. GRN clusters.** The topologies of the mean networks are shown for the 22 well-fitting clusters recovered by the fitting. Rows correspond to representative networks from each cluster, columns represent individual GRN parameters. Quantitative parameters are reduced to whether they are positive or negative for illustration purposes. This can give the impression that some networks and clusters are the same, when in fact they are quantitatively distinct. The percentage above a given parameter indicates the probability that said parameter is positive across clusters. Parameters marked with a blue circle were defined as positive by the prior. In-sample log-likelihood for each network is provided as a measure of goodness of fit.

## Conclusion

Earlier reverse-engineering frameworks have been unable to accommodate the role of cell rearrangements and tissue shape changes in the formation of developmental patterns. This limitation has heavily biased quantitative studies of pattern formation towards processes in systems where

the timing of pattern formation and morphogenesis can be separated. Unfortunately, the vast majority of patterning processes in animal development do not meet this criterion and in consequence, are largely missing from the literature on pattern formation. Furthermore, as a field, most of our collective knowledge and understanding of the generation and evolution of developmental patterns has been constructed on the omission of any role that might be played by cell movements, tissue shape changes and other morphogenetic mechanisms.

This need not be the case going forward. Thanks to recent advancements in live-imaging and spatial gene expression quantification, the data required to adopt the reverse-engineering framework presented in this paper is becoming available in an ever-increasing number of species spanning the range of animal phylogeny. This will make it possible to construct AGETs and infer GRNs in a wider range of systems. Simulation and subsequent analysis of patterning processes that are dependent on cell movements will increase our understanding of pattern formation and its evolution, and uncover previously buried general principles that weren't accessible from the restricted number of systems that we were studying. Furthermore, this methodology will find applications well-beyond beyond the study of developmental evolution. In particular, we anticipate a warm reception from fields such as bio-engineering, regenerative medicine and organoid biology, where understanding how 3D cell cultures should be shaped and constrained as they grow to obtain the desired final organisation is paramount and has proven not at all trivial.

## Materials and Methods

### Animal lines and husbandry

This research was regulated under the Animals (Scientific Procedures) Act 1986 Amendment Regulations 2012 following ethical review by the University of Cambridge Animal Welfare and Ethical Review Body (AWERB). Embryos were obtained and raised in standard E3 media at 28°C. Wild Type lines are either Tupfel Long Fin (TL), AB or AB/TL. The Tg(7xTCF-Xla.Sia:GFP) reporter line ***Moro et al.*** (***2012***) was provided by the Steven Wilson laboratory. Embryos were staged as in ***Kimmel et al.*** (***1995***).

### In Situ Hybridisation Chain Reaction (HCR)

Embryos were incubated until they reached the the desired developmental stage, then fixed in 4% PFA in DEPC treated PBS without calcium and magnesium, and stored at 4°C overnight. Once fixed, embryos were stained using HCR version 3 following the standard zebrafish protocol found in ***Choi et al.*** (***2018***). Probes, fluorescent hairpins and buffers were all purchased from Molecular Instruments. After staining, samples were stained with DAPI and mounted using 80% glycerol.

### Immunohistochemistry

Embryos were incubated until they reached the desired developmental stage, then fixed in 4% PFA in DEPC treated PBS without calcium and magnesium, and stored at 4°C overnight. The embryos were subsequently blocked in 3% goat serum in 0.25% Triton, 1% DMSO, in PBS for one hour at room temperature. Our read-out for FGF activity - Diphosphorylated ERK - was detected using the primary antibody (M9692-200UL, Sigma) diluted 1 in 500 in 3% goat serum in 0.25% Triton, 1% DMSO, in PBS. All samples were incubated at 4°C overnight and then washed in 0.25% Triton, 1% DMSO, in PBS. Secondary Alexa 647nm conjugated antibodies were diluted 1 in 500 in 3% goat serum in 0.25% Triton, 1% DMSO, 1X DAPI in PBS and applied overnight at 4°C.

### Imaging and image analysis

Fixed HCR and immunostained samples were imaged with a Zeiss LSM700 inverted confocal at 12 bit, using either 20X or 40X magnification and an image resolution of 512x512 pixels. Nuclear segmentation of whole stained embryonic tailbuds was performed using a tight mask applied around the DAPI stain using Imaris (Bitplane) with a surface detail of 0.5µm. Positional values for each

nucleus were exported as X, Y, Z coordinates relative to the posterior-most tip of the PSM where X, Y, Z were equal to (0, 0, 0). The PSM was then segmented by hand by deleting nuclear surfaces outside of the PSM, including notochord, spinal cord, anterior somites and ectoderm. PSM length was normalised individually between 0 and 1 by division of the position in X by the maximum X value measured in each embryo.

Single cell image analysis was conducted using Imaris (Bitplane) by generating loose surface masks around the DAPI stain to capture the full nuclear region and a small region of cytoplasm. Surface masks were then filtered to remove any masks where two cells joined together or small surfaces caused by background noise, or fragmented apoptotic nuclei. The intensity sum of each channel was measured and normalised by the area of the surface. Expression level was then normalised between 0 and 1 using the maximum value measured for each gene, in each experiment.

Live imaging datasets of the developing PSM were created using a TriM Scope II Upright 2-photon scanning fluorescence microscope equipped Insight DeepSee dual-line laser (tunable 710-1300 nm fixed 1040 nm line) (see details in ***Thomson et al.*** (***2021***)). The developing embryo was imaged with a 25X 1.05 NA water dipping objective. Embryos were positioned laterally in low melting agarose with the entire tail cut free to allow for normal development ***Hirsinger and Steventon*** (***2017***). Tracks were generated automatically and validated manually using the Imaris imaging software.

## Aligning point clouds with ICP

We used the Python library Open3d ***Zhou et al.*** (***2018***) and the implementation of the point-to-plane ICP (Iterative Closest Point) algorithm therein ***Rusinkiewicz and Levoy*** (***2001***) to perform the point cloud alignment. ICP algorithms can be used to align two point clouds from an initial approximate alignment. The aim is to find a transformation matrix, that rotates and moves the source point cloud in a way that achieves an optimal alignment with the target point cloud. ICP algorithms work by iterating two steps. First, for each point in the source point cloud, the algorithm will determine the corresponding closest point in the target point cloud. Second, the algorithm will find the transformation matrix that most optimally minimizes the distances between the corresponding points. The result is a transformed source point cloud that is closely aligned with the target point cloud. As a pre-processing step, the source and target point clouds have been re-scaled to have the same A-P length. Since we are working with biological tissues, point clouds will not correspond exactly, differing slightly in size and shape. This will impact the quality of the resulting alignment which had to be visually assessed and validated. In this case study, three of the thirteen source images were excluded from the analysis due to poor alignment.

## AGET construction

While the main methodology used for constructing AGETs is covered in the results section, below (Algorithm 1) we provide pseudo-code that describes the same process.

**Algorithm 1:** Mapping T-box gene expression from HCR images onto tracking data

**Result:** Cell tracks with T-box gene expression information
Create target point clouds from tracking data $Target_i$, for every time point $i \in 1, ..., 61$;
Create source point clouds with gene expression information from HCR images $Source_j$, for every source image $j \in 1, ..., 10$;
**for** *i in* $1 : 61$ **do**
  **for** *j in* $1 : 10$ **do**
    Align $Source_j$ and $Target_i$ with ICP registration;
    **for** *Every point* $Cell_k$ *in* $Target_i$ **do**
      Find n=5 closest neighbours of $Cell_k$ in $Source_j$;
      Calculate median $M_{ijk}$ of closest neighbours;
      Assign $M_{ijk}$ to $Cell_k$;
    **end**
  **end**
  **for** *Every point* $Cell_k$ *in* $Target_i$ **do**
    Calculate median $M_{ik}$ of medians $M_{ijk}$ from 10 source point clouds $Source_{1:10}$;
    Assign $M_{ik}$ to $Cell_k$;
  **end**
**end**
Extract all cell tracks with their assigned gene expression.

## Mathematical model formulation

We used a dynamical systems formulation model the T-box gene regulatory network in the zebrafish PSM. The model's aim is to recapitulate the dynamics of T-box gene expression in every cell in the developing zebrafish PSM, generating the emergence of the tissue-level T-box gene expression pattern. We use a connectionist model formulation which has been extensively used and validated to previously model other developmental patterning processes ***Mjolsness et al.*** (***1991***); ***Jaeger et al.*** (***2004***); ***Crombach et al.*** (***2012***).

The mRNA concentrations encoded by the T-box genes *tbxta*, *tbx16* and *tbx6* are represented by the state variables of the dynamical system. For each gene, the concentration of its associated mRNA $a$ at time $t$ is given by $g^a(t)$. mRNA concentration over time is governed by the following system of three coupled ordinary differential equations:

$$\frac{dg_a(t)}{dt} = R_a\phi(u_a) - \lambda_a g_a(t) \tag{1}$$

where $R^a$ and $\lambda^a$ respectively represent the rates of mRNA production and decay. $\phi$ is a sigmoid regulation-expression function used to represent the cooperative, saturating, coarse-grained kinetics of transcriptional regulation and introduces non-linearities into the model that enable it to exhibit complex dynamics:

$$\phi(u_a) = \frac{1}{2}\left(\frac{u_a}{\sqrt{(u_a)^2 + 1}} + 1\right), \tag{2}$$

where

$$u_a = \sum_{b \in G} W^{ba} g_b(t) + \sum_{s \in S} E^{sa} g_s(t) + h_a. \tag{3}$$

$G = \{tbxta, tbx16, tbx6\}$ refers to the set of T-box genes while $S = \{\mathrm{Wnt}, \mathrm{FGF}\}$ represents the set of external regulatory inputs provided by the Wnt and FGF signalling environments. The concentrations of the external regulators $g_s$ are provided directly from the AGETs into the simulation and are not themselves being modelled. Changing Wnt and FGF concentrations over time renders

the parameter term $\sum_{s \in S} E^{sa} g_s(t)$ time-dependent and therefore, the model non-autonomous ***Collier et al.*** (***1996***); ***Verd et al.*** (***2014***).

The inter-connectivity matrices $W$ and $E$ house the parameters representing the regulatory interactions among the T-box genes, and from Wnt and FGF to the T-box genes, respectively. Matrix elements $w^{ba}$ and $e^{sa}$ are the parameters representing the effect of regulator $b$ or $s$ on target gene $a$. These can be positive (representing an activation from $b$ or $s$ onto $a$), negative (representing a repression), or close to zero (no interaction). $h_a$ is a threshold parameter denoting the basal activity of gene $a$, which acknowledges the possible presence of regulators absent from our model. To perform the live-modelling simulations, the same model formulation is implemented in each cell in the time-lapse. Initial concentrations of *tbxta*, *tbx16* and *tbx6* are read out directly from the first time point of the AGET corresponding to that cell, and dynamic Wnt and FGF values are updated from the same AGET.

### Model fitting: MCMC approach

We used the Markov Chain Monte Carlo approach implemented in the Python emcee library ***Foreman-Mackey et al.*** (***2013***) to approximate the posterior distribution of the GRN parameters. A property of this implementation is the use of an ensemble of walkers, rather than a single one. We used a uniform prior over a reasonably large range of values and fitted to the time scale used in the simulation. The time scale was chosen such that 1 equals the time that the fastest cell takes to travel through the whole PSM and enter a somite. We used a Gaussian distribution with fixed standard deviations per gene to model the differences between simulated gene expression and target gene expression approximated by the AGETs, and in this way obtain a likelihood function. We ran the MCMC with 70 walkers and for a total of 50'000 steps. Although the auto-correlation time was high and the acceptance fraction with 4.1% was on the low side, the inferred parameters lead to well-fitting simulated data. Model training took approximately three days using 20 cores.

## Acknowledgments

Thanks to the Cambridge Advanced Imaging Centre (CAIC) for imaging support.

## References

**Averbukh I**, Lai SL, Doe CQ, Barkai N. A repressor-decay timer for robust temporal patterning in embryonic Drosophila neuroblast lineages. Elife. 2018; 7:e38631.

**Balaskas N**, Ribeiro A, Panovska J, Dessaud E, Sasai N, Page KM, Briscoe J, Ribes V. Gene regulatory logic for reading the Sonic Hedgehog signaling gradient in the vertebrate neural tube. Cell. 2012; 148(1-2):273–284.

**Choi HM**, Schwarzkopf M, Fornace ME, Acharya A, Artavanis G, Stegmaier J, Cunha A, Pierce NA. Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. Development. 2018; 145(12):dev165753.

**Cohen M**, Kicheva A, Ribeiro A, Blassberg R, Page KM, Barnes CP, Briscoe J. Ptch1 and Gli regulate Shh signalling dynamics via multiple mechanisms. Nature communications. 2015; 6(1):1–12.

**Collier JR**, Monk NA, Maini PK, Lewis JH. Pattern formation by lateral inhibition with feedback: a mathematical model of delta-notch intercellular signalling. Journal of theoretical Biology. 1996; 183(4):429–446.

**Crombach A**, Wotton KR, Cicin-Sain D, Ashyraliyev M, Jaeger J. Efficient reverse-engineering of a developmental gene regulatory network. PLoS computational biology. 2012; 8(7):e1002589.

**Crombach A**, Wotton KR, Jiménez-Guri E, Jaeger J. Gap gene regulatory dynamics evolve along a genotype network. Molecular biology and evolution. 2016; 33(5):1293–1307.

**D'haeseleer P**, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics. 2000; 16(8):707–726.

**El-Sherif E**, Zhu X, Fu J, Brown SJ. Caudal regulates the spatiotemporal dynamics of pair-rule waves in Tribolium. PLoS genetics. 2014; 10(10):e1004677.

**Foreman-Mackey D**, Hogg DW, Lang D, Goodman J. emcee: the MCMC hammer. Publications of the Astronomical Society of the Pacific. 2013; 125(925):306.

**Fulton T**, Speiss K, Thomson L, Wang Y, Clark B, Hwang S, Paige B, Verd B, Steventon B. Cell Rearrangement Generates Pattern Emergence as a Function of Temporal Morphogen Exposure. bioRxiv. 2022; .

**Gardner TS**, Faith JJ. Reverse-engineering transcription control networks. Physics of life reviews. 2005; 2(1):65–88.

**He F**, Balling R, Zeng AP. Reverse engineering and verification of gene networks: principles, assumptions, and limitations of present methods and future perspectives. Journal of biotechnology. 2009; 144(3):190–203.

**Hirsinger E**, Steventon B. A versatile mounting method for long term imaging of zebrafish development. JoVE (Journal of Visualized Experiments). 2017; (119):e55210.

**Huch M**, Knoblich JA, Lutolf MP, Martinez-Arias A. The hope and the hype of organoid research. Development. 2017; 144(6):938–941.

**Jaeger J**, Blagov M, Kosman D, Kozlov KN, Manu, Myasnikova E, Surkova S, Vanario-Alonso CE, Samsonova M, Sharp DH, et al. Dynamical analysis of regulatory interactions in the gap gene system of Drosophila melanogaster. Genetics. 2004; 167(4):1721–1737.

**Jaeger J**, Monk NA. Reverse Engineering of Gene Regulatory Networks. Learning and inference in computational systems biology. 2010; 9:34.

**Kawamura A**, Koshida S, Hijikata H, Ohbayashi A, Kondoh H, Takada S. Groucho-associated transcriptional repressor ripply1 is required for proper transition from the presomitic mesoderm to somites. Developmental cell. 2005; 9(6):735–744.

**Kicheva A**, Bollenbach T, Ribeiro A, Valle HP, Lovell-Badge R, Episkopou V, Briscoe J. Coordination of progenitor specification and growth in mouse and chick spinal cord. Science. 2014; 345(6204).

**Kicheva A**, Cohen M, Briscoe J. Developmental pattern formation: insights from physics and biology. Science. 2012; 338(6104):210–212.

**Kimmel CB**, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. Stages of embryonic development of the zebrafish. Developmental dynamics. 1995; 203(3):253–310.

**Liang S**, Fuhrman S, Somogyi R, et al. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In: *Pacific symposium on biocomputing*, vol. 3 Citeseer; 1998. p. 18–29.

**Manu S Surkova**, Spirov AV, Gursky VV, Janssens H, Kim AR, Radulescu O, Vanario-Alonso CE, Sharp DH, Samsonova M, Reinitz J. Canalization of gene expression and domain shifts in the Drosophila blastoderm by dynamical attractors. PLoS computational biology. 2009; 5(3):e1000303.

**Mjolsness E**, Sharp DH, Reinitz J. A connectionist model of development. Journal of theoretical Biology. 1991; 152(4):429–453.

**Mongera A**, Rowghanian P, Gustafson HJ, Shelton E, Kealhofer DA, Carn EK, Serwane F, Lucio AA, Giammona J, Campàs O. A fluid-to-solid jamming transition underlies vertebrate body axis elongation. Nature. 2018; 561(7723):401–405.

**Moro E**, Ozhan-Kizil G, Mongera A, Beis D, Wierzbicki C, Young RM, Bournele D, Domenichini A, Valdivia LE, Lum L, et al. In vivo Wnt signaling tracing through a transgenic biosensor fish reveals novel activity domains. Developmental biology. 2012; 366(2):327–340.

**Ram R**, Chetty M. MCMC based Bayesian inference for modeling gene networks. In: *IAPR International Conference on Pattern Recognition in Bioinformatics* Springer; 2009. p. 293–306.

**Rayon T**, Stamataki D, Perez-Carrasco R, Garcia-Perez L, Barrington C, Melchionda M, Exelby K, Lazaro J, Tybulewicz VL, Fisher EM, et al. Species-specific pace of development is associated with differences in protein stability. Science. 2020; 369(6510).

**Reinitz J**, Sharp DH. Gene circuits and their uses. Integrative Approaches to Molecular Biology. 1996; p. 253–272.

**Rockman MV**. Reverse engineering the genotype–phenotype map with natural genetic variation. Nature. 2008; 456(7223):738–744.

**Rusinkiewicz S**, Levoy M. Efficient variants of the ICP algorithm. In: *Proceedings third international conference on 3-D digital imaging and modeling* IEEE; 2001. p. 145–152.

**Schröter C**, Ares S, Morelli LG, Isakova A, Hens K, Soroldoni D, Gajewski M, Jülicher F, Maerkl SJ, Deplancke B, et al. Topology and dynamics of the zebrafish segmentation clock core circuit. PLoS biology. 2012; 10(7):e1001364.

**Thomson L**, Muresan L, Steventon B. The zebrafish presomitic mesoderm elongates through compaction-extension. Cells & Development. 2021; .

**Uriu K**, Morishita Y, Iwasa Y. Random cell movement promotes synchronization of the segmentation clock. Proceedings of the National Academy of Sciences. 2010; 107(11):4979–4984.

**Verd B**, Clark E, Wotton KR, Janssens H, Jiménez-Guri E, Crombach A, Jaeger J. A damped oscillator imposes temporal order on posterior gap gene expression in Drosophila. PLoS biology. 2018; 16(2):e2003174.

**Verd B**, Crombach A, Jaeger J. Classification of transient behaviours in a time-dependent toggle switch model. BMC systems biology. 2014; 8(1):1–19.

**Verd B**, Crombach A, Jaeger J. Dynamic maternal gradients control timing and shift-rates for Drosophila gap gene expression. PLOS Computational Biology. 2017; 13(2):e1005285.

**Verd B**, Monk NA, Jaeger J. Modularity, criticality, and evolvability of a developmental gene regulatory network. Elife. 2019; 8:e42832.

**Wu H**, Jiao R, Ma J, et al. Temporal and spatial dynamics of scaling-specific features of a gene regulatory network in Drosophila. Nature communications. 2015; 6(1):1–13.

**Zhou QY**, Park J, Koltun V. Open3D: A Modern Library for 3D Data Processing. arXiv:180109847. 2018; .