

RESEARCH ARTICLE

***Tbx16* and *mesogenin 1* promote presomitic mesoderm differentiation by repressing the mesodermal progenitor cell state**

Guoyu Zhu, Miriam A. Genuth¹, Yanrong Xiao¹, Abigail A. Kindberg, Kayleigh Hackett, Scott A. Holley¹*

Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut, United States of America

☞ These authors contributed equally to this work.

* scott.holley@yale.edu



OPEN ACCESS

Citation: Zhu G, Genuth MA, Xiao Y, Kindberg AA, Hackett K, Holley SA (2026) *Tbx16* and *mesogenin 1* promote presomitic mesoderm differentiation by repressing the mesodermal progenitor cell state. PLoS Genet 22(6): e1012176. <https://doi.org/10.1371/journal.pgen.1012176>

Editor: Giovanni Bosco, Geisel School of Medicine at Dartmouth, UNITED STATES OF AMERICA

Received: December 12, 2025

Accepted: May 21, 2026

Published: June 8, 2026

Copyright: © 2026 Zhu et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: Data are available at GEO with accession number GSE314056.

Funding: o Funding provided by R35GM148348 to SAH and F32HD111328 to AAK. The funders had no role in study design, data collection and

Abstract

During zebrafish embryonic body elongation, differentiation of mesodermal progenitors into presomitic mesoderm requires the transcription factors *tbx16* and *mesogenin 1*. Here, by using temporally controlled *tbx16* and *mesogenin 1* overexpression and RNAseq to identify immediate downstream changes in gene expression, we elucidate how these genes promote presomitic mesoderm differentiation. Using machine learning and game theory, we integrated differentially expressed genes with wild-type scRNAseq data and identified genes downstream of *tbx16* and *mesogenin 1* during mesoderm differentiation. This data-driven analysis indicates that *mesogenin 1* and *tbx16* primarily repress expression of genes as mesodermal progenitors differentiate. Strikingly, the genes that are most important for defining transcriptional cell states during mesoderm differentiation are most strongly repressed by *tbx16* and *mesogenin 1*. Moreover, these downstream effectors are enriched for genes with known roles in mesoderm development and body elongation such as Fgf, Wnt and Bmp pathways and the transcription factors *tbxta*, *eve1*, *hoxd12a*, *hoxd13b*, *lef1*, *cdx4*, *tbx16l*, *ved*, *vent* and *vox*. Gradients of Fgf and Wnt specify the mesodermal progenitor state in the posterior tailbud and activate many of these transcription factors indicating that *tbx16* and *mesogenin 1* promote mesoderm differentiation by repressing this progenitor state.

Author summary

Mesodermal progenitor cells continuously disperse and differentiate during vertebrate body elongation and give rise to the muscle, bone and dermis of the trunk and tail. In zebrafish, two genes, the transcription factors *tbx16* and *mesogenin 1*, function semi-redundantly during differentiation of mesodermal progenitors. However, the genes downstream of *tbx16* and *mesogenin 1* that mediate

analysis, decision to publish, or preparation of the manuscript. G.Z., M.A.G. and S.A.H. received salary from R35GM148348, and A.A.K. received salary from F32HD111328.

Competing interests: The authors have declared that no competing interests exist.

mesoderm morphogenesis remain unknown. To identify these downstream genes, we performed RNA sequencing shortly after over-expression of *tbx16* or *mesogenin 1*. These types of experiments typically provide a long list of genes, and it is challenging to identify the most important of the downstream genes for further analysis. We address this challenge by integrating the RNA sequencing data with single cell RNA sequencing data of wild-type mesoderm differentiation. We employ machine learning to sort gene expression into distinct states and game theory to identify the most important genes for defining each state. Comparing these genes with the gene expression changes after *tbx16* or *mesogenin 1* over-expression, we identify a smaller number of genes that are enriched in genes with known functions in mesodermal differentiation. This suggests that this data analysis strategy can identify important biological information from large genomic datasets.

Introduction

As the post-gastrulation vertebrate embryo elongates, it forms the posterior trunk and tail concomitant with segmentation of the paraxial mesoderm into somites. The tailbud is the posterior growth zone and contains progenitors of the spinal cord, vertebral column and skeletal muscle. Genetic analyses and live imaging experiments of the zebrafish tailbud found that posterior body elongation is largely driven by cell migration and not cell proliferation [1–5]. Neuromesodermal progenitors (NMP) in the dorsal medial posterior tailbud differentiate into either spinal cord or mesodermal progenitors [6–8]. NMPs that differentiate into mesodermal progenitors undergo an epithelial to mesenchymal transition (EMT) then migrate ventrally into the progenitor zone, where they exhibit disordered cell motion [8–11]. These cells subsequently assimilate into the presomitic mesoderm (PSM) and cell movement diminishes as the tissue solidifies [1,12,13]. This developmental progression has been demarcated by machine learning into a series of cell state transitions in both gene expression and cell motion as NMPs differentiate first into mesodermal progenitors, then into a motile progenitor zone, then into posterior PSM and then anterior PSM [14]. Here, we examine the roles of two transcription factors, *tbx16* and *mesogenin 1*, in paraxial mesoderm differentiation.

Previous studies indicate that *tbx16* and *msgn1* have both independent and redundant functions in paraxial mesoderm development. Loss of either *tbx16* or *msgn1* function leads to a failure of PSM differentiation with loss of *tbx16* having a much stronger phenotype [15–19]. *Tbx16* was first identified as a recessive lethal mutation in zebrafish called *spadetail* (*spt-1*) [15]. Embryos homozygous for this mutation exhibit a bent posterior body, an accumulation of cells in the posterior tailbud incapable of transitioning into PSM, and a corresponding lack of somitic mesoderm. Subsequently, the gene mutated in *spadetail* was identified as a t-box transcription factor, named *tbx16*, with homologs in *Xenopus* and chick [16,20–24]. In the zebrafish embryo, *tbx16* is expressed in the mesodermal progenitors, the PZ domain and the posterior PSM [16]. *Tbx16* genetically interacts with other t-box transcription

factors *tbxta* (*brachyury/no tail*) and *tbx6* within a gene network to regulate mesoderm development [16,17,25,26]. Wnt and FGF signaling activate expression of *tbxta* and *tbx16* to promote NMP differentiation into mesodermal progenitors [8,9,16,27,28]. *Tbx16* regulates *hox* gene expression in the tailbud and induces intermediate mesodermal fates by regulating Wnt, retinoic acid and FGF signaling [29–31]. As a transcription factor, Tbx16 can act as both an activator and a repressor. In the zebrafish gastrula, Tbx16 directly activates *myf5* and *myoD* transcription downstream of FGF during myogenesis [32]. In the tailbud, Tbx16 represses *sox2* transcription to drive NMPs to differentiate as mesoderm [28]. Despite these extensive analyses, we still do not know the genes downstream of *tbx16* that carryout differentiation.

Mesogenin is a basic-helix-loop-helix (bHLH) transcription factor expressed in the tailbud and first identified in chick [33]. In mouse and *Xenopus*, *mesogenin* is crucial for paraxial mesoderm specification [34,35]. Zebrafish *mesogenin 1* (*msgn1*) is expressed in the posterior tailbud, in the mesodermal progenitors, PZ domain and posterior PSM in a pattern very similar to that of *tbx16*. *Msgn1* expression is sharply reduced in *spt* mutant embryos suggesting that *msgn1* is downstream of *tbx16* [36]. Further studies in *Xenopus* and mouse showed that *msgn1*'s expression in the PSM is controlled by *tbx6* and Wnt signaling [37–41].

Embryos lacking both *tbx16* and *msgn1* function display a similar, but more severe phenotype to embryos lacking *tbx16* or *msgn1*, with a mass of mesodermal progenitors accumulating in the posterior tailbud and lack of somitic mesoderm [18,19]. These studies suggest that *tbx16* and *msgn1* repress the mesodermal progenitor state, but it remains unclear how they regulate mesoderm morphogenesis. *Tbx16;msgn1* double mutants have significantly more mesodermal progenitor cells, posterior spinal cord progenitor cells, and fewer muscle cells compared to wild type [42]. Mesodermal progenitors lacking *tbx16* and *msgn1* fail to produce functional lamellipodia to establish directional migration despite being highly mobile [43]. In mice, *tbx6;msgn1* and double mutants exhibit a similar genetic redundancy in regulating mesoderm development in the tailbud [44]. In zebrafish, *tbx6l* can partially compensate for loss of *tbx16* function [45].

It is still unclear how *tbx16* and *msgn1* regulate the morphogenesis and differentiation of the paraxial mesoderm. Indeed, it is challenging for gene expression analysis to explain phenotypes given the large number of genes and their non-linear interactions and the extensive variability observed both across expression states and among individual cells. This complexity is further compounded by the uneven functional importance of genes with similar expression patterns. Here, we propose a data-driven algorithm that applies the Shapley value [46], a well-established game theory metric, to the analysis of gene expression data. This algorithm not only identifies key genes involved in cell state transitions but also quantitatively estimates the correlation between overall gene expression levels across cells and the progression of cell differentiation. This work provides a systematic method to elucidate the relationship between gene expression patterns and developmental transitions. Moreover, the generalizability of this data-driven framework allows for its adaptation to other types of gene expression data and diverse biological processes.

In this study, we employ machine learning and game theory to analyze the gene network that governs zebrafish body elongation and paraxial mesoderm development. We perform temporally controlled overexpression of *tbx16* and *msgn1* during zebrafish body elongation and examine downstream gene expression using RNA sequencing on pooled dissected tailbuds. We analyzed the bulk RNA sequencing results by combining them with scRNAseq of the wild-type cell state transitions. This analysis reveals that *tbx16* and *msgn1* promote mesodermal differentiation primarily by inhibition of the mesodermal progenitor gene regulatory network. While *tbx16* and *msgn1* regulate many of the same genes, *tbx16* uniquely regulates more genes than *msgn1*. This study contributes new insights not only into vertebrate body elongation and mesodermal development but also into the methodology of gene expression data analysis.

Results

Transgene overexpression of *tbx16* or *msgn1*

We generated transgenic zebrafish lines using the *tol2* transposase system to overexpress either *tbx16* or *msgn1* under the *hsp70l* heat-shock promoter (Fig 1A) [47,48]. We heat-shocked transgenic embryos at the 2–3 somite stage at 38° C

for 30 minutes. There are significant phenotypic differences between heat shocked transgenic and wild-type embryos 24 hours after heat shock (Fig 1C and 1D). Compared to wild-type embryos (Fig 1B), *tg(hsp70l:tbx16)* embryos have a severely truncated tail, which is expected since *tbx16* represses *tbxta* expression [9,16,18,19,28,49]. *tbx16* overexpression also disrupts eye and head development (Fig 1C). Overall, *tg(hsp70l:msgn1)* embryos have a less severe phenotype. In these embryos, anterior morphology is similar to that of wild-type embryos, but the posterior body is truncated and lacks a notochord (Fig 1D).

We next sought to define the timing of gene expression changes that underlie these morphological phenotypes. First, we determined the earliest timepoint after heat shock at which significant alterations in gene expression can be detected. We performed RT-qPCR experiments on *tg(hsp70l:tbx16)* embryos 2, 3, 4, 5 hours after heat shock. We used the rabbit *beta globin* sequence in our transgene (Fig 1A) to measure heat shock induction of transgene mRNA expression. *her1*, *her7*, *tbx6* mRNA were assayed as markers of paraxial mesoderm differentiation and *beta-actin* was used as a control [50]. We used intron qPCR primers for *her1* and *her7* to measure nascent transcription and exon primers for *tbx6* and *beta-actin*. *tg(hsp70l:tbx16)* transgene expression is strongly induced 3 hours after heat shock and peaks 4 hours after heat shock (Fig 1H). *her1* and *her7* transcription increases most at 3 hours after heat shock, but *tbx6* expression did not exhibit a significant change in mRNA levels compared to wild type (Fig 1I). We compared the relative expression of the transgenes 3 hours after heat shock and found that the *tg(hsp70l:msgn1)* displayed higher fold induction relative to the *tg(hsp70l:tbx16)* (Fig 1J). Based on these results, we performed subsequent experiments 3 hours after heat shock to identify the initial burst of gene expression changes following transgene induction.

We examined spatial changes in gene expression using fluorescent in situ hybridization of *tbx6* and *tbxta* on wild-type, *tg(hsp70l:tbx16)* and *tg(hsp70l:msgn1)* embryos 3 hours after heat shock (Fig 1E–1G). Compared to wild-type embryos, *tg(hsp70l:tbx16)* embryos extend the expression of *tbx6* anteriorly, and *tbxta* expression is more variegated in the posterior tailbud. In *tg(hsp70l:msgn1)* embryos, *tbx6* expression is slightly extended to the anterior, and *tbxta* expression in the posterior tailbud is attenuated compared to wild type. This effect of *msgn1* over-expression is weaker than the phenotype produced by a prior *tg(hsp70l:msgn1)* transgene which induced a larger anterior expansion of *tbx6* expression [18]. Overall, the *tg(hsp70l:tbx16)* transgene has a greater effect on both body elongation and tailbud gene expression than *tg(hsp70l:msgn1)*.

RNAseq of transgenic tailbuds

Our aim was to reveal the downstream effectors of *tbx16* and *msgn1* that mediate mesodermal progenitor differentiation. The next step in this process was to identify genes that rapidly respond to over-expression of *tbx16* and *msgn1*. We used bulk RNA sequencing to maximize read depth. We performed RNA sequencing of pooled dissected tailbuds of wild type, *tg(hsp70l:tbx16)* and *tg(hsp70l:msgn1)* embryos 3 hours after heat shock. We performed three biological replicates of each genotype. To identify significant differentially expressed genes, we used DESeq2 with an adjusted p-value cutoff of 0.05 and a 1.5-fold change in expression. We identified 6976 differentially expressed genes (DEGs) in *tg(hsp70l:tbx16)* embryos and 706 in *tg(hsp70l:msgn1)* embryos compared to wild type (Fig 1K–1L and S1 Fig and S1 and S2 Tables). The larger number of DEGs in *tg(hsp70l:tbx16)* embryos is consistent with its more broad and severe phenotype compared to *tg(hsp70l:msgn1)* embryos. As expected, *tbxta* is downregulated by 3.31 fold in *tg(hsp70l:tbx16)* embryos which corresponds to the severely truncated tail phenotype. In *tg(hsp70l:msgn1)* embryos, *noto* expression is downregulated by 1.73 fold which accounts for the lack of posterior notochord [51].

Gene expression analysis via DESeq2

We proceeded to analyze the correlation between the genes misregulated by *tbx16* or *msgn1* overexpression with the genes that change expression levels during paraxial mesoderm development. To that end, we utilized the scRNAseq expression data and paraxial mesoderm cell state classification from [14]. In that study, UMAP was used to arrange the

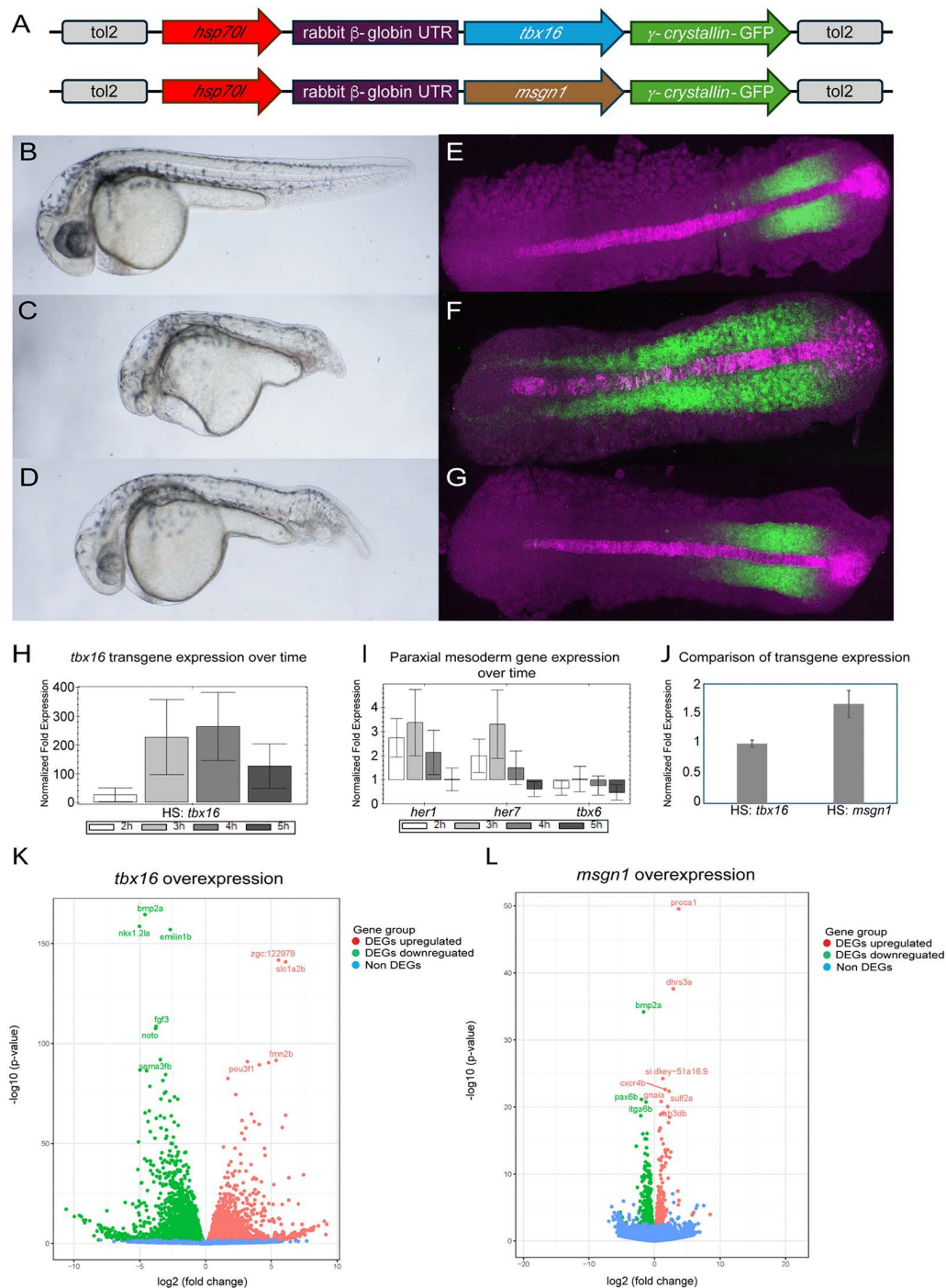


Fig 1. Heat-shock over-expression of *tbx16* or *msgn1* alters gene expression within 3 hours. (A) *hsp70l*-driven *tbx16* and *msgn1* transgenic constructs. *Tol2* sites at both ends of the construct allow for *tol2* transposase-mediated insertion into the genome. The gamma-crystallin promoter drives GFP expression in the lens to be used for screening. The rabbit beta globin sequence is used as a unique sequence to measure transgene mRNA. B-D. Embryos 1 day after 2–3 somite stage heat shock. (B) a wild-type embryo. (C) a *tg(hsp70l:tbx16)* embryo. (D) a *tg(hsp70l:msgn1)* embryo. (E–G) Fluorescent in situ hybridization of *tbx6* (green) and *tbxta* (magenta) in wild-type (E) *tg(hsp70l:tbx16)*, and *tg(hsp70l:msgn1)* embryos. (H) RT-qPCR analysis for *beta globin* sequence of the 5'UTR of transgenic constructs at different times after heat shock. (I) RT-qPCR analysis *her1*, *her7* and *tbx6* from wild-type, *tg(hsp70l:tbx16)*, and *tg(hsp70l:msgn1)* embryos. *Beta-actin* is used as a reference for to normalize for total amount of mRNA in each isolate.

Analysis was performed on *hsp70l:tbx16* embryos 2, 3, 4, 5 hours post heat shock. Fold change is relative to expression levels of WT embryos at two hours post heat shock. (J) Tg(*hsp70l:mshg1*) exhibits higher expression than tg(*hsp70l:tbx16*) 3 hours after heat shock. (K-L) Volcano plot of RNA-seq experiments to embryo dissected tailbuds of *tbx16* overexpression vs wild type (K) and *mshg1* overexpression vs wild type (L). The upregulated, down-regulated, and non DEGs are color coded as red, green and blue dots, respectively.

<https://doi.org/10.1371/journal.pgen.1012176.g001>

cells in a one-dimensional pseudotime sequence, and then the pseudotime was segmented into distinct transcriptional cell states using a variance minimization algorithm. That protocol identified six different cell states: neuronal, neuromesodermal progenitor (NMP), mesodermal progenitor (MP), progenitor zone (PZ), posterior (pPSM) and anterior PSM (aPSM), (Fig 2A and 2B). To identify genes differentially expressed during this developmental progression, we pooled gene expression in individual cells by cell state and experimental replicate, i.e., pseudobulking, and performed DESeq2 comparisons between each adjacent pair of pseudotime segments. The number of differentially expressed genes ranged from 243 at the PZ to pPSM transition to 467 at the NMP to MP transition (Fig 2C and S3 Table). This gene set includes known neuronal, NMP and mesodermal genes such as *sox2*, *sox3*, *tbxta*, *tbx16*, *mshg1*, *tbx6* and *mespaa*. *Tbx16* and *mshg1* overexpression alter expression of different subsets of genes that change at each transition.

We generated scatterplots of these data for each cell state transition (S2 and S3 Figs and S4 and S5 Tables). These plots show how expression of each gene changes (up or down) in the scRNAseq pseudotime along the x-axis. The log-fold change in gene expression after heat shock of tg(*hsp70l:tbx16*) or tg(*hsp70l:mshg1*) is plotted along the y-axis. The preponderance of gene expression is repressed by *tbx16* overexpression (S2 Fig). We reasoned that since *tbx16* is required for paraxial mesoderm differentiation, true *tbx16* target genes would either be upregulated at a particular cell state transition in wild type (WT⁺) and also upregulated after *tbx16* overexpression (tbx16⁺) or be downregulated at a particular wild-type cell state transition (WT⁻) and also downregulated by *tbx16* overexpression (tbx16⁻). These genes are in the top right and bottom left quadrants of each scatterplot, respectively. These plots show a bias in WT⁺tbx16⁻ genes, particularly at the NMP to MP, PZ to pPSM and pPSM to aPSM transitions (S2 Fig). The *mshg1* analysis also shows enrichment of WT⁺mshg1⁻ genes particularly at the PZ to pPSM transition whereas the MP to PZ transition is enriched for both WT⁺mshg1⁻ and WT⁺mshg1⁺ genes (S3 Fig).

Gene expression analysis using machine learning

The pseudotime analysis incorporates 21301 genes to calculate the pseudotime coordinates with breakpoints for gene expression state classification [14]. However, the high dimensionality of the data and the structure of this classification complicate quantification of how DESeq2 pseudobulked gene expression changes contribute to transitions between consecutive cell states (e.g., NMP to MP). To uncover the intricate relation between developmental gene expression state and individual genes, we utilized a Random Forest Classifier with Shapley Additive exPlanations (SHAP) framework customized for decision-tree models (Fig 2D) [52,53]. The Random Forest Classifier was used to reconstruct the gene expression states and provide probabilities missing from the original pseudotime states. SHAP used these probabilities to quantify the importance of each gene in defining each transcriptional state. The underlying hypothesis is that genes that are important for defining transcriptional state transitions during differentiation will be functionally important. If this hypothesis is correct, then these gene lists should be enriched for genes with known functions in paraxial mesoderm development.

The first three steps were the creation of the pseudotime-classified dataset (Fig 2D) [14]. The Random Forest Classifier was then trained on this dataset to identify a minimal subset of genes capable of accurately classifying gene expression states from the full gene set (Fig 2F). We identified a selected number of genes for each cell state with the subset size of these genes overlapping with those up- or downregulated (≥ 1.5 -fold) by *tbx16* or *mshg1* overexpression (Fig 2D). The Random Forest Classifier gene lists are more selective than the DESeq2-generated gene lists (compare Fig 2C to 2E).

The resulting model structure was then used to compute gene-specific Shapley values across individual cells [46]. Shapley values are a game theory-based metric of the contribution of each component of a group to the outcome/performance

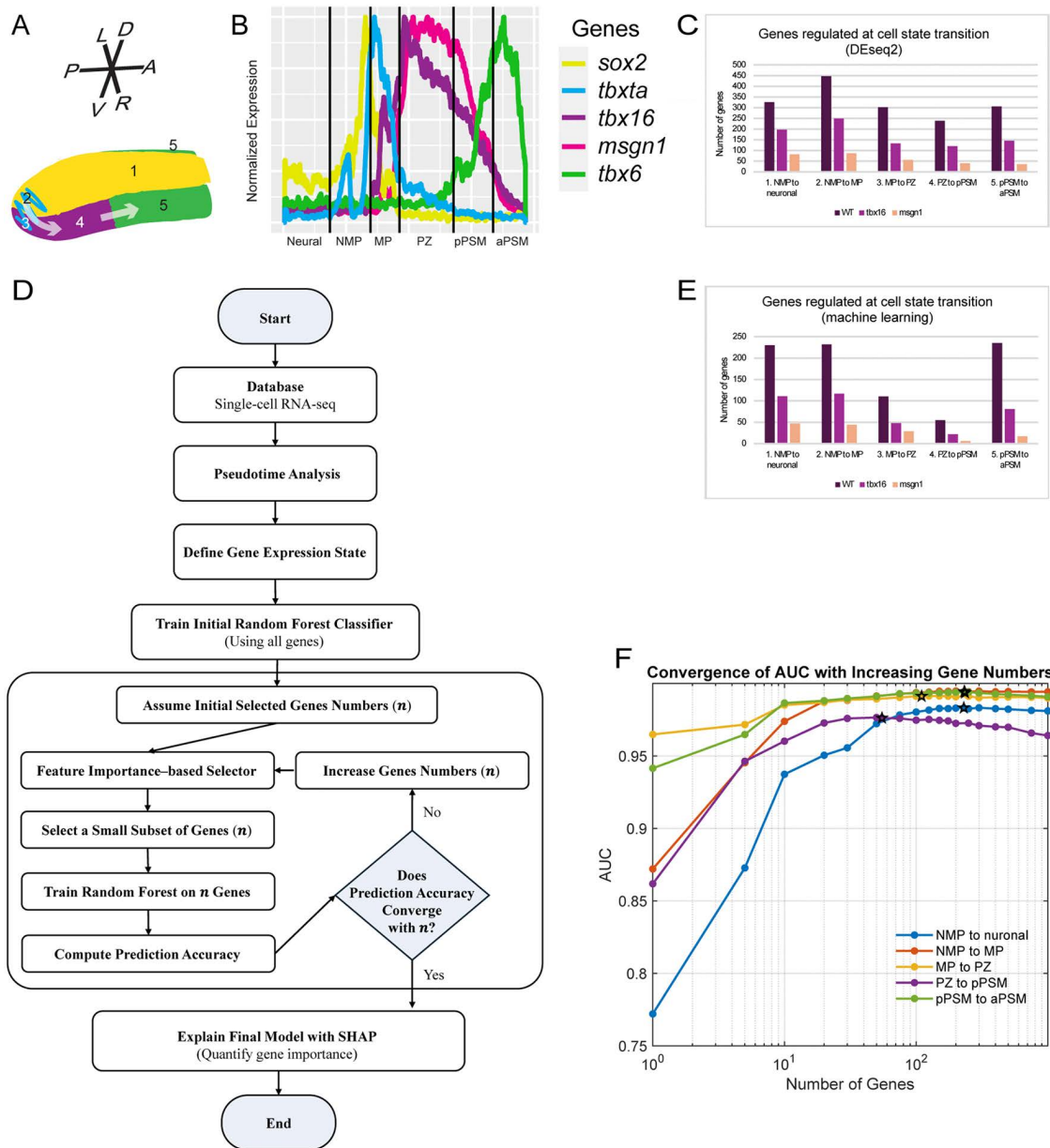


Fig 2. Overview of machine-learning-based selection of cell state transition genes. (A) A schematic showing the developmental trajectory of the paraxial mesoderm in the tailbud. Body axes are shown D (dorsal), V (ventral), A (anterior), P (posterior), R (right) and L (left). The expression patterns of *sox2* (yellow), *brachyury* (cyan), *tbx16* (purple), and *tbx6* (green) are schematized. (B) Smoothed, normalized single cell gene expression profiles of wild type tailbud cells over the pseudotime trajectory, where optimal change points delineate six gene expression states: neuronal, neuromesodermal progenitors (NMP), mesodermal progenitors (MP), progenitors zone (PZ), posterior (pPSM) and anterior PSM (aPSM). NMPs can develop either as neuronal cells (leftward) or as mesodermal progenitors (rightward). Five representative genes from 10269 identified genes are shown. (C, E) Histogram showing the number of genes exhibiting expression changes during each wild-type cell state transition. Also shown is the overlap with genes regulated by *tbx16* or *msgn1* overexpression compared to wild type (≥ 1.5 -fold change). (C) Genes identified by DESeq2 using a 1.5-fold expression change threshold during the cell state transitions. (D) The flowchart of the data driven framework to identify key genes correlated to the gene expression state transitions and to quantify each gene's importance with a Shapley value. (E) Genes identified by the machine learning algorithm. (F) Random forest model training accuracy on the test dataset versus the number of genes selected (x-axis in log scale). The star marker indicates the number of genes selected by the random forest classifier that maximizes prediction accuracy under optimal hyperparameters.

<https://doi.org/10.1371/journal.pgen.1012176.g002>

of the group. The approach has recently been extensively used to interpret machine-learning model predictions to account for intricate interactions between features [53]. Here, the Shapley value represents the expected marginal contribution of a gene's expression level to the predicted probability that a cell belongs to a particular transcriptional state, averaged over all possible combinations of other genes in the database. For each single cell, a positive Shapley value indicates that high or low expression of the gene raises the probability of the cell belonging to a particular cell state, while a negative value indicates that high or low expression of the gene reduces the probability of the cell belonging to the state, given the influence of all other genes in the random forest decision tree model. For example, high expression of *fgf8a* decreases the probability of a cell being neural while low expression increases the probability of a cell being neural (Fig 3B).

Application of SHAP analysis to all 8120 wild-type cells classified in the MP and PZ states with the 235 identified key genes yielded an 8120×235 matrix of Shapley values, representing the contributions of each gene across those wild-type cells to the decision-tree classification. Fig 3 presents summary plots that demonstrate the effect of each gene, ordered according to their importance (only the 10 most essential genes are shown). As expected, the transition between MP to PZ gene expression states is dependent upon *msgn1* in a positive relation (Fig 3D).

Lastly, we utilized the Shapley matrix of single cell gene expression to estimate how the bulk RNAseq gene expression after *tbx16* or *msgn1* overexpression influences the transitions of gene expression states during paraxial mesoderm differentiation. Here, we introduced a log-scale sensitivity parameter (*S*), representing the proportional change in a gene's Shapley value per log₂-fold change in normalized expression (S6 Table). The calculation of *S* using a linear regression method is detailed in the Materials and Methods. In scatterplots in Figs 4 and 5, *S* is plotted along the x-axis. Along the y-axis *S* is multiplied by the log fold change in bulk RNAseq data after *tbx16* or *msgn1* overexpression (Figs 4 and 5 and S7 and S8 Tables). These plots provide a measure of how strongly perturbation-induced shifts in gene expression affect the random forest model-inferred probability of cell state transitions.

This analysis indicates that *tbx16* principally represses genes important for defining each transcriptional state. *Tbx16* represses genes that are normally repressed at the MP to PZ, PZ to pPSM and pPSM to aPSM transitions (upper left quadrants in Fig 4D, 4E and 4F). Moreover, *tbx16* most strongly represses genes that are the most important for defining each cell state. *Tbx16* represses both proneural genes that are normally upregulated at the NMP to neuronal transition and promesodermal genes that are downregulated at this transition (Fig 4B). *Tbx16* also represses genes that are normally upregulated and downregulated at the NMP to MP transition. *Tbx16* is known to repress *sox2*, which promotes NMP differentiation into neuronal precursors [28], but these data suggest that *tbx16* similarly promotes mesodermal progenitor differentiation into PZ and PSM via transcriptional repression.

Msgn1 likewise represses gene expression but it affects far fewer genes (Fig 5). At the MP to PZ transition, *msgn1* almost exclusively represses genes that are normally downregulated at this transition. *Msgn1* also represses both proneural genes that are normally upregulated at the NMP to neuronal transition and promesodermal genes that are normally downregulated at this transition. Similarly, *msgn1* represses both pro-NMP genes at the NMP to MP transitions and represses pro-MP genes (Fig 5B and 5C). These results again suggest that *msgn1* promotes paraxial mesoderm differentiation by repressing genes that are important for defining the MP cell state.

The machine learning gene lists are enriched for important genes

We compared both the DESeq2 and machine learning gene lists for each cell state in wild-type tailbuds to the lists of DEGs in tg(*hsp70l:tbx16*) embryos and tg(*hsp70l:msgn1*) embryos. Since *tbx16* and *msgn1* are required for cells to mature into PSM, we focused on the MP to PZ transition and PZ to pPSM transition. Genes that are downregulated at these transitions in wild-type embryos and also repressed by overexpression of *tbx16* and *msgn1* were the most conspicuous in the scatterplots (Figs 4D, 4E, 5D and 5E), so we specifically list these genes in Table 1.

We compared genes repressed by *tbx16* with *msgn1* in the MP to PZ and PZ to pPSM transitions, and we found both common and uniquely repressed genes (Table 1). Generally, *tbx16* has more unique downstream genes than *msgn1*.

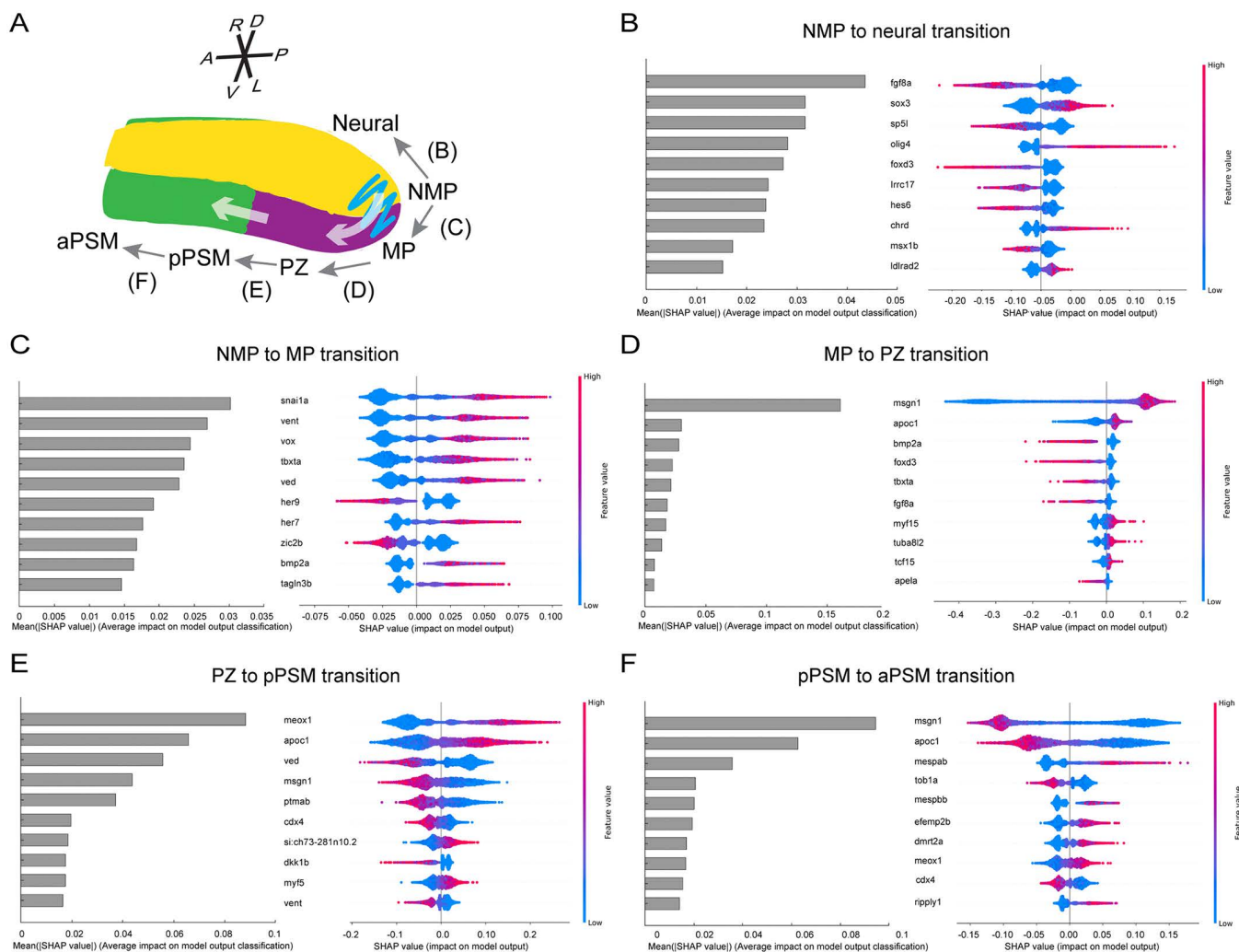


Fig 3. SHAP identifies the most important genes for each cell state transition. (A) A schematic showing the developmental trajectory of the paraxial mesoderm in the tailbud. Body axes are shown as D (dorsal), V (ventral), A (anterior), P (posterior), R (right) and L (left). Cell state transitions are shown as NMP to neural, NMP to MP, MP to PZ, PZ to pPSM, and pPSM to aPSM. Each transition is labeled with an arrow indicating the direction of transition. (B-F) The 10 most essential genes correlating to wild-type cell state transitions from NMP to neural (B), NMP to MP (C), MP to PZ (D), PZ to pPSM (E), and pPSM to aPSM (F). These genes are ranked by their overall impact as denoted by their mean absolute Shapley value (left). The summary plot (right) combines gene importance with their overall effect on the transitions: the color represents the relative expression value of each gene from low (blue) to high (red), while x axis (Shapley value) denotes the increase or decrease in probable impact on the cell state transition.

<https://doi.org/10.1371/journal.pgen.1012176.g003>

74.19% of *msgn1* regulated genes are shared with *tbx16* at the MP to PZ transition, whereas only 38.33% of *tbx16* downstream genes are shared with *msgn1*. At the PZ to pPSM transition, 58.33% of *msgn1* downstream genes are shared with *tbx16*, whereas only 17.28% of *tbx16* regulated genes are shared with *msgn1*. Many of the repressed genes have known roles in body elongation, mesoderm differentiation and somitogenesis [8,9,16,27–29,41,54–70]. During the MP to PZ transition, *apela*, *bmp2a*, *fgf3*, *fgf4*, *hoxa13b*, *hoxd12a*, *lfng*, *tbxta*, *tll1*, and *wnt2bb* signaling are repressed by both *tbx16* and *msgn1*. *Tbx16* uniquely represses *eve1*, *fgf8a*, *sp9*, *wnt11* and *wnt8a*. *Msgn1* uniquely represses *cyp26a1* and *ephb4b*. During the PZ to pPSM transition, *bmp2a*, *fgf24*, *hoxa13b*, *hoxd12a*, *lef1*, and *tbxta* are repressed by both *tbx16* and *msgn1*. *Tbx16* uniquely represses *bambia*, *cdx1a*, *cdx4*, *eve1*, *fgf10a*, *gdf3*, *hes6*, *her12*, *hoxc11a*,

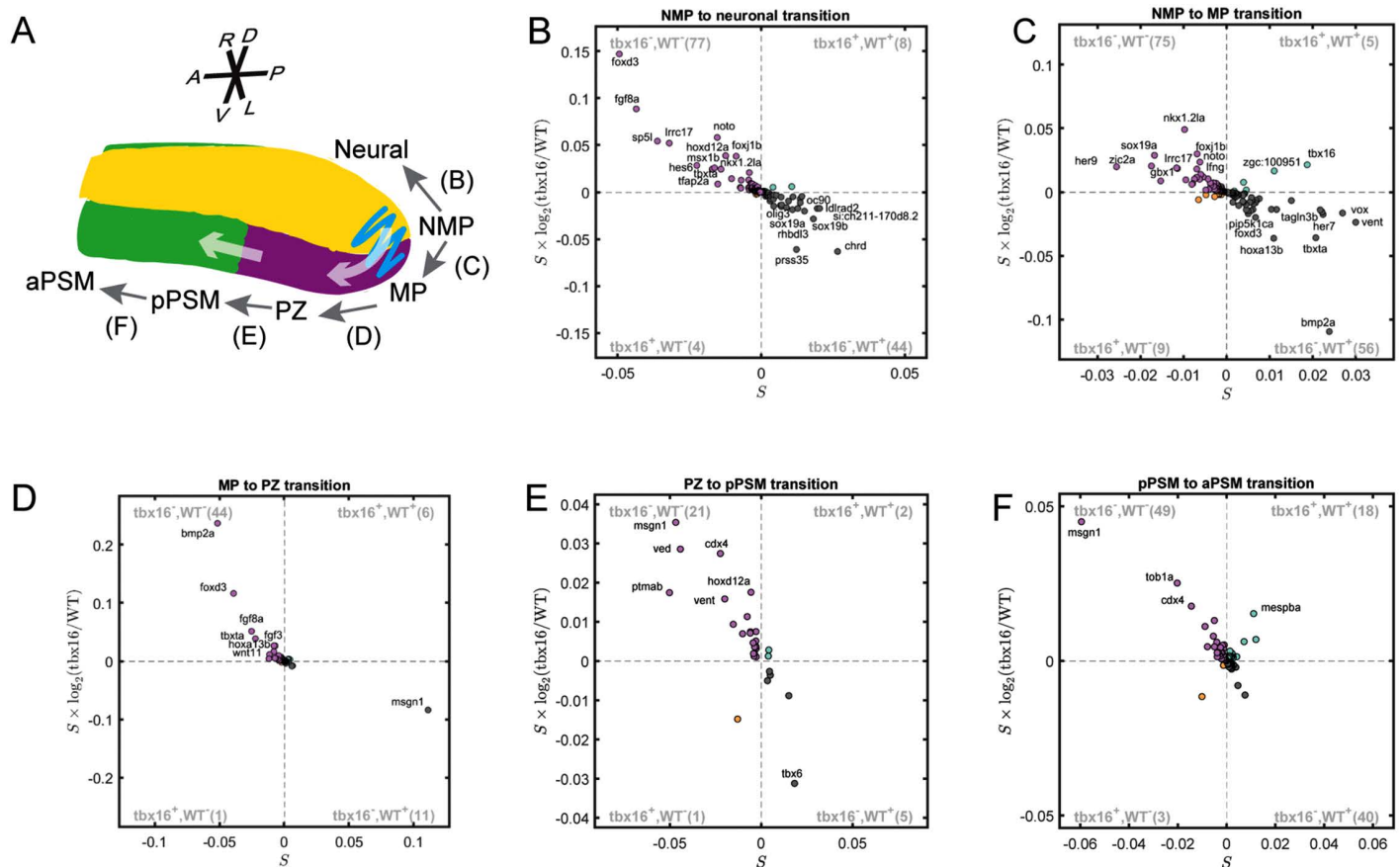


Fig 4. Transcriptional regulation by *tbx16* during cell state transitions of paraxial mesoderm differentiation. (A) A schematic showing the developmental trajectory of the paraxial mesoderm in the tailbud. Body axes are shown as D (dorsal), V (ventral), A (anterior), P (posterior), R (right) and L (left). Cell state transitions are shown as NMP to neural, NMP to MP, MP to PZ, PZ to pPSM, and pPSM to aPSM. Each transition is labeled with an arrow indicating the direction of transition. (B–F) The change in probability in decision tree model-predicted state transition under *tbx16* overexpression is estimated by a linear function ($S \times \log_2(\text{tbx16}/\text{WT})$) of each gene's log-scale sensitivity parameter (S). Key genes identified by machine learning with positive or negative correlations to wild-type gene expression states transitions (WT^+ , WT^-) and overlapping with genes up- or downregulated (≥ 1.5 -fold; tbx16^+ , tbx16^-) under *tbx16* overexpression compared to wild type are analyzed during NMP to neural (B), NMP to MP (C), MP to PZ (D), PZ to pPSM (E), and pPSM to aPSM transitions (F).

<https://doi.org/10.1371/journal.pgen.1012176.g004>

hoxc11b, *hoxd3a*, *msgn1*, *sp5l*, *szl*, *tbx16l*, *wnt11*, *ved*, *vent*, *vimr1* and *vox*. *Msgn1* uniquely represses *bmp4*, *cyp261a* and *sp5a*. This analysis indicates that *tbx16* and *msgn1* repress transcription of genes that promote the mesodermal progenitor cell state (Fig 6).

Discussion

The Random Forest Classifier and SHAPLEY aid gene expression analysis

Here, we combined analysis of RNAseq and scRNAseq data to take advantage of the read depth of the former and the single cell resolution of the latter. However, a challenge with genomics experiments is extracting useful information from the resulting gene lists. We previously utilized UMAP to create a one-dimensional pseudotime for differentiation of paraxial mesoderm in the zebrafish tailbud. We then segmented the pseudotime into discrete transcriptional cell states using a variance minimization algorithm. Changes in these cell states after perturbation of BMP, Wnt or Fgf signaling can be

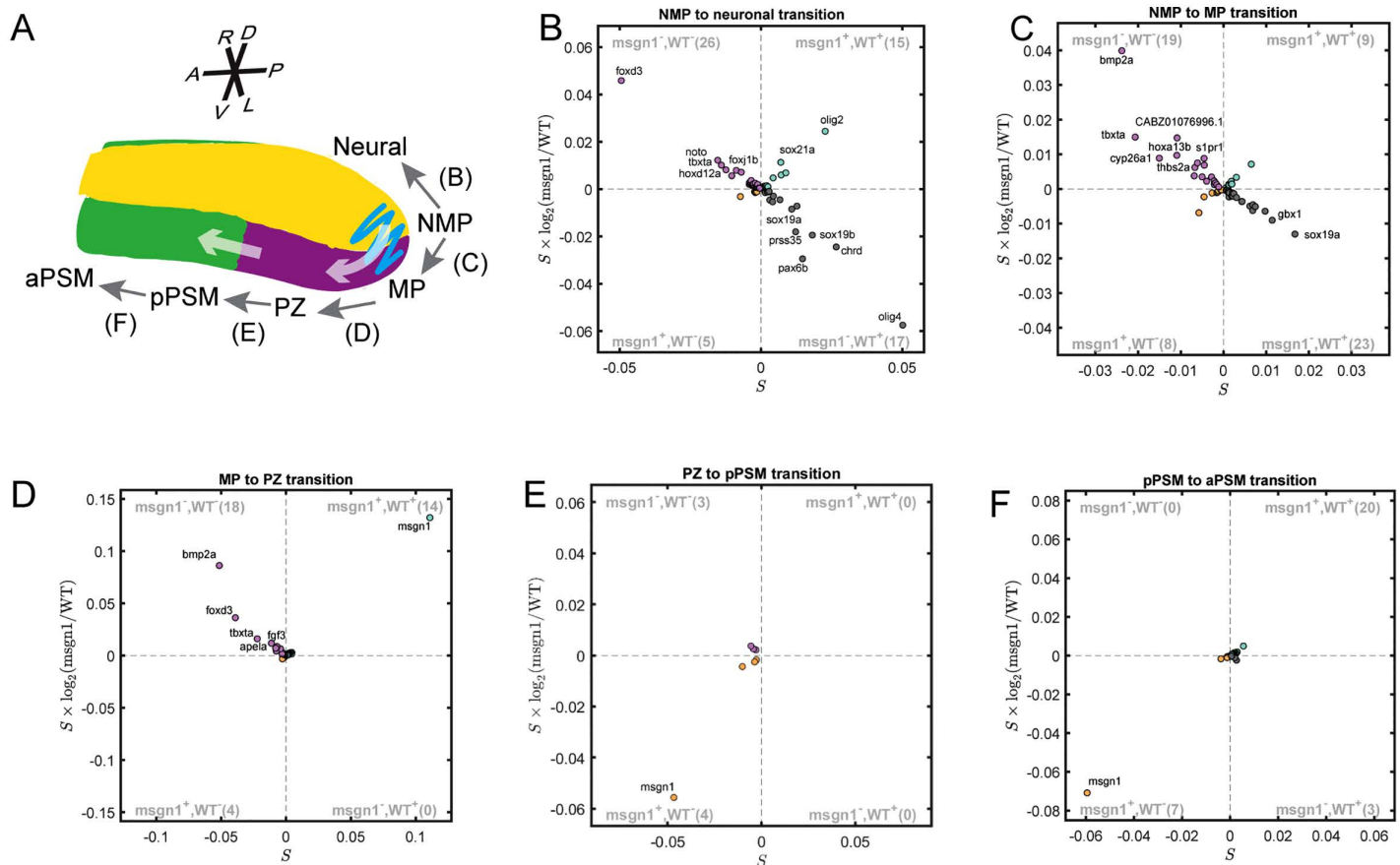


Fig 5. Transcriptional regulation by *msgn1* during cell state transitions of paraxial mesoderm differentiation. (A) A schematic showing the developmental trajectory of the paraxial mesoderm in the tailbud. Body axes are shown as D (dorsal), V (ventral), A (anterior), P (posterior), R (right) and L (left). Cell state transitions are shown as NMP to neural, NMP to MP, MP to PZ, PZ to pPSM, and pPSM to aPSM. Each transition is labeled with an arrow indicating the direction of transition. (B–F) The change in probability in decision tree model-predicted state transitions under *msgn1* overexpression, estimated by a linear function ($S \times \log_2(\text{msgn1}/\text{WT})$) of each gene's log-scale sensitivity parameter (S). Key genes identified by machine learning with positive or negative correlations to wild-type gene expression states transitions (WT^+ , WT^-) and overlapping with genes up- or downregulated (≥ 1.5 -fold; msgn1^+ , msgn1^-) under *msgn1* overexpression compared to wild type are analyzed during NMP to neural (B), NMP to MP (C), MP to PZ (D), PZ to pPSM (E), and pPSM to aPSM transition (F).

<https://doi.org/10.1371/journal.pgen.1012176.g005>

quantitatively mapped onto the tailbud using multicolor fluorescent in situ hybridization [14]. Here, we used this pseudo-time to help analyze the initial burst of gene expression after temporally controlled overexpression of *tbx16* and *msgn1*.

First, we use machine learning and DESeq2 in parallel to identify genes with significant expression level changes during cell state transitions in wild-type paraxial mesoderm differentiation. The two methods cross-validate each other because there is substantial number of genes that are identified by both methods (Table 1). Both methods support the conclusion that *tbx16* and *msgn1* promote mesodermal differentiation by inhibiting the maintenance of the mesodermal progenitor niche, though this conclusion is most evident in the machine learning analysis. Comparing the two methods with each other, we found that the machine learning method produced a shorter gene list. A premise of the machine learning analysis is that genes that display clear transitions in expression levels during paraxial mesoderm differentiation are particularly important for differentiation. This premise is validated by the observation that the machine learning approach is more enriched for genes with known functions in body elongation and mesoderm differentiation (Table 1). The DESeq2

Table 1. Commonly and uniquely repressed genes by *tbx16* and *msgn1* overexpression at the MP to PZ and PZ to pPSM transition. All genes have been filtered by a 1.5-fold change threshold of expression level change. Genes or pathways required for body elongation and/or mesoderm differentiation (bold) and somitogenesis (underlined) are indicated.

Genes repressed during the MP to PZ transition					
Genes repressed by both <i>tbx16</i> and <i>msgn1</i>			Genes repressed by <i>tbx16</i> only		
Machine learning and DESeq2	Machine learning only	DESeq2 only	Machine learning and DESeq2	Machine learning only	DESeq2 only
adam8a apela bmp2a fgf3 foxd3 hoxa13b rasl11b tbxta tlil	hoxd12a si:dkey-18j18.3 fgf4 hhl2a.1 hpcal4 itgb4 lfrg s1pr1 si:dkey-40c11.2 usp54a wnt2bb zgc:110158	cdh6 esrrga fgf4 hhl2a.1 hpcal4 itgb4 lfrg s1pr1 si:dkey-40c11.2 usp54a wnt2bb zgc:110158	ankrd50l eve1 fgf8a msx1b sox2 sp9 tuba1c wnt11	enc3 hmrpabb npm1a pcna rrm2 rtca tagln3b thbs2a timmm50 tuba1c ube2e2 zgc:110158	b3gnt7l calm1b foxb1a griffin lrrc17 lypd6b mcm3 notch3 ofml2bb pax3a prrr18 prx1b
Genes repressed during the PZ to pPSM transition					
Genes repressed by both <i>tbx16</i> and <i>msgn1</i>			Genes repressed by <i>tbx16</i> only		
Machine learning and DESeq2	Machine learning only	DESeq2 only	Machine learning and DESeq2	Machine learning only	DESeq2 only
hoxd12a tbxta	inka1b	bmp2a fgf24 foxd3 gdf11 hoxa13b klf26ab pip5k1ca si:dkey-18j18.3 sox11a thbs2a tlil	bambia cdx4 hes6 itm2cb msgn1 ptgdsb.1 rbm38 sst6 tbx16l ved vent vox	cx43.4s arhgap35b b3gnt7l blf ca8 cdx1a clip2 clstn1 col11a1a dhrs3b enc1 enc3 eve1 fgf10	adams18 arhgap35b b3gnt7l blf ca8 cdx1a clip2 clstn1 col11a1a dhrs3b enc1 enc3 eve1 fgf10
Genes repressed by <i>msgn1</i> only					
Machine learning and DESeq2	Machine learning only	DESeq2 only	Machine learning and DESeq2	Machine learning only	DESeq2 only
hoxd12a tbxta	inka1b	bmp2a fgf24 foxd3 gdf11 hoxa13b klf26ab pip5k1ca si:dkey-18j18.3 sox11a thbs2a tlil	cyp26a1	cyp26a1 thbs2a	CABZ0176996.1 ephb4b marveld1 pou5f3 zgc:92140

Bold: Genes or pathways required for body elongation. Underlined: Genes or pathways required for somitogenesis

<https://doi.org/10.1371/journal.pgen.1012176.t001>

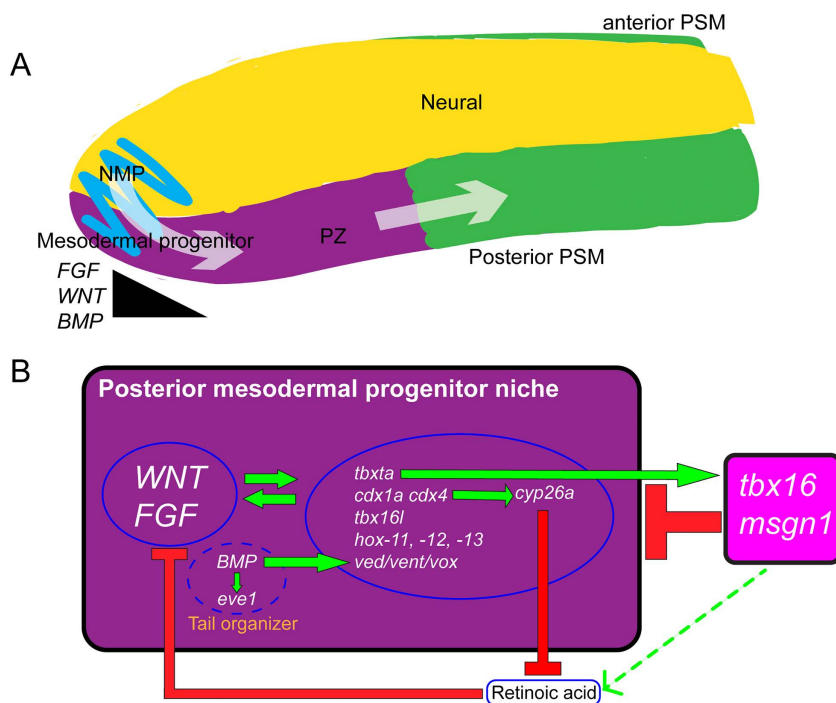


Fig 6. *Tbx16* and *msgn1* promote mesodermal differentiation by inhibiting the maintenance of the mesodermal progenitor cell state. (A) A schematic of cell state transitions in the zebrafish tailbud, from NMP to mesodermal progenitor to posterior PSM to anterior PSM. Gradients of FGF, Wnt and BMP signaling in the posterior tailbud promote body elongation, mesodermal fates and maintain the progenitor state. (B) The paraxial mesoderm gene regulatory network in which *tbx16* and *msgn1* repress expression of genes required to maintain the progenitor niche. This repression drives cell state transitions from mesodermal progenitors to PZ and PZ to posterior PSM. During body elongation, cells that are displaced away from the posterior tailbud escape the FGF, Wnt and BMP gradients enabling the negative feedback by *tbx16* and *msgn1* to predominate and commit the cells to differentiation.

<https://doi.org/10.1371/journal.pgen.1012176.g006>

pseudobulking method produced a larger gene list and included additional genes or pathways known to be required for body elongation and mesoderm development. However, the frequency of these known genes is lower than in the machine learning list suggesting that the DESeq list is noisier. Conversely, important genes with noisy expression or low expression that are poorly sampled by scRNAseq may be missed by the machine learning approach. Thus, the parallel application of these two methods is a powerful approach in gene expression analysis to identify possible target genes. The machine learning method is more selective but perhaps has more false negatives, and the DESeq2 method is likely noisier but more inclusive.

***tbx16* and *msgn1* repress the mesoderm progenitor cell state**

Previous studies indicate that *tbx16* and *msgn1* have both independent and redundant functions in paraxial mesoderm development. Loss of either *tbx16* or *msgn1* function leads to a failure of PSM differentiation with loss of *tbx16* having a much stronger phenotype [15–19]. Phenotypic differences are similarly evident after overexpression of *tbx16* or *msgn1*: *tbx16* overexpression has more severe elongation defect than *msgn1* overexpression (Fig 1C–1D) [9,18,28,49]. However, the heat-shock-inducible *msgn1* transgene in this paper produces weaker phenotypes than transgenes in these prior studies.

Several prior analyses utilized heat-shock-inducible *tbx16* and *msgn1* zebrafish transgenes to study mesodermal differentiation and gene regulatory networks [9,28,49]. In these studies, transgenic embryos were heat-shocked at 10–12

somite stage, and they exhibited less severe phenotypes than the embryos heat-shocked at 2–3 somite stage in our study. We also observe that later-stage heat-shock produces qualitatively similar but quantitatively weaker phenotypes. We chose to focus on early-stage heat shock because the tailbud dissections would isolate more tissue for RNA sequencing. These prior studies showed that Wnt and Fgf signaling induce NMPs to undergo an epithelial to mesenchymal transition and activate *tbx16* and *msgn1* expression downstream of *tbxta* as cells commit to a mesodermal fate. Meanwhile, *tbx16* and *msgn1* turn off the NMP state by repressing *sox2* and *tbxta*, making the transition to mesodermal fate irreversible. Our data show that *tbx16* and *msgn1* also repress multiple Wnt and Fgf pathway genes at the later MP to PZ and PZ to posterior PSM transitions. Therefore, *tbx16* and *msgn1* function with Wnt and Fgf signaling in a negative feedback loop that broadly regulates paraxial mesoderm differentiation. A similar regulatory network is involved in mouse NMP differentiation to mesoderm and includes *msgn1* but not *tbx16* [44,59].

Our analysis suggests that *tbx16* and *msgn1* promote mesodermal differentiation by inhibiting the maintenance of the mesodermal progenitor cell state which is consistent with prior analysis that examined a smaller number of genes (Fig 6) [18]. Wnt and Fgf signaling are key pathways required for the maintenance of mesodermal progenitor niche keeping them in an unsegmented state [71–76]. There is extensive crosstalk between the Wnt and Fgf signaling in the zebrafish tailbud [50,77]. Wnt and Fgf signaling also induce the mesodermal fates by activating t-box genes (*tbx16*, *tbxta*, and *tbx6*), *cdx* genes, and *ved/vent/vox* genes [8,9,16,27,28,41,58,59]. *Cdx* genes further activate *cyp26a* to repress retinoic acid (RA) signaling, which represses Wnt and Fgf pathway signaling in a negative feedback loop [29,59,60]. Posterior *hox* genes (*hox-11*, *-12*, *-13*) are also expressed in mesodermal progenitor cells and required for mesoderm formation and posterior body elongation [29,60]. In addition, the BMP signaling pathway also activates *ved/vent/vox* genes. It is the key pathway in the tail organizer, which regulates posterior body elongation, and cooperates with gradients of Wnt and Fgf signaling to control gene expression and cell motion [14,55,58,78–82]. All these pathways and related genes that maintain the mesodermal progenitor niche are repressed by *tbx16* and *msgn1* (Table 1 and Fig 6). This repression would drive irreversible cell fate transition from mesodermal progenitor cells to presomitic mesoderm. *Tbx16* and *msgn1* repress expression of many of the same genes, and *tbx16* represses more unique genes than *msgn1*. Thus, our data reveals both redundancy and independence of *tbx16* and *msgn1* of gene regulation in mesodermal differentiation (Table 1).

Tbx16 overexpression extends the *tbx6* expression domain anteriorly compared to wild type embryos but leads to a 3-fold decrease in overall *tbx6* mRNA level (Fig 1F and S1 Table). An anterior extension of *tbx6* expression was also observed in *rippy1*; *rippy2* double mutant zebrafish embryos [83]. However, *tbx16* overexpression increases *rippy2* expression at 1.82-fold change compared to wild type, with no significant expression level change to *rippy1* (S1 Table). In mouse, *rippy* genes are activated by *mesp2*, and *mesp* genes interact with *notch* via negative feedback to establish somite boundaries [83,84]. In zebrafish, however, *rippy* and *mesp* genes act independently to regulate somite boundary formation [83]. *Tbx16* overexpression decreases the expression of *mespaa* (*mesp1*), *notch1b* and *notch 3* at 3–4-fold change but increases *mespba* expression at more than 2-fold change compared to wild type. Thus, it is currently unclear how *tbx16* overexpression directs the anterior extension *tbx6* expression domain.

Differentiation by transcriptional repression

For both pluripotent stem cells and progenitor cells, there are several examples of differentiation by repressing genes that maintain a progenitor state. Embryonic stem cell differentiation to neural progenitor cells needs *polycomb*-mediated repressive epigenetic modification H3K27Me3 to silence *Jarid2*-sensitive genes [85]. Cortical neural stem cells require *sox10* to differentiate into oligodendrocyte precursors because *sox10* represses stem-cell programming factors such as *sox2* and *sox9* [86]. *Sox11*, expressed by oligodendrocytes progenitor cells, is epigenetically inhibited by Oligodendrocyte Transcription Factor 2 when the oligodendrocytes progenitor cells differentiate to immature oligodendrocytes [87]. Retinal progenitor cells require *Jarid2* to repress the early retinal cell gene *foxp1* to differentiate into late retinal progenitor cells [88]. Differentiation of epidermal progenitor cells necessitates a decrease in the expression of *SNAI2*, which represses

differentiation and cell adhesion genes. If *SNAI2* expression is maintained, the epidermal progenitor cells will maintain the progenitor state [89]. Compared to these examples, *tbx16* and *msgn1* repress a larger number of genes within the mesodermal progenitor niche gene regulatory network. Therefore, the role of repression is broader, regulating the general process of posterior body elongation and formation of the trunk and tail mesoderm.

Conclusion

In this study, we employ machine learning and game theory to analyze the gene network that governs zebrafish body elongation and paraxial mesoderm development. This analysis reveals that *tbx16* and *msgn1* promote mesodermal differentiation primarily by inhibition of the mesodermal progenitor gene regulatory network. This study contributes new insights not only into vertebrate body elongation and mesodermal development but also into the methodology of gene expression data analysis.

Materials and methods

Transgenic fish line preparation and heat-shock operation

Ethics Statement. Tüpfel long fin zebrafish (wild-type fish) were raised according to standard protocols and experiments approved by the Institutional Animal Care and Use Committee. Transgenic lines were created using Tol2 transposase system [48]. For the *tg(pT2_Tol2_hsp70_tbx16_OPT_yCry_GFP_Tol2)* and *tg(pT2_Tol2_hsp70_msgn1_OPT_yCry_GFP_Tol2)* constructs (abbreviated as *tg(hsp70l:tbx16)* and *tg(hsp70l:msgn1)*, the full length coding sequence of *tbx16* and *msgn1* was cloned with primers listed in S9 Table. Transgenic embryos were heat-shocked at 38°C for 30 minutes at the 2–3 somite-stage and raised for 24 hours post heat shock at 28.6°C to observe phenotypes. After robustness and consistency of phenotypes were confirmed in F1 transgenics, transgenic fish were in-crossed to produce homozygotes. F2 transgenics were out-crossed with wild type to screen by heat-shock for homozygotes and heterozygotes. For experiments, we outcrossed to wild type to either homozygotes for *tg(hsp70l:tbx16)* or heterozygotes for *tg(hsp70l:msgn1)*.

Quantitative PCR

qPCR experiments were performed for homozygous *tg(hsp70l:tbx16)* embryos 2, 3, 4, 5 hours post heat shock and wild-type embryos 2 hours post heat shock. Rabbit *beta globin* sequence in our transgene was used to measure heat shock induction of transgene expression. *Her1*, *her7*, *tbx6* were assayed as markers of paraxial mesoderm development and *beta-actin* was used as a control [50]. Primers are listed in S9 Table. 6 biological replicates, each with 3 technical replicates, were performed for each experimental condition. For each biological replicate, 10 embryos were pooled to extract RNA by QIAGEN RNeasy Plus Micro kit. To compare relative expression of the *tg(hsp70l:tbx16)* and *tg(hsp70l:msgn1)* transgenes 3 hours after heat shock, transgenic heterozygotes were crossed to TLF, heat shocked at the 2–4 somites stage and 40 embryos were pooled for RNA extraction for each of three biological replicates. A second set of β -globin primers were used for qPCR.

In Situ hybridization

Probes, embryo processing and imaging parameters are described in [14]. Images were taken by the Leica Stellaris Falcon STED confocal microscope using a 20x objective.

Tailbud dissection and RNA sequencing

Heat-shocked embryos were incubated for 3 hours after heat shock at 28.6°C and then dissected in ice-cold Hank's balanced salt solution (HBSS). The tailbud was collected by cutting immediately posterior to the last formed somite.

10 tailbuds were collected per sample and 3 samples were prepared for per genetic background (*tg(hsp70l:tbx16)*, *tg(hsp70l:mshn1)* and wild type). Since we used heterozygous *tg(hsp70l:mshn1)* fish, the anterior body of each dissected embryo was PCR genotyped using the rabbit *beta globin* UTR primers (S9 Table). Transgenic tailbuds were then pooled before RNA purification while non-transgenics tailbuds were pooled for wild-type heat shock controls. RNA was extracted for each sample with Trizol. Samples were sent to Yale Center for Genomic Analysis (YCGA) for RNA sequencing and data analysis. Low quality reads were trimmed, and adaptor contamination was removed using Trim Galore (v0.5.0). Trimmed reads were mapped to the zebrafish reference genome (GRCz11) using HISAT2 (v2.1.0) [90]. Gene expression levels were quantified using StringTie (v1.3.3b) [91]. Differentially expressed genes (DEGs) were identified using DESeq2 with a 0.05 false discovery rate (v 1.22.1) [92]. This list was additionally filtered for genes with at least a 1.5 fold change in expression.

DESeq2 analysis of expression changes

To identify genes differentially expressed over developmental pseudotime, we analyzed the scRNAseq data from [14]. Using Seurat v5 we pseudobulked by pseudotime derived cell-type and experimental replicate. We filtered the gene list to those expressed in at least 1% of cells in at least one cell-type giving a list of 10269 genes and then performed DESeq2 pairwise on each adjacent pseudotime segment. Genes with significant changes in expression at a transition were taken as those with at least a 1.5 fold change and an adjusted p-value of less than 0.05.

Data-driven analysis of gene markers

A data-driven approach was developed to correlate the change in bulk gene expression with the developmental gene expression state transitions. This analysis follows a two-step framework. In the first step, single-cell RNA sequencing (scRNA-seq) data from wild-type embryos were analyzed to identify key genes that are strongly correlated with cell state transitions using a Random Forest (RF) decision tree-based method. The contribution of each gene to single-cell state transitions was then quantified using Shapley values. The second step utilizes these Shapley values across all genes and cells to quantify how changes in bulk gene expression influence the prediction of cell gene expression states.

A flowchart summarizing the first step is shown in Fig 2D. The analysis utilizes scRNA-seq data from wild-type embryos. Gene expression state data from [14], identified by segmentation of pseudotime using a variance minimization algorithm, were used as the target labels. To identify informative genes, an initial Random Forest (RF) classifier was trained on the full gene-expression matrix, and genes were initially ranked according to their impurity-based feature-importance scores [52], also known as mean decrease impurity feature importance which reflect each gene's cumulative contribution to classification across the RF ensemble. Based on this ranking, the top n genes were selected to construct reduced-feature RF models. Because impurity-based importance can be biased when features are highly correlated in gene-expression data [93], we further examined the contribution of each selected gene by calculating its Shapley value. The dataset was randomly divided into 80% training cells and 20% testing cells.

Starting with the single highest-ranked gene ($n = 1$), the RF classifier was iteratively retrained while incrementally adding genes in descending order of impurity-based importance. For each iteration, five-fold cross-validation was performed on the training data, using four folds for training and one for validation. Predictive performance was quantified as the mean of the Area Under the receiver operating characteristic Curve (AUC) across all folds. The point of AUC convergence was used to determine the optimal number of genes required for accurate classification while mitigating sample-size imbalance effects.

During training, the RF hyperparameters, such as the number of estimators, maximum tree depth, minimum samples required for split, minimum samples per leaf, and the number of features considered at each split were jointly optimized with respect to the number of selected genes to maximize the AUC performance. In this work, when adding additional genes no longer improved the AUC, the corresponding subset of genes was selected as the final minimal gene set.

sufficient to define the respective cell state transitions. The resulting AUC convergence curves on the testing dataset and the corresponding optimal feature set sizes for each transition are presented in Fig 2F.

The final gene subsets identified by the decision tree-based approach contained 230,232,110, 55, and 235 genes for the transitions from NMP to neuronal, NMP to MP, MP to PZ, PZ to pPSM and pPSM to aPSM, respectively. Using the trained RF classifier structures and these optimized gene sets, we then applied the SHAP (Shapley Additive exPlanations) framework to compute Shapley values for each gene across all single cells in the dataset. The scatter plots in Fig 3B–F illustrate the 10 most influential genes for wild-type cell state transitions, ranked by their mean absolute Shapley values (left). The summary plots (right) integrate both feature importance and the directionality of each gene's contribution to the cell state transition.

An essential component in the analysis of bulk gene expression is how the specific changes induced by the *tbx16* or *msgn1* compared to wild type could influence the transition of gene expression states. In the second step, an algorithm is developed to quantify the feature attribution when gene expression levels are scaled up or down by a constant factor k . Importantly, this algorithm facilitates comparisons between a gene's specific fold change and influence on the gene expression states transition. Using the Shapley matrix calculated in the first step, we developed a log-scale sensitivity measure to quantify how changes in bulk gene expression affect the Shapley values in the cell state transitions.

Here, to normalize variation and reduce skewness of the expression value x_j of a given gene in cell j ($j = 1, \dots, N$), we log-transformed and standardized it into standardized expression z_j by

$$v_j = \log_2(x_j + 1)$$

$$\mu_v = \frac{1}{N} \sum_{j=1}^N v_j, \sigma_v = \sqrt{\frac{1}{N-1} \sum_{j=1}^N (v_j - \mu_v)^2}$$

$$z_j = \frac{v_j - \mu_v}{\sigma_v}$$

where v_j is the log-transform of the x_j , μ_v and σ_v are the mean and variance of v_j . x_j is added by 1 in the log2-fold change to avoid negative infinity for v_j .

We approximate the relationship between z_j and the corresponding Shapley value ϕ_j using a local linear model:

$$\phi(z) \approx \alpha + \beta z$$

where the slope of this linear regression β is calculated by

$$\beta = \frac{\text{Cov}(z, \phi)}{\text{Var}(z)}$$

The expected change in the mean Shapley value resulting from scaling the group's average expression $\bar{x}$ by a factor k is then approximated by:

$$\Delta(k) \approx \phi\left(\frac{\log_2(k\bar{x} + 1) - \mu_v}{\sigma_v}\right) - \phi\left(\frac{\log_2(\bar{x} + 1) - \mu_v}{\sigma_v}\right) \approx \left(\frac{\beta}{\sigma_v}\right) \log_2 k$$

Here, we define the log-scale sensitivity S as:

$$S = \frac{\beta}{\sigma_v}$$

This term captures the proportional change in the Shapley value per log2-fold change in expression. Under this formulation, the relationship is symmetric, which indicates that increasing or decreasing expression by the same fold produces Shapley changes of equal magnitude but opposite direction:

$$\Delta(1/k) = -\Delta(k)$$

This log-scale sensitivity S provides an interpretable, model-agnostic metric that enables comparability across genes with computational efficiency. Shapley sensitivity can be directly compared across different genes or cell states. The equation above offers a closed-form, linearized estimate of Shapley change without repeated model evaluation. This algorithm provides a fast and interpretable method for quantifying the responsiveness of model attributions to gene expression scaling in normalized log space.

In addition to the machine learning analysis of the bulk genes' expression, the direct comparison of the bulk genes' log fold changes by pseudobulk versus the genes' expression changed in the wild-type single cell differentiation transitions stage is also compared in [S2-S3 Figs.](#)

Identifying genes downstream of *tbx16* and *msgn1*

The genes list for each cell state transition in wild-type tailbuds were compared to the lists of DEGs in *tg(hsp70l:tbx16)* embryos and *tg(hsp70l:msgn1)* embryos, respectively, and common genes were found in each category. 1.5-fold-change threshold to all the gene lists was used. The common genes were then sorted by positive and negative fold change in each list. The number of genes in each category were counted, and all the gene names were listed.

Supporting information

S1 Fig. Overview of RNA sequencing of pooled dissected tailbuds from *tg(hsp70l:tbx16)*, and *tg(hsp70l:msgn1)* embryos. (A-B) Heatmap and MA plot of RNA-seq experiments comparing *tbx16* overexpression vs wild type. (C-D) Heatmap and MA plot of RNA-seq experiments comparing *msgn1* overexpression vs wild type. Each genotype has 3 biological replicates. All the identified differentially expressed genes (DEGs) have adjusted p-values < 0.05. Heatmaps are ranked by z-scores.
(TIF)

S2 Fig. Transcriptional regulation by *tbx16* during cell state transitions of paraxial mesoderm differentiation. (A) A schematic showing the developmental trajectory of the paraxial mesoderm in the tailbud. Body axes are shown as D (dorsal), V (ventral), A (anterior), P (posterior), R (right) and L (left). Cell state transitions are shown as NMP to neural, NMP to MP, MP to PZ, PZ to pPSM, and pPSM to aPSM. Each transition is labeled with an arrow indicating the direction of transition. (B-F) The logfold change of bulk gene expression induced by *tbx16* overexpression over wild type (y-axis) versus the logfold change of single-cell gene expression for the transitions in wild-type gene expression states (x-axis). Key genes identified by DESeq2 plus a 1.5-fold expression change threshold, either up- or downregulated (WT^+ , WT^-), and overlapping with genes up- or downregulated (≥ 1.5 -fold; $tbx16^+$, $tbx16^-$) under *tbx16* overexpression are analyzed across the transitions from NMP to neural (B), NMP to MP (C), MP to PZ (D), PZ to pPSM (E), and pPSM to aPSM (F).
(TIF)

S3 Fig. Transcriptional regulation by *msgn1* during cell state transitions of paraxial mesoderm differentiation.

(A) A schematic showing the developmental trajectory of the paraxial mesoderm in the tailbud. Body axes are shown as D (dorsal), V (ventral), A (anterior), P (posterior), R (right) and L (left). Cell state transitions are shown as NMP to neural, NMP to MP, MP to PZ, PZ to pPSM, and pPSM to aPSM. Each transition is labeled with an arrow as the direction of transition (B-F) The logfold change of bulk gene expression induced by *msgn1* overexpression over wild type (y-axis) versus the logfold change of single-cell gene expression during wild-type cell state transitions (x-axis). Key genes identified by DESeq2 plus a 1.5-fold expression change threshold, either up- or downregulated (WT^+ , WT^-), and overlapping with genes up- or downregulated (≥ 1.5 -fold; $msgn1^+$, $msgn1^-$) under *msgn1* overexpression are analyzed across the transitions from NMP to neural (B), NMP to MP (C), MP to PZ (D), PZ to pPSM (E), and pPSM to aPSM (F). (TIF)

S1 Table. RNAseq results for *tg(hsp70l:tbx16)* compared to WT.
(XLSX)

S2 Table. RNAseq results for *tg(hsp70l:msgn1)* compared to WT.
(XLSX)

S3 Table. DESeq2 identification of differentially expressed genes at transcriptional cell state transitions.
(XLSX)

S4 Table. *Tbx16*-regulated genes identified with DESeq2 plotted in S2 Fig.
(XLSX)

S5 Table. *Msgn1*-regulated genes with DESeq2 plotted in S3 Fig.
(XLSX)

S6 Table. Machine learning identification and S values of differentially expressed genes at transcriptional cell state transitions.
(XLSX)

S7 Table. *Tbx16*-regulated genes identified with machine learning plotted in Fig 3.
(XLSX)

S8 Table. *Msgn1*-regulated genes identified with machine learning plotted in Fig 4.
(XLSX)

S9 Table. Primer Sequences.
(XLSX)

Acknowledgments

We thank the Yale Center for Genomic Analysis for RNA sequencing services and Dörthe Jülich for help with experiments.

Author contributions

Conceptualization: Scott A Holley.

Formal analysis: Guoyu Zhu, Miriam A Genuth, Yanrong Xiao.

Funding acquisition: Scott A Holley.

Investigation: Guoyu Zhu, Miriam A Genuth, Abigail A Kindberg, Kayleigh Hackett.

Project administration: Scott A Holley.

Resources: Scott A Holley.

Software: Yanrong Xiao.

Supervision: Scott A Holley.

Visualization: Guoyu Zhu, Miriam A Genuth, Yanrong Xiao.

Writing – original draft: Guoyu Zhu, Miriam A Genuth, Yanrong Xiao, Scott A Holley.

Writing – review & editing: Abigail A Kindberg, Kayleigh Hackett.

References

1. Lawton AK, Nandi A, Stulberg MJ, Dray N, Sneddon MW, Pontius W, et al. Regulated tissue fluidity steers zebrafish body elongation. *Development*. 2013;140(3):573–82. <https://doi.org/10.1242/dev.090381> PMID: 23293289
2. Stooke-Vaughan GA, Kim S, Yen S-T, Son K, Banavar SP, Giammona J, et al. The physical roles of different posterior tissues in zebrafish axis elongation. *Nat Commun*. 2025;16(1):1839. <https://doi.org/10.1038/s41467-025-56334-7> PMID: 39984461
3. Steventon B, Duarte F, Lagadec R, Mazan S, Nicolas J-F, Hirsinger E. Species-specific contribution of volumetric growth and tissue convergence to posterior body elongation in vertebrates. *Development*. 2016;143(10):1732–41. <https://doi.org/10.1242/dev.126375> PMID: 26989170
4. Zhang L, Kendrick C, Jülich D, Holley SA. Cell cycle progression is required for zebrafish somite morphogenesis but not segmentation clock function. *Development*. 2008;135(12):2065–70. <https://doi.org/10.1242/dev.022673> PMID: 18480162
5. Kanki JP, Ho RK. The development of the posterior body in zebrafish. *Development*. 1997;124(4):881–93. <https://doi.org/10.1242/dev.124.4.881> PMID: 9043069
6. Martin BL, Steventon B. A fishy tail: Insights into the cell and molecular biology of neuromesodermal cells from zebrafish embryos. *Dev Biol*. 2022;487:67–73. <https://doi.org/10.1016/j.ydbio.2022.04.010> PMID: 35525020
7. Tzouanacou E, Wegener A, Wymeersch FJ, Wilson V, Nicolas J-F. Redefining the progression of lineage segregations during mammalian embryogenesis by clonal analysis. *Dev Cell*. 2009;17(3):365–76. <https://doi.org/10.1016/j.devcel.2009.08.002> PMID: 19758561
8. Martin BL, Kimelman D. Canonical Wnt signaling dynamically controls multiple stem cell fate decisions during vertebrate body formation. *Dev Cell*. 2012;22(1):223–32. <https://doi.org/10.1016/j.devcel.2011.11.001> PMID: 22264734
9. Goto H, Kimmey SC, Row RH, Matus DQ, Martin BL. FGF and canonical Wnt signaling cooperate to induce paraxial mesoderm from tailbud neuromesodermal progenitors through regulation of a two-step epithelial to mesenchymal transition. *Development*. 2017;144(8):1412–24. <https://doi.org/10.1242/dev.143578> PMID: 28242612
10. Attardi A, Fulton T, Florescu M, Shah G, Muresan L, Lenz MO, et al. Neuromesodermal progenitors are a conserved source of spinal cord with divergent growth dynamics. *Development*. 2018;145(21):dev166728. <https://doi.org/10.1242/dev.166728> PMID: 30333213
11. Das D, Chatti V, Emonet T, Holley SA. Patterned disordered cell motion ensures vertebral column symmetry. *Dev Cell*. 2017;42(2):170–180.e5. <https://doi.org/10.1016/j.devcel.2017.06.020> PMID: 28743003
12. Mongera A, Rowghanian P, Gustafson HJ, Shelton E, Kealhofer DA, Carn EK, et al. A fluid-to-solid jamming transition underlies vertebrate body axis elongation. *Nature*. 2018;561(7723):401–5. <https://doi.org/10.1038/s41586-018-0479-2> PMID: 30185907
13. Genuth MA, Jülich D, Ton AT, Smith SJ, Guillon E, Shattuck MD, et al. A cadherin-integrin-ECM code for presomitic mesoderm fluidity. *Development*. 2025;152(21):dev204874. <https://doi.org/10.1242/dev.204874> PMID: 40995679
14. Genuth MA, Kojima Y, Jülich D, Kiryu H, Holley SA. Automated time-lapse data segmentation reveals in vivo cell state dynamics. *Sci Adv*. 2023;9(22):eadf1814. <https://doi.org/10.1126/sciadv.adf1814> PMID: 37267354
15. Ho RK, Kane DA. Cell-autonomous action of zebrafish spt-1 mutation in specific mesodermal precursors. *Nature*. 1990;348(6303):728–30. <https://doi.org/10.1038/348728a0> PMID: 2259382
16. Griffin KJ, Amacher SL, Kimmel CB, Kimelman D. Molecular identification of spadetail: Regulation of zebrafish trunk and tail mesoderm formation by T-box genes. *Development*. 1998;125(17):3379–88. <https://doi.org/10.1242/dev.125.17.3379> PMID: 9693141
17. Garnett AT, Han TM, Gilchrist MJ, Smith JC, Eisen MB, Wardle FC, et al. Identification of direct T-box target genes in the developing zebrafish mesoderm. *Development*. 2009;136(5):749–60. <https://doi.org/10.1242/dev.024703> PMID: 19158186
18. Fior R, Maxwell AA, Ma TP, Vezzaro A, Moens CB, Amacher SL, et al. The differentiation and movement of presomitic mesoderm progenitor cells are controlled by Mesogenin 1. *Development*. 2012;139(24):4656–65. <https://doi.org/10.1242/dev.078923> PMID: 23172917
19. Yabe T, Takada S. Mesogenin causes embryonic mesoderm progenitors to differentiate during development of zebrafish tail somites. *Dev Biol*. 2012;370(2):213–22. <https://doi.org/10.1016/j.ydbio.2012.07.029> PMID: 22890044
20. Horb ME, Thomsen GH. A vegetally localized T-box transcription factor in *Xenopus* eggs specifies mesoderm and endoderm and is essential for embryonic mesoderm formation. *Development*. 1997;124(9):1689–98. <https://doi.org/10.1242/dev.124.9.1689> PMID: 9165117

21. Knezevic V, De Santo R, Mackem S. Two novel chick T-box genes related to mouse Brachyury are expressed in different, non-overlapping mesodermal domains during gastrulation. *Development*. 1997;124(2):411–9. <https://doi.org/10.1242/dev.124.2.411> PMID: 9053317
22. Lustig KD, Kroll KL, Sun EE, Kirschner MW. Expression cloning of a *Xenopus* T-related gene (Xombi) involved in mesodermal patterning and blastopore lip formation. *Development*. 1996;122(12):4001–12. <https://doi.org/10.1242/dev.122.12.4001> PMID: 9012520
23. Stennard F, Carnac G, Gurdon JB. The *Xenopus* T-box gene, Antipodean, encodes a vegetally localised maternal mRNA and can trigger mesoderm formation. *Development*. 1996;122(12):4179–88. <https://doi.org/10.1242/dev.122.12.4179> PMID: 9012537
24. Zhang J, King ML. *Xenopus* VegT RNA is localized to the vegetal cortex during oogenesis and encodes a novel T-box transcription factor involved in mesodermal patterning. *Development*. 1996;122(12):4119–29. <https://doi.org/10.1242/dev.122.12.4119> PMID: 9012531
25. Goering LM, Hoshijima K, Hug B, Bisgrove B, Kispert A, Grunwald DJ. An interacting network of T-box genes directs gene expression and fate in the zebrafish mesoderm. *Proc Natl Acad Sci U S A*. 2003;100(16):9410–5. <https://doi.org/10.1073/pnas.1633548100> PMID: 12883008
26. Jahangiri L, Nelson AC, Wardle FC. A cis-regulatory module upstream of deltaC regulated by Ntla and Tbx16 drives expression in the tailbud, presomitic mesoderm and somites. *Dev Biol*. 2012;371(1):110–20. <https://doi.org/10.1016/j.ydbio.2012.07.002> PMID: 22877946
27. Griffin KJP, Kimelman D. Interplay between FGF, one-eyed pinhead, and T-box transcription factors during zebrafish posterior development. *Dev Biol*. 2003;264(2):456–66. <https://doi.org/10.1016/j.ydbio.2003.09.008> PMID: 14651930
28. Bouldin CM, Manning AJ, Peng Y-H, Farr GH, Hung KL, Dong A, et al. Wnt signaling and tbx16 form a bistable switch to commit bipotential progenitors to mesoderm. *Development*. 2015;142(14):2499–507. <https://doi.org/10.1242/dev.124024> PMID: 26062939
29. Payumo AY, McQuade LE, Walker WJ, Yamazoe S, Chen JK. Tbx16 regulates hox gene activation in mesodermal progenitor cells. *Nat Chem Biol*. 2016;12(9):694–701. <https://doi.org/10.1038/nchembio.2124> PMID: 27376691
30. Mueller RL, Huang C, Ho RK. Spatio-temporal regulation of Wnt and retinoic acid signaling by tbx16/spadetail during zebrafish mesoderm differentiation. *BMC Genomics*. 2010;11:492. <https://doi.org/10.1186/1471-2164-11-492> PMID: 20828405
31. Warga RM, Mueller RL, Ho RK, Kane DA. Zebrafish Tbx16 regulates intermediate mesoderm cell fate by attenuating Fgf activity. *Dev Biol*. 2013;383(1):75–89. <https://doi.org/10.1016/j.ydbio.2013.08.018> PMID: 24008197
32. Osborn DPS, Li K, Cutty SJ, Nelson AC, Wardle FC, Hinitis Y, et al. Fgf-driven Tbx protein activities directly induce myf5 and myod to initiate zebrafish myogenesis. *Development*. 2020;147(8):dev184689. <https://doi.org/10.1242/dev.184689> PMID: 32345657
33. Buchberger A, Bonneick S, Arnold H. Expression of the novel basic-helix-loop-helix transcription factor cMesp in presomitic mesoderm of chicken embryos. *Mech Dev*. 2000;97(1–2):223–6. [https://doi.org/10.1016/s0925-4773\(00\)00424-x](https://doi.org/10.1016/s0925-4773(00)00424-x) PMID: 11025230
34. Yoon JK, Wold B. The bHLH regulator pMesogenin1 is required for maturation and segmentation of paraxial mesoderm. *Genes Dev*. 2000;14(24):3204–14. <https://doi.org/10.1101/gad.850000> PMID: 11124811
35. Yoon JK, Moon RT, Wold B. The bHLH class protein pMesogenin1 can specify paraxial mesoderm phenotypes. *Dev Biol*. 2000;222(2):376–91. <https://doi.org/10.1006/dbio.2000.9717> PMID: 10837126
36. Yoo K-W, Kim C-H, Park H-C, Kim S-H, Kim H-S, Hong S-K, et al. Characterization and expression of a presomitic mesoderm-specific mespo gene in zebrafish. *Dev Genes Evol*. 2003;213(4):203–6. <https://doi.org/10.1007/s00427-003-0312-1> PMID: 12684777
37. Wang J, Li S, Chen Y, Ding X. Wnt/beta-catenin signaling controls Mesp expression to regulate segmentation during *Xenopus* somitogenesis. *Dev Biol*. 2007;304(2):836–47. <https://doi.org/10.1016/j.ydbio.2006.12.034> PMID: 17266950
38. Witter L, Shin E, Grote P, Kispert A, Beckers A, Gossler A, et al. Expression of Msn1 in the presomitic mesoderm is controlled by synergism of WNT signalling and Tbx6. *EMBO Rep*. 2007;8(8):784–9. <https://doi.org/10.1038/sj.embor.7401030> PMID: 17668009
39. Tazumi S, Yabe S, Yokoyama J, Aihara Y, Uchiyama H. PMesogenin1 and 2 function directly downstream of Xtbx6 in *Xenopus* somitogenesis and myogenesis. *Dev Dyn*. 2008;237(12):3749–61. <https://doi.org/10.1002/dvdy.21791> PMID: 19035338
40. Chalamalasetty RB, Dunty WC, Biris KK, Ajima R, Iacovino M, Beisaw A, et al. The Wnt3a/β-catenin target gene Mesogenin1 controls the segmentation clock by activating a Notch signalling program. *Nat Commun*. 2011;2:390. <https://doi.org/10.1038/ncomms1381> PMID: 21750544
41. Chalamalasetty RB, Garriock RJ, Dunty WC, Kennedy MW, Jailwala P, Si H, et al. Mesogenin 1 is a master regulator of paraxial presomitic mesoderm differentiation. *Development*. 2014;141(22):4285–97. <https://doi.org/10.1242/dev.110908> PMID: 25371364
42. Saunders LM, Srivatsan SR, Duran M, Dorrity MW, Ewing B, Linbo TH, et al. Embryo-scale reverse genetics at single-cell resolution. *Nature*. 2023;623(7988):782–91. <https://doi.org/10.1038/s41586-023-06720-2> PMID: 37968389
43. Manning AJ, Kimelman D. Tbx16 and Msn1 are required to establish directional cell migration of zebrafish mesodermal progenitors. *Dev Biol*. 2015;406(2):172–85. <https://doi.org/10.1016/j.ydbio.2015.09.001> PMID: 26368502
44. Nowotschin S, Ferrer-Vaquer A, Concepcion D, Papaioannou VE, Hadjantonakis A-K. Interaction of Wnt3a, Msn1 and Tbx6 in neural versus paraxial mesoderm lineage commitment and paraxial mesoderm differentiation in the mouse embryo. *Dev Biol*. 2012;367(1):1–14. <https://doi.org/10.1016/j.ydbio.2012.04.012> PMID: 22546692
45. Morrow ZT, Maxwell AM, Hoshijima K, Talbot JC, Grunwald DJ, Amacher SL. tbx6l and tbx16 are redundantly required for posterior paraxial mesoderm formation during zebrafish embryogenesis. *Dev Dyn*. 2017;246(10):759–69. <https://doi.org/10.1002/dvdy.24547> PMID: 28691257
46. Shapley LS. *Stochastic Games**. Proceedings of the National Academy of Sciences. 1953;39(10):1095–100.
47. Lele Z, Engel S, Krone PH. hsp47 and hsp70 gene expression is differentially regulated in a stress- and tissue-specific manner in zebrafish embryos. *Dev Genet*. 1997;21(2):123–33. [https://doi.org/10.1002/\(SICI\)1520-6408\(1997\)21:2<123::AID-DVG2>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1520-6408(1997)21:2<123::AID-DVG2>3.0.CO;2-9) PMID: 9332971

48. Kawakami K. Tol2: A versatile gene transfer vector in vertebrates. *Genome Biol.* 2007;8 Suppl 1(Suppl 1):S7. <https://doi.org/10.1186/gb-2007-8-s1-s7> PMID: [18047699](#)
49. Row RH, Tsotras SR, Goto H, Martin BL. The zebrafish tailbud contains two independent populations of midline progenitor cells that maintain long-term germ layer plasticity and differentiate in response to local signaling cues. *Development.* 2016;143(2):244–54. <https://doi.org/10.1242/dev.129015> PMID: [26674311](#)
50. Stulberg MJ, Lin A, Zhao H, Holley SA. Crosstalk between Fgf and Wnt signaling in the zebrafish tailbud. *Dev Biol.* 2012;369(2):298–307. <https://doi.org/10.1016/j.ydbio.2012.07.003> PMID: [22796649](#)
51. Talbot WS, Trevarrow B, Halpern ME, Melby AE, Farr G, Postlethwait JH, et al. A homeobox gene essential for zebrafish notochord development. *Nature.* 1995;378(6553):150–7. <https://doi.org/10.1038/378150a0> PMID: [7477317](#)
52. Breiman L. Random Forests. *Machine Learning.* 2001;45(1):5–32.
53. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems.* 2017. 4768–77.
54. Williams DR, Shifley ET, Braunreiter KM, Cole SE. Disruption of somitogenesis by a novel dominant allele of *Lfng* suggests important roles for protein processing and secretion. *Development.* 2016;143(5):822–30. <https://doi.org/10.1242/dev.128538> PMID: [26811377](#)
55. Yang Y, Thorpe C. BMP and non-canonical Wnt signaling are required for inhibition of secondary tail formation in zebrafish. *Development.* 2011;138(12):2601–11. <https://doi.org/10.1242/dev.058404> PMID: [21610036](#)
56. Chng SC, Ho L, Tian J, Reversade B. ELABELA: A hormone essential for heart development signals via the apelin receptor. *Dev Cell.* 2013;27(6):672–80. <https://doi.org/10.1016/j.devcel.2013.11.002> PMID: [24316148](#)
57. Norris ML, Pauli A, Gagnon JA, Lord ND, Rogers KW, Mosimann C, et al. Toddler signaling regulates mesodermal cell migration downstream of Nodal signaling. *Elife.* 2017;6:e22626. <https://doi.org/10.7554/eLife.22626> PMID: [29117894](#)
58. Ramel M-C, Buckles GR, Baker KD, Lekven AC. WNT8 and BMP2B co-regulate non-axial mesoderm patterning during zebrafish gastrulation. *Dev Biol.* 2005;287(2):237–48. <https://doi.org/10.1016/j.ydbio.2005.08.012> PMID: [16216234](#)
59. Gouti M, et al. A gene regulatory network balances neural and mesoderm specification during vertebrate trunk development. *Dev Cell.* 2017;41(3):243–261.e7.
60. Ye Z, Kimelman D. Hox13 genes are required for mesoderm formation and axis elongation during early zebrafish development. *Development.* 2020;147(22):dev185298. <https://doi.org/10.1242/dev.185298> PMID: [33154036](#)
61. Abe G, Ide H, Tamura K. Function of FGF signaling in the developmental process of the median fin fold in zebrafish. *Dev Biol.* 2007;304(1):355–66. <https://doi.org/10.1016/j.ydbio.2006.12.040> PMID: [17258191](#)
62. Zhang J, Jiang Z, Liu X, Meng A. Eph/ephrin signaling maintains the boundary of dorsal forerunner cell cluster during morphogenesis of the zebrafish embryonic left-right organizer. *Development.* 2016;143(14):2603–15. <https://doi.org/10.1242/dev.132969> PMID: [27287807](#)
63. Valdivia LE, Young RM, Hawkins TA, Stickney HL, Cavodeassi F, Schwarz Q, et al. Lef1-dependent Wnt/ β -catenin signalling drives the proliferative engine that maintains tissue homeostasis during lateral line development. *Development.* 2011;138(18):3931–41. <https://doi.org/10.1242/dev.062695> PMID: [21862557](#)
64. McGraw HF, Drerup CM, Culbertson MD, Linbo T, Raible DW, Nechiporuk AV. Lef1 is required for progenitor cell identity in the zebrafish lateral line primordium. *Development.* 2011;138(18):3921–30. <https://doi.org/10.1242/dev.062554> PMID: [21862556](#)
65. Schröter C, Oates AC. Segment number and axial identity in a segmentation clock period mutant. *Curr Biol.* 2010;20(14):1254–8. <https://doi.org/10.1016/j.cub.2010.05.071> PMID: [20637625](#)
66. Pelliccia JL, Jindal GA, Burdine RD. Gdf3 is required for robust Nodal signaling during germ layer formation and left-right patterning. *Elife.* 2017;6:e28635. <https://doi.org/10.7554/eLife.28635> PMID: [29140250](#)
67. Montague TG, Schier AF. Vg1-Nodal heterodimers are the endogenous inducers of mesendoderm. *Elife.* 2017;6:e28183. <https://doi.org/10.7554/eLife.28183> PMID: [29140251](#)
68. Shankaran SS, Sieger D, Schröter C, Czepe C, Pauly M-C, Laplante MA, et al. Completing the set of h/E(spl) cyclic genes in zebrafish: her12 and her15 reveal novel modes of expression and contribute to the segmentation clock. *Dev Biol.* 2007;304(2):615–32. <https://doi.org/10.1016/j.ydbio.2007.01.004> PMID: [17274976](#)
69. Thorpe CJ, Weidinger G, Moon RT. Wnt/ β -catenin regulation of the Sp1-related transcription factor sp5l promotes tail development in zebrafish. *Development.* 2005;132(8):1763–72. <https://doi.org/10.1242/dev.01733> PMID: [15772132](#)
70. Hammerschmidt M, Pelegri F, Mullins MC, Kane DA, van Eeden FJ, Granato M, et al. dino and mercedes, two genes regulating dorsal development in the zebrafish embryo. *Development.* 1996;123:95–102. <https://doi.org/10.1242/dev.123.1.95> PMID: [9007232](#)
71. Aulehla A, Wehrle C, Brand-Saberi B, Kemler R, Gossler A, Kanzler B, et al. Wnt3a plays a major role in the segmentation clock controlling somitogenesis. *Dev Cell.* 2003;4(3):395–406. [https://doi.org/10.1016/s1534-5807\(03\)00055-8](https://doi.org/10.1016/s1534-5807(03)00055-8) PMID: [12636920](#)
72. Aulehla A, Wieggraebe W, Baubet V, Wahl MB, Deng C, Taketo M, et al. A β -catenin gradient links the clock and wavefront systems in mouse embryo segmentation. *Nat Cell Biol.* 2008;10(2):186–93. <https://doi.org/10.1038/ncb1679> PMID: [18157121](#)
73. Sawada A, Shinya M, Jiang YJ, Kawakami A, Kuroiwa A, Takeda H. Fgf/MAPK signalling is a crucial positional cue in somite boundary formation. *Development.* 2001;128(23):4873–80. <https://doi.org/10.1242/dev.128.23.4873> PMID: [11731466](#)

74. Dubrulle J, McGrew MJ, Pourqu   O. FGF signaling controls somite boundary position and regulates segmentation clock control of spatiotemporal Hox gene activation. *Cell*. 2001;106(2):219–32. [https://doi.org/10.1016/s0092-8674\(01\)00437-8](https://doi.org/10.1016/s0092-8674(01)00437-8) PMID: [11511349](#)
75. Akiyama R, Masuda M, Tsuge S, Bessho Y, Matsui T. An anterior limit of FGF/Erk signal activity marks the earliest future somite boundary in zebrafish. *Development*. 2014;141(5):1104–9. <https://doi.org/10.1242/dev.098905> PMID: [24504340](#)
76. Bajard L, Morelli LG, Ares S, P  craux J, J  licher F, Oates AC. Wnt-regulated dynamics of positional information in zebrafish somitogenesis. *Development*. 2014;141(6):1381–91. <https://doi.org/10.1242/dev.093435> PMID: [24595291](#)
77. Simsek MF,   zbudak EM. Spatial fold change of FGF signaling encodes positional information for segmental determination in zebrafish. *Cell Rep*. 2018;24(1):66–78.e8. <https://doi.org/10.1016/j.celrep.2018.06.023> PMID: [29972792](#)
78. Pyati UJ, Webb AE, Kimelman D. Transgenic zebrafish reveal stage-specific roles for Bmp signaling in ventral and posterior mesoderm development. *Development*. 2005;132(10):2333–43. <https://doi.org/10.1242/dev.01806> PMID: [15829520](#)
79. Connors SA, Tucker JA, Mullins MC. Temporal and spatial action of tolloid (mini fin) and chordin to pattern tail tissues. *Dev Biol*. 2006;293(1):191–202. <https://doi.org/10.1016/j.ydbio.2006.01.029> PMID: [16530746](#)
80. Stickney HL, Imai Y, Draper B, Moens C, Talbot WS. Zebrafish bmp4 functions during late gastrulation to specify ventroposterior cell fates. *Dev Biol*. 2007;310(1):71–84. <https://doi.org/10.1016/j.ydbio.2007.07.027> PMID: [17727832](#)
81. O'Neill K, Thorpe C. BMP signaling and spadetail regulate exit of muscle precursors from the zebrafish tailbud. *Dev Biol*. 2013;375(2):117–27. <https://doi.org/10.1016/j.ydbio.2012.12.002> PMID: [23246591](#)
82. Das D, et al. Organization of embryonic morphogenesis via mechanical information. *Dev Cell*. 2019;49(6):829–839.e5.
83. Yabe T, Hoshijima K, Yamamoto T, Takada S. Quadruple zebrafish mutant reveals different roles of Mesp genes in somite segmentation between mouse and zebrafish. *Development*. 2016;143(15):2842–52. <https://doi.org/10.1242/dev.133173> PMID: [27385009](#)
84. Takahashi Y, Yasuhiko Y, Kitajima S, Kanno J, Saga Y. Appropriate suppression of Notch signaling by Mesp factors is essential for stripe pattern formation leading to segment boundary formation. *Dev Biol*. 2007;304(2):593–603. <https://doi.org/10.1016/j.ydbio.2007.01.007> PMID: [17306789](#)
85. Petravicovici A, Bonasio R. Distinct PRC2 subunits regulate maintenance and establishment of Polycomb repression during differentiation. *Mol Cell*. 2021;81(12):2625–2639.e5. <https://doi.org/10.1016/j.molcel.2021.03.038> PMID: [33887196](#)
86. Castelo-Branco G, Lilja T, Wallenborg K, Falc  o AM, Marques SC, Gracias A, et al. Neural stem cell differentiation is dictated by distinct actions of nuclear receptor corepressors and histone deacetylases. *Stem Cell Reports*. 2014;3(3):502–15. <https://doi.org/10.1016/j.stemcr.2014.07.008> PMID: [25241747](#)
87. Zhang K, Chen S, Yang Q, Guo S, Chen Q, Liu Z, et al. The Oligodendrocyte Transcription Factor 2 OLIG2 regulates transcriptional repression during myelinogenesis in rodents. *Nat Commun*. 2022;13(1):1423. <https://doi.org/10.1038/s41467-022-29068-z> PMID: [35301318](#)
88. Zhang J, Roberts JM, Chang F, Schwakopf J, Vetter ML. Jarid2 promotes temporal progression of retinal progenitors via repression of Foxp1. *Cell Rep*. 2023;42(3):112237. <https://doi.org/10.1016/j.celrep.2023.112237> PMID: [36924502](#)
89. Mistry DS, Chen Y, Wang Y, Zhang K, Sen GL. SNAIL2 controls the undifferentiated state of human epidermal progenitor cells. *Stem Cells*. 2014;32(12):3209–18. <https://doi.org/10.1002/stem.1809> PMID: [25100569](#)
90. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 2019;37(8):907–15. <https://doi.org/10.1038/s41587-019-0201-4> PMID: [31375807](#)
91. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015;33(3):290–5. <https://doi.org/10.1038/nbt.3122> PMID: [25690850](#)
92. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. <https://doi.org/10.1186/s13059-014-0550-8> PMID: [25516281](#)
93. Strobl C, et al. Bias in random forest variable importance measures. Workshop on statistical modelling of complex systems. 2006.