bioRχiv

ARTICLE PREPRINT

Reconstructing signaling histories of single cells via perturbation screens and transfer learning

Nicholas T. Hutchins[1,2], Miram Meziane[2,3], Claire Lu[2,4], Maisam Mitalipova[2], David Fischer[5], Pulin Li[2,3]*

[1] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge MA, USA [2], Whitehead Institute for Biomedical Research, Cambridge MA, USA[3], Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA[4], Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA [5], Institute of Artificial Intelligence at the Medical University of Vienna, Vienna, AT

* Correspondence to: pli@wi.mit.edu

Manipulating the signaling environment is an effective approach to alter cellular states for broad-ranging applications, from engineering tissues to treating diseases. Such manipulation requires knowing the signaling states and histories of the cells *in situ*, for which high-throughput discovery methods are lacking. Here, we present an integrated experimental-computational framework that learns signaling response signatures from a high-throughput *in vitro* perturbation atlas and infers combinatorial signaling activities in *in vivo* cell types with high accuracy and temporal resolution. Specifically, we generated signaling perturbation atlas across diverse cell types/states through multiplexed sequential combinatorial screens on human pluripotent stem cells. Using the atlas to train IRIS, a neural network-based model, and predicting on mouse embryo scRNAseq atlas, we discovered global features of combinatorial signaling code usage over time, identified biologically meaningful heterogeneity of signaling states within each cell type, and reconstructed signaling histories along diverse cell lineages. We further demonstrated that IRIS greatly accelerates the optimization of stem cell differentiation protocols by drastically reducing the combinatorial space that needs to be tested. This framework leads to the revelation that different cell types share robust signal response signatures, and provides a scalable solution for mapping complex signaling interactions *in vivo* to guide targeted interventions.

INTRODUCTION

Cellular state transition driven by combinations of external signals is the foundation for building tissues in developing embryos, maintaining physiological homeostasis in adults, and triggering disease onset and progression. Identifying which combinatorial signaling conditions trigger state transition of the cells in their endogenous contexts is crucial for understanding and engineering the respective biological process. Genetic perturbations and transgenic signaling reporters are the major ways to correlate signaling pathways with cellular state transitions *in vivo*, but difficult to apply to human tissues [1,2]. Newer generations of synthetic tools for recording signals have been developed for cells cultured *in vitro*, but their application in animals remains to be tested [3–5]. Importantly, all the above-mentioned approaches are low-throughput and lacking temporal precision. In many systems, identifying the combinatorial signaling codes that drive cell state transition in time remains a challenging goal. The recent explosion of single-cell RNA sequencing (scRNAseq) and spatial transcriptomics data ushered in numerous computational algorithms to infer cell-cell communication based on gene expression [6–9]. These methods often rely on the presence of ligand-receptor pairs in the tissue, which does not guarantee signaling pathway activation [10–12]. Therefore, these methods estimate the signaling potential rather than the signaling state.

Signal activation induces transcriptional responses inside a cell, which is one of the most direct indicators of a cell's signaling state and compatible for multiplexed high-throughput measurements. However, given the number of signaling pathways, the diversity of cell types, and the combinatorial nature of signaling pathway usage, it is impractical to experimentally measure the transcriptional responses of all cell types upon stimulation by each defined signal. This raises the question of how many signal perturbation measurements are needed before we can accurately infer signaling states of most cell types in their endogenous contexts. A conceptual barrier for approaching this question has been the widely-held belief that transcriptional responses to signal activation are unique for each cell type [13,14]. If this is true, one would need to perform

signal perturbations on the exact cell type of interest, which is often not possible for *in vivo* cell types and certainly not scalable to most cell types. Therefore, we challenged the conventional notion by asking whether each signaling pathway has its own transcriptional signature that is shared across diverse cell types. If such signatures exist, it offers a new path for mapping cell signaling states and histories in their endogenous contexts.

Here, we developed a novel experimental-computational framework to learn response signatures from high-throughput *in vitro* perturbations and use these signatures to map the signaling activities in *in vivo* cell types (**Fig. 1a**). This framework revealed the existence of response signatures shared across diverse cell types, acting as a "fingerprint" for each signaling pathway. Using these "fingerprints", we mapped the signaling states of over 40 defined cell types in mouse embryos for the first time, revealing patterns of combinatorial code usage with unprecedented temporal resolution and biologically meaningful heterogeneity within each cell type. The framework relies on the design of sequential combinatorial signal screens on human embryonic stem cells (hESC) and modeling the data with a deep neural-network-based algorithm, namely Intracellular Response to Infer Signaling State (IRIS). The inference recovered signaling histories along well-studied cell lineages and made accurate predictions on understudied lineages, such as organ-specific mesenchyme, which was experimentally validated. IRIS greatly reduced the combinatorial signaling space that needs to be experimentally tested to develop hESC differentiation protocols *in vitro*. This new framework represents a conceptual and technical leap in predictive modeling of cell-cell communication that can enable discovery in fundamental biology and design of therapeutic interventions.

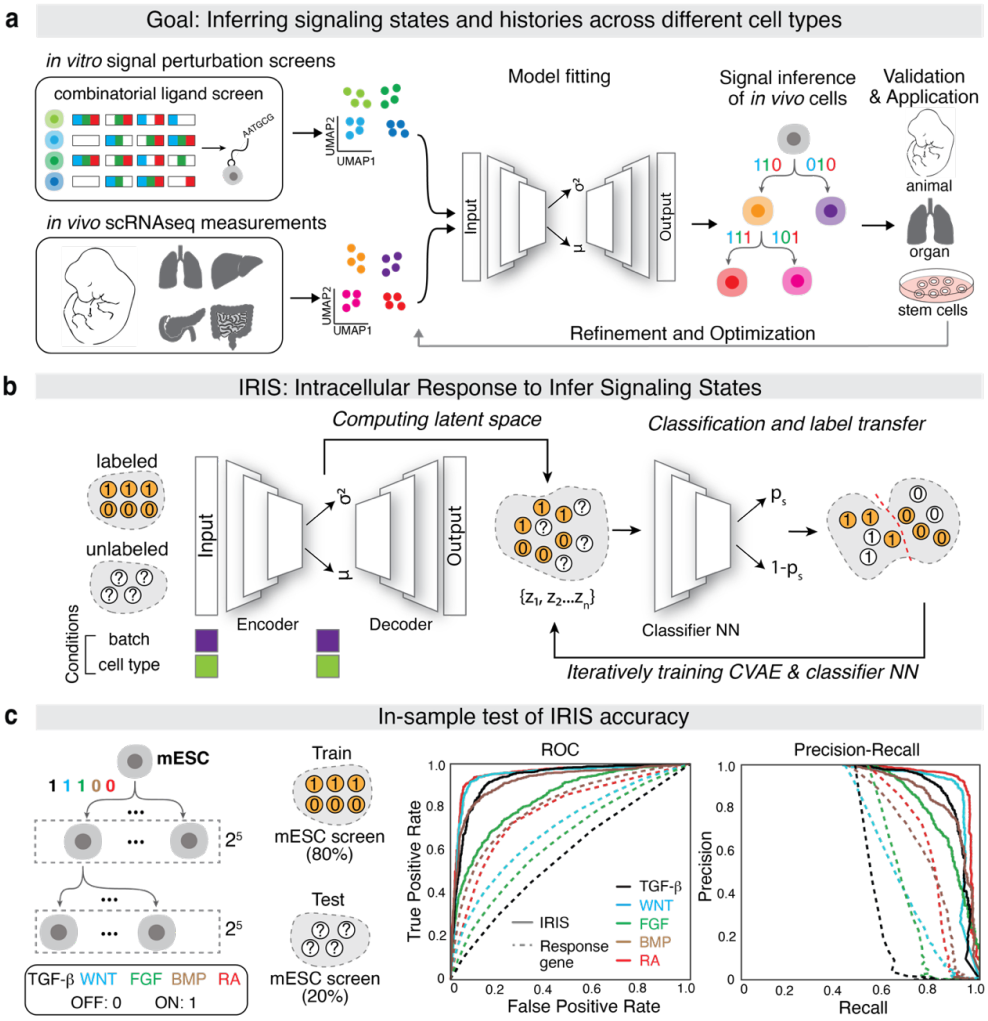

Fig 1. IRIS learns response signatures for accurate signaling state inference. a, Schematic of the proposed workflow for inferring signaling histories and optimizing hESC differentiation protocols. *In vitro* high-throughput signal perturbation data sample transcriptional responses to specific signaling pathways, and these response signatures can be learnt by transfer-learning models and used for inferring signaling states among *in vivo* cell types of interest. The inference can iteratively be used to propose new directed stem cell differentiation protocols. **b**, Schematic representation of the IRIS model. IRIS uses a CVAE (conditional variational autoencoder) architecture that includes a feedforward NN (neural network) for classification. This jointly computes a latent space and classifiers for signaling states in the space. **c**, IRIS improves accuracy of signaling state inference over the response-gene method on an 80/20 train-test split of the mESC differentiation data.

Learning generalizable signal response signatures from the ground up using IRIS

Despite the widely-held belief in the cell-type specificity of signal responses, a small number of genes have been used to indicate signal activation across cell types, such as *Axin2* for Wnt pathway and *Id1* for bone morphogenic protein (BMP) pathway [15–17]. Therefore, we first asked whether these empirically annotated response genes are sufficient for inferring the pathway activation states across diverse cell types [17]. We focused on the major developmental signaling pathways, including transforming growth factor beta (TGF-β), fibroblast growth factor (FGF), BMP, Hedgehog (HH), retinoic acid (RA), and Wnt. The activation and inhibition of these pathways occur repeatedly throughout development

and in adult tissue regeneration, and thus, they are the major pathways extensively used for pluripotent stem cell differentiation [18–20]. Using a rare large-scale signal perturbation dataset, in which mouse embryonic stem cells (mESCs) were stimulated with all possible combinations of these six pathways, we created response scores for each pathway in each cell, inferred signaling states, and quantified the inference accuracy (**Fig. S1**) [21]. The response genes had reasonable classification power on RA, WNT, and BMP pathways (>0.75 AUROC), but performed much worse for TGF-β and FGF. The HH pathway was not included because there were too few stimulated cells to render enough statistical power.

When applied to a gastrulating mouse embryo dataset (E6.5-E8.5), the response-gene method recovered some ground-truth knowledge, but in general, it was difficult to convert the response scores into a binary classification (**Fig. S2**) [33]. This is partly because there are no clear cross-cell type heuristics for determining what level of response gene expression constitutes a responding cell. Technical effects due to batch corrupt gene expression measurements, exacerbating the unreliability of the threshold. The small number of response genes used makes them sensitive to the well-known issue of dropouts for scRNAseq data [22]. Furthermore, these genes were compiled from limited biological contexts that have been extensively studied and may not be generalizable to other cell types. Altogether, these results suggest that shared response signatures might exist across cell types, but such a linear method based on a small number of genes has limited utility for classifying signaling states.

We hypothesized that this cross-cell type and cross-batch classification problem might be better solved by nonlinear models of gene expression that are fit to the entire transcriptome as genes do not operate independently of each other [23]. Models that incorporate the often-nonlinear relationship among genes across the transcriptome might be more likely to capture the response signatures, be robust to dropouts, and facilitate the choice of thresholds across cell types due to its increased flexibility. Our classification challenge is conceptually similar to the "label propagation" problem in the machine learning field, the goal of which is to learn data features associated with a label and then use those features to annotate unlabeled datasets [24]. Classic label propagation algorithms have had much success in transferring cell-type labels between different batches of scRNAseq data. In particular, algorithms based on conditional variational autoencoders (CVAEs), such as scANVI and trVAE, can remove the highly nonlinear batch effects more comprehensively and embed cells with similar gene expression patterns together more robustly than comparatively simpler models [25–27]. They also show success in leveraging small annotated datasets, projecting those onto much larger atlases and learning complicated thresholding rules for diverse cell types.

We asked whether we can apply CVAE-based neural networks to learn generalizable signal response signatures from the ground up. Using the mESC differentiation scRNAseq dataset as a starting point [21], we randomly split all the cells into a training set (80%) with signal conditions visible and a test set (20%) with signal conditions hidden. We first applied the CVAE to project all the data across technical batches into a single latent space without awareness of the signal conditions. We then jointly trained the latent space and a classifier network to learn signaling response features, predicted on the test set and compared the prediction with the true signal condition (**Fig. 1b**). When we fit the model using the default parameters of scANVI, we observed higher classification accuracy than the aforementioned response-gene method (**Fig. 1c**). Importantly, no threshold needed to be determined arbitrarily. We named this algorithm IRIS (Intracellular Response Inferred Signaling States).

Sequential combinatorial signal perturbation screen enables rapid sampling of diverse cell and signaling states

To test the cross-cell-type and cross-dataset performance of IRIS, we collected an orthogonal signal perturbation dataset on human embryonic stem cell (hESC) differentiation. Even though the mESC dataset includes over 100 final cell states, they were all derived from two initial cell states, namely mESC and anterior primitive streak (PS) [21]. To maximize the cell type/state diversity in the perturbation data, we applied a sequential combinatorial signal screening strategy (**Fig. 2a**). We first differentiated hESCs into the middle PS state using published protocols (Step 1), which is distinct from the anterior PS state in the mESC dataset [17,19]. We then created 46 distinct starting cell states by stimulating the middle PS cells with different combinations of the 6 signaling pathways for one day (Step 2). Finally, cells from each of the 46 starting cell states were further differentiated by a single or multiple signal combinations for another two days, totaling 96 final cell states (Step 3). Each of the 96 final cell states was barcoded and pooled for multiplexed scRNA-sequencing [28].

With 6 signaling pathways acting in two sequential steps, the number of possible combinations is 2^12, which exceeds the current throughput within a single experiment of hESC differentiation. To make the data as informative as possible, we represented each Step 2-Step 3 sequential combination as a 12-digit binary sequence and employed a few strategies to spread out the conditions in the combinatorial space (**Fig. 2a)**. We included 4 combinations that represent published protocols as controls, and 38 combinations that represent a single 'bit' flip away from these 4 positive controls [17]. The rest were randomly chosen

combinations with the constraint that every combination has to be away from all other combinations by at least a Hamming Distance of 4. For the most part, cells responded to all the signals provided and each sequential combination induced a relatively pure population (**Fig. S3a-d**). The screen produced more diverse final cell states overall, compared to the mESC screen, as measured by correlation of gene expression across the barcoded conditions (**Figs. 2b, S3e, f).** Together, this sequential combinatorial screen design optimizes for the diversity of cell types/states observed, which was not reported in other signal perturbation screens to our knowledge.

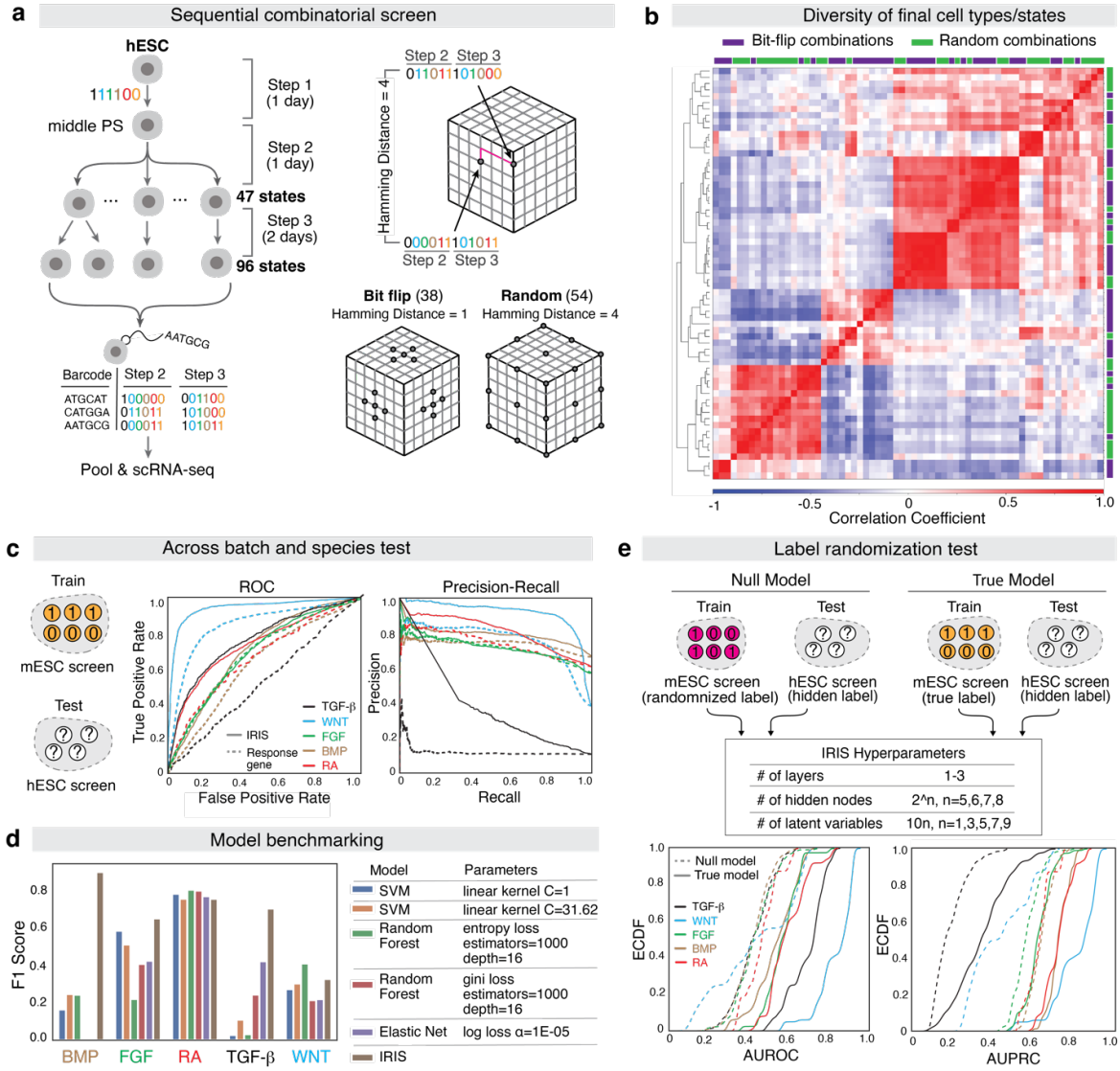

Figure 2. IRIS can generalize across batches, cell types and species. a, Design of the sequential combinatorial signal screen. hESCs were first differentiated into mPS (middle primitive streak) cells and then randomly stimulated with signal combinations for one day (Step 2). Cells from Step 2 were further randomly stimulated for two additional days in distinct randomized combinations, resulting in a total of 96 final states, which were barcoded and pooled for scRNA sequencing (*left*). Signal combinations are represented as points on a 12-dimensional space where the distance between each combination is computed with the Hamming distance. In the bit-flip strategy, 36 combinations were generated by choosing combinations of Hamming distance 1 away from one of the three control combinations. In the random strategy, 54 combinations were generated by randomly choosing combinations such that they all have a Hamming distance of at least 4 away from each other (*right*). **b**, The diversity of final cell states in the screen, quantified by the correlation coefficient averaged across the transcriptome between distinct signal combinations. **c**, Generalization of IRIS inference across cell types and species. IRIS was trained on the mESC screen and evaluated on the hESC screen. Hyperparameter selection was performed on the mESC screen only (**Fig. S5**). **d**, Comparison between a Null Model and a True Model, both trained on the mESC screen and evaluated for their inference accuracy on the hESC screen data across a range of hyperparameters. The True Model (*solid lines*) consistently outperforms the Null model (*dashed lines*). ECDF, empirical cumulative distribution function. AUROC, area under the receiving operator curve. AUPRC, area under the precision recall curve. **e**, Model benchmarking was performed by comparison to hyperparameter-optimized machine learning models. IRIS performance exceeds or compares to alternate models in terms of F1 score for all signaling pathways.

IRIS can generalize across batches, cell types and species

With the new perturbation data, we asked if IRIS can truly learn response signatures that generalize across batches, cell types and species. Given the well-known batch effect between datasets that are collected separately, we first optimized IRIS to minimize the batch effect. Taking advantage of the fact that the mESC data was collected as three independent batches, we performed a hyperparameter search to evaluate cross-batch performance by holding out one batch at a time. Notably, we found that shallow architectures with fewer layers performed the best in this test **(Fig. S4)**.

To test the cross-cell-type and cross-species performance of IRIS, we trained IRIS on all the data from the mESC screen using the optimal hyperparameters, and predicted on the hESC screen. We observed that IRIS improved the prediction for all pathways by both reducing false-positive and false-negative rates, compared to the response-gene method (**Fig. 2c**). Furthermore, we compared IRIS with simpler machine learning models, for which the hyperparameters were optimized in a similar way (**Figs. 2d, S5).** Using optimized hyperparameters for each model in several cross-validation tests, we observed that IRIS performed the best on average, especially achieving substantially better F1 scores (**Figs. 2e, S6)**. This suggests that IRIS may be learning how to threshold for classification from functions that these classic models cannot learn effectively. To understand if this improvement was due to learning true response signatures or spurious relationships in the data, we provided mESC data either with the ground-truth signaling labels (true model) or with the signaling labels completely randomized (null model), and found the true model outperformed the null model over a broad range of hyperparameters (**Fig. 2e**). Together, these results support the idea that there is sufficient information of signal activation encoded in the transcriptome, which is shared across cell types and species for the models to generalize.

Signaling response signatures reflect transcriptome-wide features

The generalizability of IRIS piqued our interest in understanding what constitutes these response signatures associated with each signaling pathway and why they are robust across different cell types and species. First, we asked whether the response signatures rely on a small number of genes and to what extent the previously annotated response genes contribute to the signature **(Fig. S1a)**. To quantify the importance of individual genes in an unbiased way, we performed a gene ablation test (**Figs. 3a, S7**). We first ranked genes either based on maximal dispersion or mutual information, and then randomized the expression level of each gene within the entire hESC screen, which was used as test data, to ask how the randomization of each gene impacts the prediction accuracy. We observed that thousands of genes need to be randomized before the predictive power falls to random chance. The canonical response genes are ranked highly by the mutual information but only contribute a small fraction of the model's predictive power. Although indirect, these results suggest the majority of the transcriptome contributes to the response signatures, which likely reflects the interconnectedness of the underlying gene regulatory network.

As an orthogonal way of interpreting response signatures learnt by IRIS, we combined saliency maps with gene set enrichment analysis (GSEA) [29]. We first asked whether IRIS just re-learnt gene sets associated with each signaling pathway. For each signaling pathway, to rank the genes based on their contribution to the response signature learnt by IRIS, we computed the gradients of the output with respect to each gene at different points of training (**Fig. 3b**). Inputs associated with larger gradients are likely to have a bigger contribution to the prediction. We then performed GSEA on the weights associated with each predictor and found the classic signaling pathway response gene sets are enriched or nearly enriched in their respective predictors, suggesting IRIS draws from genes classically annotated to be part of these response modules (**Fig. 3b**). However, we observed that the response signatures learnt by IRIS are broadly distributed across a variety of gene sets during the supervised learning steps, and the gene sets annotated to be associated with each pathway are not preferentially enriched (**Fig. 3c, d**). Therefore, IRIS learns response signatures across a large collection of annotated gene sets, including but not limited to those associated with signaling response.

IRIS reveals combinatorial signaling code usage in developing mouse embryos

The ultimate utility of IRIS is to accurately infer signaling states of cells in a tissue or organism and reconstruct the signaling histories that drive the fate decisions. Mouse embryos are ideal for this purpose due to the available scRNAseq atlas across developmental timepoints, and importantly, the feasibility of performing experimental validation. Although we had two perturbation datasets representing mouse and human respectively, it was unclear whether including the human dataset in training would be advantageous for inference i n mouse embryos. We performed two tests to evaluate this question and found that including the hESC screen further improved the accuracy of IRIS in both tests **(Fig. S8)**. The improvement by adding

cells that are more distant from the test cells, compared to the cells that are already in the training data, further suggests that IRIS learns response signatures that are generalizable across cell types/states.

Having trained IRIS on both screens, we inferred the activation states of all five signaling pathways and the 32 possible combinations amon g ~40 annotated cell-type clusters from gastrulating mouse embryos (E6.5-E8.5) (**Figs. 4a, b, S9)** [33]. This analysis revealed highly dynamic switches of the combinatorial signaling codes over this critical period of development. At the global level, there is a distinct switch around E7.0 when many of the combinations associated with earlier development disappear and are replaced by a different set of combinations that are present through E8.5. We observed that TGF-β is associated with combinations present in early time points while RA is associated with later time points, suggesting that RA and TGF-β convey temporal information across cell types. In contrast, WNT, FGF and BMP are present across time points and annotated clusters, but only in a fraction of cells in those clusters, suggesting these pathways are associated with specifying heterogeneity of cell states within a cell type cluster.

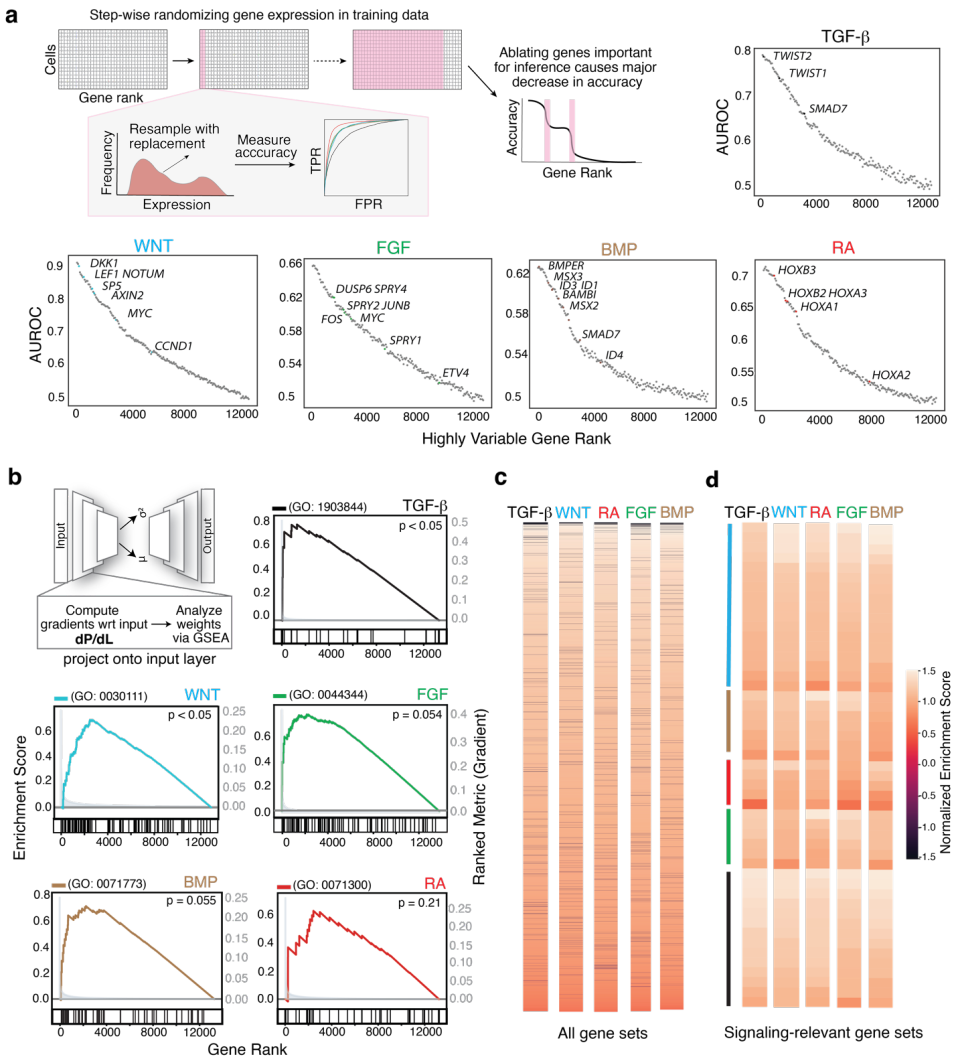

Figure 3. Signaling response signatures reflect transcriptome-wide features. a, Gene ablation test. Genes were ranked according to their dispersions and the expression level of each gene was progressively corrupted through randomization, which was performed by resampling with replacement. The inference accuracy of IRIS was computed after each round of corruption. Many genes seem to contribute to the inference accuracy for each pathway, including but not limited to the classic response genes as highlighted in each plot. Models with above average performance were selected. **b**, Gene saliency test. To estimate the contribution of each gene to the model performance, gradients of the loss with respect to the input variables (dP/dL) were computed for each pathway during training. Genes were ranked based on the gradient and a GSEA (Gene Set Enrichment Analysis) was performed on each pathway by choosing the gene set annotated as "response to" the corresponding signal (GO number labeled). The p-values were computed with a permutation test (n=1000). **c**, GSEA for all gene sets present in the GO Biological Pathways 2021 database showed broad enrichment of many gene sets for each pathway. NES, normalized enrichment scores. **d**, GSEA on signaling-relevant gene sets showed no preferential enrichment of the gene sets annotated to be related to the pathway of interest. Color lines denote the gene sets assigned to the corresponding signaling pathway.

Furthermore, within a single annotated cell cluster, IRIS reveals heterogeneous signaling states that reflect true cell-state diversity. For example, within the cardiomyocyte population, IRIS identified two populations that differ in their RA activation states, with the RA+ population enriched for atrial markers and RA- population enriched for ventricular markers, consistent with the role of RA in specifying atrial fate (**Fig. 4c**) [31]. Similarly, IRIS identified several major signal combinations in the "gut" cell population, among which cells annotated as FGF+WNT+BMP+ express mid-hindgut markers and cells negative for all three pathways express foregut markers (**Fig. 4d, e**). This result is consistent with the observation that high levels of FGF4, BMP2 and Wnt ligands in the posterior of E7.5-E8.5 mouse embryos promote hindgut and inhibit foregut fate [30]. Unlike other cell-cell communication inference algorithms which use pseudo-bulk analysis, IRIS does not rely on accurate cell clustering which is an ongoing challenge for interpreting scRNAseq data [6–8]. These results highlight the capability of IRIS to deal with heterogeneous cell populations and reveal meaningful biological information.

IRIS reconstructs spatiotemporal signaling dynamics of cell fate lineages

Next, we asked if IRIS can recover signaling histories along diverse cell fate lineages, represented by the sequential usage of signal combinations. The anterior foregut and cardiomyocyte lineages have relatively well annotated signaling histories and are well covered in the mouse gastrulation dataset **(Fig. 5a, b)** [32,33]. By manually selecting the relevant cell populations to construct the two lineages, we found IRIS produced much more interpretable prediction on each pathway than the response-gene method **(Figs. S10 & 2)**. Importantly, the inferred signaling histories aligned well with existing knowledge, such as the preferential activation of TGF-β along the endodermal lineage, activation of BMP along the mesodermal lineage, and inactivation of Wnt prior to the differentiation of cardiomyocytes [18,19,31]. When focusing on specific signal combinations annotated in either lineage, we found the inferred combinations lined up in temporal orders that correspond to what would be expected for each lineage **(Fig. 5c)**. For example, the signal combination associated with nascent mesoderm precedes the signal combination associated with cardiac differentiation. These results suggest that IRIS can correctly infer and order signal combinations along cell fate trajectories.

Finally, we asked if IRIS can identify signal combinations associated with spatial position, which would allow us to infer the location of the cells. Somitogenesis is a classic example in which cells' positional information is

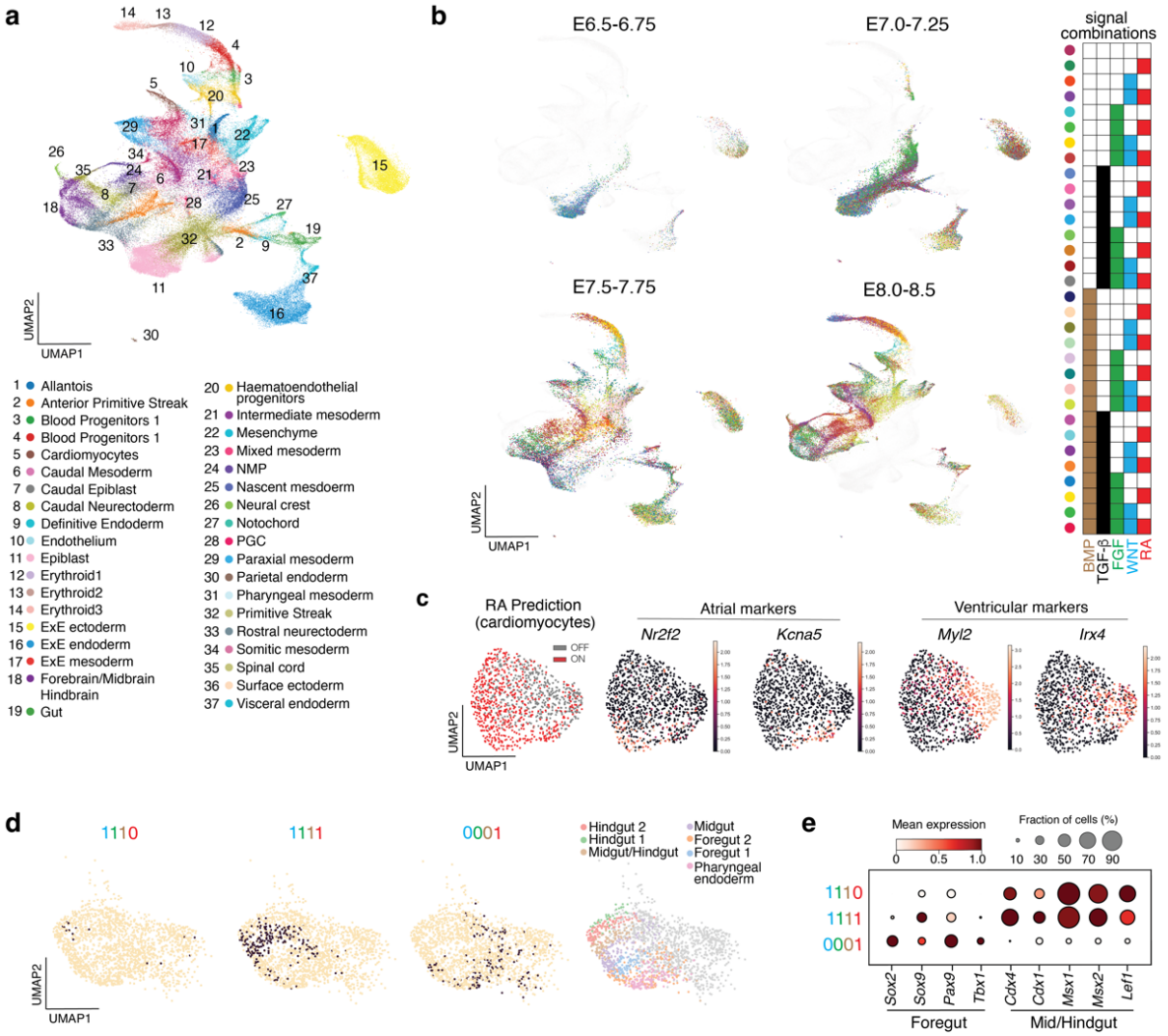

Figure 4. IRIS reveals heterogeneity of combinatorial signaling code usage in developing mouse embryos. a, Annotated cell populations in gastrulating mouse embryos (E6.5-E8.5) [33]. **b**, Dynamic switches of signal combinations across time points in gastrulating mouse embryos. Signal combinations are color coded. Cells with the corresponding signal combination are highlighted at each time point. **c**, IRIS revealed heterogenous states of RA activation within the annotated cardiomyocyte cluster. Cells predicted to be RA+ were enriched for atrial markers, and cells predicted to be RA- were enriched for ventricular markers. **d**, IRIS identified several main combinatorial signaling states that corresponded to cell fate states within the gut cluster. Cells inferred to have the denoted combinations are highlighted (*black dots*). The gut population was re-clustered and annotated by the original paper based on cell fate markers (*right panel*) [33]. **e**, Gut cells associated with different combinatorial signaling states are enriched for different cell-fate markers that correspond to the foregut or mid/hindgut.

determined by three known morphogen gradients along the anterior-posterior axis, with RA coming from the anterior and WNT/FGF from the posterior (**Fig. 5d**) [34]. We selected the population annotated as somitic mesoderm and inferred their signaling states with respect to these three pathways. We found that five combinations, out of the eight possible ones, described most of the cells (**Figs. 5e, S11a, b**). By placing cells belonging to each signal combination along the anterior-posterior axis based on the three morphogen gradients, we found the arrangement of the cells recovered the spatial patterns of gene expression (**Figs. 5f, S11c**). Specifically, markers for differentiating somites (*Mesp2, Ripply2*) were enriched anteriorly and markers for progenitor states (*Tbx6*) were enriched posteriorly [34–36]. Cells in the small RA-Wnt+FGF- population (4% of the cells) expressed either anterior or posterior markers, and we suspected that this population represented FGF or RA false negatives (**Fig. S11d**). These results together showed that IRIS is capable of both correctly ordering signal combinations in time and space in *in vivo* atlases.

IRIS helps optimize hESC differentiation protocol for organ-specific mesenchyme

Finally, we asked whether IRIS can help reveal signaling histories along less-studied cell lineages and inform directed hESC/hiPSC differentiation. Organ-specific mesenchymal cells are a family of enigmatic cell types that are essential for organ development, such as instructing the branching morphogenesis in the lung [37]. Their developmental histories are poorly understood and thus, efforts in deriving them from hESC/hiPSC to build more realistic organoids have received limited success. Taking advantage of a published dataset on mesenchymal diversity in early mouse organogenesis (E8.5-E9.5), we used IRIS to predict signaling activity during the diversification of organ-specific mesenchyme in the foregut [17,38].

We found that WNT activation distinguished the respiratory mesenchyme from other types of mesenchyme at this time point more than any other signals (**Figs. 6a, S12**). Although the role of canonical WNT signaling has been studied in the specification of the lung endoderm and differentiation of lung mesenchymal progenitors into various stromal cell types [39,40,41], its role in the initial specification of the respiratory mesenchymal identity was not reported. To test the prediction made by IRIS, we cultured the E9.0 mouse foregut *ex vivo* with either a WNT agonist or antagonist for 24 hrs. WNT activation spatially expanded the expression of key respiratory mesenchymal markers (e.g. *Tbx4* and *Foxf1*), whereas WNT inhibition completely eliminated *Tbx4* expression (**Fig. 6b**). The expansion of these markers is restricted to the anatomical region immediately adjacent to the lung, where the mesenchymal cells likely share common progenitors with the respiratory mesenchyme, suggesting WNT might regulate the choice

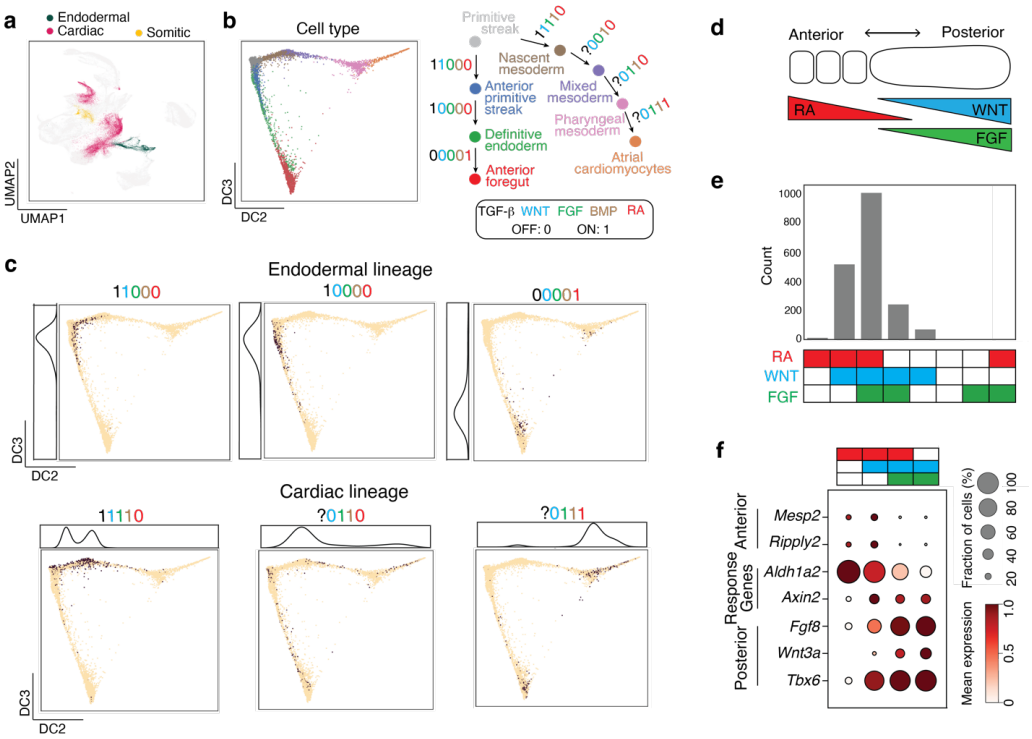

Figure 5. IRIS reconstructs spatiotemporal combinatorial signaling dynamics of cell fate lineages. a, Cell populations belonging to the three developmental lineages for which signaling histories are well characterized. Data and population annotations from scRNAseq atlas of E6.5-8.5 mouse embryos [33]. **b**, Diffusion map of cell populations associated with the endodermal and cardiac lineages (*left*), and annotated signal combinations at each step of differentiation (*right*), with the activation status of TGF-β along the cardiac lineage unknown. **c**, IRIS recovered the correct temporal order of signal combinations. Cells inferred to have the denoted combinations are highlighted on the diffusion maps (*black dots*). Kernel density estimates of signal combinations along either the endodermal or cardiac trajectories are plotted alongside the diffusion maps. **d**, Schematic representation of the three morphogen gradients that spatially regulate the differentiation of the somitic mesoderm. **e**, Counts and distribution of cells annotated with each possible signal combination with respect to the three morphogen signals. Top five combinations accounted for 99.5% of the cells. **f**, Dot plot of marker genes associated with spatial positions or signaling responses. Somitic mesoderm cells with inferred signaling states were placed into the spatial coordinates specified by the morphogen gradients, and this spatial reconstruction recovered the spatial expression patterns of cell fate markers.

between alternative organ-identity of mesenchymal progenitor cells **(Fig. S13a)**.

We sought to apply this knowledge to improve the hESC/hiPSC differentiation protocols for the respiratory mesenchyme, a step towards building lung organoids with branched structure [38]. The published protocol only applied WNT stimulation at the end of the 7-day differentiation for 24 hrs to promote trachea fate, which occurs after the commitment to respiratory identity (**Fig. 6c**). IRIS predicted that WNT was involved in the commitment to the respiratory fate, a step presumably prior to the trachea fate commitment. We decided to activate WNT earlier and for a longer duration compared to the published protocol, starting from day 4 of *in vitro* differentiation, which corresponds to the stage of splanchnic mesoderm progenitors at roughly E8.5. We found that earlier WNT activation dramatically increases *TBX4* expression in a concentration-dependent manner, and an intermediate WNT activity is optimal for preserving all key marker genes of respiratory mesenchyme, as high Wnt activity decreases *FOXF1* expression (**Fig. S13b, c**). The same effect of pro-respiratory fate by early WNT activation holds true across different hESC/hiPSC lines, suggesting that the modified protocol brings us closer to deriving respiratory mesenchyme (**Fig. S13d, e**). This shows the utility of IRIS for proposing and optimizing new directed differentiation protocols for stem cells, without having to explore the entire combinatorial space.

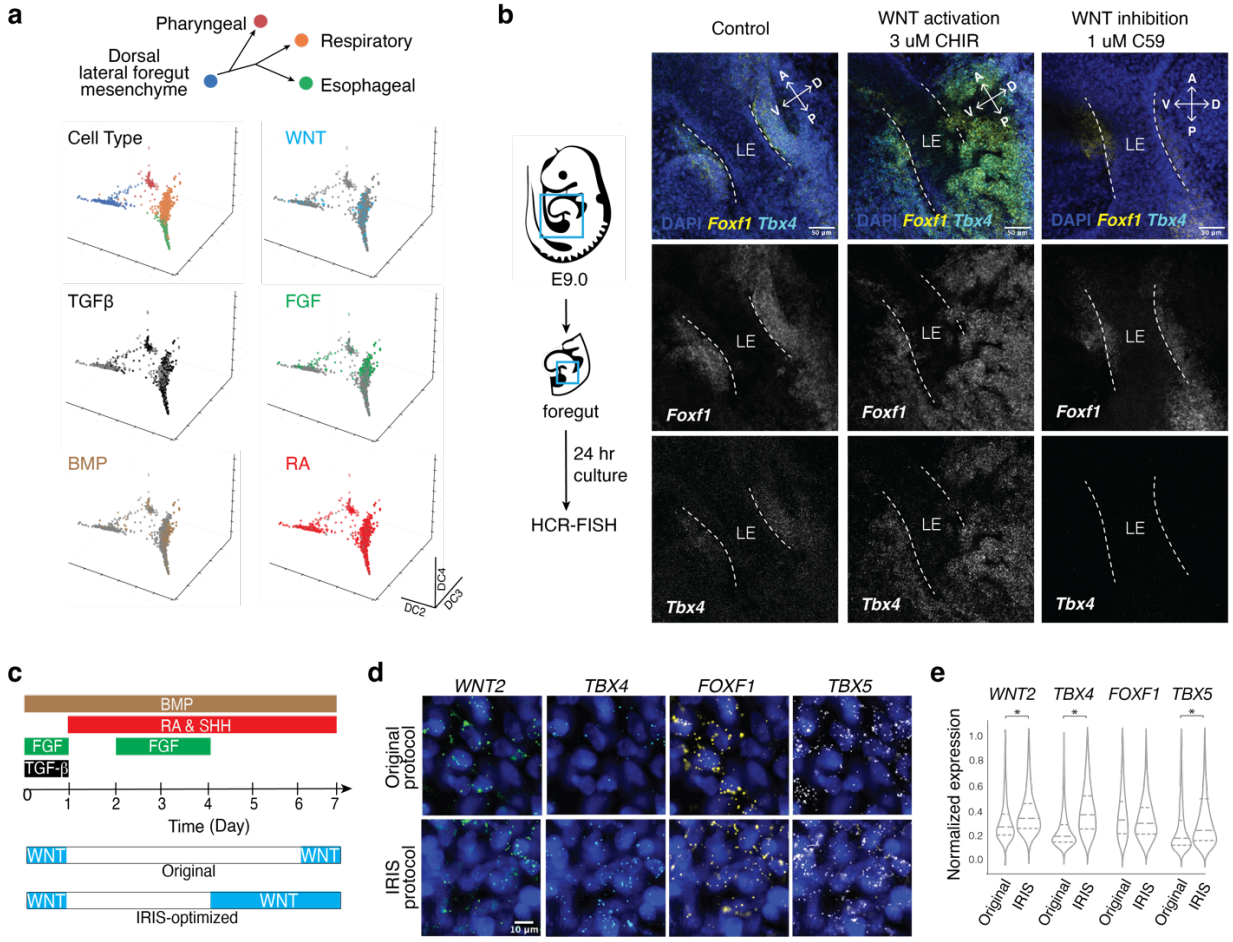

Figure 6. IRIS helps optimize hESC differentiation protocol for organ-specific mesenchyme. a, IRIS predicted that WNT activation distinguishes between the early respiratory and alternative mesenchymal fates, derived from the shared progenitor state. The respective cell populations are from E9-E9.5 mouse embryo foregut [17]. Cells inferred to have the denoted signaling pathway activated are highlighted in the corresponding color. **b**, Validating IRIS prediction in respiratory mesenchymal fate commitment using mouse foregut. Foregut tissues were dissected from E9.0 mouse embryos and cultured *ex vivo* for 24 hrs with a WNT agonist (CHIR) or inhibitor (C59). HCR-FISH for key respiratory mesenchymal markers was performed and the region of the putative lung mesenchyme was imaged. **c**, Comparison of original protocol and IRIS-optimized protocol. IRIS-optimized protocol applies an earlier and extended WNT stimulation, while keeping the other pathway conditions identical to the original protocol. **d**, HCR-FISH staining of respiratory mesenchymal marker genes in cells differentiated from hESCs with either the original protocol or IRIS-optimized protocol at Day 7. 3 μM CHIR was used in both protocols. **e**, Quantification of the HCR-FISH data shown in (**d**). IRIS significantly improved the efficiency of differentiating into respiratory mesenchyme fate. * denotes significant p value of <0.05 according to Student's T-test with the alternative hypothesis of a greater mean.

DISCUSSION

Mapping cells' signaling states and histories is crucial for understanding how cells make fate decisions, but challenging to do due to the expensive and low-throughput nature of such experiments in animal models, and lack of tools in human tissues. We demonstrated a novel experimental-computational framework that performs combinatorial signal perturbation screens on a collection of cell types *in vitro,* and uses the signal response signatures computed at the transcriptome level to infer the activity of the same signaling pathways *in vivo*. The signal response signatures serve as accurate classifiers on diverse cell types, including those not represented in the perturbation screens. This cross-cell-type, cross-batch, and cross-species signal classification is enabled by IRIS, a neural network-based transfer learning algorithm. Although our work focused on six recurring developmental pathways which determine most of cell fate decisions in early embryos, the same framework can be applied to other developmental signaling pathways, cytokines and hormones.

The observation that signaling response signatures are transferable between cell types is a surprising discovery, because transcriptional responses to signal activation are often considered unique to each cell type [42,43]. We demonstrated that the response signatures learnt from mESC and hESC screens can accurately classify signaling states of cells from mouse embryos at equivalent and later stages. To what extent the response signatures learnt by IRIS can be generalized between cells that are further apart in the transcriptomic space remains to be tested. Increasing cell type diversity in the training data may enable broader generalization to distant cell types. The model itself may be improved by integration with recent advances in transformers and foundational models that are able to learn generalizable principles on large, unlabeled datasets [44–47]. We envision a cycle of perturbation screens, model training, prediction on new data, and experimental testing, the results of which can be used subsequently to refine the model.

Furthermore, our results support the idea that signaling information can be broadly stored throughout the genome and read out through RNA. Our analysis suggests that many genes carry at least some information about signaling response and classifiers relying on many genes are more robust to cellular contexts and technical noise associated with scRNAseq data. However, we observed noticeable differences in the generalizability of the response signatures amongst the signaling pathways studied. Specifically, response signatures to WNT and RA seem to transfer well across cell types whereas the response signature to FGF seems more challenging to transfer. This may suggest that the cell-type nonspecific response for some pathways might be more confounded by cell-type specific responses, making it more difficult to learn. Additionally, our models were fit to data collected after 24 hours of signal stimulation, which might introduce an artificial time delay to the predicted signal activation/inhibition. How the response signatures evolve over time within developmentally-relevant periods remains to be rigorously characterized.

Overall, IRIS is a tool that enables learning signaling states and histories from scRNAseq data. Reconstruction of signaling histories by IRIS relies on accurate lineage inference, which is an active area of algorithm development. Nevertheless, IRIS can act on imperfect cell clusters that include heterogeneous cell types/states, due to its complete reliance on cell-autonomous information at the single-cell level. As a result, IRIS' response-centric approach is complementary to other signal inference methods relying on ligand-receptor pairing, such as CellPhoneDB, NicheNet and CellChat [6–8]. These two types of approaches can be used jointly to infer potential cell-cell communication, especially in the case of analyzing spatial transcriptomics data. This will expand opportunities in using scRNAseq and spatial transcriptomics for scientific discoveries and translational applications, such as engineering stem cells and identifying signaling pathways that drive pathological states of the cells.

Acknowledgements: We would like to thank Qiu Chang Wu, Douglas Lauffenburger, Na Sun, Alexandra Cao, Lauren Tso, Mark Greenwood, and the rest of the Li Lab for helpful discussions. We would also like to thank Jennifer Love, Sumeet Gupta, and Stephan Mraz from the Whitehead Institute Genome Technology core and George Bell from the Bioinformatics and Research Computing core for their advice and assistance throughout the project. N.H. is supported by the National Science Foundation Graduate Research Fellowship program. This work was supported by National Institute of Health grants DP2HD108777 (P.L.), Allen Distinguished Investigator Award, and a Paul G. Allen Frontiers Group advised grant of the Paul G. Allen Family Foundation (P.L.).

Competing interests: Authors declare that they have no competing interests.

Author contributions: N.H. and P.L. conceived the study. D.F. and P.L. supervised the work. N.H. led the computational and experimental work. C.L. developed the code base. M. Meziane performed experimental validations in mice. M. Mitalipova assisted with stem cell validation experiments.

REFERENCES

1. Maretto, S. *et al.* Mapping Wnt/beta-catenin signaling during mouse development and in colorectal tumors. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 3299–3304 (2003).
2. Shibata, M. *et al.* Regulation of prefrontal patterning and connectivity by retinoic acid. *Nature* **598**, 483–488 (2021).
3. Frieda, K. L. *et al.* Synthetic recording and in situ readout

of lineage information in single cells. *Nature* **541**, 107–111 (2017).
4. Chen, W. *et al.* Symbolic recording of signalling and cis-regulatory element activity to DNA. *Nature* **632**, 1073–1081 (2024).
5. McNamara, H. M., Solley, S. C., Adamson, B., Chan, M. M. & Toettcher, J. E. Recording morphogen signals reveals mechanisms underlying gastruloid symmetry breaking. *Nat Cell Biol* **26**, 1832–1844 (2024).
6. Jin, S., Plikus, M. V. & Nie, Q. CellChat for systematic analysis of cell-cell communication from single-cell transcriptomics. *Nat. Protoc.* (2024) doi:10.1038/s41596-024-01045-4.
7. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2020).
8. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).
9. Armingol, E., Baghdassarian, H. M. & Lewis, N. E. The diversification of methods for studying cell-cell interactions and communication. *Nat Rev Genet* **25**, 381–400 (2024).
10. Klumpe, H. E. *et al.* The context-dependent, combinatorial logic of BMP signaling. *Cell Syst.* **13**, 388–407.e10 (2022).
11. Antebi, Y. E. *et al.* Combinatorial signal perception in the BMP pathway. *Cell* **170**, 1184–1196.e24 (2017).
12. Müller, P. & Schier, A. F. Extracellular movement of signaling molecules. *Dev. Cell* **21**, 145–158 (2011).
13. Miller-Jensen, K., Janes, K. A., Brugge, J. S. & Lauffenburger, D. A. Common effector processing mediates cell-specific responses to stimuli. *Nature* **448**, 604–608 (2007).
14. Cui, A. *et al.* Dictionary of immune responses to cytokines at single-cell resolution. *Nature* **625**, 377–384 (2024).
15. Miyazono, K., Kamiya, Y. & Morikawa, M. Bone morphogenetic protein receptors and signal transduction. *J. Biochem.* **147**, 35–51 (2010).
16. MacDonald, B. T., Tamai, K. & He, X. Wnt/beta-catenin signaling: components, mechanisms, and diseases. *Dev. Cell* **17**, 9–26 (2009).
17. Han, L. *et al.* Single cell transcriptomics identifies a signaling network coordinating endoderm and mesoderm diversification during foregut organogenesis. *Nat. Commun.* **11**, 4158 (2020).
18. Loh, K. M. *et al.* Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations. *Cell Stem Cell* **14**, 237–252 (2014).
19. Loh, K. M. *et al.* Mapping the pairwise choices leading from pluripotency to human bone, heart, and other mesoderm cell types. *Cell* **166**, 451–467 (2016).
20. Wandzioch, E. & Zaret, K. S. Dynamic signaling network for the specification of embryonic pancreas and liver progenitors. *Science* **324**, 1707–1710 (2009).
21. Yeo, G. H. T. *et al.* A multiplexed barcodelet single-cell RNA-seq approach elucidates combinatorial signaling pathways that drive ESC differentiation. *Cell Stem Cell* **26**, 938–950.e6 (2020).
22. Linderman, G. C. *et al.* Zero-preserving imputation of single-cell RNA-seq data. *Nat. Commun.* **13**, 192 (2022).
23. Rood, J. E., Hupalowska, A. & Regev, A. Toward a foundation model of causal cell and tissue biology with a Perturbation Cell and Tissue Atlas. *Cell* **187**, 4520–4545 (2024).
24. Iscen, A., Tolias, G., Avrithis, Y. & Chum, O. Label Propagation for Deep Semi-supervised Learning. (2019).
25. Xu, C. *et al.* Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
26. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022).
27. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
28. McGinnis, C. S. *et al.* MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).
29. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
30. Zorn, A. M. & Wells, J. M. Vertebrate endoderm development and organ formation. *Annu. Rev. Cell Dev. Biol.* **25**, 221–251 (2009).
31. Lee, J. H., Protze, S. I., Laksman, Z., Backx, P. H. & Keller, G. M. Human pluripotent stem cell-derived atrial and ventricular cardiomyocytes develop from distinct mesoderm populations. *Cell Stem Cell* **21**, 179–194.e4 (2017).
32. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
33. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
34. Aulehla, A. & Pourquié, O. Signaling gradients during paraxial mesoderm development. *Cold Spring Harb. Perspect. Biol.* **2**, a000869 (2010).
35. Oginuma, M., Niwa, Y., Chapman, D. L. & Saga, Y. Mesp2 and Tbx6 cooperatively create periodic patterns coupled with the clock machinery during mouse somitogenesis. *Development* **135**, 2555–2562 (2008).
36. Biris, K. K., Dunty, W. C., Jr & Yamaguchi, T. P. Mouse Ripply2 is downstream of Wnt3a and is dynamically expressed during somitogenesis. *Dev. Dyn.* **236**, 3167–3172 (2007).
37. Morrisey, E. E. & Hogan, B. L. M. Preparing for the first breath: genetic and cellular mechanisms in lung development. *Dev. Cell* **18**, 8–23 (2010).
38. Kishimoto, K. *et al.* Directed differentiation of human pluripotent stem cells into diverse organ-specific mesenchyme of the digestive and respiratory systems. *Nat. Protoc.* **17**, 2699–2719 (2022).
39. Kishimoto, K. *et al.* Bidirectional Wnt signaling between endoderm and mesoderm confers tracheal identity in mouse and human cells. *Nat. Commun.* **11**, 4159 (2020).
40. De Langhe, S. P. *et al.* Formation and differentiation of

multiple mesenchymal lineages during lung development is regulated by beta-catenin signaling. *PLoS One* **3**, e1516 (2008).
41. Jiang, M. *et al.* Gpr177 regulates pulmonary vasculature development. *Development* **140**, 3589–3594 (2013).
42. Hoppler, S. & Kavanagh, C. L. Wnt signalling: variety at the core. *J. Cell Sci.* **120**, 385–393 (2007).
43. Mullen, A. C. *et al.* Master transcription factors determine cell-type-specific responses to TGF-β signaling. *Cell* **147**, 565–576 (2011).
44. Yang, F. *et al.* scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* **4**, 852–866 (2022).
45. Heimberg, G. *et al.* A cell atlas foundation model for scalable search of similar human cells. *Nature* 1–3 (2024).
46. Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
47. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
48. Choi, H. M. T. *et al.* Mapping a multiplexed zoo of mRNA expression. *Development* **143**, 3632–3637 (2016).
49. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).

Materials and Methods

Single-cell Data Processing

Unless otherwise stated, single-cell data was processed using the Scanpy package (version 1.10.0). For visualization, data was log normalized, scaled according to z-scores, and nearest neighbor graphs were constructed from a PCA computed from these scaled expression values. These nearest neighbor graphs were the basis for the UMAP projections. For developmental lineage visualization, scanpy.tl.diffmap function was used to visualize developmental trajectories. This function takes data, $X \in R^{n \times d}$, and constructs a pairwise similarity matrix between observations using a Gaussian kernel, $S_{ij} = exp\ exp\ (-\frac{\left\|x_i - x_j\right\|^2}{2\epsilon})$. Epsilon is used as the bandwidth parameter. This similarity matrix is then converted to a Markov transition matrix using the partition function $P_{ij} = \frac{S_{ij}}{\sum_k S_{ik}}$. The eigenvectors are then computed from this transition state matrix and used as the diffusion components. Optimal diffusion components were those which, when used in pairs, correctly placed the lineages in order based on known cell type markers and annotations. Human and mouse gene names were converted between each other using mousipy (version 0.1.5). For consistency in the analysis, when mouse and human data was integrated, human gene names were used.

Signaling state inference with the response-gene method

Response gene scores were computed by summing the log normalized gene expression values for the response gene sets for each pathway. For a set of j response genes (G) and expression r.

$$S(i) = \sum_{i \in G} r(j)$$

The performance of the inference method was evaluated using the receiver operator characteristic (ROC), precision recall curve (PRC), and F1 score. The ROC was computed and plotted using sklearn.metrics.roc_curve. Specifically, for each signaling pathway, a variety of arbitrary thresholds for the response gene scores were set to classify the pathway activity. Each cell in the test group was classified to be ON or OFF for the pathway activity based on the corresponding response score. By comparing with the ground-truth label, the true positive rate (TPR) and false positive rate (FPR) were calculated for each threshold and plotted against each other to generate the ROC:

$$TPR = \frac{TP}{TP + FP}$$

$$FPR = \frac{FP}{TP + FP}$$

where TP is true positive and FP is false positive. The area under the receiving operating curve (AUROC) was computed using sklearn.metrics.roc_auc_score. The PRC was computed and plotted similarly using sklearn.metrics.prc_curve by setting a variety of arbitrary thresholds for response gene scores. The precision (same as TPR) was plotted against recall at each threshold.

$$Recall = \frac{TP}{TP + FN}$$

where FN is false negative. The area under the precision recall curve (AUPRC) was computed using sklearn.metrics.auc, which computes the integral over these two scores using the trapezoid rule. F1 score was evaluated using sklearn's metrics functions (version 1.4.1.post1):

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Calculating F1 score requires a single threshold of response gene scores for each signaling pathway, which was computed by finding the threshold which maximally separates true positive (TP) and false positive (FP) according to the ROC curve.

IRIS model and hyperparameter selection

IRIS is a semi-supervised model that was implemented using the scANVI architecture[25] from the scvi-tools package (version 0.20.3), and trained on an NVIDIA-A6000 GPU using the Cuda library (version 11.3). The model architecture is a variant of the conditional variational autoencoder (CVAE) model. It includes an encoder neural network which learns the mapping of the input data X and the conditions S to an n-dimensional latent space Z. A decoder neural network takes as input the conditions S, and latent space encoding in Z. The CVAE model minimizes the evidence lower bound: $L_{CVAE} = E_{q_\phi(Z|X,S)}[\ log\ log\ p_\theta(Z,S)\ - \alpha D_{KL}(|p_\theta(S))]$, where $q_\phi(X,S)$, represents the variational posterior, $p_\theta(Z,S)$ is the reconstructed distribution from the latent space, is a hyperparameter which determines the strength of the Kullback-Leibler regularization term (D_{KL}). For the purposes of this analysis, gene expression counts were assumed to follow a zero-inflated negative binomial distribution (ZINB). The default value of 0.01 was used. Simultaneously a feedforward neural network takes as input Z and S for each data point and minimizes the cross-entropy (CE) loss $= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} Y_{ik}\ log\ log\ P_{ik}$ with respect to the target labels Y. Here, k are the individual classes, K is the total number of classes, and P is the probability of being assigned either class label. Each node in the neural network uses a rectified linear activation function $ReLU = max\ (0, x)$. The two neural networks are trained jointly to minimize the loss $L = L_{CVAE} + CE$.

Data was stored in the anndata format with observations as individual cells and variables as gene names. $X \in R^{n \times d}$ is the input scRNAseq with raw counts as the values. $S \in R^{n \times c}$ are the conditions from the covariates c (e.g. batch, cell type, and species), with a one-hot encoding. The target labels $Y \in R^{1 \times d}$ is presented as a vector in the anndata object with binary indicator variables. X was partitioned into training and test data as specified in each trial. Models were fit by minimizing loss on training data and then evaluated on held-out test data by freezing the model weights and evaluating the prediction output. Model performance was evaluated using the ROC, PRC, and F1 scores computed on the test data (see above).

Hyperparameter selection was performed by fitting distinct architectures on the training data (Layers=1, 2, 3; Hidden nodes=32, 64, 128, 256, 1024; Latent Variables=10, 20, 30, 50, 60). Their performance was evaluated using AUPRC. The best performing architectures were then used in subsequent analysis and evaluated via k-fold cross-validation over the batches. Models were trained for 200 epochs because this was found to provide optimal performance across a range of trials.

Setup of the model tests

In-sample tests were performed on random partitions of mESC screen data or the combined mESC and hESC screen data. The data was partitioned using the np.random.choice function to select 20% of the data as test data and the remaining 80% for model fitting. Model performance was evaluated with ROC, PRC, and F1. Out-of-sample test was performed by fitting the model with the complete mESC screen data and evaluating on the hESC screen data. The cross-validation tests were performed by splitting the mESC screen data into 3 technical batches and treating the hESC screen data as an additional technical batch. In the first test that performs the hyperparameter screen, the model was fit to two of the three mESC batches at a time and tested on the third mESC batch. The test was done iteratively by holding out each batch and averaging the performance metrics across all three held-out batches (Figs. S5, 8). The second cross-validation test was performed by including the hESC screen data and iterating through all four possible held-out batches (Fig. S9).

Comparison with classic machine-learning models

Classic machine learning models such as support vector machines (SVM), random forests (RF) and elastic net (EN) were implemented using the sklearn (version 1.4.1.post1) library and run on the same machine. Hyperparameter selection was performed over multiple orders of magnitude for each of the tunable hyperparameters (Fig. S8).

The support vector machine minimizes the following objective:

$$L(w,b) = \frac{C}{2}\|w\|_2^2 + \frac{1}{n}1^T\big(0, Y(Xw + b_1)\big)$$

where w is the vector of weights, $w \in R^n$. N is the number of data points, $1 \in R^d$ is a vector of ones, $Y \in R^n$is a vector of the binary labels, $b \in R$ is the bias term. The scaling factor, C, was chosen as stated in each analysis.

The random forest minimizes the following objective:

$$L_{classification} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} Y_{ik}\ log\ log\ P_{ik}$$

where K is the number of classes (in this case for binary classification, 1), n is the number of data points, and P_{ik} is the log probability of the correct classification. The classification loss minimizes the cross-entropy loss.

The elastic net model minimizes the following objective:

$$L(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda(\alpha\|\beta\|_1 + \frac{1-\alpha}{2}\|\beta\|_2^2)$$

where $\beta \in R^{1\times d}$ and are the weights of the model. N is the number of data points $\|\cdot\|_2^2$ is the euclidean norm and the $\|\cdot\|_1$ is the regularization penalty which is set to values as specified in each trial. α is the weighting parameter for L1 and L2 penalties. For all trials, it was set to 0.5.

Computing correlation matrix of transcriptional states Hierarchical clustering

Correlation between cells from two signaling conditions was computed as the averaged gene expression for each gene across both conditions using Pearson's correlation:

$$R = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

where x_i and y_i are individual genes in the two conditions x and y respectively. $\bar{y}$ and $\bar{x}$ are the average expression over all genes in the respective conditions. Correlation matrices were then generated with scanpy.pl.correlation_matrix. The distance metric used was $1 - R$.

All signaling conditions were hierarchically clustered on the dendrogram using Ward's method. In Ward's method, two groups are clustered if they result in the smallest increase in total intra-cluster variance, computed as the error sum of squares:

$$ESS = \sum_{x\in C_i \cup C_j}\|x - \mu_{C_i \cup C_j}\|^2 - \left(\sum_{x\in C_i}\|x - \mu_{C_i}\|^2 + \sum_{x\in C_j}\|x - \mu_{C_j}\|^2\right)$$

where C_i and C_j represent clusters that are to be merged, and μ represents the centroid of a given cluster. This process was continued iteratively until all clusters were merged, representing the root of the dendrogram.

Inferring signaling histories along developmental lineages

IRIS was used to infer signaling histories in the mouse gastrulation dataset and the foregut mesenchymal dataset [17, 33]. Cells were first clustered using Leiden clustering in scanpy with default parameters. Cell type annotations were determined using canonical marker genes and annotations provided by each of the datasets. The diffusion components were computed using sc.tl.diffmap and visually inspected for those that placed the cell types and marker gene expression in the expected order consistent with literature. For the cardiac lineage analysis, this included the primitive streak, nascent mesoderm, mixed mesoderm, pharyngeal mesoderm and cardiomyocytes. The endodermal lineage analysis included the primitive streak, anterior primitive streak, definitive endoderm, gut and visceral endoderm. A single diffusion map was made for the endoderm and cardiac lineages. Endodermal cells were sub clustered and visualized with a UMAP. The presomitic mesoderm analysis included only cells from the somitic mesoderm cluster and visualized with a UMAP. The respiratory mesenchyme lineage analysis included only cells from

the dorsal lateral foregut, respiratory mesenchyme, esophageal mesenchyme and pharyngeal mesenchyme, which were projected on a single diffusion map.

To infer the signaling histories, two methods were applied. For the response-gene method, response-gene scores were overlaid on diffusion maps to serve as a relative estimate of the pathway activity. Thresholds were chosen to maximize the separation between true and false positives. For the IRIS method, the model was fit to the raw count matrix using optimized hyperparameters for each pathway. Signaling pathway activities were predicted independently for each pathway and the combinatorial signaling state of a cell was composed by concatenating all of the pathway predictions.

To temporally align inferred combinatorial signaling states with the developmental trajectories, kernel density estimates (kde) were calculated using seaborn.kdeplot (version 0.13.2), a $\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h})$, using a Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$. $\hat{f}(x)$ is the estimated kernel density function, n is the number of samples and h is the selected bandwidth parameter. The calculation uses a bandwidth of 0.75 for the diffusion component and subset of cells that were annotated with the corresponding lineage.

Feature importance analysis

Feature importance was determined by the reduction of inference accuracy when the expression level of individual genes was sequentially resampled with replacement in the test data. Resampling with replacement was performed using the numpy.choices function on a vector containing all of the original gene expression values. Genes were ranked from high to low either by their dispersion or by mutual information. Dispersion was computed using scanpy.pp.highly_variable_genes: $D = \frac{\sigma^2}{\mu}$, where σ^2 is the variance of the expression for a given gene and is the averaged expression for a given gene. Mutual information was calculated for a given gene X and a given signaling state vector Y, which is a vector of 0 and 1 binary values representing signaling stimulation. The entropy (H) and mutual information (I) were approximated using scikit-learn.mutual_info_classif which implements a k-nearest neighbor estimation of entropy between continuous variables and categorical variables as follows: $I(X;Y) = H(X) + H(Y) - H(X|Y)$.

Gradients-GSEA Analysis

Gradients were computed natively by Pytorch (version 2.2.2+cu121). Neural networks were trained via backpropagation and the gradient with respect to the input layer was computed as:

$$\frac{\partial L}{\partial \hat{X}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \hat{X}}$$

where L is the loss function, y is the target vector of signaling labels, and X is the input data. Genes were ranked according to the magnitude of the gradient. These rankings were used to perform gene set enrichment analysis using GSEApy (version 1.1.3). The normalized enrichment score S is computed as the maximum value of a running sum according to the following equation:

$$S(i) = \sum_{j=1}^{i} \left(\frac{\delta(j) - |r(j)|^p}{\sum_{j \in G} |r(j)|^p} - \frac{1 - \delta(j)}{N - |G|}\right)$$

where the score S is computed for a given gene set of interest G, N is the number of genes in the ranked list and $|G|$ is the number of genes in the gene set of interest. The weight parameter p indicates how heavily to weigh each gene's ranking. $\delta(j)$ is the indicator function for whether or not a gene j is in the gene set G. $r(j)$ is the ranking statistic for a given gene. The normalized enrichment score is then computed as $NES = \frac{(S(i))}{max\,(S_{null}(i))}$. The p value was computed by a permutation test with 1000 samples. The gene set of interest was considered to be enriched if $p < 0.05$ and nearly enriched if $0.05 < p < 0.06$.

Stem cell maintenance and directed differentiation

H1 Human embryonic stem cells (hESCs) (WiCells WA01) or induced pluripotent stem cells (iPSCs) (WIBR iPSC66) were maintained in mTESR plus (STEMCELL Technologies 100-0276) and StemFlex media (Gibco A3349401), passaged every four days at 1:20 using ReLesR (STEMCELL Technologies 100-0483) and checked for the undifferentiated morphology. To differentiate hESCs or iPSCs, cells were dissociated as single cells with accutase (STEMCELL Technologies 07920), counted and plated at a density of 10^5 cells/mL on matrigel-coated plates (Corning 354277) with 10 μM Rock Inhibitor Y-27632 (STEMCELL Technologies 72308). All differentiation was performed in a basal media that includes Advanced DMEM/F12 (Gibco 12634010), N2 Supplement (Gibco 17502048), B27 Supplement (Gibco 17504044), and Penicillin-Streptomycin-Glutamine (Gibco 10378016), with the addition of signal agonists and antagonists specified below. The initial differentiation of hESCs or iPSCs into middle primitive streak (mPS) state (Day 1) and subsequently various lateral plate mesoderm (LPM) states (Day 2) follows published protocols [17,38]. Specifically, hESCs or iPSCs were differentiated into mPS with 6 μM CHIR99021 (Fisher Scientific 44-231-0), 40 ng/mL BMP4 (R&D Systems 314-BP-050/CF), 30 ng/mL Activin A (R&D Systems 338-AC-010/CF), 20 ng/mL FGF2 (R&D Systems 3718-FB-010), and 100 nM PIK90 (EMD Millipore 528117). mPS was induced into cardiac LPM with 30 ng/mL BMP4 (R&D Systems 314-BP-050/CF), 1 μM A83-01 (Fisher Scientific 29-391-0), and 1 μM WNT-C59 (Fisher Scientific 51-481-0). mPS was induced into anterior foregut (aFG)-LPM with 30 ng/mL BMP4, 1 μM A83-01, 1 μM WNT-C59, 2 μM RA (Fisher Scientific AC207341000), and 1 μM PMA (Fisher Scientific 50-197-5893). mPS was induced into posterior foregut (pFG)-LPM with 30 ng/mL BMP4, 1 μM A83-01, 1 μM WNT-C59, and 2 μM RA. LPM cells were further differentiated into their respective splanchnic mesoderm (SM) fate by cardiac SM induction media (30 ng/mL BMP4, 1 μM A83-01, 1 μM WNT-C59, 20 ng/mL FGF2), aFG-SM induction media (30 ng/mL BMP4, 1 μM A83-01, 1 μM WNT-C59, 20 ng/mL FGF2, 2 μM RA, 1 μM PMA), or pFG-SM induction media (30 ng/mL BMP4, 1 μM A83-01, 1 μM WNT-C59, 20 ng/mL FGF2, 2 μM RA) for 2 days.

Sequential combinatorial signal screen with pooled scRNAseq

For the combinatorial signal screen, hESCs were plated at 5104 cells/well as single-cells into 24-well cell-culture plates. Plates were coated with matrigel (Corning 354277) at room temperature for an hour before use. After 24 hrs of recovery in StemFlex feeder free media, cells were differentiated into the mPS state with the mPS induction media for 1 day (Step 1). The mPS cells were then used for the sequential combinatorial signal screen that differentiated the cells for 3 more days (Step 2 for 1 day + Step 3 for 2 days). The sequential combinations were designed either based on published protocols (3 combinations). random combinations of the five signaling pathways (54 combinations) or bit-flip combinations of the five signaling pathways (38 combinations). Media was changed every day On day 4, cells were incubated in Rocki for 30 minutes, dissociated as single cells in accutase, and barcoded with multiSeq oligos (Millipore Sigma LMO-001). Each well in the screen received a unique barcode. Multiplexing was performed in accordance with the manufacturer's protocol [28]. Cells were resuspended in PBS + 1% BSA (Sigma Aldrich A9418), pooled and sorted for live cells using a DAPI stain. Cells were processed using a 10X Chromium flow-cell and sequenced on a Illumina Novaseq sequencer, and 19,155 cells passed quality controls.

Count matrices were generated using CellRanger (v7.1.0) and aligned to the reference human genome GRCh38. Barcodes were aligned and demultiplexed using the R library deMULTIplex (version 1.0.2) [28]. Barcodes were visually inspected using tSNEs of cells clustered in barcode space. This was performed in accordance with the criteria specified in deMULTIplex. Once a list of observed barcodes was identified, each barcode was normalized to the number of barcodes observed per cell. Each barcode then had outliers trimmed and its probability density function computed. Lastly, thresholds were identified by defining the local maxima, with the mode for negatives and maximum for positive values. Barcodes were then assigned independently for each cell. Cells with one barcode are designated as singlets, more than one as doublets, and those with none as negatives. The optimal threshold was chosen which maximized the number of singlets relative to doublets and negatives. Barcodes representing 65 unique combinations were recovered, with up to 975 cells per combination.

Directed differentiation into respiratory mesenchyme

hESCs or iPSCs were first differentiated into aFG-SM using the aforementioned media condition. Then they were induced with either IRIS-optimized media protocol (30 ng/mL BMP4, 2 μM RA, 2 μM PMA, 3 μM CHIR) for 3 days, or published protocol (30 ng/mL BMP4, 2 μM RA, 2 μM PMA) for 3 days with 2 μM CHIR added only for the last day [38] . On Day 7 of differentiation, samples were fixed in 4% PFA for 15 minutes, then processed for HCR.

Mouse foregut explant culture

To perform the signaling perturbation in explants, E9.0 mouse embryos were harvested from wild type C57BL/6J timed matings. Foreguts were isolated by dissecting the anatomical region between the pharyngeal arches and liver, excluding additional tissues including the somites. Foreguts were cultured in DMEM/F12 (Gibco 10565042) + 5% FBS (Takara 631368) + 1% Penicillin/Streptomycin/Glutamate (Gibco 10378016) supplemented with either 3 μM of CHIR or 1 μM C59. After 24 hours in culture at 37°C, samples were fixed in 1% PFA overnight at 4°C and processed for subsequent analysis.

Hybridization Chain Reaction (HCR)

HCR-FISH probes were designed in-house and tested on samples with known expression to validate probe efficiency [48]. Mouse foregut explant or differentiated hESC samples were dehydrated with 100% methanol and subsequently processed according to the HCR-FISH protocols from Molecular Instruments. In brief, samples were hybridized overnight at 37°C with 4 μM of probes and amplified overnight at room temperature with B1, B3, B4, B5 hairpin amplifiers (Molecular Instruments). Afterwards, hESC-derived samples were imaged on a Nikon TI2 widefield fluorescence microscope. Explant samples were imaged on a Zeiss LSM980 Airyscan 2 confocal microscope. Analysis of the hESC-derived cells was performed by first segmenting in cellpose using the 'cyto2' model (parameters: diameter = 35, flow-threshold = 0.7) [49]. Total nuclear fluorescence signal was then measured in each channel with these segmentation masks. Individual fluorescence measurements for each cell were grouped based on the treatment condition and a student's T-test was performed to determine statistical significance using scipy.stats.ttest_ind (version 1.12.0). This test assumes the expression level following a T-distribution and values are considered significant if $p < 0.05$. Data was visualized using seaborn.violinplot.

Code Availability

Code is available at https://github.com/Pulin-Li-Lab/IRIS-signaling-inference.

Data availability

All data analyzed in this article is available publicly through online sources. The following data are available at the Gene Expression Omnibus database with their corresponding accession numbers: raw data for the hESC screen (GSE289836), the mESC screen (GSE122009), and mouse foregut organogenesis dataset (GSE136689). The gastrulation atlas is available on ArrayExpress (E-MTAB-6967).

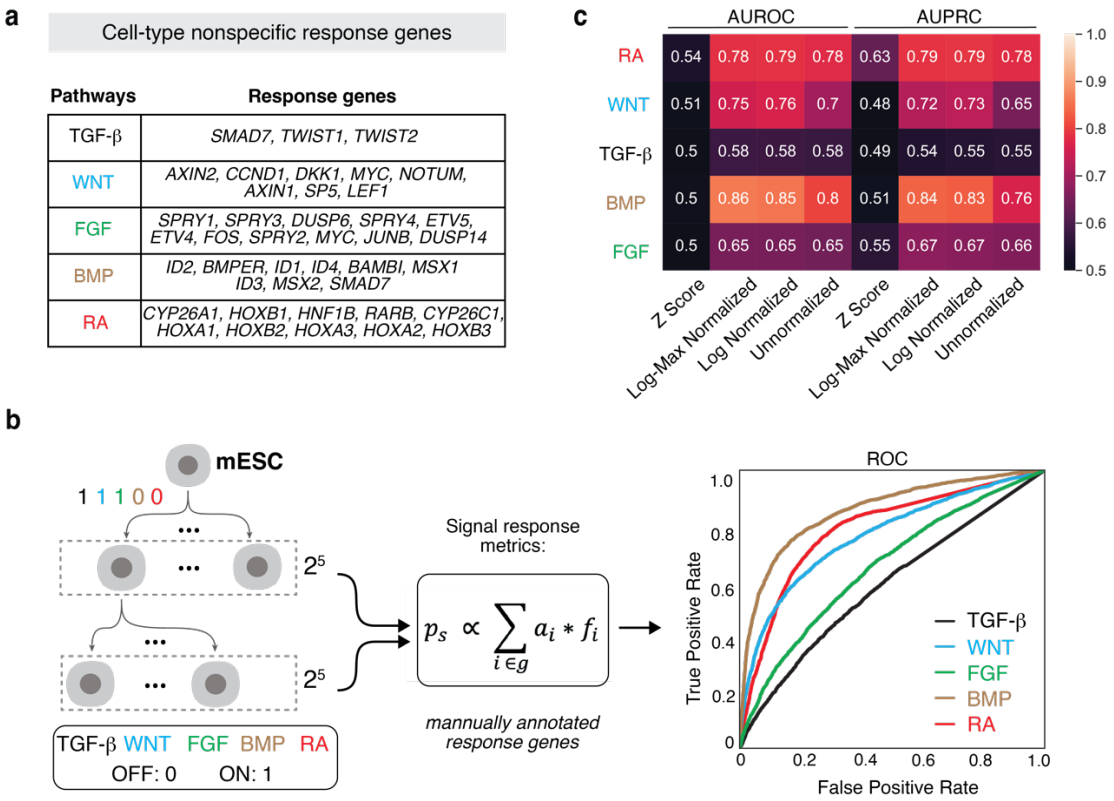

Figure S1. Signal inference using the response-gene method.
a, A list of response genes chosen for each signaling pathway was modified from Han *et al*. 2020 [17]. **b**, Response genes inferred signaling pathway activities with limited accuracy, as tested with a ground-truth dataset of mESC differentiation [21]. Signaling states are represented as binary strings. The probability of being a responding cell (p_s) was computed as summed expression of response genes (f_i) with weights (a_i). ROC, receiver operating characteristic. **c**, The performance of the response-gene method for signal inference, computed based on different normalization strategies for gene expression. AUROC, area under the receiver operating curve; AUPRC, area under the precision recall curve.

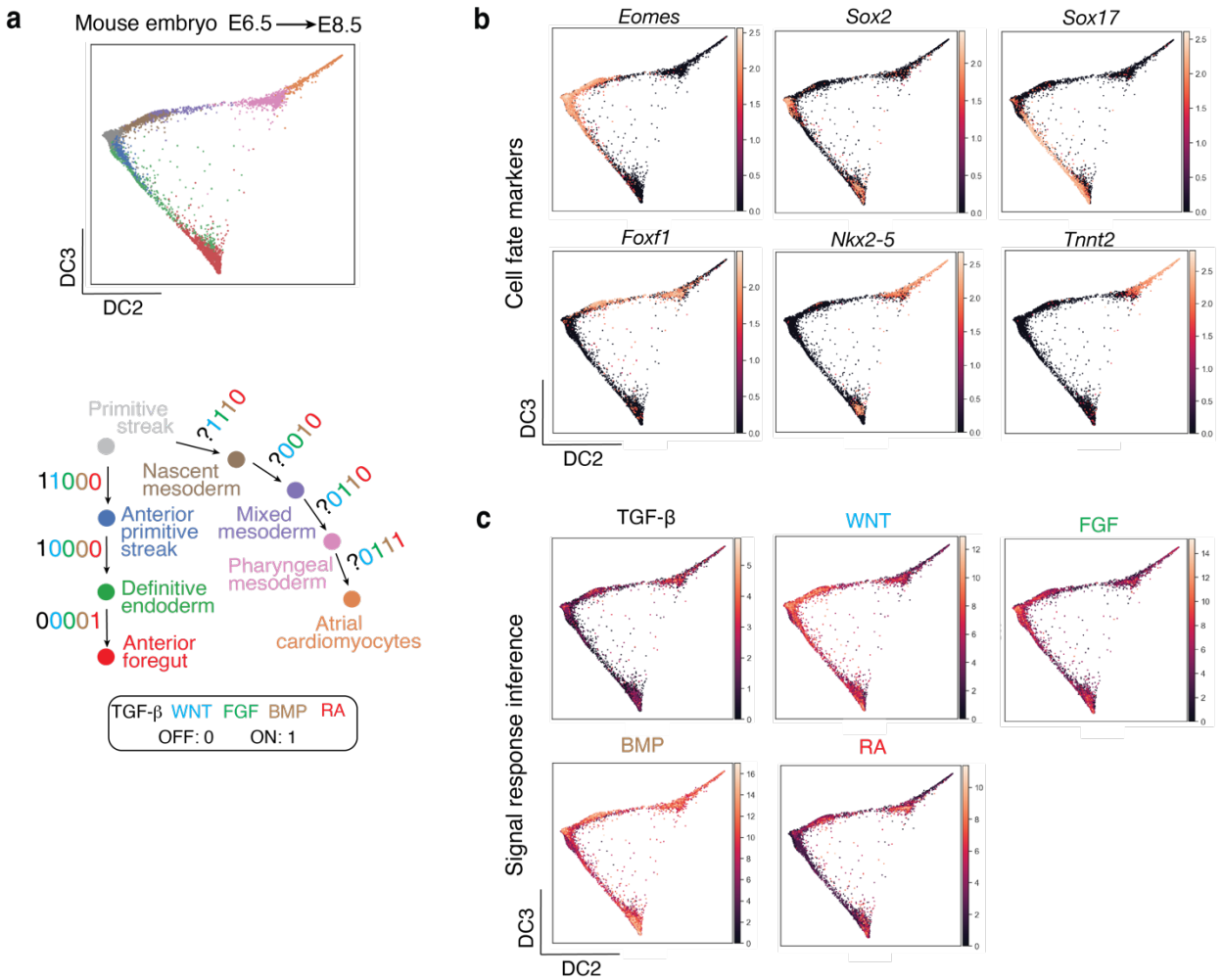

Figure S2. Signal inference of *in vivo* cell types using the response-gene method.
a, Construction of the anterior foregut and atrial cardiomyocyte trajectories, using cell populations from the mouse embryo (E6.5-E8.5) scRNAseq dataset [33]. The colors of the cell populations in the diffusion map correspond to those in the schematics. Signal combinations at each step of differentiation were annotated based on the literature. Due to lack of knowledge in the literature, the activity of TGF-β along the cardiac lineage was left out in the annotation. **b**, The expression of key cell fate markers associated with each of the two lineages (normalized to counts per 10^4). **c**, Signal response-gene scores for all five signaling pathways. Consistent thresholds for binary classification from these scores are difficult to calibrate, limiting their interpretability.

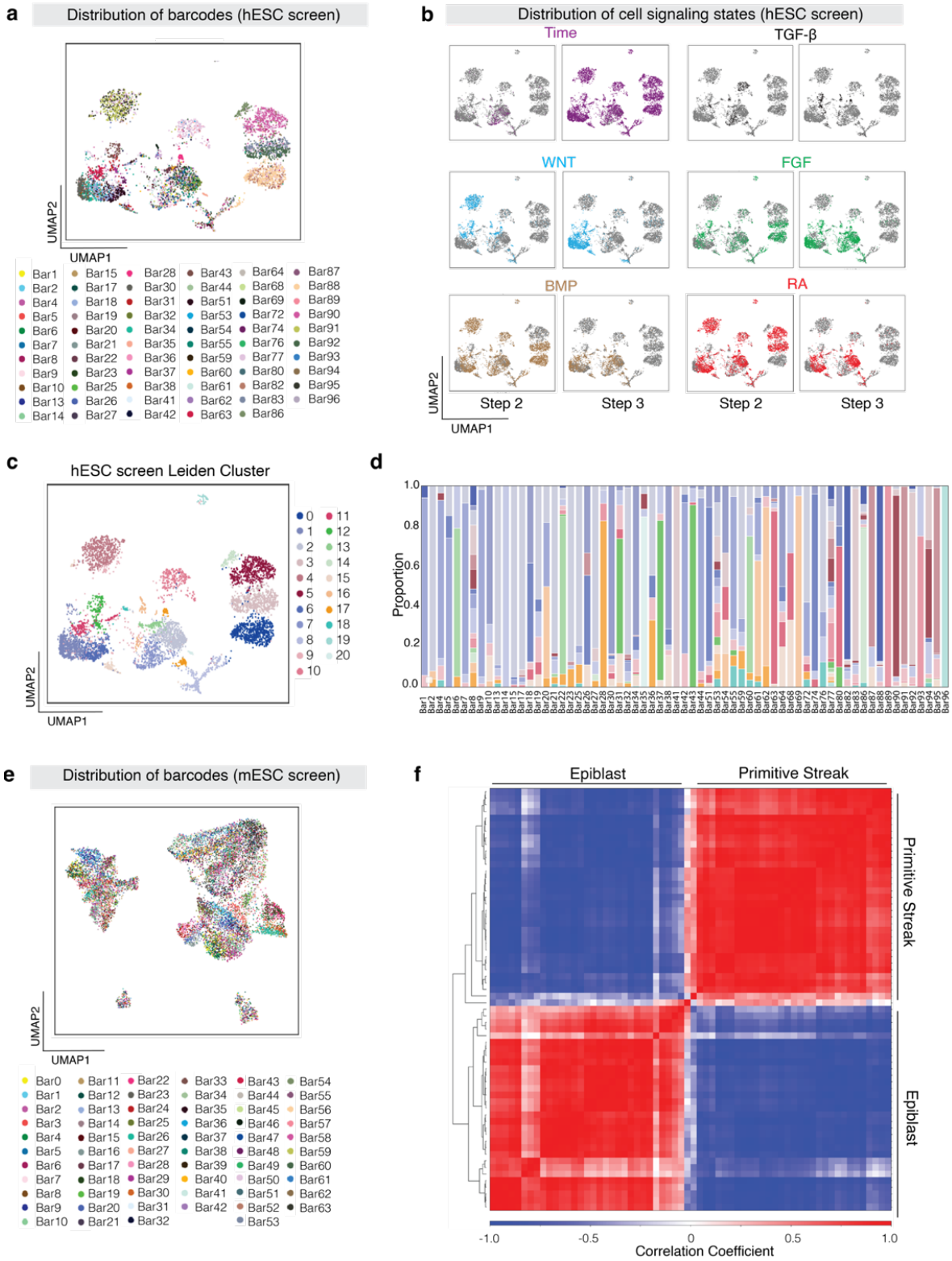

Figure S3. Analysis on cell state diversity in the mESC and hESC screens.
a, Sequential signal combinations in the hESC screen generated unique and diverse cell states, as visualized by the distribution of the barcodes on the UMAP. **b**, The sequential order of signal

activation impacts the final cell states. Cells receiving the annotated signal at the corresponding step are highlighted on the UMAP. In the "Time" plots, cells sequenced after Step 2 vs Step 3 are highlighted separately. **c**, Leiden clustering of the hESC screen data on UMAP. The number of clusters is defined using default Scanpy parameters. **d**, Cells associated with each barcode predominantly belong to a single Leiden cluster, suggesting the relative uniformity of cell states induced by each signal combination in the hESC screen. **e**, Signal combinations in the mESC screen generated diverse cell states, as visualized by the distribution of the barcodes on the UMAP. **f**, The correlation coefficient averaged across the transcriptome between distinct signal combinations shows limited diversity of final cell states in the mESC screen, compared to the hESC screen (see **Fig. 2b**).

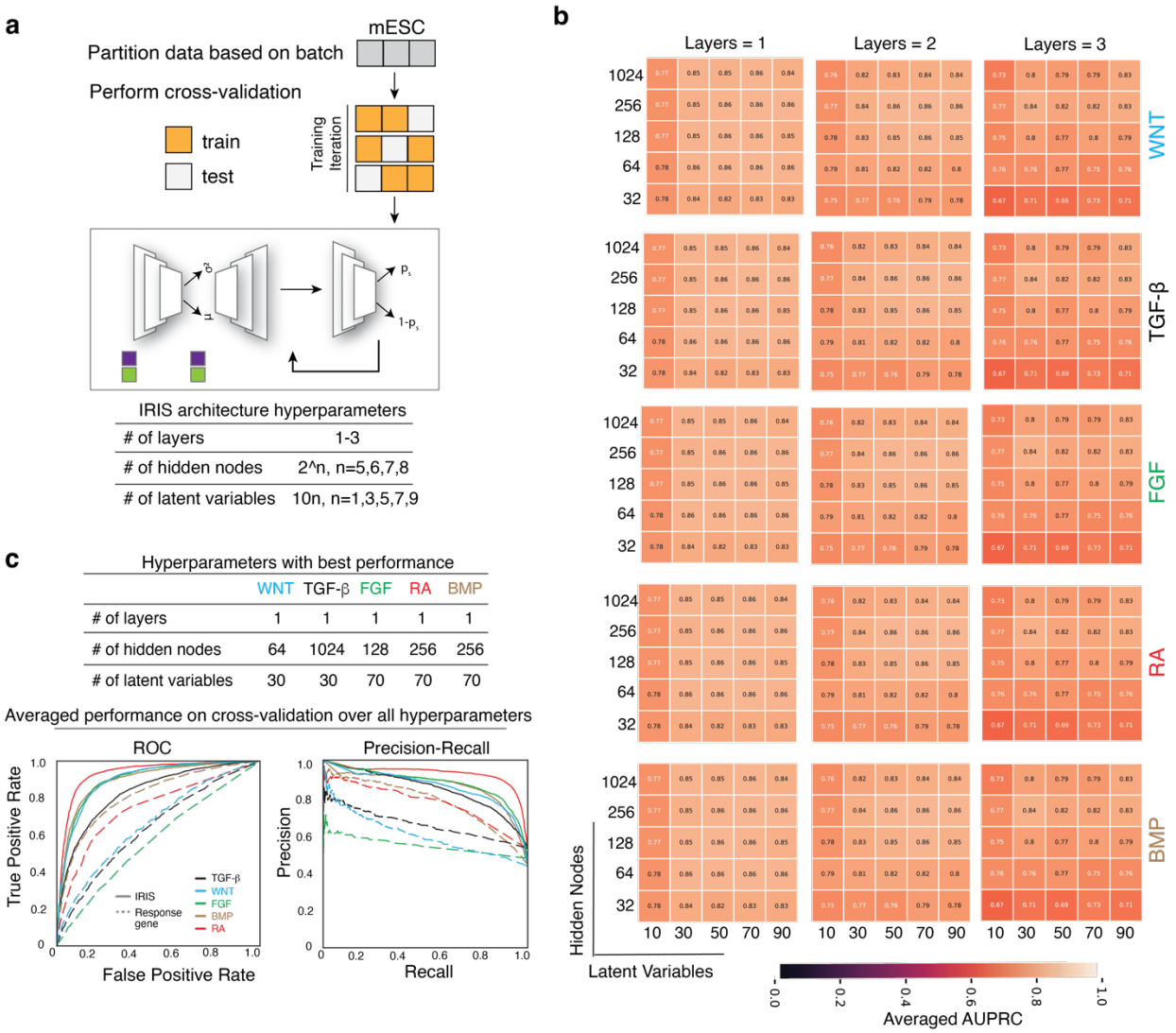

Figure S4. A hyperparameter screen of IRIS reveals optimal architecture parameters. **a**, Schematic of hyperparameter optimization. The mESC screen was partitioned into three batches during data collection. In the hyperparameter screen that aims to optimize for cross-batch generalizability, two of the three batches were used for training at a time and the remaining batch was used as a test. Multiple parameters were screened for layers, hidden nodes, and latent variables. **b**, The performance for each combination of the tested hyperparameters was evaluated based on the AUPRC (Area Under the Precision-Recall Curve) averaged across all three training iterations. **c**, Averaged cross-batch performance metrics of the ideal hyperparameters. The optimal parameters were chosen from the hyperparameter screen and all pathways showed improvement over the response gene approach. Performance was averaged across all three iterations.

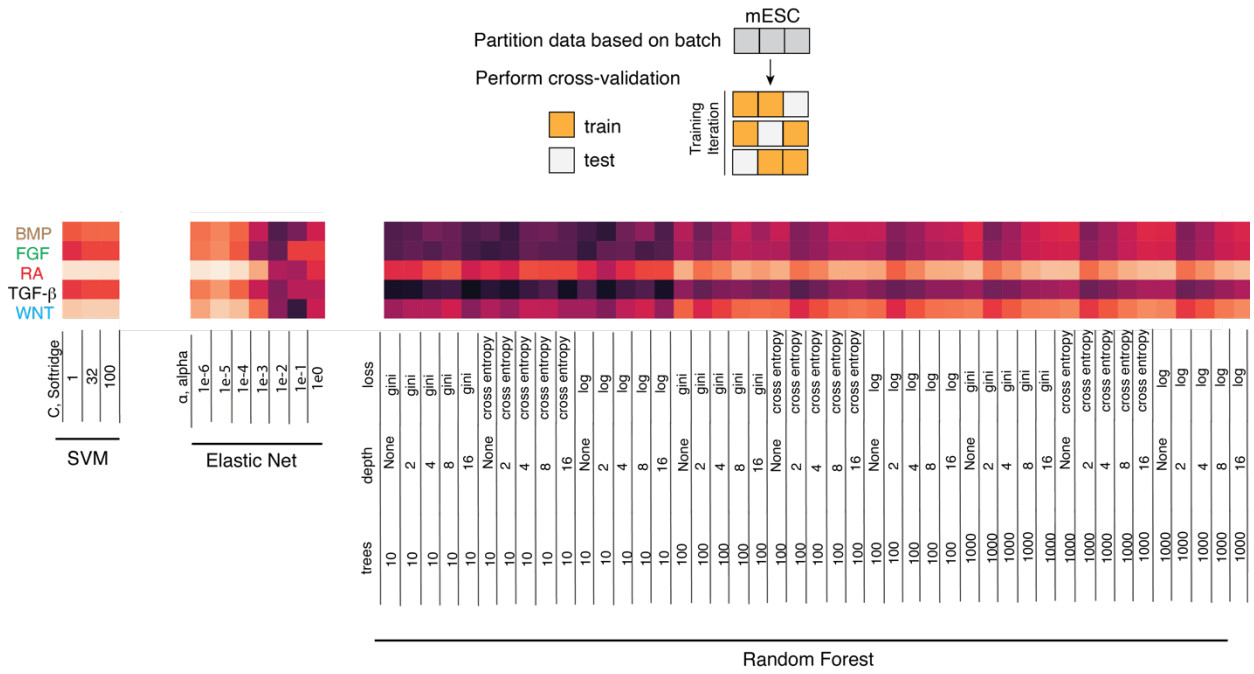

Figure S5. Hyperparameter optimization for white-box machine learning models using an in-sample cross-validation test.
PCA was first performed on the data in accordance with Scanpy's default parameters. The mESC screen was partitioned into three batches during data collection. Individual batches were held out one at a time and the average AUPRC is presented in the heat map for a variety of hyperparameters for each model: SVM (softridge, C=1, 32, 100), Elastic net (alpha, 10^-n, n=1, 2, 3, 4, 5, 6), and Random forest (loss=gini, cross entropy, log; number of trees= 10, 100, 1000; max depth=2, 4, 8, 16, None).

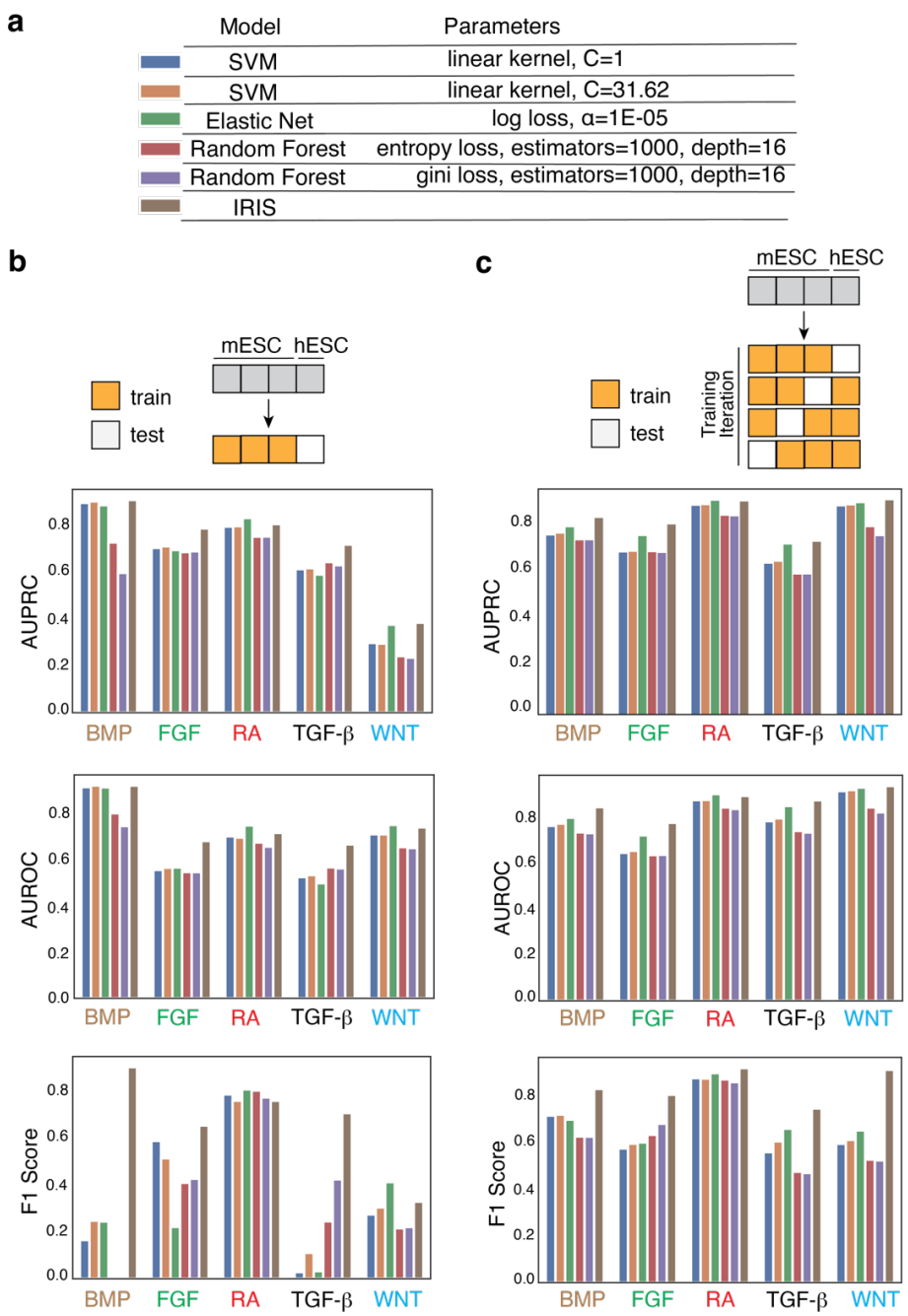

Figure S6. Comparison between IRIS and white-box machine learning models.
a, Optimal parameters were chosen from the in-sample cross validation test for each model. **b**, **c**, Performance of different models was evaluated in two ways: Trained on all three batches of mESC screen and tested on hESC screen, which evaluates generalization across cell types and species (**b**); and trained on three out of the four batches and tested on the held-out batch, with the average performance across all possible train-test splits displayed (**c**). The AUPRC, F1 score and AUROC were all evaluated for each of the signaling pathways. The outperformance of IRIS over white-box models with regard to the F1 scores suggests that improved generalization of IRIS has to do with improved thresholding. The F1 score plot in (b) is the same as **Fig. 2e**.

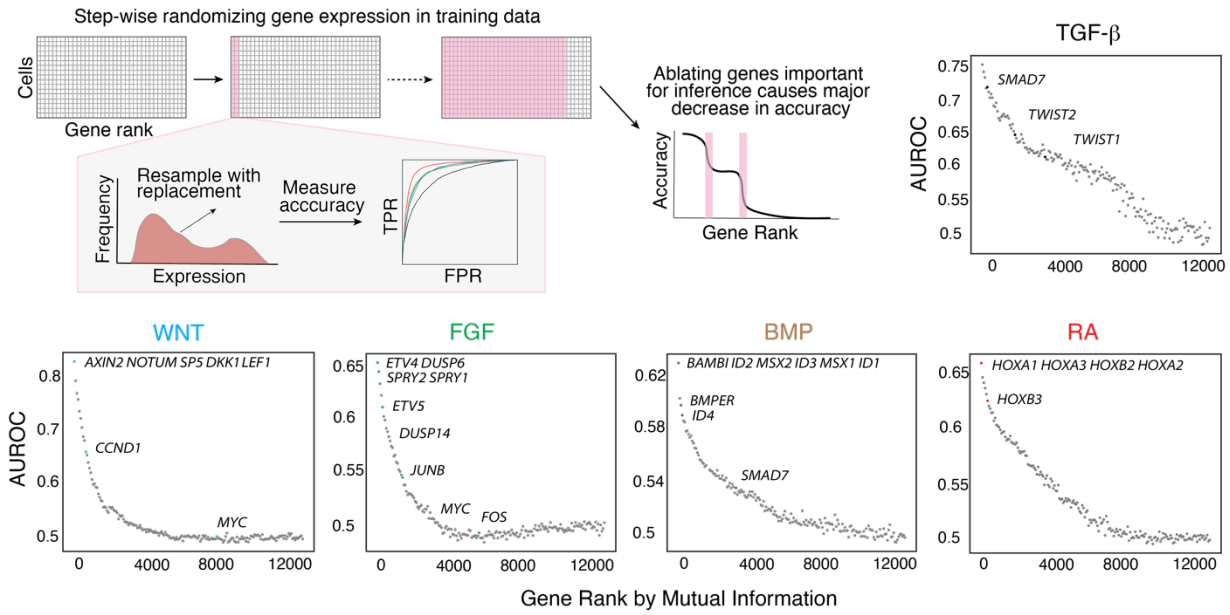

Figure S7. Gene ablation test.
Genes were ranked according to their mutual information with the predicted signaling pathway activity and the expression level of each gene was progressively corrupted through randomization, which was performed by resampling with replacement. The inference accuracy of IRIS was computed after each round of corruption. As in (**Fig. 4a**), many genes seem to contribute to the inference accuracy for each pathway, including but not limited to the classic response genes as highlighted in each plot. Models with above average performance were selected.

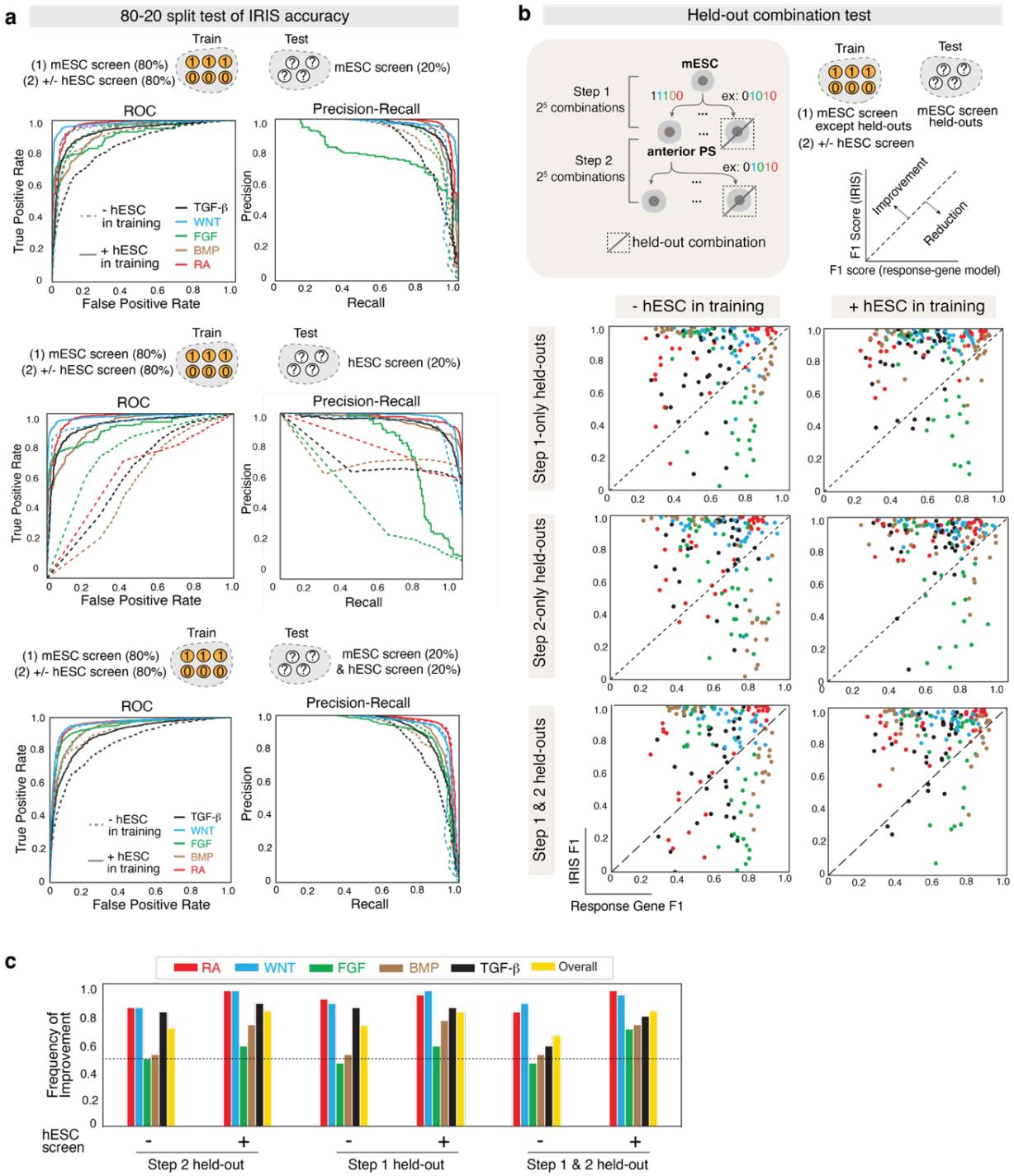

Figure S8. Increasing diversity of cell types in training data broadly improves inference accuracy.

a, Including the hESC screen in the training data improved in-sample inference accuracy. IRIS was trained on 80% mESC screen with or without 80% of the hESC screen, and tested on the remaining 20% of the mESC screen, the remaining 20% of the hESC screen, or both. **b**, **c**, Held-out signal combination test showed consistent improvement of IRIS over the response-gene

method and including hESC screen in training further improved the accuracy of IRIS. Samples with the same combination in Step 1 and Step 2 were held out together or separately. Comparison of IRIS and Response Gene F1 scores for each pathway. Individual points represent the performance of IRIS on each signaling pathway in a given held-out condition (**b**). The frequency of improvement over the response-gene method for each signaling pathway was calculated and presented as barplot (**c**).

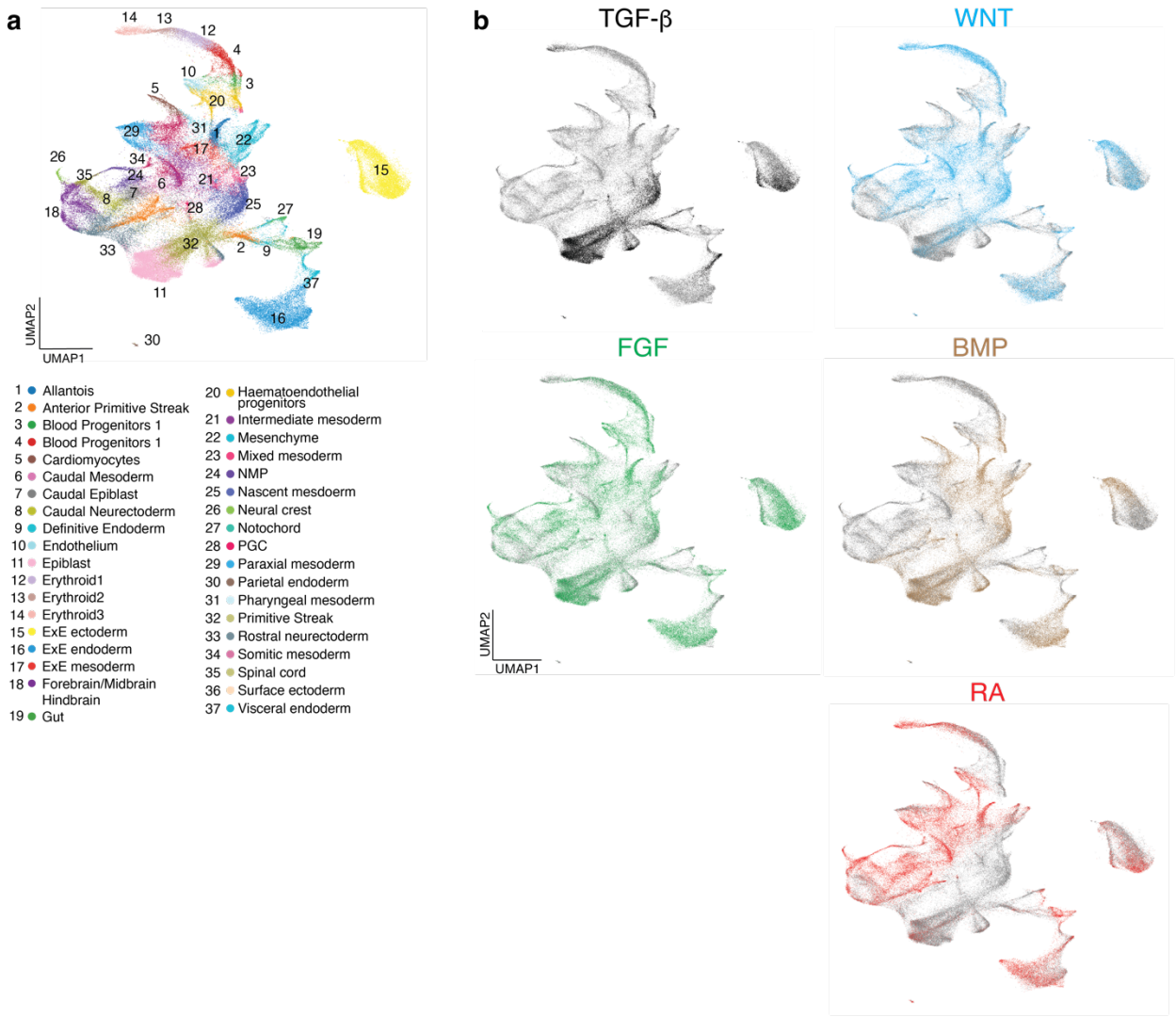

Figure S9. Inferring the activity of individual signaling pathways across the mouse gastrulation dataset.
a, Annotated cell populations in gastrulating mouse embryos (E6.5-E8.5) (same as **Fig. 4a**) [33]. **b**, The five signaling pathway activation predictions annotated on the mouse gastrulation atlas revealed patterns of response across diverse cell types. Cells inferred to have the denoted signaling pathway activated are highlighted in the corresponding color.

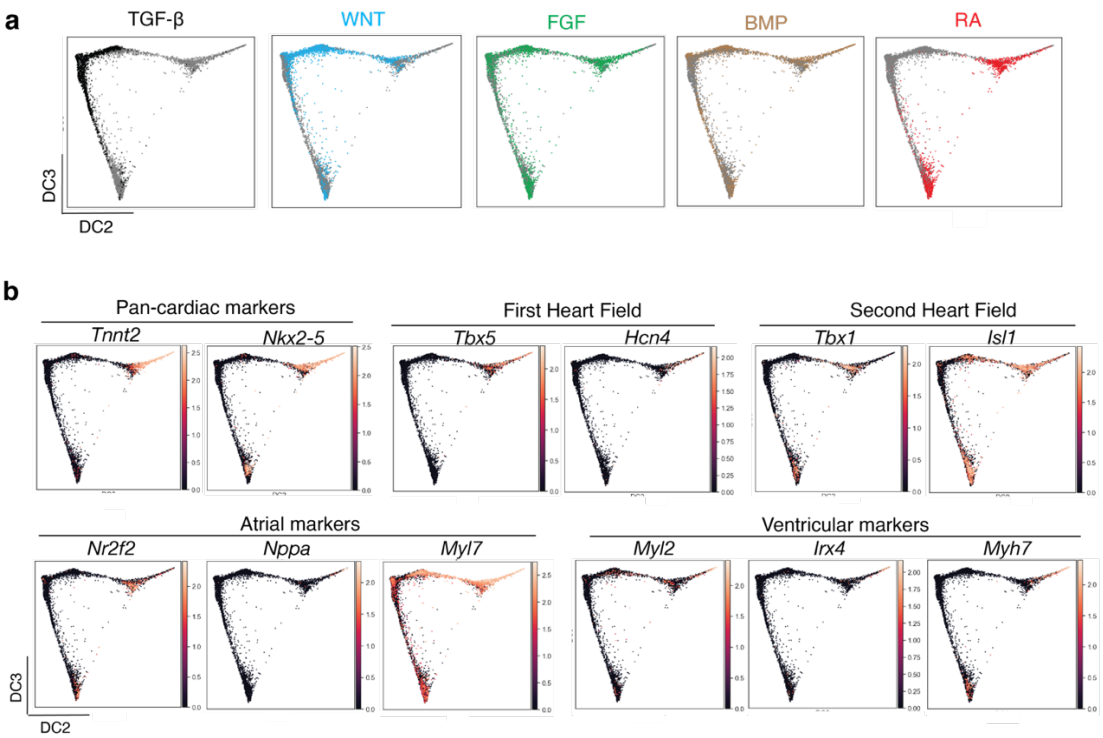

Figure S10. IRIS prediction for the endodermal and cardiac lineages.
a, Inferred signaling states for individual pathways across both the endodermal and cardiac lineages. Cells are arranged in the same way as they are in **Fig. 5b**. Cells inferred to have the denoted signaling pathway activated are highlighted in the corresponding color. **b**, Markers associated with different developmental stages of cardiac mesoderm align well with the diffusion component position. Both atrial and ventricular markers are expressed in the annotated cardiomyocyte population, revealing the heterogeneity of cell types.

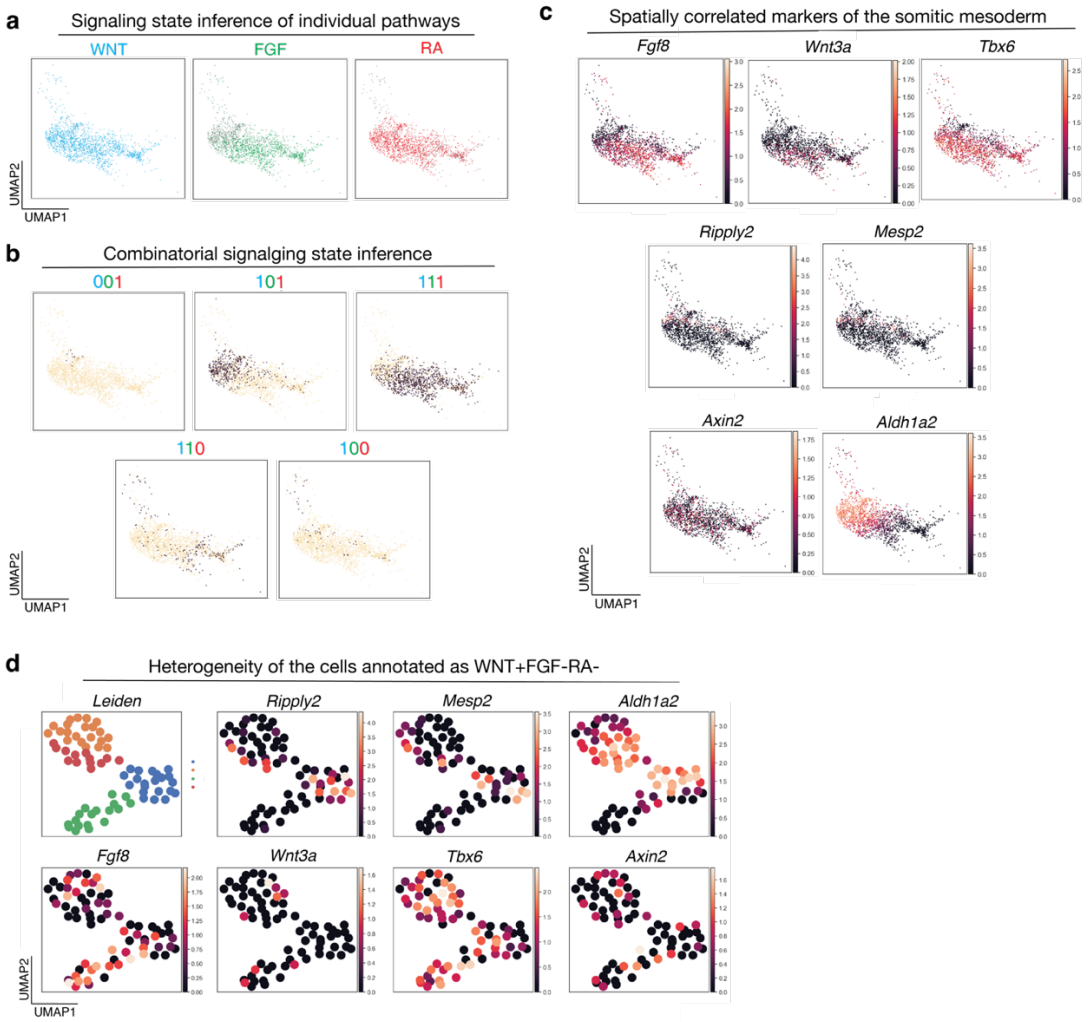

Figure S11. Spatial reconstruction of heterogeneous somitic mesoderm cells.
a, IRIS prediction on the somitic mesoderm population annotated in the atlas with respect to the activity of the RA, WNT, and FGF pathways individually [33]. Cells inferred to have the denoted signaling pathway activated are highlighted in the corresponding color. **b**, Inferred combinatorial signaling states overlaid on UMAP of the somitic mesoderm. Cells inferred to have the denoted combinations are highlighted on the diffusion maps (*black dots*). **c**, Expression of marker genes that are associated with spatial positions of the cells in the somitic mesoderm (Anterior: *Ripply2, Mesp2*; Posterior: *Tbx6, Wnt3a, Fgf8*; Signaling State: *Axin2*). **d**, Re-clustering of the predicted WNT+FGF-RA- cells show heterogeneous expression of both anterior and posterior markers. WNT+FGF-RA- cells fall into distinct clusters with anterior or posterior identities, suggesting WNT+FGF-RA- cells are most likely a collection of cells with false-negative and/or false-positive annotation of the signaling states.

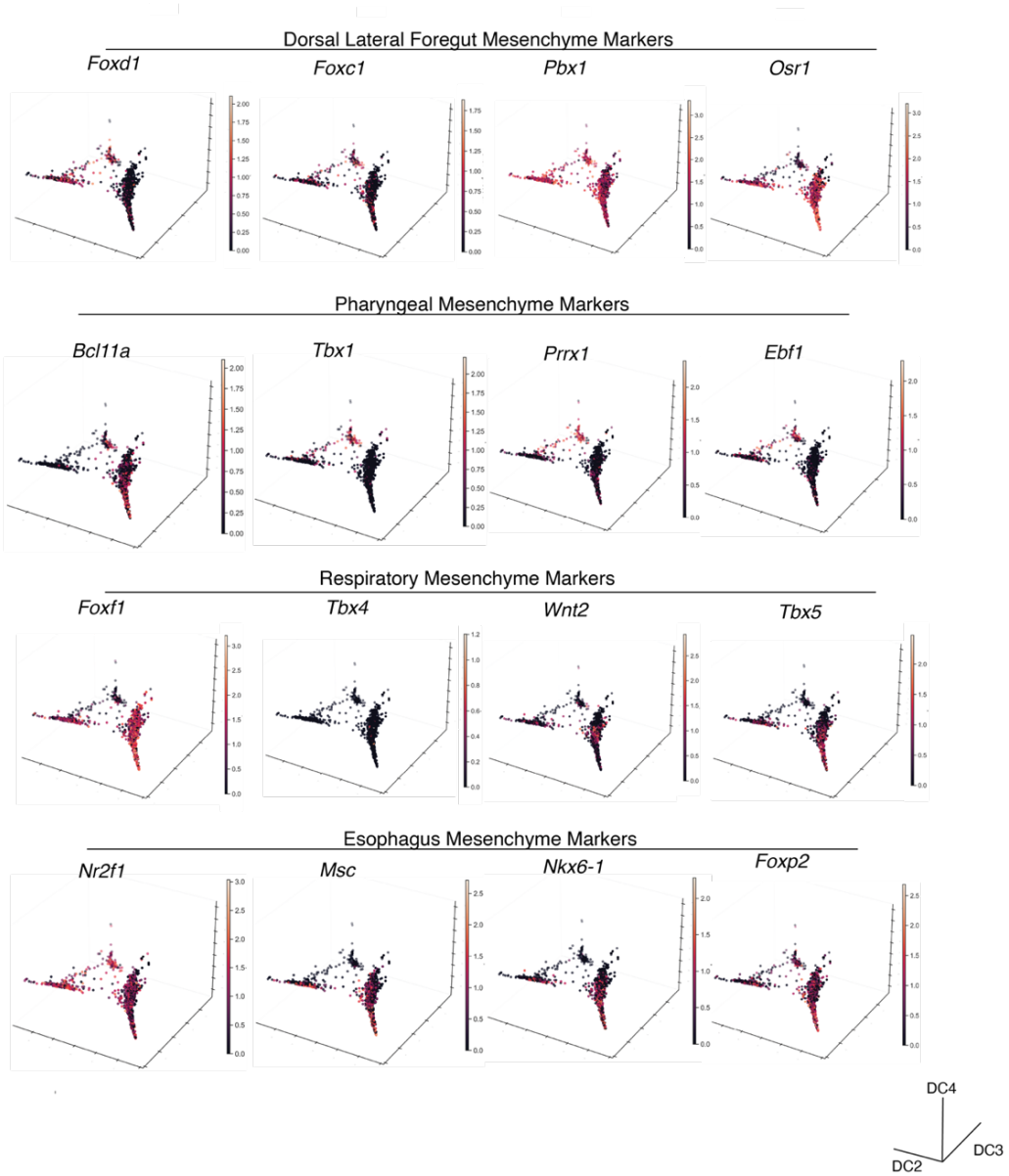

Figure S12. Fate divergence of organ-specific mesenchyme.
Marker genes corresponding to the dorsal lateral foregut mesenchyme, esophageal, pharyngeal and respiratory mesenchyme are overlaid on the diffusion maps. The respective cell populations are from E9-E9.5 mouse embryo foregut [17], and the 3D diffusion map was constructed in **Fig. 6a**.

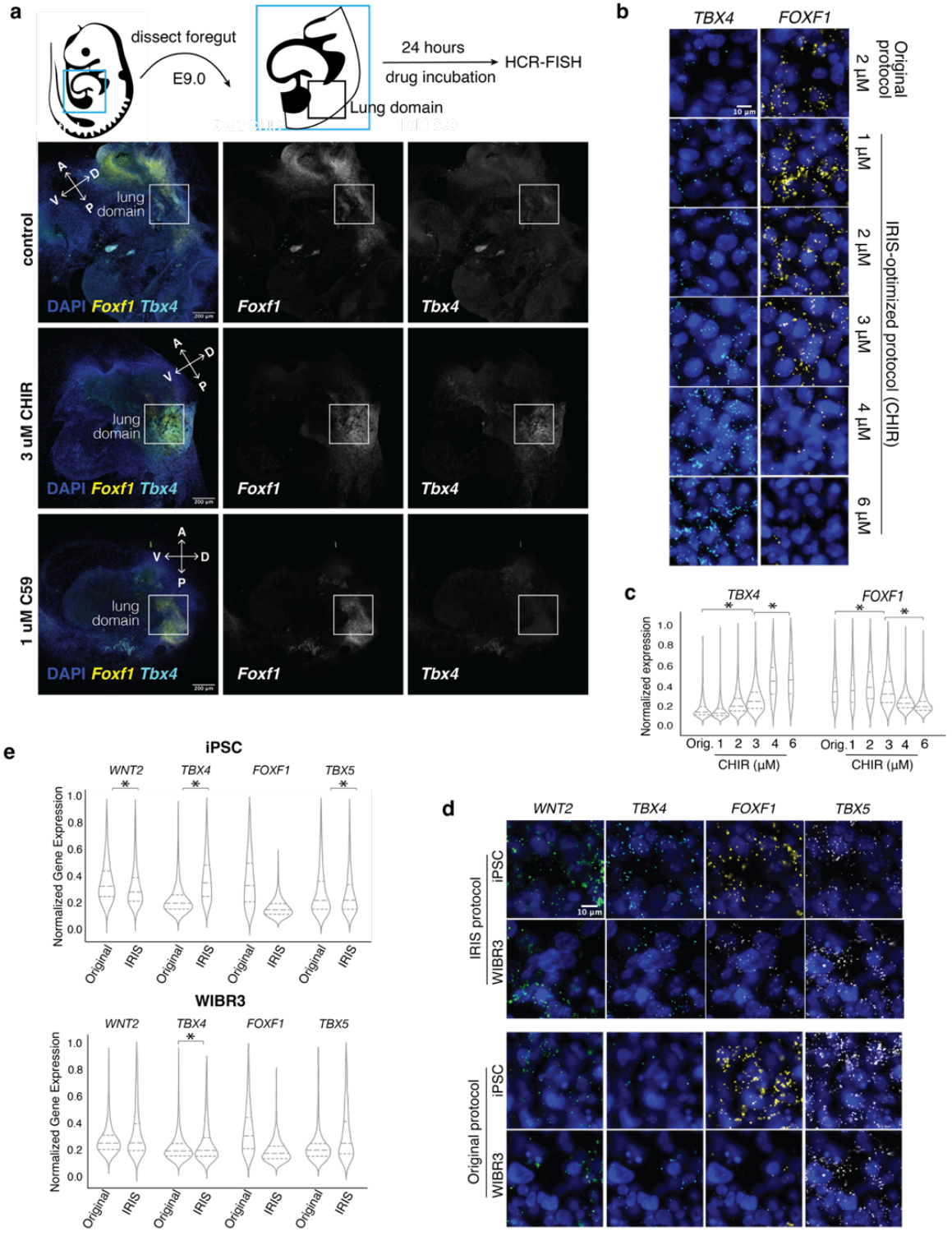

Figure S13. Validation of IRIS prediction in respiratory mesenchyme specification.
a, Zoomed out images of **Fig. 6b.** Whole mouse foregut explant images revealed restricted expansion of *Tbx4+* domain to the tissues adjacent to the putative lung domain upon WNT

stimulation. **b**, **c**, Titrating the level of WNT activation in the IRIS-optimized hESC differentiation protocol (WiCells WA01) modulated the expression levels of key respiratory mesenchyme markers. Multiplexed HCR-FISH was performed on Day 7 with representative images shown in (**b**) and quantification of transcripts per cell in (**c**). Number of cells for each condition: 6 μM, n=3228; 4 μM, n=2986; 3 μM, n=2990; 2 μM, n=3040; 1 μM, n=3417; Original, n=3913). **d**, **e**, IRIS-optimized protocol increased the expression of lung mesenchyme markers at the single-cell level, a result replicated across multiple human pluripotent cell lines. Multiplexed HCR-FISH was performed on Day 7 with representative images shown in (**d**) and quantification of gene expression levels per cell in (**e**). Number of cells for each condition: iPSC with IRIS protocol, n=2421; iPSC with original protocol, n=3200; WIBR3 with IRIS protocol, n=1637; WIBR3 with original protocol, n=2843).