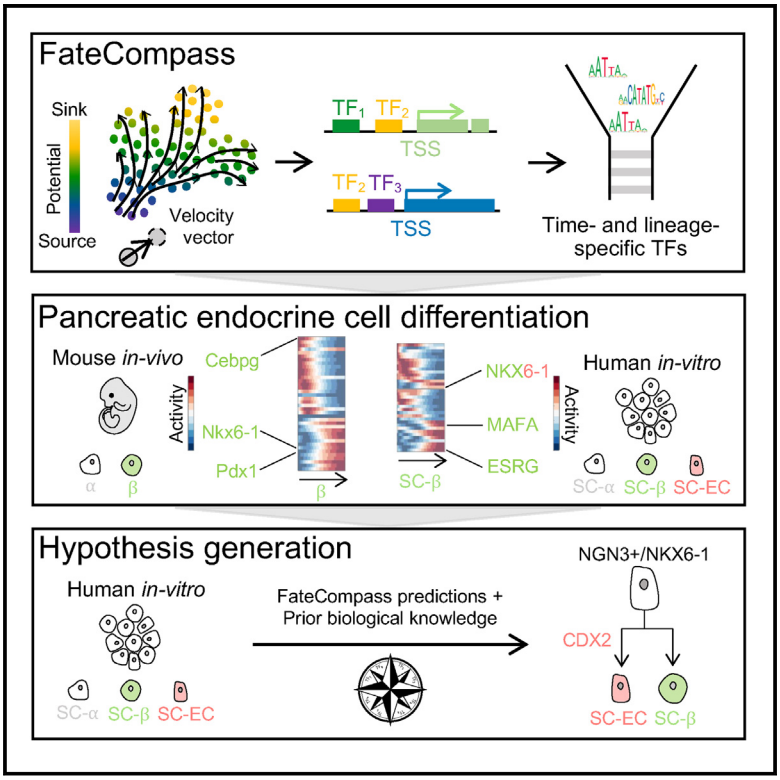
Cell Reports
Methods

Article

Characterization of cell-fate decision landscapes by estimating transcription factor dynamics

Graphical abstract



Highlights

- FateCompass is an integrative pipeline to estimate TF activity dynamics
- FateCompass learns trajectories using RNA velocity or a differentiation potential
- FateCompass reveals known and novel regulators in pancreatic datasets
- Lineage tracing validates FateCompass predictions for *in vitro* β cell differentiation

Authors

Sara Jiménez, Valérie Schreiber, Reuben Mercier, Gérard Gradwohl, Nacho Molina

Correspondence

sarajcorrea@gmail.com (S.J.), gradwohl@igbmc.fr (G.G.), molinan@igbmc.fr (N.M.)

In brief

Jiménez et al. present a flexible pipeline called FateCompass to identify time- and lineage-specific transcription factors and apply it to pancreatic endocrine cell differentiation datasets from mouse embryonic development and human *in vitro* differentiation. FateCompass predicted known and novel cell-subtype-specific regulators, for which they provided experimental support.

Jiménez et al., 2023, Cell Reports Methods *3*, 100512
July 24, 2023
https://doi.org/10.1016/j.crmeth.2023.100512

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Characterization of cell-fate decision landscapes by estimating transcription factor dynamics

Sara Jiménez,[1,2,3,4,*] Valérie Schreiber,[1,2,3,4] Reuben Mercier,[1,2,3,4] Gérard Gradwohl,[1,2,3,4,*] and Nacho Molina[1,2,3,4,5,*]
[1]Université de Strasbourg, Strasbourg, France
[2]CNRS, UMR 7104, 67400 Illkirch, France
[3]INSERM, UMR-S 1258, 67400 Illkirch, France
[4]IGBMC, Institut de Génétique et de Biologie Moléculaire et Cellulaire, 67400 Illkirch, France
[5]Lead contact
*Correspondence: sarajcorrea@gmail.com (S.J.), gradwohl@igbmc.fr (G.G.), molinan@igbmc.fr (N.M.)
https://doi.org/10.1016/j.crmeth.2023.100512

MOTIVATION Identifying lineage-specific regulators during dynamic biological processes, such as differentiation from single-cell transcriptomics data, remains challenging. Current methods rely on correlations between the expression of transcription factors (TFs) and their target genes without considering the dynamics of their regulatory activity, often resulting in complex lists or networks of TFs, from where the dynamic cascade of interactions is difficult to disentangle. Hence we developed FateCompass, an integrative pipeline that estimates TF activity dynamics and predicts lineage-specific regulators. FateCompass facilitates hypothesis generation to advance our understanding of the gene-regulatory networks underlying cell-subtype specification processes.

SUMMARY

Time-specific modulation of gene expression during differentiation by transcription factors promotes cell diversity. However, estimating their dynamic regulatory activity at the single-cell level and in a high-throughput manner remains challenging. We present FateCompass, an integrative approach that utilizes single-cell transcriptomics data to identify lineage-specific transcription factors throughout differentiation. By combining a probabilistic framework with RNA velocities or differentiation potential, we estimate transition probabilities, while a linear model of gene regulation is employed to compute transcription factor activities. Considering dynamic changes and correlations of expression and activities, FateCompass identifies lineage-specific regulators. Our validation using *in silico* data and application to pancreatic endocrine cell differentiation datasets highlight both known and potentially novel lineage-specific regulators. Notably, we uncovered undescribed transcription factors of an enterochromaffin-like population during *in vitro* differentiation toward ß-like cells. FateCompass provides a valuable framework for hypothesis generation, advancing our understanding of the gene regulatory networks driving cell-fate decisions.

INTRODUCTION

Gene regulation is pivotal during many biological processes, including development, cell cycle, regeneration, reprogramming, and cancer, and it usually occurs in a cell- and stage-dependent mode.[1] Notably, cells transition from a less to a more differentiated state during differentiation via the interplay of transcriptional regulation events in a highly dynamic manner.[1,2] Transcription factors (TFs) are essential proteins able to bind specific DNA regulatory regions and link signaling transduction networks to gene-specific transcriptional regulation;[3] hence they are commonly used as pathway readouts.[4] Currently, there are no high-throughput techniques to measure the dynamic regulatory activity of TFs; instead, their direct product, gene expression levels, can be measured with an unprecedented high resolution using single-cell transcriptomics.

Single-cell RNA sequencing (scRNA-seq) techniques allow the identification of different cell types and, more importantly, the study of lineage specification at the single-cell resolution,[5] enabling a quantitative study of differentiation. However, the challenge remains to devise a robust integrative workflow that infers time- and cell-type-specific regulators. The inherent asynchrony of scRNA-seq data has allowed the development of several approaches to reconstruct differentiation trajectories, which rely on variation among cell types within the captured population.[6] The developed computational techniques include pseudotime methods[7,8] and RNA velocity.[9,10] Noteworthy, pseudotime algorithms depend on the previous knowledge of the initial state, and

it is limited to the analysis of general trends of biological progressions rather than the precise dynamics of individual cells. Conversely, RNA velocity overcomes the limitation of the directionality by leveraging the splicing kinetics and predicting the RNA expression states in the near future. Nevertheless, it has intrinsic limitations, e.g., when the spliced-to-unspliced mRNA ratio is trendless or in predicting the continuous evolution of cells over a long period of time.[11,12]

Typically, differentiation trajectories are used together with differential gene expression analysis to identify TFs specific to a given cell type.[13] However, this approach ignores the fact that even lowly expressed TFs can have high regulatory activity, and it does not consider direct regulatory interactions with target genes. In contrast, several methods to derive mechanistic signatures in cell-fate decisions from transcriptomics data have been proposed, including ISMARA[14] and DOROTHEA[2] for bulk RNA-seq, and SCENIC[15] and metaVIPER[16] for scRNA-seq. Although previous studies attempted to compare these methods,[4,17] it is difficult to perform a benchmark given that most of them have different underlying assumptions. Notably, except for ISMARA, they are based on correlations between the expression of TF transcripts and the TF target genes or the expressed genes in general. Using correlations requires further assumptions or perturbation assays to distinguish causal relationships. Furthermore, none describes the dynamic change of TF activity throughout the cell-fate decision process, which is pivotal in time-dependent systems.

Here, we present FateCompass, an integrative workflow that aims to identify time- and lineage-specific TFs for a system undergoing differentiation. First, we outlined differentiation trajectories from progenitor cells to final states using a discrete Markov Process on a network. This allows us to describe stochastic gene expression dynamics during the cell-fate decision process incorporating RNA velocity or differentiation potentials to infuse the differentiation direction. Then, we modeled the observed gene expression as a linear combination of the regulatory sites and the TF activities; from this, we could infer TF activities. Finally, we performed a differential TF activity analysis using statistical criteria and validated the predictions using an *in silico* dataset. We applied FateCompass to a pancreatic endocrine differentiation system, where endocrine progenitors, marked by the transient expression of the TF Neurog3 (Ngn3), differentiate toward glucagon-producing α cells and insulin-producing β cells, among others.[18] We analyzed a well-characterized scRNA-seq dataset from the developing mouse pancreas[19] recapitulating known lineage-specific regulators and nominating novel regulators. Further, to demonstrate the capabilities of FateCompass, we used a scRNA-seq experiment from the differentiation of human stem cells toward pancreatic β-like cells.[20] Of note, this complex population includes, besides the expected endocrine cells, a population of intestinal-like cells called enterochromaffin (EC). FateCompass identified not only α- and β-specific but also EC-specific known and novel factors such as CDX2, for which we present further experimental support. The integrative approach of FateCompass boosts the ability to identify lineage-specific regulators by estimating TF activity dynamics, revealing time-specific transcriptional regulatory interactions underlying cell-subtype specification.

RESULTS

Integrative methodology to infer dynamic TF activities during cell-subtype specification

FateCompass aims to identify lineage- and time-specific TFs for a cellular system undergoing differentiation. To this end, we designed a three-step pipeline using scRNA-seq (Figure 1). First, we characterized the dynamics of the system by sampling differentiation trajectories from progenitor to final states using a discrete Markov process on a network representing cell states and possible transitions. Next, we focused on TFs as readouts of pathway activity. Hence, we inferred TF activities using a linear model of gene regulation, which involved modeling the observed gene expression as a linear combination of the regulatory sites and the TF activities, similarly to the original framework implemented in ISMARA.[14] Finally, we defined a differential TF activity analysis considering dynamic changes and correlations to coarse-grain the list of TFs and identify lineage-specific regulators.

To infer differentiation trajectories and describe stochastic gene expression profiles along the cell-fate decision process, we used cell-to-cell similarity based on transcriptomic profiles together with RNA velocity[9] or a differentiation potential from progenitor to mature cells.[21] Similar to other differentiation trajectory inference methods,[6–8,22] FateCompass models differentiation trajectories based on similarities in the expression patterns, except that we implemented a cell-dependent drift that biases the trajectories toward the direction of differentiation.

First, we represented the phenotypic manifold in a low-dimensional space and modeled it using a Markov chain on a network (Figure 1A; STAR Methods). To that end, we first embedded the gene expression data in a significant low-dimensional manifold using either principal-component analysis (PCA)[23] or Uniform Manifold Approximation and Projection (UMAP).[24] Next, we built a nearest-neighborhood graph in the low-dimensional space connecting each cell with the *k* most similar neighbors. Notably, FateCompass flexibility allows the input of the graph when computed using alternative methods (Figure S1); for instance, Sanity calculates cell-to-cell distances by removing Poissonian noise from scRNA-seq data by implementing a rigorous Bayesian approach.[25] Currently, RNA velocity is a well-accepted method to infer differentiation dynamics; however, some of the method's limitations lead to inconclusive velocity fields in some biological systems. For instance, some datasets might have, to name just a few, time frames out of the initial modeling framework, insufficient unspliced counts in the key biological driver genes, and multiple kinetic regimens.[11] Therefore, to make FateCompass flexible and applicable to any differentiating system, we used either the RNA velocity vector (STAR Methods) or the gradient of potential energy from progenitor cells to mature cells (STAR Methods) to bias the transitions between states in the Markov process (Figures 1B and S1). The resulting transition probabilities reflect both transcriptional and directional similarities. To ultimately describe the time evolution of the differentiating system, we used a Monte Carlo sampling algorithm where the transition matrix of the Markov chain gives the next-jump probability (STAR Methods). This approach is instrumental in estimating quantities of interest, e.g., gene expression or TF

CellPress
OPEN ACCESS

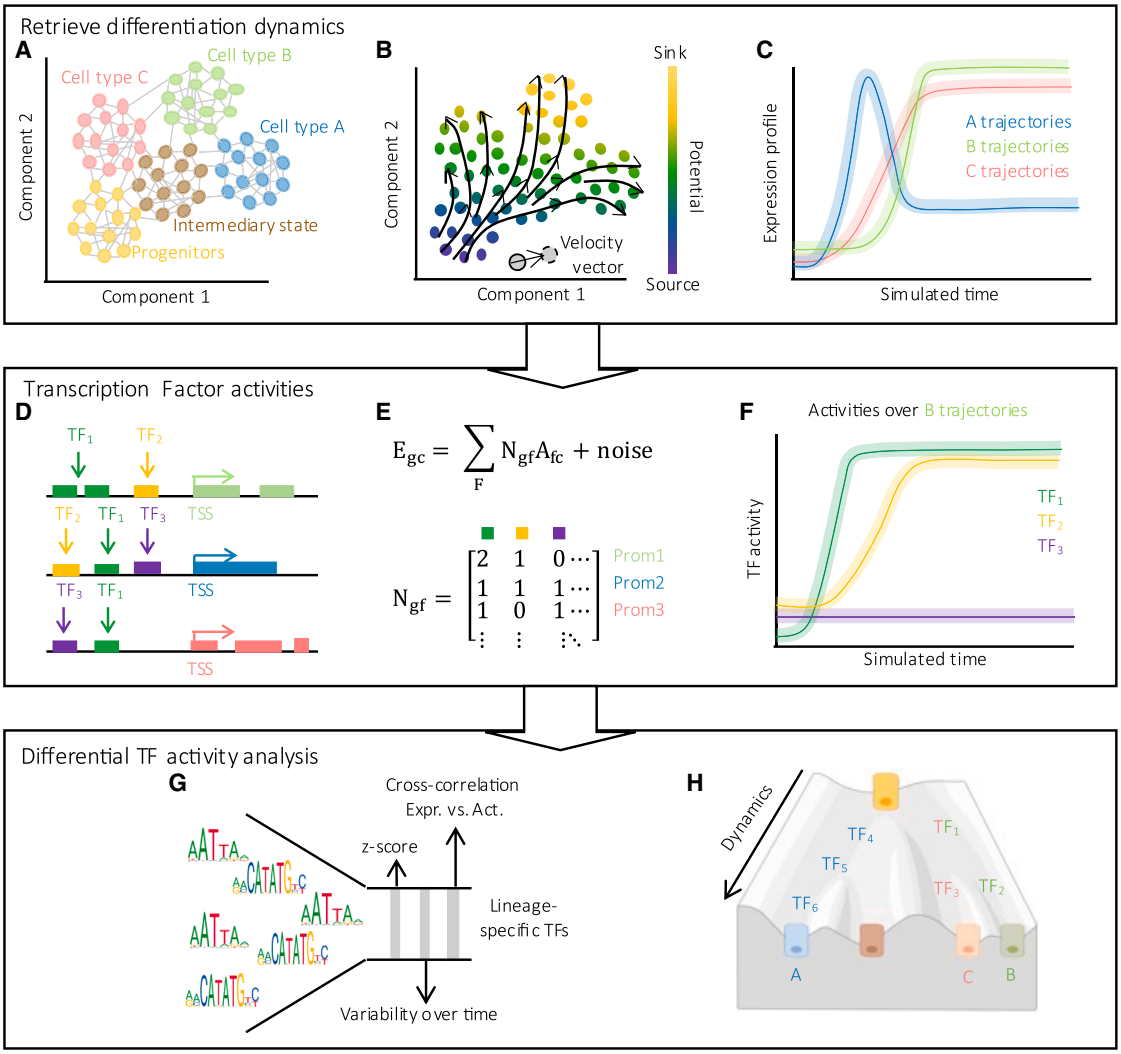


Figure 1. FateCompass workflow to identify transcriptional regulators during cell-subtype specification

(A) scRNA-seq data annotated for different cell types and embedded in a low-dimensional space showing connections to the *k* nearest neighbors.

(B) The direction of differentiation, either from RNA velocity or differentiation potential, is the underlying force to compute transition probabilities between cell states.

(C) The dynamic profile of a given gene over the differentiation trajectories ending in a respective final state (A, B, and C).

(D) TF binding sites for different motifs across promoters.

(E) Linear model of gene regulation where the observed gene expression (E_{gc}) is modeled as a linear combination of the regulatory sites (N_{gf}) and the TF activity (A_{fc}).

(F) The dynamic profile of the TF activity for TF_1, TF_2, and TF_3 over the differentiation trajectories ending in the final state B.

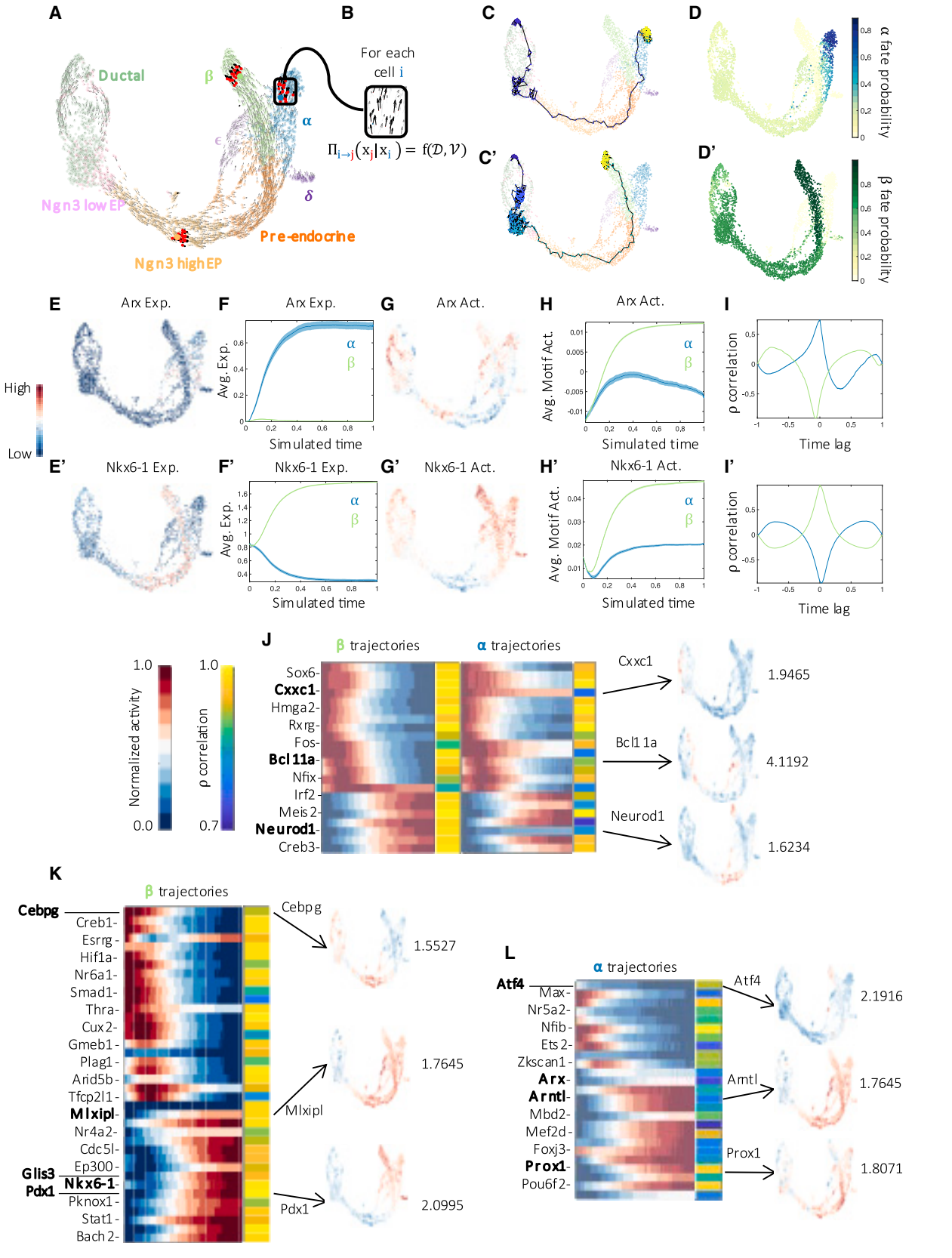
(G) Differential TF activity analysis. Lineage-specific TFs were identified according to three criteria: *Z* score, variability of TF activity over time, and dynamic correlation between the TF activity and the expression of its mRNA.

(H) Lineage-specific TFs identified by FateCompass ordered according to who is active first over the dynamic profiles and constitute the differentiation TF landscape.

See also Figures S1 and S2.

activity over differentiation trajectories (Figures 1C and 1F; STAR Methods).

To decipher the transcriptional interactions driving cell-subtype specification, we used TFs as a proxy because of their direct role in gene-specific transcriptional regulation.[26] We reasoned that changes in the transcriptional state, in response to developmental cues, are conditioned by conserved regulatory mechanisms, such as the interaction between TFs and *cis*-regulatory regions (Figure 1D). Similar to Balwierz et al.,[14] we used a linear model to infer TF activities (STAR Methods). The primary assumption is that the transcription rate is controlled by the TF binding sites in the promoters (Figure 1E; STAR Methods); we considered promoters because of their direct assignation to the target genes based on their proximity to the transcriptional



(legend on next page)

start site (TSS). We defined a promoter region as the 2-kb region centered in the TSS. Currently, there is no standard high-throughput way to assign reliably long-distant regulatory interactions such as enhancers to target genes. However, we designed FateCompass such that the regulatory interactions, summarized in a binding-site matrix, could be customized to include further information when available (Figure S1), for example, coming from single-cell ATAC sequencing (scATAC-seq). Importantly, we implemented a new regularization technique using data diffusion to control the model's complexity and avoid overfitting (STAR Methods). Shortly, we used the *k*-nearest neighborhood graph to smooth the learned activities correcting for dropout and other noise sources.[27] In the data diffusion regularization, the cells share information through the local neighbors, which is analogous to diffusing the data over the network. The *t*-step is akin to raising the diffusion operator to the t^{th} power. We fitted the optimal value of *t* using a cross-validation scheme.

Finally, to identify lineage-specific TFs, we defined a differential TF activity analysis based on three metrics (Figure 1G; STAR Methods). First, we reasoned that activities that are important to explain the expression variation across cells should be relevant in the cell-fate decision; we summarize this using the *Z* score (STAR Methods). Next, we looked for TFs whose activity profile over the differentiation trajectories was highly changing, intuiting that these will have a crucial role in the state-transition process (STAR Methods). Lastly, we acknowledge that for a TF to be active, it has first to be expressed; hence, we sought TFs with high and positive dynamical cross-correlation (STAR Methods). The strength of FateCompass lies in its integrative approach, which boosts the ability to identify time- and lineage-specific regulators by estimating TF activity dynamics.

We tested FateCompass performance with an *in silico* dataset generated using the simulation engine Dyngen (Figure S2A).[28] The synthetic data has an underlying ground truth for developmental trajectories, RNA velocity profiles (Figure S2B), and cell-specific regulatory networks (Figure S2C). We applied FateCompass using the differentiation potential mode (Figure S2D), and we used the prior biological knowledge on the initial and final fates of the ground-truth system. FateCompass accurately estimated fate probabilities (Figure S2E) for each final fate (sEndC and sEndD). Also, when exploring the expression and the TF activities computed using FateCompass of fate-specific TFs for each lineage, according to the underlying gene-regulatory network, we observed a coherent behavior in the dynamic profiles (Figures S2F and S2G). Finally, the differential TF activity analysis identified TFs that drive each cell subtype at the decision point: B11 for the sEndC fate and B14 for the sEndD fate (Figure S2H). Altogether, these results showed the potential of FateCompass to identify time-specific regulatory interactions, thereby providing the opportunity to generate hypotheses on transcriptional regulation cascades underlying cell-subtype specification processes.

Delineating transcriptional regulators during mouse islet cell formation

To assess the robustness of FateCompass, we applied it to a well-characterized scRNA-seq dataset from the developing mouse pancreas at 15.5 days post-coitum (E15.5).[19] In the pancreas, endocrine cells differentiate from endocrine progenitors marked by the transient expression of the TF Ngn3.[18] Bastidas-Ponce et al.[19] profiled pancreatic epithelial cells using a Ngn3-Venus fusion reporter mouse line, sequencing both Venus-positive and Venus-negative ($Epcam^{+}$) cells using droplet-based scRNA-seq (10X genomics chromium). We tested the capabilities of our workflow with the data from E15.5 (3,696 cells), when endocrine cell commitment ends in four major cell types: glucagon-producing α cells, insulin-producing β cells, somatostatin-producing δ cells, and ghrelin-producing ε cells (Figure 2A). Moreover, this dataset presents a strong directional velocity flow toward the final endocrine fates (Figure 2A).[10]

To retrieve the dynamic profiles toward the final endocrine fates, we embedded the data in a low-dimensional manifold using UMAP, and we built the nearest-neighborhood graph in the reduced space. Next, we leveraged the robust RNA velocity profile to direct the edges of the Markov chain (Figure 2B) and estimated transition probabilities using a velocity-driven kernel (see STAR Methods). We used the transition matrix with a Monte Carlo sampling algorithm to simulate stochastic gene expression profiles along the differentiation trajectories, which allowed us to plot gene expression trends. The time-evolution simulation

Figure 2. Islet cell formation landscape in the mouse

(A) UMAP plot of 3,696 cells at E15.5;[19] colors highlight clustering into eight main cell types. Arrows indicate the direction of cell transitions, which was estimated using RNA velocity. Red dots represent the possible neighborhood a cell can explore when modeled using a Markov chain.

(B) Propagator (Π_{ij}) of the Markov chain as a function of the direction of differentiation ($\mathcal{V}$) and the stochasticity of gene expression ($\mathcal{D}$). Π_{ij} represents the probability of transitioning to the state *j* when being at the state *i*.

(C and C′) Stochastic differentiation trajectories starting in a Sox9 bipotent cell and ending in α (C) and β (C′) fates.

(D and D′) Fate probabilities for the α (D) and β (D′) fates.

(E and E′) UMAP plots with normalized gene expression of known lineage-specific markers: Arx for α (E) and Nkx6-1 for β (E′) cells.

(F and F′) Average gene expression of Arx and Nkx6-1 over α and β differentiation trajectories. Shaded areas represent standard errors of the mean.

(G and G′) UMAP plot with TF activity distribution for Arx (G) and Nkx6-1 (G′).

(H and H′) Average TF activity of Arx and Nkx6-1 over α and β differentiation trajectories. Shaded areas represent standard errors of the mean.

(I and I′) Dynamic Pearson correlation between mRNA expression and TF activity over α and β trajectories for Arx (I) and Nkx6-1 (I′).

(J–L) Heatmaps showing the average TF activity over differentiation trajectories and dynamic Pearson correlation. Not all TF names are shown (see Table S1 for full list). Right: activity distribution on the UMAP embedding of selected examples and respective *Z* score. TFs are sorted according to which was active first in each trajectory.

(J) TFs predicted as drivers of both β and α fates.

(K) TFs predicted as β specific.

(L) TFs predicted as α specific.

See also Figures S3–S5.

showed that, after two thousand iterations, the simulations mainly ended in three of the final endocrine fates, α, β, and ε, with higher frequencies for α (12.4%) and β (77%). We next focused the analysis on α and β lineages. In Figures 2C and 2C′, we presented an example of the simulated trajectories for the final fates α and β. These fates represent sinks during the endocrine cell-subtype specification. We computed fate probabilities utilizing the information from stochastic simulations (STAR Methods) and summarized this information in a fate-probability distribution (Figures 2D and 2D′). In summary, FateCompass can integrate RNA velocity information to estimate gene expression dynamics and identify final fate states from scRNA-seq data.

To assess the influence of the kernel on our dynamic inference approach, we computed fate probabilities using the differentiation potential as the drift of the stochastic differentiation process. Notably, no major changes were observed in the fate-probability distribution (Figure S3A). This suggests that differentiation potential can be used when prior biological knowledge is available on final cell fates and RNA velocity fails to provide useful dynamic information. Furthermore, we tested the robustness of FateCompass by examining whether the number of nearest neighbors (*k*) or the dimensionality reduction method could influence the inference dynamics. To this end, we recomputed fate probabilities for different values of *k* or using PCA to reduce the dimensionality of the data. As shown in Figure S3B for three random cells, we did not observe any significant change in their fate probabilities. These findings demonstrate that FateCompass is robust to different parameter settings and dimensionality reduction methods.

Importantly, we compared the fate probabilities inferred with FateCompass and CellRank, another computational method that uses RNA velocity to estimate fate probabilities.[29] In the dataset we used, which differs from the one used in the original CellRank paper in the fact that we included the Ductal and Ngn3 low EP populations, CellRank identified ε and β cells as terminal hormone-producing fates failing to identify the glucagon-producing α cells as a terminal state (Figure S3C). Importantly, for all the variations in parameters and kernels, FateCompass accurately identified the β fate as the most likely terminal fate for endocrine progenitors at E15.5, consistent with the previous biological knowledge.[19,30]

Next, to evaluate the dynamic profile of transcriptional regulators, we estimated TF activities from the behavior of the predicted target genes. To assess the inferred transcriptional dynamics objectively, we examined the expression and activity profiles of known lineage-specific regulators. The TF Arx is essential for α cell formation,[31] and unsurprisingly, it had high expression in α cells (Figure 2E), and the expression profile over the α trajectories increased, whereas over the β trajectories it was flat (Figure 2F). Similarly, the TF Nkx6-1, necessary for β cell development,[32] was highly expressed in developing β cells (Figure 2E′), and its expression profile increased only over β trajectories (Figure 2F′). Strikingly, we found high Arx and Nkx6-1 activity in both α and β cells (Figures 2G and 2G′). According to the TF activity definition, there will be increased activity for a TF when its targets show, on average, an increase in expression that cannot be explained by the presence of sites for other TFs in their promoters.[14] We reasoned that for a TF to have an effect, it must be present, and we assumed that if the TF is expressed in the population of interest, its protein will be present and can have regulatory activity. Hence, TFs that show a positive correlation between their activity and their mRNA expression are predicted to be activators, and when the correlation is negative, the TFs are predicted as repressors. To assess the role of Arx and Nkx6-1 in both α and β cells, we extracted the dynamic profile of Arx (Figure 2H) and Nkx6-1 (Figure 2H′) activities and then performed dynamical correlation over α and β trajectories between the expression and the activity. We found that Arx is an activator of α cell identity (positive correlation without any time lag), whereas it is a repressor of the β lineage (negative correlation) (Figure 2I). This cell-dependent role has been well documented.[31,33] In contrast, Nkx6-1 behaves as an activator during β cell differentiation, whereas it has a repressor role in α cells (Figure 2I′). This antagonistic behavior is supported by the findings of Schaffer et al.[34] showing that repression of *Arx* by Nkx6-1 is important for determining and maintaining β cell identity. Importantly, we tested the robustness of the TF activity profiles against the number of nearest neighbors and the dimensionality reduction method used in the first step of graph construction, and we found consistent activities for several values of the parameters; in Figure S4A we showed Nkx6-1 activities as a relevant example.

The differential TF activity analysis identified 86 TFs (Table S1), from which 22 were predicted to be specific for both α and β fates (Figure 2J), 38 were β specific (Figure 2K), and 25 were α specific (Figure 2L). Of note, we also checked the robustness of the differential TF activity analysis against the number of nearest neighbors and the dimensionality reduction method. Specifically, when we focused on the β lineage, we found that the TFs identified by FateCompass were largely overlapping (Figure S4B). Interestingly, although Neurod1 was identified for both fates, the profile over the differentiation trajectories was higher in the β cells; this can also be observed in the distribution of activities in the UMAP plot (Figure 2J). The prediction of Neurod1 for both fates is consistent with previous publications about the role of Neurod1 in murine α and β cell specification, where the authors found a cell-type-dependent role of Neurod1 in combination with Nkx2-2,[35] whose activity profile is similar to Neurod1's profile (Figures S5C and S5C′). Bcl11a was the factor with the highest *Z* score (4.1192) (Figure 2J), indicating the significance of this TF to explain the variance of the linear model of gene regulation. Remarkably, Bcl11a has an active role as a potent suppressor of insulin secretion in adult islets;[36] however, its role during islet cell-subtype specification remains to be studied. Some of the identified factors for both α and β fates have not yet been reported to have a function during pancreatic endocrine differentiation; thus, they are potential novel regulators. For instance, we found Cxxc1 early on during differentiation (Figure 2J); it has already been pointed out as a critical factor during other differentiation processes, such as in thymocyte development.[37] Also, previous studies from our group reported it to be a direct target of Ngn3.[38]

Regarding the identified β-specific factors, we identified several TFs known for playing a role in β cell subtype specification or identity, including Nkx6-1,[32] Glis3,[39] Mlxipl,[40,41] and

Pdx1,[42] as well as novel candidates such as Cebpg (Figure 2K). Similarly, among the factors designated as α-specific (Figure 2L), we found known ones such as Arx[31] and novel candidates such as Prox1.[43] Of note, Prox1 needs to be downregulated in β cells for their expansion and maturation.[43] Hence our predictions might suggest a role for Prox1 in α cell differentiation, where this gene, in contrast with β cells, is highly expressed. We also identified core clock factors such as Arntl as α specific. Previous studies have shown that the distinct characteristics of α cell and β cell clocks harbor different circadian properties resulting in differential gene expression and functional regulation.[44] Notably, Arntl was previously identified with an advanced phase in α cells.[44]

Finally, we compared the TF activities inferred with FateCompass and Scenic, a computational approach that infers regulatory activity using the AUC method.[15] We performed a Wilcoxon test on the estimated activities to highlight the top five TFs per cluster (Figures S4C and S4D). Of note, this statistical test is not part of the Scenic workflow, and although through it TFs such as Pdx1 and Arx are pinpointed for β and α cells, respectively (Figure S4C), Scenic does not identify relative-time-lineage-specific regulators. We would like to point out that just a statistical test on the estimated activities by FateCompass is not enough to retrieve TFs involved in differentiation dynamics (Figure S4D). Taken together, by integrating dynamics and TF activity inference, FateCompass predicted well-known TFs, which serve as a positive control of the method's performance. Beyond that, we identified novel potential β and α fate regulators with clearly distinct dynamic behaviors. This information can be harnessed further to characterize the regulatory interactions behind pancreatic endocrine cell formation.

FateCompass identifies transcriptional dynamic profiles beyond RNA velocity

To test whether the FateCompass workflow retrieves differentiation trajectories and identifies lineage-specific TFs in more complex experimental designs, with several harvesting points, we considered a scRNA-seq experiment from an *in vitro* differentiation of human stem cells toward pancreatic β-like cells.[20] In this study, the authors used the SC-β cell protocol to mimic β cell development[45] in six stages using specific inducing factors to produce "stem-cell-derived islets" (SC-islets) that contained SC-β cells (Figure S6A). We applied FateCompass to a dataset of 25,299 cells profiled using In-Drops sequencing across eight time points throughout stage five. Notably, at the beginning of stage five, there were NKX6-1$^+$ pancreatic progenitors, as well as the first SC-α cells, and by the end of it, there were three classes of endocrine cells: SC-β cells expressing INS, NKX6-1, ISL1, PDX1, and other β cell markers; SC-α cells expressing GCG, ARX, IRX2, and also INS; and SC-EC cells expressing CHGA, TPH1, LMX1A, and SLC18A1 that resembled intestinal EC cells (Figure 3A). We were interested in seeing how well FateCompass retrieved lineage-specific TFs in this setting, where there is a directed differentiation (toward SC-islets) with an undesired by-product (SC-EC cells).

To infer the differentiation trajectories toward the final endocrine cell types, we first computed RNA velocities using scVelo[10] and visualized them using two-dimensional UMAP embedding (Figure S6B). Notably, the projected velocities did not have a conclusive pattern toward the final fates, probably because of the high proportion of unspliced transcripts (30%; Figure S6C). The above, together with the inherent limitation coming from the batch effect introduced by harvesting at different time points (Figure S6A),[11] made the use of RNA velocity as a drift for the transition probabilities a liability. Therefore, we estimated FateCompass transition probabilities using the differentiation potential gradient from NKX6-1$^+$ progenitors to each terminal fate (SC-α, SC-β, and SC-EC) (Figures 3B and 3C; see STAR Methods). To summarize the information of the stochastic simulations, we plotted the fate-probability distribution for each final cell type. Notably, the likelihood of having SC-α as final fate strongly decreases for NEUROG3-mid and NEUROG3-late progenitors, suggesting that this cell type comes mainly from early endocrine precursors (Figure 3D), consistent with previous reports.[46] Conversely, SC-β and SC-EC are the prevalent endpoints for trajectories passing through late NEUROG3 progenitors (Figures 3E and 3F), consistent with previous studies,[20,47] validating our drift-dependent Markov chain approach to infer differentiation dynamics. Of note, alternative methods to infer fate probabilities, such as the pseudotime-based kernel available in a recent update of CellRank, failed to distinguish between the SC-β and SC-EC fates (Figures S6D–S6F).

Next, we checked the dynamic profile of known regulators. NKX6-1 is pivotal at several differentiation stages to giving rise to β-like cells.[46] Congruently, Veres et al.[20] reported high expression of NKX6-1 in early endocrine precursors (named NKX6-1 progenitors), NEUROG3-late progenitors, SC-β cells, and SC-EC cells (Figure 3G). In agreement, the dynamic profile of NKX6-1 expression (Figure 3H) started at a high value that corresponds to the trajectories passing through NKX6-1 progenitors, and then a decreasing profile is followed by an expected burst on SC-β and SC-EC cells, corresponding to its influence in endocrine cell-subtype specification.[20] Similar to the *in vivo* situation in the mouse embryonic pancreas, we found high NKX6-1 activity in SC-β and SC-α cells and, in this context, also in SC-EC cells (Figures 3I and 3J). After checking the dynamical correlation of mRNA expression and TF activity (Figure 3K), we consistently predicted NKX6-1 as an activator of the SC-β identity. Also, we provide evidence of the possible role of NKX6-1 protein as an activator of the EC fate specification during pancreatic endocrine differentiation *in vitro* (positive correlation without time lag with the mRNA expression in Figure 3K). In contrast, the high NKX6-1 activity in SC-α cells that negatively correlated with the mRNA expression of the NKX6-1 transcript points to a possible repressor role in the SC-α cells, suggesting a similar function to that reported by Schaffer et al.[34] in the mouse. In contrast, ISL1 is a well-known marker of β cells,[48] and it functions as a regulator of ARX during α cell development.[49] Indeed, Veres et al.[20] reported it as differentially expressed in the SC-β branch (Figure 3G′), and the dynamic expression profile showed an increasing pattern over SC-α and SC-β differentiation trajectories (Figure 3H′). Moreover, the activity profile of the ISL1 motif was higher in the expected populations, SC-α and SC-β cells (Figures 3I′ and 3J′). Also, the dynamic correlation between expression and activity (Figure 3K′) confirmed the expected activator role over SC-α and SC-β trajectories. Altogether, these

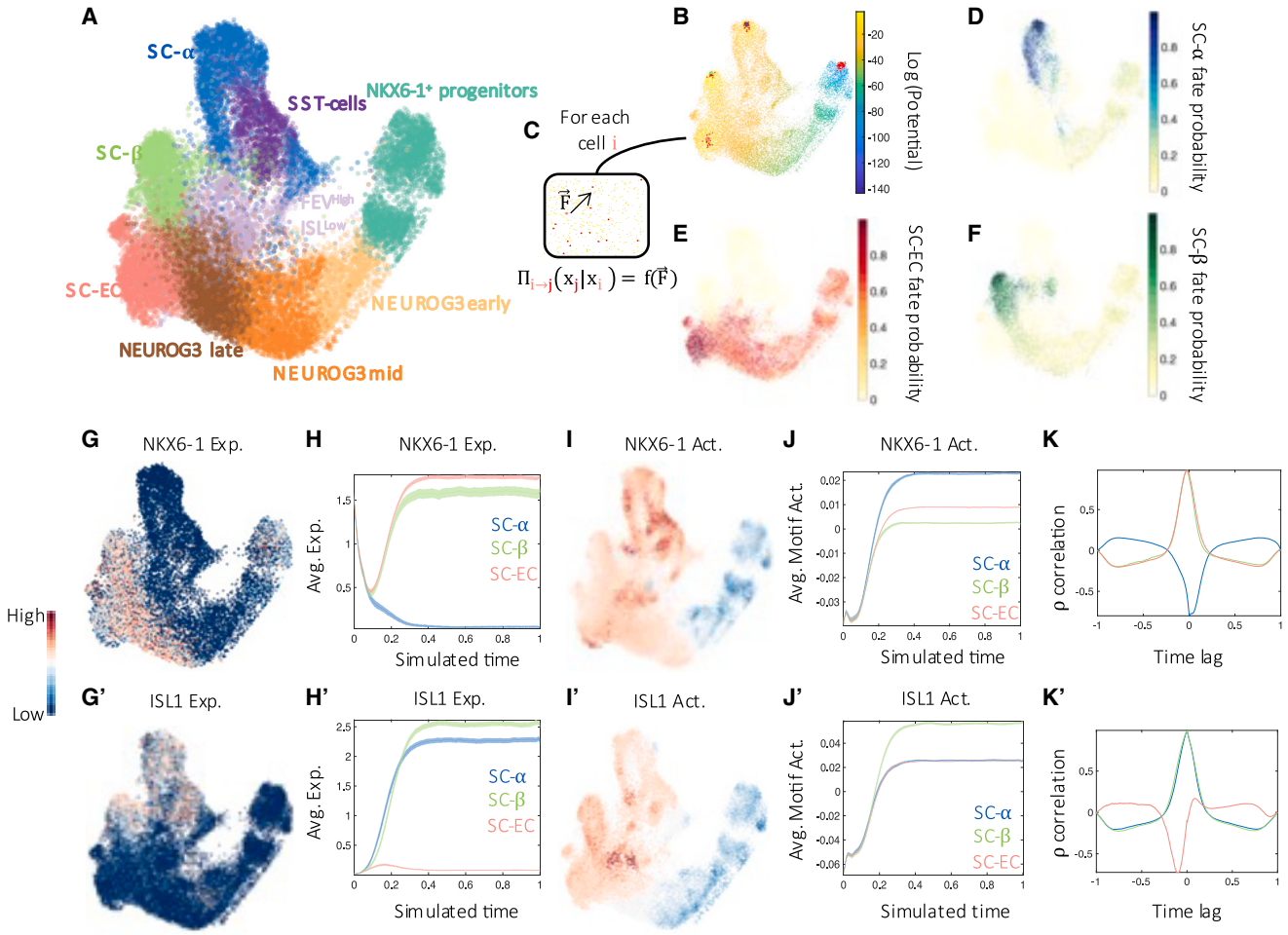


Figure 3. FateCompass retrieves dynamic profiles of *in vitro* β cell differentiation beyond RNA velocity

(A) UMAP plot of 25,299 cells profiled during a 7-day time course at stage 5 of differentiation toward pancreatic β-like cells;[20] colors highlight clustering into nine main cell types.

(B) UMAP plot colored according to the differentiation potential; the gradient goes from NKX6-1$^+$ progenitors to the hormone-producing cell types (SC-α, SC-β, and SC-EC). Red dots represent the possible neighborhood a cell can explore when modeled using a Markov chain.

(C) Propagator (Π_{ij}) of the Markov chain as a function of the differentiation potential ($\vec{F}$). Π_{ij} represents the probability of transitioning to the state *j* when being at the state *i*.

(D–F) Fate probabilities for the SC-α (D), SC-EC (E), and SC-β (F) fates.

(G and G′) UMAP plots with normalized gene expression of known lineage-specific markers; NKX6-1 for SC-β and SC-EC (G) and ISL1 for SC-α and SC-β (G′) cells.

(H and H′) Average gene expression of NKX6-1 and ISL1 over SC-α, SC-β, and SC-EC differentiation trajectories. Shaded areas represent standard errors of the mean.

(I and I′) UMAP plot with TF activity distribution for NKX6-1 (I) and ISL1 (I′).

(J and J′) Average TF activity of NKX6-1 and ISL1 over SC-α, SC-β, and SC-EC differentiation trajectories. Shaded areas represent standard errors of the mean.

(K and K′) Dynamic Pearson correlation between mRNA expression and TF activity over SC-α, SC-β, and SC-EC for NKX6-1 (K) and ISL1 (K′).

See also Figure S6.

results further validate the FateCompass integrative approach to infer regulators in challenging experimental setups.

Differential motif activity analysis predicts driving factors during *in vitro* β cell differentiation protocols

To check FateCompass performance in identifying lineage-specific regulators on the *in vitro* β cell differentiation dataset, we applied the differential TF activity analysis. We identified 126 differentially active TFs (Table S2), 14 for the three endocrine lineages (Figure 4A), 14 for both SC-β and SC-α (Figure 4B), 15 for SC-β and SC-EC (Figure 4C), 10 for SC-α and SC-EC, 20 were SC-β specific (Figure 4D), 25 were SC-EC specific (Figure 4E), and 28 were SC-α specific (Figure 4F). Interestingly, the TF CDX2 was differentially active for SC-β, SC-EC, and SC-α; this finding was puzzling because CDX2 is well known for its role in intestinal specification of the gut endoderm during development.[50] Different organs, such as the stomach, pancreas, liver, and intestine, derive from the gut endoderm during mammalian

CellPress
OPEN ACCESS

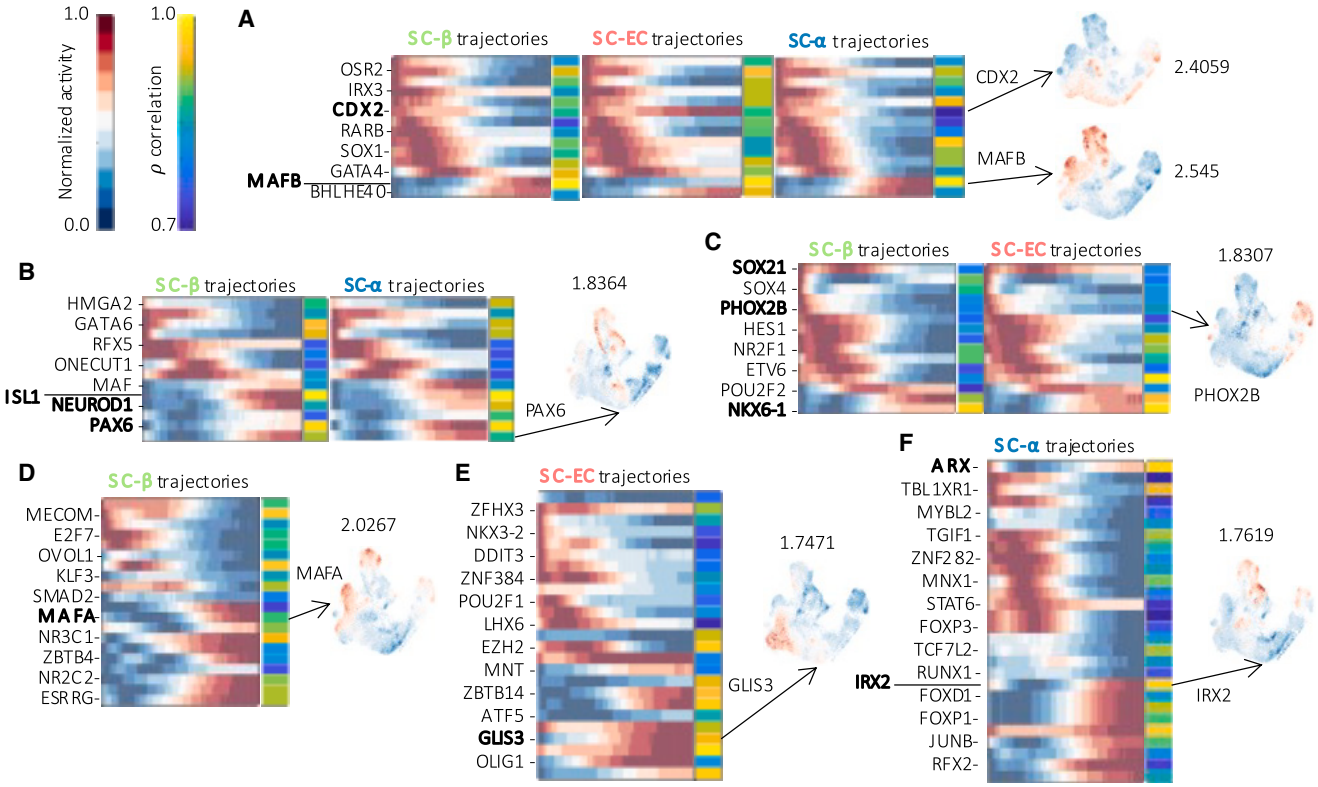


Figure 4. Differential TF activity analysis during *in vitro* pancreatic β cell differentiation

Heatmaps showing the average TF activity over differentiation trajectories and dynamic Pearson correlation. Not all TF names are shown (see Table S2 for full list). Right: activity distribution on the UMAP embedding of selected examples and respective *Z* score. TFs are sorted according to which was active first over the respective trajectory.

(A) TFs predicted as drivers of the three lineages: SC-β, SC-EC, and SC-α.

(B) TFs predicted as drivers of both SC-β and SC-α fates.

(C) TFs predicted as drivers of both SC-β and SC-EC fates.

(D) TFs predicted as SC-β specific.

(E) TFs predicted as SC-EC specific.

(F) TFs predicted as SC-α specific.

See also Figures S7 and S8.

development; particularly, the repression of CDX2 by the positive cross-regulatory loop between PDX1 and SOX9 promotes the expression of pancreas-specific factors.[51] Thus, our predictions suggest that, during *in vitro* differentiation, the early endodermal progenitors might still be plastic and have the potential to activate other fates that will be repressed on pancreatic commitment; Ramond et al.[52] also found a distinct population characterized by CDX2 when comparing pancreatic cells generated *in vitro* with *in vivo* fetal cells. Of note, CDX2 activity is high at the beginning of the three endocrine fates trajectories, then it decreased in SC-β and SC-α cells while remaining high in SC-EC cells, which resemble intestinal EC cells (Figure 4A). For its part, MAFB was also identified for the three endocrine fates. However, its activity was higher through SC-β and SC-α trajectories (Figure 4A); this predicted behavior agrees with the known role of MAFB during islet α and β cell development.[53]

FateCompass predicted some factors to be specific for two lineages simultaneously. As expected, ISL1 and NEUROD1 were classified as SC-β and SC-α specific (Figure 4B).[35,48,49] Contrary to what we observed in the mouse *in vivo*, where NEUROD1 and NKX2-2 had a similar activity profile, in this case, during human *in vitro* β cell differentiation, the activity of NKX2-2 was restricted to NEUROG3 endocrine progenitors and SC-EC cells (Figure S7). This observation points to differences in the regulatory programs of endocrine cell differentiation in mice versus humans. PAX6 was also identified as SC-β and SC-α specific (Figure 4B). Previous chromatin analysis and short hairpin RNA (shRNA)-mediated gene suppression experiments showed that PAX6 has a key role in the identity and function of β cells by activating specific markers and repressing alternative islet genes. Interestingly, using RNA-seq and luciferase assay, the authors found that PAX6 represses NKX2-2.[54] We observed mutually exclusive behavior for PAX6 and NKX2-2 activities (Figure S7), which supports antagonistic roles for PAX6 and NKX2-2. Regarding the factors classified as SC-β and SC-EC specific (Figure 4C), we found NKX6-1, supporting our previous observation of the possible activator role of NKX6-1 for both the pancreatic β-like cells and SC-EC cells. Along the same line, a recent

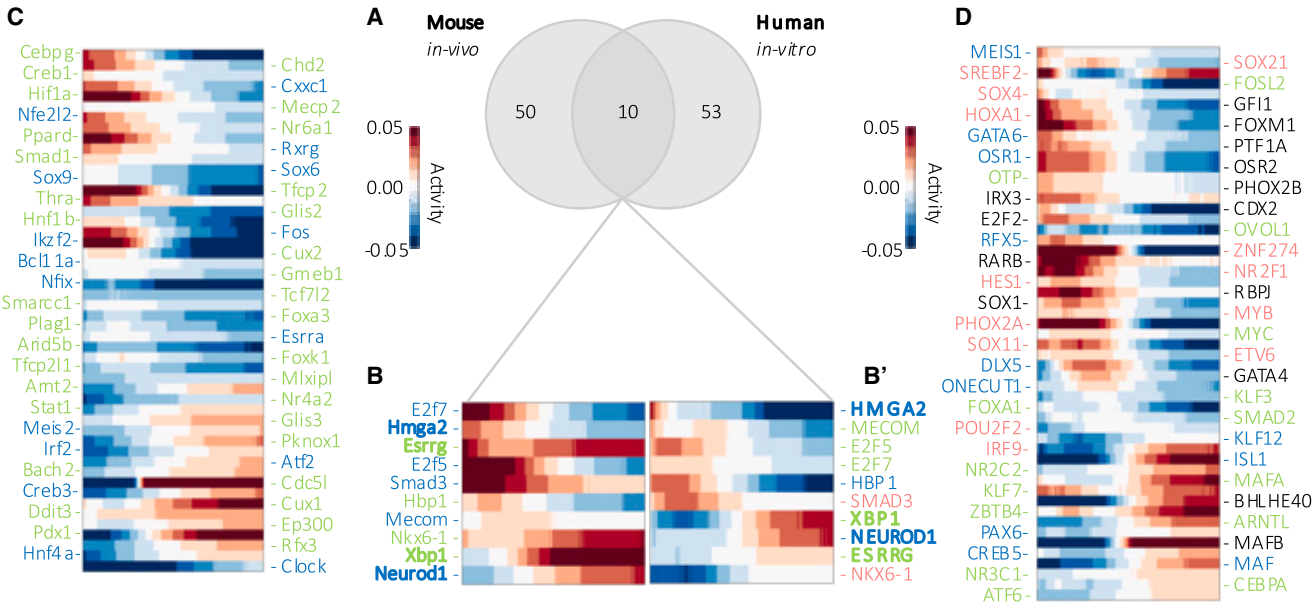


Figure 5. Cross-species comparison of the β-specific factors

(A) Venn diagram showing the number and overlap of β-specific TFs predicted by FateCompass.

(B and B′) Heatmaps showing the average TF activity over β trajectories of the 10 overlapping factors in the mouse *in vivo* system (B) and in the human *in vitro* system (B′).

(C) Heatmaps showing the average TF activity over β trajectories of the 50 mouse *in vivo*-specific factors; TFs are sorted according to which was active first.

(D) Heatmaps showing the average TF activity over β trajectories of the 53 human *in vitro*-specific factors, TFs are sorted according to which was active first. Color code for the name of the TFs in the heatmaps: green for the β specific; blue for the β and α specific; pink for the β and EC specific; and black for the β, α, and EC specific.

See also Figure S9.

study found an enrichment of the NKX6-1 motif on EC-like cells using single-cell ATAC-seq.[55] Notably, we identified PHOX2B, which belongs to the same motif family as LMX1A (Figure S8); LMX1A has a known role as a regulator of the EC fate in the adult small intestine downstream of NKX2-2.[56] Thus, our findings suggest that promoters with the binding site for PHOX2B/LMX1A are, on average, highly expressed on the SC-β and SC-EC trajectories. Whether the differentiation of SC-EC cells has similar regulatory mechanisms to those in the murine small intestine remains largely elusive.

To learn more about SC-β, SC-α, and SC-EC development, we focused on FateCompass predictions for each lineage. MAFA, a hallmark of β cells,[57] was among the SC-β-specific factors (Figure 4D). GLIS3, which we previously found as a direct target of NEUROG3 in pancreatic endocrine progenitor (PEP) cells,[38] was classified as SC-EC (Figure 4E). Notably, it has been reported both as a β cell marker in the pancreas[39] and as an EC marker in the adult small intestine.[58] Thus, our data might imply a tissue- and cell-dependent role for GLIS3. ARX is the first TF to become highly active during SC-α trajectories, corroborating its function during glucagon-producing cell development in mice.[31] IRX2 was highly active later on during SC-α specification (Figure 4F); importantly, Gage et al.[59] found it was downregulated in human pluripotent stem cell (hPSC)-derived human islet cells lacking ARX,[59] and Schreiber et al.[38] found it as a direct target of NEUROG3. Hence our predictions and previous evidence position IRX2 as a potential α-specific novel TF acting downstream of NEUROG3. Taken together, by integrating dynamics and TF activity inference, FateCompass systematically predicts known and novel potential regulators during a complex differentiation system, highlighting the possible use of the pipeline in the improvement of differentiation protocols.

Comparison between *in vivo* and *in vitro* β cell-specific regulators

The design of stepwise directed differentiation protocols to produce islet-like cells has relied heavily on mouse pancreas developmental biology knowledge. We compared the differentially active TFs involved in β cell specification in mouse *in vivo* and human *in vitro*. Interestingly, we found only around 8% of the differentially active TFs at the intersection (Figure 5A). We reasoned that having few overlapping factors could be because of significant differences at the expression level that translate in different TFs driving such an expression pattern. We performed hierarchical clustering among both populations to test this. We found that mouse β cells are more similar to the rest of murine hormone-producing cells (α, ε, and δ), with a Pearson correlation higher than 0.7 (Figure S9). Although human-derived SC-β cells clustered with most of the endocrine-committed murine cells, their relationship with mouse β cells was not that high, Pearson correlation of 0.6 (Figure S9).

Next, we focused on the common TFs to see whether there were relative time-specific profiles. To that end, we sorted the TFs according to which was active first for each system

independently (Figures 5B and 5B′). Estrogen-related receptor γ (Essrg) had a different dynamic profile despite being β specific in both systems. Although it was constantly highly active in the mouse *in vivo* and appeared third on the dynamic ranking, it had a slowly increasing profile in the human *in vitro* and showed next to last. This result was striking because Esrrg is a hallmark of adult and not developing β cells, with a known function for metabolic maturation.[60] Then, our prediction might indicate that Esrrg has a stage-dependent role that remains to be explored in the mouse embryo. Notably, Neurod1 was β and α specific, appearing at later stages in both datasets (Figures 5B and 5B′), indicating a similar role in both systems. Similarly, Nkx6-1 was a common factor that became active progressively during the β-differentiation trajectories. Of note, NKX6-1 was also involved in the EC cell-subtype specification in the human *in vitro* dataset, which raises flags about its organism-dependent role and the different programs it is activating.

To unravel possible time-specific regulatory interactions, we plotted the dynamic profiles of the species-specific TFs (Figures 5C and 5D). Regarding the mouse-specific factors (Figure 5C), Sox9 was active early on, supporting its role in inducing Ngn3 in the progenitor cord.[61] In addition, we found Meis2 with an increasing profile toward the end of the β trajectories; previous studies reported it to be enriched in the second wave of murine fetal α cells,[62] but its specific role in β cell differentiation remains unknown. We observed two dynamic waves regarding the human-specific factors (Figure 5D). The first includes factors active early on during the SC-β trajectories with fetal-like functions, such as SOX4, PTF1A, and CDX2; the second with factors active later on during β cell differentiation resembling maturation and maintenance roles, such as MAFA and PAX6. This suggests that the human *in vitro* regulatory programs differentiate β-like cells activating adult-like factors to produce insulin-responsive cells. Interestingly, FateCompass identified GATA4 as human specific; this factor represents a well-known human-mouse difference. Indeed, its expression is delayed during human development, appearing simultaneously as PDX1.[63] Altogether, these comparisons showed that significant differences in gene expression profiles lead to inferring different lineage-specific TFs.

FateCompass guides hypothesis generation to understand SC-EC cell-fate determination during *in vitro* β cell differentiation protocols

FateCompass identified NKX6-1 as a regulator of the SC-EC fate during human *in vitro* β cell differentiation, which suggests this TF activates different programs in an organism-dependent manner. Thus, we looked at the targets of the NKX6-1 motif in the mouse and the human (Figure 6A). Remarkably, CDX2, a small intestinal epithelial marker,[64] and NKX2-2, a well-documented factor for having a role in developing endocrine cells in the small intestine and the pancreas,[56,65–67] are human-specific targets. Hence, we hypothesized that CDX2 and NKX2-2 drive the specification of the SC-EC cells downstream of NKX6-1 during human *in vitro* β cell differentiation (Figure 6B).

To test this hypothesis (Figure 6B), we lineage-traced NEUROG3-NKX6-1-positive cells at the PEP stage and examined the heterogeneity of this population. First, we generated a human induced pluripotent stem cell (hiPSC) line where NEUROG3 is fused to a cleavable mCherry reporter and NKX6-1 to a GFP reporter (NKX6.1-GFP/NEUROG3-HA-mCherry) (Figure 6C). Next, we differentiated the NKX6.1-GFP/NEUROG3-HA-mCherry hiPSCs toward the islet lineage until the PEP stage at day 13, FACS-sorted the differentiated cells, and sequenced the different fluorescent populations using SortSeq[68] (Figures 6C, S10A, and S10B). The sequenced cells clustered according to the reporter for which they were sorted (Figure S10C). Negative cells expressed gut endoderm markers such as GATA4 and RFX6, GFP^+ cells were more pancreatic epithelial progenitors expressing SOX9 and PTF1A, and $mCherry^+$ cells were endocrine committed cells expressing CHGA, GCG, INS, etc. We focused on the double-positive population, which clearly expressed markers of both SC-β cells and SC-EC cells (Figure S10D). Indeed, after clustering the double-positive population, we identified two groups, SC-β and SC-EC precursors (Figures 6D, S10D, and S10E), confirming that both cell types differentiate from similar progenitor cells. Furthermore, we computed TF activities to see the potential regulators of each fate. We found NKX6-1 activity in both SC-β precursors and SC-EC precursors; NKX2-2 followed a similar pattern, whereas CDX2 activity was higher only in SC-EC precursors (Figure 6E). SC-β-known drivers, such as PDX1, PAX6, and NEUROD1, had high activity on the SC-β precursors. Likewise, the well-described EC markers LMX1A and FEV were highly active in the SC-EC precursors; interestingly, GLIS3, a factor that FateCompass predicted to be a driver of the SC-EC fate, also had high activity on the SC-EC precursor cells (Figure 6E).

Taken together, by exploring the behavior of common TFs and digging further into the transcriptional interactions of NKX6-1, we opened the question of whether the SC-EC differentiation is regulated by the TFs NKX2-2 and CDX2 acting downstream of NKX6-1. We tested this hypothesis by lineage-tracing NEUROG3/NKX6-1-positive cells and found a population of precursors with clear SC-β and SC-EC characteristics. Interestingly, CDX2 activity was high exclusively on the SC-EC-precursors, indicating that promoters with a binding site for CDX2 were, on average, highly expressed in this population. This type of hypothesis merits further investigation to identify key targets to improve β cell differentiation protocols.

DISCUSSION

Here, we have introduced FateCompass, an integrative workflow that robustly estimates time- and lineage-specific TFs. FateCompass pipeline integrates a flexible framework to infer gene expression dynamic profiles with a linear model of gene regulation based on interactions between TFs and promoters to predict regulators implicated in fate choice during development in different contexts (*in vivo* and *in vitro*), across sequencing platforms (10X and InDrops) and across organisms (mouse and human). We designed an innovative differential TF activity analysis that considers the significance of the TF to explain the variability of the linear model of gene regulation, the change of the regulatory activity throughout the cell-fate decision process, and the dynamical correlation of the TF activity with the TF mRNA level. We tested FateCompass with a ground-truth system generated using the simulation engine

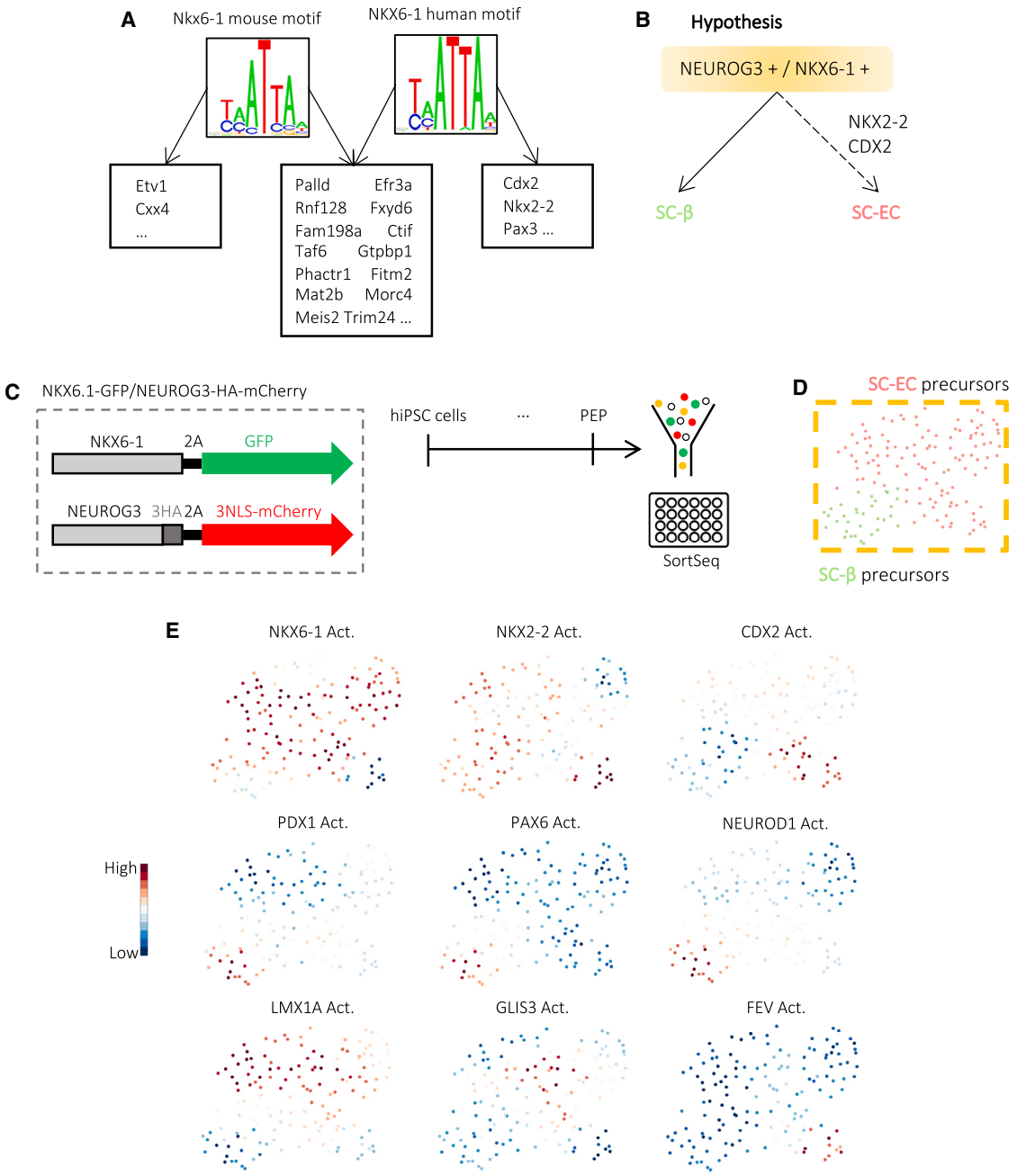


Figure 6. FateCompass guides hypothesis generation to understand SC-EC cell-fate determination during *in vitro* β cell differentiation protocols

(A) Nkx6-1 common and diverging targets for the mouse and human motif.

(B) Diagram of the hypothesis generated using FateCompass predictions.

(C) Scheme of the experimental design to test the generated hypothesis. NKX6-1-GFP/NEUROG3-HA-mCherry cell line to lineage-trace the NEUROG3^{+} and NKX6-1^{+} population. Cell differentiation, sorting, and sequencing strategy.

(D) UMAP plot of 150 cells from the double-positive population.

(E) UMAP plot with TF activity profile of lineage-specific regulators.

See also Figure S10.

Dyngen[28] and validated its capabilities. Applied to pancreatic islet cell-subtype specification, we predicted time- and fate-specific known and novel TFs; the former serves as ground truth, whereas the latter represents an advance in the current understanding of the transcriptional interactions underlying endocrine cell differentiation.

In the inference of differentiation trajectories, we assumed, like other studies, that the process of a cell changing states along a trajectory until it reaches a final fate can be understood as a particle diffusing on a volume;[8,69] but, unlike them, we infused the direction of the differentiation as a drift to bias the transition probabilities. When RNA velocity profiles are robust, FateCompass uses them to direct the edges of the Markov chain. On this line, similar existing methods[29] did not recover glucagon-producing α cells as a terminal state. During development, when the starting cell and the final fates are clear, and the RNA velocity profiles are inconclusive, FateCompass infers differentiation trajectories beyond RNA velocity, biasing the transition probabilities using the gradient of differentiation potential from the starting cells toward the final states. We validated this approach using a dataset from an *in vitro* differentiation toward β-like cells, where we accurately recovered that β-like cells differentiate from NEUROG3-late progenitors, whereas α-like cells start to differentiate from NEUROG3-early progenitors.[20,46,47]

FateCompass uses TFs as the leading players in the gene regulation model; they are well known for their direct role in gene-specific transcriptional regulation; hence, they are commonly used as readouts of pathway activities.[4] Other approaches attempting TF activity inference from transcriptomic data, both bulk and single cell, do not consider their dynamic nature.[2,15,16,70] Some studies have based their predictions merely on correlations between mRNA level of the TF and expressed genes.[16,70] Other more advanced studies rely on known regulons and inferred TF activities using the correlation of the mRNA level of the TF and the group of genes that it can potentially regulate, based on the presence of binding sites on a given regulatory region.[2,15] In contrast with the previously cited methods, ISMARA, initially developed for bulk RNA-seq data, does not rely on correlations; it modeled the expression levels as a linear combination of TF binding-site predictions and unknown TF activities.[14] Here, we extended the use of ISMARA to single-cell transcriptomics. The original ISMARA model proposed a symmetric Gaussian to avoid overfitting; however, in that way, all the parameters are regularized equally, which might not be suitable on single-cell data, where different regions of the manifold represent, usually, different phenotypes associated with changing TF activities. FateCompass addresses the multicollinearity problem in linear regression using a newly developed regularization approach. We defined a data-diffusion-based regularization, where we enforced the smoothness and stability of the inferred activities across cells. This approach has been widely used for imputation methods.[27]

In the embryonic mouse pancreas (*in vivo*) dataset, FateCompass recovered well-documented regulatory interactions, such as the antagonistic role of Arx and Nkx6-1[34,35] and the cell-type-dependent interaction between Neurod1 and Nkx2-2.[35] We also identified putative driver factors with interesting known roles in β cell function[71] and a circadian pattern in α cells,[44] but whether they are also involved in endocrine cell differentiation remains to be tested. A recently published study aimed to identify lineage-specific drivers during pancreatic endocrine differentiation, where they focused on the differential gene expression of TFs, and also identified some of the known regulators.[72] In contrast, we steered on regulation principles by considering interactions between TFs and promoters, which provide a more accurate picture of gene-specific regulation. We anticipate applying our framework to guide experiment design to test the function of the identified lineage-specific factors. In the stem-cell-derived human pancreatic islet (*in vitro*) dataset, we identified TFs acting early on during the differentiation trajectories that confirmed the plasticity of the less mature cells in differentiation protocols.[52,64] Moreover, we retrieved cell-type-specific drivers for the pancreatic endocrine cells and the intestinal-like EC population. Importantly, our differential TF activity analysis pinpointed NKX6-1 as a potential regulator of the SC-EC cells; we generated the hypothesis that NKX2-2 and CDX2 drive the SC-EC fate downstream of NKX6-1, leveraging FateCompass predictions. Further experimental validation supported the role of CDX2 as a potential regulator of SC-EC cells; this observation agrees with recent studies on stem-cell-derived pancreatic islets.[73] Comparing the *in vivo* and *in vitro* predictions for the β cell trajectories, we found known TFs at the intersection, such as NEUROD1 and NKX6-1;[34,35] also, we were able to recapitulate mouse- and human-specific differences.[63] In summary, we foresee the use of FateCompass to generate hypotheses targeted to provide means to optimize differentiation protocols.

The fast evolution of high-throughput methods and generation of large-scale datasets impose the need for robust computational approaches not only to characterize genome-wide patterns but also to extract information and mechanistically model biological phenomena that, in the end, will provide predictions aimed at increasing the current state of the knowledge. As with any inference method, aspiring to reconstruct the exact interactions underlying a complex biological process, such as endocrine cell formation, is a futile task. In this study, we rely on computationally predicted regulatory sites, summarized in a binding-site matrix; this represents a bias on the structure of the gene-regulatory network. Moreover, we are considering only interactions between TFs and promoters, and it is well known that some essential regulatory interactions occur at distal regulatory sites.[74] We have designed our pipeline such that the limitations mentioned above could be addressed by extending the binding-site matrix; indeed, with the advent of single-cell chromatin accessibility assays such as single-cell ATAC-seq, FateCompass could be extended to consider binding sites only in open regions and to perform *de novo* binding-site predictions at proximal and distal *cis*-regulatory regions. However, including this information in the binding-site matrix, as well as incorporating appropriate regularization terms to solve the linear model, may not be a straightforward task. As a framework for identifying lineage-specific drivers, we forecast FateCompass to be used as a tool to explore scRNA-seq data, guide hypothesis generation, and direct experiment design. Further experimental validation of the generated hypothesis will increase the current understanding of a given process and provide means to improve existing translational experiments aimed at cell therapy.

Limitations of the study

FateCompass builds differentiation trajectories in a nearest-neighbor graph computed in a reduced gene expression space. Except for PCA, commonly used dimensionality reduction

methods such as UMAP should be treated carefully, given the lack of biological interpretability. Moreover, when using differentiation potential as a drift for guiding the transition probabilities, FateCompass depends on the input of prior biological knowledge, which is difficult to provide in some cases, such as cancer studies. FateCompass estimates TF activities considering binding sites in the promoter regions of the target genes, an extension considering long-range interactions will provide a more accurate picture of the regulatory landscape. To that end, we designed FateCompass so that the binding-site matrix could be updated. Finally, the pipeline depends on input parameters that must be hand tuned. However, we provide detailed tutorials to guide the user on the parameter setup.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHODS DETAILS
  - FateCompass pipeline
  - Simulated data from dyngen
  - Mouse *in-vivo* dataset from endocrinogenesis
  - Human *in-vitro* dataset from differentiation toward β-like cells
  - Lineage tracing of NEUROG3+/NKX6-1+ cells
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2023.100512.

ACKNOWLEDGMENTS

We thank Erik van Nimwegen and Gioele la Manno for stimulating discussions on pipeline development and Leland McInnes for valuable insights on UMAP implications, interpretation, and limitations. We thank Christian Honoré (Novo Nordisk A/S) for providing the NKX6.1-GFP iPSC line[75] through the IMI/EU-sponsored StemBANCC consortium. We would further like to thank the IGBMC Flow Cytometry and Cell Culture facilities for cell sorting and the members of the Gradwohl lab and the Molina lab. Research in the Gradwohl and Molina labs is supported by a grant from ANR (ANR-21-CE14-0003-01), IdEx Unistra (ANR-10-IDEX-0002), SFRI-STRAT'US project (ANR 20-SFRI-0012), and EUR IMCBio (ANR17-EURE-0023). S.J. is an IGBMC international PhD program fellow supported by LabEx INRT.

AUTHOR CONTRIBUTIONS

Conceptualization, S.J., G.G., and N.M.; methodology, S.J., V.S., R.M., G.G., and N.M.; software, S.J. and N.M.; formal analysis, S.J., V.S., G.G., and N.M.; investigation, S.J., V.S., R.M., G.G., and N.M.; resources, V.S., G.G., and N.M.; data curation, S.J.; writing – original draft, S.J.; writing – review & editing, S.J., V.S., G.G., and N.M.; visualization, S.J.; supervision, G.G. and N.M.; project administration, S.J., G.G., and N.M.; funding acquisition, G.G. and N.M.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: September 6, 2022
Revised: March 23, 2023
Accepted: June 1, 2023
Published: June 22, 2023

REFERENCES

1. Lee, T.I., and Young, R.A. (2013). Transcriptional regulation and its misregulation in disease. Cell *152*, 1237–1251. https://doi.org/10.1016/j.cell.2013.02.014.
2. Garcia-Alonso, L., Holland, C.H., Ibrahim, M.M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. *29*, 1363–1375. https://doi.org/10.1101/337915.
3. Alon, U. (2006). An Introduction to Systems Biology: Design Principles of Biological Circuits Second, C. Press, ed. (Taylor & Francis Group).
4. Trescher, S., Münchmeyer, J., and Leser, U. (2017). Estimating genome-wide regulatory activity from multi-omics data sets using mathematical optimization. BMC Syst. Biol. *11*, 1–18. https://doi.org/10.1186/s12918-017-0419-z.
5. Rostom, R., Svensson, V., Teichmann, S.A., and Kar, G. (2017). Computational approaches for interpreting scRNA-seq data. FEBS Lett. *591*, 2213–2225. https://doi.org/10.1002/1873-3468.12684.
6. Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554. https://doi.org/10.1038/s41587-019-0071-9.
7. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386. https://doi.org/10.1038/nbt.2859.
8. Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F., and Theis, F.J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nat. Methods *13*, 845–848. https://doi.org/10.1038/nmeth.3971.
9. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498. https://doi.org/10.1038/s41586-018-0414-6.
10. Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nat. Biotechnol. *38*, 1408–1414. https://doi.org/10.1038/s41587-020-0591-3.
11. Bergen, V., Soldatov, R.A., Kharchenko, P.V., and Theis, F.J. (2021). RNA velocity—current challenges and future perspectives. Mol. Syst. Biol. *17*, e10282–e10289. https://doi.org/10.15252/msb.202110282.
12. Qiu, X., Zhang, Y., Martin-Rufino, J.D., Weng, C., Hosseinzadeh, S., Yang, D., Pogson, A.N., Hein, M.Y., Hoi Joseph Min, K., Wang, L., et al. (2022). Mapping transcriptomic vector fields of single cells. Cell *185*, 690–711.e45. https://doi.org/10.1016/j.cell.2021.12.045.
13. Byrnes, L.E., Wong, D.M., Subramaniam, M., Meyer, N.P., Gilchrist, C.L., Knox, S.M., Tward, A.D., Ye, C.J., and Sneddon, J.B. (2018). Lineage dynamics of murine pancreatic development at single-cell resolution. Nat. Commun. *9*, 3922–4017. https://doi.org/10.1038/s41467-018-06176-3.

14. Balwierz, P.J., Pachkov, M., Arnold, P., Gruber, A.J., Zavolan, M., and Van Nimwegen, E. (2014). ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. Genome Res. *24*, 869–884. https://doi.org/10.1101/gr.169508.113.

15. Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086. https://doi.org/10.1038/nmeth.4463.

16. Ding, J., Aronow, B.J., Kaminski, N., Kitzmiller, J., Whitsett, J.A., and Bar-Joseph, Z. (2018). Reconstructing differentiation networks and their regulation from time series single-cell expression data. Genome Res. *28*, 383–395. https://doi.org/10.1101/gr.225979.117.

17. Holland, C.H., Tanevski, J., Perales-Patón, J., Gleixner, J., Kumar, M.P., Mereu, E., Joughin, B.A., Stegle, O., Lauffenburger, D.A., Heyn, H., et al. (2020). Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. Genome Biol. *21*, 1–19. https://doi.org/10.1186/s13059-020-1949-z.

18. Gradwohl, G., Dierich, A., LeMeur, M., and Guillemot, F. (2000). Neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. Proc. Natl. Acad. Sci. USA *97*, 1607–1611. https://doi.org/10.1073/pnas.97.4.1607.

19. Bastidas-Ponce, A., Tritschler, S., Dony, L., Scheibner, K., Tarquis-Medina, M., Salinno, C., Schirge, S., Burtscher, I., Böttcher, A., Theis, F.J., et al. (2019). Massive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. Development *146*, dev173849. https://doi.org/10.1242/dev.173849.

20. Veres, A., Faust, A.L., Bushnell, H.L., Engquist, E.N., Kenty, J.H.-R., Harb, G., Poh, Y.-C., Sintov, E., Gürtler, M., Pagliuca, F.W., et al. (2019). Charting cellular identity during human in vitro β-cell differentiation. Nature *569*, 368–373. https://doi.org/10.1038/s41586-019-1168-5.

21. Brackston, R.D., Lakatos, E., and Stumpf, M.P.H. (2018). Characteristics during cell differentiation. PLoS Comput. Biol. *14*, e1006405.

22. Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. *20*, 1–9. https://doi.org/10.1186/s13059-019-1663-x.

23. Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. Dublin Philos. Mag. J. Sci. *2*, 559–572. https://doi.org/10.1080/14786440109462720.

24. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.03426.

25. Breda, J., Zavolan, M., and van Nimwegen, E. (2021). Bayesian inference of gene expression states from single-cell RNA-seq data. Nat. Biotechnol. *39*, 1008–1016. https://doi.org/10.1038/S41587-021-00875-X.

26. Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676. https://doi.org/10.1016/j.cell.2006.07.024.

27. van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. Cell *174*, 716–729.e27. https://doi.org/10.1016/j.cell.2018.05.061.

28. Cannoodt, R., Saelens, W., Deconinck, L., and Saeys, Y. (2021). Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. Nat. Commun. *12*, 3942–3949. https://doi.org/10.1038/s41467-021-24152-2.

29. Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H.B., et al. (2022). CellRank for directed single-cell fate mapping. Nat. Methods *19*, 159–170. https://doi.org/10.1038/s41592-021-01346-6.

30. Johansson, K.A., Dursun, U., Jordan, N., Gu, G., Beermann, F., Gradwohl, G., and Grapin-Botton, A. (2007). Temporal control of Neurogenin3 activity in pancreas progenitors reveals competence windows for the generation of different endocrine cell types. Dev. Cell *12*, 457–465. https://doi.org/10.1016/j.devcel.2007.02.010.

31. Collombat, P., Mansouri, A., Hecksher-Sørensen, J., Serup, P., Krull, J., Gradwohl, G., and Gruss, P. (2003). Opposing actions of Arx and Pax4 in endocrine pancreas development. Genes Dev. *17*, 2591–2603. https://doi.org/10.1101/gad.269003.

32. Sander, M., Sussel, L., Conners, J., Scheel, D., Kalamaras, J., Dela Cruz, F., Schwitzgebel, V., Hayes-Jordan, A., and German, M. (2000). Homeobox gene Nkx6.1 lies downstream of Nkx2.2 in the major pathway of β-cell formation in the pancreas. Development *127*, 5533–5540. https://doi.org/10.1242/dev.127.24.5533.

33. Courtney, M., Gjernes, E., Druelle, N., Ravaud, C., Vieira, A., Ben-Othman, N., Pfeifer, A., Avolio, F., Leuckx, G., Lacas-Gervais, S., et al. (2013). The inactivation of Arx in pancreatic α-cells triggers their neogenesis and conversion into functional β-like cells. PLoS Genet. *9*, 110039344–18. https://doi.org/10.1371/journal.pgen.1003934.

34. Schaffer, A.E., Taylor, B.L., Benthuysen, J.R., Liu, J., Thorel, F., Yuan, W., Jiao, Y., Kaestner, K.H., Herrera, P.L., Magnuson, M.A., et al. (2013). Nkx6.1 controls a gene regulatory network required for establishing and maintaining pancreatic beta cell identity. PLoS Genet. *9*, e1003274. https://doi.org/10.1371/journal.pgen.1003274.

35. Mastracci, T.L., Anderson, K.R., Papizan, J.B., and Sussel, L. (2013). Regulation of Neurod1 contributes to the lineage potential of Neurogenin3+ endocrine precursor cells in the pancreas. PLoS Genet. *9*, 110032788–14. https://doi.org/10.1371/journal.pgen.1003278.

36. Peiris, H., Park, S., Louis, S., Gu, X., Lam, J.Y., Asplund, O., Ippolito, G.C., Bottino, R., Groop, L., Tucker, H., and Kim, S.K. (2018). Discovering human diabetes-risk gene function with genetics and physiological assays. Nat. Commun. *9*, 3855–3911. https://doi.org/10.1038/s41467-018-06249-3.

37. Cao, W., Guo, J., Wen, X., Miao, L., Lin, F., Xu, G., Ma, R., Yin, S., Hui, Z., Chen, T., et al. (2016). CXXC finger protein 1 is critical for T-cell intrathymic development through regulating H3K4 trimethylation. Nat. Commun. *7*, 11687. https://doi.org/10.1038/ncomms11687.

38. Schreiber, V., Mercier, R., Jiménez, S., Ye, T., García-Sánchez, E., Klein, A., Meunier, A., Ghimire, S., Birck, C., Jost, B., et al. (2021). Extensive NEUROG3 occupancy in the human pancreatic endocrine gene regulatory network. Mol. Metabol. *53*, 101313. https://doi.org/10.1016/j.molmet.2021.101313.

39. Kang, H.S., Kim, Y.-S., ZeRuth, G., Beak, J.Y., Gerrish, K., Kilic, G., Sosa-Pineda, B., Jensen, J., Pierreux, C.E., Lemaigre, F.P., et al. (2009). Transcription factor Glis3, a novel critical player in the regulation of pancreatic β-cell development and insulin gene expression. Mol. Cell Biol. *29*, 6366–6379. https://doi.org/10.1128/mcb.01259-09.

40. Gao, N., Lay, J.L., Qin, W., Doliba, N., Schug, J., Fox, A.J., Smirnova, O., Matschinsky, F.M., and Kaestner, K.H. (2010). Foxa1 and Foxa2 maintain the metabolic and secretory features of the mature beta-cell nan. Mol. Endocrinol. https://doi.org/10.1210/me.2009-0513.

41. Soggia, A., Flosseau, K., Ravassard, P., Szinnai, G., Scharfmann, R., and Guillemain, G. (2012). Activation of the transcription factor carbohydrate-responsive element-binding protein by glucose leads to increased pancreatic beta cell differentiation in rats. Diabetologia *55*, 2713–2722. https://doi.org/10.1007/s00125-012-2623-0.

42. Gao, T., McKenna, B., Li, C., Reichert, M., Nguyen, J., Singh, T., Yang, C., Pannikar, A., Doliba, N., Zhang, T., et al. (2014). Pdx1 maintains β cell identity and function by repressing an α cell program. Cell Metabol. *19*, 259–271. https://doi.org/10.1016/j.cmet.2013.12.002.

43. Paul, L., Walker, E.M., Drosos, Y., Cyphert, H.A., Neale, G., Stein, R., South, J., Grosveld, G., Herrera, P.L., and Sosa-Pineda, B. (2016). Lack of Prox1 downregulation disrupts the expansion and maturation of postnatal murine β-Cells. Diabetes *65*, 687–698. https://doi.org/10.2337/db15-0713.

44. Petrenko, V., Saini, C., Giovannoni, L., Gobet, C., Sage, D., Unser, M., Heddad Masson, M., Gu, G., Bosco, D., Gachon, F., et al. (2017). Pancreatic α- and β-cellular clocks have distinct molecular properties and impact on islet hormone secretion and gene expression. Genes Dev. *31*, 383–398. https://doi.org/10.1101/gad.290379.116.

45. Pagliuca, F.W., Millman, J.R., Gürtler, M., Segel, M., Van Dervort, A., Ryu, J.H., Peterson, Q.P., Greiner, D., and Melton, D.A. (2014). Generation of functional human pancreatic β cells in vitro. Cell *159*, 428–439. https://doi.org/10.1016/j.cell.2014.09.040.

46. Petersen, M.B.K., Azad, A., Ingvorsen, C., Hess, K., Hansson, M., Grapin-Botton, A., Honoré, C., and Honore, C. (2017). Single-cell gene expression analysis of a human ESC model of pancreatic endocrine development reveals different paths to β-cell differentiation. Stem Cell Rep. *9*, 1246–1261. https://doi.org/10.1016/j.stemcr.2017.08.009.

47. Weng, C., Xi, J., Li, H., Cui, J., Gu, A., Lai, S., Leskov, K., Ke, L., Jin, F., and Li, Y. (2020). Single-cell lineage analysis reveals extensive multimodal transcriptional control during directed beta-cell differentiation. Nat. Metab. *2*, 1443–1458. https://doi.org/10.1038/s42255-020-00314-2.

48. Ediger, B.N., Du, A., Liu, J., Hunter, C.S., Walp, E.R., Schug, J., Kaestner, K.H., Stein, R., Stoffers, D.A., and May, C.L. (2014). Islet-1 is essential for pancreatic β-cell function. Diabetes *63*, 4206–4217. https://doi.org/10.2337/db14-0096.

49. Liu, J., Hunter, C.S., Du, A., Ediger, B., Walp, E., Murray, J., Stein, R., and May, C.L. (2011). Islet-1 regulates Arx transcription during pancreatic islet α-cell development. J. Biol. Chem. *286*, 15352–15360. https://doi.org/10.1074/jbc.M111.231670.

50. Gao, N., White, P., and Kaestner, K.H. (2009). Establishment of intestinal identity and epithelial-mesenchymal signaling by Cdx2. Dev. Cell *16*, 588–599. https://doi.org/10.1016/j.devcel.2009.02.010.

51. Shih, H.P., Seymour, P.A., Patel, N.A., Xie, R., Wang, A., Liu, P.P., Yeo, G.W., Magnuson, M.A., and Sander, M. (2015). A gene regulatory network cooperatively controlled by Pdx1 and Sox9 governs lineage allocation of foregut progenitor cells. Cell Rep. *13*, 326–336. https://doi.org/10.1016/j.celrep.2015.08.082.

52. Ramond, C., Beydag-Tasöz, B.S., Azad, A., van de Bunt, M., Petersen, M.B.K., Beer, N.L., Glaser, N., Berthault, C., Gloyn, A.L., Hansson, M., et al. (2018). Understanding human fetal pancreas development using subpopulation sorting, RNA sequencing and single-cell profiling. Development *145*, dev165480. https://doi.org/10.1242/dev.165480.

53. Artner, I., Blanchi, B., Raum, J.C., Guo, M., Kaneko, T., Cordes, S., Sieweke, M., and Stein, R. (2007). MafB is required for islet beta cell maturation. Proc. Natl. Acad. Sci. USA *104*, 3853–3858.

54. Swisa, A., Avrahami, D., Eden, N., Zhang, J., Feleke, E., Dahan, T., Cohen-Tayar, Y., Stolovich-Rain, M., Kaestner, K.H., Glaser, B., et al. (2017). PAX6 maintains β cell identity by repressing genes of alternative islet cell types. J. Clin. Invest. *127*, 230–243. https://doi.org/10.1172/JCI88015.

55. Augsornworawat, P., Marquez, E., Maestas, M.M., Ishahak, M., Gale, S.E., Schmidt, M.D., Veronese-Paniagua, D.A., Miller, J.R., Velazco-Cruz, L., and Millman, J.R. (2022). Multiomic profiling defines cell fate plasticity of in vitro-derived islets. Preprint at bioRxiv. https://doi.org/10.1101/2022.02.26.482126.

56. Gross, S., Garofalo, D.C., Balderes, D.A., Mastracci, T.L., Dias, J.M., Perlmann, T., Ericson, J., and Sussel, L. (2016). The novel enterochromaffin marker Lmx1a regulates serotonin biosynthesis in enteroendocrine cell lineages downstream of Nkx2.2. Development *143*, 2616–2628. https://doi.org/10.1242/dev.130682.

57. Hang, Y., and Stein, R. (2011). MafA and MafB activity in pancreatic β cells. Trends Endocrinol. Metabol. *22*, 364–373. https://doi.org/10.1016/j.tem.2011.05.003.

58. Piccand, J., Vagne, C., Blot, F., Meunier, A., Beucher, A., Strasser, P., Lund, M.L., Ghimire, S., Nivlet, L., Lapp, C., et al. (2019). Rfx6 promotes the differentiation of peptide-secreting enteroendocrine cells while repressing genetic programs controlling serotonin production. Mol. Metabol. *29*, 24–39. https://doi.org/10.1016/j.molmet.2019.08.007.

59. Gage, B.K., Asadi, A., Baker, R.K., Webber, T.D., Wang, R., Itoh, M., Hayashi, M., Miyata, R., Akashi, T., and Kieffer, T.J. (2015). The role of ARX in human pancreatic endocrine specification. PLoS One *10*, 1–24. https://doi.org/10.1371/journal.pone.0144100.

60. Yoshihara, E., Wei, Z., Lin, C.S., Fang, S., Ahmadian, M., Kida, Y., Tseng, T., Dai, Y., Yu, R.T., Liddle, C., et al. (2016). ERRγ is required for the metabolic maturation of therapeutically functional glucose-responsive β cells. Cell Metabol., 622–634. https://doi.org/10.1016/j.cmet.2016.03.005.ERR.

61. Shih, H.P., Wang, A., and Sander, M. (2013). Pancreas organogenesis: from lineage determination to morphogenesis. Annu. Rev. Cell Dev. Biol. *29*, 81–105. https://doi.org/10.1146/annurev-cellbio-101512-122405.

62. Szlachcic, W.J., Ziojla, N., Kizewska, D.K., Kempa, M., and Borowiak, M. (2021). Endocrine pancreas development and dysfunction through the lens of single-cell RNA-sequencing. Front. Cell Dev. Biol. *9*, 629212–629222. https://doi.org/10.3389/fcell.2021.629212.

63. Nair, G., and Hebrok, M. (2015). Islet formation in mice and men: lessons for the generation of functional insulin-producing β cells from human pluripotent stem cells Gopika. Curr. Opin. Genet. Dev. *32*, 171–180. https://doi.org/10.1016/j.gde.2015.03.004.Islet.

64. Grainger, S., Savory, J.G.A., and Lohnes, D. (2010). Cdx2 regulates patterning of the intestinal epithelium. Dev. Biol. *339*, 155–165. https://doi.org/10.1016/j.ydbio.2009.12.025.

65. Doyle, M.J., Loomis, Z.L., and Sussel, L. (2007). Nkx2.2-repressor activity is sufficient to specify α-cells and a small number of β-cells in the pancreatic islet. Development *134*, 515–523. https://doi.org/10.1242/dev.02763.

66. Churchill, A.J., Gutiérrez, G.D., Singer, R.A., Lorberbaum, D.S., Fischer, K.A., and Sussel, L. (2017). Genetic evidence that Nkx2.2 acts primarily downstream of Neurog3 in pancreatic endocrine lineage development. Elife *6*, 1–18. https://doi.org/10.7554/eLife.20010.

67. Desai, S., Loomis, Z., Pugh-Bernard, A., Schrunk, J., Doyle, M.J., Minic, A., McCoy, E., and Sussel, L. (2008). Nkx2.2 regulates cell fate choice in the enteroendocrine cell lineages of the intestine. Dev. Biol. *313*, 58–66. https://doi.org/10.1016/j.ydbio.2007.09.047.

68. Muraro, M.J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M.A., Carlotti, F., de Koning, E.J.P., and van Oudenaarden, A. (2016). A single-cell transcriptome atlas of the human pancreas. Cell Syst. *3*, 385–394.e3. https://doi.org/10.1016/j.cels.2016.09.002.

69. Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L., and Pe'er, D. (2019). Characterization of cell fate probabilities in single-cell data with Palantir. Nat. Biotechnol. *37*, 451–460. https://doi.org/10.1038/s41587-019-0068-4.

70. Schacht, T., Oswald, M., Eils, R., Eichmüller, S.B., and König, R. (2014). Estimating the activity of transcription factors by the effect on their target genes. Bioinformatics *30*, 401–407. https://doi.org/10.1093/bioinformatics/btu446.

71. Szabat, M., Page, M.M., Panzhinskiy, E., Skovsø, S., Mojibian, M., Fernandez-Tajes, J., Bruin, J.E., Bround, M.J., Lee, J.T.C., Xu, E.E., et al. (2016). Reduced insulin production relieves endoplasmic reticulum stress and induces β cell proliferation. Cell Metabol. *23*, 179–193. https://doi.org/10.1016/j.cmet.2015.10.016.

72. Duvall, E., Benitez, C.M., Tellez, K., Enge, M., Pauerstein, P.T., Li, L., Baek, S., Quake, S.R., Smith, J.P., Sheffield, N.C., et al. (2022). Single-cell transcriptome and accessible chromatin dynamics during endocrine pancreas development. Proc. Natl. Acad. Sci. USA *119*, 122012671199–12. https://doi.org/10.1073/pnas.2201267119.

73. Zhu, H., Wang, G., Nguyen-Ngoc, K.V., Kim, D., Miller, M., Goss, G., Kovsky, J., Harrington, A.R., Saunders, D.C., Hopkirk, A.L., et al. (2023). Understanding cell fate acquisition in stem-cell-derived pancreatic islets using single-cell multiome-inferred regulomes. Dev. Cell *58*, 727–743.e11. https://doi.org/10.1016/j.devcel.2023.03.011.

74. Alvarez-Dominguez, J.R., Donaghey, J., Rasouli, N., Kenty, J.H.R., Helman, A., Charlton, J., Straubhaar, J.R., Meissner, A., and Melton, D.A. (2020). Circadian entrainment triggers maturation of human in vitro islets. Cell Stem Cell *26*, 108–122.e10. https://doi.org/10.1016/j.stem.2019.11.011.

75. Gupta, S.K., Wesolowska-Andersen, A., Ringgaard, A.K., Jaiswal, H., Song, L., Hastoy, B., Ingvorsen, C., Taheri-Ghahfarokhi, A., Magnusson, B., Maresca, M., et al. (2018). NKX6.1 induced pluripotent stem cell reporter lines for isolation and analysis of functionally relevant neuronal and pancreas populations. Stem Cell Res. *29*, 220–231. https://doi.org/10.1016/j.scr.2018.04.010.

76. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). Scanpy : large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

77. Petukhov, V., Guo, J., Baryawno, N., Severe, N., Scadden, D.T., Samsonova, M.G., and Kharchenko, P.V. (2018). dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. Genome Biol. *19*, 1–16. https://doi.org/10.1186/s13059-018-1449-6.

78. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

79. Waddington, C.H. (1957). The Strategy of the Genes (Routledge).

80. Gillespie, D.T. (1992). Continuous Markov processes. In Markov Processes. An Introduction for Physical Scientists (Academic Press, Inc.), pp. 111–219. https://doi.org/10.1016/b978-0-08-091837-2.50008-9.

81. Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M.I. (2003). An introduction to MCMC for machine learning. Mach. Learn. *50*, 5–43. https://doi.org/10.1023/A:1020281327116.

82. Arnold, P., Erb, I., Pachkov, M., Molina, N., and Van Nimwegen, E. (2012). MotEvo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. Bioinformatics *28*, 487–494. https://doi.org/10.1093/bioinformatics/btr695.

83. Blondel, V.D., Guillaume, J.L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. *2008*, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Sheep anti NEUROG3	R&D systems	Cat#AF3444; RRID:AB_2149527
Chicken anti GFP	Abcam	Cat#ab13970; RRID:AB_300798
Alexa Fluor 647 Mouse anti-Nkx6-1	BD Biosciences	Cat#563338; RRID: AB_2738144
Donkey anti Sheep-Alexa Fluor 647	Jackson ImmunoResearch	Cat#713-605-147; RRID:AB_2340751
Donkey anti Chicken-DyLight 488	Jackson ImmunoResearch	Cat#703-486-155; RRID:AB_2736851
Chemicals, peptides, and recombinant proteins		
TrypLE Select	Thermo Fischer Scientific	Cat#12563011
mTeSR1	STEMCELL Technologies	Cat#85850
Matrigel Growth Factors Reduced	Corning	Cat#356231
Matrigel hESC-qualified	Corning	Cat#354277
DMEM/F12	Thermo Fischer Scientific	Cat#21331020
PBS (without Ca^{2+} and Mg^{2+})	Thermo Fischer Scientific	Cat#14190094
MCDB 131	Thermo Fischer Scientific	Cat#10372019
GlutaMAX	Thermo Fischer Scientific	Cat#35050038
Penicillin-Streptomycin	Thermo Fischer Scientific	Cat#15140130
Sodium bicarbonate	Thermo Fischer Scientific	Cat#25080060
D-Glucose	Sigma	Cat#G7021
Bovine Serum Albumin Fraction V, fatty acid free, Roche	Sigma	Cat#10775835001
ITS-X	Thermo Fischer Scientific	Cat#51500056
Y-27632	STEMCELL Technologies	Cat#72304
Activin A	STEMCELL Technologies	Cat#78001.1
CHIR99021	Axon Medchem	Cat#1386
L-Ascorbic Acid	Sigma	Cat#A4544
FGF7	Peprotech	Cat#100-19
SANT-1	Sigma	Cat#S4572
LDN193189	Stemgent	Cat#04-0074
TBP (PKC act V)	Millipore EMD	Cat#565740
Retinoic Acid	Sigma	Cat#R2625
ZnSO4	Sigma	Cat#Z0251
ALK5i II	ENZO	Cat#ALX 270 445
heparine	Sigma	Cat#H3149
T3	Sigma	Cat#T6397
Donkey serum	Jackson Immunoresearch	Cat#017-000-121
Paraformaldehyde 16%	EUROMEDEX	Cat#15710
Triton X-100	Sigma	Cat#T8787
Critical commercial assays		
384-well plates for scRNA SortSeq, containing ERCC spike-ins, reverse transcription primers and dNTPs, provided by Single Cell Discoveries	Single Cell Discoveries	N/A
Human Stem Cell Nucleofector Kit 2	LONZA	Cat#VPH-5022
12-well plates CellBind	Corning	Cat#3336

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
50 μm nylon mesh NITEX	Dutscher	Cat#074010
50 μM sterile Filcon	BD Biosciences	Cat#340602
Deposited data		
Mouse embryonic pancreatic endocrine differentiation scRNAseq data at E15.5	Bastidas-Ponce et al., 2018[19]	GSE132188
Time course of human *in-vitro* differentiation toward b-like cells scRNAseq data at stage 5	Veres et al., 2019[20]	GSE114412
Raw scRNAseq data at PEP stage from NKX6-1-GFP/NEUROG3-HA-mCherry hiPS cell line	This paper	GSE202092
Experimental models: Cell lines		
Human: NKX6-1-GFP/NEUROG3-HA-mCherry clone#10 hiPS cells	This paper	N/A
Human: NKX6-1-GFP (1a-21) hiPS cells	Provided by C. Honoré (NovoNordisk) IMI/EU sponsored StemBANCC	Gupta et al., 2018[75]
Oligonucleotides		
sgRNA Fwd CACCGTGAAAGGAC CTGTCTGTCGC	This paper	N/A
sgRNA Rev AAACGCGACAGACA GGTCCTTTCAC	This paper	N/A
PCRwt_F CTTTGTCCGGAATCCAGCTG	This paper	N/A
PCRwt_R CCTTACCCTTAGCACCCACA	This paper	N/A
PCRki_F AAAGAGGGATCCTCTGACCCA	This paper	N/A
PCRki_R ACATGAACTGAGGGGACAGG	This paper	N/A
Recombinant DNA		
pX458 plasmid	Addgene	48138
pBS-hNEUROG3-3HA-2A-3NLS-mCherry-pA	This paper	N/A
Software and algorithms		
Dyngen v1.0.5	Cannoodt et al., 2021[28]	https://github.com/dynverse/dyngen
Scanpy v1.8.2	Wolf et al., 2018[76]	https://scanpy.readthedocs.io/en/stable/index.html
scVelo v0.2.4	Bergen et al., 2020[10]	https://scvelo.readthedocs.io/en/stable/
dropEst	Petukhov et al., 2018[77]	https://github.com/kharchenkolab/dropEst
STAR 2.7.9a	Dobin et al., 2013[78]	https://github.com/alexdobin/STAR
FateCompass	This paper	https://github.com/sarajimenez/fatecompass https://zenodo.org/record/7974789
Other		
BD Fortessa LSR II Cell analyser	BD Biosciences	N/A
FACSAria Fusion cell sorter	BD Biosciences	N/A
Nucleofector 2b	AMAXA	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Nacho Molina (molinan@igbmc.fr).

Materials availability

The NKX6-1-GFP/NEUROG3-HA-mCherry iPSC line is available upon request to the Gradwohl lab.

Data and code availability

- Single cell RNA sequencing data of this study have been deposited in the Gene Expression Omnibus (GEO) under accession code GSE202092.
- Scripts to reproduce our analysis and run the FateCompass pipeline are available at https://github.com/sarajimenez/fatecompass.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

The NKX6-1-GFP hiPSC female line (1a-21) in which the NKX6-1 coding region is fused to a T2A-eGFP reporter is described in Gupta et al.[75] and was obtained through the IMI/EU sponsored StemBANCC consortium via Christian Honoré (Novo Nordisk A/S). This line was gene edited to knock in a 3xHA-P2A-3xNLS-mCherry cassette in fusion with NEUROG3 coding sequence. The heterozygous NKX6-1-GFP/NEUROG3-HA-mCherry (clone 10) was used in the present study. The heterozygocity of the NEUROG3 locus was controled by ddPCR (Institut Clinique de la Souris, Illkirch) and sequencing. All hiPSCs were maintained in mTeSR1 medium (STEMCELL Technologies) and cultured on hESC-qualified Matrigel (Corning) coated plates in a humidified incubator at 5% CO_2 at 37°C, with everyday medium change. Cells were routinely passaged at 80% confluency either by manual colony scrapping, or as single cells with TryPLE Select (Thermo Fisher Scientific); in that latter case cells were reseeded in mTeSR1 supplemented with 10 mM Y-27632 (STEMCELL Technologies) for the first 24 h. The hiPSC lines used in this study were negative for mycoplasma tests performed by the IGBMC Cell Culture facility.

METHODS DETAILS

FateCompass pipeline

The FateCompass workflow aims to identify lineage-specific transcription factors (TFs) during a cellular system undergoing differentiation. To mechanistically understand the dynamic transcriptional interactions underlying the cell subtype specification, we reasoned that inherent asynchrony of the cells, coming from single-cell RNA sequencing (scRNAseq) experiments, provides a temporal resolution of how the transcriptome changes during the differentiation process. In addition, we assumed that *cis*-regulatory regions of the expressed genes contain essential information of the TFs that regulate their transcription. Thus, we integrated both state-of-the-art methods and newly developed algorithms in a coherent and flexible pipeline (https://github.com/sarajimenez/fatecompass). The main input of FateCompass is a pre-processed data object (adata), with annotations of at least cell types (`'clusters'`) and optionally the RNA velocity field, and a TFs binding sites matrix, where the columns are the TFs or motifs and the rows are the expressed genes. We suggest performing basic pre-processing using Scanpy and scVelo.[10,76] Importantly, the FateCompass pipeline can be generalized to include epigenetic information coming from chromatin accessibility and interactions between promoters and enhancers by extending the binding site matrix. The FateCompass pipeline consists of three main steps.

i. Retrieve gene expression dynamics of cell differentiation.
ii. Estimate TF activities along the cell-fate decision process.
iii. Identify lineage-specific regulators.

Cell differentiation dynamics from single-cell RNA sequencing

The first step of FateCompass aims to describe the trajectory a cell follows to arrive at its final state. A single cell, whose phenotype is represented by a point in the multidimensional space, will move along a specific trajectory as its transcriptome changes. Considering a regionalized scenario such as the Epigenetic Landscape of Waddington,[79] we reasoned that trajectories converge to end-states that are essentially different from one another; also, that if a cell moving along a specific trajectory is pushed slightly out of its way, then the canalization of the landscape will compensate, and eventually, the cell will arrive in the stable state it would typically have done.[79] The process of a cell changing states along a trajectory until it reaches a final fate can be understood as a particle diffusing on a volume, where the diffusing particle is a single cell. To delineate the differentiation trajectories, we considered a biased diffusion process employing two approaches to model the drift component. The first approach involves a single cell following a random walk under the influence of a vector field, represented by the RNA velocity[9] until it becomes trapped in an attractor. The second approach entails a single cell following a random walk under the influence of a differentiation potential from progenitor cells, defined as sources, toward mature cells, defined as sinks. In this case, sources and sinks are determined using prior biological knowledge.

Nearest neighbor graph representing the phenotypic manifold

To model the differentiation trajectories, we assumed that scRNAseq data is a relevant sample of representative gene expression states that cells visit during the differentiation process. Furthermore, FateCompass restricts possible state changes to those consistent with a *k*-nearest neighbor graph based on cell-cell distances on the gene expression space. Due to the high sparsity and noise in scRNAseq data, finding nearest neighbors directly from raw expression data is likely to accumulate spurious connections and

obscure the structure of the phenotypic landscape. To build the neighbor graph based on solid data trends and define more reliable distance metrics, we used dimensionality reduction approaches such as Principal Component Analysis (PCA) or Uniform Manifold Approximation and Projection (UMAP). PCA is a linear dimensionality reduction algorithm that preservers the global variability of the data[23]; hence, it is broadly used for detecting among-sample heterogeneity. On the other hand, UMAP is a non-linear dimensionality reduction algorithm that estimates the topology of the high dimensional data and uses this information to build a low-dimensional representation that better preserves the local structure of the data over the global variability.[24]

Formally, given a data object with single-cell transcriptomes, the *k*-nearest neighbor graph is built in the reduced space using the Euclidean distance on the D first components of the dimensionality reduction method of choice (`'pca'` or `'umap'`), where $2 < D < K$, and K is the initial number of neighbors used for the basic pre-processing. Then, to build the adjacency matrix (M) the same number of neighbors for each cell (n_neighbors) are kept. These are the default procedures in FateCompass:

```
from fatecompass import graph_fatecompass
graph_fatecompass(adata,mode='velocity',basis='umap',components=10,n_neighbors=10)
```

Alternatively, the graph could be computed using recent methods such as Sanity,[25] which calculate cell-to-cell distances by removing Poissonian noise from scRNAseq data using a rigorous Bayesian approach. When the graph is computed using an alternative approach, the data object should be annotated with the cell-to-cell connectivities.

Modeling transition probabilities using a Markov process

Next, to model transition probabilities we reasoned that single-cell transcriptomics provides a static picture of a time-evolving system whose possible states are represented by points in a manifold, i.e., $X(t) \equiv$ state point of the system at time t. The value of X at some initial time t_0 is fixed, $X(t_0) = x_0$, and for successive instants $t_1, t_2, \ldots, t_n$; where $t_1 < t_2 < \ldots < t_n$; there are n corresponding random states $X(t_1), X(t_2), \ldots, X(t_n)$. To determine how the cells move from progenitors to mature cells, we assumed that they traverse the manifold in small steps under the influence of an external drift in the direction of differentiation, e.g., RNA velocity or differentiation potential from progenitor cells (sources) to mature cells (sinks); in the FateCompass pipeline this is represented by the *mode* parameter, which could be either `'velocity'` or `'potential'`. The transition probabilities can be modeled using a Markov chain to represent cell fate choices in a probabilistic manner as follows[80]:

$$\frac{\partial}{\partial t}p(x,t) = \frac{\partial^2}{\partial x^2}[\mathcal{D}(x,t)p(x,t)] - \frac{\partial}{\partial x}[F(x,t)p(x,t)] \quad \text{(Equation 1)}$$

where the left-hand-side of the equation represents the change with respect to time of the probability of being at the state x at the time t, $p(x,t)$; the first term on the right-hand-side is the flux through the state x due to diffusion, $\mathcal{D}(x,t)$, and the second term on the right-hand-side is the flux through the state due the external drift, $F(x,t)$.

To outline the differentiation trajectories, we assumed that the continuous process described in Equation 1 can be represented in a discrete space where jumps are allowed only between observed states connected by the *k*-nearest graph introduced above. As a result, gene expression dynamics are described by a discrete Markov process on a network. This simplification circumvents the complex problem of inferring the high-dimensional drift field, $F(x, t)$, with only a few thousand observations. The jumps between observed states are then characterized by a propagator $\Pi(x_j|x_i, F(x_i))$, representing the normalized transition probabilities where the function $F(x_i)$ is the state-dependent drift. Below, we introduce the form of the propagator when the drift is derived based on RNA velocity or a differential potential energy (`'velocity'` and `'potential'` mode, respectively).

RNA velocity as driving force

To obtain the transition probabilities using RNA velocity information, we reasoned that the drift directing the transition probabilities in Equation 1 is represented by the direction of the RNA velocity vector.[9] To implement this idea, we further assumed that the drift is a time-independent vector field $V(x)$ locally constant around a given state x, and that the diffusion coefficient $\mathcal{D}$ is homogeneous and constant over time. Under these assumptions, and for a Δt sufficiently small, the solution of Equation 1 can be approximated by the following Gaussian propagator:

$$\Pi(\xi, t|x_0, t_0) = \frac{1}{(2\pi\mathcal{D}\Delta t)^{\frac{1}{2}}}\exp\left(-\frac{(\xi - V(x)\Delta t)^2}{2\mathcal{D}\Delta t}\right) \quad \text{(Equation 2)}$$

where ξ is the distance between the current state and the next possible state, $\xi = |x_t - x_0|$. A row normalization is applied to transform the Gaussian distribution into transition probabilities over the network of discrete observed states:

$$\Pi_{ij} = \frac{1}{z_i(2\pi\mathcal{D}\Delta t)^{\frac{1}{2}}}\exp\left(-\frac{(|x_j - x_i| - V(x_i)\Delta t)^2}{2\mathcal{D}\Delta t}\right) \quad \text{(Equation 3)}$$

with row normalization factors $z_i \equiv \sum_j \Pi_{ij}$. The vector filed $V(x_i)$ are the RNA velocities obtained using scVelo.[10] It is important to note that we used Equation 1 to justify the functional form of the propagator in (3). However, we do not claim that the discretization of the

stochastic process on a network can be derived and proven from Equation 1. As mentioned before, this simplification circumvents the characterization of the force field in continuous space and the resolution of the general Fokker-Planck Equation 1.

We fitted the diffusion coefficient, $\mathcal{D}$, and Δt heuristically based on the number of neighbors. Shortly, we set Δt such that on average the number of nearest neighbors can be reached, and $\mathcal{D}$ such that the average number of connections is twice the number of nearest neighbors. In this way, the distance traveled until the next state is close to the distance to the nearest neighbors, and the explored neighborhood is within the velocity gradient. Therefore, we make sure that the progression of stable-states follows the direction of differentiation.

Potential energy as driving force

When RNA velocity fails, we can alternatively estimate transition probabilities using a differentiation potential energy, provided that prior biological knowledge is available regarding what are the progenitor and mature cells. We reasoned that the regulatory program driving the differentiation process could be phenomenologically modeled as the combination of a repulsive force, which causes a cell to move away from the progenitor state, plus an attractive force, which draws the cell toward its final fate. We further assumed that the strength of these forces should decay with the distance separating the cell from the progenitor and final states, respectively. Then, by defining the set of initial progenitor states (sources) and the final fate states (sinks), we characterized the total force exerted at each observable state i by the following potential energy (W):

$$W_i = -\sum_{k \in \text{sources}} \frac{Q}{[\text{dist}(x_i, x_k)+1]^n} + \sum_{k \in \text{sinks}} \frac{Q}{[\text{dist}(x_i, x_k)+1]^n} \quad \text{(Equation 4)}$$

where the coefficient Q is the strength of the attraction or repulsion; $\text{dist}(x_i, x_k)$ is the shortest network distance between the state i and k, and n is an exponent. Finally, we used a heuristic criterion to set the values of the parameters Q and n based on the distribution profile of the energy landscape. We reasoned that the potential energy gradient should be high enough for the sources to be phenotypically distinct from the sinks; hence, using a value of 100 for |Q| and 0.5 for n guaranteed such a gradient. Having the potential energy for each state, we defined the transition kernel (propagator) as[81]:

$$\Pi_{ij} = q(x_j|x_i)\mathcal{A}(x_i|x_j) + I_G(x_j)r(x_i) \quad \text{(Equation 5)}$$

where the left-hand-side of the equation, Π_{ij}, is the transition probability from x_i to x_j, the first term on the right-hand-side represents the probability of jumping from state x_i to state x_j given by the multiplication of the proposal distribution q and the acceptance distribution $\mathcal{A}$, and the second term on the right-hand-side is the probability of not jumping represented by the rejection distribution r. Having a symmetric proposal distribution, $q(x_j|x_i) = q(x_i|x_j) = {}^1/_{\text{n\_neighbors}}$, the acceptance probability is

$$\mathcal{A}(x_i|x_j) = \min\left\{1, \frac{p(x_j)q(x_i)}{p(x_i)q(x_j)}\right\} = \min\left\{1, \frac{p(x_j)}{p(x_i)}\right\} \quad \text{(Equation 6)}$$

where p is the invariant distribution,

$$p(x_i) = e^{W_i} \quad \text{(Equation 7)}$$

Finally, the rejection distribution reads

$$r(x_i) = \sum_{\text{n\_neighbors}} q(x_j|x_i)(1 - \mathcal{A}(x_i|x_j)) \quad \text{(Equation 8)}$$

Prior biological knowledge

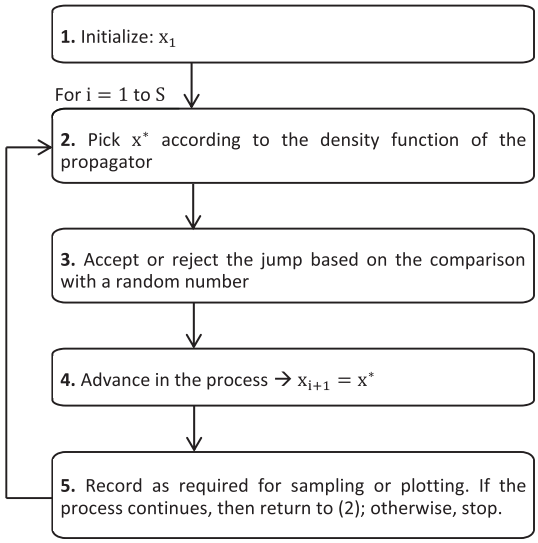
The differentiation potential approach to infer transition probabilities relies on prior biological knowledge on the initial and final fates (sources and sinks, respectively). This information can be provided to FateCompass in three different ways via the `'mode'` parameter.

i. `'cell_types'`: using the clustering annotation to provide the name of the initial and final cell types,
ii. `'marker_genes'`: via marker genes to specify gene names with a clear distinctive expression profile in the initial and final fates, or
iii. `'prior_knowledge_indices'`: by annotating the data object with the indices of the cells that correspond to the initial and final fates.

Stochastic simulations

To describe the time evolution of the previously described Markov processes, we used a numerical approach called Monte Carlo sampling algorithm. Shortly, the idea of a Monte Carlo simulation is to draw an i.i.d. set of samples $\{x^{(i)}\}_1^S$ from a target density

$P(x)$[81]. To estimate the values of X without knowing the density function $P(x)$, we used sampling methods that essentially mimic the real-time evolution of the process. The pseudo-code for the simulations is below:



The default procedures in FateCompass depend on the driving force selected for the transition probabilities. When using RNA velocity as drift they are as follows:

```
from fatecompass import rna_velocity_driven_stochastic_simulations
rna_velocity_driven_stochastic_simulations(adata,root=[],cell_types_key='clusters')
```

on the other hand, when using differentiation potential as drift, FateCompass default procedures are:

```
from fatecompass import diff_potential_driven_stochastic_simulations
diff_potential_driven_stochastic_simulations(adata,cell_types_key='clusters',mode='cell_types',initial_fate=['cluster_initial_cell_type'],final_fate=['cluster_final_cell_type_1',...])
```

The final output of the stochastic simulations' functions is differentiation trajectories stored in the data object as `'states'` and the counting of how many trajectories end in a given final fate (or sink) stored in `'num_trajectories.'`

Average profiles over stochastic trajectories

We used the previously generated S samples (stochastic trajectories) with the following empirical point-mass function to approximate the expected value of the final quantities of interest, mean and standard deviation, for the gene expression profiles ($g(X)$) as follows,

$$\langle g(X)\rangle_S = \frac{1}{S}\sum_{i=1}^{S} g(x_i) \qquad \text{(Equation 9)}$$

$$\text{and } \langle g^2(X)\rangle_S = \frac{1}{S}\sum_{i=1}^{S} g^2(x_i) \qquad \text{(Equation 10)}$$

Of note, the estimates (9) and (10) will become exact in the limit $S \to \infty$. This procedure is effectively implemented in FateCompass:

```
from fatecompass import avg_profiles_over_trajectories
avg_profiles_over_trajectories(adata,cell_types_key='clusters')
```

Fate probabilities

We defined the fate probabilities based on the information of the stochastic trajectories. Thus, we estimated the fate probabilities by counting how often a random walk that visits cell i terminated in any of the terminal index sinks.

```
from fatecompass import fate_probabilities
fate_probabilities(adata,cell_types_key='clusters')
```

Modeling regulatory interactions between transcription factors and cis-regulatory regions

To model the regulatory interactions underlying cell-fate decisions, we considered TFs as the central drivers of transcriptional regulation. TFs are usually designed to transit rapidly between active and inactive molecular states at a rate modulated by a specific

environmental signal. Each active TF can bind the DNA to regulate the rate at which specific target genes are transcribed.[3] This section describes the model we used to infer TF activates in single cells from gene expression profiles.

Binding site matrix

The binding site matrix is one of the main inputs of FateCompass; it must be formatted in a way such that the columns are the TFs or motifs and the rows are the expressed genes; then, each entry represents the binding of a TF in the *cis-regulatory* region of a given gene. As default, we provided TF binding sites predictions reported in[82] and available in the SwissRegulon Portal as "motevo predictions of binding sites" (https://swissregulon.unibas.ch/sr/downloads) for mouse (mm10:FANTOM5) and human (hg19) systems. The authors used a Bayesian framework to estimate the posterior probability that a binding site for a given weight matrix (associated with a motif) occurs in an interval. We summarized the TF binding sites in a matrix of site-counts by summing the posterior probabilities for each motif in the promoter of each gene. We defined a promoter as the TSS +/− 1kb.

Linear model to estimate TF activities

We hypothesize that the expression level of each gene is proportional to the activity of the TFs that can potentially bind to its promoter. Therefore, as in the original framework,[14] we modeled the log-expression level of a gene as the linear combination of motif activities weighted by their number of the binding sites present in its promoter, that is,

$$E_{gc} = \sum_{F} N_{gf} A^{*}_{fc} + \text{noise} \qquad \text{(Equation 11)}$$

where the left-hand-side of the equation, E_{gc}, are the cell- and gene- normalized log-expression values, and the right-hand-side is the summation over the number of TFs, F, of the multiplication of the TF-normalized site-counts, N_{gf}, and the cell-normalized TF activities A^{*}_{fc}. The noise term is related to multiple sources, namely technical, biological, and error in the model. To estimate the unknown TF activities A^{*}_{fc}, we first used minimum norm least-squares solution to linear equations to fit the best estimates of A^{*}_{fc} from Equation 11. Next, to control the model's complexity and avoid overfitting, we applied regularization as explained in the following section.

Regularization using data diffusion

A model's ability to reproduce intricate patterns in data is typically related to its number of parameters and complexity. However, the higher the complexity of a model, the higher the risk of overfitting, i.e., fitting spurious noise in the data leading to poor generalizing performance when applied to new observations. Single-cell transcriptomic data present large technical and biological noise introducing variability in the gene expression profiles across cells that do not reflect true variability in the physiological cellular state. Not accounted for, this variability propagates to the inferred TF activities leading to non-functional cell-to-cell differences in TF activity levels. To control for this, we penalized the model's complexity by introducing a regularization term that enforces smoothness and stability for the fitted activities across cells. We embedded the cells in a low-dimensional manifold that faithfully represents the phenotypic similarities (using a nearest neighbor graph). Next, we imputed a cell's TF activities as the weighted average of the activities across the neighboring cells. This strategy, which is mathematically akin to diffusing heat through the data, has been used to correct for dropout and other noise sources on transcriptomic data.[27] It reads,

$$A^{Reg}_{fc} = (M^{*})^{t} A^{*}_{fc} \qquad \text{(Equation 12)}$$

where A^{Reg}_{fc} are the regularized activities, M^{*} is the adjacency matrix, and A^{*}_{fc} are the maximum-likelihood estimates of the activities. Raising M^{*} to the power of t results in a matrix where each entry represents the probability that a random walk of length t starting at cell i will reach cell j, a process similar to diffusion.[27] Importantly, we want a cell's own estimated activities to have the highest impact on the imputation of its final regularized activities; therefore, our adjacency matrix M^{*} allows for self-loops, and these are the most probable steps in the random walk. Thus,

$$M' = M + 10 * I_c$$

$$M^{*}(i,j) = \frac{M'(i,j)}{\sum_{c} M'(i,c)} \qquad \text{(Equation 13)}$$

where I is the identity matrix of size *C*, and *c* the total number of cells. To find the optimal t, we evaluated the impact of t on the final imputed data. We used an 80/20 cross-validation scheme, where the set of promoters was divided randomly into two sets, the training set containing 80% of all promoters and the test set with the remaining 20%. We used the training set to fit the TF activities and the test set to evaluate the quality of the fit. Then, we choose the value of t that minimizes the mean square error (MSE) between the observed expression levels and those predicted by the model in the test set.

Differential TF activity analysis

To identify key lineage-specific regulators during the cell-fate decision process, we defined a differential TF activity analysis based on the following criteria:(i) TFs with high positive *Z* score, i.e., TFs that significantly varied across cells compared with their estimated errors. (ii) TFs with high variability across the lineage-specific differentiation trajectory. (iii)TFs with a high dynamic correlation between their activity and their own mRNA expression within a specific window of time lags. Finally, to identify the lineage-specific TFs, we performed a filtering using thresholds for each one of the criteria.

Z score

To estimate the importance of a TF, we reasoned that activities that fluctuate the most across conditions should be the more important. Therefore, we used the number of standard-deviations that the activity of TF f is away from its average of zero corrected by the precision of the estimation (error bar), also known as *Z* score (z_f), as an indicator of the importance of each TF

$$z_f = \sqrt{\frac{1}{C}\sum_{C}\left(\frac{\mu_{A_{fc}^{Reg}}}{\delta A_{fc}}\right)^2} \quad \text{(Equation 14)}$$

where C is the number of cells, $\mu_{A_{fc}^{Reg}}$ is the mean of the TF activity distribution, and δA_{fc} is the reliability of the fitting of A_{fc}^{Reg} (error-bar).[14] Since we do not know the posterior distribution of the A_{fc}^{Reg}, there is no analytical way to estimate the standard deviations δA_{fc}. Therefore, we used bootstrapping, as explained below.

Bootstrapping

Important to the concept of bootstrapping is that inference about a population from sample data can be modeled by resampling the sample data and performing inference about a sample from the resampled data. Then, we built the distribution of the estimate for the TF activities using random sampling with replacement following the steps below.

i. The activity of a TF f in cell c is a function of the gene expression on that cell and the binding-site matrix; in other words, $A_{fc}^{Reg} = f(E_{gc}, N_{gf})$. Similar to the cross-validation scheme, here we sampled taking the promoters/genes as observations.
ii. Resample from E_{gc} and N_{gf} taking randomly 80% of the observations. Importantly, the bootstrap resample has the same number of observations as the original data used in the training.
iii. Compute the estimate of A_{fc}^{Reg}.
iv. Repeat (ii) and (iii) a large number of times, B, to get $A_{fc}^{Reg,1}$, $A_{fc}^{Reg,2}$, $A_{fc}^{Reg,3}$, …, $A_{fc}^{Reg,B}$.
v. Use the estimates in (iv) to build the empirical bootstrap distribution of the estimate for the TF activities.
vi. Infer $\mu_{A_{fc}^{Reg}}$ and δA_{fc} from the empirical bootstrap distribution of the estimate.

To set the threshold for the *Z* score, we explored the density distribution and selected the most common value. We have made this flexible such that a more or less stringent cutoff can be chosen.

```
from fatecompass import ksdensity_fatecompass
ksdensity_fatecompass(adata,criterion=['z_score'],cell_types_key='clusters',trajectory=
['name_final_fate'])
```

Variability over time

We seek TFs with a high rate of change in the activity with respect to time, which is, clearly, the definition of the first-order derivative. However, since we do not know the exact distribution of the TF activities, there is no easy way to get the analytical solution. To get a proxy of the rate of change, we estimate the standard deviation over time of the mean activity profile along the simulated trajectories (computed via Equation 9), as follows:

$$Std_f = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left|\widehat{A_{f,i}} - \mu_f\right|^2} \quad \text{(Equation 15)}$$

where n is the number of iterations, $\widehat{A_f}$ is the average activity of TF f over the simulated trajectories, and μ_f is the mean of $\widehat{A_f}$

$$\mu_f = \frac{1}{n}\sum_{i=1}^{n}\widehat{A_{f,i}} \quad \text{(Equation 16)}$$

To set the threshold for the variability over time, Std_f, we explored the density distribution and selected the most common value. We have made this flexible such that a more or less stringent cutoff can be chosen.

```
from fatecompass import ksdensity_fatecompass
ksdensity_fatecompass(adata,criterion=['std_tf_time'],cell_types_key='clusters',trajectory=
['name_final_fate'])
```

Dynamic correlation

Next, we used cross-correlation to identify dynamical correlations between average TF activities and their average mRNA expression along the differentiation trajectories. *Cross-correlation* is defined as a similarity measure between two series as a function of the displacement of one relative to the other.

$$R_{\hat{E}_f\hat{A}_f}(t_1, t_2) \triangleq E\left[\widehat{E}_{f t_1}, \overline{\widehat{A_{f t_2}}}\right] \quad \text{(Equation 17)}$$

the left-hand-side reads as the cross-correlation between times t_1 and t_2 for the average mRNA expression over the simulated trajectories of f, $\widehat{E_f}$, and $\widehat{A_f}$. Next, we converted cross-correlation to Pearson correlation to facilitate the comparisons (1: maximum correlation, 0: no correlation, and −1: maximum anti-correlation).

To set the threshold for the dynamic correlation, we assumed that by having Pearson correlations, any value higher than 0 implies correlation. We reasoned that a value of 0.5 or higher could be considered a default threshold. We have made this flexible such that a more or less stringent cutoff can be chosen.

The thresholds of the three criteria can be summarized using a dictionary data structure:

```
thresholds = {'variability': integer, 'z_score': integer, 'correlation': integer}
```

Finally, to identify the lineage-specific TFs, we performed a filtering using thresholds for each one of the criteria and formatted the results in a data frame data structure:

```
from fatecompass import differential_tf_activity, get_df_differential_tf_activity
differential_tf_activity(adata,cell_types_key='clusters')
df = get_df_differential_tf_activity(adata,fates=['final_fate'],thresholds=thresholds)
```

Simulated data from dyngen

We used synthetic scRNAseq data to test FateCompass performance. We obtained the *in-silico* data from the simulation engine dyngen (https://github.com/dynverse/dyngen), which allowed the generation of a dataset with underlying ground truth for developmental trajectories, cell-specific regulatory network and RNA velocity profiles.[28] We used dyngen v1.0.5, which depends on tidyverse v1.3.2, both implemented in R (v4.2.1), to generate a dataset with 1000 cells, 1035 genes from which 130 were TFs, we used most of the default parameters. Next, we saved the annotated data object using anndata v0.7.5.5 to continue the analysis in python.

Data pre-processing and velocity computation

The data object was already filtered, normalized, and embedded in a low dimensional space. We used scVelo[10] in the dynamical mode with most of the default parameters to compute the RNA velocity profile.

FateCompass specific computations

For downstream analysis of the FateCompass pipeline, we embedded the *in-silico* gene expression data in the PCA space. Next, we computed a neighborhood graph in the reduced gene expression space with k = 10. This setting was the graph structure for the Markov chain operations of the FateCompass pipeline. The edges of the Markov chain were directed using the differentiation potential gradient, where we used as prior biological knowledge the marker genes `['Target62']` and `['C5-TF1','D6-TF1']` which were distinctively expressed in the initial and final populations, respectively. We outlined stochastic trajectories using Monte Carlo Sampling algorithm. We retrieved the binding site matrix from the underlying ground truth interactions of the simulated data, and we formatted it as a data frame where the columns were TFs and rows were genes. Last, the differential TF activity analysis thresholds were: minimum *Z* score of 0.2, minimum standard deviation over trajectories of 0.006, and minimum Pearson correlation of 0.5.

Mouse *in-vivo* dataset from endocrinogenesis

We used scRNAseq data from the developing pancreas during the secondary transition, i.e., from embryonic day 12.5–15.5, published by Bastidas-Ponce et al.,[19] and available in the Gene Expression Omnibus under accession number GSE132188. In particular, we used data from the last experimental time point, embryonic day 15.5 (E15.5). We retrieved an annotated object directly from https://scvelo.org with the expression matrix and the annotations for unspliced/spliced reads using the following command: scvelo.datasets.pancreatic_endocrinogenesis(). Our final subset for Figure 2 contained 3696 cells. We kept the original cluster annotation reported by Bastidas-Ponce et al.,.[19]

Data pre-processing and velocity computation

We used SCANPY[76] and scVelo[10] with most of the default parameters. We filtered out genes with less than 20 counts in both spliced and unspliced layers. Next, we normalized by total counts per cell, log-transformed the data, and kept the top 2000 highly variable genes. We embedded the data in the PCA space and used the top 30 principal components to compute a k-nearest neighbor graph with k = 30. For visualization, we used UMAP embedding with two dimensions with default parameters. To compute RNA velocity, we used scVelo's dynamic model of splicing kinetics.

FateCompass specific computations

For downstream analysis of the FateCompass pipeline, we embedded the mouse *in-vivo* data of gene expression and RNA velocity on ten dimensions in the UMAP space. Next, we computed a neighborhood graph in the reduced gene expression space with k = 10. This setting was the graph structure for the Markov chain operations of the FateCompass pipeline. The edges of the Markov chain were directed using the RNA velocity information and Equation 3. We outlined stochastic trajectories using a Gibbs Sampling algorithm. Last, the differential TF activity analysis thresholds were: minimum *Z* score of 1.5, minimum standard deviation over trajectories of 0.003, and minimum Pearson correlation of 0.7.

Human *in-vitro* dataset from differentiation toward β-like cells

We used a scRNAseq time-series dataset from a differentiation protocol from human embryonic stem cells toward pancreatic β-like cells profiled using inDrops.[20] The differentiation protocol consists of six stages, with pancreatic endocrine cells appearing throughout stage five. Veres et al.[20] performed sequencing at the end of each stage and daily sampling across stage five. The data is available in the Gene Expression Omnibus under accession number GSE114412. We restricted the data to the endocrine lineage, from NKX6-1+ progenitors to hormone-producing cells. Our final subset for Figure 3 contained 25299 cells. We kept the original cluster annotations.

Data pre-processing and velocity computation

Sequencing reads were preprocessed according to the dropEst pipeline[77] (https://github.com/kharchenkolab/dropEst). A reference index was built from the Ensembl GRCh38 human genome assembly and the GRCh38.88 transcriptome annotation to run the pipeline. Shortly, we first extracted the cell barcodes and UMIs from the library using the dropTag command. Next, we used STAR 2.7.9a[78] to map the reads to the human transcriptome. Finally, we used the dropEst command with the option -V, which allows the output of three separate matrices containing only UMIs of a specific type: intronic, exonic, or exon/intron spanning. These matrices were used to build an annotated h5ad object with the unspliced layer equal to the sum of intronic and spanning UMIs and the spliced layer corresponding to the exonic UMIs. We used SCANPY and scVelo with mostly default parameters. We filtered genes to be expressed in at least three cells. Next, we normalized by total counts per cell, log-transformed the data, and kept the top 2000 highly variable genes. We embedded the data in the PCA space and used the top 50 principal components to compute a k-nearest neighbor graph with k = 50. For visualization, we used UMAP embedding with two dimensions with default parameters. To compute RNA velocity, we used scVelo's dynamic model of splicing kinetics.

FateCompass specific computations

For downstream analysis of the FateCompass pipeline, we embedded the human *in-vitro* data of gene expression and RNA velocity on ten dimensions in the UMAP space. Next, we computed a neighborhood graph in the reduced gene expression space with k = 50. This setting was the graph structure for the Markov chain operations of the FateCompass pipeline. The edges of the Markov chain were directed using the potential energy landscape described in Equation 5. We outlined stochastic trajectories using a Monte Carlo Sampling algorithm. Last, the differential TF activity analysis thresholds were: minimum Z score of 1.5, minimum standard deviation over trajectories of 0.006, and minimum Pearson correlation of 0.7.

Lineage tracing of NEUROG3+/NKX6-1+ cells

Generation of the NKX6-1-GFP/NEUROG3-HA-mCherry hiPSC line

The NKX6-1-GFP hiPSC line (1a-21)[75] was gene edited by CRISPR/Cas9 to knock in a 3xHA-P2A-3xNLS-mCherry cassette in fusion with NEUROG3 coding sequence, following exactly the strategy detailed in Schreiber et al.[38] The NKX6-1-GFP hiPSC line was nucleofected with a pX458-plasmid (Addgene) expressing the sgRNA targeting NEUROG3 14 bp downstream of the STOP codon and the Cas9 fused to GFP, and the targeting vector pBSII-KS-hNEUROG3-3HA-2A-3NLS-mCherry-pA containing a 5′ homology arm of 1112 bp and a 3′ homology arm of 807 bp flanking the 3HA-2A-3NLS-mCherry-pA cassette (detailed information on the generation of plasmids is available upon request). Nucleofection was performed with 8 x 10^5 NKX6-1-GFP hiPSC cells and 2.5 μg of each plasmid DNA in nucleofection mix (Human Stem Cell Nucleofector Kit 2, LONZA), using a Nucleofector 2b (AMAXA) according to the manufacturer instructions. The day after, cells were harvested with TryPLE Select (Thermo Fischer Scientific), Cas9-GFP+ cells were sorted and 1000 cells seeded into a Corning Matrigel-coated 35mm dish with mTeSR1 (STEMCELL Technologies), supplemented with 10 μM Y-27632 (STEMCELL Technologies) for the first 24h. Medium was replaced everyday with mTeSR1 for 11 days. Individual clones were picked and characterized by PCR genotyping, sequencing and ddPCR and expanded for banking. The heterozygous NKX6-1-GFP/NEUROG3-HA-mCherry clone#10 was selected for this study. Sequence of sgRNA oligos and PCR primers used to amplify and sequence DNA can be found in key resources table.

Differentiation of hiPSC toward pancreatic endocrine progenitors

The heterozygous NKX6-1-GFP/NEUROG3-HA-mCherry clone#10 was differentiated to pancreatic endocrine progenitors (Stage 5 day 3, or day 13) following the protocol detailed in Schreiber et al., based on Petersen et al.[38,46] Cells were harvested with TrypLE Select and seeded at 5.5 x 10^5 cells/cm^2 on 12-well CellBind plates (Corning) coated with Growth Factor Reduced Matrigel (Corning) diluted 1:30 in DMEM/F12 (Thermo Fischer Scientific), in mTeSR1 (STEMCELL Technologies) supplemented with 10 mM Y-27632 (STEMCELL Technologies). Differentiation was initiated 24 h after seeding. Cells were rinsed with 1x PBS, then exposed daily to freshly prepared differentiation medium as described below.

- Base media composition:

Stage 1/2 base medium: MCDB 131 medium (Thermo Fisher Scientific), 1.5 g/L sodium bicarbonate (Thermo Fisher Scientific), 1X GlutaMAX (Thermo Fisher Scientific), 10 mM D-Glucose (Sigma), 0.5% BSA (Sigma), Penicillin-Streptomycin 0.1% (Thermo Fisher Scientific).

Stage 3/4 base medium: MCDB 131 medium, 2.5 g/L sodium bicarbonate, 1X GlutaMAX, 10 mM D-Glucose, 2% BSA, Penicillin-Streptomycin 0.1%, 1:200 ITS-X (Thermo Fisher Scientific).

Stage 5 base medium: MCDB 131 medium, 1.5 g/L sodium bicarbonate, 1X GlutaMAX, 20 mM D-Glucose, 2% BSA, Penicillin-Streptomycin 0.1%, 1:200 ITS-X (Thermo Fisher Scientific).

- Differentiation media composition:

Stage 1 (definitive endoderm): Stage 1/2 base medium supplemented with 100 ng/mL Activin A (STEMCELL Technologies) and 3 μM CHIR99021 (Axon Medchem) at day 0, 100 ng/mL Activin A and 0.3 μM CHIR99021 at day 1, and 100 ng/mL Activin A at day 2.

Stage 2 (primitive gut tube): Stage 1/2 base medium supplemented with 250 μM L-Ascorbic Acid (Sigma) and 50 ng/mL FGF7 (Peprotech), at days 3 and 4.

Stage 3 (posterior foregut): Stage 3/4 base medium supplemented with 250 μM L-Ascorbic Acid, 50 ng/mL FGF7, 250 nM SANT-1 (Sigma), 100 nM LDN193189 (Stemgent), 200 nM TBP (Millipore EMD), and 1 μM Retinoic Acid (Sigma), at days 5 and 6.

Stage 4 (pancreatic progenitor): Stage 3/4 base medium supplemented with 250 μM L-Ascorbic Acid, 2 ng/mL FGF7, 250 nM SANT-1, 200 nM LDN193189, 100 nM TBP, and 100 nM retinoic acid, at days 7, 8 and 9.

Stage 5 (pancreatic endocrine progenitor): Stage 5 base medium supplemented with 10 μM $ZnSO_4$ (Sigma), 250 nM SANT-1, 100 nM LDN193189, 10 μM ALK5i II (ENZO), 50 nM retinoic acid, 1 μM T3 (Sigma), and 10 μg/mL heparine (Sigma) at days 10, 11 and 12.

Flow cytometry analyses

Flow cytometry analyses were performed on differentiated cells as described in Schreiber et al.[38] Cells differentiated at day 13 were dissociated to single cells with TrypLE Select for 4 min at 37°C, resuspended in Stage 5 base medium, washed once with PBS and fixed with 4% paraformaldehyde in PBS for 20 min at room temperature. After 2 washes with PBS, cells were permeabilized 30 min with 0.2% Triton X-100, 5% Donkey serum (Jackson Immunoresearch) in PBS (permeabilization buffer) and incubated overnight at +4°C in the dark with primary antibodies diluted in permeabilization buffer. After 2 washes with 0.1% Triton X-100, 0.2% BSA in PBS (PBSTB), cells were incubated for 1–2 h at room temperature in the dark with fluorophore-conjugated secondary antibodies diluted in permeabilization buffer. After 2 washes with PBSTB in the dark, cells were resuspended at 1 M/mL in 1% BSA in PBS, filtered on 50 μm nylon mesh (NITEX, Dutscher, FR) and analyzed on a BD Fortessa LSR II Cell analyser (BD Biosciences). Analyses were performed using FlowJo software v10. Antibodies used were: sheep anti-NEUROG3 (1:400 dilution, R&D), chicken anti-GFP (1:8,000 dilution, Abcam), Alexa Fluor 647 coupled anti-NKX6-1 (1:20 dilution, BD), Alexa Fluor 647 Donkey anti Sheep (1:500 dilution, Jackson ImmunoResearch) and Alexa Fluor 488 Donkey anti Chicken (1:1,000 dilution, Jackson ImmunoResearch). Refer to key resources table for antibodies details.

FACS sorting and single-cell RNA sequencing

Cells differentiated at day 13 were harvested with TrypLE Select as described above, resuspended in Stage 5 medium supplemented with 10 μM Y-27632, filtered using a 50 μM sterile Filcon (BD Biosciences) and sorted using a FACSAria Fusion cell sorter (BD) directly into 2 x 384-well plates with ERCC spike-ins (Agilent), reverse transcription primers and dNTPs (both Promega), according to the gating shown in Figure S10B. Repartition was as follow: 216 cells for each NKX6-1-GFP+, NEUROG3-mCherry+ and NKX6-1-GFP+/NEUROG3-mCherry+ population and 104 cells for the negative population. Single cell sequencing was performed according to the Sort-seq method[68] by Single Cell Discoveries. Briefly, Sequencing libraries were generated with TruSeq small RNA primers (Illumina) and sequenced paired-end at 60 and 26 bp read length, respectively, on the Illumina NextSeq. Reads were mapped to the human GRCh38 genome assembly. Sort-seq read counts were filtered to exclude reads with identical library-, cell- and molecule barcodes. UMI counts were adjusted using Poisson counting statistics.[68]

Data pre-processing

We considered 522 cells for downstream analysis. We used SCANPY with mostly default parameters. For quality control, we removed cells with a high fraction of mitochondrial gene counts (>50%), and a high percentage of ERCC spike-in reads (>50%); also, cells with more than 100000 and less than 10000 counts were excluded. We filtered genes to be expressed in at least three cells and remove spike-in genes for downstream analysis. Next, we normalized by total counts per cell, log-transformed the data. We embedded the data in the PCA space and used the top 15 principal components to compute a k-nearest neighbor graph with k = 15. For visualization, we used UMAP embedding with two dimensions with default parameters. We used Louvain-based clustering[83] for clustering and cell-type identification. Cell types were annotated based on the expression of known marker genes.

TF activity estimation

TF activities were estimated using FateCompass. We embedded the data of gene expression on ten dimensions in the UMAP space. Next, we computed a neighborhood graph in the reduced gene expression space with k = 10. This setting was the graph structure for the regularization of the TF activities.

QUANTIFICATION AND STATISTICAL ANALYSIS

All figures report the value of the mean and the standard error of the mean calculated as described in the methods section. We used Wilcoxon rank-sum test implemented in the Scanpy toolkit to compare transcription factor activities in groups of at least three members. All statistical analyses performed are included within the Supplementary Code (key resources table).