IQCELL: A platform for predicting the effect of gene perturbations on developmental trajectories using single-cell RNA-seq data

Tiam Heydari[1,6], Matthew A. Langley[2,3], Cynthia Fisher[1,6], Daniel Aguilar-Hidalgo[1,6], Shreya Shukla[2], Ayako Yachie-Kinoshita[2,4], Michael Hughes[5], Kelly M. McNagny[1,5] & Peter W. Zandstra[1,6,*]

[1] School of Biomedical Engineering, University of British Columbia, Vancouver, British Columbia, Canada, V6T 1Z3
[2] Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, ON, Canada
[3] Current Address: Division of Biology and Bioengineering, California Institute of Technology, Pasadena, CA 91125, USA
[4] Current Address: The Systems Biology Institute, Shinagawa, Tokyo, Japan
[5] Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada
[6] Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada, V6T 1Z4
[*]Corresponding Author and Lead Contact (peter.zandstra@ubc.ca)

ABSTRACT

The increasing availability of single-cell RNA-sequencing (scRNA-seq) data from various developmental systems provides the opportunity to infer gene regulatory networks (GRNs) directly from data. Herein we describe IQCELL , a platform to infer, simulate, and study executable logical GRNs directly from scRNA-seq data. Such executable GRNs provide an opportunity to inform fundamental hypotheses in developmental programs and help accelerate the design of stem cell-based technologies. We first describe the architecture of IQCELL. Next, we apply IQCELL to a scRNA-seq dataset of early mouse T-cell development and show that it can infer *a priori* over 75% of causal gene interactions previously reported via decades of research. We will also show that dynamic simulations of the derived GRN qualitatively recapitulate the effects of the known gene perturbations on the T-cell developmental trajectory. IQCELL is applicable to many developmental systems and offers a versatile tool to infer, simulate, and study GRNs in biological systems. (https://gitlab.com/stemcellbioengineering/iqcell)

KEYWORDS

gene regulatory networks; single-cell RNA sequencing; regulatory network inference; development

INTRODUCTION

Stem cell fate decisions are made via dense arrays of interacting transcription factors (TFs) forming gene regulatory networks (GRNs) (Semrau & van Oudenaarden, 2015). Information gleaned from GRNs in stem cell differentiation can lead to more effective design-based cell cultures, applicable to cell therapies (Lipsitz et al., 2016; Prochazka et al., 2017). As a prominent example, the effect of transcription factors on GRNs has been widely utilized in the reprogramming of embryonic and adult somatic cell GRNs for the establishment of a pluripotent state via induction of driver TFs (Takahashi & Yamanaka, 2006). Stem cell reprogramming and differentiation can be modeled as executable and

logical (Boolean) GRNs undergoing state transition (Sara-Jane Dunn et al., 2019; Peter et al., 2012; Yachie-Kinoshita et al., 2018). Executable GRNs provide information about both the topology and the regulatory rules of gene interactions that can be simulated as time-evolving (dynamical) systems. However, deriving informative, executable, and predictive GRNs for stem cell differentiation has proven to be a challenging task. Specifically, developing executable GRNs by piecing together evidence from gene perturbation experiments has shown to be an effective strategy (Peter et al., 2012) but is extremely time-consuming, labor intensive, and expensive. In a notable advancement, automated formal reasoning successfully identified a set of minimal GRNs underlying naive pluripotency in mice. Gene expression observations across multiple culture conditions were used to logically constrain possible GRN configurations, and the resulting set was able to accurately predict the outcome of 70% of new experiments(S.- J. Dunn et al., 2014; Yordanov et al., 2016). Yet, these methods are not based on high-throughput data.

More recently, the emergence of single-cell profiling technologies has provided an unprecedented archive of information regarding cells undergoing fate determination and maturation. . Deriving more accurate GRNs based on sc data is at the center of many recent efforts (Babtie et al., 2017; Fiers et al., 2018; Pratapa et al., 2020). Formal reasoning has been used to infer executable GRNs directly from high throughput single-cell quantitative PCR (sc qPCR) data (Hamey et al., 2017; Moignard et al., 2015). However, using single-cell RNA-sequencing (scRNA-seq) data has many advantages in terms of coverage, availability, flexibility in gene selection, and accuracy in clustering and pseudo-time inference compared to sc qPCR. These benefits exist alongside the disadvantage of dropout effects and low sensitivity in profiling TFs. Despite the availability of this data resource, this emerging field is still missing an integrated platform to infer, study, and simulate executable GRNs directly from scRNA-seq.

Herein we report an effective strategy implemented in a Python software package (IQCELL) for reconstructing GRNs directly from scRNA-seq data, called IQCELL. Our method includes steps for correcting dropout effects, selecting desired genes, building logical GRNs directly from pseudo-time with respect to interaction hierarchy and mutual information between gene pairs, and simulating developmental trajectories under normal and perturbed conditions. We demonstrate the utility of IQCELL by reconstructing a GRN for early mouse T-cell development, a well-characterized mammalian developmental system (Longabaugh et al., 2017), using published scRNA-seq data (Zhou et al., 2019). Our resulting GRN recovers over 75% of experimentally validated causal gene-gene interactions spanning years of research. Dynamic simulations of the inferred GRN resemble experimentally observed gene expression dynamics and capture the effects of knocking out or forcibly expressing various genes during early T-cell development. Our method is generally applicable to scRNA-seq data of differentiating cells and should serve as a useful resource for the community.

RESULTS

Integrative Method for Predicting the Qualitative Effect of Gene Perturbations on Developmental Trajectories of Cells (IQCELL)

IQCELL infers logical regulatory networks directly from existing information in the scRNA-seq data of cells during development and uses these regulatory networks to

simulate and predict the behavior of developing cells under perturbed conditions (**Figure 1**). IQCELL works with quality controlled and pre-processed scRNA-seq gene expression data (Butler et al., 2018; Wolf et al., 2018). The second input of the IQCELL platform is the inferred pseudo-time ordering of cells based on scRNA-seq data (**Figure 1**). The temporal dynamics of genes helps with the inference of causal gene interactions. Pseudo-time ordering of cells using scRNA-seq data has been shown to be informative for capturing temporal and developmental dynamics (Haghverdi et al., 2016; Qiu et al., 2017).

Since gene dropout is common among many scRNA-seq datasets, particularly for transcription factors with low mRNAs copy numbers, IQCELL employs a recently developed graph-based algorithm (MAGIC) to recover gene expression (van Dijk et al., 2018) (**Figure S1A**). After selecting genes of interest based on literature curation (see **Figure S1B** for more details), we generate a set of possible interactions between genes. Information-based metrics such as mutual information are well-suited for quantifying relationships between genes (Song et al., 2012). IQCELL scores gene-gene interactions according to the mutual information between gene pairs (Krishnaswamy et al., 2014) and assigns a regulatory sign (activation or repression) to each interaction based on the significance and sign of their correlation (**Figure S1C**). These steps result in a dense weighted network of gene-gene interactions that needs to be filtered into a functional GRN. In a functional GRN, interactions are not necessarily biophysically direct but capture the consequence of regulatory relations.

To reduce the number of possible gene interactions, IQCELL forms a gene interaction hierarchy in which higher ranked genes influence lower ranked ones. To form this hierarchy, IQCELL binarizes the gene expression counts by clustering them into expressed and non-expressed states (Macqueen, 1967). The binarization process divides the pseudo-time axis into regions with compact and sparse expression densities for the genes, reflecting the pseudo-time domains where a gene is expressed at a higher or lower level (**Figure S1D**). Next, the platform identifies the transition points between expression regions for all genes and uses the order of transitions to form a gene interaction hierarchy, with highly ranked genes (with earlier transition points) having greater potential to influence those downstream. This acts as an additional filter on gene-gene interactions along with the mutual information (**Figure S1E**). The resulting directional network serves as the foundation for inferring executable GRNs.

To obtain an executable GRN model, IQCELL models interactions between genes as Boolean logic functions (Yachie-Kinoshita et al., 2018). IQCELL uses a satisfiability modulo theory engine (Z3) (**Figure S1F**), (de Moura & Bjørner, 2008) to identify logic functions that are compatible with the pseudo-time dynamics of binarized gene expression states.(Hamey et al., 2017). Finally, it selects the GRN with the highest average mutual information as the most probable constrained model. The result is a functional and executable GRN that optimally fits the input scRNA-seq data. IQCELL has built-in capabilities to simulate GRN dynamics via random asynchronous Boolean simulation (Yachie-Kinoshita et al., 2018) and compare the results with experimental data under normal and perturbed conditions. In summary, IQCELL processes scRNA-seq data inputs to infer an executable logical GRN that best fits the data.

IQCELL sorts genes based on transition points and places them in a biologically relevant order

To assess the functional capabilities of IQCELL, we evaluated its performance using a well characterized mammalian developmental system, mouse early T-cell development (Hosokawa & Rothenberg, 2020; Yui & Rothenberg, 2014). The T-cell developmental program takes place in the thymus. It drives pre-thymic progenitors differentiated within the bone marrow toward T lineage commitment , and involves dense network of genes (Kueh & Rothenberg, 2012). Sustained exposure to Notch signaling drives early thymic progenitors (ETP) to the *CD4/8* double-negative 2A (DN2A) and DN2B stages, where upregulation of T-cell lineage-specific genes and progressive loss of potential for other blood cell fates occurs. Once committed to the T-cell fate, the double-negative 3 (DN3) T-cell progenitors begin recombining the β-chain of the pre-T-cell receptor (TCR). Cells are selected for functional β-chain rearrangements through pre-TCR signaling and proceed toward *CD4/8* double-positive state (DP) (**Figure 2A**). We used a publicly available scRNA-seq dataset (Zhou et al., 2019), where the authors used fluorescence-activated cell sorting to capture mouse thymocytes at ETP-DN2 and DN3 stages based on cell surface markers. After processing the data in a manner consistent with the original publication (**Figure S4** and **Table S2**), the gene expression profiles and the pseudo-time orders were used as inputs to IQCELL.

After expression recovery, selecting genes of interest based on expression variation and biological relevance (**Table S1** and **Figure 2B**), and finding possible gene-gene interactions, IQCELL binarized gene expression values and calculated the expression density over pseudo-time (**Figure 2C**) (see STAR Methods). Sorting the genes based on their transition points placed *Notch1* and *Hes1* at the top of the gene interaction hierarchy (since their expression level stayed relatively high consistently) followed by *Lmo2*, *Tcf7*, *Myb*, and *Runx1* which agrees with their position in the regulatory hierarchy during T-cell lineage establishment (Yui & Rothenberg, 2014); whereas DN3 associated genes such as *Cd3e*, *Lef1*, and *Ptcra* (Masuda et al., 2007; Yui et al., 2010) appeared at the bottom of the hierarchy (**Figure 2C**).

Next, IQCELL used this order of genes (**Figure 2C**) as a hierarchical filter of possible interactions, with the genes at the top having the most regulatory potential in terms of number of genes they can regulate. The combination of the regulatory potential of individual genes, the mutual information between gene pairs, and interaction signs, led to a directed gene interaction network comprising the set of possible interactions (**Figure 2D**). This network then constitutes a foundation for further constraints and analyses at next steps.

IQCELL is highly predictive for functionally regulatory interactions

Following our *in silico* analysis, we compared predictions from our initial inferred interaction network to validated regulatory interactions in mouse T-cell development. The initial interaction network (**Figure 2D**) was simplified with additional constraints. These constraints enforce the gene interactions to follow the expression patterns throughout the pseudo-time axis (**Figure 2B**) when executed as a logical network. This resulted in a set of possible update logical rules for each gene, selecting the most probable interactions as scored by mutual information leading to a provisional executable GRN for early T-cell development (**Figure 3A** and **Table S1**).

To benchmark this GRN, we compared our predicted directional interactions with a recent comprehensive GRN model of mouse T cell development based on experimentally validated gene interactions (Longabaugh et al., 2017). This network consists of 38 reported interactions between the genes of interest, of which 29 (over 75%) are de novo captured directly by our simulated functional regulatory network (**Figure 3B** and **Figure 3C**). For example, it is well known that *Bcl11b*, the gene that marks T lineage commitment, is activated by Notch signaling, *Gata3*, *Tcf7*, and *Runx1* (Kueh et al., 2016). Our model predicted four activators for *Bcl11b, Notch1* (as a Notch signaling mediator (Radtke et al., 1999) and target gene (Weerkamp et al., 2006) functionally represents the presence of Notch signaling), *Gata3*, *Tcf7*, and *Runx1*. The presence of *Runx1* in our model is a notable result. *Runx1* is present in developing cells, however it only gains access to the *Bcl11b* locus after chromatin restructuring during the DN2a stage (Ng et al., 2018). Notably, more than half of the interactions that were not captured by our model are related to the *Spi1* gene which is a T-cell lineage suppressor (Yui & Rothenberg, 2014).

To test the IQCELL performance with another source of data of T-cell development under different condition, we performed scRNA-seq analysis of *in-vitro* differentiation of fetal liver hematopoietic progenitor cells toward the T-cell lineage. Application of IQCELL to this second scRNA-seq data set provided further validation of its ability to predict gene-gene interactions (**Figure S3**).

One advantage of logical GRN models is that they can not only provide information about gene interactions, but can also be simulated to predict how the system evolves in time. To demonstrate this capability, we simulated our inferred logical GRN model and compared its output to experimental observations of mouse T cell development. The scRNA-seq expression data (Zhou et al., 2019) was binarized by grouping the gene expression count into on and off states. This data was then used in principle component analysis (**Figure 3D**) and the simulated trajectories overlaid on top of the binarized scRNA-seq gene expression data (**Figure 3F**). As the initial states of the simulations (representing the starting expression state of simulations), we used the binarized representation of cells at the beginning of the pseudo-time trajectory. These cells resemble the known expression state of ETP cells (Yui & Rothenberg, 2014). However, given the noisy expression of *Notch1* and *Hes1* at the earlier pseudo-time points (**Figure 2C**), we considered the expression states of these two genes to be random which results in four distinct initial states in total (**Figure 3E**). Two steady states have been obtained for the given initial cell states (**Figure 3F**), with one of them matching the DN3a cell profile (noted by star * in Fig 3). The simulated gene expression dynamics from ETP state towards this steady state shows a similar trajectory compared to the one observed from scRNA-seq data (**Figure 3G,** compare with **Figure 2C**). The other steady state shares similarities with common lymphoid progenitors (CLP) and ETP cells (**Figure 3H** and **I**). Overall, this analysis demonstrates that our GRN model is informative about both gene interactions and the behavior of genes at the system level. Such a model has the potential to predict the effect on gene perturbations at the system level as well.

IQCELL predicts the effect of gene perturbations on developmental trajectories

Next, we tested the effect of gene perturbations on simulated developmental trajectories (**Figure 4A**). In particular, we tested the effect of gene perturbations known

to result in halting or promoting T-cell development during the ETP-DN3 stages (reviewed in (Yui & Rothenberg, 2014)).

Notch1, a cell surface receptor that mediates Notch signaling, is known to play an essential role in early T-cell development. *Notch1* deficiency leads to blocked T-cell development and accumulation of other hematopoietic lineages (Radtke et al., 1999). Using our inferred executable network, we simulated the developmental trajectory of ETP cells in the absence of *Notch1*. Simulations predicted the presence of two possible steady state attractors, localized near the earlier section of the pseudo-time domain (**Figure 4B**). Comparing the expression states of the simulation attractors (**Figure 4C**) with the binarized expression of known cell states extracted from microarray data (Jojic et al., 2013), we found that the attractor states are more similar to ETP or CLP, and none show significant similarities to later stages of T-cell development (**Figure 4D**). This agrees with previous reports (Radtke et al., 1999) that lack of *Notch1* blocks T-cell development.

Tcf7 is a crucial transcription factor for T-cell specification and differentiation that is upregulated by Notch signaling. Lack of *Tcf7* results in premature arrest of T-cell development before the DN2 stage (Weber et al., 2011). Our model predicts a single attractor state in the absence of *Tcf7* (**Figure 4B**). This attractor precedes the DN2 stage and does not express *Gata3*, *Bcl11b*, *Ets1*, *Cd3e*, or *Cd3g* (**Figure 4C**), in agreement with experimentally reported analysis of *Tcf7*-/- lymphoid-primed multipotent progenitors cultured *in vitro* (on OP9-DL4) at day 4 (Weber et al., 2011). The simulated *Tcf7* knockout steady state attractor also show more similarity to microarray profiles for ETP cells than either DN2 and DN3 cells (**Figure 4D**), which is in agreement with previous reports (Weber et al., 2011).

Next, we investigated the effect of simulated knock-out of *Bcl11b*, a crucial gene for T-cell commitment (Hosokawa et al., 2018). It has been shown experimentally that *Bcl11b* deficient cells cannot proceed beyond the DN2 stage (L. Li et al., 2010). Our *in silico* results predict one attractor in the absence of *Bcl11b* (**Figure 4B**). This attractor resembles the DN2A cell state (**Figure 4D**) which recapitulates the aforementioned experimental result of *Bcl11b* knockout (L. Li et al., 2010).

We also simulated the effects of perturbing *Runx1*, *Tcf12* and *Myb*. Knock-out of *Runx1* stops T-cell development before the DN3 stage (Egawa et al., 2007); knock-out of *Tcf12* results in developmental blockage before DP stage (Wojciechowski et al., 2007), whereas *Tcf12* overexpression enhances T-cell development(Braunstein & Anderson, 2012); knock-out of *Myb* causes multiple blocks during T-cell development (Bender et al., 2004). We have simulated these gene perturbations, and found qualitative agreement with the experimentally reported results (**Figure 4B**, **Figure 4C** and **Figure 4D**).

Next, we tested our model for a simultaneous perturbation of two genes. Interestingly, while the absence of Notch signaling results in loss of T-cell development, forced expression of *Tcf7* is able to partially rescue T-cell development in the absence of Notch (Weber et al., 2011). To test our model against this observation, we simulated shutting off the expression of *Notch1* (the mediator of Notch signaling) and forcing expression of *Tcf7* to the 'on' state simultaneously. This resulted in two attractors (**Figure 4B**), one of them localized at earlier stages, and one of them close to the DN2 stage, which reflects the limited but not complete development of T-cells when compared with the knock-out of *Notch1* (**Figure 4C** and **Figure 4D**).

In addition to these perturbations, we have performed full systematic GRN perturbations for one and two gene perturbations (**Figure S2A** and **Table S3** and **Table S4**) and sensitivity analysis of gene interactions (**Figure S2B**, **Table S3**, and **Table S4**). Taken together, we showed that our model can predict the effect of single and double gene perturbations on the developmental trajectory of early T-cell development.

DISCUSSION

There is increasing availability of scRNA-seq datasets for different developmental systems. To date, the inference, analysis, and simulation of logical GRNs directly from scRNA-seq data have not been integrated together. Here we present IQCELL, an integrated strategy implemented as a Python package to infer, analyze, and simulate GRNs directly from scRNA-seq data and pseudo-time order of the cells.

IQCELL is able to capture, directly from scRNA-seq data, over 75% of the reported gene interactions in early T-cell development (Longabaugh et al., 2017). These interactions were obtained and characterized by decades of research and experiments (Yui & Rothenberg, 2014). For example, regulators of *Bcl11b*, an essential gene for T lineage commitment, were successfully identified by IQCELL. More than half of the interactions that have not been captured were *Spi1* effector genes, which is a Notch signaling antagonist (Rothenberg et al., 2019). However, Notch signaling contains *Spi1* inhibitory effect on T-cell regulators (Rothenberg et al., 2019) and potentially masks some of *Spi1* negative regulatory roles in early T-cell development.

We also tested the dynamics of the obtained GRN. We showed that when this logical GRN is simulated from ETP cell state, its dynamics evolves to the cell state associated with the DN3 stage, in agreement with experimental observations. Importantly, we showed that our platform can produce GRN models with high predictive power for the effect of genetic perturbations. For example, simulated knock-out of *Bcl11b* caused the developmental trajectory to halt at the DN2 stage, in agreement with experimental studies. We identified eight gene perturbations that halt T-cell development at different points between the ETP and DN3 stages, and IQCELL showed satisfactory agreement for all perturbations with experimental studies (**Figure 4C**).

These results show that the multi-step strategy implemented in IQCELL is effective for reconstructing functional GRNs from existing information in scRNA-seq data. Because its methodology is not specific to a single developmental system, IQCELL may be broadly useful in understanding how GRNs contribute to cell development in a variety of developmental contexts. IQCELL results may help uncover functional relations between genes and thereby help design more effective gene manipulation strategies to drive stem cell cultures toward fates of interest. Synthetic gene interactions can be added to the GRNs outputted by IQCELL to predict the effect of novel synthetic gene circuits on native cell GRNs. A major goal in systems biology is the creation of multi-scale models that connect the decisions of individual cells within a multicellular system to emergent properties of the whole tissue (Qu et al., 2011; Swat et al., 2012). IQCELL can fill an important layer in such multi-scale models. By exposing the intracellular decision-making machinery of single cells, IQCELL could interface with other methods that connects these cellular decisions to tissue-level dynamics.

To date, there have been many methods introduced for reconstructing GRNs from single-cell data (Pratapa et al., 2020) and many of them focus on finding some type of

correlation between genes. In one case, binarized sc qPCR data was used to decode logical GRNs for embryonic blood development (Moignard et al., 2015). In another recent study, sc qPCR data and its pseudo-time order was used to decode GRNs of blood stem cells (Hamey et al., 2017). However, to the best of our knowledge, prior to IQCELL there has not been any existing method or platform to infer executable and logical GRNs from scRNA-seq data, nor have previous methods dealt with the associated challenges of lower sensitivity and dropout effects.

Although the IQCELL framework allowed us to effectively model regulatory modules as logical (Boolean) gates where no extra parameters are required, logical models are limited in some respects. Firstly, logical modeling cannot effectively capture dose-dependency in gene interactions; for example, it is known that the downstream responses to *Gata3* are dose-dependent (Rothenberg, 2019). We suggest in future that this aspect be captured by multilevel models (Collombet et al., 2017). Multilevel modeling requires more sensitive measurements of TF expression levels, which may become feasible with emerging TF profiling methods (Moffitt et al., 2016). Secondly, capturing biophysical timescales in the logical framework is not trivial; one solution would be assigning a weighted time scale (Sun et al., 2017) to each simulation update step of the logical model. This can potentially help to include some time-scale sensitive events in cell GRN dynamics, such as stochastic chromatin restructuring events (Ng et al., 2018).

Since IQCELL provides users with a flexible framework, future studies could integrate other sources of information such as binding of TFs to DNA via ChIP-seq (Johnson et al., 2007) and CRISPR screening on the effect of gene perturbations on developmental trajectories (Gilbert et al., 2014) to potentially improve GRN reconstruction. In a recent study, the combination of scATAC-seq and scRNA-seq with machine learning methods have been used to infer a set of informative transcription factors during differentiation (Kamimoto et al., 2020). In addition, new opportunities are arising to investigate the decision-making machinery of the cells in their native environment (via in-situ cell profiling) (Lee et al., 2015). The combination of these methods, prior knowledge of cell-cell interactions (Browaeys et al., 2019; Kirouac et al., 2009), and emerging theoretical knowledge and computational technologies for capturing and quantifying spatio-temporal information content of cell signaling (Cepeda-Humerez et al., 2019; Dubuis et al., 2013; Maity & Wollman, 2020; Ostblom et al., 2019) can be used as invaluable resources for the next generation of GRN inference methods. These next-generation methods would ideally integrate cell signaling (P. Li & Elowitz, 2019) with GRNs directly from multi-omics sc data. In conclusion, the results presented here suggest that IQCELL will be a broadly useful tool to study cellular decision making in a variety of developmental systems.

AUTHORS CONTRIBUTIONS

Conceptualization, T.H. and P.W.Z; Methodology, T.H; Software, T.H, M.A.L, and A.Y.K Investigation, M.A.L, C.F, M.H, and S.S; Provision of study animals, K.M.M; Writing – Original Draft, T.H, D.A.H, and P.W.Z; Writing –Review & Editing, , T.H, D.A.H, M.A.L, K.M.M, C.F and P.W.Z; Supervision, P.W.Z;

ACKNOWLEDGEMENTS

We thank Wen Zhou and Ellen V. Rothenberg for generating a publicly available high-quality scRNA-seq data set of early mouse T-cell development. We thank Sara-Jane

Dunn, Boyan Yordanov, and Ellen V. Rothenberg for our fruitful discussions and Yale S. Michaels and John M. Edgar for critically reading the manuscript. We also thank Microsoft Research (Cambridge, UK) and Sara-Jane Dunn for facilitating the opportunity for the author to deepen his understanding of the Z3 reasoning engine. Funding: This work has been supported by CIHR Foundation and NSERC Discovery grants to P.W.Z.

MAIN FIGURES

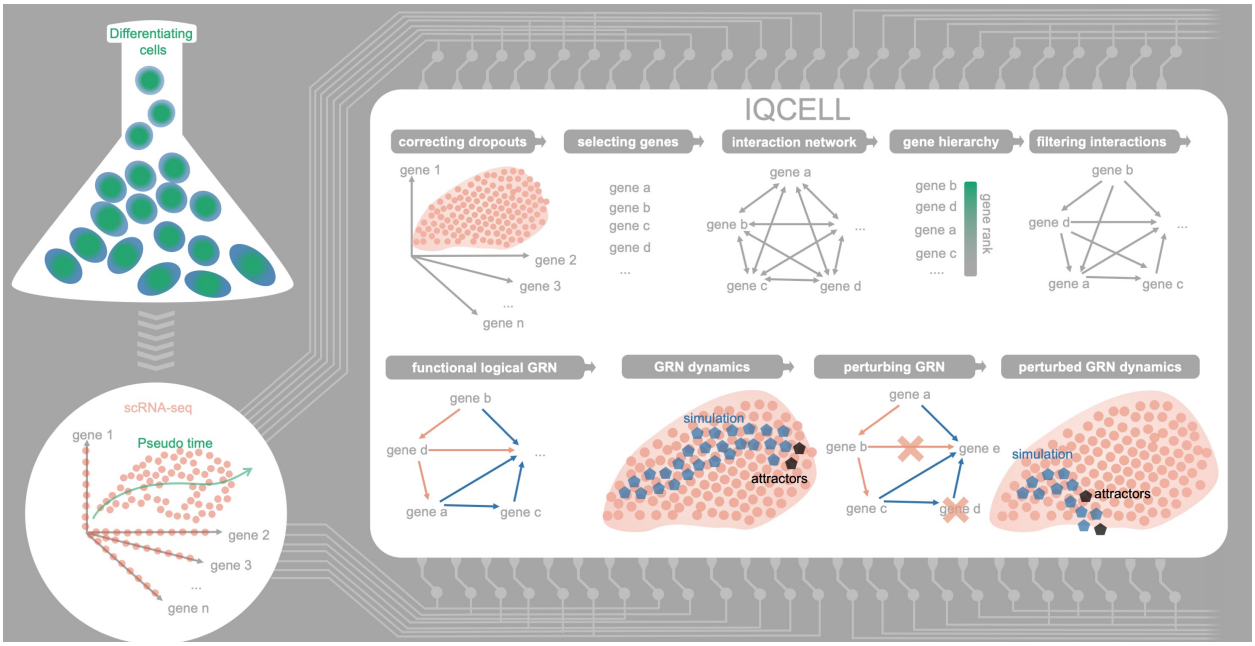


Figure 1. Overview of IQCELL.
IQCELL infers logical GRNs directly from sc-RNA seq data and allows the simulation and analysis of *insilico* developmental trajectories in normal and perturbed conditions. The typical inputs of IQCELL are sc-RNA seq expression data along with the pseudo-time ordering of the cells. After correction of dropout effects and gene selection steps, gene-gene interactions are calculated and weighted based on mutual information. Binarized gene expression values are used to constrain possible gene-gene interactions and obtain a functional GRN for the data. IQCELL can be used to analyze the GRN and simulate possible developmental trajectories under normal and perturbed conditions.

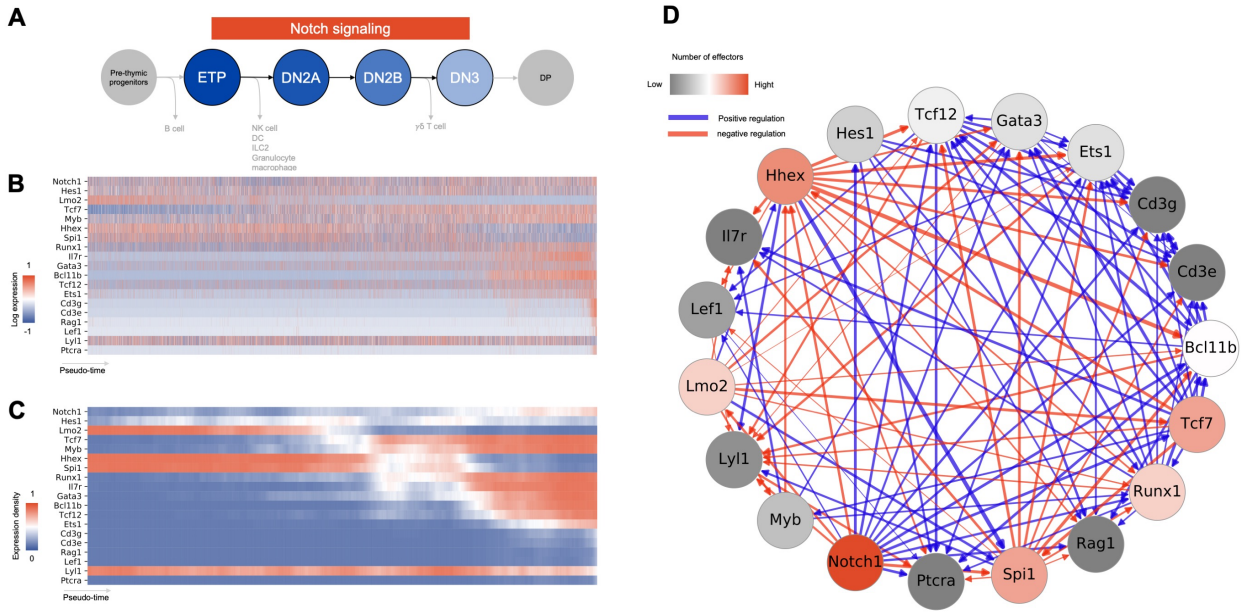


Figure 2. IQCELL initial processing of early T-cell development scRNA-seq data.
(A) Summary of the scope of the scRNA-seq data used as an input to IQCELL (Zhou et al., 2019). ETPs originated from pre-thymic progenitors progress toward DN2A, DN2B (coincides with upregulation of Bcl11b and lineage commitment), DN3 stages and eventually lead to DP cells (not covered here).
(B) Log transformed expression matrix for selected genes from scRNA-seq data along the pseudo-time axis. Gene expression is corrected for dropout effects using MAGIC (van Dijk et al., 2018).. Red indicates high expression, blue indicates low expression.
(C) Smoothed binarized gene expression matrix (expression density). Gene expression values were binarized by clustering, averaged over a pseudo-time window, then sorted based on transition points from early to late. Red indicates high expression, blue indicates low expression.
(D) The set of all possible gene-gene interactions, filtered by interaction hierarchy and mutual information (and signed by correlation. Positive and negative interactions are represented by blue and red edges, respectively. Edge width represents the relative amount of mutual information of the interaction. Nodes colored red have higher out-degrees.

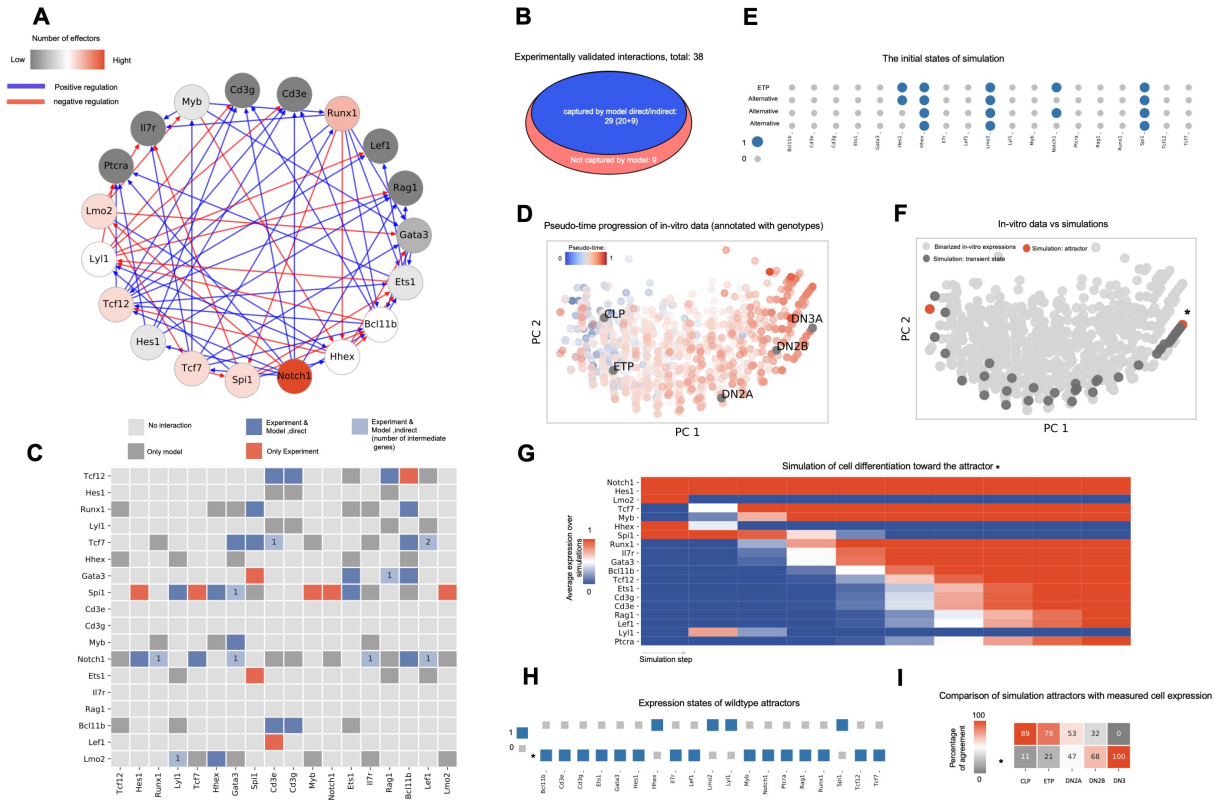


Figure 3. The provisional GRN for mouse early T-cell development inferred by IQCELL captures essential gene interactions and accurately simulates T-cell developmental trajectories.

(A) The provisional GRN for early mouse T-cell development. The GRN is obtained by constraining the possible interactions to both follow the *in vitro* data progression when executed as a logical network and maximize mutual information between gene pairs. Positive and negative interactions are represented by blue and red edges, respectively. Nodes colored red have higher out-degrees.

(B) Out of 38 experimentally reported gene interactions of early mouse T-cell development (Longabaugh et al., 2017), 29 of them are captured by the functional GRN model proposed by IQCELL.

(C) Detailed representation of the proposed interactions by IQCELL and experimentally reported ones. Rows and columns represent regulators and effector genes, respectively. Blue indicates that the interaction is captured by the model directly (dark blue) or indirectly (light blue); in the latter case, the numbers indicate the number of intermediate genes. Dark gray indicates that the interaction is only proposed by IQCELL. The red color indicates the experimentally validated interaction is not present in the model. Light gray cells indicate no interaction. Genes downstream of *Spi1* comprise 50% of the experimentally-reported interactions not captured by IQCELL.

(D) The PCA plot of the binarized scRNA-seq data color-coded with the pseudo-time values attributed to each cell. The binarization is performed by clustering the scRNA-seq expressions into expressed or not expressed levels. On top of that, the binarized

expressions of CLP, ETP, DN2A, DN2B, and DN3A cells have been calculated from the immgen microarray data (Jojic et al., 2013) and overlaid on RNA-seq data.
(E) The four initial states that have been used in simulations. Three variations of the state representing ETP are due to the noisy expressions of Notch1 and Hes1 genes in recovered sc-RNA seq data with early pseudo-time. Genes that are expressed (1) and not expressed (0) are represented with blue and grey circles, respectively.
(F) The PCA plot of the simulated developmental trajectories are overlaid on the binarized scRNA-seq. The two detected attractors are colored red, and the attractor that matches the DN3A state is marked by star (*).
(G) Average gene expression at each simulation step. All simulations started from the same initial condition (ETP) and move toward the same attractor (*).
(H) Expression states of the GRN model steady state attractors. Genes that are expressed (1) and not expressed (0) are represented with blue and grey squares, respectively.
(I) Percentage of similarity between the two attractors (vertical axis) and binarized microarray expression profiles of CLP, ETP, DN2A, DN2B, and DN3A cells (horizontal axis) (Jojic et al., 2013). The average agreement between two random states is 50%.

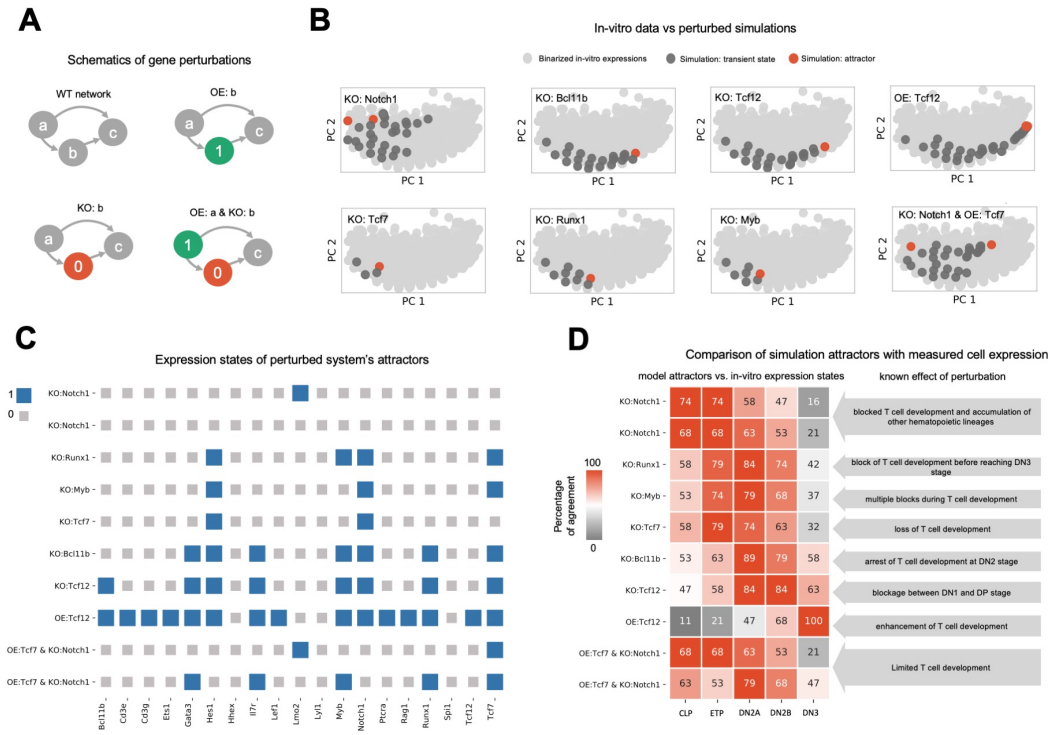


Figure 4. Testing the known effect of eight gene perturbations on in-silico developmental trajectories.

(A) Schematic of performed gene perturbations. In overexpression (OE), the gene is always expressed (represented with 1) and in knockout(KO), the gene is always silent (represented with 0).

(B) PCA plot of the simulated developmental trajectories under perturbed conditions are overlaid on the binarized scRNA-seq. The perturbations include knock-out of *Notch1*, knock-out of *Tcf7*, knock-out of *Bcl11b*, knock-out of *Runx1*, knock-out of *Tcf12*, knock-out of *Myb*, overexpression of *Tcf12* and the double perturbation, overexpression of *Tcf7* and knock-out of *Notch1* at the same time.

(C) Expression states of the model attractors under perturbations. Genes that are expressed (1) and not expressed (0) are represented with blue and grey squares, respectively.

(D) Percentage of similarity between the model attractors under perturbations (vertical axis) and the binarized expressions of CLP, ETP, DN2A, DN2B, and DN3A cells (horizontal axis) (Jojic et al., 2013) (left). Description of known effect of the gene perturbation on T-cell development (right).

METHODS

RESOURCES

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
IQCELL	This paper	https://gitlab.com/stemcellbioengineering/iqcell
MAGIC	(van Dijk et al., 2018)	https://github.com/KrishnaswamyLab/MAGIC
DREMI	(Krishnaswamy et al., 2014)	https://dpeerlab.github.io/dpeerlab-website/dremi.html
Z3	(de Moura & Bjørner, 2008)	https://github.com/Z3Prover/z3
Garuda-Boolean	(Yachie-Kinoshita et al., 2018)	https://gitlab.com/stemcellbioengineering/garuda-boolean
Scanpy	(Wolf et al., 2018)	https://github.com/theislab/scanpy
Monocole2	(Qiu et al., 2017)	http://cole-trapnell-lab.github.io/monocle-release/
Deposited Data		
will be available to reviewers upon request		
Experimental Models: Organisms/Strains		
8-week old adult male CD1 mice	Charles River Laboratories	Crl:CD1(ICR)
Biological Samples		
Fetal thymocytes	This paper	N/A
Other		
scRNA-seq of mouse thymic ETP populations	(Zhou et al., 2019)	Gene Expression Omnibus GSE137165

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to the Lead Contact, Peter Zandstra (peter.zandstra@ubc.ca).

Data and Code Availability

The source code of IQCELL python package generated during this study is available on Gitlab: (https://gitlab.com/stemcellbioengineering/iqcell)

METHOD DETAILS

Gene expression recovery

scRNA-seq data is usually affected by dropouts, which is a technical term that is used to describe the false-negative reads of messenger RNAs. Dropout effects cause the expression profile of genes to be underrepresented. Usually, genes with low copy numbers (e.g transcription factors) are more affected by this effect. IQCELL applies a recent method (MAGIC) that uses a graph-based imputation method to infer the expression of dropouts (van Dijk et al., 2018) (**Figure S1A**). This imputation is important as dropouts can affect the inference of gene relations (**Figure S1A**).

Gene selection

Selecting a small subset of genes to include in a functional GRN from the entire set of genes detected by scRNA-seq is, in general, a challenging task. Fortunately in mouse T-cell development, many relevant genes are known (Longabaugh et al., 2017). Here we describe a possible general approach for a selecting smaller subset of genes from a large set.

Selecting relevant genes for a functional/causal GRNs is generally a multilayered process (**Figure S1B**) which ideally combines multiple sources of information. One possible source is prior knowledge of important genes in the process, which can be obtained through literature review or systematic gene perturbation experiments (such as genome-wide CRISPR screening). . Alternatively, genes of interest can be selected directly from scRNA-seq data via various information theory metrics (Ang et al., 2016). Many scRNA-seq data analysis packages set an initial filter to only select highly variable genes (HVGs) for downstream analysis. HVGs are genes whose mean-scaled variance exceeds an automatic threshold. Although this is generally a useful filter, it does not necessarily select for genes whose expression levels vary significantly across pseudo-time. Therefore, IQCELL has a built-in function to visualize expression dynamics along pseudo-time and calculate the degree of variation, which can be used as an additional input for gene selection.

Beside these, there are other network-based approaches to select informative genes. These methods typically prioritize genes with connections to many other genes (high degree). Finally, enrichment analysis and ChIP-seq data can be another source of gene selection; however, these methods are generally low-throughput, noisy, and prone to false positives/negatives. For this study, we manually curated a list of genes based on biological significance (Longabaugh et al., 2017) and dynamics along the pseudo-time (**Figure S1B** and **Table S1**).

Establishing the initial gene-gene interaction network

To form the initial gene-gene interactions network from scRNA-seq data (**Figure S1C**), IQCELL first forms a list of all possible pairwise gene-gene interactions. This list does not include autoregulation by default (optional). Next, it uses a recent method (DREMI) to calculate the resampled and conditional mutual information between gene pairs (Krishnaswamy et al., 2014) (**Figure S1C**). In general, the mutual information (I) between a pair of variables X and Y (where X and Y represent two genes) is calculated as below:

$$\mathrm{I}(\mathrm{X};\mathrm{Y}) = \mathrm{H}(\mathrm{Y}) - \mathrm{H}(\mathrm{Y}|\mathrm{X})$$

where H(X) is called the entropy of distribution X and P(X) is the probability density of the variable X:

$$\mathrm{H}(\mathrm{X}) = \sum \mathrm{P}(\mathrm{X})\log\ \mathrm{P}(\mathrm{X})$$

The mutual information score is symmetric and unsigned. Next, IQCELL applies the Pearson correlation coefficient to assign a sign (+/- which represents activation/repression) to the interactions, based on the sign of the correlation between two genes (**Figure S1C**):

$$c(X,Y) \geq 0 \quad \rightarrow +$$
$$c(X,Y) < 0 \quad \rightarrow -$$

Where c(X,Y) is the Pearson correlation coefficient of the pair of gene X and Y:

$$c(X,Y) = \frac{E(XY) - E(X)E(Y)}{\sigma_X \, \sigma_Y}$$

E(X) is the expected value and σ_X is the standard deviation of X. Finally, IQCELL eliminates interactions with mutual information values smaller than a pre-defined threshold (with the default value of one standard deviation below the mean). The result is an undirected and signed interaction network.

Binarization of gene expressions

IQCELL binarizes the expression for each gene individually. To do this, it can apply two possible methods that can be chosen by the user (**Figure S1D**). The first method finds the mean of expression and assigns 'off' (0) to the genes (for each cells) with the expressed numbers of mRNAs that are smaller than the prefixed threshold and 'on' (1) if they are larger. The second method (the default method) binarizes the expression based on the cut-off identified by k-means clustering (k=2) method (**Figure S1D**)(Macqueen, 1967). In general, we group the expression of each gene in each cell g_i ($i \in N_{cells}$) into two possible groups $S_i = \{S_{off}, S_{on}\}$ by finding:

$$\arg\min_S \sum_{i \in off,on} \sum_{g \in S_i} ||g - \mu_i||$$

In which μ_i (centroids) is the mean of expression values in 'on' or 'off' cluster. The threshold (τ) is obtained as the average of two centroids:

$$\tau = (\mu_{off} + \mu_{on})/2$$

The binarized gene expression value (G_i) is obtained similar to the mean method:

$$if(g_i \leq \tau) \rightarrow G_i = 0$$
$$if(g_i > \tau) \rightarrow G_i = 1$$

Establishing the gene hierarchy via pseudo-time

After binarization of expression levels, IQCELL calculates an interaction hierarchy to further filter gene-gene interactions, thereby eliminating interactions that are less likely to be causal. First, it averages the binarized expression values in a sliding window over pseudo-time (**Figure S1E**). To do this, IQCELL first sorts the cells based on their pseudo-time values c_i ($i \in N_{cells}$). Next, it averages the values of binarized gene expressions over an averaging window (with the default length of $L = N_{cells}/N_{genes}$). This results in a density representation of the binarized gene expression values along pseudo-time (t) (**Figure S1E**):

$$D(t) = \sum_{i \in (t-L/2, t+L/2)} G_i / L$$

Next, it calculates transition points between low-to-high or high-to-low density regions for all genes (**Figure S1E**) and sorts genes based on their transition points (**Figure S1E**). Finally, IQCELL includes autoregulation (autoactivation) as a possible self-interaction for genes with less than two possible activators. This step is optional and is used in the current study, leading to inclusion of autoactivation interactions for *Hes1*, *Notch1*, *Lmo2*, and *Spi1* in the final interaction network.

Implementing the Z3 reasoning engine to infer logical GRNs

To generate functional GRNs, IQCELL implements a modified network inference strategy in the Z3 engine (Hamey et al., 2017). This method effectively finds optional logical rules for each gene based on the possible list of interactions obtained from previous steps. The optimal rules are those that when executed as logical gates for each gene (given the state updates of other genes along the pseudo-time as an input), follow the experimental data. This is quantified with the percentage of similarity (based on Hamming distance) between the two (**Figure S1F**). Similar to (Hamey et al., 2017), we allow up to four possible activators and up to two repressors for each gene. In contrast to (Hamey et al., 2017) and similar to (S.- J. Dunn et al., 2014), for gene activation, we assume that all the activators are necessary (which is implemented with the 'and' logic gate), but only one repressor is enough for repression (which is implemented with the 'or' logic gate). In summary, the most general logical rule for the regulation of a gene (g_j) by (the maximum number of) six regulators including (maximum) four activators (A_1, A_2, A_3, A_4) and (maximum) two repressors (R_1, R_2) is:

$$g_j = (A_1 \text{ and } A_2 \text{ and } A_3 \text{ and } A_4) \text{ and not } (R_1 \text{ or } R_2)$$

Which indicates that all activators and none of the repressor should be expressed for the gene to be expressed. Next, for each gene, the rule with the highest average mutual information for interactions is selected by IQCELL for the final GRN (**Table S1**).

Asynchronous simulator of GRN under normal and perturbed condition

To analyze the system-level behavior of the obtained GRN models and predict the effect of gene perturbations on developmental trajectories, IQCELL uses (asynchronous) Boolean simulations. Boolean GRNs contain a set of genes $\mathbf{G} = \{G_1, G_2, \dots G_{N_{genes}}\}$ and their update function which encodes the gene regulatory details (one update function per gene). The update function of a gene (F_{G_i}) implies what will be the activity state of that gene (on or off) at the next discrete time point $G_i(t+1)$ given the state of all the genes at the current discrete time point $\{G_1(t), G_2(t), \dots G_{N_{genes}}(t)\}$:

$$G_i(t+1) = F_{G_i}(G_1(t), G_2(t), \dots G_{N_{genes}}(t))$$

IQCELL, uses the asynchronous update strategy (Yachie-Kinoshita et al., 2018). In this strategy, at each discrete time step, only one random gene is selected and updated. This results in stochastic dynamics. This lets us average the expression states ($\mathrm{average\_exp}$) at each discrete time point (j) over the ensemble of stochastic states that started from the same initial point and are now at that particular time point ($\mathbf{EX}(\mathrm{j})$) (**Figure 3G**):

$$\mathrm{average\_exp_j} = \sum_{S \in EX(j)} S\,/||\mathrm{EX(j)}||$$

Where $S = \{0,1\}$ therefore:

$$0 \leq \mathrm{avrage\_exp_j} \leq 1$$

IQCELL also has a built-in function to perturb the GRNs in two ways. It has the capability to perform KO (setting the gene to be always 'off') and OE (setting the gene to be always 'on') experiments simultaneously for one or multiple genes systematically. Also it can perturb the gene-gene interactions systematically (**Figure S2**).

Software architecture of IQCELL

IQCELL is implemented as a Python package. It is modular and scalable, to help researchers to expand, optimize, and customize it for their future studies. Also, it has a minimal Python interface, which allows the users to use it with minimal computational skills, and implement it to their system of interest.

Pre-processing the scRNA-seq data

To dissect the developmental trajectory of T-cell development (from ETP to DN3 stages), we used a scRNA-seq dataset from the Rothenberg Lab (Zhou et al., 2019). We used the analysis pipeline provided by the Theis Lab (Luecken & Theis, 2019). After the quality control and normalizing expression values, single-cell transcriptional states were visualized in reduced dimensional space using UMAP (**Figure S4A**). To understand the underlying structure of data we perform clustering based on the Louvain method (**Figure S4B**) which yields 14 sub-clusters. Next, we evaluate the expression pattern of developmentally important genes in blood and particularly T-cell development. ETP associated genes (*Flt3*, *Lmo2*, *Mef2c*) are all expressed in the cell clusters in the left side (**Figure S4C**). DN2-a stage is marked by *Il2ra*, this gene along with the gene associated with committed DN2 cells (*Bcl11b*) and DN3 associated genes (*Cd3e* and *Rag1*) expressions are localized on the right side (**Figure S4C**). Granulocyte lineage marker (*Elane*) and macrophage lineage marker (*Mpo*) expression are high at cluster 13 (not shown) and we excluded this cluster for future analysis. This can be due to alternative lineage decisions in development or contamination. Altogether we conclude that cluster 0 includes many of the cells at the ETP stage and the developmental progression is toward clusters 9 and 11 as the endpoints (**Figure S4C**).

As previously reported (Zhou et al., 2019), we use a supervised approach in pseudo-time ordering on the subset of genes that are known to be developmentally important in T-cell development or are alternate lineage markers. The selected genes were similar to (Zhou et al., 2019) and DDRtree pseudo-time ordering is performed on the data (Qiu et al., 2017). As established in the cluster analysis, step cluster number 0 is the best

candidate as the cluster with the earliest developmental stage and is used as the root for the algorithm. The result shows a single trajectory starting at cluster 0 and progressing toward later stages (**Figure S4D**). Pseudo-time ordering shows the dynamics of genes as a function of differentiation (**Figure S4E**).

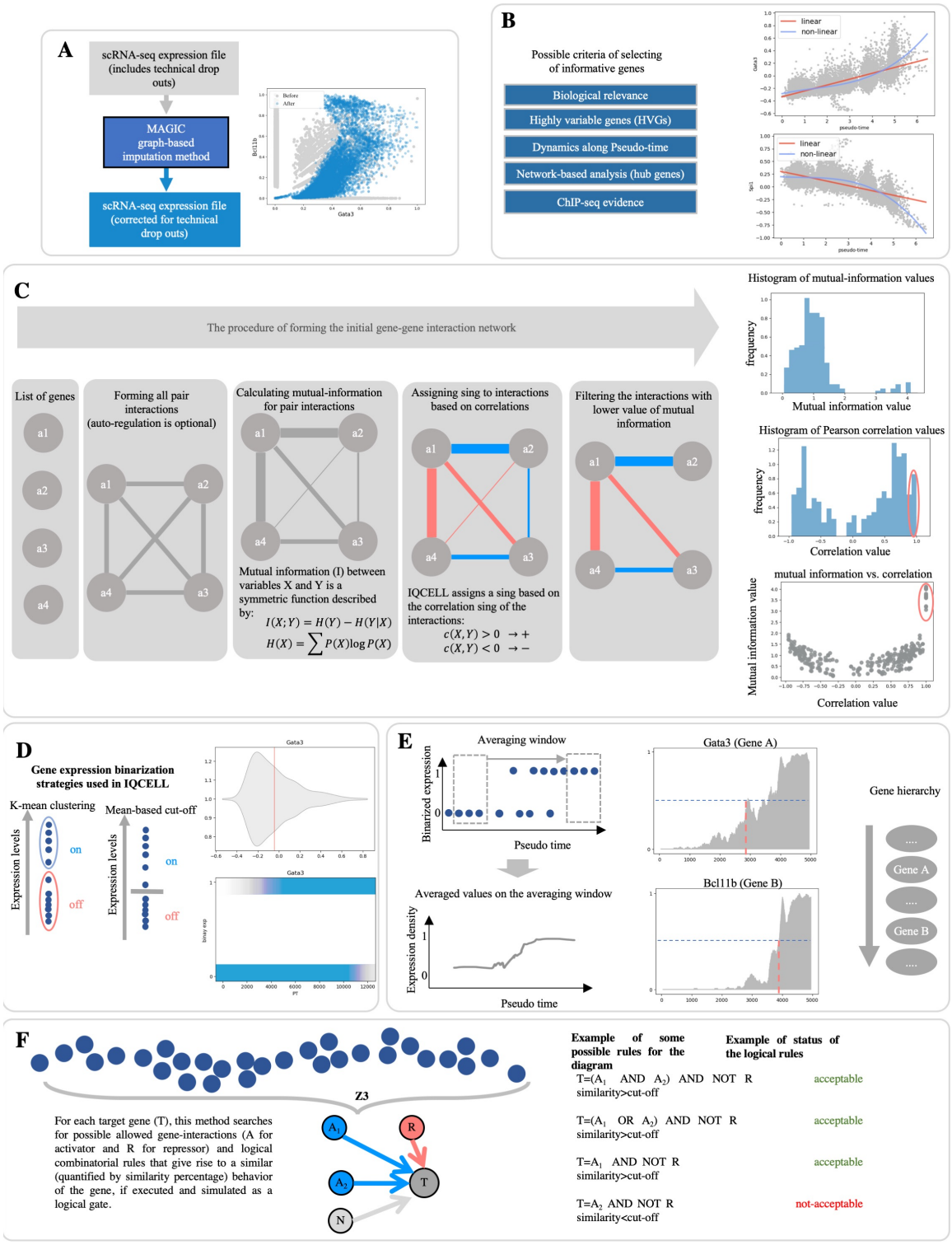
Relationship between the quality of single cell data and the GRN inference

As a short note, there is a direct relationship between the quality of the scRNA-seq data and inferred pseudo-time with the quality of the inferred GRN. A suitable scRNA-seq data has high resolution with regard to the underlying expression dynamics during the developmental process of interest, sufficient read depth, and yields a pseudo-time trajectory that qualitatively resembles the known gene expression state progression of the developmental system to be modeled (Zhou et al., 2019).

Sample preparation and Single-cell RNA-sequencing of in vitro T-cell differentiation

Isolated fetal liver cells from decapitated E13.5 CD1 mouse embryos were subjected to TER-119 depletion by EasySep magnetic sorting (STEMCELL Technologies). Next, sorted HSPCs (Sca-1+ cKit+) cultured at 3.1 x103 HSPCs/cm2 (corresponding to 1000 cells/well) in DL4 (10 µg/mL) and VCAM-1 (2.32 µg/mL) coated 96-well plates (Shukla et al., 2017).10X Chromium was used to prepare single-cell cDNA libraries, and Illumina Nextseq was used to 3’ sequence the samples. Gene-barcode expression matrices were calculated from the raw data via CellRanger (10X Genomics).

SUPPLEMENTAL FIGURES

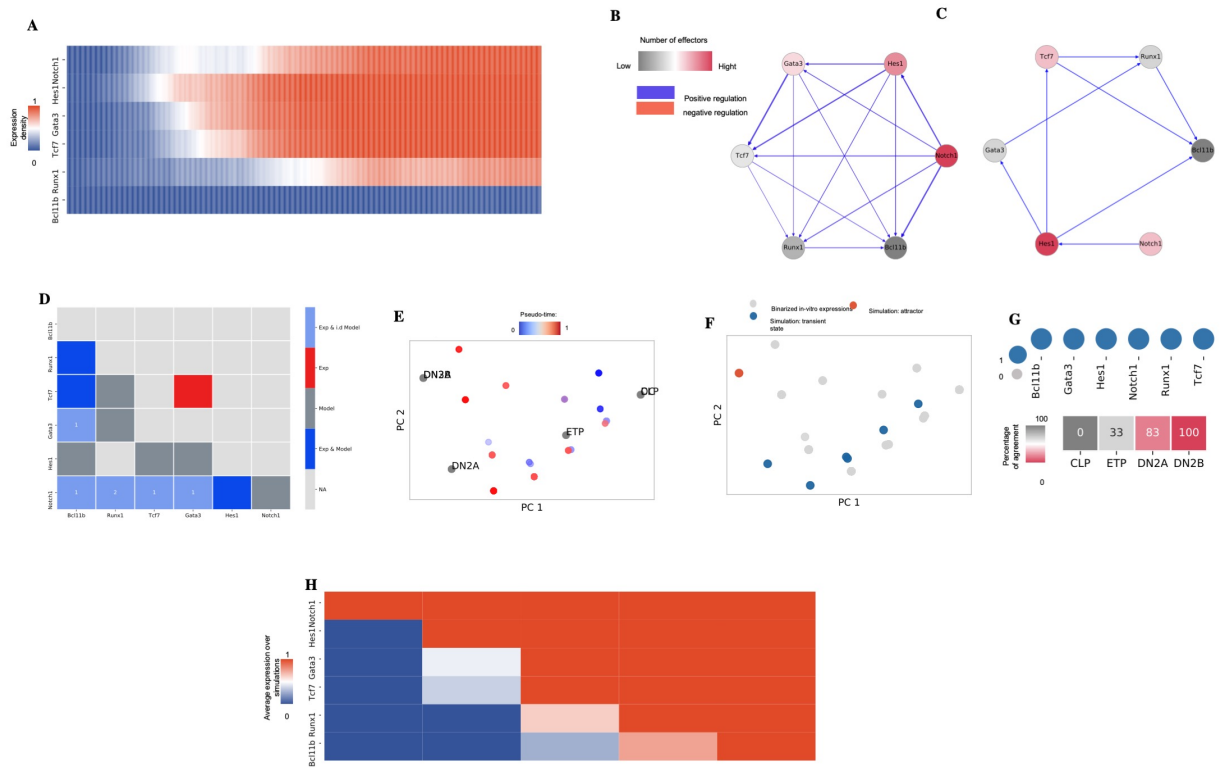


Supplementary Figure S1. **Overview of IQCELL parts and algorithms** (related to Fig. 1).

(A) Overview of gene expression recovery step. The scRNA-seq data is corrected for dropout effect via a borrowed library 'MAGIC' from literature (left). Raw (grey) and recovered (blue) expression of Bcl11b vs. Gata3 (right).
(B) Overview of gene selection step. The most important criteria for gene selection is biological relevance and literature curation. However, the user can benefit from observing the dynamics along the variability pseudo-time and variability of genes. Other network-based methods and chip-seq evidence can be used as supplementary methods for gene selection (left). To visualize the dynamics of genes along the pseudo-time (right), IQCELL has a built-in function to visualize and fit linear and non-linear functions to the gene. vs time function. Users can use this information to select genes as well.
(C) Overview of generating the initial interaction network. The steps toward obtaining the directional and signed interaction network from an initial list of genes (left column). The histogram of mutual information between gene pairs, the histogram of Pearson correlation between gene pairs (the correlation value of 1 is for the correlation of genes with themselves, marked by red), and mutual information vs. Correlation values (right column).
(D) Overview of expression binarization step. There are two implemented binarization methods in IQCELL. K-means clustering (default) and binarization based on the mean value of expression of the gene between all the cells (left). Example of binarization of genes and their expression along the pseudo time (right).
(E) Overview of generating the gene hierarchy step from the binarized gene expressions. First, the expression levels are averaged with a sliding window along the pseudo-time. This results in the density profile of binarized genes along the pseudo-time (left). Next, based on clustering the density, the transition point (from high to low or low to high) are captured (center). Finally, genes are sorted based on transition points. Genes can interact with genes with a lower rank (right).
(F) Overview of implementing the Z3 reasoning engine. At this step, the filtered set of interactions are used to make provisional update rules for each gene.

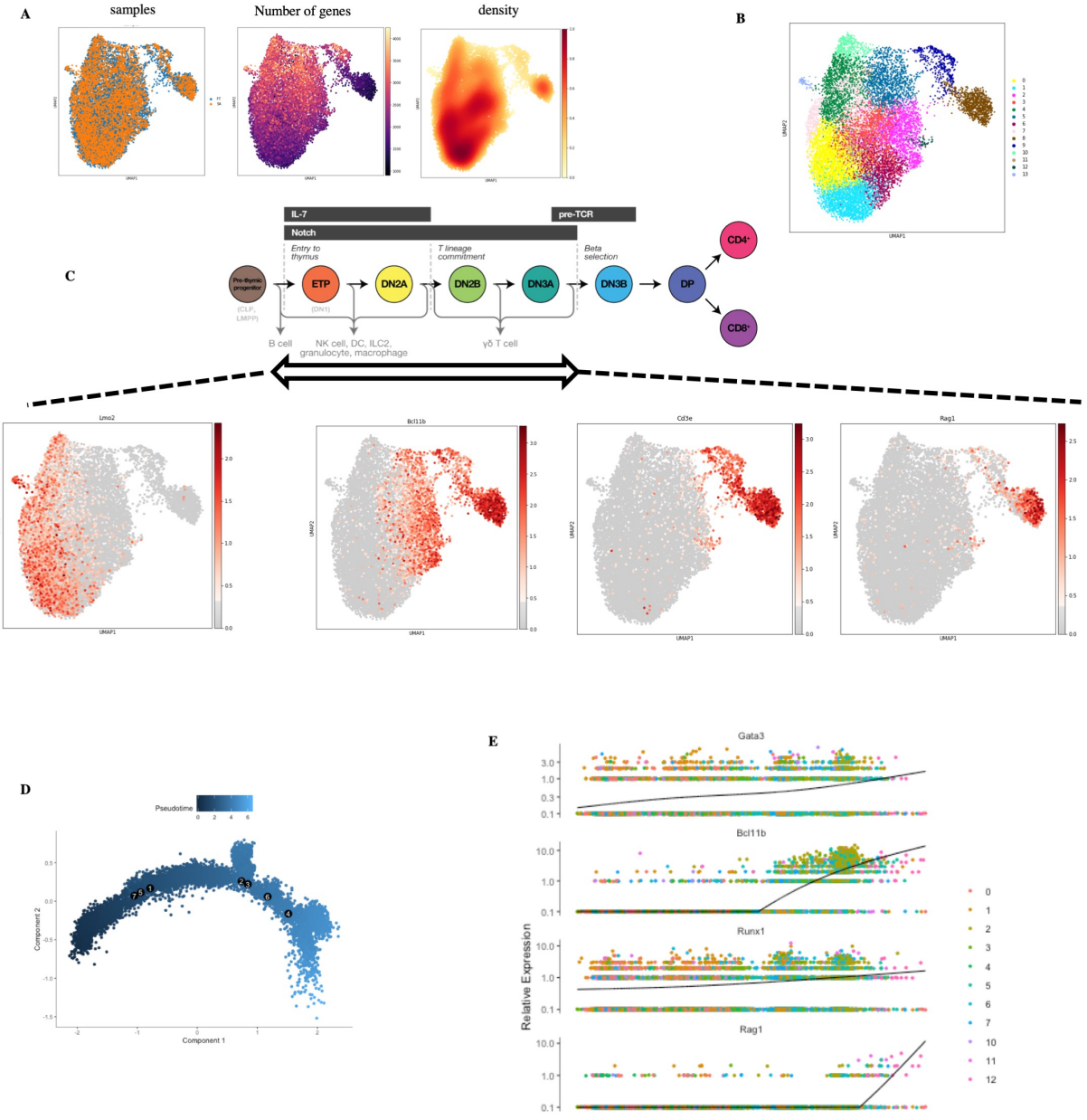
Supplementary Figure S2. GRN perturbations via IQCELL(related to Fig. 4).
(A) Systematic gene perturbations. The expression states of the model attractor (left). The percentage of similarity between the attractors (vertical axis) and the binarized expressions of CLP, ETP, DN2A, DN2B, and DN3A cells (horizontal axis) (Jojic et al., 2013) (right) for systematically perturbed GRN with single gene KO and OE.
(B) Systematic gene-gene interaction perturbations. Overview of GRN link perturbation (top). The percentage of similarity between the attractors (vertical axis) and the binarized

expressions of CLP, ETP, DN2A, DN2B, and DN3A cells (horizontal axis) (Jojic et al., 2013) (bottom) for systematically perturbed GRNs.



Supplementary Figure S3. Implementation of IQCELL with another early T-cell development scRNA-seq data set. (related to Fig. 3)
(A) To demonstrate the universality of IQCELL, we have tested this planform with another in-house scRNA-seq data. The IQCELL tutorial in the IQCELL website is based on this dataset. We used 10X scRNAseq by performing whole-genome transcriptional analysis. These experiments are performed with the mouse T-cell progenitor populations from fetal liver (FL) hematopoietic stem and progenitor cells (HSPCs) differentiated in vitro using the DL4+VCAM platform(Shukla et al., 2017). FL HSPCs were seeded on DL4+VCAM coated plates and cultured for 4 or 7 days prior to analysis, or immediately sorted and captured for library preparation. Pooling cells from multiple differentiation time points enabled sampling of cells from the entire T-cell lineage progression, rather than just endpoint transcriptional states. Here we have selected a small set of 6 genes that are important in the early T-cell development (from ETP to DN2 stages). The heat map shows the expression matrix of smoothed binarized expressions along the pseudo-time.

(B) The set of all possible gene-gene interactions, filtered by interaction hierarchy and mutual information cut-off (the thickness of lines represents the mutual information between genes), and signed by correlation.
(C) Provisional GRN for early mouse T-cell development obtained by Z3 step with additional maximize mutual information criteria.
(D) Detailed representation of the proposed interactions provided by IQCELL and experimentally reported ones.
(E) PCA plot of the binarized scRNA-seq data color-coded with the pseudo-time values attributed to each cell. The binarization is performed by clustering the cs RNA-seq expressions into expressed or not expressed levels. On top of that, the binarized expressions of CLP, ETP, DN2A, DN2B, cells have been calculated from the immgen microarray data (Jojic et al., 2013) and overlaid on RNA-seq data.
(F) PCA plot of the simulated developmental trajectories are overlaid on the binarized scRNA-seq.
(G) Expression states of the model attractors (top). The percentage of similarity between attractor and known cell states (Jojic et al., 2013).
(H) Averaged gene expression of the simulated data at each simulation step. All simulations started from the same initial condition.



Supplementary Figure S4. The overall workflow of preprocessing the scRNA-seq data (related to Fig. 3).

(A) UMAP representation of Data overlaid with sample id, number of genes per cell, and density plot. (B) Clustering of scRNA-seq data. (C) Gene expression of stage-specific gene overlaid on top of UMAP. (D) The pseudo-time trajectory of the data inferred by Monocle platform. (E) Example expression of genes along the pseudo-time trajectory.

SUPPLEMENTAL TABLES

Gene	Update Rule
Bcl11b	((Gata3 and Notch1) and (Runx1 and Tcf7)) and not (Hhex or Spi1)
Cd3e	((Bcl11b and Tcf12) and (Hes1 and Notch1)) and not Lyl1
Ets1	((Bcl11b and Gata3) and (Tcf12 and Runx1)) and not (Hhex or Spi1)
Cd3g	((Bcl11b and Notch1) and (Tcf12 and Hes1)) and not Lyl1
Gata3	((Myb and Tcf7) and Runx1) and not (Hhex or Lmo2)
Hes1	Notch1
Hhex	(Lmo2 and Spi1) and not (Myb and Runx1)
Il7r	((Runx1 and Tcf7) and Myb) and not (Lmo2 and Spi1)
Lef1	(Ets1 and Tcf12) and not Lyl1
Lyl1	(Hhex and Spi1) and not (Bcl11b and Ets1)
Lmo2	Lmo2 and not Notch1
Myb	Tcf7 and not Lmo2
Notch1	Notch1
Ptcra	((Ets1 and Hes1) and (Tcf12 and Notch1)) and not Lyl1
Rag1	((Ets1 and Notch1) and (Tcf12 and Hes1)) and not Lyl1
Runx1	(Myb and Tcf7)
Spi1	Spi1 and not (Runx1 and Tcf7)
Tcf7	Notch1 and not Lmo2
Tcf12	((Notch1 and Runx1) and Bcl11b) and not Hhex

Supplementary Table S1 (related to Fig. 2 and 3). The final gene list and provisional logical GRN for early T-cell development. The rules are picked from possible rules in Z3 step based on maximizing the average mutual information per gene.

Gene	
Bcl11a	Itgax
Bcl11b	Kit
Ccr9	Lef1
Cd34	Lmo2
Cd3e	Ly6d
Cd3g	Lyl1
Cd44	Mef2c
Cd82	Meis1
Cebpa	Mpo
Cxcr4	Myc
Dtx1	Mycn
Erg	Nfil3
Ets1	Notch1
Ets2	Nrarp
Flt3	Nt5e
Gata1	Pdgfrb
Gata2	Pgk1
Gata3	Pim1
Gfi1	Ptcra
Gfi1b	Rag1
Hes1	Rag2
Hhex	Runx1
Hoxa9	Runx2
Id2	Runx3
Id3	Sox13
Ikzf1	Spi1
Ikzf2	Spib
Il2ra	Tcf12
Il4ra	Tcf7
Il7r	Tlr7
Irf8	Zbtb16
Itga2b	Zfpm1
Itgam	

Supplementary Table S2 (related to Fig. 2). Genes used for supervised trajectory inference.

Supplementary Table S3 (related to Fig. 4). Comparison of attractors of perturbed GRN with the known cell states from microarray data. (Online)

Supplementary Table S4 (related to Fig. 4). The attractor states of perturbed GRN. (Online)

REFERENCES

Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. A. (2016). Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *13*(5), 971–989. https://doi.org/10.1109/TCBB.2015.2478454

Babtie, A. C., Chan, T. E., & Stumpf, M. P. H. (2017). Learning regulatory models for cell development from single cell transcriptomic data. *Current Opinion in Systems Biology*, *5*, 72–81. https://doi.org/10.1016/j.coisb.2017.07.013

Bender, T. P., Kremer, C. S., Kraus, M., Buch, T., & Rajewsky, K. (2004). Critical functions for c-Myb at three checkpoints during thymocyte development. *Nature Immunology*, *5*(7), 721–729. https://doi.org/10.1038/ni1085

Braunstein, M., & Anderson, M. K. (2012). HEB in the Spotlight: Transcriptional Regulation of T-Cell Specification, Commitment, and Developmental Plasticity. *Clinical and Developmental Immunology*, *2012*, 1–15. https://doi.org/10.1155/2012/678705

Browaeys, R., Saelens, W., & Saeys, Y. (2019). NicheNet: Modeling intercellular communication by linking ligands to target genes. *Nature Methods*. https://doi.org/10.1038/s41592-019-0667-5

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*(5), 411–420. https://doi.org/10.1038/nbt.4096

Cepeda-Humerez, S. A., Ruess, J., & Tkačik, G. (2019). Estimating information in time-varying signals. *PLOS Computational Biology*, *15*(9), e1007290. https://doi.org/10.1371/journal.pcbi.1007290

Collombet, S., van Oevelen, C., Sardina Ortega, J. L., Abou-Jaoudé, W., Di Stefano, B., Thomas-Chollier, M., Graf, T., & Thieffry, D. (2017). Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. *Proceedings of the National Academy of Sciences*, *114*(23), 5792–5799. https://doi.org/10.1073/pnas.1610622114

de Moura, L., & Bjørner, N. (2008). Z3: An Efficient SMT Solver. In C. R. Ramakrishnan & J. Rehof (Eds.), *Tools and Algorithms for the Construction and Analysis of Systems* (Vol. 4963, pp. 337–340). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-78800-3_24

Dubuis, J. O., Tkacik, G., Wieschaus, E. F., Gregor, T., & Bialek, W. (2013). Positional information, in bits. *Proceedings of the National Academy of Sciences*, *110*(41), 16301–16308. https://doi.org/10.1073/pnas.1315642110

Dunn, S.- J., Martello, G., Yordanov, B., Emmott, S., & Smith, A. G. (2014). Defining an essential transcription factor program for naive pluripotency. *Science*, *344*(6188), 1156–1160. https://doi.org/10.1126/science.1248882

Dunn, Sara-Jane, Kugler, H., & Yordanov, B. (2019). Formal Analysis of Network Motifs Links Structure to Function in Biological Programs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1–1. https://doi.org/10.1109/TCBB.2019.2948157

Egawa, T., Tillman, R. E., Naoe, Y., Taniuchi, I., & Littman, D. R. (2007). The role of the Runx transcription factors in thymocyte differentiation and in homeostasis of naive T cells. *Journal of Experimental Medicine*, *204*(8), 1945–1957. https://doi.org/10.1084/jem.20070133

Fiers, M. W. E. J., Minnoye, L., Aibar, S., Bravo González-Blas, C., Kalender Atak, Z., & Aerts, S. (2018). Mapping gene regulatory networks from single-cell omics data. *Briefings in Functional Genomics*, *17*(4), 246–254. https://doi.org/10.1093/bfgp/elx046

Gilbert, L. A., Horlbeck, M. A., Adamson, B., Villalta, J. E., Chen, Y., Whitehead, E. H., Guimaraes, C., Panning, B., Ploegh, H. L., Bassik, M. C., Qi, L. S., Kampmann, M., & Weissman, J. S. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*, *159*(3), 647–661. https://doi.org/10.1016/j.cell.2014.09.029

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., & Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, *13*(10), 845–848. https://doi.org/10.1038/nmeth.3971

Hamey, F. K., Nestorowa, S., Kinston, S. J., Kent, D. G., Wilson, N. K., & Göttgens, B. (2017). Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. *Proceedings of the National Academy of Sciences*, *114*(23), 5822–5829. https://doi.org/10.1073/pnas.1610609114

Hosokawa, H., Romero-Wolf, M., Yui, M. A., Ungerbäck, J., Quiloan, M. L. G., Matsumoto, M., Nakayama, K. I., Tanaka, T., & Rothenberg, E. V. (2018). Bcl11b sets pro-T cell fate by site-specific cofactor recruitment and by repressing Id2 and Zbtb16. *Nature Immunology*, *19*(12), 1427–1440. https://doi.org/10.1038/s41590-018-0238-4

Hosokawa, H., & Rothenberg, E. V. (2020). How transcription factors drive choice of the T cell fate. *Nature Reviews Immunology*. https://doi.org/10.1038/s41577-020-00426-6

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, *316*(5830), 1497–1502. https://doi.org/10.1126/science.1141319

Jojic, V., Shay, T., Sylvia, K., Zuk, O., Sun, X., Kang, J., Regev, A., & Koller, D. (2013). Identification of transcriptional regulators in the mouse immune system. *Nature Immunology*, *14*(6), 633–643.

Kamimoto, K., Hoffmann, C. M., & Morris, S. A. (2020). *CellOracle: Dissecting cell identity via network inference and in silico gene perturbation* [Preprint]. Genomics. https://doi.org/10.1101/2020.02.17.947416

Kirouac, D. C., Madlambayan, G. J., Yu, M., Sykes, E. A., Ito, C., & Zandstra, P. W. (2009). Cell–cell interaction networks regulate blood stem and progenitor cell fate. *Molecular Systems Biology*, *5*. https://doi.org/10.1038/msb.2009.49

Krishnaswamy, S., Spitzer, M. H., Mingueneau, M., Bendall, S. C., Litvin, O., Stone, E., Pe'er, D., & Nolan, G. P. (2014). Conditional density-based analysis of T cell signaling in single-cell data. *Science*, *346*(6213), 1250689. https://doi.org/10.1126/science.1250689

Kueh, H. Y., & Rothenberg, E. V. (2012). Regulatory gene network circuits underlying T cell development from multipotent progenitors: Regulatory gene network circuits

underlying T cell development. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *4*(1), 79–102. https://doi.org/10.1002/wsbm.162

Kueh, H. Y., Yui, M. A., Ng, K. K. H., Pease, S. S., Zhang, J. A., Damle, S. S., Freedman, G., Siu, S., Bernstein, I. D., Elowitz, M. B., & Rothenberg, E. V. (2016). Asynchronous combinatorial action of four regulatory factors activates Bcl11b for T cell commitment. *Nature Immunology*, *17*(8), 956–965. https://doi.org/10.1038/ni.3514

Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Ferrante, T. C., Terry, R., Turczyk, B. M., Yang, J. L., Lee, H. S., Aach, J., Zhang, K., & Church, G. M. (2015). Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nature Protocols*, *10*(3), 442–458. https://doi.org/10.1038/nprot.2014.191

Li, L., Leid, M., & Rothenberg, E. V. (2010). An Early T Cell Lineage Commitment Checkpoint Dependent on the Transcription Factor Bcl11b. *Science*, *329*(5987), 89–93. https://doi.org/10.1126/science.1188989

Li, P., & Elowitz, M. B. (2019). Communication codes in developmental signaling pathways. *Development*, *146*(12), dev170977. https://doi.org/10.1242/dev.170977

Lipsitz, Y. Y., Timmins, N. E., & Zandstra, P. W. (2016). Quality cell therapy manufacturing by design. *Nature Biotechnology*, *34*(4), 393–400. https://doi.org/10.1038/nbt.3525

Longabaugh, W. J. R., Zeng, W., Zhang, J. A., Hosokawa, H., Jansen, C. S., Li, L., Romero-Wolf, M., Liu, P., Kueh, H. Y., Mortazavi, A., & Rothenberg, E. V. (2017). Bcl11b and combinatorial resolution of cell fate in the T-cell gene regulatory network. *Proceedings of the National Academy of Sciences*, *114*(23), 5800–5807. https://doi.org/10.1073/pnas.1610617114

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: A tutorial. *Molecular Systems Biology*, *15*(6). https://doi.org/10.15252/msb.20188746

Macqueen, J. (1967). SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS. *MULTIVARIATE OBSERVATIONS*, 17.

Maity, A., & Wollman, R. (2020). Information transmission from NFkB signaling dynamics to gene expression. *PLOS Computational Biology*, *16*(8), e1008011. https://doi.org/10.1371/journal.pcbi.1008011

Masuda, K., Kakugawa, K., Nakayama, T., Minato, N., Katsura, Y., & Kawamoto, H. (2007). T Cell Lineage Determination Precedes the Initiation of *TCR* β Gene Rearrangement. *The Journal of Immunology*, *179*(6), 3699–3706. https://doi.org/10.4049/jimmunol.179.6.3699

Moffitt, J. R., Hao, J., Wang, G., Chen, K. H., Babcock, H. P., & Zhuang, X. (2016). High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences*, *113*(39), 11046–11051. https://doi.org/10.1073/pnas.1612826113

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A. J., Tanaka, Y., Wilkinson, A. C., Buettner, F., Macaulay, I. C., Jawaid, W., Diamanti, E., Nishikawa, S.-I., Piterman, N., Kouskoff, V., Theis, F. J., Fisher, J., & Göttgens, B. (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnology*, *33*(3), 269–276. https://doi.org/10.1038/nbt.3154

Ng, K. K., Yui, M. A., Mehta, A., Siu, S., Irwin, B., Pease, S., Hirose, S., Elowitz, M. B., Rothenberg, E. V., & Kueh, H. Y. (2018). A stochastic epigenetic switch controls the dynamics of T-cell lineage commitment. *ELife*, *7*, e37851. https://doi.org/10.7554/eLife.37851

Ostblom, J., Nazareth, E. J. P., Tewary, M., & Zandstra, P. W. (2019). Context-explorer: Analysis of spatially organized protein expression in high-throughput screens. *PLOS Computational Biology*, *15*(1), e1006384. https://doi.org/10.1371/journal.pcbi.1006384

Peter, I. S., Faure, E., & Davidson, E. H. (2012). Predictive computation of genomic logic processing functions in embryonic development. *Proceedings of the National Academy of Sciences*, *109*(41), 16434–16442. https://doi.org/10.1073/pnas.1207852109

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., & Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*. https://doi.org/10.1038/s41592-019-0690-6

Prochazka, L., Benenson, Y., & Zandstra, P. W. (2017). Synthetic gene circuits and cellular decision-making in human pluripotent stem cells. *Current Opinion in Systems Biology*, *5*, 93–103. https://doi.org/10.1016/j.coisb.2017.09.003

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., & Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, *14*(10), 979–982. https://doi.org/10.1038/nmeth.4402

Qu, Z., Garfinkel, A., Weiss, J. N., & Nivala, M. (2011). Multi-scale modeling in biology: How to bridge the gaps between scales? *Progress in Biophysics and Molecular Biology*, *107*(1), 21–31. https://doi.org/10.1016/j.pbiomolbio.2011.06.004

Radtke, F., Wilson, A., Stark, G., Bauer, M., van Meerwijk, J., MacDonald, H. R., & Aguet, M. (1999). Deficient T Cell Fate Specification in Mice with an Induced Inactivation of Notch1. *Immunity*, *10*(5), 547–558. https://doi.org/10.1016/S1074-7613(00)80054-0

Rothenberg, E. V. (2019). Causal Gene Regulatory Network Modeling and Genomics: Second-Generation Challenges. *Journal of Computational Biology*, *26*(7), 703–718. https://doi.org/10.1089/cmb.2019.0098

Rothenberg, E. V., Hosokawa, H., & Ungerbäck, J. (2019). Mechanisms of Action of Hematopoietic Transcription Factor PU.1 in Initiation of T-Cell Development. *Frontiers in Immunology*, *10*, 228. https://doi.org/10.3389/fimmu.2019.00228

Semrau, S., & van Oudenaarden, A. (2015). Studying Lineage Decision-Making In Vitro: Emerging Concepts and Novel Tools. *Annual Review of Cell and Developmental Biology*, *31*(1), 317–345. https://doi.org/10.1146/annurev-cellbio-100814-125300

Shukla, S., Langley, M. A., Singh, J., Edgar, J. M., Mohtashami, M., Zúñiga-Pflücker, J. C., & Zandstra, P. W. (2017). Progenitor T-cell differentiation from hematopoietic stem cells using Delta-like-4 and VCAM-1. *Nature Methods*, *14*(5), 531–538. https://doi.org/10.1038/nmeth.4258

Song, L., Langfelder, P., & Horvath, S. (2012). Comparison of co-expression measures: Mutual information, correlation, and model based indices. *BMC Bioinformatics*, *13*(1), 328. https://doi.org/10.1186/1471-2105-13-328

Sun, N., Yu, X., Li, F., Liu, D., Suo, S., Chen, W., Chen, S., Song, L., Green, C. D., McDermott, J., Shen, Q., Jing, N., & Han, J.-D. J. (2017). Inference of differentiation

time for single cell transcriptomes using cell population reference data. *Nature Communications*, *8*(1), 1856. https://doi.org/10.1038/s41467-017-01860-2

Swat, M. H., Thomas, G. L., Belmonte, J. M., Shirinifard, A., Hmeljak, D., & Glazier, J. A. (2012). Multi-Scale Modeling of Tissues Using CompuCell3D. In *Methods in Cell Biology* (Vol. 110, pp. 325–366). Elsevier. https://doi.org/10.1016/B978-0-12-388403-9.00013-8

Takahashi, K., & Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, *126*(4), 663–676. https://doi.org/10.1016/j.cell.2006.07.024

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., Bierie, B., Mazutis, L., Wolf, G., Krishnaswamy, S., & Pe'er, D. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, *174*(3), 716-729.e27. https://doi.org/10.1016/j.cell.2018.05.061

Weber, B. N., Chi, A. W.-S., Chavez, A., Yashiro-Ohtani, Y., Yang, Q., Shestova, O., & Bhandoola, A. (2011). A critical role for TCF-1 in T-lineage specification and differentiation. *Nature*, *476*(7358), 63–68. https://doi.org/10.1038/nature10279

Weerkamp, F., Luis, T. C., Naber, B. A. E., Koster, E. E. L., Jeannotte, L., van Dongen, J. J. M., & Staal, F. J. T. (2006). Identification of Notch target genes in uncommitted T-cell progenitors: No direct induction of a T-cell specific gene program. *Leukemia*, *20*(11), 1967–1977. https://doi.org/10.1038/sj.leu.2404396

Wojciechowski, J., Lai, A., Kondo, M., & Zhuang, Y. (2007). E2A and HEB Are Required to Block Thymocyte Proliferation Prior to Pre-TCR Expression. *The Journal of Immunology*, *178*(9), 5717–5726. https://doi.org/10.4049/jimmunol.178.9.5717

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1), 15. https://doi.org/10.1186/s13059-017-1382-0

Yachie-Kinoshita, A., Onishi, K., Ostblom, J., Langley, M. A., Posfai, E., Rossant, J., & Zandstra, P. W. (2018). Modeling signaling-dependent pluripotency with Boolean logic to predict cell fate transitions. *Molecular Systems Biology*, *14*(1). https://doi.org/10.15252/msb.20177952

Yordanov, B., Dunn, S.-J., Kugler, H., Smith, A., Martello, G., & Emmott, S. (2016). A method to identify and analyze biological programs through automated reasoning. *Npj Systems Biology and Applications*, *2*(1), 16010. https://doi.org/10.1038/npjsba.2016.10

Yui, M. A., Feng, N., & Rothenberg, E. V. (2010). Fine-Scale Staging of T Cell Lineage Commitment in Adult Mouse Thymus. *The Journal of Immunology*, *185*(1), 284–293. https://doi.org/10.4049/jimmunol.1000679

Yui, M. A., & Rothenberg, E. V. (2014). Developmental gene networks: A triathlon on the course to T cell identity. *Nature Reviews Immunology*, *14*(8), 529–545. https://doi.org/10.1038/nri3702

Zhou, W., Yui, M. A., Williams, B. A., Yun, J., Wold, B. J., Cai, L., & Rothenberg, E. V. (2019). Single-Cell Analysis Reveals Regulatory Gene Expression Dynamics Leading to Lineage Commitment in Early T Cell Development. *Cell Systems*, *9*(4), 321-337.e9. https://doi.org/10.1016/j.cels.2019.09.008