Methods for multiomic data analysis

**Archetypal Analysis**

**Fitting a transcription factor network**

DIRECT NET analysis resulted in a list of 1216 TF motif-regulatory region pairs, each with an associated “motif score.” To build a transcription factor network from these results, we add all connections to a directed graph of TF-target gene interactions, which resulted in a network with 66 vertices and 852 edges. This reduced number of connections is due to the redundancy of some TF motif-regulatory region pairs (i.e. one TF binds multiple regulatory regions for a single target gene). Connections between TFs from a single family with the same TF binding motif and their target genes were manually added to the network (e.g. RORA and RORB shared target genes). For any nodes without target genes (i.e. sink nodes), such as ASCL1, we used the FIGR analysis to add data-supported regulatory relationships. Using the motif scores from DIRECT NET, the top eight regulators of each node were kept. Finally, remaining sink nodes with no target genes were removed since their presence does not affect the dynamics of network simulations. This resulted in a network of 49 vertices and 329 edges.

Using BooleaBayes version 0.1.XXX, we fit this resulting network as described in Wooten et al., 2019 XXX to the preprocessed combined scRNA-seq data. Briefly, we split the dataset into a training set (80%) and test set (20%). The scRNA-seq data is then scaled between 0 and 1 and then binarized using a cutoff of 0.5. For each node in the network, we use this data to fit a probabilistic Boolean function based on its parent nodes while accounting for sparsity of data (Wooten et al., 2019). The result gives a probabilistic rule for updating a particular node in the network based on the state of its parent nodes and is used in simulations as described below.

To determine how well our predictive rules fit the data, we compute several metrics comparing the predicted probability of a node turning ON given a parent state and the actual expression (scaled between 0 and 1) of that node in the test data. As shown in Figure SXXX, most of the nodes of the network have accuracies between XXX and 1, with precision between XXX and XXX, and recall between XXX and XXX. Overall, the area under the curve of the receiver operating characteristic (ROC) curve for each node ranged from XXX and XXX, with an average of XXX.

**Network simulations with BooleaBayes**

The transcription factor network was then simulated using the probabilistic rules determined above with BooleaBayes version 0.1.XXX. Simulations were run are described in Wooten et al., 2019. Briefly, each simulation starts at the states associated with each datapoint. The nodes are updated asynchronously by choosing one at random and using the current state and that node’s rule. The simulations to determine attractors were run by searching the 2-TF basin of each sample datapoint to search for states with a higher probability of entering the state (P > 0.5 for all in-edges) than leaving the state (P < 0.5 for all out-edges). This analysis found XXX attractors, which are labeled by the closest average state of the clusters in the data. Attractor stability was determined by running 500 random-walk simulations starting at each attractor and counting the number of steps it takes to leave a basin of different radii around the attractor. For computational efficiency, random walks that did not leave a basin after 500 steps were truncated. *In silico* perturbations were run as 500 random walks from each attractor with a single node held ON (activation) or OFF (silencing) during the simulation. The number of steps it takes to leave a 2-TF basin were counted and compared to the count without the perturbation. The difference in averages gives a destabilization score, which was used to rank the perturbations by most destabilizing to a particular attractor (or group of attractors).

**RNA Velocity Analysis**