BIOINFORMATICS

Defining developmental grammar

An iterative clustering approach finds the few genes that mark discrete cell states and their transitions during early mouse development.

Language needs grammar. You may know a lot of words, but you cannot speak in intelligible sentences without adhering to grammatical rules. For Sharad Ramanathan from Harvard University, a similar principle applies to developmental biology. To go from complicated gene expression patterns to understanding how cells transition from one state to another along a trajectory, one needs to know the grammar or code behind lineage development.

Ramanathan explains that to find cell types and the lineage relationships between them, some knowledge of distance in gene expression space is needed. He thinks that previous tools to measure distances included too many data points, which led to uninformative distance calculations. “We know from decades of developmental biology that there are a few key factors that control development,” he says. “The core is sparse; few things matter, many things don’t.” The key problem is of course how to know which data should be ignored and which kept.

Leon Furchtgott and Samuel Melton, graduate students in the Ramanathan lab, decided to reverse the common trend of taking feature-rich data and considering the expression patterns of all genes at all time points; instead, they compared the expression levels of only one gene at a time in three cell types. Their goal was to develop a statistical framework to determine which genes one should use to measure distances.

They first used gene expression data from 41 cell types along B and T cell development. Most genes did not have distinct expression patterns in any given trio of cell types, but a few showed either high or low expression in one cell type. Genes highly expressed in only one cell type were

designated as marker genes and were useful for characterizing this cell type. Genes with significantly weaker expression in only one cell type, termed transition genes, proved informative for determining developmental relationships between cells. The cell type with minimal transition gene expression cannot reside in the middle but has to be at either end of the lineage. By determining transition genes of all possible cell triplets, Furchtgott and Melton ultimately reconstructed the entire hematopoietic lineage tree *de novo* (Furchtgott *et al.*, 2017).

The framework was put to a challenging test by graduate student Sumin Jang and postdoctoral fellow Sandeep Choubey in Ramanathan’s group when they applied it to early mouse development.

Taking single-cell RNA-seq data from mouse embryonic stem cells during early differentiation, obtained in collaboration with Boaz Levi from the Allen Institute, the researchers started by specifying twelve seed clusters based on transcription-factor expression representing potential cell states. Then they considered every possible group of three clusters, and for each gene they determined the probability of it being a transition or marker gene. In multiple iterations they revised their clusters based on the transition genes they found, until the results converged on nine distinct cell states connected in a lineage tree. These matched known embryonic cell types from naïve mouse embryonic stem cells (mESCs) to primed epiblast cells, mesendoderm and ectoderm progenitor cells (Jang *et al.*, 2017).

To confirm that these cell states were discrete, the team sampled cells every 24 h during a 5-d differentiation protocol and labeled the product of one marker gene characteristic for each cell type. They saw a clear separation of flow cytometry peaks that corresponded with their nine cell states.

Clustering of single-cell gene expression data allows the identification of transition genes which are the basis of lineage trees. Illustration inspired by Jang *et al.*, *eLife* (2017).

The next question was whether the cell states were computational artifacts or could also be distinguished based on their function. The researchers combined marker and transition genes into modules from which they inferred gene regulatory networks specific to each state. They then experimentally tested each network’s response to stimuli. “These cells all have unique physiological responses to perturbation,” says Ramanathan. “This is the validation of the computational algorithm. We define them functionally.” He foresees that they could take single-cell gene expression data and predict the lineage and developmental potential of a cell.

Since cell states can be defined by only a few transition genes, it will also be possible to follow these genes in real time. “You can start seeing single-cell decision making in live cells,” predicts Ramanathan. Observing transition genes can begin to address basic questions such as whether a cell divides before making a decision, or whether mother and daughter or sister cells make the same decision. Exploiting sparsity will be the key to the full answers.

Nicole Rusk

RESEARCH PAPERS

Furchtgott, L.A. *et al.* Discovering sparse transcription factor codes for cell states and state transitions during development. *eLife* 6, e20488 (2017).

Jang, S. *et al.* Dynamics of embryonic stem cell differentiation inferred from single-cell transcriptomics show a series of transitions through discrete cell states. *eLife* 6, e20487 (2017).