Robust sub-population discovery using self-pruning decision tree

Single-cell RNA sequencing enables unbiased analysis of expression patterns but researchers don't have the tools for appropriate decision making during the analysis. Our general aim is to introduce unbiased data-driven strategies to identify the appropriate number of robust subpopulations, their discriminatory defining markers and the relationships between populations. I outlined a solid plan below but it is open for innovative ideas, improvements and suggestions.

Challenges

- What is the highest resolution that can be defined reproducibly?
- What is the relationship between populations (similar/distinct lineage)?
- What are the cells found in 'undefined'/mixed clusters? Can they be reassigned to well defined clusters?
- Which clusters are detected robustly and which are just experiment/batch specific?
- What are the discriminatory and robust markers defining each population? Are they discriminatory in terms of both detection?

Work plan

1. Introduce hierarchy to clustering [unsupervised ML]

- identifying the appropriate hierarchical clustering methodology
- test sensitivity to confounders and zero-inflation
- test reproducibility across experiments
- compare to established clustering methods

2. **Build a decision tree robust for expression false negatives [supervised ML]**
   - develop a robust strategy for decision tree rules using bootstrap analysis
   - consider giving higher importance to cells proximal to cluster centroids and less to “not important” genes
   - test generalizability across experiments
   - try to reassign cells based on rules and include undefined cells

3. **Assess hierarchy reproducibility on an independent experiment [Stats/CS]**
   - define cluster and tree similarity
   - measure hierarchy reproducibility
   - identify irreproducible subpopulations
   - find the reproducible common ancestor

4. **Prune and merge decision trees [CS]**
   - identify the optimal pruning point for each subtree by comparing the two experiments
   - measure cluster homogeneity
   - merge decision tree rules to the reproducible rules

5. **Define gene signatures and visualize tree on tSNE [Visualization]**
   - define robust gene signatures defining each cluster beyond the decision tree rules
   - measure signature similarity across experiments
   - find the appropriate tSNE parameters
   - overlay tree hierarchy on tSNE annotated by (1) cluster membership (2) decision tree genes (3) robust cluster gene signatures activation
   - generating additional figures for the manuscript (heatmaps, illustrations etc)

6. Writing

Feel free to read about the problems I mentioned here so you’re familiar with the terminology on Monday. We’ll split into about 3 groups working on the different subsets of this project in parallel. It would be great if you to think where would you fit best and let me know so I can plan accordingly.

ML = machine learning.
Stats = statistics
CS = computer science

Assaf Magen - Team Lead