2017-06

Cloud-based highly parallel execution of t-SNE and SPADE with metaclustering for analysis and visualization of large single-cell datasets

Cloud-based Highly Parallel Execution of t-SNE and SPADE with Metaclustering for Analysis and Visualization of Large Single-cell Datasets

Abstract

The use of machine learning techniques, in particular unsupervised clustering and dimensionality reduction algorithms, is quickly becoming a standard workflow for identifying and visualizing biological populations from within high-dimensional data. These methods allow researchers to approach data analysis without the bias and subjectivity that has traditionally been standard in the field.

Algorithms have context-dependent strengths and weaknesses. Across algorithms, an inability to scale computation to large datasets is a common theme. Most algorithms are designed and distributed to run on individual computers where memory and CPU are quickly exhausted by large datasets. Even when high-performance compute resources are available, algorithms often don't scale to large datasets as a fundamental property of their design. If they do, it might result in an untenable increase in runtime or diminished quality of results.

t-SNE and SPADE are two well-published algorithms that suffer problems as discussed above after datasets exceed a number of observations on the order of 1 million. This study introduces an alternative approach to the use of SPADE and t- SNE whereby a dataset is divided and distributed across numerous compute nodes in the cloud to process independently in parallel. The results of each computation are then combined in a metaclustering step for final visualization and analysis. The improvement in execution speed as a function of degree of parallelization is established. The method is validated against a non-parallel analysis of the same dataset to establish concordance of identified populations. The workflow is executed on Cytobank for portability to other researchers.