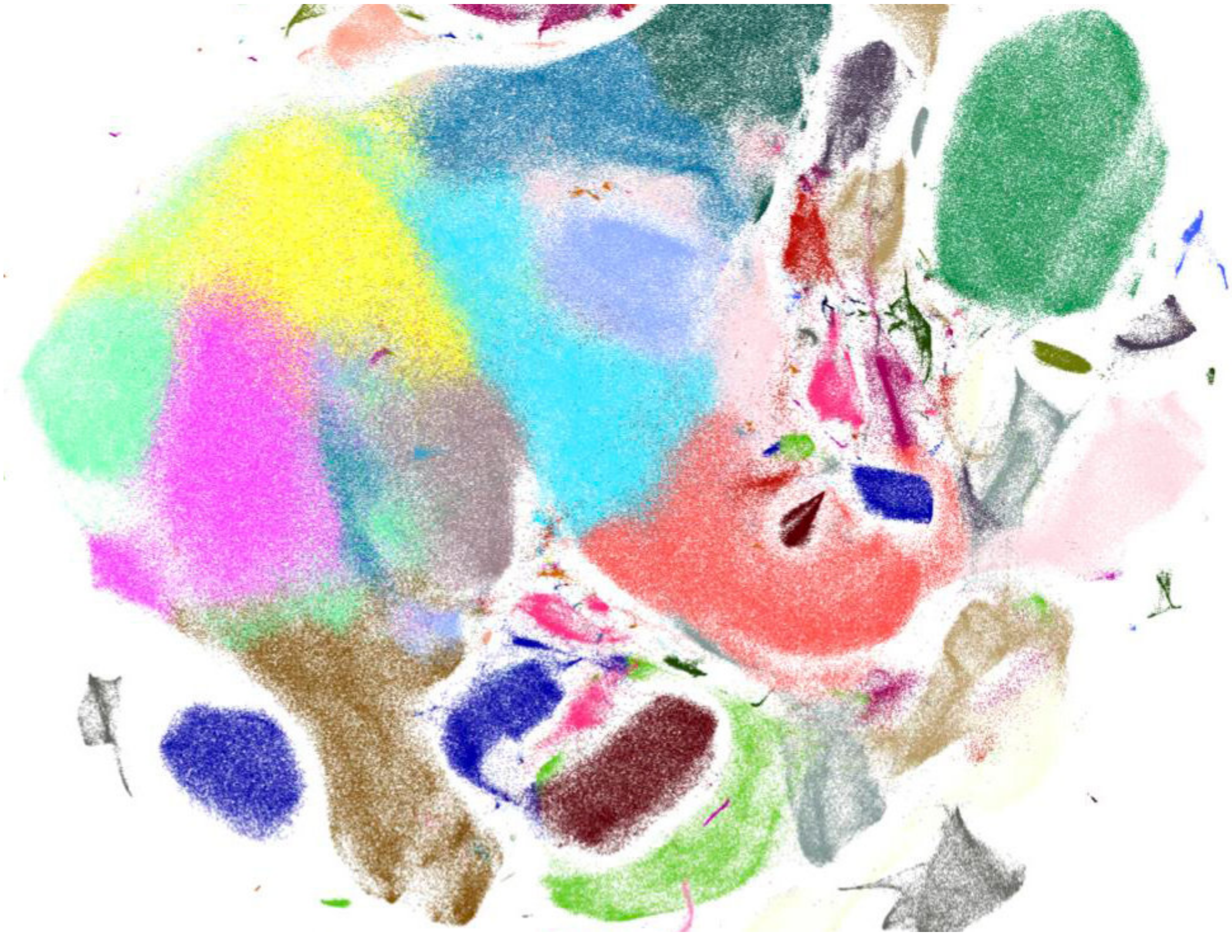
PHYS.ORG

Software package processes huge amounts of single-cell data

February 13 2018



Visualization of gene expression patterns of murine brain cells generated with Scanpy. Credit: Helmholtz Zentrum München

Scientists from the Helmholtz Zentrum München have developed a program that for managing enormous datasets. The software, called Scanpy, is a candidate for analyzing the Human Cell Atlas, and has recently been published in *Genome Biology*.

"It's about analyzing gene-expression data of a large number of individual cells," explains lead author Alex Wolf of the Institute of Computational Biology (ICB) at Helmholtz Zentrum München. He developed Scanpy together with his colleague Philipp Angerer in the Machine Learning Group of Prof. Dr. Dr. Fabian Theis. In addition to his position at Helmholtz Zentrum, Theis is also a professor of mathematical modelling of biological systems at the Technical University of Munich. "New technical advances generate several orders of magnitude more data with a correspondingly greater information content," Theis says. "However, the historically evolved software infrastructure for gene-expression analysis simply wasn't designed to cope with the new challenges. New analytic methods are therefore needed."

The race for the Human Cell Atlas

According to Theis, a major international research project could also benefit from the software. A team of international scientists is compiling a reference database, called the Human Cell Atlas, which holds data on the gene activity of all human cell types. "For this project, and in a growing number of other projects in which databases are combined, it is important to have scalable software," says Theis. It is therefore no surprise that Scanpy is currently a candidate for helping to analyze the Human Cell Atlas.

"The publication of Scanpy marks the first software that allows comprehensive analysis of large gene-expression datasets with a broad range of machine-learning and statistical methods," explains Wolf, describing the achievement. "The software is already being used by a number of groups around the world, notably at the Broad Institute of Harvard University and the Massachusetts Institute of Technology, MIT."

Technologically, the application is a trailblazing development: Whereas biostatistics programs are traditionally written in the programming language R, Scanpy is based on the Python language, the dominant language in the machine learning community. Another new feature is that graph-based algorithms lie at the heart of Scanpy. Unlike the usual approach of regarding cells as points in a coordinate system within gene-expression space, the algorithms use a graph-like coordinate system. Instead of characterizing a single cell by the expression value for thousands of genes, the system simply characterizes cells by identifying their closest neighbors - very much like the connections in social networks. In fact, to identify cell types, Scanpy uses the same algorithms as Facebook does for identifying communities.

More information: F. Alexander Wolf et al, SCANPY: large-scale single-cell gene expression data analysis, *Genome Biology* (2018). DOI: 10.1186/s13059-017-1382-0

Provided by Helmholtz Association of German Research Centres

Citation: Software package processes huge amounts of single-cell data (2018, February 13) retrieved 2 May 2024 from https://phys.org/news/2018-02-software-package-huge-amounts-single-cell.html