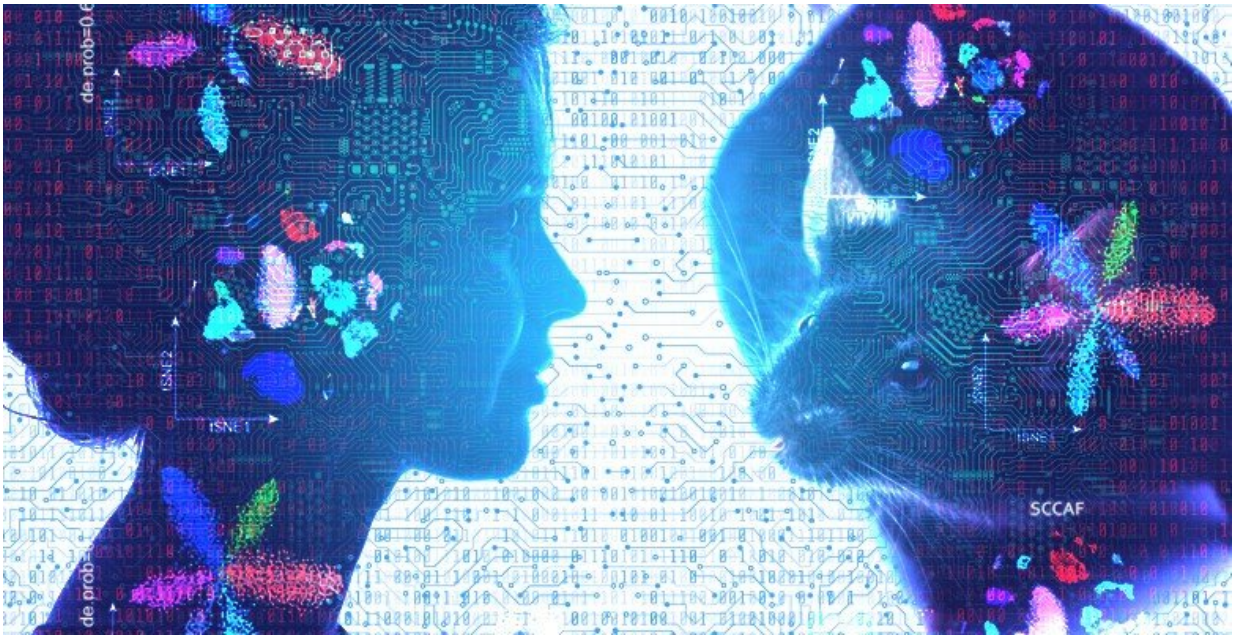
Identifying cell types from single-cell RNA sequencing data automatically

May 19 2020



Credit: Spencer Phillips/ EMBL-EBI

Identifying different types of cells within a tissue or an organ can be very challenging and time-consuming. Methods to identify cell types from single-cell RNA sequencing data have been proposed, but they all fall short in discovering potentially new cell types. Researchers from the Wellcome Sanger Institute and EMBL's European Bioinformatics Institute (EMBL-EBI) have created a new method called Single Cell Clustering Assessment Framework (SCCAF) that bridges this gap.

Published today in *Nature Methods*, this automated method uses machine learning and can replicate manual, expert annotations that are normally used for this task, and can characterise new cell types.

All somatic cells in a multicellular organism have the same genome, yet they perform a variety of functions. This functional diversity occurs between cells of different types (skin cells and neurons, for instance), but also between states of the same cell lineage as it differentiates.

Historically, researchers have identified cell types or states based on visible features or the expression of a handful of genes. Single-cell RNA sequencing (scRNA-seq) has brought high-throughput gene expression data into the picture.

A cell's gene expression pattern (which genes are expressed at what level) serves as a proxy for its function and allows scientists to classify or "cluster" that cell with others that have the same function. Until now, annotating cells from scRNA-seq data has required time-consuming human intervention, with automated methods unable to identify cell types or states that had not been previously annotated by human experts.

The researchers came up with a method that uses machine learning to address these challenges.

Single Cell Clustering Assessment Framework (SCCAF) starts by using a clustering algorithm to group the cells of a sample into many clusters, based on their gene expression patterns. Each cell cluster is split into a "training set" and a "testing set" for the second stage of the analysis. A classifying model then takes over, using the training set to learn to distinguish cell clusters, and predicting likely clusters in the testing set. The model's accuracy is assessed by comparing its prediction with the original clusters.

"The model repeats the training and testing steps for each cell cluster, gradually merging indistinguishable clusters, until its accuracy reaches a good enough level. Finally, our Single Cell Clustering Assessment Framework lists a set of feature genes to characterise each annotated cluster,"
says Dr. Zhichao Miao, the first author on the paper from EMBL-EBI and the Wellcome Sanger Institute.

This Single Cell Clustering Assessment Framework method has been shown to be highly reliable and fast.

"We've tested the method on many existing large-scale datasets of human and mouse gene expression, treating human annotation as a gold standard. Our method can reproduce human annotation in an automated manner. By minimising human involvement in data processing, we solve the most important bottleneck in high-throughput projects, such as the Human Cell Atlas,"
says Dr. Alvis Brazma, a senior author and Functional Genomics Senior Team Leader at EMBL-EBI.

Not only does SCCAF reproduce and refine existing cell type classification, it also helps reveal new cell types and states from unannotated samples. The new method will be implemented in large-scale projects, including the Single Cell Expression Atlas and the Human Cell Atlas.

"The Human Cell Atlas initiative is a global consortium to map every cell type in the human body, to understand health and disease. The new automated cell-clustering method will enable us to identify cell types much more easily than before, helping us expand our understanding of cellular function and diversity," says Dr. Sarah Teichmann, a senior author from the Wellcome Sanger Institute, and co-chair of the Human Cell Atlas Organising Committee.

More information: Zhichao Miao et al. Putative cell type discovery from single-cell gene expression data, *Nature Methods* (2020). DOI: 10.1038/s41592-020-0825-9

Provided by Wellcome Trust Sanger Institute