Computational Biology

Using artificial intelligence for error correction in single cell analyses

Modern technology makes it possible to sequence individual cells and to identify which genes are currently being expressed in each cell. These methods are sensitive and consequently error prone. Devices, environment and biology itself can be responsible for failures and differences between measurements. Researchers at Helmholtz Zentrum München joined forces with colleagues from the Technical University of Munich (TUM) and the British Wellcome Sanger Institute and have developed algorithms that make it possible to predict and correct such sources of error. The work was published in 'Nature Methods' and 'Nature Communications'.

A visionary project of enormous scope, the Human Cell Atlas aims to map out all the tissues of the human body at various time points with the goal of creating a reference database for the development of personalized medicine, i.e. the ability to distinguish healthy from diseased cells. This is made possible by a technology known as single-cell RNA sequencing, which helps researchers understand exactly which genes are switched on or off at any given moment in these tiny components of life. "From a methodological point of view, this represents an enormous leap forward. Previously, such data could only be obtained from large groups of cells because the measurements required so much RNA," Maren Büttner explains. "So the results were always only the average of all the cells used. Now we're able to get precise data for every single cell," says the doctoral student at the Institute of Computational Biology (ICB) of the Helmholtz Zentrum München.

The increased sensitivity of the technique, however, also means increased susceptibility to the batch effect. "The batch effect describes fluctuations between measurements that can occur, for example, if the temperature of the device deviates even slightly or the processing time of the cells changes," Maren Büttner explains. Although several models exist for the correction of these deviations, those methods are highly dependent on the actual magnitude of the effect. "We therefore developed a user-friendly, robust and sensitive measure called kBET that quantifies differences between experiments and therefore facilitates the comparison of different correction results," Büttner says.

"Our new method, kBET, is a powerful tool for comparing batch-effect correction schemes, allowing researchers to study different single-cell RNA sequencing datasets. This has important implications in the future for data-integration, which is central to major initiatives such as the Human Cell Atlas," said Dr Sarah Teichmann, a corresponding author on the paper from the Wellcome Sanger Institute, UK.

Besides the batch effect, a phenomenon known as dropout events poses a major challenge in single-cell sequencing. "Let's say we sequence a cell and observe that a particular gene in the cell does not emit any signal at all," explains Dr. Dr. Fabian Theis, ICB Director and professor of Mathematical Modeling of Biological Systems at the TUM. "The underlying cause of this can be biological or technical in nature: either the gene is not being read by the sequencer because it is simply not expressed, or it was not detected for technical reasons," he explains.

To recognize these cases, bioinformaticians Gökcen Eraslan and Lukas Simon from Theis's group used a large number of sequences of many single cells and developed what is known as a deep learning algorithm, i.e. artificial intelligence which simulates learning processes that occur in humans (neural networks).*

Drawing on a new probabilistic model and comparing the original and reconstructed data, the algorithm determines whether the absence of a gene signal is due to a biological or technical failure. "This model even allows cell type-specific corrections to be determined without two different cell types becoming artificially similar," Fabian Theis says. "As one of the first deep learning methods in the field of single-cell genomics, the algorithm has the added benefit that it scales up well to handle data sets containing millions of cells."

But there is one thing the method is not– and this is important to emphasize: "We're not developing software to smooth out results. Our chief goal is to identify and correct errors," Fabian Theis explains. "We're able to share these data, which are as accurate as possible, with our colleagues worldwide and compare our results with theirs," – for example when the Helmholtz researchers contribute their algorithms and analyses to the Human Cell Atlas, because reliability and comparability of the data are of paramount importance.

Further information

* The new method, known as the deep count auto-encoder, learns to simplify the presentation of complex data by compressing them and reconstructing them afterwards.

Background:
The work published in 'Nature Methods' was developed in close collaboration with Dr. Sarah Amalia Teichmann from the Wellcome Sanger Institute. She is also the co-chair of the Human Cell Atlas Organizing Committee and in 2017 was the recipient of a Helmholtz International Fellow Award, which aims to promote cooperation between Helmholtz scientists and first-rate colleagues abroad, which in this case certainly proved successful.

Original Publications:
Büttner, M. et al. (2019): A test metric for assessing single-cell RNA-seq batch correction. Nature Methods, DOI: 10.1038/s41592-018-0254-1

Eraslan, G. & Simon, L.M. (2019): Single cell RNA-seq denoising using a deep count autoencoder. Nature Communications, DOI: 10.1038/s41467-018-07931-2

Related Articles:
Tackling Big Data – new graduate school for Data Science in Munich
Call for standards of epigenetic data
The Scanpy software processes huge amounts of single-cell data

The Helmholtz Zentrum München, the German Research Center for Environmental Health, pursues the goal of developing personalized medical approaches for the prevention and therapy of major common diseases such as diabetes, allergies and lung diseases. To achieve this, it investigates the interaction of genetics, environmental factors and lifestyle. The Helmholtz Zentrum München is headquartered in Neuherberg in the north of Munich and has about 2,300 staff members. It is a member of the Helmholtz Association, a community of 19 scientific-technical and medical-biological research centers with a total of about 37,000 staff members. www.helmholtz-muenchen.de/en

The Institute of Computational Biology (ICB) develops and applies methods for the model-based description of biological systems, using a data-driven approach by integrating information on multiple scales ranging from single-cell time series to large-scale omics. Given the fast technological advances in molecular biology, the aim is to provide and collaboratively apply innovative tools with experimental groups in order to jointly advance the understanding and treatment of common human diseases. www.helmholtz-muenchen.de/icb

The Technical University of Munich (TUM) is one of Europe's leading research universities, with around 550 professors, 41,000 students, and 10,000 academic and non-academic staff. Its focus areas are the engineering sciences, natural sciences, life sciences and medicine, combined with economic and social sciences. As TUM is an entrepreneurial university that actively promotes young talent and creates value for society, it benefits from having strong partners in science and industry. It is represented worldwide via the TUM Asia campus in Singapore and also has liaison offices in Beijing, Brussels, Cairo, Mumbai, San Francisco, and São Paulo. Nobel Prize winners and inventors such as Rudolf Diesel, Carl von Linde, and Rudolf Mössbauer have conducted research at TUM. In 2006 and 2012, it was recognized as a German "University of Excellence". In international rankings, TUM regularly places among the best universities in Germany. www.tum.de

Contact for the Media:
Communication Department, Helmholtz Zentrum München - German Center of Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany – Tel. +49 89 3187-2238 - E-mail: presse@helmholtz-muenchen.de

Scientific Contact:
Prof. Dr. Dr. Fabian Theis, Helmholtz Zentrum München - German Research Center for Environmental Health, Institute of Computational Biology, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany - Tel. +49 89 3187 4030 – E-mail: fabian.theis@helmholtz-muenchen.de