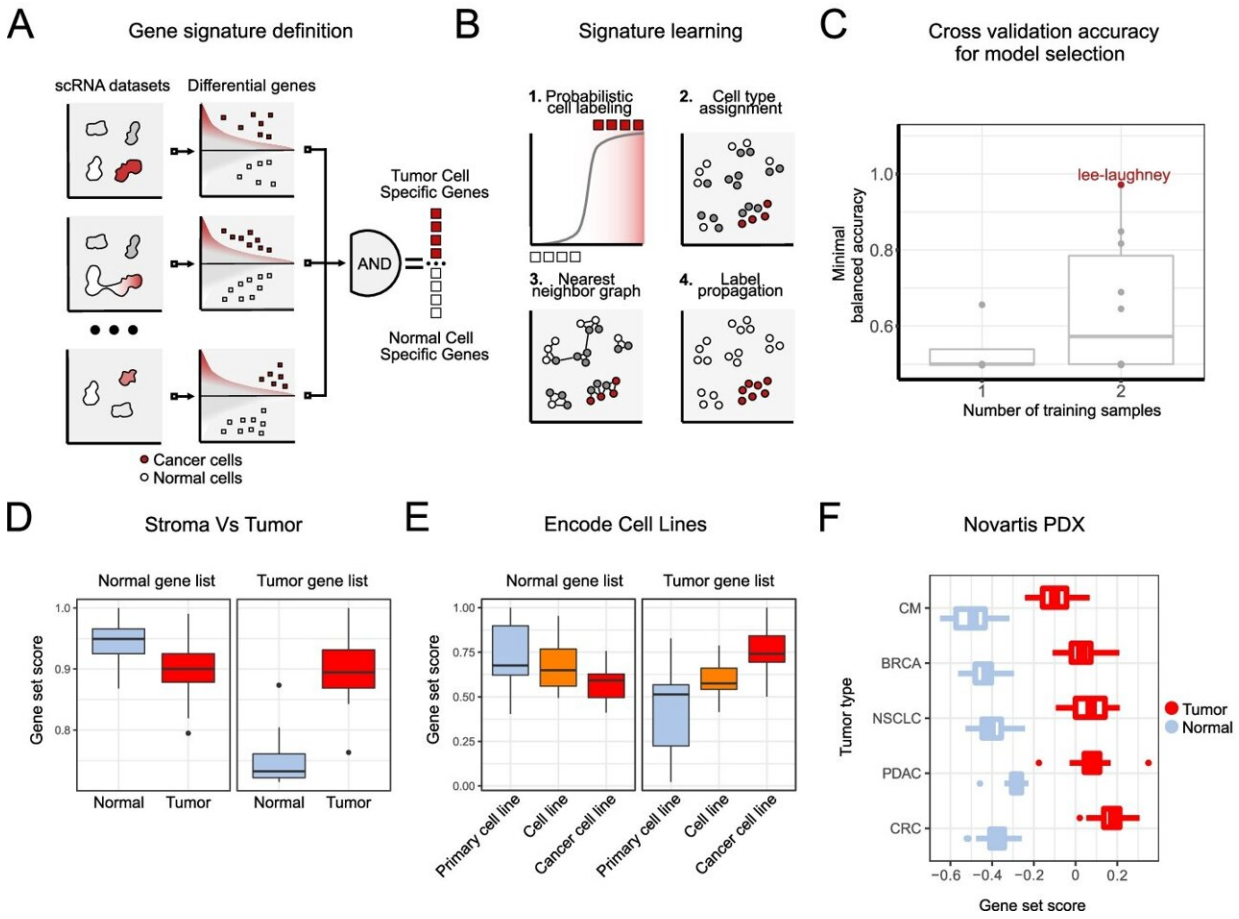
AI identifies cancer cells

June 10 2022



Integration of multiple datasets enables robust extraction of informative gene sets. A, B ikarus workflow. ikarus is a two-step procedure for classifying cells. In the first step, integration of multiple expert labeled datasets enables the extraction of robust gene markers. The gene markers are then used in a composite classifier consisting of logistic regression and network propagation. C Comparison of cross validation accuracy for signature derivation and model selection. Minimal balanced accuracy on the validation set was chosen as the metric of choice (i.e., worse performance on the test set). Models trained on just

one dataset achieved lower balanced accuracy than models trained on two datasets (p value given by the two sided Wilcoxon test is 0.063). The combination of colorectal cancer from Lee et al. and lung cancer from Laughney et al. achieved the highest minimal balanced accuracy of 0.97. D Comparison of gene signature scores in laser microdissected gastric cancer data. The normal gene list shows lower signature scores in cancer samples (p value 0.052, N = 8, Mood's median test), when compared to the cancer-associated normal tissue. The tumor gene signature is significantly higher for cancer samples than the normal tissue (p value 0.003, N = 8, Mood's median test). E Primary cells and cancer cell lines have significantly different gene signature distributions. The normal-cell gene signature shows a gradual reduction in gene signature score distribution when compared in primary cells, cell lines, and tumor cell lines. The gene signature shows the complete opposite effect. Cancer cell lines have the higher gene signature score distribution, followed by cell lines, and primary cells. Distributions were compared using pairwise Wilcoxon tests with BH-FDR correction. All adjusted p values were lower than 0.01. F Patient-derived xenografts (PDX) show significantly higher tumor gene signature score, than the normal gene signature score. The same pattern is observed in multiple cancer types. Normal and tumor signature distributions were compared using Wilcoxon tests, for each cancer type, followed by BH-FDR correction. All adjusted p values were lower than 0.01. Credit: *Genome Biology* (2022). DOI: 10.1186/s13059-022-02683-1

How do cancer cells differ from healthy cells? A new machine learning algorithm called "ikarus" knows the answer, reports a team led by MDC bioinformatician Altuna Akalin in the journal *Genome Biology*. The AI program has found a gene signature characteristic of tumors.

When it comes to identifying patterns in mountains of data, human beings are no match for artificial intelligence (AI). In particular, a branch of AI called machine learning is often used to find regularities in data sets—be it for stock market analysis, image and speech recognition, or the classification of cells. To reliably distinguish cancer cells from

healthy cells, a team led by Dr. Altuna Akalin, head of the Bioinformatics and Omics Data Science Platform at the Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), has now developed a machine learning program called "ikarus."

The program found a pattern in tumor cells that is common to different types of cancer, consisting of a characteristic combination of genes. According to the team's paper in the journal *Genome Biology*, the algorithm also detected types of genes in the pattern that had never been clearly linked to cancer before.

Machine learning essentially means that an algorithm uses training data to learn how to answer certain questions on its own. It does so by searching for patterns in the data that help it to solve problems. After the training phase, the system can generalize from what it has learned in order to evaluate unknown data. "It was a major challenge to get suitable training data where experts had already distinguished clearly between 'healthy' and 'cancerous' cells," relates Jan Dohmen, the first author of the paper.

A surprisingly high success rate

In addition, single-cell sequencing data sets are often noisy. That means the information they contain about the molecular characteristics of individual cells is not very precise—perhaps because a different number of genes is detected in each cell, or because the samples are not always processed the same way. As Dohmen and his colleague Dr. Vedran Franke, co-head of the study, reports, they sifted through countless publications and contacted quite a few research groups in order to get adequate data sets. The team ultimately used data from lung and colorectal cancer cells to train the algorithm before applying it to data sets of other kinds of tumors.

In the training phase, ikarus had to find a list of characteristic genes which it then used to categorize the cells. "We tried out and refined various approaches," Dohmen says. It was time-consuming work, as all three scientists relate. "The key was for ikarus to ultimately use two lists: one for cancer genes and one for genes from other cells," Franke explains. After the learning phase, the algorithm was able to reliably distinguish between healthy and tumor cells in other types of cancer as well, such as in tissue samples from liver cancer or neuroblastoma patients. Its success rate tended to be extraordinarily high, which surprised even the research group. "We didn't expect there to be a common signature that so precisely defined the tumor cells of different kinds of cancer," Akalin says. "But we still can't say if the method works for all kinds of cancer," Dohmen adds. To turn ikarus into a reliable tool for cancer diagnosis, the researchers now want to test it on additional kinds of tumors.

AI as a fully automated diagnostic tool

The project aims to go far beyond the classification of "healthy" versus "cancerous" cells. In initial tests, ikarus already demonstrated that the method can also distinguish other types (and certain subtypes) of cells from tumor cells. "We want to make the approach more comprehensive," Akalin says, "developing it further so that it can distinguish between all possible cell types in a biopsy."

In hospitals, pathologists tend only to examine tissue samples of tumors under the microscope in order to identify the various cell types. It is laborious, time-consuming work. With ikarus, this step could one day become a fully automated process. Furthermore, Akalin notes, the data could be used to draw conclusions about the tumor's immediate environment. And that could help doctors to choose the best therapy. For the makeup of the cancerous tissue and the microenvironment often indicates whether a certain treatment or medication will be effective or

not. Moreover, AI may also be useful in developing new medications. "Ikarus lets us identify genes that are potential drivers of cancer," Akalin says. Novel therapeutic agents could then be used to target these molecular structures.

A remarkable aspect of the publication is that it was prepared entirely during the COVID pandemic. All those involved were not at their usual desks at the Berlin Institute for Medical Systems Biology (BIMSB), which is part of the MDC. Instead, they were in home offices and only communicated with one another digitally. In Franke's view, therefore, "The project shows that a digital structure can be created to facilitate scientific work under these conditions."

More information: Altuna Akalin et al, Identifying tumor cells at the single-cell level using machine learning, *Genome Biology* (2022). DOI: 10.1186/s13059-022-02683-1. genomebiology.biomedcentral.co … 6/s13059-022-02683-1

Provided by Max Delbrück Center for Molecular Medicine