genomeweb
A CRAIN FAMILY BRAND

Participant Health Privacy at Risk From Single-Cell RNA Sequencing Data, Study Suggests

Oct 03, 2024 | staff reporter

Save for later

NEW YORK – New research suggests that it is possible to use "linking attacks" to identify health and phenotypic information about individuals based on their single-cell RNA sequencing data, even when expression quantitative trait locus (eQTL) data is not available or is drawn from another dataset.

"[W]e demonstrate that individuals in single-cell gene expression datasets are vulnerable to linking attacks, where attackers can infer their sensitive phenotypic information using publicly available tissue or cell type-specific expression quantitative trait loci (eQTLs) information," senior author Gamze Gürsoy, a biomedical informatician and computer scientist affiliated with Columbia University and the New York Genome Center, and her colleagues wrote in a paper published in *Cell* on Wednesday.

In contrast to past studies, which have tended to focus on bulk RNA-seq datasets and privacy breaches that can arise when individuals' eQTL data is available, the team showed that it is also possible to tease out genotype and phenotype profiles from relatively variable or "noisy" single-cell RNA-seq data in the presence or absence of corresponding eQTL data.

The researchers started by scrutinizing single-cell RNA-seq data on peripheral blood mononuclear cells profiled for the OneK1K project or for studies focused on systemic lupus erythematosus, putting together synthetic mixtures of individual cells from the datasets.

Similar to analyses focused on bulk RNA-seq samples, they were able to predict genotypes and related phenotypes based on single-cell count matrices — tables showing expression counts for individual genes, particularly with the help of cell type-specific eQTL data from the Genotype-Tissue Expression (GTEx) resource.

Taking the analysis one step further, the team next tested a computational method for predicting genotypes and related phenotypes in the absence of direct eQTL data.

Their results suggested that "linking attacks" can be accomplished using variant profiles unearthed in one eQTL study to link single-cell gene expression data to individual genotypes in another cohort, thus predicting their phenotypes.

"Our study suggests a clear conflict between the need for sharing single-cell gene expression matrices, which are immensely useful for understanding complex biological processes in health and disease, and protecting the privacy of research participants," the study authors explained.

Given such concerns, the team highlighted the need for secure storage for RNA-seq data, noting that some situations may warrant the use of statistical mitigation methods to introduce additional noise into some single-cell count matrix sets.

They also pointed to the need for appropriate laws and legislation aimed at protecting research participant data, along with informed consent policies that carefully spell out privacy risks to participants.

"The linking attacks in our study require only a quantitative molecular phenotype and its associated QTLs," the authors reported, adding that the method "can be applied to datasets involving not only gene expression but also chromatin accessibility or methylation."

Even so, they noted that the low genotyping accuracy of the approach "poses a limitation for attackers who aim to predict the genomes of study individuals without access to an external database, because they may not be able to fully predict genomes solely from count matrices."

Filed Under + Sequencing + Informatics + single-cell gene expression analysis + North America + Journal Study + RNA-seq + algorithm