

# Crowdsourcing for scientific discovery: Researchers find novel ways to analyze data for drug and target discovery

September 26 2016

---

In a unique project, researchers at the Icahn School of Medicine at Mount Sinai have crowdsourced the annotation and analysis of a large number of gene expression profiles from the National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO). More than 70 volunteers from 25 countries helped Mount Sinai researchers analyze the data, enabling the identification of new associations between genes, diseases, and drugs – something that a smaller number of unaided researchers, or an automated computer program, would not be able to achieve. An article published today in the journal *Nature Communications* describes the crowdsourcing project.

Omics repositories, which are virtual storehouses for raw [gene expression data](#), contain thousands of studies. Such an abundance of data opens opportunities for integrative analyses that can uncover new knowledge that was missed or was not possible in the initial publication of the data. For example, while a dataset from a given study may have been used for a particular published article, that same dataset may contain evidence whose value can only become realized when combined with data from another study. Then, it might become apparent that a drug can be repurposed to treat a different disease. Several computerized search engines have been designed to comb through this data. However, for these tools to be effective they require heavy, time-consuming human curation to ensure accuracy.

That is where crowdsourcing can be useful. For this project, the 70 volunteers were recruited through a massive open online course (MOOC), which was being taught on the Coursera MOOC platform by Avi Ma'ayan, PhD, Professor of Pharmacological Sciences and Director of the Mount Sinai Center of Bioinformatics at the Icahn School of Medicine at Mount Sinai. The student volunteers were asked first to identify relevant studies in the NCBI GEO database – in this case, studies in which single-gene or single-drug perturbations were applied to mammalian cells, or in which normal versus diseased tissues were compared. Once the studies were selected, the volunteers extracted metadata from the studies, and then computed differential expression using a custom-designed Chrome browser extension developed by the Mount Sinai researchers.

This process extracted information about gene signatures – observations of groups of genes whose combined expression is associated with a particular condition or drug action – which were stored in a new database. Then, Dr. Ma'ayan and his team used the database to visualize and analyze the signatures on a web portal known as Crowd Extracted Expression of Differential Signatures, or CREEDS, which was developed by the Ma'ayan Lab at Mount Sinai. Over the course of the project, over 3,100 single-gene perturbations from more than 2,300 studies were submitted, as well as 1,238 single-drug perturbations from nearly 450 studies.

"There is an incredible amount of data stored in these databases, but much of it has not been fully explored," said Dr. Ma'ayan. "The profiling and extrication of [gene expression](#) signatures is time-consuming and labor-intensive, and cannot be completely automated. By utilizing volunteers, so called 'citizen-scientists,' we were able to bring a much greater scale of human curation and quality control than we could have performed alone. By combining that human touch with automated programs, we could process much more data than would have been

otherwise possible."

Ultimately, the manually extracted signatures were used as a training set to help a program that uses machine learning to process all the data currently available in GEO for adding more drug, gene, and disease signatures to the CREEDS database. While researchers generally find that the quality of automatically generated signatures is subpar compared to those created by humans, such crowdsourced efforts can be integrated with machine learning to help refine the data. Instances that the computer programs find more difficult can be presented to the crowdsourced human curators with suggestions; this allows for higher quality data, while reducing effort required of the volunteer.

"We are grateful to the volunteers who helped demonstrate that citizen-scientists, working with researchers towards a common goal, can achieve remarkable results that have a real impact," said Dr. Ma'ayan. "Such collective efforts can help us discover new drugs, new causes of diseases, and new scientific knowledge."

While many new relationships between genes, drugs, and diseases were identified, further hypotheses can be formed through additional analysis of the data, which Dr. Ma'ayan and his team have made available to the public on the CREEDS portal. To interact with the portal, visit <http://amp.pharm.mssm.edu/creeds>.

Provided by The Mount Sinai Hospital

Citation: Crowdsourcing for scientific discovery: Researchers find novel ways to analyze data for drug and target discovery (2016, September 26) retrieved 10 April 2024 from <https://medicalxpress.com/news/2016-09-crowdsourcing-scientific-discovery-ways-drug.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private

study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.