Team taps the wisdom of the crowd to impact breast cancer prognosis

April 17 2013

Two new reports issuing in *Science Translational Medicine* (STM) today showcase the potential of teams of scientists working together to solve increasingly complex medical problems. The results demonstrate that better predictors of breast cancer progression than those currently available can be rapidly evolved by running open Big Data Challenges such as The Sage Bionetworks/DREAM Breast Cancer Prognosis Challenge (BCC).

In breast cancer, a key undertaking is determining those patients whose disease is most likely to progress rapidly and therefore tailor the best course of treatment for them. Currently oncologists are using gene-expression based assays such as MammaPrint and Oncotype Dx, that are based on 10 year old science, and both do better with breast cancer risk prediction than models based only on clinical data.

Dr. Stephen Friend, the Founder of Sage Bionetworks and one of the organizers of the BCC reflects, "Ten years ago, members of our research group used gene expression profiling to build one of the first breast cancer predictors. Mammaprint and Oncotype Dx were developed off of that but further improvement seems to have stalled. We wondered if running a Challenge like BCC would motivate lots ofdifferent groups to tackle this problem, some working collaboratively, and if that might be more fruitful than the current "go it alone" single researcher approach."

To push the envelope on all the innovations that could be incorporated into the BCC, Sage partnered with the DREAM Project, a visionary

distributed systems biology group that has run 24 successful open computational challenges over the last five years.

DREAM's founder and leader, Dr. Gustavo Stolovitzky saw the BCC as an opportunity to, "… refocus our efforts to create a collaborative research environment that fosters a complementary way of doing science, which accelerates the pace of discovery with the goal of contributing to a faster reduction of suffering due to disease. This seems to me like an ethical imperative."

The goal of the BCC was to build a computational model that accurately predicts breast cancer survival. To do this, participants of the Challenge used genomic and clinical information from 2000 women diagnosed with breast cancer (the METABRIC data set). They accessed this data on Synapse, Sage Bionetworks' open compute platform for data sharing and analysis: Google donated cloud-based standardized virtual machines that each participant used to train their models against the data. Individual participants and/or teams submitted their computational models to Synapse as open source code made viewable to all: their models were assessed against a hidden dataset and their scores were reported on a real-time leaderboard. The combination of immediate feedback and code-sharing allowed participants to improve their leaderboard ranking by adjusting their own models or by borrowing the code of others to forge new models.

Throughout the July-October 2012 model-training phase, a crowd of 350 players from 35 countries across the globe joined the Challenge and submitted a total of 1700 computational models for scoring. The winning model was determined by scoring the predictive accuracy of players' models against a newly generated data set: for this, the Avon Foundation For Women funded the generation of gene expression and copy number data as well as collection of corresponding clinical information from 180 breast cancer patients. Finally, the BCC organizers recognized that the

basic science community might be most energized to participate if the Challenge prize were not money but the invitation to publish an article about the winning model in a top tier journal. The editors of STM saw the unique opportunity to run their own experiment on how to structure the peer-review process for competition-based crowdsourcing studies such as the BCC. Today's issue of STM features not only the winner's article (the BCC Challenge prize) and a report from the BCC organizers on the Challenge's conception, execution and insights—STM also chose to highlight the BCC with an Editorial Summary and an iconic cover of "Rosie the Riveter," intended to symbolize the power of women and their data to transform health.

Quipped Challenge participant Richard Savage (MRC Fellow in Biostatistics at the University of Warwick) on the prospect of winning the opportunity to publish in STM, "This is huge and a genuinely new way to do some great science. I really think the organizers are onto something with this."

The winner turned out not to be a breast cancer doctor, or even a breast cancer researcher: the winning team ("Attractor Metagenes") hails from Professor Dimitris Anastassiou's laboratory at Columbia University's School of Engineering and Applied Science. Anastassiou, now a member of the Columbia Initiative in Systems Biology, funded this research from his own inventor's research allocation of patent royalties related to his previous work on digital television, which is now used in all DVDs and TV broadcasting systems worldwide. Working with two of his Ph.D. students, they developed the winning model underpinned by so-called "attractor metagenes," gene signatures that they had identified as behaving similarly in multiple cancer types. They refer to attractor metagenes as "bioinformatic hallmarks of cancer:" Remarks Professor Anastassiou, "We had discovered these 'pan-cancer' gene signatures previously, and so we hypothesized that they play important roles in cancer in general. The BCC allowed us to prove that they are indeed

highly prognostic at least in breast cancer". Indeed, the winning model's predictive accuracy for breast cancer survival outperformed the best 60 models of a pre-competition group of expert programmers and bested current clinical standards. He is now excited with the prospect of collaborating with medical researchers to make good use of these signatures of cancer for potential use in diagnostic, prognostic and eventually therapeutic productsapplicable in multiple cancer types.

Based on the success of the BCC, Sage Bionetworks and DREAM announced earlier this year that they would merge to run open science computational Challenges which foster the broader collaboration of the research community and provide a meaningful impact to both discovery and clinical research. Their merger provides a collaborative framework that will bring the ideals of open science one step closer to reality.

The BCC demonstrated the wisdom of the crowd to develop predictive models but also highlighted that the value of those models is limited by the questions being posed and by the data being utilized. Even as the BCC reports in this week's issue of STM, Sage Bionetworks and DREAM are announcing five DREAM8 Challenges at Sage's 4th Commons Congress taking place in San Francisco and working with the Avon Foundation For Women, Susan G. Komen, the Breast Cancer Research Foundation to develop the next BCC which will start by mobilizing breast cancer patients to donate their data to drive the solving of a clinically relevant question inbreast cancer with the potential to transform patient treatment.

Provided by The Sage Bionetworks