metagenome profiling to characterize the gut microbiota based initially on mouse models. Human studies will be more complex, involving large-scale screening of human samples, sophisticated computational analysis, bacterial culturing and phenotyping methods, together with comprehensive whole-genome sequencing. They will identify not only candidate bacteria but also potential problems, such as virulence factors or antibiotic-resistance genes. It will also be necessary to discriminate between pathogenic and benign bacterial strains, for which 16S rRNA gene profiling is insufficiently sensitive. Because the bacterial species in the mouse and human gut are different, preclinical mouse models with a humanized microbiota will be useful for deciphering protective mechanisms and ensuring safety and efficacy before phase 1 clinical trials.

In principle, the selection approach of Buffie *et al.*[1] could be used to identify bacteriotherapy cocktails to treat other diseases that are currently undergoing clinical trials with fecal microbiota transplant, including Crohn's disease, ulcerative colitis, necrotizing enterocolitis, type 2 diabetes and obesity.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

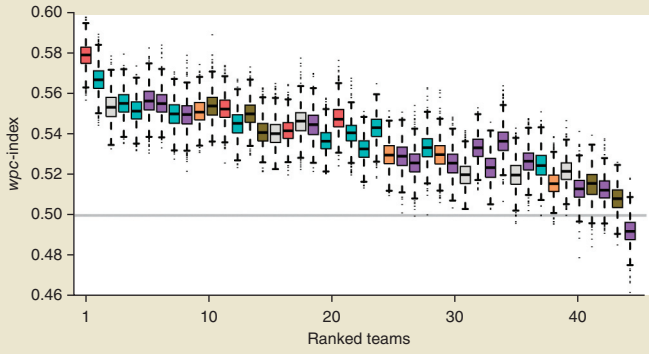
1. Buffie, C.G. *et al. Nature* doi:10.1038/nature13828 (22 October 2014).
2. Borody, T.J. *et al. J. Clin. Gastroenterol.* **38**, 475–483 (2004).
3. van Nood, E. *et al. N. Engl. J. Med.* **368**, 407–415 (2013).
4. Tvede, M. & Rask-Madsen, J. *Lancet* **1**, 1156–1160 (1989).
5. Petrof, E.O. *et al. Microbiome* **1**, 3 (2013).
6. Lawley, T.D. *et al. PLoS Pathog.* **8**, e1002995 (2012).
7. Reeves, A.E., Koenigsknecht, M.J., Bergin, I.L. & Young, V.B. *Infect. Immun.* **80**, 3786–3794 (2012).
8. Theriot, C.M. *et al. Nat. Commun.* **5**, 3114 (2014).
9. Wilson, K.H., Kennedy, M.J. & Fekety, F.R. *J. Clin. Microbiol.* **15**, 443–446 (1982).

DREAMing of benchmarks

A series of recent papers[1–4] showcases an emerging approach to data analysis in which open 'challenges' are issued to the research community to solicit computational methods that can solve important scientific questions. These studies illustrate some of the strengths of the community-challenge approach and provide new ways of addressing complex research problems.

The concept of carrying out a community challenge to solve a problem in biomedical research was popularized by an initiative called DREAM (Dialogue for Reverse Engineering Assessments and Methods). DREAM was launched by IBM's (Armonk, NY, USA) Gustavo Stolovitzky and collaborators, who in 2006 began brainstorming about computational platforms that could facilitate research in systems biology. Out of those conversations (in DREAM 1.0) came the first set of challenges (DREAM 2.0).

DREAM challenges focus on problems amenable to solution by a community-wide effort. Data, algorithms and software associated with the challenge are hosted on a technology platform (called Synapse) that is provided by the nonprofit Sage Bionetworks (Seattle), headed by Stephen Friend. Each year, at the DREAM annual meeting, which takes place jointly with the Annual RECOMB/ISCB Conference on Regulatory and Systems Genomics (in San Diego in 2014) and is open to anyone, participants and organizers discuss problems that would be well-suited to the challenge concept. Once the question is framed, "we do a dry run in which we analyze the data set using off-the-shelf methods to see if there is a signal," says Stolovitzky. "We need to know if it is possible to make a prediction with the available data before issuing a challenge," he explains. The submissions are scored to identify top-performing algorithms, and the results—and the code of the top performers—are made publicly available.



Since its inception, DREAM has successfully launched more than two dozen challenges, including efforts to benchmark methods in cancer genomics[1,5], to identify which patients will respond to certain treatments[2], to predict the synergistic activity of pairs of drug compounds[3] and to facilitate disease prognosis[4,6].

The competitions bring together members of different communities, such as those working on machine learning, DNA microarrays or biomarkers, who might otherwise not interact or work on large and complex biological data sets. The challenges are well suited to attracting newcomers into a field "because the problem is defined and there's an evaluation, which is easier than figuring out on your own what problem to study," says Christina Leslie, a machine learning scientist at Memorial Sloan-Kettering Cancer Center in New York. DREAM competitions "are democratizing," she adds, giving anyone a chance to analyze the data and submit an algorithm that could win. Encouraging the involvement of scientists with diverse backgrounds brings new perspectives to long-standing problems. The competitions can also highlight a need in an area, bringing together a community that may want to work collaboratively in the future.

"These are tasks that people are trying to do anyway," says University of California, San Diego, researcher Trey Ideker. If they weren't collected through these efforts, he says, "these methods would end up scattered in various publications, and it would be very hard to really understand where the progress in the field is." By providing the same data to all participants and evaluating them using the same criteria, DREAM challenges produce not only algorithms but also benchmarks. Scoring the algorithms is difficult, Stolovitzky acknowledges, because there are many ways to evaluate performance. "We seek expertise in statistics to innovate and refine scoring systems for each new challenge," he says.

Peer review provides another form of evaluation, which is sometimes embedded in the challenge. For one challenge[6], *Science Translational Medicine* offered to publish the competition results provided that the criteria for the winning model, which the editors helped define, were met. The editors then chose referees who were given access to the data and the algorithms and who participated in the challenge as organizers. Similarly, editors at *Nature Genetics* were closely involved in the creation and implementation of a rheumatoid arthritis challenge[7], which sought to identify genetic predictors of response in rheumatoid arthritis patients

receiving drugs targeting tumor necrosis factor alpha.

In many of the DREAM challenges, the data sets used are specifically generated for that purpose by researchers who are keen to tap into the broad base of analytic skills offered by the crowdsourcing format. "We find large numbers of groups who recognize that someone else might help crack the nut open, and they are very excited about that," says Sage's Friend. "Synapse keeps track of who did what using version control, providing an important measure of reproducibility, because you can go back and see exactly what was done," says Friend. Synapse also hosts the top-performing algorithms as open source code in R format, which are available to anyone and can be improved after the challenge is over.

In some cases, contest participants are not allowed to collaborate until the challenge submissions have been scored[4], whereas other challenges are completely open. In a recent competition to find predictive models of breast cancer prognosis[6], for example, the submitted algorithms were visible to all participants throughout the competition. Synapse featured a live leaderboard showing the scores of the submissions, which allowed participants to see how they were performing, to go back and improve on their own and others' approaches, and to then submit another version. However, challenge organizers keep a watchful eye on this because "if people work on a large set of slightly varying solutions, overfitting can happen," says Friend. "It's very important to design a challenge so that the validation sets don't become contaminated from the potential insights that come from working together," he explains.

A potential drawback of collaboration during a challenge is that rather than independent groups submitting new approaches in subsequent rounds, groups will improve incrementally the algorithms that ranked high in the first round.To encourage the development of innovative methods, one possibility is "a situation where you are judged by a panel on how innovative your approach is, rather than just numerical metrics," says Ideker. More creative approaches may ultimately perform better overall on multiple data sets even if they do not win on a particular data set. As important as new algorithm development is, also critical is being able to assess which types of data are most valuable and, as Friend says, to "discover robust data sets that will allow translational insights to emerge."

The ability to compare multiple algorithms is also particularly effective in identifying underlying principles that work well for particular research questions. As Leslie points out, "in some cases, the results of a challenge show that methods at the top of the ranking share components." For example, Costello *et al.*[2] found that modeling nonlinearities in the data was a common feature of top-performing methods to predict drug sensitivity. Other actionable outcomes may be unearthed by a challenge, such as an algorithm for reducing the number of patients needed in clinical trials for amyotrophic lateral sclerosis (ALS) drugs[4].

DREAM challenges have an important role in data science communities because they stimulate collaboration, whether during or after a challenge. Indeed, building communities may be their most enduring contribution. The hope, says Friend, is that DREAM will lead to more collaborative work in areas where there is a need for it, and that scientists will continue to join forces, perhaps to organize new competitions. The approach is also likely to help young scientists who work as part of teams to be recognized for their contributions, for example, through Synapse's tracking of algorithm development. "This is the Olympic arena in this field," says DREAM participant Yang Zhang.

In the coming years, DREAM-Sage wants to "learn how to scale the challenges so that instead of running 6 to 8 a year, we can run 25 or 50 a year," says Friend. "We are thinking about how to build a network that can do this." So far, challenges have tackled quantifiable problems, and whether they can address other types of questions remains unclear. InnoCentive, Kaggle and other platforms use crowdsourcing to find solutions to the problems clients need addressed, and in certain cases, this approach may be the most suitable. Good, old-fashioned, lab-to-lab collaborations will, of course, continue to be the *modus operandi* when the question being addressed is highly specific or the data set small. But the DREAM-Sage challenge network, by building communities and stimulating cooperation and openness, is poised to transform the way large data sets are analyzed. As big data become ever bigger, organizing people to find the best ways to mine the data will be the key.

Irene Jarchum, Associate Editor, &
Susan Jones, Senior Editor

1. Boutros, P.C. *et al. Nat. Genet.* **46**, 318–319 (2014).
2. Costello, J.C. *et al. Nat. Biotechnol.* **32**, 1202–1212 (2014).
3. Bansal, M. *et al. Nat. Biotechnol.* **32**, 1213–1222 (2014).
4. Küffner, R. *et al. Nat. Biotechnol.* **33**, 51–57 (2015)
5. Bilal, E. *et al. PLoS Comput. Biol.* **9**, e1003047 (2013).
6. Margolin, A.A. *et al. Sci. Transl. Med.* **5**, 81re1 (2013).
7. Plenge, R.M. *et al. Nat. Genet.* **45**, 468–469 (2013).