Leveraging crowdsourcing to accelerate global health solutions

To the editor:

Crowdsourced data science challenges can achieve in months what would take years for conventional research approaches. However, they remain largely untapped for underserved or critical biomedical challenges, such as treatment of malaria, where computational modeling lags and accelerated innovation is urgently needed to combat the emergence of drug resistance.

The diversity and quality of computational solutions obtained in data science challenges[1] combined with the rapid pace at which they are developed (i.e., typically weeks to months for a single challenge) could in principle accelerate research in fields where data science expertise is highly underrepresented. When it comes to disease-related research questions, data science crowdsourcing platforms have only been used in a limited way and for a very restricted number of diseases. For example, out of the 19 ongoing challenges on Kaggle at the time of writing, only 1 is for a disease, cancer. And since its establishment in 2007, not a single DREAM challenge has involved an infectious disease, whereas 15 have involved cancer. In this context, we believe there is an untapped opportunity for the data science challenges model of innovation to expand beyond the areas to which it has been traditionally applied into new areas of need, such as neglected diseases. Here, we consider the case of emerging drug resistance to the anti-malarial drug artemisinin and present a call for participation in the Malaria DREAM challenge for computational models of drug resistance.

The need for crowdsourced data science challenges in malaria

Malaria can serve as a clear example of a disease area where data science challenges could accelerate the pace of research. First, despite numerous eradication efforts, malaria remains a global health challenge. Artemisinin**,** the last line of defense against multi-drug resistant malaria, has had multiple reports of resistance within Southeast Asia in recent years[2]. If artemisinin resistance was to spread to Africa, where most cases and deaths due to malaria occur, this could erase decades of malaria eradication progress and efforts. Before artemisinin resistance reaches Africa, a concerted effort must be made to understand the mechanistic changes that malaria

parasites undergo to obtain resistance and to determine what drugs may be most effective in combination with artemisinin derivatives to deter or counter further resistance. This calls for accelerated innovation for alternative drugs. Collaborative crowdsourcing through data science challenges can accelerate innovative solutions with modest funding investment.

Second, although investment in global health research, the availability of large-scale datasets and overall sharing of data have increased, the field of malaria research is lacking in data scientists and job openings. For example, a search on LinkedIn for data science jobs in malaria yields only 11 jobs compared with 1601 jobs for cancer-focused data scientists. As has happened in other challenges, such as the ALS DREAM challenge[3], a malaria-focused data science challenge could enable the field to tap into data scientists who do not ordinarily work on malaria and, with the right incentives, elicit interest in the disease. In addition to this, malaria research has historically not kept pace with modeling advances in machine learning as it has failed to attract expertise outside the field, a trend also seen across other research areas of global health. For example, many data scientists are turning to cancer modeling but it is hard to identify any who have turned to malaria. Crowdsourced data science challenges open to a wider community of modelers are bound to inject new modeling ideas into the field where real-time discoveries can guide refined disease control strategies. Finally, even though data generation has grown over the past decade, many areas in malaria research lack what would qualify as 'big' data. Developing computational solutions that could work on 'small' biological datasets through crowdsourcing malaria challenges could drive innovation on data science with limited data and convince more communities that a good avenue for unpublished, difficult-to-create clinical datasets are crowdsourced data science challenges[4].

Time for a malaria DREAM challenge

Recently, we held the 'DREAM of Malaria' hackathon, a 1-week effort that brought together young scientists from various African countries, many with no expertise in malaria, to assess whether malaria transcriptomic data can be used to predict artemisinin resistance[5]. The hackathon demonstrated the potential for data scientists outside of the current malaria research community to make valuable contributions in pre-publication exploratory data analysis[5], and

highlighted an opportunity to bring a level of rigor inherent in community analysis that cannot be captured in a typical research project run by one group.

In late April, we launched the Malaria DREAM Challenge, which is open to anyone interested in contributing to the development of computational models that address important problems in advancing the fight against malaria[6]. The overall goal of this Malaria DREAM Challenge is to predict artemisinin drug resistance level for a test set of malaria parasites using their *in vitro* transcription data and a training set consisting of published *in vivo* and unpublished *in vitro* transcriptomes (Fig. 1). The *in vivo* dataset consists of ~1000 transcription samples from various geographic locations covering a wide range of life cycles and resistance levels, with other accompanying data such as patient age, geographic location, artemisinin combination therapy used, etc.[7]. The *in vitro* transcription dataset consists of 55 isolates, with transcription collected at two timepoints (6 and 24 hours post-invasion), in the absence or presence of an artemisinin perturbation, for two biological replicates using a custom microarray designed at the Ferdig lab in the University of Notre Dame. Using these transcription datasets, participants will be asked to predict three different resistance states of a subset of the 55 *in vitro* isolate samples: 50% inhibitory concentration values (IC_{50}), patient clearance half-life (PC1/2), and categorical resistance state (resistant/sensitive). The Malaria DREAM challenge could enable the field to gain insights into the mechanisms that underlie a lack of correlation between *in vitro* (IC50) and *in vivo* (clearance rate) measures of artemisinin resistance, a major drawback for laboratory studies of artemisinin resistance that are essential for studying mechanisms of resistance[8].

Potential for impact in malaria and beyond

The internet is a powerful enhancer of human collaboration globally. Crowdsourced data science challenges make it possible to engage a wider community via the internet, including those not traditionally involved in a field, by lowering the entry barriers to a subject through well curated datasets and a community of interested data scientists that compete collaboratively. Furthermore, data science challenges marshal a large number of solutions to a single well-defined problem within a field increasing the chance for new and better solutions to emerge. Crowdsourced data science challenge platforms, such as DREAM challenges, also promote sharing of open source code, which anybody can access and reuse after the challenge. A Malaria DREAM challenge on

emerging artemisinin drug resistance will engage perhaps the largest community of modelers to work simultaneously on the problem of drug resistance in the disease. Our experience suggests that this will inspire new ways of problem solving in the field.

We also contend that a Malaria DREAM challenge could serve as a model for the potential of data science challenges to provide solutions to underserved global health challenges in biomedicine, such as in neglected tropical diseases[9]. Indeed, the significance of this challenge approach is further shown by the funding support for this challenge from the Bill and Melinda Gates Foundation as well as the Foundation's exploration on crowdsourcing global health data as a mechanism to vastly accelerate learning and impact on tough global health problems where data and modeling can function as key drivers.

Acknowledgments

The data collection for the Malaria DREAM challenge is funded by an NIH R21 grant to MTF (AI103872-01AI). The Malaria DREAM Challenge has received funding support from the Bill and Melinda Gates Foundation.

Author contributions

GHS, MTF and GS conceived the challenge. TA, FN and XS provided malaria parasite isolates. SD, KBS, LC, KV, GJF and SKK collected transcriptional data for the challenge. GHS, PM, GS, MTF, KBS, NM, AG, EMA, SK, TN, TB, JB, TM and DM were all involved in organizing the challenge. All authors read and approved the paper.

Competing interests

The declare no competing interests

Michael T. Ferdig[1], Gustavo Stolovitzky[2], Geoffrey H. Siwo[1,6], Sage Davis[1], Katrina Button-Simons[1], Steven Kern[12], Thea Norman[12], Pablo Meyer[2], Nicola Mulder[7], Amel Ghouila[14], Sumir Panji[7], Thomas Yu[9], Julie Bletz[9], Timothy JC Anderson[3], Xinzhuan Su[8], Francois Nosten[4], Lisa Checkley[1], Gabriel J. Foster[1], Katelyn Vendrely[1], Sok Kean Khoo[5], Taoufik Bensellak[13], Eren Mehmet Ahsen[10], Darlington Mapiye[11] & Ahmed Moussa [13]

[1] Eck Institute for Global Health, Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA

[2] Thomas J. Watson Research Center, IBM, NY, USA

[3] Texas Biomedical Research Institute, San Antonio, TX, USA

[4] Shoklo Malaria Research Unit, Mahidol-Oxford Tropical Medicine Research Unit, Mahidol University, Mae Sot, Thailand

[5] Grand Valley State University, Grand Rapids, MI, USA

[6] Center for Research Computing, University of Notre Dame, Notre Dame, IN, USA

[7] Computational Biology Division (H3ABioNet), University of Cape Town, South Africa

[8] National Institutes of Health, Bethesda, MD, USA

[9] Sage Bionetworks, WA, USA

[10] Icahn School of Medicine at Mount Sinai, NY, USA

[11] IBM Research Africa, Johannesburg, South Africa

[12] Bill and Melinda Gates Foundation, Seattle, WA, USA

[13] Abdelmalek Essaadi University, ENSATg, System and Data Engineering Team, Tangier, Morocco

[14] Institut Pasteur de Tunis, LR11IPT02, Tunis-Belvédère, 1002, Tunisia.

Correspondence should be addressed to G.H.S. (e-mail: geoffrey@unitedgenomes.org; siwomolbio@gmail.com)

1. Saez-Rodriguez, J. *et al.* Crowdsourcing biomedical research: Leveraging communities as innovation engines. *Nature Reviews Genetics* (2016).

doi:10.1038/nrg.2016.69

2. Das, D. *et al.* Artemisinin Resistance in Plasmdium falciparum Malaria. *N. Engl. J. Med.* (2009). doi:10.1086/657120

3. Küffner, R. *et al.* Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nat. Biotechnol.* (2015). doi:10.1038/nbt.3051

4. Longo, Dan L., Drazen, J. Data Sharing. *Med. Sci. Publ. Author, Ed. Rev. Perspect.* (2017). doi:10.1016/B978-0-12-809969-8.00024-3

5. Ghouila, A. *et al.* Hackathons as a means of accelerating scientific discoveries and knowledge transfer. *Genome Res.* **28,** (2018).

6. Malaria DREAM Challenge. Available at: https://www.synapse.org/#!Synapse:syn16924919/wiki/583955.

7. Mok, S. *et al.* Population transcriptomics of human malaria parasites reveals the mechanism of artemisinin resistance. *Science (80-.).* (2015). doi:10.1126/science.1260403

8. Amaratunga, C. *et al.* Artemisinin-resistant Plasmodium falciparum in Pursat province, western Cambodia: A parasite clearance rate study. *Lancet Infect. Dis.* (2012). doi:10.1016/S1473-3099(12)70181-0

9. Feasey, N., Wansbrough-Jones, M., Mabey, D. C. W. & Solomon, A. W. Neglected tropical diseases. *British Medical Bulletin* (2010). doi:10.1093/bmb/ldp046

Figure 1. Malaria DREAM challenge. In objective 1 of the challenge, participants will be provided with a training set composed of: transcription data from 30 isolates at 6- or 24-hours post-invasion under Art exposure for 2 hours, transcription data of controls at the same time points and corresponding IC50s of each isolate as determined by *in vitro dose* response assays. Art perturbations were performed with dihydroartemisinin, DHA. Participants will be challenged to use the provided data to predict the IC50 of 25 isolates whose actual IC50s will be hidden from them but have been determined in the lab. Participants will submit their predicted IC50s alongside any code used and the DREAM challenge organizers will score the submissions for how well they correlate with lab determined IC50s using various metrics. In objective 2 of the challenge, participants will be provided with a training set composed of published transcriptional profiles of 1043 isolates alongside the corresponding clearance rate for each parasite and patient metadata[12]. Participants will be challenged to predict the absolute clearance rate in hours and binary resistance status (i.e. resistant or sensitive) based on a threshold. The challenge will be open for submissions for 3 months after which submissions will be scored and best performing teams identified. All code and underlying datasets will be made publicly available via the Synapse data platform and at least one publication detailing the challenge results is expected. Best performing teams will also be invited to present their models at the annual DREAM challenges meeting and we will seek avenues to share the results more specifically with the malaria community.