METHODS TO ENSURE THE REPRODUCIBILITY OF BIOMEDICAL RESEARCH

KONRAD J. KARCZEWSKI
Massachusetts General Hospital, Boston, MA; Broad Institute, Cambridge, MA
Email: konradjkarczewski@gmail.com

NICHOLAS P. TATONETTI
Columbia University, New York, NY
Email: nick.tatonetti@columbia.edu

ARJUN K. MANRAI
Harvard Medical School, Boston, MA
Email: manrai@post.harvard.edu

CHIRAG J. PATEL
Harvard Medical School, Boston, MA
Email: chirag_patel@hms.harvard.edu

C. TITUS BROWN
University of California , Davis, CA
Email: ctbrown@ucdavis.edu

JOHN P. A. IOANNIDIS
Stanford University, Stanford, CA
Email: jioannid@stanford.edu

Science is not done in a vacuum – across fields of biomedicine, scientists have built on previous research and used data published in previous papers. A mainstay of scientific inquiry is the publication of one's research and recognition for this work is given in the form of citations and notoriety -- ideally given in proportion to the quality of the work. Academic incentives, however, may encourage individual researchers to prioritize career ambitions over scientific truth. Recently, the *New England Journal of Medicine* published a commentary calling scientists who repurpose data "research parasites" who *mis*use data generated by others to demonstrate alternative hypotheses[1]. In our opinion, the concept of data hoarding not only runs contrary to the spirit of, but also hinders scientific progress. Scientific research is meant to seek objective truth, rather than promote a personal agenda, and the only way to do so is through maximum transparency and reproducibility, no matter who is using the data.

To maintain the integrity of the scientific process, it is necessary to cultivate practices that ensure reproducibility, especially as large and public heterogeneous databases proliferate. Many of these paradigms can be likened to open-source practices already adopted by much of the computer

science community. These include, but are not limited to, version control, code review, and containerization. There are many benefits to improving reproducibility: aside from the general benefit to science through increased transparency, releasing code enables additional peer review and is educational and efficient as it reduces duplications of efforts. Of course, these approaches require additional time for investigators to document and clean up code and data for release, which is the top reason for not sharing data and code[2] (in addition to managing the intricacies of tools for version control, for example). Various incentive structures have been proposed to improve reproducibility rates across scientific fields, including creation of requirements by funding agencies or establishment of reward systems[3]. Additionally, like many computational skills, these require some initial effort, but have long-term benefits and will eventually become ingrained. Finally, public release of code can enable public code review, which improves programming habits: efforts such as Software Carpentry have been established to teach these skills and have met with recent success[4].

Reproducibility can take a number of forms and the desired extent of reproducibility has been debated in other fora: whatever the ideal solution, there is room for improvement in ensuring that research is reproducible. A growing number of researchers have begun to share their code and processed data, where possible. For instance, the ENCODE project released a virtual machine image that contained the code and data to reproduce the figures in their manuscript[5] [http://encodeproject.org/ENCODE/integrativeAnalysis/VM]. Similarly, the ExAC consortium deposited the figure generating code for their recent papers[6,7] on Github [https://github.com/macarthur-lab/exac_papers; https://github.com/ericminikel/prnp_penetrance]. Some have gone even further as to publicly release a full manuscript under version control[8,9] and document the process for others to do so [http://ivory.idyll.org/blog/2014-our-paper-process.html].

In this session, we feature five papers that explore research on the topic of reproducibility. This year, we required submissions to strive for reproducibility by depositing data and code on public repositories. The authors have stepped up to the challenge and are practicing what they preach: where possible, they have released applicable code and/or data to make their own research as reproducible as possible.

Session Contributions

Cohain, Divaraniya, and colleagues[10] address an important challenge for reproducibility of Bayesian networks. While frequentist approaches can rely on p-values to predict replication, the construction of a Bayesian network is a data-dependent and heuristic process, and consistency between multiple analyses has not been rigorously performed. This paper explores the replication of Bayesian networks, particularly in relation to key driver nodes and hubs, as well as edge reproducibility.

Hundreds of studies have used publicly available data to predict adverse drug reactions and drug indications and have reported seemingly exceptional predictive accuracy: Guney[11] investigates the

issue of performance overestimation for drug side effect and indication, and finds that major assumption of these methods (independence) is violated, which overestimates their performance. Haynes et al[12] present a pipeline for expression meta-analysis, which fills an unmet need for systematic processing and visualization of results from such analyses. Kaushik and colleagues[13] describe a workflow engine that uses graph theory approaches to optimize and ensure reproducible data analyses. Finally, Yang et al[14] provide a detailed look on the reproducibility of clinical genetics data: concordance across variant classifications is reasonably high, but more work will be required to resolve differences and accurately classify all variants as pathogenic or benign. In summary, these exemplar papers demonstrate how to enhance research reproducibility across a variety of biomedical domains critical in this era of "big data" and precision medicine.

References

1. Longo, D. L. & Drazen, J. M. Data Sharing. *http://dx.doi.org.ezp-prod1.hul.harvard.edu/10.1056/NEJMe1516564* **374,** 276–277 (2016).
2. Stodden, V. The Scientific Method in Practice: Reproducibility in the Computational Sciences. *SSRN Journal* (2010). doi:10.2139/ssrn.1550193
3. Ioannidis, J. P. A. How to Make More Published Research True. *PLoS Med* **11,** e1001747 (2014).
4. Wilson, G. Software Carpentry: lessons learned. *F1000Research* **3,** (2014).
5. ENCODE Project Consortium *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
6. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
7. Minikel, E. V. *et al.* Quantifying prion disease penetrance using large population control cohorts. *Science Translational Medicine* **8,** 322ra9–322ra9 (2016).
8. Pell, J. *et al.* Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proc Natl Acad Sci USA* **109,** 13272–13277 (2012).
9. Zhang, Q., Pell, J., Canino-Koning, R., Howe, A. C. & Brown, C. T. These Are Not the K-mers You Are Looking For: Efficient Online K-mer Counting Using a Probabilistic Data Structure. *PLoS ONE* **9,** e101271 (2014).
10. Cohain A, Divaraniya AA, Zhu K, Zhu J, Chang R, Dudley JT, Schadt EE. "Exploring the reproducibility of probabilistic causal molecular network models" *Pac. Symp Biocomput* (2017).
11. Guney E. "Reproducible Drug Repurposing: When Similarity Does Not Suffice" *Pac. Symp Biocomput* (2017).
12. Haynes WA, Vallania F, Liu C, Bongen E, Tomczak A, Andres-Terrè M, Lofgren S, Tam A, Deisseroth CA, Li MD, Sweeney TE, Khatri P. "Empowering Multi-Cohort Gene Expression Analysis to Increase Reproducibility" *Pac. Symp Biocomput* (2017).
13. Kaushik G, Ivkovic S, Simonovic J, Tijanic N, Davis-Dusenbery B, Kural D. "Graph Theory Approaches For Optimizing Biomedical Data Analysis Using Reproducible Workflows" *Pac. Symp Biocomput* (2017).
14. Yang S, Cline M, Zhang C, Paten B, Lincoln SE. "Data Sharing and reproducible Clinical genetic testing: successes and challenges" *Pac. Symp Biocomput* (2017)