Bioinformatics, 35(4), 2019, 709–710
doi: 10.1093/bioinformatics/bty666
Advance Access Publication Date: 7 August 2018
Letter to the Editor

OXFORD

Databases and ontologies

Genetic Simulation Resources and the GSR Certification Program

Bo Peng[1], Man Chong Leong [2], Huann-Sheng Chen[3], Melissa Rotunno[3], Katy R. Brignole[3], John Clarke[4] and Leah E. Mechanic[3,*]

[1]Department of Bioinformatics and Computational Biology, MD Anderson Cancer Center, Houston, TX 77030, USA, [2]Children's Environmental Health Initiative, Rice University, Houston, TX 77005, USA, [3]Division of Cancer Control and Population Sciences, National Cancer Institute (NCI), National Institutes of Health (NIH), Bethesda, MD 20892, USA and [4]Cornerstone Systems, Lynden, WA 98264, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Contact: mechanil@mail.nih.gov

Received and revised on March 27, 2018; editorial decision on July 16, 2018; accepted on August 6, 2018

Dear editor

With recent explosion in the diversity and volume of genetic data generated, an increasing number of genetic simulation programs have been developed to aid the development of statistical methods for the analysis of such data in all genetic-related disciplines (Escalona *et al.*, 2016; Hoban *et al.*, 2012; Peng *et al.*, 2015; Ritchie and Bush, 2010). Despite the recognized importance of genetic simulation tools, it is often difficult to discover and select the right simulation tool for a particular study due to differences in the type of genetic data of interest, simulation methods, features, terminologies and assumptions (Mechanic *et al.*, 2012), and lack of external evaluations of the usability and maintenance status of published simulators. To address these issues, we want to encourage use of the Genetic Simulation Resources (GSR) online catalog and search tool (https://popmodels.cancercontrol.cancer.gov/gsr/) and participation in the GSR Certification Program.

GSR was created in 2013 to help researchers sort through and compare a large number of genetic simulation programs to identify the best tool for their research topic (Peng *et al.*, 2013). The GSR catalog currently features more than 100 simulators that use various simulation techniques (coalescent and forward-time simulations are the most common methods), and with application areas such as linkage analysis, genome wide association and sequencing analyses. GSR records basic information, features of the simulation programs and lists typical applications of the software.

After reviewing the software listed in GSR, we realized that the quality and usefulness of published simulation tools varied greatly due to inaccessible source code, lack of or incomplete documentation, difficulties in installation and execution, lack of support from authors and lack of programs maintenance (Peng *et al.*, 2015). An approach to address questions about the quality and utility of tools was suggested in a workshop in 2013 (Chen *et al.*, 2015), to evaluate genetic simulation tools based on a defined checklist of features that may benefit end users. Therefore, the GSR Certification Program was developed and implemented to appraise tools based on criteria in the categories of accessibility, documentation, application and support (https://popmodels.cancercontrol.cancer.gov/gsr/certification/). It is hoped that evaluating software based on these criteria will encourage improvements in reporting and documentation of simulation software tools. Moreover, the GSR certification criteria ensure that these programs are more readily findable and usable by the research community and consistent with the FAIR (Findability, Accessibility, Interoperability and Reusability) principles for scientific data management and stewardship (Wilkinson *et al.*, 2016).

The GSR Certification Program is overseen by the GSR Certification Committee, which currently consists of NIH staff and extramural researchers. The manuscript authors invite volunteers who wish to participate in the evaluation process to contact the corresponding author of this letter. The GSR Certification Committee is in process of reviewing all packages cataloged in GSR, beginning with newer software packages and packages that are considered actively maintained and broadly used in the scientific community. The review process consists of two stages; an initial review where a GSR moderator completes a questionnaire based on publicly available information (e.g. publication, web site, documentation, Pubmed citation, issue tracker), and a panel review during which three members of the GSR Certification Committee decide about award of any certificates (Accessibility, Documentation, Application and Support). Simulators with one or more GSR certificates are prominently listed on the GSR web site. Simulators which obtain all 4 GSR certificates are invited to place a GSR Certified tile on the homepages for their software.

The GSR Certification Program provides a service to the community by evaluating genetic simulation software, with the ultimate goal of promoting the development and application of genetic simulation software. With *Bioinformatics* being the target journal for many simulation programs, it would be beneficial to authors, the journal and the scientific community if package authors are encouraged to register their simulation programs at the GSR where they are evaluated to conform to basic guidelines for accessibility, documentation, application and support for GSR certifications. We believe this policy could facilitate dissemination of software and ensure the quality and usability of genetic simulation software.

Funding

The development and maintenance of GSR has been supported by contracts from the National Cancer Institute, National Institutes of Health [HHSN261201100558P] and [HHSN261201300002I].

Conflict of Interest: none declared.

References

Chen,H.S. *et al*. (2015) Genetic simulation tools for post-genome wide association studies of complex diseases. *Genet. Epidemiol*., **39**, 11–19.

Escalona,M. *et al*. (2016) A comparison of tools for the simulation of genomic next-generation sequencing data. *Nat. Rev. Genet*., **17**, 459–469.

Hoban,S. *et al*. (2012) Computer simulations: tools for population and evolutionary genetics. *Nat. Rev. Genet*., **13**, 110–122.

Mechanic,L.E. *et al*. (2012) Next generation analytic tools for large scale genetic epidemiology studies of complex diseases. *Genet. Epidemiol*., **36**, 22–35.

Peng,B. *et al*. (2013) Genetic Simulation Resources: a website for the registration and discovery of genetic data simulators. *Bioinformatics*, **29**, 1101–1102.

Peng,B. *et al*. (2015) Genetic data simulators and their applications: an overview. *Genet. Epidemiol*., **39**, 2–10.

Ritchie,M.D. and Bush,W.S. (2010) Genome simulation approaches for synthesizing in silico datasets for human genomics. *Adv. Genet*., **72**, 1–24.

Wilkinson,M.D. *et al*. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.