Editorial: Opportunities and challenges in reusing public genomics data

Mahmoud Ahmed and Deok Ryong Kim*

Department of Biochemistry and Convergence Medical Sciences, Institute of Health Sciences, College of Medicine, Gyeongsang National University, Jinju, Republic of Korea

KEYWORDS

reusing public data, genomics, data sharing, metadata, data curation and integration

Editorial on the Research Topic
Opportunities and challenges in reusing public genomics data

Genomics data is accumulating in public repositories at an ever-increasing rate. Large consortia and individual labs continue to probe animal and plant tissue and cell cultures, generating vast amounts of data using established and novel technologies. The human genome project kick started the era of systems biology (Lander et al., 2001; Gates et al., 2021). Ambitious projects followed to characterize non-coding regions, variations across species, and between populations (Feingold et al., 2004; Sabeti et al., 2007; Auton et al., 2015). The cost reduction allowed individual labs to generate numerous smaller high-throughput datasets (Edgar et al., 2002; Parkinson et al., 2007; Metzker, 2010; Leinonen et al., 2011). As a result, the scientific community should consider strategies to overcome the challenges and maximize the opportunities to use these resources for research and the public good. In this Research Topic, we have elicited opinions and perspectives from researchers in the field on the opportunities and challenges of reusing public genomics data. The articles in this Research Topic converge on the need for data sharing while acknowledging the challenges that come with it. Two articles defined and highlighted the distinction between data and metadata. The characteristic of each should be considered when designing optimal sharing strategies. One article focuses on the specific issues surrounding the sharing of genomics interval data, and another on balancing the need for protecting pediatric rights and the sharing benefits.

The definition of what counts as data is itself a moving target. As technology advances, data can be produced in more ways and from novel sources. Events of recent years have highlighted this fact. "The pandemic has underscored the urgent need to recognize health data as a global public good with mechanisms to facilitate rapid data sharing and governance," Schwalbe et al. The challenges facing these mechanisms could be technical, economic, legal, or political. Defining what data is and its type, therefore, is necessary to overcome these barriers because "the mechanisms to facilitate data sharing are often specific to data types." Unlike genomics data, which has established platforms, sharing clinical data "remains in a nascent phase." The article by Patrinos et al. considers the strong ethical imperative for protecting pediatric data while acknowledging the need to avoid over protections. The authors discuss a model of consent for pediatric research that can balance the need to protect participants and generate health benefits.

Xue et al. focus on reusing genomic interval data. Identifying and retrieving the relevant data can be difficult, given the state of the repositories and the size of these data. Similarly,

integrating interval data in reference genomes can be hard. The author calls for standardized formats for the data and the metadata to facilitate reuse.

Sheffield et al. highlight the distinction between data and metadata. Metadata describes the characteristics of the sample, experiment, and analysis. The nature of this information differs from that of the primary data in size, source, and ways of use. Therefore, an optimal strategy should consider these specific attributes for sharing metadata. Challenges specifics to sharing metadata include the need for standardized terms and formats, making it portable and easier to find.

We go beyond the reuse issue to highlight two other aspects that might increase the utility of available public data in Ahmed et al. These are curation and integration. Despite being generated using different protocols, combining the datasets from separate groups could help to fill the gaps in the design and increase the statistical power of the analysis. Integrating data types can be beneficial to either verify or complement the observations made based on a single data type. We also emphasize the critical requirements for these strategies to be successful. We draw on our experience and others in using publicly available datasets to support, develop, and extend our research interest.

The articles in this Research Topic converge on the importance of data sharing. In addition, the articles present the challenges facing data sharing and reuse and propose models to increase the utility of public data.

Author contributions

MA and DK wrote and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

Acknowledgments

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi:10.1038/nature15393

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids Res.* 30, 207–210. doi:10.1093/nar/30.1.207

Feingold, E. A., Good, P. J., Guyer, M. S., Kamholz, S., Liefer, L., et al. (2004). The ENCODE (ENCyclopedia of DNA elements) project. *Science* 306, 636–640. doi:10.1126/science.1105136

Gates, A. J., Gysi, D. M., Kellis, M., and Barabási, A. L. (2021). A wealth of discovery built on the human genome project — By the numbers. *Nature* 590, 212–215. doi:10.1038/d41586-021-00314-6

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi:10.1038/35057062

Leinonen, R., Sugawara, H., and Shumway, M. (2011). The sequence read archive. *Nucleic Acids Res.* 39, D19–D21. doi:10.1093/nar/gkq1019

Metzker, M. L. (2010). Sequencing technologies the next generation. *Nat. Rev. Genet.* 11, 31–46. doi:10.1038/nrg2626

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., et al. (2007). ArrayExpress - a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* 35, D747–D750. doi:10.1093/nar/gkl995

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918. doi:10.1038/nature06250