A Community Assessment of Data Perturbation Techniques on Privacy Protection for Human Genome Data

Xiaoqian Jiang[1], Lucila Ohno-Machado[1], Bradley Malin[3], Haixu Tang[2],
Shuang Wang[1], Xiaofeng Wang[2], Yongan Zhao[2]

[1]Division of Biomedical Informatics, University of California, San Diego, 9500 Gilman Drive, San Diego, CA, 92093, USA.
[2]School of Informatics and Computing, Indiana University, 150 S. Woodlawn Avenue, IN 47401 Bloomington, USA.
[3]Departments of Biomedical Informatics and Electrical Engineering & Computer Science, Vanderbilt University, 2525 West End Avenue, Suite 1030, Nashville, TN 37232, USA.

The biomedical community is evolving to incorporate, and heavily rely upon, “big data" science. This is due, in part, to advances in high throughput and computing technologies (Howe, Costanzo et al. 2008). Every day, a large amount of biomedical data (e.g., human DNA sequences and biomedical images) is being collected in research and clinical laboratories across the world, which can be stored and shared rapidly. However, to transform such data into knowledge that is applicable to biomedicine, novel approaches for analysis, integration and mining need to be developed (Marx 2013). The technical challenge of sharing and analyzing the massive, often heterogeneous datasets that comes from this demand is further complicated by the need to preserve the privacy and confidentiality of data donors, from whom the data were collected. Recent studies showed that the identity of a donor can be inferred from their biomedical data (e.g., their health records or genomic sequences) (Craig et. al., 2011). Effective privacy preserving techniques need to be developed to strike a balance between data sharing and privacy protection.

Over the past several years, there has been a growing interest in methodologies that perturb genomic data, or associated summary statistics. Different from the cryptographic protocols such as secure multi-party computation, which protect privacy of inputs to some computing operation, data perturbation techniques achieve privacy protection on both the input and output data. They can be used to disseminate sensitive data or to publish computing results that contain sensitive information. To investigate the extent to which the data perturbation technologies are mature, we organized the Critical Assessment of Data Privacy and Protection (CADPP) challenge at the 3rd iDASH Privacy Workshop as a community effort to evaluate the effectiveness of these methodologies (e.g., the differential privacy approach) for genomic data.

This workshop was organized around two specific problems. In the first problem, teams focused on the challenge of sharing aggregate human genomic data (i.e., allele frequencies) in a way that preserves the privacy of the data donors, without undermining the utility of the data in genome-wide association studies (GWAS); i.e., the most significant genomic regions identified by GWAS are preserved after data perturbation. In the second problem, teams were challenged to publish GWAS results (i.e., the most significant genomic regions) that achieved differential privacy under a specific privacy budget ($\varepsilon=1$). We devised a task for each of these two problems and used publicly available data from the International HapMap Project and the Personal

Genomics Project (PGP). A total of six teams from two countries participated in the challenges, and submitted their results to one of the two tasks.

The final results were presented in a workshop on March 24, 2014 in San Diego, California. Based on the results of the competition, we observed that it remains a challenge to disseminate data in a privacy-preserving way, while maintaining their utility in GWAS. Specifically, for a dataset involving a few hundreds of Single Nucleotide Polymorphisms (SNPs), the utility of the data was largely damaged after adding enough noise so that the data donors could be inferred using the likelihood ratio statistic, which was subject to the strongest inference attack on allele frequencies to date (Sankararaman, Obozinski et al. 2009). Although the most significant SNPs (detected using GWAS-like statistics) were preserved, over 8 times more non-significant SNPs would be reported as significant, resulting in a very high false positive rate.

The problem of preserving privacy while disclosing useful information will become even more challenging when a larger volume of human genomic data are to be shared. On the other hand, the results show that privacy-preserving techniques can work well on sharing analytical results (by using GWAS-like statistics) rather than raw genomic data. High utility can be preserved when only a small number (e.g., 5-10) of the most significant genomic regions needs to be published: on average, 85% of the top 10 and 90% of the top 5 most significant regions remain in the published results when a high differential standard is applied (i.e., privacy budget = 1.0). These results suggest that the differential privacy-based approach is ready to be deployed on the server of a data center for disseminating human genomic data.

A more detailed report on the challenge can be found on its website at http://www.humangenomeprivacy.org/. We plan to organize this event annually, and hereby welcome suggestions and comments to make the challenge better in the future years.

Acknowledgements

This work is supported by iDASH (grant no: U54HL108460), NLM (grant no: R00LM011392), NCBC-collaborating R01 (grant no: HG007078-01) from National Institutes of Health.

References

Craig, David W., et al. "Assessing and managing risk when sharing aggregate genetic variant data." *Nature Reviews Genetics* 12.10 (2011): 730-736.

Howe, D., M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. Schaeffer, S. St Pierre, S. Twigger, O. White and S. Y. Rhee (2008). "Big data: The future of biocuration." Nature **455**(7209): 47-50.

Marx, V. (2013). "Biology: The big challenges of big data." Nature **498**(7453): 255-260.

Sankararaman, S., G. Obozinski, M. I. Jordan and E. Halperin (2009). "Genomic privacy and limits of individual detection in a pool." Nat Genet **41**(9): 965-967.