National Human Genome Research Institute
Workshop on Sharing Aggregate Genomic Data
May 19-20, 2016
Rockville, MD
Workshop Summary

Background

In light of the high value of aggregate genomic data from genome-wide association studies and evolving ideas in the public and research communities about privacy risks and data sharing, the National Human Genome Research Institute (NHGRI) hosted a two-day workshop where participants were asked to weigh the benefits and risks of sharing aggregate genomic data with secondary users.

For this workshop, the term "aggregate genomic data" was defined as calculated summary statistics, including genotype counts, allele frequencies, effect size estimates and standard errors, and p-values calculated from a study sample. These statistics are calculated based on many individuals' genotypes generated either on arrays or sequences in a study and are also called "genomic summary statistics". Since "aggregate genomic data" has different meanings in different contexts, workshop participants agreed instead to use the term "genomic summary statistics," and that term is used for the remainder of this document.

Prior to 2008, genomic summary statistics were publicly available in the National Institutes of Health (NIH) Database of Genotypes and Phenotypes (dbGaP). However, when Homer et al. (2008)[1] demonstrated a method with the potential to extract individual-level information from genomic summary statistics NIH proactively responded to this vulnerability by moving all genomic summary statistics information into controlled access.

This 2008 NIH policy change in the management of genomic summary statistics was intended to be a temporary[2] response to a new and unexpected potential vulnerability of information in an NIH database, to be followed by a comprehensive assessment of the implications. A 2012 NIH workshop on data aggregation[3] addressed the NIH policy for access to genomic summary statistics as part of a wider scope of recommendations on genomic data aggregation. Therefore, this workshop was convened specifically to initiate a contemporary reevaluation of the access model for genomic summary statistics, and assess the balance of risks and benefits of sharing summary statistics from genomics research.

Value of Genomic Summary Statistics

Finding 1: Genomic summary statistics provide extremely valuable information regarding which variants contribute to biological function and disease.

[1] N. Homer, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet*., 4 (8) (2008), pp. 1–9
[2] Zerhouni, E.A. and Nabel, E.G. Protecting aggregate genomic data. Science. 2008 October 3; 322(5898): 44. Published online 2008 September 4. doi: 10.1126/science.1165490
[3] Workshop on Establishing a Central Resource of Data from Genome Sequencing Projects. 2012. National Institutes of Health. https://www.genome.gov/27552142/workshop-on-establishing-a-central-resource-of-data-from-genome-sequencing-projects/

One major theme of discussion at this and the 2012 workshop was that genomic summary statistics are extremely valuable for scientific and clinical discovery. Genomic summary statistics facilitate the interpretation of one person's genome or exome by using hundreds or thousands (and in the future millions) of others for comparison. The information in genomic summary statistics can help answer critical questions about which genetic variants contribute to disease. For example, allele frequencies provide information related to pathogenicity, because common variants are unlikely to cause rare, serious diseases or lead to tumors. This information can also aid in the diagnosis of rare diseases. P-values provide information about which genomic regions and variants are robustly associated with diseases.

<u>Finding 2</u>: Public access to genomic summary statistics through central resources maximizes their value by enabling vastly more researchers to use genomic data in biomedical studies.

Policies that allow for the public sharing of genomic summary statistics would enable broad use of the results from association studies to address a wide range of scientific questions. Many research and clinical questions can be addresses using summary information but do not require the individual-level data. Since it is much faster to look up public data than to obtain permission to use data in controlled-access databases, only a limited number of researchers request access to controlled-access datasets. Orders of magnitude more investigators use public data sets, such as 1000 Genomes and ExAC, to obtain information on allele frequencies than request data from closed access databases[4,5] than use controlled access resources. In addition, summary information on allele frequencies and disease associations from large disease studies provide information that many more researchers could use than currently gain access. For example, scientists beyond the genomics community could make use of the genomic summary statistics, diversifying the ways the information is used and increasing the benefit to the public. Public sharing of genomic summary statistics would maximize discovery from this information and produce a much greater return on the funding investment and the contributions of research participants in each study.

Tens of thousands of human genomes and even more exomes have been sequenced, but summary statistics from this body of work are currently sequestered in individual project and institutional databases. While some of this sequestration is attributable to logistical and ethical issues, workshop participants identified policy guidelines set by the NIH and other major funding agencies after the Homer et al. paper as a major obstacle to unifying these resources.

The value of genomic data in biomedical research increases with scale, and genomic summary statistics enable the synthesis and interpretation of large amounts of genomic data. Enormous aggregations of data enable more powerful analyses and increase the opportunities for discovery, including detecting the effects of rare variants, small effects of common variants, and interactions among variants and with environmental factors.

[4] [Thousands] Data Access Request (DAR) Approvals and Disapprovals by Data Access Committee (DAC). https://gds.nih.gov/19dataaccesscommitteereview_dbGaP.html

[5] [Millions] MacArthur, D. https://macarthurlab.org/2016/08/17/announcing-the-exome-aggregation-consortium-paper/

Finding 3: **Sharing genomic summary statistics publicly would improve the ease of access to this information while reducing the need for access to individual-level datasets.**

Accessing genomic summary statistics held in dbGaP requires investigators to submit a Data Access Request to an NIH Data Access Committee, and often has limitations on how the information can be used. Since the NIH policy change, genomic summary statistics are generally available only in conjunction with the underlying individual-level data, meaning that investigators only interested in allele frequencies and disease association statistics within a particular study also have to be granted access to corresponding the individual-level data. However, obtaining the ability to download individual-level data to their computers is an unnecessary—and from a privacy-protective perspective, an undesirable—byproduct, because such superfluous access to individual-level data increases the potential for inadvertent release of the data or for a malicious actor to gain access to individual-level data. In contrast, publicly accessible genomic summary statistics browsers, such as the Exome Sequencing Project (ESP) and Michigan Imputation Server, can be queried through web searches, enabling access to allele frequencies and disease associations without requiring or even allowing access to the individual-level data in the controlled-access databases.

Finding 4: **A number of institutions have approved sharing of summary statistics and these resources are highly utilized in the clinical and research communities.**

Several presenters described how their institutions have allowed the sharing of genomic summary statistics in public-access databases, including the Exome Aggregation Consortium (ExAC) at the Broad Institute, Exome Variant Server (University of Washington), resulting from the National Heart Lung and Blood Institute's GO-ESP program, Accelerating Medicine's Partnership (AMP) Type 2 Diabetes Knowledge Portal (T2DKP) (Broad/University of Michigan/Oxford University), Michigan Imputation Server (University of Michigan), and AmbryShare[6] (Ambry Genetics). While these resources provide various types of information, all provide genomic summary statistics, because the institutions decided that they are of such scientific value as to warrant release relative to the minimal privacy risks to research participants. For example, ExAC (which includes data from 1000 Genomes, ESP, T2DKP, and other projects) provides allele frequencies that are widely used, with about 75,000 page views per week. To minimize the risk of identification raised by the Homer et al. methods, ExAC provides only summary information about exomes by high-level ancestry group; users can freely download the summary statistics without login or institutional signoff. AmbryShare makes available allele frequencies from groups of breast and ovarian cancer patients, and there are plans to expand to allele frequencies in conjunction with phenotype information for cohorts with other cancer types and diseases in the near future. Likewise, the disease-specific variant browser T2DKP provides allele frequency statistics from studies of type 2 diabetes and related traits.

Potential Risks and Harms

Finding 5: Privacy and confidentiality risks posed by genomic summary statistics are distinct from those posed by individual-level data.

[6] Ambry Share is a database maintained by the privately-held healthcare company Ambry Genetics and is not supported by NIH funding.

<u>Finding 6</u>: The degree of privacy harm that might occur related to inappropriate use of genomic summary statistics depends on what additional information is revealed by determining whether an individual participated in a particular research study.

The Homer et al. study identified a hypothetical risk from the sharing of genomic summary statistics where, if an individual's genome sequence is already known, it may be possible by statistical analysis of the study's allele frequencies for someone with ill intentions to determine whether the individual participated in the study. Workshop speakers described how the confidence of any such determination (still a hypothetical risk) would depend on multiple factors about the genomic summary statistics, research participants, and study design, with the risk lower for studies with: fewer genetic markers, more participants, broad population groups, or studies with less phenotype information.

For the institutions that are providing public genomic summary statistics, some factors, such as number of genetic markers or amount of phenotype information, were limited to decrease the risk of re-identifying the presence of an individual's sequence in their databases. Making these adjustments decreases risk, but also reduces richness and, therefore, the scientific value of the genomic summary statistics. Workshop participants acknowledged that no single threshold for such a tradeoff exists; however, there was unanimous support for promoting an appropriate balance between available and masked data to enable sharing while protecting participants' privacy.

Compared to an individual's genome sequence, which could provide information about a person's potential health issues and those of their family, the hypothetical risk posed by public access to genomic summary statistics is limited to the ability to deduce whether an individual participated in a particular research study. While determining such information could potentially provide previously unknown clinical status about an individual to the antagonist, such as the diagnosis of a particular disease in a case-control disease study, this information is bounded by the phenotypes studied and the need for the ill-doer to already have that individual's genome sequence or information on many genotypes (the possession of which could already be used to ascertain information about that individual's health status or risks independent of whether such genomic summary statistics were available). Therefore, any harm caused by possible phenotype revelations from genomic summary statistics could be a secondary breach of privacy. However, as one workshop participant observed, increasing access to genomic information with few protections against surreptitious testing mean that the barriers to obtaining an individual's sample for sequencing are likely to decrease in the future.

Workshop participants also discussed how the high level of technical knowledge required to use genomic summary statistics to re-identify an individual as being part of a study also constrains the likelihood of their misuse. As an example, participants noted that while GWAS studies in dbGaP remain under controlled access following the NIH policy change, genomic summary statistics are publicly available as an element of published results from many genome-wide association studies and no privacy breaches have been reported nor privacy harms identified to date.

<u>Finding 7</u>: There is greater privacy concern when studying potentially stigmatizing traits or vulnerable populations, because the outcomes of any privacy breach could cause greater

harm. Studies of this nature merit additional considerations about appropriate protections.

One important caveat to the risk estimate for sharing genomic summary statistics regards studies including individuals with sensitive phenotypes (e.g., mental illnesses, substance abuse) or vulnerable populations in which risks related to potential stigma might be relevant. The risk of identification from a Homer et al. type analysis is still hypothetical and small, but the harms potentially could be greater. Developing criteria regarding which phenotypes are sensitive and which populations should be considered vulnerable with respect to potential stigmatization risks or other concerns will require further discussion. Workshop participants noted that studies that involve these issues are a small proportion of genomics research, and a simple accommodation might be not to release genomic summary statistics from such studies.

Finding 8: There is "institutional risk" in public release of genomic summary statistics related to the potential to damage public trust if expectations related to sharing are not clear.

Even if an individual does not suffer harm from being revealed as a participant in a study through genomic summary statistics, NIH and the scientific community may suffer institutional harm to their reputations as stewards of participant research information should such a breach occur. Loss of participant trust in the NIH could have a major impact on the willingness of participants to volunteer for studies.

Additionally, workshop participants expressed some concerns about the consequences from the public perception of a reevaluation of the NIH policy for the management of genomic summary statistics. The general lack of understanding among members of the public as to how genomic summary statistics are related to their individual information could cause public anxiety. Some workshop participants noted that the distinction between individual-level data and genomic summary statistics has very often been blurred during discussions about genomic data sharing. Research participants who have contributed to genomics research studies may be alarmed by conversations about increasing sharing of genomic information, even if that information is simply summary descriptions of data aggregated across hundreds to thousands of individuals through controlled access. The public also may not be aware of the value of genomic summary statistics to advance biomedical research and genomic medicine. Precise terminology in discussions of sharing genomic summary statistics is important for clarity among researchers and transparency with the public.

Public Engagement

Finding 9: The research community needs to better explain the difference between genomic summary statistics and individual-level data with regard to the type of information and the associated risks.

Finding 10: Transparency is needed for participants regarding plans to share genomic summary statistics from research studies.

For many genomics research studies, privacy risks have long been a major concern and efforts have been taken by researchers and institutions to minimize these risks. Some workshop attendees suggested that the risk of harm to research participants and to an institution's

reputation could potentially be mitigated if participants better understood and were better informed about how their information was to be aggregated and analyzed, and how those findings would ultimately be shared in the scientific literature (and otherwise) to contribute to new discoveries about health and disease. While a few suggested that participants should specifically consent for sharing of genomic summary statistics separate from consent to participate in research, most felt that in most cases the research community should instead clearly and effectively communicate to participants that genomic summary statistics are derived by analyzing aggregated individual data contributed by all participants in a study and, in fact, are the results of the studies for which they volunteer. Therefore, consent to participate in a study means consent to aggregate and analyze the information generated, in the form of genomic summary statistics, to be shared in the scientific literature and through other public scientific resources. In situations where sensitive traits or vulnerable populations are included in the study design, this default position should be considered for its appropriateness and additional protection strategies employed if warranted.

Studies have demonstrated that research participants understand the tradeoff between assuring personal privacy and maximizing the utility of their data to achieve scientific progress. Providing potential participants a more transparent opportunity to consider that tradeoff could increase understanding related to study participation and mitigate potential institutional risk of losing the trust of research participants if a loss of privacy were to occur. Workshop participants pointed to the results of several surveys that indicated research participants' desire to advance research generally outweighs their privacy concerns, resulting in increased consent for public sharing.[7,8] Other work in the literature finds that individuals who enroll in research studies usually do so to contribute to science and help individuals like themselves. Therefore, researchers have a moral obligation in situations when risks to individual privacy are minimal to facilitate genomic data sharing[9] on behalf of the participants who volunteer to contribute to research.

Privacy Enhancing and Security Technologies

Finding 11: Privacy enhancing and security technologies can provide useful protections as part of a risk-mitigation strategy. Such technologies might be appropriate in the context of studies involving stigmatizing traits or vulnerable populations.

Numerous privacy enhancing strategies have become available in the past several years that might be used to mitigate privacy leaks and risk to participants. These technologies are being engaged to protect individual-level data in several settings and some workshop participants mentioned these approaches could provide added protection to genomic summary statistics as well. One presenter explained how the Beacon Project of the Global Alliance for Genomics and Health is applying privacy technologies in combination with a strong privacy policy to provide protection for genomic variant information shared.

[7] McGuire, A. L., Hamilton, J. A., Lunstroth, R., McCullough, L. B., & Goldman, A. (2008). DNA data sharing: Research participants' perspectives. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *10*(1), 46–53.

[8] Sandra Soo-Jin Lee. (Workshop Presentation 5/19/16; *unpublished data*) Patient VALUES Study.

[9] Knoppers, B.M., Harris, J.R., Budin-Ljøsne, I., Dove, E.S. (2014). A human rights approach to an international code of conduct for genomic and clinical data sharing. *Hum Genet*. 2014; 133(7): 895–903.

However, while privacy technologies provide additional protection for information, they have costs, both in lost utility and added resource burdens. Workshop participants expressed concern that the computational burden of encryption methods, especially for the enormous amount of genomic summary statistics being generated, could ultimately inhibit sharing by creating an impractical standard of protection. However, they noted that privacy- enhancing technology could be appropriate to mitigate risks for information where higher levels of protection should be considered (e.g., sensitive phenotypes or vulnerable populations).

One prominent discussion point during the workshop was whether some form of an intermediate model between fully public access and controlled access could provide a useful degree of protection. Other organizations are considering a registration system to lessen the burden of access, but provide greater protection than fully open models for certain individual-level data.[10, 11, 12] Lower levels of registration and validation were proposed, including "light" registrations such as requiring users to submit a username and password. However, many participants argued that the protections provided by a light registration system would be superficial, because it would provide no proof of user identity. It was also argued that a light registration would provide little protection against leakage or malicious abuse, while inhibiting the ability of other resources to embed or link to the registered resource (e.g., dbSNP could not include allele frequencies from ExAC) and would greatly hinder the availability of resources through APIs.

The group agreed that any barrier will have a negative effect on usage of genomic summary statistics, but there was substantial debate about how to determine the appropriate degree of utility lost to achieve the protections gained through a registration system. In light of the high utility of genomic summary statistics and low risk of harm from misuse of the information, workshop participants generally agreed that it is inappropriate to restrict genomic summary statistics with the same level of protection as individual-level data.

Most workshop attendees also agreed that a more elaborate system of protections could provide greater protections when warranted, so while it can introduce more "burden," it could be appropriate when sharing genomic summary statistics from studies including more sensitive phenotypes or vulnerable populations.

Recommendations

In consideration of the findings listed above, the workshop participants provided the following recommendations:

1. NHGRI should recommend that NIH reconsider the policy for maintaining all genomic summary statistics under controlled access, and develop a default public access model based on transparent policy considerations for most genomics studies.

[10] Philippakis, A. A., Azzariti, D. R., Beltran, S., Brookes, A. J., Brownstein, C. A., Brudno, M., ... & Dumitriu, S. (2015). The Matchmaker Exchange: a platform for rare disease gene discovery. Human mutation, 36(10), 915-921.http://onlinelibrary.wiley.com/doi/10.1002/humu.22858/full

[11] Kosseim, Patricia, et al. "Building a data sharing model for global genomic research." Genome biology 15.8 (2014): 1. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0430-2

[12] Genetic Alliance. Platform for Engaging Everyone Responsibly. http://www.geneticalliance.org/programs/biotrust/peer

2. NHGRI should work with NIH to engage the public on discussions of the access policy and the relative risks and benefits to participants and science from public sharing of genomic summary statistics, including institutional risks resulting from any changes to the current policy.

3. NHGRI should work with NIH and public stakeholders to define “sensitive” phenotypes and vulnerable populations with regard to sharing genomic summary statistics in order to develop alternative sharing models as appropriate.

4. NHGRI and NIH should work with the public to anticipate education or resources needed to explain any relative risks of public access to genomic summary statistics, any policy modifications, and any risk mitigation strategies employed in those modifications.