Original Paper

Public Health Genomics

Public Health Genomics 2011;14:319–324
DOI: 10.1159/000324705

Received: April 27, 2010
Accepted after revision: January 26, 2011
Published online: March 24, 2011

Public Perspectives Regarding Data-Sharing Practices in Genomics Research

S.B. Haga J. O'Daniel

Institute for Genome Sciences & Policy and Sanford School of Public Policy, Duke University, Durham, N.C., USA

Key Words
Data-sharing · Genomics research · Public attitudes

Abstract
Background: Genomics research data are often widely shared through a variety of mechanisms including publication, meetings and online databases. Re-identification of research participants from sequence data has been shown possible, raising concerns of participants' privacy. ***Methods:*** In 2008–09, we convened 10 focus groups in Durham, N.C. to explore attitudes about how genomic research data were shared amongst the research community, communication of these practices to participants and how different policies might influence participants' likelihood to consent to a genetic/genomic study. Focus groups were audio-recorded and transcripts were complemented by a short anonymous survey. Of 100 participants, 73% were female and 76% African-American, with a median age of 40–49 years. ***Results:*** Overall, we found that discussants expressed concerns about privacy and confidentially of data shared through online databases. Although discussants recognized the benefits of data-sharing, they believed it was important to inform research participants of a study's data-sharing plans during the informed consent process. Discussants were significantly more likely to participate in a study that planned to deposit data in a restricted access online database compared to an open access database ($p < 0.00001$). ***Conclusions:*** The combination of the potential loss of privacy with concerns about data access and identity of the research sponsor warrants disclosure about a study's data-sharing plans during the informed consent process.

Introduction

One of the major themes characterizing the genome era has been a policy of open data-sharing. Open data-sharing promises to more rapidly advance research by providing the opportunity for other researchers to validate a study's findings, combine multiple datasets for analysis and test new hypotheses. Several studies, however, have shown the possibility of identifying participants based on analysis of these data [1–3]. The issue has garnered a range of opinions on how best to balance data-sharing with privacy and protection of research participants [4, 5].

In addition to traditional methods of data-sharing such as presentations or posters at professional meetings and publication in peer-reviewed journals, genomics data are often deposited in online databases. In some cases, participant characteristics (demographic and clinical) may accompany sequence or genotype data. Genomic databases may be established per study and operated by the researcher, company, research institute, or be maintained

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

Accessible online at:
www.karger.com/phg

Susanne B. Haga, PhD
Institute for Genome Sciences & Policy and Sanford School of Public Policy
Duke University, 304 Research Drive, Room 228, Box 90141
Durham, NC 27708 (USA)
Tel. +1 919 684 0325, E-Mail susanne.haga@duke.edu

by a third party such as a government agency. Access to these databases varies from completely open access to limited access for approved researchers only [6]. For example, some databases maintained by government agencies are completely open to anyone with Internet access (i.e. GenBank, dbGaP). Other government databases require consent of the user to some type of data use certification agreement (e.g. Genetic Association Information Network (GAIN)) and/or approval from a data access committee. Further, data-sharing may be a requirement for publication or of the funding source.

In 2008, Homer et al. [1] developed a metric to determine the presence of an individual of known genotype from pooled genomic data. Though limitations have been noted and further refinement is needed to improve identification of the proband and his/her relatives within a pooled dataset, follow-up studies have validated the ability to re-identify individuals [7–10]. Identification of disease states of participants in genome-wide association studies has also been demonstrated [11], further raising medical privacy concerns. The work by Homer et al. [1] and others have instigated changes in data-access policies for some genomic databases to reduce potential identification of research participant [12]. While such policy changes will limit access to some data, limiting individual harm, these policies are not standard for all genomic databases. Furthermore, efforts to balance social benefits to individual risks by the research community raise concerns about their ability to impartially achieve such a balance given researchers' inherent conflict of interest.

Despite the common practice of data-sharing in genomics research, investigators are just beginning to understand the public's views and the potential impact of these practices on research participation [13–15]. In contrast, several studies have examined the public's attitudes toward biobanking, a related issue involving storage and multiple use (or sharing) of DNA specimens and potentially personal or clinical information, as well as opt-out policies and appropriate informed consent [16–20]. In general, the public's concerns regarding biobanking center on issues of privacy, autonomy and underlying mistrust of how samples may be used and by whom. In this article, we describe findings from a focus group study of predominantly African-Americans about attitudes toward data-sharing in genetic/genomic research and the potential impact of possible practices on research participation. Based on public concerns about privacy, confidentiality and future uses and users with biobanks, we hypothesize that research participants will express similar concerns and may believe researchers should disclose information about who maintains the genetic/genomic database as well as who has access to the data as it may influence their perception of risk/harm. Certain entities (e.g. government, law enforcement or pharmaceutical industry) may raise more concern than others (e.g. researchers) if access is permitted or possible.

This study provides additional insight into concerns and expectations regarding data-sharing from the 'potential' participant's perspective. The results will be useful for development of future policies related to both data-sharing and disclosure of data-sharing plans to participants prior to enrollment.

Methods

Study Population

We were particularly interested in gathering perspectives from minority communities, particularly African-Americans, given that this population group is frequently underrepresented in genetic/genomic research. Participants were recruited from community locations across Durham, N.C., USA, through advertisements in predominantly African-American churches, flyers posted in community centers and libraries in predominantly African-American neighborhoods and advertising in newsletters and radio stations targeted toward the African-American community as well as word-of-mouth. A meal and USD 25 were provided as compensation for their participation in the focus group. The study was approved by the Duke University Health System Institutional Review Board.

Focus Group Design

A three-part questioning route guide employing a 'funneling' approach and utilizing 3 hypothetical research vignettes was designed to engage participants in an open discussion about returning research results and data-sharing practices. This article will focus on the findings related to data-sharing; the data on returning research results is described elsewhere (O'Daniel and Haga, in press). Initial questions were intentionally broad regarding genetic and genomic research, encouraging discussants in free expression. A brief overview of genetic/genomic research was presented by the moderator, highlighting the various options for data-sharing and returning research results. Potential risks and benefits for the various options, including no data-sharing, were also presented. For data-sharing practices, we noted that identifiability of research participants was possible based on sequence data alone. Subsequent questions were more targeted and issue-specific, following the presentation of 3 hypothetical research vignettes illustrating combinations of different options for returning genetic/genomic research results as well as data-sharing practices amongst other researchers. In brief, Vignette 1 described a large-scale, case-control study in which the research findings would be disseminated through publications and presentations at professional meetings. The researchers would also share their research data with other scientists studying the same disease upon request. Vignette 2 described a large-scale study of the prevalence of a candidate gene variant within the African-American population.

The research findings would be disseminated through publications and presentations at professional meetings, and research data would be shared via open-access, online database maintained by a government entity. Vignette 3 described a large-scale, multi-site genome study of individuals with a family history of heart disease in which research findings would be disseminated through publications and presentations at professional meetings, and research data would be shared via centralized, online database accessible only to researchers involved in the study.

Two pilot focus groups were held to obtain feedback on the understandability of the content and questions as well as meeting logistics (e.g. food, time and location). Materials were revised accordingly.

Focus Groups

Ten focus groups were held between February 2008 and February 2009 at various locations throughout the Durham community. Each focus group discussion was audio-recorded and transcribed.

Data Analysis

The focus groups yielded 3 datasets: (1) socio-demographic data, (2) digitally recorded and transcribed text from each focus group session and (3) anonymous responses to a short survey. Qualitative analysis of transcripts was performed including coding and semantic content analysis. Initial coding categories of responses were developed by the authors and a research assistant; areas of intercoder disagreement were resolved by consensus. Descriptive and analytic statistics were utilized for assessment of the socio-demographic and survey data using the STATA statistics package. The data gathered from this study may not be generalizable given the small sample size and recruitment from one region. In addition, it is possible that the moderators (the authors) of the focus groups may have biased participants as we acknowledged that we conduct genetics research and are from a major university that conducts a lot of research in the area.

Results

Focus Group Discussants

One hundred individuals participated in 10 focus groups. In summary, discussants were predominantly female (73%), African-American (76%), between the ages of 40 and 49 years (36%), and had some college education but no degree (26%) (O'Daniel and Haga, in press).

Responses to Research Vignettes

Three hypothetical vignettes were discussed to illustrate the various ways researchers may share and disseminate data. Following each vignette, discussants were asked to consider the scenario and the data-sharing practices and whether they would participate in a similar type of study if approached. In general, discussants did not express concerns regarding dissemination of research findings through traditional methods such as scientific publication and/or meeting presentation. Significant concerns were raised, however, about being able to link research data back to personal information or identity. This worry about loss of privacy and/or confidentiality appeared to be linked to several factors including the perceived security of the data storage and the ability to link the research data back to the individual through codes or other means.

Several discussants seemed to be especially skeptical of the security of an open, government database as described in Vignette 2. While some of the concerns were linked to the prospect of open-access, many voiced significant misgivings about the integrity of a government entity to protect the privacy of their data and the lack of trust.

'Well, I believe that if a government official wanted my DNA, that they would break that code and get it … So, for me, it's a trust issue.' (focus group (FG) 2, female)

'It's just something about that word, "government," because there were other studies in the past on the African American population. And I probably would not [participate].' (FG 2, female)

'That did concern me … that it's, you know a government internet … and the world, you know, can have access to it. But if they really did separate it from, you know, the data, away, then maybe.' (FG 3, female)

'Well, I don't know what they are going to do with my DNA. [The study] could be a fluke. You going to get my stuff and then there's paranoia, I got the FBI looking for me for something.' (FG 6, male)

'As long as I knew that my sample was going to be coded and not in any way traced back to me, I would feel comfortable. If I didn't know that, I feel uncomfortable … I'm sorry, but I don't trust the powers that be.' (FG 8, male)

For some, a coded link to personal information raised worry about potential harm of third-party access to data, particularly discrimination.

'… well, they could trace my information back to me. I would want to know if someone disclosed that.' (FG 1, female)

'Because if you release your [information], and you don't mind someone sharing your information, it could go to the insurance companies, your rates could go up.' (FG 8, female)

Support for Data-Sharing

Despite potential concerns, many discussants recognized the benefits of data-sharing and stated scientific data-sharing would not affect their decision to participate in a study.

'It wouldn't bother me, because, for one, I know scientists aren't really, you know, looking for personal stuff. They're not – they don't care about what the name is. They want the bottom-line raw numbers.' (FG 3, female)

'Oh, I think that's the important thing that they ought to share. That's the purpose of it.' (FG 4, female)

'I don't think it would affect my decision one way or the other, because it would be based on various other factors. I don't think that would enter into it. As several people have said, the more widely this information is [shared] the better.' (FG 8, male)

'Peer review is so important in science, [for] the credibility for the study …'. (FG 4, male)

Discussants were not able to identify circumstances when research data should not be shared within the scientific community, though some sought assurance that it would be anonymous.

'As long as it's not violating our privacy, if we're doing codes and all that other stuff, that other guy's not going to know who we are. If they receive our full names and are publicizing [them], that's another story.' (FG 7, female)

However, some also acknowledged situations when access to research data should be restricted.

'I really don't think the drug company or – somebody who has a – who can stand to benefit from the results the way it's presented. I don't think that should be allowed. I mean, especially if it's presented as scientific information.' (FG 4, female)

Informing Research Participants about Data-Sharing Methods

Although many expressed that the data-sharing method would have little effect on their decision to participate in a study, the vast majority of discussants believed researchers *should* disclose it before they consented to participate in the study. In particular, discussants believed they should be informed about how research participants' confidentiality would be protected.

'So, you know, you can make an informed choice.' (FG 3, female)

'I think however they plan to [share the data] – they should inform so that you know what they are doing, and [where] it's going to go – any method that they use.' (FG 6, female)

'You could just put [it] like in the consent form "This may or may not be published; this may or may not be …" – the uses, not like [a] promise.' (FG 9, male)

Discussants also raised the possibility that they would come across a media report of a study in which they were a participant but were unaware the study findings would be publicly announced. One of the discussants expressed concern about her unawareness of secondary uses of her data, however, it is highly unlikely that she would come to know which studies her data were used in if they were publicly available, highlighting some confusion about the harms of data-sharing.

'Because you know, you suppose – okay, they didn't tell you, and all of a sudden you go to the library or something, and this magazine, and you say, oh, researchers have found – have discovered this and that and that, so who knows, you might be part of that study, you know? But they didn't tell you that they were sharing it.' (FG 8, female)

Survey Results

To further explore the impact of data-sharing via online databases on decisions to participate in a research study, we administered a short anonymous survey at the conclusion of the focus group. Ninety-nine of the 100 participants completed the survey. The majority of discussants indicated that they would be very/somewhat likely to participate in a genomic research study if the data were to be shared through a restricted online database (84%). This was significantly greater than those who indicated that they would participate in a study in which data were to be shared through an open online database (52%; $p < 0.00001$).

Discussion

The field of genomics has taken a unique approach to data-sharing. Few other types of clinical research are required to deposit data in public databases, potentially accessible by any person or entity. While this practice may be advantageous to the scientific community, some disadvantages to research participants deserve further attention [5, 21, 22]. Our study showed that while focus group discussants recognized the importance of data-sharing, they desired to be informed of how the data would be shared due to concerns about privacy and confidentiality. These data, gathered from a primarily African-American population, validate other findings from smaller studies of predominantly White research participants and the general public about the trade-off between data-sharing and concerns of privacy along with the desire to have some control/input regarding data-sharing practices [13–15].

Discussants recognized both the personal risks of data-sharing as well as the larger benefits to biomedical research. In particular, some discussants were concerned about potential harms that could arise if their participation and individual results could be revealed from databases. Government involvement in data-sharing (as the sponsor or users of the database) raised substantial concerns in our study population, particularly as a number of participants recalled past abuses by government researchers in studies of vulnerable or minority communi-

ties. Mistrust, insecurity and suspicion have similarly been noted as concerns with respect to biobanking [23]. Concerns about potential negative implications for some ethnic groups have led to the development of restricted data-sharing policies vs. an open-access policy [24]. In regards to company-sponsored databases, others have expressed concern about the security of the data in the event of the sale of private databases if a company is purchased or becomes insolvent [25]. As none of the vignettes included commercial involvement, little discussion about commercial groups occurred, and thus, we were unable to elucidate if discussants' perspectives might differ from those related to government entities. Some concerns were spontaneously raised about drug companies benefiting from data-sharing, and thus, the potential for commercial profit from research participants' data may be a significant influencing factor for participants, similar to other findings with respect to data-sharing [13, 15].

Similar to national policy [26] and recommendations made by others to obtain consent for future uses of genetic data [27] and stored tissue [28], our data support informing research participants of the method(s) of data-sharing, data access policies and database privacy protections. Specifically, the National Institutes of Health (NIH) policy on data-sharing of prospective genome-wide association studies (GWAS) recommended that investigators obtain consent from participants for data to be shared through the NIH-GWAS database [26]. Re-consent may be needed for sharing of data generated from existing collections before the data can be deposited in public databases. This policy aligns with our findings of discussants' desire to be informed of data-sharing options, but limits data-sharing as re-consent may not be possible depending on the age of the collection, ability to contact participants and the possibility that only a subset will consent to data-sharing. However, for other non-GWAS studies, no standard policy is available to address notifying or informing research participants of how data may be shared. The potential loss of privacy and confidentiality would appear to meet the threshold of 'reasonably foreseeable risks (45 CFR 46.116)' to the participant, another reason to consider disclosing information about a study's data-sharing plan during the informed consent process.

As noted in a prior study about patient preferences to learn of research using anonymized or coded DNA samples [28], a potential benefit to disclosing modes of data-sharing is to maintain trust of participants and transparency of research practices [13, 15], a particularly important aspect to minority research participants [29]. Although several elements have been recommended to be disclosed to participants regarding the benefits and risks of data-sharing [13, 30], we suggest a concise section with simple but explicit language on the benefits and risks, access policies, privacy and confidentiality protections, and information about the sponsor of the database. Too much information may lead to confusion about the study's data-sharing plans. Highlighting the need for concise and simple language, the study by McGuire [14] reported participants' confusion about the data-sharing plan described in the informed consent.

In addition, consideration of use of a tiered consent with respect to data-sharing merits further study regarding feasibility and impact on overall goals of data-sharing given discussants' higher likelihood to participate in a genomic research study if the data were to be shared through a restricted database rather than an open-access database. These findings are similar to participant preferences about restricted use of DNA samples in biobanking, although very few participants have withdrawn consent because of broad use of samples [17] suggesting the impact on research participation may be less than indicated for the issue of data-sharing as well. Similar to other findings [13], our discussants emphasized, however, that disclosure of data-sharing practices was important in order to make a truly informed decision and fulfill the fundamental ethical principles of participant autonomy and respect.

Although the ability to identify a research participant through analysis of aggregate data or through a combination of demographic information, disease status and genetic sequence may be rare, prospective participants may be justly concerned about data-sharing practices as was observed in this study. While participant concerns are only one aspect to be considered in the development of policies regarding disclosure of data-sharing practices, the apparent consensus from this and other studies strongly supports careful consideration by institutional review boards, policy-makers and researchers regarding the establishment of standards for disclosure of data-sharing plans to research participants. Overall, the scientific community must strive to define a balance between protecting and showing respect to research participants and advancing scientific discovery for the potential benefit of all.

Acknowledgement

We would like to thank our research assistant, Ms. Genevieve Tindall, for her kind assistance on this project and the manuscript. This study was supported by NIH grant #1R03-HG-004312.

References

1 Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 2008;4:e1000167.

2 Lin Z, Owen AB, Altman RB: Genetics. Genomic research and human subject privacy. Science 2004;305:183.

3 McGuire AL, Gibbs RA: Genetics. No longer de-identified. Science 2006;312:370–371.

4 Church G, Heeney C, Hawkins N, de Vries J, Boddington P, Kaye J, Bobrow M, Weir B: Public access to genome-wide data: five views on balancing research with privacy and protection. PLoS Genet 2009;5:e1000665.

5 Lowrance WW, Collins FS: Ethics. Identifiability in genomic research. Science 2007; 317:600–602.

6 Resnik DB: Genomic research data: open vs. restricted access. IRB 2010;32:1–6.

7 Sankararaman S, Obozinski G, Jordan MI, Halperin E: Genomic privacy and limits of individual detection in a pool. Nat Genet 2009;41:965–967.

8 Visscher PM, Hill WG: The limits of individual identification from sample allele frequencies: theory and statistical analysis. PLoS Genet 2009;5:e1000628.

9 Braun R, Rowe W, Schaefer C, Zhang J, Buetow K: Needles in the haystack: identifying individuals present in pooled genomic data. PLoS Genet 2009;5:e1000668.

10 Jacobs KB, Yeager M, Wacholder S, Craig D, Kraft P, Hunter DJ, Paschal J, Manolio TA, Tucker M, Hoover RN, Thomas GD, Chanock SJ, Chatterjee N: A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. Nat Genet 2009;41:1253–1257.

11 Lumley T, Rice K: Potential for revealing individual-level information in genome-wide association studies. JAMA 2010;303:659–660.

12 National Institutes of Health: NIH modifications to genome-wide association studies (GWAS). http://grants.nih.gov/grants/gwas/data_sharing_policy_modifications_20080828.pdf (accessed August 28, 2008).

13 Lemke A, Wolf W, Hebert-Beirne J, Smith M: Public and biobank participant attitudes toward genetic research participation and data sharing. Public Health Genomics 2010;13: 368–377.

14 McGuire AL: Identifiability of DNA data: the need for consistent federal policy. Am J Bioeth 2008;8:75–76.

15 Trinidad SB, Fullerton SM, Bares JM, Jarvik GP, Larson EB, Burke W: Genomic research and wide data sharing: views of prospective participants. Genet Med 2010;12:486–495.

16 Hoeyer K, Olofsson BO, Mjorndal T, Lynoe N: Informed consent and biobanks: a population-based study of attitudes towards tissue donation for genetic research. Scand J Public Health 2004;32:224–229.

17 Johnsson L, Hansson MG, Eriksson S, Helgesson G: Patients' refusal to consent to storage and use of samples in swedish biobanks: cross sectional study. BMJ 2008;337:a345.

18 Murphy J, Scott J, Kaufman D, Geller G, LeRoy L, Hudson K: Public perspectives on informed consent for biobanking. Am J Public Health 2009;99:2128–2134.

19 Kaufman DJ, Murphy-Bollinger J, Scott J, Hudson KL: Public opinion about the importance of privacy in biobank research. Am J Hum Genet 2009;85:643–654.

20 Tupasela A, Sihvo S, Snell K, Jallinoja P, Aro AR, Hemminki E: Attitudes towards biomedical use of tissue sample collections, consent, and biobanks among finns. Scand J Public Health 2010;38:46–52.

21 Foster MW, Sharp RR: Share and share alike: deciding how to distribute the scientific and social benefits of genomic data. Nat Rev Genet 2007;8:633–639.

22 Caulfield T, McGuire AL, Cho M, Buchanan JA, Burgess MM, Danilczyk U, Diaz CM, Fryer-Edwards K, Green SK, Hodosh MA, Juengst ET, Kaye J, Kedes L, Knoppers BM, Lemmens T, Meslin EM, Murphy J, Nussbaum RL, Otlowski M, Pullman D, Ray PN, Sugarman J, Timmons M: Research ethics recommendations for whole-genome research: consensus statement. PLoS Biol 2008; 6:e73.

23 Melas PA, Sjoholm LK, Forsner T, Edhborg M, Juth N, Forsell Y, Lavebratt C: Examining the public refusal to consent to DNA biobanking: empirical data from a swedish population-based study. J Med Ethics 2010;36: 93–98.

24 Parker M, Bull SJ, de Vries J, Agbenyega T, Doumbo OK, Kwiatkowski DP: Ethical data release in genome-wide association studies in developing countries. PLoS Med 2009; 6:e1000143.

25 Tavani HT: Genomic research and data-mining technology: implications for personal privacy and informed consent. Ethics Inf Technol 2004;6:15–28.

26 National Institutes of Health: Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS). 2007. http://grants.nih.gov/grants/guide/notice-files/not-od-07-088.html.

27 McGuire AL, Gibbs RA: Meeting the growing demands of genetic research. J Law Med Ethics 2006;34:809–812.

28 Hull SC, Sharp RR, Botkin JR, Brown M, Hughes M, Sugarman J, Schwinn D, Sankar P, Bolcic-Jankovic D, Clarridge BR, Wilfond BS: Patients' views on identifiability of samples and informed consent for genetic research. Am J Bioeth 2008;8:62–70.

29 Corbie-Smith G, Thomas SB, Williams MV, Moody-Ayers S: Attitudes and beliefs of African Americans toward participation in medical research. J Gen Intern Med 1999;14: 537–546.

30 National Institutes of Health: NIH points to consider for IRBS and institutions in their review of data submission plans for institutional certifications under NIH's policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS). http://grants.nih.gov/grants/gwas/gwas_ptc.pdf (accessed July 30, 2008).