November 30, 2006

NIH GWAS RFI Comments
National Institutes of Health
Office of Extramural Research
6705 Rockledge Dr., Rm. 350
Bethesda, MD 20890-7963

Comments of the American Civil Liberties Union (ACLU)

Re: Department of Health and Human Services, National Institutes of Health Request for Information (RFI): Proposed Policy for sharing of data obtained in NIH-supported or conducted genome-wide association studies (GWAS), 71 Federal Register 51629 - 51631.

The American Civil Liberties Union (ACLU) welcomes the opportunity to comment on the proposed policy of the National Institutes of Health (NIH) to create a centralized repository for Genome-Wide Association Study (GWAS) data (71 Federal Register 51629-51631). The ACLU is a nationwide, non-partisan organization of more than 500,000 members dedicated to protecting the principles of liberty, freedom and equality as set forth in the Bill of Rights to the United States Constitution. For almost 80 years, the ACLU has sought to preserve and strengthen privacy in all aspects of American life.

The NIH defines a genome-wide association study as "any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition."[i] The NIH proposes to create a centralized repository of genotype and phenotype datasets arising from NIH-supported genomic research that can be made available to a wide range of scientific investigators.

The ACLU commends the NIH for recognizing that information generated about individuals through genetic research may be "sensitive and substantial" and that "it is critically important that the privacy and confidentiality of the participants be protected."[ii] We believe, however, that these privacy concerns have not been adequately aired or addressed, and that this proposal should not proceed without far greater attention to these issues in the form of broad public discussion and consultation with privacy experts, in addition to adoption of the recommendations contained in this document.

The ability of an individual to exercise control over the collection, maintenance, and use by the government or commercial entities of his or her sensitive personal information is central to personal integrity and human dignity. Genetic information is highly sensitive information in that it has the potential to reveal a virtually unlimited amount of personal

and familial information about a person. The storage, sharing and use of this information create new possibilities for privacy infringements, stigmatization and discrimination. For example, employers or insurance companies who might gain access to this information could discriminate against individuals on the basis of a genetic predisposition to a particular condition or trait.

Scientific freedom of inquiry is also an important societal goal that must be respected. A repository such as the one being proposed by NIH and the facilitation of "rapid and broad data access" could enhance opportunities for advances in medical research. Nevertheless, the efficiency and convenience that may be offered by a central repository provide insufficient justification for overriding basic and well-accepted protections for research participants in the form of privacy and informed consent, which are the appropriate basic models for conducting scientific research on human subjects.

The NIH should not move forward with the GWAS repository without a baseline of protection such as that contained in the Genetic Information Non-Discrimination Act of 2005 (S. 306, as passed unanimously by the U.S. Senate in February, 2005). Public trust in biomedical research cannot be maintained without this basic degree of protection. At a minimum, these protections should be firmly established as law before research participants are exposed to additional risks of discrimination through the construction of a large database of genetic and phenotypic information.

Below we describe our primary concerns associated with the NIH proposed policy as falling within the following five areas: 1) De-Identification and Data Inclusion; 2) Data Access; 3) Informed Consent; 4) Public Benefits; and 5) Intellectual Property. Overall, we believe that far more rigorous public discussion is warranted to determine whether a national repository can be created with protections that are sufficient for safeguarding the privacy of research participants. Therefore, beyond the concerns we've outlined below, we recommend strongly that the NIH follow the comment period with a broader public discussion to include researchers, privacy advocates and other stakeholders.

1) <u>De-Identification and Data Inclusion</u>

The NIH proposal states: "In order to minimize the risks to study participants, data will be submitted to the GWAS data repository without identifiable information and using a random, unique code." At the same time, the NIH "strongly encourages the submission of curated and coded phenotype, exposure, genotype, and pedigree data."[iii]

We find the NIH's apparent method for protecting individual privacy while allowing for the creation of a large database of genotypic and phenotypic data to be insufficient to achieve adequate privacy protection. First, it is questionable as to whether a database containing genotype information can ever be considered truly anonymous, or "de-identified" when DNA itself is considered to be a "unique identifier." In 2004, Lin and colleagues estimated that an individual could be uniquely identified with access to as few as 75 single-nucleotide polymorphisms (SNPs) from that person.[iv] While individual identification is currently limited to cases where a reference sample is available, recent

developments in genetic testing to predict ethnicity and facial characteristics[v] indicate that it is conceivable that genotype alone may one day be sufficient for identification.[vi] The presence and linkage of phenotypic information as provided by the repository will also increase the risk of individual identification.

Next, the NIH proposal does not define "identifiable information." Is "identifiable information" simply name, address, telephone number, and social security number? Some phenotypic information that could be considered useful to a wide array of researchers, such as zip code, ethnicity, type or location of employment, could be used to identify subjects. NIH should spell out precisely what information fields are barred from the repository and which are not contemplated for inclusion.

Similarly, while the proposal does not specify "identifiable information" it simultaneously appears to encourage the inclusion of large amounts of "curated and coded" information. The more information contained in the database, the greater the risk to participants. In addition, a lack of standardized data collection requirements means that individuals who participate in multiple studies may be at a greater risk of being readily identifiable; data considered "non-identifiable" in the context of one particular study may become identifiable when it is joined with other "non-identifiable" information from one or more additional studies.

The NIH appears to acknowledge that identification of individuals in the database is a real possibility. The proposal states that researchers seeking data from the GWAS data repository will be asked to stipulate that they will, among other things, "not attempt to identify individual participants from whom data within a dataset were obtained."[vii]

We acknowledge that the proposal's requirement that data be submitted in an encoded, anonymized fashion is an important step toward protecting individual privacy; if properly orchestrated, the coding of identifiable information can offer significant privacy protection. However, the NIH proposal does not specify the coding scheme or clarify how the sharing of encryption keys will work. For example, the proposal simply states that "keys to codes will be held by submitting institutions."[viii] The proposal fails to specify whether standard rules for oversight and dissemination of keys might be issued for all submitting institutions, whom at those institutions will hold the key, and how many people might have access.

Finally, under the section entitled, "Data Management," the proposal states that "the repository will also accept GWAS datasets contributed from other sources."[ix] It is not clear as to whether this data will be subject to the same requirements as data submitted by NIH-sponsored institutions and we recommend that protections be explicitly stated for the datasets prior to any implementation of the proposal.

In sum, there is a well-recognized tension in the scientific and policy communities between the desire to both protect privacy and allow for the collection and storage of genetic and phenotypic information for research.[x] Calling for encryption and asking researchers to simply withhold identifiable information are grossly insufficient for

addressing this fundamental challenge. **The ACLU recommends that the NIH develop far greater clarity as to what constitutes "identifiable information" including whether and under what circumstances genetic data itself may be "identifiable." The ACLU also recommends that the NIH solicit encryption coding advice from numerous cryptographers and privacy specialists to determine whether personally identifiable information can be protected under a shared repository scenario.**

2) Data Sharing, Access and Use

The purpose of creating a centralized repository is to allow data to be accessible to researchers. At the same time, accessibility renders a centralized databank – already by its nature susceptible to hacking and theft – even more vulnerable to these problems. Even without personal identifying information in the repository, a security breach would be a disaster, with data sharers potentially selling genetic information to the highest bidder. Insurance companies, for example, might be interested in securing this information, which could lead to discrimination against individuals seeking health insurance, who have donated their DNA for the advancement of science.

A centralized, highly accessible repository also runs the risk of becoming an irresistible resource for law enforcement. The extraordinary expansion of forensic DNA databanks over the past decade is evidence of law enforcement's increasing interest in and reliance on genetic information in the tracking of suspects and the resolution of crime.[xi] The notion that the NIH repository is vulnerable to law enforcement access is not far-fetched; we have already seen proposals to co-mingle forensic and public health DNA banks or to create all-population databanks that can serve both ends[xii]. While NIH cannot be expected to curb law enforcement's broad appetite to access government-held data, the agency should impose standards for that access that make meaningful its stated commitment to protecting the confidentiality of research participants wherever possible.

The ACLU recommends that the NIH take every step possible to prevent unauthorized access and use of the data stored in the repository. Unlawful sharing of information or other abuses of information by any researcher should be severely sanctioned, and such researchers should be barred from future NIH funding. In addition, the agency should make it explicitly clear that under no circumstances will law enforcement be granted access to the data in the repository without first obtaining a court order based on a showing that probable cause exists to believe that a crime has been committed, and that relevant information may be obtained from the disclosure of the data.

The NIH proposal appears to make access to submitted data a requirement for NIH-sponsored researchers: "All investigators who receive NIH support to conduct genome-wide analysis of genetic variation in a study population are expected to submit to the GWAS data repository descriptive information about their studies for inclusion in an open access portion of the GWAS data repository."[xiii] But what happens if a researcher does not wish to provide data or feels that it is not consistent with the expectations of his or her research participants? Will that researcher be penalized in some way for

withholding information from the repository or otherwise compelled to turn over the data?

The proposal allows secondary users of the coded data access to the identities of research participants with "appropriate institutional approvals."[xiv] Sharing of identifying information between institutions allows a privacy loophole that the NIH should close if it is serious about protecting research participants. **The ACLU recommends that under no circumstances should the identities of individual participants be disclosed to secondary users without the further express affirmative consent of those individuals.**

The NIH proposal contains ambiguous guidelines for approving access to repository data that lack any enforcement or monitoring provisions. Under the proposal, investigators seeking access to data in the repository would be required to submit a "Data Use Certification" that includes a brief description of the proposed research use of the requested dataset(s).[xv] The only other requirement is that the investigators will stipulate that they will: "use the data only for the approved research use; protect data confidentiality; follow all applicable laws and any local institutional policies and procedures for handling GWAS data; not attempt to identify individual participants from whom data within a dataset were obtained; not sell or share any of the data elements from datasets obtained from the GWAS data repository with third parties; and provide annual progress reports on research."[xvi]

These requirements are vague. For example, what does it mean to "protect data confidentiality?" What specifically is to be protected? Can an investigator store the information on a laptop computer the investigator takes home? What kind of protections must s/he place on the datasets? What happens if s/he suspects that a third party has accessed the information? In addition, the NIH has not proposed any monitoring or oversight mechanism of this self-certification. How will the NIH know if these requirements are being met? How would a potential misuse of the data be recognized? What would be the repercussions to an investigator for breaching any of these requirements? **The ACLU recommends that the NIH establish clear guidelines for holding researchers accountable for protecting the privacy of research subjects and that the overly general provisions be clarified.**

Finally, the proposal states that access to the datasets will be approved by an "NIH Data Access Committee (DAC)."[xvii] It is impossible to evaluate whether these committees will be effective in protecting the interests of research participants or patients, since the makeup of these committees is not specified. Will these committees be comprised primarily of NIH staff and research scientists or will they also include ethicists and privacy experts? **The ACLU recommends that the role and powers of the DACs be clearly spelled out, and that each DAC include privacy advocates.**

3) <u>Informed Consent</u>

The NIH proposal fails to address the complex and highly important issue of informed consent in the context of a centralized data repository. The creation of a multi-use

genetic repository raises unique challenges to the traditional model of "informed consent" in medical research. For one, it has been noted that consent cannot be truly "informed" in regards to the creation of a multi-use data repository, since there would be no way to inform the individual providing the source material about all future uses of his or her material.[xviii] This predicament has been widely discussed as multi-use gene banks have proliferated over the last decade and around the world.[xix]

The NIH proposal largely avoids the question of whether informed consent is required as a condition of including an individual's data in the repository and how it can be obtained. At best, this responsibility appears to be left to the Internal Review Boards (IRBs), and, in part, investigators who must "submit certification by the responsible IRB that it has reviewed and approved submission to the NIH, noting any limitations on data use based on the relevant informed consents."[xx]

An absence of any discussion in the proposal about informed consent of research participants indicates that the authors of this proposal have in mind one of two alternative consent models for secondary uses of research participants' data: 1) "blanket consent" – where subjects are expected to consent to an open-ended and indefinite number of future uses of their data; or 2) "presumed consent" – where removal of identifiable data is considered sufficient for protecting privacy of individuals and no further consent for use of their data is necessary.

The ACLU is opposed to either of these methods of "consent."[xxi] Sensitive information, such as genetic information, should not be disclosed, made available, or used for purposes other than those explicitly specified in the original consent authorization for collecting the information, except with the further express affirmative consent of the individual. Individuals must be fully informed about how their genetic data may be broadcast and maintain the authority to decide how they want their information to be used or shared. As stated above, we are not convinced that the mechanisms proposed to ensure that only "de-identified" data is included in the repository are sufficient. So long as potentially identifiable sequenced data is contained in the repository, informed consent should be mandated for the release of that information for secondary use.

An additional problem is that the proposal is not clear as to whether ongoing studies or past studies would be "grandfathered in" to this repository. Inclusion in the repository of data collected from individuals in the past for the purpose of a particular study would likely violate – in most cases – the informed consent of those individuals and should not be allowed.

The ACLU recommends that the NIH directly address the fundamental challenge of maintaining informed consent for individual research participants in the establishment of a national repository of genetic and phenotypic information by requiring express affirmative consent. This crucial issue should not be cast aside or left to the discretion of individual IRBs. Public trust in genomic science will not be maintained unless individuals retain control over the use of their information and they are fully informed about how their information is being used.

4) Public Benefits

The stated goal of the NIH proposal is "to advance science for the benefit of the public through the creation of a centralized NIH GWAS data repository."[xxii] But while the benefits to researchers are fairly clear – e.g. rapid and broad access to genetic and phenotypic data; exclusive rights to publish an analysis of their own data for a limited period of time; open access to genotype/phenotype association data through discouragement of patenting – the benefits to the public are far less concrete.

What the proposal makes clear is that individuals who contribute their data to the repository cannot expect any direct benefit at all: "Research participants should not expect the return of individual research results derived from analyses of submitted data."[xxiii] The ACLU appreciates that participants could not be contacted without breaching the very confidentiality and anonymity that the NIH is trying to achieve in the repository. At the same time, individuals who have donated their tissues for research that results in the development of a promising or even life-saving therapy should have access to that therapy. This issue underscores the need for participants to be fully informed of the scope of research projects they are a part of; potential benefits are far more likely to be realized if they are in a position where they can monitor the availability of new drugs or clinical trials for which they may be suited.

The NIH proposal further states that: "Maximizing the availability of resources facilitates research and enables medical science to better address the health needs of people based on their individual genetic information."[xxiv] This implies that the ultimate goal of the research is to develop individualized therapies. But who will be able to afford such therapies? How does the agency reconcile the issue that select individuals are being asked to undergo privacy risks for the benefit of research that may not be widely accessible?

The long-term success of a national repository can only be realized if the public is confident that they will ultimately benefit from the research. **The ACLU recommends that the NIH be clearer as to the expected public benefits of the creation of the repository so that it does not risk discouraging public participation in biomedical research.**

5) Intellectual Property

The ACLU applauds the NIH for its efforts to ensure that basic genotype-phenotype associations identified through NIH-supported GWAS datasets remain available to all investigators, unencumbered by intellectual property claims. Associations of these types are basic scientific facts that should not be considered as patentable subject matter. **The ACLU recommends that, as a condition for access to the GWAS repository, investigators be required to certify to the NIH that patents will not be sought that would restrict the use of genotype/phenotype associations in a way that would**

substantially diminish the utilization of information and the potential public benefit they could provide.

The NIH has an opportunity to set a major precedent for privacy protection and to establish the gold standard for conducting biomedical research while safeguarding the rights of research subjects. The GWAS policy, as currently proposed, falls short of that standard. Privacy and informed consent are necessary prerequisites for public trust in the scientific enterprise. We hope that before issuing any final rules, the NIH will adopt the recommendations outlined above, and initiate a broad public discussion of the range of scientific, ethical, and civil liberties issues, followed by a fully transparent implementation process. The ACLU looks forward to working with the NIH and participating further in this process.

Sincerely,

Tania Simoncelli
Science Advisor
Technology & Liberty Project
American Civil Liberties Union
125 Broad Street, 18th Fl.
New York, NY 10004
tel: 212-519-7809
fax: 212-549-2629
email: tsimoncelli@aclu.org

Noam Biale
Technology & Liberty Project
American Civil Liberties Union
125 Broad Street, 18th Fl.
New York, NY 10004
tel: 212-549-2500 x4205
fax: 212-549-2629
email: tlp_nb@aclu.org

Caroline Fredrickson
Director, Washington Legislative Office
American Civil Liberties Union
915 15th St. NW
Washington DC 20005
tel: 202-675-2304
fax: 202-546-2934
email: cfredrickson@dcaclu.org

[i] Federal Register, Vol. 71, no. 168 (30 August 2006) p. 51629.
[ii] *Ibid*.
[iii] Federal Register, p. 51630.
[iv] Lin, Z., A.B. Owen and R.B. Altman, "Genomic Research and Human Subject Privacy," *Science*, Vol. 305, no. 5681 (9 July 2004), p. 183.
[v] *See* Cho, M.K. and P. Sankar, "Forensic Genetics and Ethical, Legal and Social Implications Beyond the Clinic, *Nature*, Vol. 36, no. 11 (November 2004), pp. S8-S12. *See also* "Retinome," DNA Witness, available at: http://dnaprint.humid.e-symposium.com/dnawitness/retinome.html
[vi] McGuire, A.L. and R.A. Gibbs, "No Longer De-Identified," *Science*, Vol. 312 (21 April 2006).
[vii] Federal Register, p. 51630.
[viii] *Ibid*.
[ix] *Ibid*.
[x] *See for example* Roche, P.A. and G.J. Annas, "DNA Testing, Banking, and Genetic Privacy," *NEJM*, Vol. 355, no. 6 (August 2006), pp. 545-546.
[xi] Rothstein, M.A. and M.K. Talbott, "The Expanding Use of DNA in Law Enforcement: What Role for Privacy?" *The Journal of Law, Medicine & Ethics*, Vol. 34, no. 2 (Summer 2006), pp. 153-164.
[xii] Simoncelli, T., "Dangerous Excursions: The Case Against Expanding Forensic DNA Databases to Innocent Persons," *The Journal of Law, Medicine & Ethics*, Vol. 34, no. 2 (Summer 2006), pp. 390-397.
[xiii] Federal Register, p. 51630.
[xiv] *Ibid*.
[xv] *Ibid*.
[xvi] Federal Register, p. 51631.
[xvii] Federal Register, p. 51630.
[xviii] Greely, H.T., "Breaking the Stalemate: A Prospective Regulatory Framework for Unforeseen Research Uses of Human Tissue Samples and Health Information," *Wake Forest Law Review*, Vol. 737 (1999).
[xix] *See* Clayton, E.W., K. K. Steinberg, M. J. Khoury, E. Thomson, L. Andrews, M. J. Kahn, L. M. Kopelman and J. O. Weiss, "Informed Consent for Genetic Research on Stored Tissue Samples," *JAMA*, Vol. 274, no. 22 (December 1995), pp. 1786-1792. *See also* Bereano, P.L., "The National Bioethics Advisory Commission Report on the Use of Human Biological Materials and Research: Ethical Issues and Policy Guidelines," *Biotechnology Law Report*, Vol. 18, no. 4 (August 1999), pp. 322-325. *See also* H.T. Greely, 1999.
[xx] Federal Register, p. 51631.
[xxi] For a discussion on these alternative methods of consent, *see* Trouet, C., "Informed Consent for the Research Use of Human Biological Materials," *Medicine and Law*, 2003. *See also* Gulcher, J.R. and K. Stefansson, "The Icelandic Healthcare Database and Informed Consent," *NEJM*, Vol. 342 (June 2000), pp. 1827-1830. *See also* Winickoff, D., "Governing Population Genomics: Law, Bioethics and Biopolitics in Three Case Studies," *Jurimetrics*, Vol. 43 (Winter 2003), pp. 187-228.
[xxii] Federal Register, p. 51631.
[xxiii] Federal Register, p. 51630.
[xxiv] Federal Register, p. 51631.