NHGRI 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 NHGRI-Supported or Conducted Medical Sequencing Studies (NHGRI MSP)

INTRODUCTION

This document closely mirrors the Points to Consider for NIH-funded Genome-Wide Association Studies (GWAS) available at http://grants.nih.gov/grants/gwas/gwas_ptc.pdf. This is appropriate for several reasons:

- Genome sequence data that will be produced under NHGRI MSP, and extensive genotype data produced in GWAS, raise nearly identical issues for informed consent and risk to participants.
- The proposed deposition into a controlled-access repository, data access terms, conditions, and access procedures for NHGRI MSP data will be essentially identical (with differences outlined below) to that used for GWAS.
- As with GWAS data, data from NHGRI MSP will be available via the National Center for Biotechnology Information (NCBI) in dbGaP with access controlled by an NIH Data Access Committee (DAC).
- In view of these similarities, the National Advisory Council on Human Genome Research recommended making NHGRI MSP policies as similar as possible to NIH GWAS policies.

This document is therefore nearly the same as the GWAS policy document, and where it is not, generally refers to that document as a point of comparison.

MSP DATA SUBMISSION CERTIFICATION

The NHGRI will accept MSP data into the NIH GWAS data repository after receiving appropriate certification by the responsible Institutional Official(s) of the submitting institution that they approve submission to the NHGRI MSP data repository. The certification should assure that:

The use of samples including data submission to the data repository is consistent with all applicable laws and regulations, as well as institutional policies;

The appropriate research uses of the data and the uses that are specifically excluded by the informed consent documents are delineated;

The identities of research participants will not be disclosed to the data repository; and

An IRB and/or Privacy Board, as applicable, reviewed and verified that:

-The submission of data to the data repository and subsequent sharing for research purposes are consistent with the informed consent of study participants from whom the data were obtained;
-The investigator's plan for de-identifying datasets is consistent with the standards outlined above;
-It has considered the risks to individuals, their families, and groups or populations associated with data submitted to the data repository;
-It has considered specific questions raised by the NHGRI staff, if any, and
-The genotype and phenotype data to be submitted were collected in a manner consistent with 45 C.F.R. Part 46

The purpose of this document is to assist Institutional Review Boards (IRBs) and/or, as appropriate, Privacy Boards in their review, and institutions in their certification, of investigator applications and proposals involving the submission of MSP data to the NIH under this policy.[1]

This information is being provided in two parts: Part I provides information about the:

a) policy;
b) benefits of broad sharing of GWAS data through a central data repository at NIH;
c) risks associated with the submission and subsequent sharing of such data; and,
d) safeguards that will be in place at NIH to protect the data.

Part II is intended to provide specific points to consider for institutions and IRBs in their review and certification of an investigator's plans for submission of data to the MSP data repository, including the adequacy of consent forms for data submission.

The NIH recognizes the complex and evolving nature of the ethical issues related to this policy and will issue additional guidance as may be needed at http://www.genome.gov/20019650.

.

[1] In the GWAS "Points to Consider" document, there is the following footnote: "The NIH recognizes that this review and certification process goes beyond regulatory requirements under 45 CFR part 46 as outlined in an August 2004 policy guidance of the Office for Human Research Protections entitled 'Guidance on Research Involving Coded Private Information and or Biological Specimens.' Following discussions with NIH staff, OHRP advised NIH that the GWAS repository does not currently involve human subjects research because the data being submitted will be collected solely for other research studies, and because the data will be coded and the identity of individuals from whom the data were obtained will not be readily ascertainable to the investigators maintaining the repository. This determination also means that IRB review and approval of the submission of GWAS data to dbGaP is not required under the regulations. Nonetheless, for the reasons outlined in this document, NIH, as a policy matter, will not accept data into the MSP repository without the appropriate certifications from the institution and verification by an IRB and/or Privacy Board that the submission criteria stipulated in the policy have been met." Based on this, NHGRI will follow similar policy regarding deposition of sequence data in the MSP database.

PART I: BACKGROUND INFORMATION

A. NHGRI Policy for Sharing of MSP Data

1. Data types to be shared

The MSP policy facilitates the sharing of large datasets containing coded[2], de-identified[3] genome sequence and phenotype data obtained in NHGRI supported or conducted research. The policy applies to data obtained prospectively as well as to studies using existing specimens and phenotype data. A key element of the policy is the expectation that data from NHGRI-funded MSP will be deposited into the MSP data repository, currently designated as the database of Genotypes and Phenotypes (dbGaP), at the National Center for Biotechnology Information (NCBI), a component of the National Library of Medicine (NLM). Note that dbGaP is able to accommodate genome sequence data as well, for example in the Trace Archive or the Short Read Archive. The data submitted for inclusion in the data repository will be coded and de-identified by the submitting investigator, but the investigator may retain the key to the code that would link to specific individuals. NCBI will never receive the code or any other information that would enable the identification of the individuals who are the source of the data.

The sequence data will likely be produced by the NHGRI large-scale sequencing centers (http://www.genome.gov/10001691). Data will be produced on a number of different sequencing platforms, which may use different formats for sequence data. However, in general, such sequence data within the repository will consist of a set of sequence "reads". Individual reads originating from an individual sample can be linked within the database, and in turn those can be linked to phenotype information from that individual. As with genotype data, sequence data in sufficient amounts are equivalent to genotype data in that they can distinguish the likely ethnic[4] origin of the sample, and comparisons of sufficient DNA sequence information from individuals can enable the recognition of family relationships. Identification of a specific individual through sequence data in the data repository will require comparison with a genome sequence from another identifiable DNA sample from the same person. It is anticipated that technological and analytical capacity available to the public is likely to enhance the feasibility of such identification in the future.

The phenotype data deposited in the NIH GWAS data repository may include information about disease status and characteristics that are not individually identifiable; however, some characteristics may be shared within families or common among population subgroups.

There are certain differences between genotype and sequence data that are important to mention here. Because of technical limitations, for the time being, all MSP projects so far will produce sequence from a limited number of loci from each sample. In contrast, GWAS data include hundreds of thousands to a million SNPs over the entire genome from each individual, which may actually provide more information that could serve as a basis to determine ethnic origin or family relationship than is possible with limited genome sequencing. As sequencing technologies improve, more and more sequence from each individual will be produced, diminishing this difference. NHGRI considers that larger contiguous, or linkable discontinuous, amounts of sequence from an individual carry a greater likelihood of being

---

[2] *Coded* means that any identifying information (such as name or social security number) that would enable the investigator to readily ascertain the identity of the individual to whom the private information or specimens pertain has been replaced with a number, letter, symbol, or combination thereof (i.e., the code); and a key to decipher the code exists, enabling linkage of the identifying information to the private information or specimens. From http://www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.htm

[3] *De-identified*, for purposes of this document, means that the identities of data subjects cannot be readily ascertained or otherwise associated with the data by the repository staff or secondary data users (45 CFR 46.102(f)), the 18 identifiers enumerated at section 164.514(b)(2) of the HIPAA Privacy Rule are removed and the submitting institution has no actual knowledge that the remaining information could be used alone or in combination with other information to identify the subject of the data.

[4] The original GWAS document used the term "ethnic origin". NHGRI believes that a more accurate term is "biogeographical population origin".

identifying as to family relationship, ethnic origin, or being matched to another sample sequenced from the same individual.

2. Essential role of Institutional Officials and IRBs and/or Privacy Boards in implementation of the policy

The nature of MSP data about participants and the broad data distribution goals of the MSP data repository highlight the importance of IRBs and institutions in reviewing plans for data submission, as well as the adequacy of the informed consent process and documents through which the data were obtained. Because the sequence and phenotype information generated about individuals will be substantial and, in some instances, sensitive (such as data related to the presence or risk of developing particular diseases or conditions and information regarding family relationships or ancestry), the confidentiality of the data and the privacy of participants must be protected.

In order to minimize risks to study participants, data submitted to the MSP data repository will be de-identified and coded using a random, unique code. Data should be de-identified according to the following criteria: the identities of data subjects cannot be readily ascertained or otherwise associated with the data by the repository staff or secondary data users (45 C.F.R. 46.102(f)); the 18 identifiers enumerated at section 45 C.F.R. 164.514(b)(2) (the HIPAA Privacy Rule) are removed; and the submitting institution has no actual knowledge that the remaining information could be used alone or in combination with other information to identify the subject of the data.

Institutional Officials of the submitting institution and IRBs and/or Privacy Boards play a key role in making sure that the submission of data to the MSP data repository is consistent with the NHGRI MSP policy. The NHGRI will only accept MSP data into the data repository after receiving appropriate certification by the responsible Institutional Official(s) of the submitting institution that they approve submission to the MSP data repository.

The certification should assure that:

- The use of samples including data submission to the data repository is consistent with all applicable laws and regulations[5], as well as institutional policies;
- The appropriate research uses of the data and the uses that are specifically excluded by the informed consent documents are delineated[6];
- The identities of research participants will not be disclosed to the data repository; and

- An IRB and/or Privacy Board, as applicable, reviewed and verified that:
    - The submission of data to the data repository and subsequent sharing for research purposes are consistent with the informed consent of study participants from whom the data were obtained;
    - The investigator's plan for de-identifying datasets is consistent with the standards outlined above;
    - It has considered the risks to individuals, their families, and groups or populations associated with data submitted to the data repository;
    - It has considered specific questions raised by NHGRI staff, if any, and

[5] Applicable federal regulations may include HHS human subjects protection regulations (45 CFR Part 46), FDA human subjects protection regulations (21 CFR Parts 50 and 56), and the Health Insurance Portability and Accountability Act Privacy Rule (45 CFR Part 160 and Part 164, Subparts A and E).

[6] Any limitations of the consent will be honored by NHGRI and carried through as the data are released to requesting investigators. For example, if an individual consent is for research only on a specific disease or condition, NIH will not release that data for research on another disease or condition. However, NHGRI will prefer studies that do not have disease-specific restrictions.

- The genotype and phenotype data to be submitted were collected in a manner consistent with 45 C.F.R. Part 46

B. Benefits of Broad Sharing of MSP Data through an NIH Central Data Repository

1. Nature of medical sequencing studies

The overall aim of medical sequencing studies is to discover genetic variants that contribute to the development, progression, or treatment options for a particular disease or trait (such as high blood pressure or obesity). There are multiple different types of MSP that can be envisioned, but each has in common the genomic sequencing (and sometimes transcriptomic sequencing) of multiple, sometimes thousands, of individuals. Usually, a comparison is made between individuals with (or at risk) for a disease, and controls. Examples include targeted gene sequencing of candidate genes or regions identified functionally or through association studies to identify all variations associated with a particular phenotype or to discover novel alleles of genes already known to be involved in disease; whole exome sequencing of disease cohorts to create a catalog of variants that can subsequently be used by an entire disease research community e.g. for following up on association studies; and eventually whole genome sequencing to identify all variants as a list of candidates for association with disease. When combined with clinical and other phenotypic data, analysis of genome sequence information offers the potential for increased understanding of basic biological processes affecting human health and improvement in the prediction of disease and treatment options.

2. Reasons for making data accessible to multiple investigators

The NHGRI is promoting and facilitating the sharing of data generated by MSP because the volume of data that will be generated in even one study is far greater than any individual or small group of collaborators can fully explore and because of the potential to gain important scientific knowledge and tools through the analysis of aggregated data. The MSP data repository enhances the capacity to make MSP data available to a wide range of scientific investigators in order to facilitate genetic research and enable research discoveries for the benefit of the public health.

Medical sequencing projects are most informative when the study population is large. The larger the population, the greater the statistical power to determine that observed associations are real and not due to chance. Although the costs associated with sequencing have been decreasing and are expected to continue to decline over time, the costs in terms of research resources (in terms of participant samples and funding) are high because of the large number of study samples required to produce high quality data. The very nature of medical sequencing projects allows the data to be used to address multiple research hypotheses. Given the resources involved and the potential for public benefit, it is prudent to create a database that facilitates the use of these data to address as many hypotheses as are ethically appropriate.

NHGRI will prefer to undertake MSP studies in which the consent does not limit the disease that can be studied using the data. The major reason for this is to encourage the use of the sequence data for development of research and informatics tools that will enable advances in the use and interpretation of sequence data for all MSP studies. In addition, it is very likely that multiple diseases may be influenced by variants in an overlapping set of genes. For example, the inflammation pathway, the lipid pathway, and the coagulation pathway each have been shown to be involved in more than a single disorder.

C. Risks Associated with Submission and Broad Sharing of MSP Data

The main concerns associated with submitting data to the MSP data repository are those entailed with other genetic research, i.e., those relating to participant privacy and confidentiality. Privacy and confidentiality concerns associated with wide data sharing through the MSP data repository stem from the

nature and magnitude of the sequence and phenotype data involved; the storage of those data in a central, Federal government repository; and the distribution of these data for secondary research.

Described below are risks associated with submission and broad sharing of MSP data. Section D describes measures to minimize such risks. As in the review of any research, it is important to consider any risks in the context of the protections in place to minimize those risks as well as in the context of the expected benefits of the proposed research.

Risks of Identification. The MSP database will NOT contain information that is typically used to identify individuals such as name, address, telephone number, birth date or social security number (see *De-identification of Data*, below). Although the data will be coded and the NIH will not hold direct identifiers to individuals whose data are included within the data repository, we recognize the personal and potentially sensitive nature of the genome sequence and phenotype data. Additionally, technologies available within the public domain today, and technological advances expected over the next few years, make the identification of specific individuals from sufficient amounts of raw sequence data feasible and increasingly straightforward. For example, someone might be able to compare information in the MSP database with sequence, genotype or phenotype information obtained from other, unrelated activities and be able to identify the individual who is the source of the data (or a blood relative of that individual). In a case in which data come from a discrete population (e.g., one small community), it could be more straightforward to cross classify individuals on several variables and make inferences about the source of a given sample.

In addition, discussions are occurring in the scientific community and among privacy experts about the uniqueness of individual genome-wide data and the possibility that in the future such data may by itself become identifiable. See NHGRI's Workshop on Privacy, Confidentiality and Identifiability in Genomic Research (http://www.genome.gov/19519197 and as further discussed in Science, 3 Aug. 2007, vol. 317, p. 600).

The NHGRI is committed to the protection of research participant privacy and the preservation of the confidentiality of individual-level data submitted to the MSP data repository. The NHGRI is, therefore, implementing a number of measures to protect the confidentiality and security of all data submitted to the data repository (see below). However, as in any system of protections, there are limitations to the protections afforded by these measures.

Risks Associated with Inadvertent or Inappropriate Use or Disclosure of Individually Identifiable Information. The NHGRI MSP data repository will not contain individually identifiable information (as defined in *De-identification of Data*, below) and, therefore, such data cannot be released to secondary users. However, the primary study may involve individually identifiable information. Submitting institutions should understand that potential harms to research participants or their family members can occur if individually identifiable information is inadvertently or inappropriately used or disclosed. These harms could include denial of employment or insurance of a research participant (or a relative). Other harms that may occur from inadvertent or inappropriate disclosure or use of individually identifiable information include psychosocial harms, such as stress, anxiety, stigmatization, or embarrassment resulting from inadvertent disclosure of information on family relationships, ethnic heritage, or potentially stigmatizing conditions.

Risks Associated with FOIA. The datasets submitted to NHGRI will be maintained in an NIH data repository and will, thereby, become U.S. government records that are subject to the Federal Freedom of Information Act (FOIA). As an agency of the Federal government, the NIH is required to release government records in response to requests under the Federal Freedom of Information Act (FOIA), unless the records are exempt from release under one of the FOIA exemptions. The NIH believes that release of individual-level genomic information in response to a FOIA request would constitute an

unreasonable invasion of personal privacy under FOIA Exemption 6, 5 U.S.C. § 552 (b)(6). Therefore, among the safeguards that the NIH foresees using to preserve the privacy of research participants and confidentiality of genomic data in NIH data repositories is the redaction of individual-level genomic data from any disclosures made in response to FOIA requests and the denial of requests for unredacted datasets. It is important to note, however, that FOIA affords requesters an opportunity to contest an agency's determination.

Risks Associated with Law Enforcement Access. The NHGRI will not possess direct identifiers within the MSP data repository, nor will the NHGRI have access to the link between the data keycode and the identifiable information that may reside with the primary investigators and institutions for particular studies. However, it is conceivable that law enforcement agencies could request access to the de-identified sequence and phenotype data within the MSP data repository and, for example, search for matches to DNA specimens collected for forensic purposes[7]. While expected to be rare, such requests may be fulfilled by the NIH. Law enforcement officials might then seek to compel disclosure of identifying information from the institution holding the identifying information. However, the release of identifiable information from the institution holding the identifying information may be protected from compelled disclosure if a Certificate of Confidentiality is or was obtained for the original study.

Risks to Specific Populations, Groups, and Communities. Medical research has already shown that some populations demonstrate a higher predisposition to develop certain medical diseases or disorders than others. MSP will provide insight into how certain genome sequence variants contribute to health and disease and will also increase knowledge of how such variants differ in frequency between and among populations. Genetic variants associated with physical disorders, diseases, and behavioral traits are expected to be found. Causative variants will be found in all populations with differing frequencies. Higher or lower frequencies that contribute to observed health patterns, particularly those that tend to be viewed negatively, can lead to genetic stereotypes that can stigmatize all members of a population group whether they possess a given genetic variant or not. In the absence of genetic non-discrimination laws, such information may also affect the insurability or employability of populations or groups. Persons sharing ethnic heritage may similarly be affected by results obtained from sharing of MSP data.

Return of Individual Research Results. For reasons explained later in this document, the return of individual research results to participants from MSP studies is expected to be a rare occurrence. Nevertheless, as in all research, the return of individual research results to participants must be carefully considered because the information can have a psychological impact (e.g., stress and anxiety) and implications for the participant's health and well-being. While clinically valid and meaningful results may have a positive impact on an individual's health, harms can occur if unvalidated research results are provided back to participants or used for medical decision-making. The ethical protections for MSP data that have been developed to address these and other issues are discussed in the next section.

[7] Law enforcement officials routinely obtain DNA specimens as part of their investigative work and collect DNA from convicted offenders. Every state has established a DNA database, and these databases are linked through the Federal Combined DNA Index System (CODIS) program.

D. Protections for MSP Data

The NHGRI acknowledges that the practical and ethical questions relevant to the NHGRI MSP Policy are the subject of considerable discussion in the research community. The NHGRI remains committed to participating in the on-going dialog on these topics and to addressing the evolving scientific, ethical and societal issues within the policy and practices as appropriate. NHGRI intends to accomplish this primarily by reflecting the NIH-wide GWAS policy, but also by soliciting ongoing advice form the National Council on Human Genome Research and other advisors specific to the handling of sequence data.

Operating Policies. As NIH-wide policies fro genomic data, especially GWAS policies, are developed, they will be incorporated as appropriate into the NHGRI MSP policies. As the GWAS policy states, NIH is establishing policies and procedures for the NIH GWAS data repository that address, among other matters, the privacy of GWAS research participants and confidentiality of their data, the interests of participants, families and groups, data access procedures, and data security mechanisms. They will be reviewed periodically and updated as necessary by several GWAS oversight bodies discussed in the GWAS policy itself: http://grants.nih.gov/grants/gwas/gwas_ptc.pdf.

De-identification of Data. Before data are submitted to the MSP repository, submitting investigators will be expected to de-identify the data according to the following criteria: 1) the identities of data subjects cannot be readily ascertained or otherwise associated with the data by the repository staff or secondary data users (45 CFR § 46.102(f)); and 2) the following identifiers enumerated at section 164.514(b) (2) of the HIPAA Privacy Rule are removed:

1. Names.

2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP Code, and their equivalent geographical codes, except for the initial three digits of a ZIP Code if, according to the current publicly available data from the Bureau of the Census: a. The geographic unit formed by combining all ZIP Codes with the same three initial digits contains more than 20,000 people. b. The initial three digits of a ZIP Code for all such geographic units containing 20,000 or fewer people are changed to 000.

3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.

4. Telephone numbers.

5. Facsimile numbers.

6. Electronic mail addresses.

7. Social security numbers.

8. Medical record numbers.

9. Health plan beneficiary numbers.

10. Account numbers.

11. Certificate/license numbers.

12. Vehicle identifiers and serial numbers, including license plate numbers.

13. Device identifiers and serial numbers.

14. Web universal resource locators (URLs).

15. Internet protocol (IP) address numbers.

16. Biometric identifiers, including fingerprints and voiceprints.

17. Full-face photographic images and any comparable images.

18. Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification

In addition, the submitting institution should have no actual knowledge that the remaining information could be used alone or in combination with other information to identify the individuals who are the subject of the information. In reviewing data submission plans, the relevant IRB and/or Privacy Board should consider the extent to which the sequence and other phenotype information associated with the participants could be used to identify an individual or his or her family members by matching the sequence/phenotype datasets to other sources of information. The IRB and/or Privacy Board should also consider that genotype data may be available for the research participants from other studies where data are in the NIH GWAS repository.

Coding of Data. Before data are submitted to the MSP data repository, submitting investigators will be expected to assign a random, unique code to the data to protect participant privacy and confidentiality. As a further protection, submission of MSP data must be accompanied by a written certification by the submitting institution stating that the identities of research participants will not be disclosed to the MSP data repository.

Certificates of Confidentiality. Prior to submitting data to the MSP data repository, investigators and their IRBs may want to determine whether a Certificate of Confidentiality has been obtained for their research or, if one has not been obtained, to consider whether or not it would be appropriate to do so. Certificates of Confidentiality may provide an additional safeguard with regard to compelled disclosure in any civil, criminal, administrative, legislative, or other proceeding, whether at the federal, state, or local level, of information that could be used to identify individual research participants. Certificates of Confidentiality are issued to help achieve research objectives and promote participation in research. They can be granted for studies collecting genetic and other information that, if disclosed, could have adverse consequences for participants or damage their financial standing, employability, insurability, or reputation. Further information on when Certificates of Confidentiality may be appropriate and application instructions, can be obtained at the NIH Certificate of Confidentiality kiosk: http://grants2.nih.gov/grants/policy/coc/

MSP Data Repository Security Measures. To secure the data, the MSP data repository will include multiple tiers of data security such as sequential firewalls, independent networks, and encryption based on the content and level of risk associated with the data. All data and information will be submitted to a high security network within NIH through a secure transmission process. Details on security measures can be found on the NCBI website, http://www.ncbi.nlm.nih.gov.

Controlled Access to Individual Data. Access to individual-level sequence and phenotype data will be tightly controlled (with the exception discussed immediately below). Individual genotype and phenotype data will only be available for research through a controlled access procedure. Only basic descriptive information about each MSP study, such as the measures that it used, and the composition of the study population will be publicly available. Selected aggregate statistical calculations[8] will also be made publicly available.

Public release of some fragmentary sequence data. NHGRI has determined that some types of fragmentary sequence data bear minimal risk of identifyability to individual participants: specifically, small fragments of sequence data that cannot be linked within the MSP database to other small fragments originating from a single individual. This is in explicit contrast to GWAS data, where releasing hundreds of thousands of SNPs from a single individual would carry some risk. Sequence fragments may be useful to the community of researchers developing informatics tools for handling medical sequencing data. In addition, they will provide an ability to understand population frequencies of sequence variants. Therefore, NHGRI has, together with its advisors, decided that if possible, fragmentary MSP sequence data could be publicly released. In no case will NHGRI publicly release sequence fragments over 1 Mb in length, the fragment size that NHGRI advisors have said bears, in most cases, a negligible risk of being identifiable[9]. In practice, the MSP will publicly release fragments that are about 500 to 800 bp in length. It is important to note that this policy was established for the ABI 3730 sequencing platform, which produces reads in the range of 500-800 bp. Newer sequencing platforms produce many more, shorter read lengths from a single sample. At the current time, due to computer storage costs and the consequent architecture of the NCBI repository for "short read" data, there is no cost effective way to make fragmentary read data from these platforms public without information attached to each read that links the reads together as originating from the same machine "run" (and therefore, sample). Thus, for projects using newer sequencing platforms, all sequence data will only be available via the Controlled Access Repository.

Assuring Appropriate Data Use. Researchers eligible for access to individual-level data include, but are not limited to, qualified investigators from academic institutions and commercial organizations, both domestic and foreign. Researchers will have to apply for access to data included in the MSP data repository through the submission of a Data Access Request that will include a brief description of the proposed research use. Requests will be approved by a researcher's home institution and then routed to an NIH Data Access Committee (DAC). In general, the DAC will be organized by NHGRI. However, NHGRI recognizes that many sample sets that will be the subject of MSP studies will have been part of studies funded by other NIH institutes. In these cases, NHGRI will attempt to include a staff member of that institute in the DAC decision-making process, if that other institute is interested. A DAC consists of Federal staff with expertise in relevant scientific disciplines and ethical issues related to protecting the privacy of research participants and the confidentiality of their data. Outside experts may be consulted as necessary. DACs review requests for access to determine that the proposed use of a dataset is scientifically and ethically appropriate and does not conflict with any constraints or informed consent limitations identified by the submitting institution. If a data request raises concerns related to privacy and confidentiality, risks to populations or groups, or other concerns, the relevant DAC may consult with other experts as appropriate. Only after approval by the relevant DAC will data be available for download in a secure and encrypted format by a recipient investigator.

[8] The particular considerations for a given dataset may vary by project.

[9] The ability to match two samples varies with location in the genome, and varies from individual to individual. It is therefore not possible to provide an absolute probability of matching any two samples.. Other NIH policies dealing with release of genotype data have determined that the threshold for release of genotype data should be 60 SNPs. Sequence data released in contiguous segments is less informative because it is from a single location in the genome. However, in rare cases, even a single SNP can be identifying if it is rare enough. For the time being, variations this rare are largely beneath the practical level of detection of sequencing technologies used at high throughput.

Investigators and institutions seeking data from the NHGRI MSP data repository will submit to the NIH a Data Access Request along with a Data Use Certification that will stipulate a number of protections for research participants. Both the Data Access Request and the Data Use Certification must be co-signed by the investigator and by the appropriate designated Institutional Official to document their joint agreement to follow NHGRI policy for the use of MSP data obtained from the data repository. The Data Use Certification will stipulate that, subject to applicable law, the investigator and institution will:

- Use the data only for the approved research;
- Protect data confidentiality;
- Follow appropriate data security protections;
- Follow all applicable laws, regulations and local institutional policies and procedures for handling MSP data;
- Not attempt to identify individual participants from whom data within a dataset were obtained;
- Not sell any of the data elements from datasets obtained from the data repository;
- Not share with individuals other than those listed in the request any of the data elements from data sets obtained from the data repository;
- Agree to the listing of a summary of approved research uses within the data repository along with his or her name and organizational affiliation;
- Agree to report violations of the MSP policy to the appropriate DAC;
- Acknowledge the MSP policy with regard to publication and intellectual property; and
- Provide annual progress reports on research using the GWAS dataset.

The recipient investigator will be expected to protect the data by following best practices for data security posted on the NIH GWAS data repository website at http://www.ncbi.nlm.nih.gov/projects/gap/pdf/dbgap_2b_security_procedures.pdf, or other dataset-specific recommendations as detailed for a given MSP dataset within the repository. In addition, progress reports will be reviewed by the relevant DAC to verify continued appropriate use of the data.

Alternative methods for data access. All MSP data will be made available via the MSP data repository. However, NHGRI will consider requests by other NIH institutes for alternative means of managing access to the MSP repository. Specifically, if another NIH institute can guarantee access on an equal basis to all requestors, using the same criteria for all applicants, can show that it has the infrastructure to manage requests for access to the data (a DAC), and has procedures and safeguards that are at least as stringent as those outlined above, is able to report usage statistics and potential violations of policy to NHGRI, and the institute is interested in managing access to the data, NHGRI will consider passing responsibility for access to the institute. This alternative will be decided on a case-by-case basis, and is included to encourage use of samples that have a considerable prior investment by other NIH institutes, and where conditions for use of the data may be substantially different from those outlined by NHGRI's policies.

Withdrawal of Consent. The data repository has developed policies with regard to removal of individual data records if consent is withdrawn. Submitting investigators and their institutions may request removal of data on individual participants from the data repository in the event that a research participant withdraws consent. Such data sets will be removed from the repository records at the time of the next repository update. However, data that have already been distributed for approved research use will not be able to be retrieved.

Return of Research Results. The NHGRI anticipates that MSP will generate an unprecedented number of associations between particular genetic variants and diseases, or conditions or treatments. These associations constitute one step in a multistep process between uncovering the mechanism of action of a genetic locus and developing therapies or diagnostics that can be used in patient care. Initial findings will need to be confirmed and validated by further research before their potential clinical significance is

understood. In addition, many technical and statistical challenges in this area of research must be overcome in order to avoid false positive or false negative results, and to establish clinically meaningful relationships between particular variants and disease. In these cases, the argument for returning such uncertain results is not strong.

However, in rare cases, MSP sequence data will reveal sequence variants in individual samples that are already known to cause or strongly contribute to disease. As our knowledge of disease-causing variants grows, and as more sequence can be produced efficiently from an individual, these cases will grow more common. In these cases, there is a strong argument that results should be returned to research participants. There are two important considerations:

- As in any research, harms may result if individual research findings that have not been clinically validated are returned to subjects or are used for clinical decision-making prematurely. NHGRI sequencing centers are not CLIAA-approved laboratories and so their results should not be used to make clinical decisions.
- Neither the NHGRI sequencing centers, nor any secondary investigators that have access to the data in the MSP repository are in a position to return results, because they can have no links between the data and the identities of the original research participants, and because they have no direct research or informed consent relationship with the participants.

If a secondary investigator does generate results of immediate clinical significance, he or she can only facilitate their return by contacting the contributing investigator who holds the key to the code that identifies the participants. In such cases, the contributing investigator would be expected to comply with all applicable laws and regulations and consider the benefits and risks associated with the return of individual research results to participants and follow established institutional procedures (e.g., consultation with and approval by the IRB) to determine whether return of the results is appropriate and, if so, how it should be accomplished.

If they have not already done so, contributing institutions and their IRBs may wish to establish policies for determining when it is appropriate to return individual findings from research studies.

Oversight of GWAS Activities. The NHGRI will establish policies for oversight of the MSP data repository and for monitoring MSP data use practices. They include an annual review process for MSP activities that will include monitoring of:

- Information about Data Access Requests (number received, number granted, etc.)
- Reports of policy violations, if any
- Review of data access policies with regard to both protection of research subjects and appropriate research community access

Oversight will be conducted by NHGRI staff members together with a board of outside advisors including experts in, for example, human subjects protections, database security, and human genetics.

PART II: DATA SHARING PLANS, INSTITUTIONAL CERTIFICATION, AND POINTS TO CONSIDER REGARDING INFORMED CONSENT

A. Data Sharing Plans

Many individual projects that are part of MSP will be solicited directly from the community by NHGRI. Some, but not all, solicitations will request an individual proposal. Such proposals will be expected to include a data sharing plan as part, or to provide an appropriate explanation as to why submission to the repository is not possible (see *Exemptions from Data Release Requirement* http://www.genome.gov/Pages/Research/SequenceMapsBAC/MedicalSequencing/MSPExemptionsfromDataReleaseRequirement.pdf). Data sharing plans are expected to describe how the expectations of the policy will be met, including the consistency of the informed consent for submission to the MSP data repository and subsequent sharing, how informed consent will be obtained (for prospectively collected samples and data), and how data will be subsequently de-identified in accord with the specific criteria for data submission. IRBs should be cognizant of the MSP data sharing plans at the time of IRB review of the application in order to assess their appropriateness for a specific dataset and to provide the relevant analysis called for within the policy under the institutional certification expectations.

B. Institutional Certification

Institutions submitting data to the MSP data repository are responsible for certifying that data submission plans meet the following expectations defined in the MSP policy:

- The use of samples including data submission to the data repository is consistent with all applicable laws and regulations[10], as well as institutional policies;
- The appropriate research uses of the data and the uses that are specifically excluded by the informed consent documents are delineated[11];
- The identities of research participants will not be disclosed to the data repository; and

- An IRB and/or Privacy Board, as applicable, reviewed and verified that:
    - The submission of data to the data repository and subsequent sharing for research purposes are consistent with the informed consent of study participants from whom the data were obtained;
    - The investigator's plan for de-identifying datasets is consistent with the standards outlined above;
    - It has considered the risks to individuals, their families, and groups or populations associated with data submitted to the data repository;
    - It has considered specific questions raised by NHGRI staff, if any, and
    - The genotype and phenotype data to be submitted were collected in a manner consistent with 45 C.F.R. Part 46

C. Points to Consider Regarding Informed Consent

[10] Applicable Federal regulations may include HHS human subjects protection regulations (45 CFR Part 46), FDA human subjects protection regulations (21 CFR Parts 50 and 56), and the Health Insurance Portability and Accountability Act Privacy Rule (45 CFR Part 160 and Part 164, Subparts A and E).

[11] Any limitations of the consent will be honored by NHGRI and carried through as the data are released to requesting investigators. For example, if an individual consent is for research only on a specific disease or condition, NIH will not release that data for research on another disease or condition. However, NHGRI will prefer studies that do not have disease-specific restrictions.

The NHGRI recognizes that the issues related to determining the appropriateness of informed consent for submission of data to the MSP data repository and subsequent sharing for research are quite complex. The MSP policy applies to genome sequence data utilizing samples and phenotype data collected both prospectively and retrospectively and the applicable considerations regarding informed consent may vary depending upon which type of study is being proposed.

Prospective Studies. For prospective studies, in which sequencing is included within the study design at the time research participants provide their consent, the consent form and process must comply with the requirements of 45 C.F.R. Part 46 and any other applicable law. From an ethical standpoint, the informed consent process and document should make it clear that participants' DNA will undergo genomic analysis and that sequence and phenotype data will be shared for research purposes through a controlled-access data repository, available to biomedical researchers on the Internet. The consent should also discuss risks of sharing genomic data. See http://www.genome.gov/Pages/Research/SequenceMapsBAC/MedicalSequencing/MSPModelLanguageforConsent.pdf for model consent language.

Retrospective Studies. For retrospective studies performed using existing genetic materials and previously collected data, the NIH anticipates considerable variation in the extent to which future genetic research and data sharing have been addressed within the informed consent documents. In all such cases, IRBs are expected to determine whether the initial consent under which existing genetic materials and data were obtained is consistent with the submission of data to the MSP repository and the sharing of that data in accord with the MSP policy.

The NHGRI anticipates that for studies that propose to use pre-existing data or samples, IRBs may conclude in some cases that the original consent is not adequate for submission to the MSP data repository and subsequent sharing for research. In these cases, the IRB may decide that it is appropriate and necessary for the investigator to seek explicit consent of the research participants for submission to the MSP repository and subsequent sharing. Programmatic consideration to requests from investigators for funding to support efforts to seek re-consent from participants will be provided on a case-by-case basis.

It should be noted that the criteria for a waiver of consent under 45 CFR part 46 are inapplicable to such IRB considerations since the MSP database does not currently involve human subjects research. The criteria that are expected to be applied in making the determination that submission is consistent with the consent are set forth in the MSP policy and explained in this document.

The IRB also may determine that re-consent is not feasible or appropriate for a given study. Moreover, the IRB may determine that it cannot verify that the other criteria described in the policy[12] have been met for submission to the MSP repository. In all these cases, the researcher's data sharing plan should explain the IRB's determination that submission to the MSP repository is not appropriate. NHGRI will consider these issues on a case-by-case basis when making programmatic decisions to proceed with MSP studies for which the submission criteria cannot be met.

[12] As outlined elsewhere in this document, in addition to verifying that submission to the GWAS repository and subsequent sharing for research purposes is consistent with the informed consent of study participants from whom the data were obtained, the IRB is also expected to verify that:

- The investigator's plan for de-identifying datasets is consistent with the standards outlined in the policy;
- It has considered the risks to individuals, their families, and groups or populations associated with data submitted to the NIH GWAS data repository;
- It has considered any other issues raised by NHGRI staff; and
- The genotype and phenotype data to be submitted were collected in a manner consistent with 45 C.F.R. Part 46.

The following points to consider may be helpful to IRBs in determining the consistency of existing consents with the MSP data sharing policy, as well as to investigators in preparing new consent documents for this purpose. They are not intended to be proscriptive, nor are they all of the issues that may be appropriate for IRBs to consider in specific scenarios. Each research project and consent document is unique and local IRBs are in the best position to evaluate the potential benefits and risks of data submission and the consistency of consent with submission to the MSP data repository.

Scope of Written Consent.
Is the informed consent consistent with the anticipated research activities under the MSP policy? For instance:

- Does the consent form either allow or preclude:
  - genetic research or analysis?
  - future use and broad sharing of the participant's coded phenotype and genotype data for research?
  - submission of the participant's coded phenotype and genotype data to a government health research database for broad sharing to qualified investigators?

- Does the consent form have any restrictions, such as:
  - types of subsequent research using the participant's phenotype and genotype data?
  - location of such research?
  - types of medical conditions or diseases studied?
  - duration of storage and use of phenotype and genotype data?
  - limitations on who can use the participant's phenotype and genotype data (e.g. some consents may state that only non-commercial researchers can use the data)?

For studies that are found to be acceptable for submission to the MSP data repository, the certification provided to the NIH should delineate the appropriate research uses of the data and any uses that are specifically excluded by the informed consent documents. NHGRI will prefer studies that allow sharing of data for any legitimate biomedical use, that is, without restriction to a specific disease. However, NHGRI can accommodate such restrictions, especially where sample sets are uniquely valuable.

Potential Benefits
Does the consent form discuss that potential benefits may accrue broadly to the public through the advancement of science and understanding of health and disease, rather than resulting in direct benefits to individuals?

Risks
Does the consent form discuss risks associated with genetic or genomic research? Are these risks consistent with the risks involved in MSP activities? For example:

- Does the consent form discuss risks of broad sharing of phenotype and sequence or other genomic data?
- Does the consent form discuss privacy risks of data sharing (e.g., the possibility that the coded data may be released to members of the public, insurers, employers, and law enforcement agencies)?
- Does the consent form discuss the risks of computer security breaches relevant to maintaining data in an electronic format?
- Does the consent form discuss relevant risks to relatives or identifiable populations or groups?

Return of Research Results
Does the consent form include a discussion of whether or not research results will be returned to subjects, and under what conditions? Are those representations consistent with the MSP policy that research results may only be returned in rare instances following established procedures at the contributing institutions?

Privacy and Confidentiality Protections
Does the consent form address how individual privacy and data confidentiality will be protected? Is the manner in which privacy and confidentiality measures are described consistent with the MSP policies?

Withdrawal of Consent
Does the consent form address whether a subject can withdraw his/her phenotype and genotype data from research use? Is this language consistent with MSP policies?

Commercial Use
Does the consent form allow for or preclude commercial use of the subject's phenotypic and genotypic data? If specific restrictions are specified, they should be included within the institutional certification to the NHGRI. NHGRI will prefer samples that do not have restrictions on commercial use.

Other
Is there any other information in the consent form that is inconsistent with the information provided about the NHGRI MSP data repository and the MSP policies and procedures?

Does the study involve children? If so, has the IRB considered the appropriateness of the continued maintenance and sharing of the data when the child reaches the legal age of consent?

Does the study involve proxy consent? If so, are there any special ethical issues that should be considered?

Does the study involve vulnerable populations, and if so, have any special ethical concerns related to the study population been addressed?

Have any special cultural considerations or requirements been addressed with regard to the study population (e.g., the need for tribal consent from Native American populations)?

Are any issues of group harm relevant and have they been considered?