RI-INBRE Bioinformatics Core

Data type: human genomic data and non-human genomic data

The GDS Policy applies to all NIH-funded research that generates large-scale human or non-human genomic data as well as the use of these data for subsequent research. Large-scale data include genome-wide association studies (GWAS), single nucleotide polymorphisms (SNP) arrays, and genome sequence, transcriptomic, metagenomic, epigenomic, and gene expression data, irrespective of funding level and funding mechanisms (e.g. grant, contract, cooperative agreement, or intramural support). NIH Institute or Centers (IC) may expect submission of data from smaller scale research projects based on the state of the science, the programmatic priorities of the IC funding the research, and the utility of the data for the research community.

Data generated by the RI-INBRE Core Facilities or by individual INBRE investigators will be shared in full compliance with NIH protocols for sharing of large-scale human and non-human genomic data, including genome-wide association studies, metagenomic sequencing, functional genomics data (RNA-seq/microarray, ChIP-seq/epigenomics, proteomics, etc). RI-INBRE will also require users and investigators to follow established public standards for sharing of small-scale bioinformatic data. This will include deposition of sequences into public databases, collection of detailed metadata for samples, and release of custom code or pipelines used for bioinformatics analysis.

The data expected to be generated or analyzed by INBRE researchers will include but is not limited to nucleotide sequences (including genomic and/or metagenomic data), protein sequences, functional genomics data (RNA-seq, ChIP-seq, etc.) and protein structure data. The Bioinformatics Core will dissemninate data sharing protocols to all researchers to ensure compliance with NIH, INBRE and journal requirements.

It is anticipated that most data will be from non-human organisms, in particularly mouse and microbes. If data is generated or analyzed from human sources, the NIH Genomic Data Sharing Policy for human sequence data will be followed. This will include deposition of experimental/clinical data to the proper repositories and anonymization of the data to comply with the HHS Regulations for the Protection of Human Subjects and the Health
Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

In addition to standard bioinformatics data, the Bioinformatics Core is also compiling a student tracking database to allow for rapid and accurate assessment of outcomes for students working on INBRE projects. This database will be compliant with the The Family Educational Rights and Privacy Act (FERPA) to protect student privacy. The full database limits the type of personal data collected on the students (no SSN, BoD, address, etc.) and will only be available fully to RI-INBRE administrative staff and selectively to RI-INBRE researchers. If the data is shared with the public or other INBRE states, the data will be anonymized to remove identifying information about the students, including removal of non-critical meta data that could be used to infer a student's identity.

Data repositories

Identify the data repositories to which the data will be submitted, and for human data, whether the data will be available through unrestricted or controlled-access.

Investigators should register all studies with human genomic data that fall within the scope of the GDS Policy in dbGaP by the time that data cleaning and quality control measures begin. After registration in dbGaP, investigators should submit the data to the relevant NIH-designated data repository (e.g., dbGaP, GEO, SRA, the Cancer Genomics Hub). NIH-designated data respositiories need not be the exclusive source for facilitating the sharing of genomic data, that is, investigators may also elect to submit data to a non-NIH-designated data respository in addition to an NIH-designated data respository. However, investigators should ensure that appropriate data security measures are in place, and that confidentiality, privacy, and data use measures are consistent with the GDS Policy.

Non-human data may be made available through any widely used data repository, whether NIH- funded or not, such as GEO, SRA, Trace Archive, Array Express, Mouse Genome Informatics, WormBase, the Zebrafish Model Organism Database, GenBank, European Nucleotide Archive, or DNA Data Bank of Japan.

Non-Human Sequence Data

All nucleotide sequence data generated by the Bioinformatics Core, analyzed by the Bioinformatics Core, or generated by RI-INBRE researchers will be deposited in the appropriate NIH-compliant databases. Small-scale (e.g. genetic) sequencing or assembled sequences will be deposited in GenBank. Raw sequence data from large-scale projects will be deposited in the Sequence Read Archive. Protein sequence data will be deposited in the Protein database and protein structure data will be deposited in the RCSB protein data bank. Experimental data for functional genomics analyses including transcriptomics (microarray/RNA-seq) and epigenomics (ChIP-seq) will be deposited in the Gene Expression Omnibus database. Researchers may also deposit their data in additional repositories. In order to promote reproducibility of research, all code generated by the Bioinformatic Core or RI-INBRE researchers for analysis of data will be open source and made available to the public through standard sources (e.g. GitHub, SourceForge, etc.) within nine months of validation.

Human Sequence Data

Any sequence data generated from humans will comply with the above requirements. In addition, any data regarding individual human subjects (GWAS, SNP data, precision medicine, etc.) must be anonymized prior to deposition to ensure patient privacy. Researchers should consider the eighteen data elements defined by the Health Insurance Portability and Accountability Act of 1996 (HIPAA) in such cases. Data should be

deposited in a controlled access database such as the database of Genotypes and Phenotypes (dbGaP) within 45 days of data generation.

Data submission expectations and timeline

Investigators should submit large-scale genomic data as well as relevant associated data (e.g. phenotype and exposure data) to an NIH-designated data repository in a timely manner. Investigators should also submit any information necessary to interpret the submitted genomic data, such as study protocols, data instruments and survey tools. Genomic data undergo different levels of data processing, which provides the basis for NIH's expectations for data submission and timelines for the release of the data for access by investigators. These expectations and timelines are provided in the Supplemental Information. In general, NIH will release data submitted to NIH-designated data repositories no later than six months after the initial data submissions begins, or at the time of acceptance of the first publication, whichever occurs first, without restrictions on publication or other dissemination.

RI-INBRE will require all researchers publishing data using RI-INBRE resources to submit their data as described above prior to publication regardless of the individual data sharing policies of the journals publishing the data. Data should be deposited no later than six months after data submission is initiated.

The Bioinformatics Core and RI-INBRE researchers will also comply with the MIxS (Minimum Information about any (x) Sequence) standards implemented by the Genome Standards Consortium. These protocols include the MIGS (genomes), MIMS (metagenomes) and MIMARKS (marker genes) standards and govern the quality of the sequencing data, methods used to generate and/or assemble the data, and collection of relevant metadata (e.g. environment, biogeographical data, etc.).

Informed consent and institutional certification

Respect for, and protection of the interests of, research participants are fundamental to NIH's stewardship of human genomic data. The informed consent under which the data or samples were collected is the basis for the submitting institution to determine the appropriateness of data submission to NIH-designated data repositories, and whether the data should be available through unrestricted or controlled access.

For research that falls within the scope of the GDS Policy, submitting institutions, through their Institutional Review Boards (IRB's), privacy boards, or equivalent bodies, are to review the informed consent materials to determine whether it is appropriate for data to be shared for secondary research use. Specific considerations may vary with the type of study and whether the data are obtained through prospective or retrospective data collections. NIH provides additional information on issues related to the respect for research participant interests its "*Points to Consider for IRB"s and Institutions in their Review of Data Submission Plans for Institutional Certifications*" (updated in 2016 to "*Points to Consider for Institutions and Institutional Review Boards in Submission and Secondary Use of Human Genomic Data under the National Institutes of Health Genomic Data Sharing Policy").*

The Bioinformatics Core and RI-INBRE researchers will comply with all standards of informed consent for data generated for human subjects as outlined by the NIH and the RI-INBRE institutions and their institutional review boards (IRBs) or equivalent bodies.

Exceptions to data submission expectations

In cases where data submission to an NIH-designated data repository is not appropriate, that is, the Institutional Certification criteria cannot be met, investigators should provide a justification for any data submission exceptions requested in the funding application or proposal. The funding IC may grant an exception to submitting relevant data to NIH, and the investigator would be expected to develop an alternate plan to share data through other mechanisms.

Question not answered.

Intellectual Property

NIH encourages patenting of technology suitable for subsequent private investment that may lead to the development of products that address public needs without impeding research. However, it is important to note that naturally occurring DNA sequences are not patentable in the U.S. Therefore, basic sequence data and certain related information (e.g. genotypes, haplotypes, *p*-values, allele frequencies) are pre-competitive. Such data made available through NIH-designated data repositories, and all conclusions derived directly from them, should remain freely available, without any licensing requirements.

NIH encourages broad use of NIH-funded genomic data that is consistent with a responsible approach to management of intellectual property derived from downstream discoveries, as outlined in the NIH *Best Practices for the Licensing of Genomic Inventions* and Section 8.2.3. Sharing Research Resources, of the NIH Grants Policy Statement. NIH discourages the use of patents to prevent the use of or to block access to genomic or genotype-phenotype data developed with NIH support.

Question not answered.