NIAID Data and Reagent Sharing and Release Guidelines

Version 12 September 22, 2009

The NIAID has made a significant investment in genomic-related activities that provide comprehensive genomic, functional genomics, bioinformatics, structural genomics, proteomics and integrated "omics" data sets, resources and reagents to the scientific community for basic and applied research in infectious diseases. This wealth of genomics and other data sets, as well as the availability of the human genome, provides a valuable and critical resource for the scientific community. The functional genomic analysis of DNA sequences from microbial pathogens is enhancing the understanding of a pathogen's biology and its ability to cause disease; these efforts may lead to new strategies for diagnostics, prevention and treatment. The knowledge of the human genome sequence is enhancing the understanding of the host immune response and an individual's genetic susceptibility to microbial pathogens; these efforts may provide insights regarding how an individual may respond to drugs, treatments and vaccines.

This document serves to provide general guiding principles and specific guidelines to prepare and establish consistent data release plans across NIAID/DMID Omics centers including Genome Sequencing Centers for Infectious Diseases (GSCID). NIAID acknowledges the variety of projects among the Centers, but at the same time, considers it of the highest importance to develop guidelines that are flexible enough to achieve rapid data release and to be sensitive to the aims of the Centers and their individual projects. Continued discussions of the data release guidelines will be an ongoing function of the Centers, Scientific Working Groups, scientific community and the NIAID to achieve the overall goals of the NIAID Omics Centers. Modifications to NIAID Data Release Guidelines and data release plans may be needed as a result of these discussions and other data release guidelines developed by NIH including those for NIH Human Microbiome Project.

General Data Release Guiding Principles and Guidelines

Rapid and unrestricted sharing of data and research resources is essential for advancing research on human health and infectious diseases. The utility of the generated data to the scientific community is largely dependent on how quickly these data can be deposited into public databases. NIAID is committed to rapid, pre-publication release of genomic and other data types and in addition, recognizes that clinical data and other metadata associated with the genomic data are valuable research resources. For these reasons, NIAID endorses rapid release of all these data sets. The users of any released data are expected to act responsibly to recognize the scientific contribution of the data producers by following normal standards of scientific etiquette and fair use of unpublished data. Such guidelines can be found in "Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility" (http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf), and the Toronto Data Release Workshop (Nature 461, 168-170 (10 September 2009) | doi: 10.1038/461168a; Published online 9 September 2009).

Data release plans for NIAID-funded Omics Centers should be based on the guiding principle that genomic and other omic data types should be released to the scientific community as rapidly as possible via deposition into a searchable, publicly accessible, international database to facilitate use of the data by the scientific community. Therefore, it is anticipated that genomic sequence data and other data generated at the NIAID Omics Centers (including NIAID GSCID) will be made freely available, as rapidly as possible, via deposition into publicly accessible and

searchable international databases and to the NIAID funded Bioinformatics Resource Center or other web sites approved by NIAID. Data release plans should also describe, when appropriate, plans for release of project data and analytical results to the GSCID's web sites. GSCIDs are encouraged to coordinate these activities between their collaborators and other NIAID-funded Omics Centers.

Specific Data Release Guidelines:

Sequence Data:

All raw genome, expression-generated sequence data, and metagenome sequence (which includes 16S rDNA sequence) data and next generation sequencing data will be submitted as rapidly as possible (e.g., on a *weekly* basis) to either the Trace Archive or, as appropriate, to the Short Read Archive at NCBI/NLM/NIH. These data should also include information on templates, vectors, and quality values for each sequence, as appropriate.

Clinical data and other metadata

NIAID also expects that relevant metadata (clinical data or any other type of data) that are essential for the biological interpretation of genome sequence data will be made available to the scientific community as rapidly as possible through a publicly accessible database such as the NIAID Bioinformatics Resource Center, or other databases designated by NIAID such as NCBI's dbGAP. Metadata will provide a detailed phenotype of the strain enabling a multitude of population genetic, epidemiological, disease association, and other studies. Notably, much of the metadata that will be associated with genome sequencing projects could be analyzed independent of the genome sequence. NIAID anticipates that the release of metadata associated with the genomic data will necessitate continued discussion between the collaborators, GSCID, and other relevant parties. It is expected that a data release plan for both the genomic and metadata be defined in the project plan for each project prior to the initiation of data generation. A final plan including: i) a list of metadata to be released, ii) the database(s) they will be released to, and iii) timelines of data release will be agreed upon by NIAID, GSCID and the collaborators.

It is anticipated that metadata will be submitted by the participating investigators/collaborators to the GSCID at that time of sample submission and the GSCID in coordination with collaborators will submit these data to NIAID Bioinformatics Resource Center or other databases designated by NIAID when the genomic sequencing data are submitted to GenBank. Because many of the metadata could be analyzed independent of the genome sequences, NIAID will consider up to a 9 month embargoed release of all or a subset of the metadata with the embargo period starting after the genomic data is released. The metadata embargo will be managed by the NIAID Bioinformatics Resource Centers or other designated databases such as dbGAP.

Release of patient/donor identifying data: The rights and privacy of human subjects who participate in clinical research studies shall be protected at all times. It is recognized that a genomic, other data sets, or a subset of the clinical and other metadata may be potentially identifying of the donor and should be deposited in a controlled access database as designated by NIAID such as the NIAID Bioinformatics Resource Center or another controlled access database, or in rare cases, not made available in controlled access database.

NIAID has funded Bioinformatics Resource Centers (BRCs) that collect, make publicly available and provide analysis tools for a variety of data types related to microbial pathogens and human infectious diseases. These Centers will assist the NIAID omics Centers such as the GSCID to store, visualize and manage genomic data and clinical and other metadata. The NIAID funded

Bioinformatics Resource Centers will also provide a controlled access database to NIAID funded omics Centers, if needed. As part of the NIAID GSCID data release guidelines, it is expected that Centers will collaborate with the NIAID funded Bioinformatics Resource Centers to facilitate the standardization, transfer and dissemination of omics data sets as well as clinical and other metadata from the Center and their collaborators to the broad scientific community.

Genome and Metagenome Assembly and Annotation:
Genome and metagenomic full and partial assemblies and their annotations should be deposited in GenBank at NCBI after verification by the center. Assuming no specific errors are detected during the validation process, assemblies and annotations will be submitted to GenBank as rapidly as possible and no later than 45 calendar days of being generated, followed by release to other web sites, as approved by NIAID.

The NIAID Omics Center as the GSCID is expected to prepare GenBank records with genome assemblies and annotation that contain language to acknowledge the funding source and the joint ownership of the Genbank records by the NIAID funded GSCID and the Bioinformatics Resource Centers or database reviewed and approved by NIAID.

NIAID recommends the following language to be added to the COMMENT field of Genbank records:
"This work was supported by the National Institute of Allergy and Infectious Diseases (NIAID), Genome Sequencing Centers for Infectious Diseases (GSCID) program."

All GenBank submissions must be submitted with co-ownership by the GSCID and BRC, or ownership can be assigned to a designated group with the approval of the NIAID Program Officer.

SNP Data:
Single nucleotide polymorphisms (SNP) should be submitted as rapidly as possible to NCBI dbSNP and not later than 45 days from validation. Non-identifying clinical and other metadata should follow the release guidelines above.

Genome Wide Association Studies Data (GWAS):
Data generated from human genome wide association studies should be submitted as rapidly as possible to NIH dbGAP following the NIH policy on GWAS data deposition.
http://grants.nih.gov/grants/gwas/

Release of Other Data
Other data types not specifically addressed above including expression data, immunological data, proteomic data, and other omics data, including unpublished data, from Centers must be rapidly deposited into a publicly accessible web site(s) to include the appropriate NIAID BRC or another site designated by the NIAID. In some cases, NIAID will consider minimal delay in data release of other data types. Delayed data release for these other types of data should be discussed in the data release plan submitted to NIAID as part of Project Plan.

Analysis performed by the Centers should be made available to the public upon acceptance of a manuscript for publication. The data release plan should discuss public accessibility of the analysis data and site where such data will be housed. It is anticipated that the NIAID Bioinformatics Resource Centers should house these data sets and should be discussed in the data release plans as part of the Project Plan.

Sharing of Reagents and Other Resources:
Reagents, such as microbial strains to be sequenced, should be deposited at the Biodefense & Emerging Infections Research Resources Repository (http://www.beiresources.org/Default.aspx?base) or other approved public repositories before the strain is sequenced by the GSCID. It is anticipated that the collaborator providing the strain to the GSCID will contact and submit deposition forms to BEI prior to sequencing of the strain by the GSCID. Although all viral strains sequenced need not be submitted to BEI, it is expected that strains representing key lineages will be deposited. The strategy and criteria for selection these strains must be outlined in the Project Plan. Other resources and reagents to be shared should be released rapidly to promote the principles expressed above and documented in the data and reagent and resources release plan.

Data Release Plan Preparation

- A data release plan for each NIAID-funded omics project from the Center is required, and a final plan will be negotiated between NIAID, the GSCID and collaborators to ensure that genomics data and associated data releases will support the guiding principles stated above.
- The data release plan will include specification of the data types and datasets that will be released, including chromatograph files, raw genome and metagenome sequence, next generation sequence data, other metadata, genome assemblies, annotation, SNP and GWAS data sets, microarray data sets, etc. The specifications should also include the target database(s), data ownership and the submission format to the pre-defined repository.
- It is expected that NIAID data release guidelines will be made available to any potential collaborator prior to the submission of a white paper, if possible, and posted on Center's web sites. A data release plan will be prepared and sent to NIAID as part of the Project Plan (organization and management plan) for each individual Center's project.
- Final approval for the data release plan will be given by NIAID prior to beginning of the project.

The plan shall detail:

1. What data types, reagents and resources will be generated under this collaboration and shared with scientific community,
2. An estimated timeline for release of each data set generated,
3. A detailed list (excel) of the clinical and other metadata fields collected with the samples to be sequenced or genotyped and proposed timeline for sharing clinical and other meta data with scientific community based on the data release guidelines.
4. Site(s) where the data will be released: NIAID funded Bioinformatics Resource Centers, GenBank or Center's web sites and other database sites.