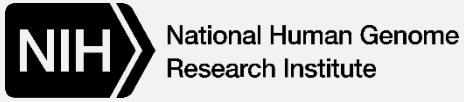
Themes from Journal Participants' Responses

Chris Gunter, Ph.D.

Senior Advisor to the NHGRI Director on Genomics Engagement

Head, Engagement Methods Unit

November 30, 2020



The Forefront of Genomics®

NHGRI posed the following questions:

1. What are challenges in the genomic data sharing arena that you are currently facing? What do you worry about or expect might come up related to genomic data sharing?
2. Are there examples of successes? What actions led to these successes?
3. Could you share with us an example of when things did not go smoothly? We are looking for ~4-5 case studies of issues that were handled well or for which there isn't a solution.
4. Are there ways that funders and journals can work together to improve the "FAIRness" of genomic data sharing?
5. Please share a link to your current genomic data sharing policies, if possible.

1. Challenges and Worries

Consistency:

- Country-specific barriers
- Consistent review across editors
- Private sector wishes to determine access
- “Rich get richer,” where only privileged can access data

Delays / Tracking Data Sharing:

- Submission challenges can delay establishment of an accession number
- Controlled-access datasets can not be evaluated in-depth by reviewers
- Authors may have to go back and redo analyses to remove data that don’t meet journal/repository standards

Differing or Missing Standards:

- Differing standards for human datasets
- Lack of standards for sharing code and/or reproducible pipelines for data analyses
- “Asks” of reviewers are too big, which can allow manipulation or fraud to go undetected
- Linking genotype and clinical phenotype data increases risk of patient identifiability

Enforcement:

- Editorial staff may have to go multiple rounds with authors to get them to make data available
- Changes in repositories over time: non-human genomic data previously deposited/publicly available in dbVar now harder to find/obtain.

Lack of Information About Permission to Share:

- Particularly in the clinical space, authors are sometimes unsure about whether samples were consented for broad data sharing
- Unclear whether sharing summary statistics of an unpublished dataset is sufficient if the author is not the primary steward of the dataset

2. Successes / Best Practices

Requiring that authors use accepted **standards** for variant description

Jan Higgins, *Genetics in Medicine*

- Pilot project at two journals determined that requiring authors to verify that variants comply with the Human Genome Variation Society standards is a reasonable first step towards standardizing the worldwide inventory of human variation.
- *Human Mutation* publication forthcoming.

Reluctance from authors to share data can be overcome by **emphasizing reasons** to share data

Rabia Begum, *Genome Medicine*

- Reiterate the importance of reproducibility of the conclusions drawn and for their work to be of interest.
- Emphasize that data can be de-identified, and explain how they may do so (e.g., categorize into ranges, etc.).
- Where appropriate, suggest controlled-access as a possibility for sharing data, if consented accordingly.

3. Case studies

- Working with Chinese authors – challenging to keep country-specific policies straight
- Use of data without consent for data sharing – authors often realize the requirement to share data, and the lack of informed consent for such sharing, at the end of the research endeavor
- Analyses based on proprietary, commercial datasets (e.g., Ancestry.com) require exceptions to Journals' data sharing policies (only summary statistics).

4. Ideas for Funder / Journal Collaboration

Funders:

- Data sharing and management plans should be required at the outset of a study, to save time at the publication stage
- Support for upload, download, and analysis at reasonable cost for less privileged scientists (while preserving privacy and security)
- Input on appropriate balance of sharing “raw” and “processed” data
- Infrastructure and incentives for peer review of deposited datasets

Journals:

- Have a robust, and enforced, data sharing policy
- Encourage data citations to support findability of the data
- Continuously improve submission systems to make data and paper co-submission easier for authors, and checking easier for editors & staff
- Update guides to authors as new policies come out
- Ongoing discussion with community on reasonable analysis/reanalysis expectations for peer reviewers

Both:

- Endorse genomics standards, recommended file types, best practices for reporting computational pipelines, and nomenclature (and support for evolution over time)
- Support an international code of conduct for genomic data sharing (*Nature* paper – https://doi.org/10.1038/d41586-020-00082-9)
- Outline a set of standard, FAIR, supported repositories dependent on the data type and restrictions
- Researcher education on FAIR data sharing, inc. other fields

Discussion

Chris Gunter, Ph.D.

Senior Advisor to the NHGRI Director on Genomics Engagement

Head, Engagement Methods Unit

Veronique Kiermer, Ph.D.

Chief Scientific Officer, PLOS

Board Member, ORCID

Board Member, Keystone Symposia

Topic 1: Data Standards for Metadata and Code

Topic 2: Challenges with International Data Sharing

Topic 3: Improving FAIRness (Education and Repositories)

Topic 4: OPEN DISCUSSION

NIH
NHGRI

Topic 1: Data Standards for Metadata and Code

Topic 1: Data Standards for Metadata and Code

Topic 2: Challenges with International Data Sharing

Topic 1: Data Standards for Metadata and Code

Topic 2: Challenges with International Data Sharing

Topic 3: Improving FAIRness (Education and Repositories)

Topic 1: Data Standards for Metadata and Code

Topic 2: Challenges with International Data Sharing

Topic 3: Improving FAIRness (Education and Repositories)

Topic 4: OPEN

Journal Data Policies Ensure Data are in the Public Domain

Open Data & Software

- "All data […] **must be made publicly available**."
- "Data **should be in a public repository or database managed by a third-party**."
- **"**We **require deposit of the underlying sequence data necessary to call variants**. We further require the deposit of mass spectrometry-based proteomics data, as well as individual-level data including phenotypic data and genotypes, **under controlled access, if necessary."**
- "Additionally, **any new software reported in a manuscript must be made publicly available**."
- **"Must share the "minimal data set"** for their submission […] the minimal data set to consist of **the data required to replicate all study findings reported in the article, as well as related metadata and methods**."

Availability Statements

- "[A]uthors must include a **statement on reagent, software, and data availability."**
- **"**Accession numbers for data should be included within the manuscript **prior to acceptance to avoid delays in publication**."
- "[We] require authors to make all data necessary to replicate their study's findings publicly available without restriction **at the time of publication**."

Standardization

- "For studies that include genotype and DNA sequences, the genotypes or sequences for all individuals **should be provided in common formats** […]and raw sequences reads should be deposited in a public repository"
- **"Encourage authors to follow FAIR Data Principles."**
- "[We require] **compliance with the Human Genome Variation Society (HGVS) recommendations for describing sequence variants** before manuscripts can be accepted and published."
- "[…] deposited in publicly available repositories […] or presented in the main manuscript or additional supporting files, **in machine-readable format whenever possible.**"

NIH NHGRI


The Forefront
of Genomics®