Genetics Society of America

9650 Rockville Pike, Bethesda, MD 20814-3998
Phone: (301) 634-7300 • Fax: (301) 634-7079
society@genetics-gsa.org
www.genetics-gsa.org

**Genetics Society of America Response to Request for Information**
**Input on Sustaining Biomedical Data Repositories (NOT-ES-15-011)**

*March 2015*

The Genetics Society of America (GSA) is pleased to submit this response on the economic, technical, policy, and administrative approaches toward enhancing long-term sustainability of biomedical data repositories. This topic is of significant interest to the GSA because our members depend on such resources for their research, which increasingly generates, uses and accesses large amounts of data.

Among our more than 5,000 members are many researchers who work with a number of model organisms that extend from *Arabidopsis* to zebrafish and who utilize key community resources—including biomedical data repositories—that provide a central locus for information and data relevant to their system of study.

## Importance of Model Organism Data Repositories

The GSA community depends upon a number of model organism data repositories that not only help provide central access to research results and interactions, but also enable the community to use the same research infrastructure, helping ensure consistency and reproducibility among different laboratories. These resources are also often interconnected with each other—and with other related databases—helping facilitate interactions across experimental systems.

The importance of these model organism data repositories cannot be overstated. Members of the research community often make daily use of these resources, which serve as the locus for those working with that model system around the world.

The long-term and consistent support for these model organism databases—as well as organismal stock centers—has been a crucial component of the strength and success of biomedical research in the United States and will be necessary for its future vigor. Centralized stock centers and databases provide optimal resource sharing that maximizes the return on the investments made by the NIH and other government agencies. These community resources provide "off-the-shelf" research tools and thus increase the efficiency and speed of hypothesis-driven research supported by other grants. In addition, public support for these community resources allows them to operate on an open access model, thus assuring that all researchers have the tools they need for discovery. These databases

serve to preserve data well beyond the length of the original grant in a format that makes it easily accessed by other researchers.

The alternative to community data repositories would be for laboratories to store their own data in isolated servers without an expectation of centralization, interoperability, or consistency. This alternative would dramatically limit the utility of the data for others and, therefore, reduce the return on investments in research, since data would remain inaccessible to the research community at large.

**Technology**

Model organism data repositories such as FlyBase, the *Saccharomyces* Genome Database, and WormBase—for the *Drosophila*, yeast, and *C. elegans* communities, respectively—have pioneered innovations that increase the utility of such resources and the effectiveness of bioinformatics tools for advancing science more generally.

For example, many of these databases use advanced text-mining technology to assist in the efficient indexing and identification of documents and to associate published papers with data elements in the repositories. This technology helps increase the utility of the databases and better connects the peer-reviewed literature with the underlying data. Providing easy and open access to original data is a key focus of the NIH and the research community—and data repositories such as those that support model organisms have proven to be an effective mechanism for doing so.

Such resources also enhance the ability of the community to improve annotation of stored data, thus helping maximize the utility and discovery of key results and datapoints. Although we stress the crucial role that professional biocurators play in validating and annotating data in these repositories—and developing ontologies which help define relationships between data—model organism databases are increasingly calling upon the expertise of those who work with this information daily to further increase the completeness of these data resources.

One may wonder if the effort spent to curate data is an efficient use of resources, but we stress that these investments will pay off over many years. A few minutes spent by a professional biocurator when the data are originally deposited may save the time of tens or hundreds of researchers over the next decade or more.

We believe there may also be opportunities for the NIH to work with other partners to support the development of a common framework or platform that might increase the overall efficiency of such model organism databases over time by enabling them to share aspects of technology development.

**Financial Models**

GSA appreciates the challenge of sustaining biomedical data repositories, but does not believe that all such resources should be self-sustaining. We feel that model organism databases are essential to the overall research enterprise as they serve the collective needs of the community. As such, they perform a public good for the community that is significantly greater than the sum of the value to individual investigators.

It may be tempting to suggest that costs should be borne by the most active researchers. However, we point to the experience with The Arabidopsis Information Resource (TAIR), which had been supported by the National Science Foundation but has more recently been forced to move toward a self-sustaining model. Academic and commercial entities are now asked to pay an annual subscription—often many thousands of dollars per year—for their investigators to have access. If an institution does not maintain its subscription, its researchers may not have access to the most current and validated information, thus limiting the quality of research conducted. One troubling aspect from the change in the financial model for TAIR has been a splintering of data resources; while the community had previously enjoyed a central, publicly supported data repository and integrated stock center, there has recently been a proliferation of several competing resources that are not fully integrated or even consistent.

We fear that a similar fate may befall other model organism communities if federal agencies curtail or limit public support with the expectation of self-sustainability.

The GSA cautions that charging users for accessing data would have negative consequences on the exchange of information. Even a small fee would be a disincentive for accessing validated and current data. It would also have a negative impact on the use of such resources by more casual users or those who lack significant funds, such as researchers at under-resourced institutions and those using such data repositories for educational purposes.

**Partnerships**

Although the NIH has a crucial role to play in supporting and sustaining biomedical data repositories, it is not the only important stakeholder. Indeed, we urge the NIH to work closely with other funders—including other government agencies, private sponsors, and international partners—as well as organizations that play a critical role in fostering the success of the research community—such as publishers, scientific societies, and others.

**Evaluation**

GSA recognizes that it is not possible to support all biomedical data repositories indefinitely, and that it will be necessary to evaluate which resources have the greatest need for sustained funding models. Making these choices will require an open and transparent evaluation process for assessing the value and utility of each resource.

We urge the NIH to work with the research community and other partners to develop consistent standards and metrics that will assist in performing these evaluations. We stress that these metrics should not be limited to the size of the userbase and the frequency of access, but also consider the importance and uniqueness of the resource and the role that it plays in advancing research.

In the event that there is a collective decision to curtail future funding for a resource, we urge the NIH to work with its partners and the data repository to maintain access to the original data even if further development will end.

On behalf of the Genetics Society of America and our more than 5,000 members throughout the United States and around the world, thank you again for the opportunity to provide input into your conversations. We look forward to working closely with you and other partners in continuing the discussion of these important issues and helping ensure sustainable, efficient, and effective access to biomedical data.

Genetics Society of America

**ABOUT GSA**: Founded in 1931, the Genetics Society of America (GSA) is a professional scientific society with more than 5,000 members worldwide working to deepen our understanding of the living world by advancing the field of genetics, from the molecular to the population level. GSA promotes research and fosters communication through a number of GSA-sponsored conferences including regular meetings that focus on particular model organisms. GSA publishes two peer-edited scholarly journals: *GENETICS*, which has published high quality original research across the breadth of the field since 1916, and *G3: Genes|Genomes|Genetics*, an open-access journal launched in 2011 to disseminate high quality foundational research in genetics and genomics. The Society also has a deep commitment to education and fostering the next generation of scholars in the field. For more information about GSA, please visit www.genetics-gsa.org. Also follow GSA on Facebook at facebook.com/GeneticsGSA and on Twitter @GeneticsGSA.